Molecular Diversity
https://doi.org/10.1007/s11030-025-11352-x

ORIGINAL ARTICLE

ACP-EPC: an interpretable deep learning framework for anticancer peptide prediction utilizing pre-trained protein language model and multi-view feature extracting strategy

Jingwei Lv[1] · Kexin Li[1] · Yike Wang[1] · Junlin Xu[2] · Yajie Meng[3] · Feifei Cui[1] · Leyi Wei[4] · Qingchen Zhang[1] · Zilong Zhang[1]

Received: 13 July 2025 / Accepted: 1 September 2025

Abstract
Cancer remains a major global health challenge, as conventional chemotherapy often causes extensive damage to healthy cells and leads to severe side effects. Anticancer peptides (ACPs) have emerged as a promising therapeutic alternative, capable of selectively targeting and eliminating cancer cells while improving patient quality of life and treatment outcomes. Nevertheless, identifying ACPs through traditional biological experiments is both labor-intensive and time-consuming. To address this limitation, we developed ACP-EPC, a deep learning framework which predicts ACPs directly from protein sequences. ACP-EPC integrates contextual representations from Evolutionary Scale Modeling 2 (ESM-2) with handcrafted physicochemical descriptors and employs a Cross-Attention mechanism for multimodal feature fusion. The model was rigorously evaluated using tenfold cross-validation and two test sets, ACP135 and ACP99, achieving accuracy of 0.935 and 0.984, respectively. These results substantially outperform existing models, underscoring the advantages of combining diverse feature representations. To promote accessibility, we have also deployed ACP-EPC as a publicly available web server at http://www.bioai-lab.com/ACP-EPC.

Keywords Anticancer peptide · Cross-Attention · Multimodal feature fusion · Classification · Deep learning

Introduction

Cancer remains a formidable challenge in modern medicine, with its numerous types and high propensity for metastasis rendering it a malignant disease that humanity has yet to fully conquer [1]. Despite significant progress, a universal, effective, and well-tolerated cancer therapy remains elusive. Current treatment modalities—including chemotherapy, radiotherapy, surgery and targeted therapy—each have inherent limitations [2]. Chemotherapy and radiotherapy impose substantial burdens on the body and are often accompanied by severe side effects such as alopecia and nausea. Surgical interventions typically fail to eliminate cancer cells that have metastasized to distant sites; although the primary tumor can be excised, residual solitary cells in surrounding tissues may remain undetected. Furthermore, targeted therapies are expensive and only effective against specific types of cancer. In recent years, ACPs have attracted considerable attention from researchers. ACPs are widely distributed across in mammals, amphibians, insects, plants, and microorganisms and can also be synthetically produced [3]. They interact with the phospholipid bilayer of cancer cell membranes, altering membrane permeability, which leads to cellular content leakage and ultimately cell death [4]. ACPs offer several advantages as anticancer agents, including low molecular weight, simple structures, high anticancer activity, and strong selectivity. Moreover, they produce fewer side effects, can be administered via multiple routes, and are less likely to induce multidrug resistance. However,

✉ Zilong Zhang
zhangzilong@hainanu.edu.cn

1 School of Computer Science and Technology, Hainan University, Haikou 570228, China

2 School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, Hubei, China

3 School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, Hubei, China

4 Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao SAR, China

Published online: 13 September 2025

identifying ACPs from large peptide libraries typically relies on conventional approaches such as in vitro cell assays or animal experiments [5]. These methods are time-consuming, require meticulously designed experimental protocols, appropriate control group selection, and substantial financial support [6]. Additionally, animal experiments are increasingly regarded as ethically problematic due to concerns about cruelty. Consequently, there is an urgent need for rapid and accurate methods to identify ACPs. Such approaches would enable researchers to efficiently exclude non-ACP candidates, thereby reducing experimental scale, time, and financial cost.

The rapid advancement of protein sequencing technology now allows researchers to obtain high-precision sequence data swiftly [7–11]. Simultaneously, developments in machine learning and deep learning have provided robust support for creating ACP prediction models [12, 13]. To date, numerous models for ACP prediction have been proposed [14]. For instance, in 2018, Wei et al. [15] introduced the ACPred-FL model, which relied on sequence feature descriptors and a two-step feature selection method to extract a five-dimensional feature vector rich in information, using a support vector machine (SVM) for ACP sequence prediction. In 2019, Schaduangrat et al. [16] developed a model combining sequence composition information with physicochemical properties, employing SVM and random forest (RF) classifiers and analyzing the effects of different feature combinations on predictive performance. In the same year, Boopathi et al. [17] proposed the mACPpred model, which extracted features based on seven types of sequence composition and physicochemical properties, applied a two-step feature selection to reduce dimensionality and used SVM for prediction; their results demonstrated that feature sets refined through two-step feature selection outperformed single features in predictive performance. Additionally, Wei et al. [18] developed the PEPred-Suite model, combining commonly used features with mRMR [19] feature selection and sequential forward search, selecting the highest-scoring features for input and using RF as the base classifier to predict eight different peptide types. Yi et al. [20] were the first to apply deep learning to ACP prediction, using two manually extracted, high-efficiency features to achieve strong results. In 2020, Ge et al. [21] integrated multiple feature types, trained them using LightGBM on various feature sets and employed SVM for the final predictions, resulting in the EnACP model. This approach, combining ensemble learning classifiers with traditional methods, significantly improved prediction accuracy. In 2021, Agrawal et al. [22] proposed the AntiCP 2.0 method, integrating SVM, artificial neural networks (ANN), RF, and Extra-Tree (ET) classifiers; their findings indicated that ET-based ensemble learning achieved the best performance in ACP prediction. Chen et al. [23] introduced the ACP-DA model, which combined ACP-DL's feature extraction techniques with AAindex features, used a multilayer perceptron (MLP) for prediction, and employed data augmentation techniques to address the challenge of limited training samples. That same year, Huang et al. [24] developed the iDACP model, exploring differences in the physicochemical properties of ACPs and proposing a two-step approach: identifying ACPs in the first step and classifying subtypes in the second, providing a foundation for studying ACP mechanisms. Ahmed et al. [25] proposed the ACP-MHCNN model, integrating features from multiple sources into a convolutional neural network (CNN), achieving strong generalization capabilities on independent test datasets. In 2022, Han et al. [26] developed the ACPred-BMF model, encoding peptide sequences using amino-acid one-dimensional encodings and physicochemical properties, employing bidirectional long short-term memory (BiLSTM) with an attention mechanism for prediction and using SHapley Additive exPlanations (SHAP) [27] analysis to interpret feature importance. Wu et al. [28] proposed the ACP-MCAM model, which used convolutional layers with various kernel sizes for feature extraction and combined them with position embeddings and attention mechanisms to enhance representation capabilities. Phan et al. [29] developed the MLACP 2.0 model, integrating seven classifiers to construct 119 baseline models and selecting 67 generated feature vectors as neural network inputs, significantly improving predictive performance. In 2024, Bian et al. [30] introduced the ACP-ML model, exploring the predictive capabilities of different feature combinations. Using Maximum Relevance Minimum Distance (MRMD) [31] and Recursive Feature Elimination (RFE) [32] feature selection methods, they obtained a 183-dimensional optimal feature subset. Subsequently, they evaluated nine machine-learning models via 10-fold cross-validation and selected the six best performers as baseline classifiers to build a voting-based ensemble for precise ACP identification. Despite the various methods employed to build ACP identification models, many existing approaches—even those utilizing deep learning—have limitations in feature engineering. These limitations include extracting single-dimensional features, producing discrete representations, or lacking interpretability. To address these challenges, we developed a novel ACP prediction model, ACP-EPC.

Our study presents a novel anticancer peptide prediction framework, termed ACP-EPC. The acronym ACP denotes the model's focus on anticancer peptide prediction, 'E' denotes the incorporation of the ESM-2 model [33], 'P' represents physicochemical properties, and 'C' indicates the application of Cross-Attention. The framework integrates physicochemical features extracted through traditional methods with structural and evolutionary features derived from ESM-2. Traditional discrete features are integrated with ESM-2 extracted features through Cross-Attention, capturing

the physicochemical information embedded within the structural data. Evaluated through tenfold cross-validation and two test sets (ACP135 and ACP99), our model demonstrates substantial improvements over previous approaches, particularly in ACC, MCC, and Sp. On the ACP135 dataset, ACP-EPC achieved improvements of 2.6% and 6.5% in ACC and MCC, respectively, compared to the latest existing model. On the ACP99 dataset, our model outperformed the best-performing ACP prediction model by 5.8%, 4.3%, 6.9%, and 15% in ACC, Sn, Sp, and MCC, respectively. The model, source code, and datasets are publicly available at https://github.com/EuclidLv/ACP-EPC, and a user-friendly web server has been deployed at http://www.bioai-lab.com/ACP-EPC to facilitate community access.

Materials and methods

Overall framework of ACP-EPC

As shown in Fig. 1, the workflow comprises three stages: (A) dataset collection and assembly, (B) model development, and (C) web-server deployment. For (A), this study employed training and test sets curated from prior studies. In (B), the model adopts a two-stream multimodal design. Each sequence is encoded with ESM-2 to obtain contextual embeddings and simultaneously represented by traditional descriptors, including amino-acid composition (AAC), dipeptide composition (DPC), and physicochemical properties (PCP). The PCP set includes hydrophobicity, polarity, isoelectric point (pI), molecular weight, bulkiness, charge, average flexibility, α-helix propensity, and β-sheet propensity. Traditional descriptors are linearly projected to match the dimensionality of the ESM-2 stream and serve as the Query (Q), while ESM-2 embeddings provide the Key/Value (K/V). A Cross-Attention module integrates the two streams, and the fused representation is subsequently classified by an MLP. To improve efficiency and reproducibility, we pre-extract and cache both ESM-2 embeddings and traditional descriptors in RAM prior to training, thereby avoiding redundant feature computation at each epoch. In (C), the trained model is deployed via a web interface to enable convenient, interactive use by researchers. The detailed hyperparameters of the model are shown in Table S1, and the config running environment running time is shown in Table S2 and Figure S1.

Dataset

To ensure a fair comparison with previous models, we employed the same training set and two independent test sets, ACP135 and ACP99. The training set and ACP135 were originally curated by Bian et al. [30], who compiled 1,350 experimentally validated ACPs from the CancerPPD [34], APD3 [35], and SATPdb [36] databases. To mitigate homology bias and prevent artificially inflated recognition accuracy, CD-HIT [37] was utilized with a 0.9 similarity threshold to remove sequences sharing more than 90% identity. Subsequently, Seqkit [38] was employed to extract sequences ranging from 5 to 50 amino acids in length, retaining only those for which Position-Specific Scoring Matrix (PSSM) [39] profiles could be generated using the PSI-BLAST [40]. This filtration process yielded 622 ACPs. For negative samples, Bian et al. [30] randomly selected peptides and applied the same filtering procedure, yielding 1839 non-ACPs. From these, 487 ACPs and 1,479 non-ACPs were randomly selected to construct the training dataset.

ACP135 comprises 135 ACPs and 360 non-ACPs, while the training set includes 487 ACPs and 1479 non-ACPs. ACP194, developed by Agrawal et al. [22], contains 388 sequences, including 194 ACPs and 194 non-ACPs. Bian et al. [30] mitigated homology bias and applied PSSM, yielding the ACP99 dataset with 99 ACPs and 157 non-ACPs. Because ACP99 is more stringently processed and recently curated and to ensure comparability with Bian et al. [23], we adopt ACP99 in this study.

Detailed characteristics of the datasets are provided in Table 1. Figure 2 summarizes the amino-acid composition and peptide length distributions for the training set, ACP135 and ACP99. In Fig. 2A, ACP135 closely mirrors the training set across most residues, whereas ACP99 shows modest shifts in several amino acids, indicating slight compositional differences relative to the other two cohorts. Figure 2B shows that peptide lengths in all three datasets are concentrated around 25–30 residues. ACP99 exhibits the narrowest and most concentrated distribution, the training set shows moderate dispersion, and ACP135 displays the greatest variability, with a long tail of extended sequences. These patterns suggest that ACP135 is compositionally similar to the training set but encompasses greater length diversity, while ACP99 is more tightly constrained in sequence length.

Feature encoding

Physicochemical property-based traditional feature extraction

Physicochemical motivation. Feature set, and physicochemical meaning Optimized matching hydrophobicity (OMH) [41, 42] captures the role of hydrophobicity in driving peptide partitioning into the lipid phase. Together with charge segregation (hydrophobic moment), this property enables amphipathic helices to insert into membranes or form pores. Fine-tuning hydrophobicity balances on-target membranolytic activity with off-target hemolysis. Molecular weight [43] reflects peptide size; ACPs are generally

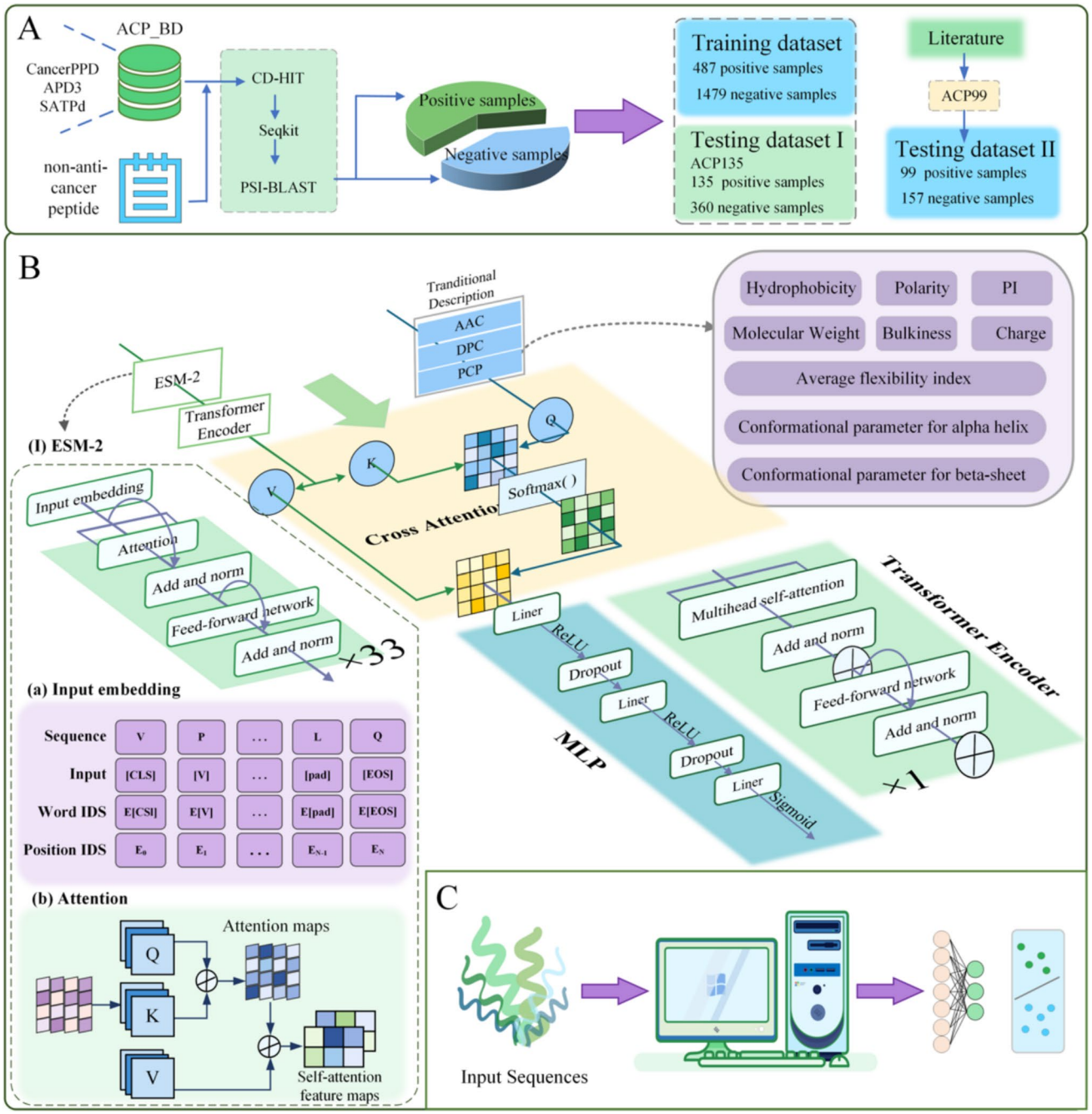

Fig. 1 Overview of ACP-EPC. **A** Dataset construction. ACPs were collected from multiple public databases, while non-ACPs were used as negative controls. Redundancy was reduced using CD-HIT and PSI-BLAST clustering, resulting in a training dataset of 487ACPs and 1479 non-ACPs. Two test sets were employed: ACP135 (ACPs and 360 non-ACPs) and ACP99 (ACPs and 157 non-ACPs from literature), ensuring robust evaluation. **B** Model architecture. The ACP-EPC framework integrates embeddings from ESM-2 with handcrafted physicochemical descriptors. ESM-2 encodes sequences through tokenization, positional embeddings, and one Transformer encoder layers, producing contextual residue-level representations. Cross-attention modules fuse these embeddings with traditional descriptors, including AAC, DPC, and PCP. The fused representations are processed by a ML, followed by a sigmoid classifier for final prediction. **C** Web server deployment. The trained model was implemented as a web server, enabling users to input peptide sequences and obtain ACP prediction results along with interpretability outputs such as attention maps, facilitating accessibility to the research community

short sequences, with most reported examples ranging from 5 to 40 amino acids, though some extend to 60. Size influences diffusion, cellular uptake, protease susceptibility, and pore-forming geometry; notably, many designed short helices achieve selective cancer cell killing. Polarity, quantified by the Grantham scale [44], modulates side-chain interac-

Table 1 Training and testing datasets for this research

Dataset	Negative samples	Positive samples
Training dataset	1479	487
ACP135	360	135
ACP99	157	99

tions at the membrane interface and determines the balance between aqueous solubility and bilayer affinity. Appropriate polarity distribution across a helical face supports amphipathicity and enhances selectivity. PI [45] correlates with net cationic charge at physiological pH, promoting electrostatic capture by the anionic outer leaflet of tumor cells (e.g., exposed phosphatidylserine and glycosaminoglycans), a key basis of ACP selectivity. The average flexibility index (Bhaskaran–Ponnuswamy) [46] reflects conformational flexibility, which facilitates membrane adaptation and insertion and, in some families, enhances potency and selectivity. Gly and proline residues can introduce adaptive bends that promote membrane activity. The conformational parameter for α-helix (helix propensity) [47] reflects the likelihood of peptides forming amphipathic α-helices upon membrane contact; higher helix propensity supports stable interfacial binding, insertion, and pore formation. The conformational parameter for β-sheet (β-sheet propensity) [48] reflects the tendency of peptides to adopt β-sheet structures. A subset of ACPs, such as defensins, are β-sheet-rich, disulfide-stabilized peptides whose segregated polar and hydrophobic faces facilitate membrane disruption or lipid-specific binding. Bulkiness, defined by Zimmerman [49], describes side-chain steric volume. Aromatic residues (Trp, Phe, Tyr) enhance interfacial anchoring and stabilize peptide–bilayer contacts. While bulky hydrophobic residues strengthen partitioning, excessive bulk may increase nonspecific toxicity. Charge [50] is another defining feature of ACPs. Positive residues (Lys, Arg, and partially protonated His) enable electrostatic attraction to negatively charged cancer membranes (e.g., phosphatidylserine exposure, sulfated proteoglycans). Optimal charge, in concert with hydrophobicity, tunes selectivity. The detailed figures of the properties and reference are shown in Table S3.

Per-residue mapping and sequence-level aggregation Let the sequence be $r_{1:L}$ and let each property k define a mapping $s_k : \{20\text{AA}\} \to \mathbb{R}$ assigning a numerical value to each amino-acid type. We aggregate to sequence level using the length-normalized mean

$$\bar{s}_k = \frac{1}{L}\sum_{t=1}^{L} s_k(r_t) \tag{1}$$

which yields a nine-dimensional vector $\bar{\mathbf{s}} \in \mathbb{R}^9$ per sequence. Mean aggregation is length-invariant and retains the original interpretability (units and scale) of each index, aligning with recommended practices for amino-acid indices in sequence profiling.

Normalization and leakage control Because the nine aggregates differ in scale and units, we apply z-score standardization based on statistics computed from the training set,

$$\hat{S}_k = \frac{\bar{s}_k - \mu_k}{\sigma_k} \tag{2}$$

and apply (μ_k, σ_k) for validation and test to prevent information leakage. This harmonization places heterogeneous

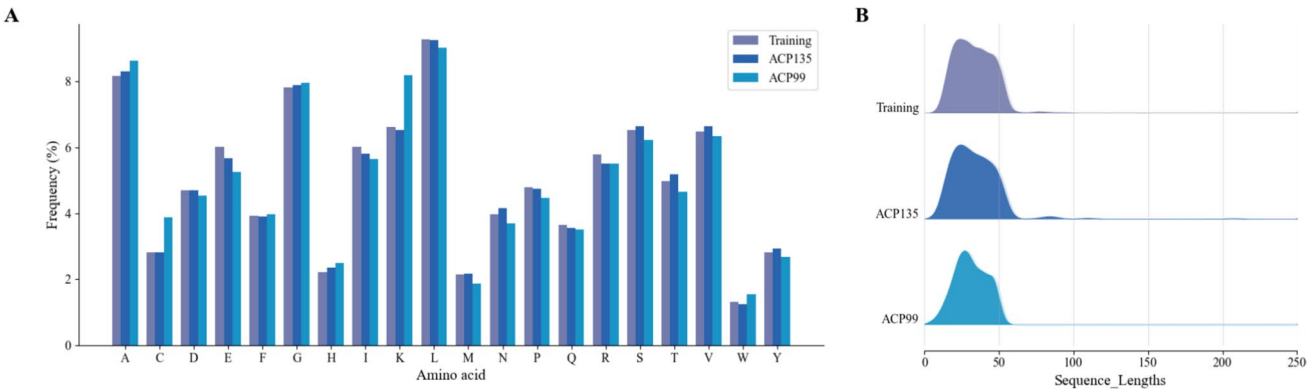

Fig. 2 Dataset profiles. **A** Amino-acid composition. The frequency distribution of the 20 standard amino acids across the training dataset and two test sets (ACP135 and ACP99) is shown. Although the three datasets share broadly consistent compositional trends, ACP135 and ACP99 exhibit greater fluctuations owing to their smaller sample sizes, with ACP99 showing more pronounced deviations in certain residues. **B** Peptide length distributions. Density plots illustrate the sequence length distributions of peptides in the training dataset, ACP135, and ACP99. While most peptides in the training set and ACP135 fall within the 15–60 residue range, the training dataset additionally displays a long-tail distribution with longer peptides. By contrast, ACP99 peptides are generally shorter and lack this long-tail pattern. These differences underscore ACP99 as a more stringent and representative benchmark for assessing model generalization

properties on a comparable footing during learning and improves optimization stability.

Complementary composition features In parallel, we compute AAC as normalized frequencies $a \in \mathbb{R}^{20}$ with $a_i = \frac{1}{L}\sum_t 1\{r_t = \mathrm{AA}_i\}$ and DPC as normalized frequencies $d \in \mathbb{R}^{400}$ over ordered residue pairs with $d_{ij} = \frac{1}{L-1}\sum_t 1\{(r_t, r_{t+1}) = (\mathrm{AA}_i, \mathrm{AA}_j)\}$. The minimal "traditional" feature vector therefore concatenates $\hat{\mathbf{s}}$ (9), AAC (20), and DPC (400) for a total of 429 dimensions.

Structural and evolutionary feature extraction using ESM-2

We adopt ESM-2 [51], a state-of-the-art protein language model trained with self-supervised learning on UniRef50 [52] as the upstream feature extractor for all sequences. ESM-2 captures rich contextual dependencies and evolutionary regularities directly from raw amino-acid sequences and has demonstrated strong transfer across diverse downstream tasks, including structure prediction, functional annotation, and binding-site identification [53, 54]. In this work, we instantiate the 650 M-parameter variant comprising 33 transformer blocks and use its final block (Layer 33) to obtain token-level embeddings with a hidden dimensionality of 1,280. These representations complement handcrafted descriptors by encoding structural and physicochemical cues that are difficult to recover from traditional features alone.

Because our task-specific dataset is comparatively small, we avoid end-to-end fine-tuning of ESM-2 to reduce overfitting risk and preserve out-of-distribution generalization. Specifically, all ESM-2 parameters are frozen, and the model is set to evaluation mode to disable dropout. Feature extraction is performed under *no_grad* to prevent gradient accumulation. This strategy stabilizes the upstream representation while allowing the downstream modules to adapt to task semantics without perturbing the pre-trained backbone.

For integration, we expose a single-entry point, *extract_esm2_features*(), which takes a list of amino-acid sequences and returns per-residue embeddings in a shape-consistent, order-preserving manner. Formally, given input $sequences = [s_1, s_2, \ldots s_N]$, the function produces $esm2\_features\_list = [E_1, E_2, \ldots, E_N]$ with $E_i \in \mathbb{R}^{L_i \times 1280}$. Special tokens introduced by the tokenizer (e.g., BOS/EOS) are excluded so that L_i matches the raw sequence length exactly. This design avoids padding at the extraction stage and yields a list of variable-length tensors suitable for per-residue downstream modules or for subsequent pooling if sequence-level features are required.

Transformer encoder layer

The Transformer encoder layer encodes input sequences into high-dimensional feature representations. It effectively captures both local and global dependencies within the data. Each encoder layer comprises two primary submodules: the multi-head self-attention mechanism and a feed-forward neural network (FFN). The multi-head self-attention mechanism enables each position in the sequence to attend to all other positions, thereby capturing long-range dependencies and contextual information. This mechanism involves computing a query vector (Q), a key vector (K), and a value vector (V) for each position. The scaled dot-product attention is then applied to compute a weighted representation, where multiple attention heads operate in parallel to capture diverse features of the input data. Mathematically, the scaled dot-product attention can be expressed as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V\,, \tag{3}$$

where d_k is the dimensionality of the key vectors. By employing multiple heads, the model can focus on different aspects of the input features, thereby enhancing its representational capacity.

Following the self-attention mechanism, the feed-forward neural network (FFN) applies position-wise, nonlinear transformations to the features. The FFN typically consists of two linear transformations separated by an activation function, such as ReLU or GELU. This structure allows the model to capture more complex feature interactions and improve its expressiveness:

$$FFN(x) = GELU\left(xW_1 + b_1\right)W_2 + b_2\,, \tag{4}$$

where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors.

In this model, a Transformer encoder is employed to refine the per-residue representations obtained from the ESM-2 protein language model before performing Cross-Attention with traditional physicochemical features. Specifically, each protein sequence is first encoded by ESM-2 into a sequence of embeddings with shape $(B, L, 1280)$, where B is the batch size and L is the padded sequence length. These embeddings are linearly projected into a shared embedding space of dimension 256 to reduce dimensionality and computational cost. The projected embeddings are then passed through a single-layer Transformer encoder ($num\_encoder\_layers = 1$) with 8 attention heads ($num\_heads = 8$), a feed-forward hidden size of 1024 ($dim\_feedforward = 1024$), and a dropout rate of 0.3. The encoder processes the sequence in batch-first mode ($batch\_first = True$) and outputs contextualized representations of the same shape $(B, L, 256)$. These refined embeddings serve as the key and value in a subsequent

Cross-Attention module, allowing the model to selectively align ESM-based structural information with query vectors derived from normalized physicochemical descriptors. This design enables effective integration of discrete attribute-based features and deep contextual representations.

Cross-attention

Cross-attention is an attention variant that explicitly separates the roles of the query and the key–value memory across two inputs. Given projected tensors $Q \in \mathbb{R}^{B\times T_q\times d}$, $K \in \mathbb{R}^{B\times T_k\times d}$ and $V \in \mathbb{R}^{B\times T_V\times d}$

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V, \tag{5}$$

where d_k is the per-head key dimension. Intuitively, Q specifies the selection signal, while K and V provide the content-addressable memory. In contrast to self-attention, which derives Q, K and V from the same sequence, Cross-Attention lets one representation actively retrieve and reweight context from another, making it well suited for multimodal fusion, multi-sequence conditioning, and aligning upstream encoders with downstream attributes.

In our model, we instantiate a single Cross-Attention block to fuse standardized physicochemical descriptors with contextual protein embeddings from ESM-2 (650 M). Let B be the batch size and L the sequence length. For each sequence, ESM-2 provides per-residue embeddings$X^{\mathrm{ESM}} \in \mathbb{R}^{B\times L\times 1280}$. The physicochemical descriptor is a single vector $x^{\mathrm{phy}} \in \mathbb{R}^{B\times 429}$, standardized feature-wise. We adopt a shared embedding width $d_{\mathrm{model}} = 256$ with $h = 8$ heads, yielding $d_k = d_{\mathrm{model}}/h = 32$ per head. The query pathway maps x^{phy} through a linear layer Linear$(429 \to 256)$ to $\widetilde{Q} \in \mathbb{R}^{B\times 256}$ which is *unsqueezed* to $Q \in \mathbb{R}^{B\times 1\times 256}$ (one query token per sequence). The key–value pathway maps ESM-2 features through *Linear*$(1280 \to 256)$ to$\widetilde{X} \in \mathbb{R}^{B\times L\times 256}$, reused as$K, V \in \mathbb{R}^{B\times L\times 256}$. Multi-head attention produces$O \in \mathbb{R}^{B\times 1\times 256}$, which we squeeze to $z \in \mathbb{R}^{B\times 256}$ as the fused representation fed to the classifier. ESM-2 remains frozen and in evaluation mode; when variable-length padding is present, we apply a key mask so attention ignores padded residues. This configuration keeps the trainable footprint confined to lightweight projection and classifier layers while preserving the full contextual capacity of the pre-trained backbone.

We assign Q to the physicochemical descriptor and K, V to ESM-2 embeddings to realize attribute-guided retrieval from a rich contextual "memory." With this design, a low-dimensional, interpretable attribute vector issues the query, and the attention weights $\alpha = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) \in \mathbb{R}^{B\times 1\times L}$ reveal which residues and local contexts are most relevant to those attributes. This improves interpretability—one can visualize, for a given property, the residue-level hotspots that the model attends to—while leveraging ESM-2's well-documented capacity to encode fine-grained structural and functional signals. Standardizing the physicochemical features before projection further stabilizes optimization by placing heterogeneous attributes on comparable scales inside the scoring term $\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}$, which typically yields crisper attention distributions and more reliable training dynamics.

From an engineering standpoint, this allocation of roles confers two practical advantages under limited data. First, it enables frozen-backbone training: the large ESM-2 encoder serves as a fixed repository, while only the compact query and mapping layers are trained. This reduces overfitting risk and computational cost, conceptually resembling lightweight querying modules that extract task-specific information from frozen encoders. Second, Cross-Attention naturally supports reading a large memory with a small query, creating an effective information bottleneck and computational down-projection reminiscent of latent-array designs that attend from a small set of queries to long inputs. Together, these properties allow us to preserve the information content of both modalities while achieving stable, interpretable, and sample-efficient fusion.

MLP

The MLP serves as the final decision module that maps the fused representation to a calibrated probability of the ACP class. Its input is a 256-dimensional context vector per peptide, obtained by querying the Transformer-refined ESM-2 token features with the handcrafted descriptor query through Cross-Attention. The MLP outputs $p(y = 1 \mid \mathbf{x}) \in [0,1]$ through a sigmoid, which we threshold at 0.5 for class labels while also using the raw probabilities for AUC and related metrics. In our model, the MLP is a three-layer perceptron: $256 \to 128 \to 64 \to 1$. Each hidden layer uses ReLU activations and dropout (rate 0.3) to improve generalization. Concretely,

$$\mathrm{h}_1 = \mathrm{ReLU}\left(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1\right), \mathrm{h}_1 = \mathrm{Dropout}\left(\mathrm{h}_1\right) \tag{6}$$

$$\mathrm{h}_2 = \mathrm{ReLU}\left(\mathrm{W}_2\mathrm{h}_1 + \mathrm{b}_2\right), \mathrm{h}_2 = \mathrm{Dropout}\left(\mathrm{h}_2\right) \tag{7}$$

$$p = \sigma(\mathrm{w}_3^{\mathrm{T}}\mathrm{h}_2 + b_3) \tag{8}$$

This design provides sufficient nonlinearity to capture interactions among fused features without introducing excessive depth or parameters that could overfit the relatively short peptide inputs.

A shallow, regularized MLP is well suited to our upstream encoder–fusion stack: since the Cross-Attention already

yields a compact, task-aware summary, the classifier mainly separates classes in latent space rather than learning new high-level abstractions. Two hidden layers (128 and 64 units) strike a balance between capacity and stability, while dropout and weight decay (used during optimization) mitigate overfitting. The sigmoid output supports probability-based evaluation and class-imbalance–aware training. However, in training, we optimize the entire network end-to-end with Adam (learning ratel $\times 10^{-4}$, weight decay, 1×10^{-5}). To handle class imbalance and hard examples, we couple the sigmoid output with focal loss($\gamma = 2.0, \alpha = 0.25$), which down-weights easy negatives and sharpens gradients on difficult samples. This pairing improves robustness without altering inference-time behavior of the MLP.

Performance evaluation strategies

To ensure a fair comparison with previous models, we employed five common statistical metrics [55]: accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity (Sn), and specificity (Sp). Additionally, considering the imbalance in our training and testing datasets, we included area under the curve (AUC) and the F1 score in our evaluations. The equations for calculating these metrics are presented below:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

$$SN = \frac{TP}{TP+FN} \tag{11}$$

$$Sp = \frac{TN}{TN+FP} \tag{12}$$

$$F1Score = \frac{2\times Precision\times Recall}{Precision+Recall}, \tag{13}$$

where TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively. Additionally, we calculated the Precision-Recall (PR) curve to further assess the model's performance. These metrics provide a comprehensive evaluation of the model's classification capabilities, particularly in the context of imbalanced datasets where traditional accuracy may not fully capture performance nuances.

Results

Ablation study

To elucidate the influence of each major component of our architecture on predictive performance, we performed ablation studies on four modules—ESM-2, Traditional Descriptors (TD), the Transformer encoder, and the Cross-Attention block—as well as on different combinations of features within the TD set. We adopted tenfold cross-validation [56, 57] partitioning the dataset into ten non-overlapping folds. In each round, ninefold were used for training and the remaining fold for evaluation, with results reported as the mean ± standard deviation across folds.

To further elucidate the impact of each traditional descriptor subset, we evaluated DPC alone as well as its pairwise and full fusions with AAC and PCP, the results are shown in Table 2. The full fusion (ACP-EPC) concatenates AAC, DPC, and PCP into a 429-dimensional vector; pairwise fusions yield 420-dimensional (DPC + AAC) or 409-dimensional (DPC + PCP) vectors. DPC alone establishes a strong baseline (ACC = 0.9166 ± 0.0175, Sn = 0.8028 ± 0.0559, Sp = 0.9540 ± 0.0272, MCC = 0.7754 ± 0.0452). ACP-EPC was therefore adopted as the default descriptor set for all subsequent experiments, owing to its consistent improvements in ACC and MCC and its negligible runtime overhead. Lower-dimensional fusions required linear projection to 256 dimensions, which risked distorting feature correlations and amplifying noise. These combinations were therefore excluded. When fused with AAC, DPC achieves the highest sensitivity of all configurations (Sn = 0.8236 ± 0.0359), together with modest gains in ACC (ACC = 0.9187 ± 0.0251) and MCC (0.7827 ± 0.0599). In contrast, the DPC + PCP fusion attains the highest Sp (Sp = 0.9628 ± 0.0223), improving false-positive control while maintaining ACC (0.9181 ± 0.0185) and MCC (0.7762 ± 0.0496) comparable to the DPC baseline. The ACP-EPC configuration delivers

Table 2 Ablation of traditional descriptor blocks (AAC, DPC, PCP) and their pairwise fusions

Feature combination	ACC	Sn	Sp	MCC	Time/Epoch
DPC (400)	0.9166 ± 0.0175	0.8028 ± 0.0559	0.9540 ± 0.0272	0.7754 ± 0.0452	**0.74 ± 0.05**
DPC + AAC (420)	0.9187 ± 0.0251	**0.8236 ± 0.0359**	0.9500 ± 0.0304	0.7827 ± 0.0599	0.81 ± 0.19
DPC + PCP (409)	0.9181 ± 0.0185	0.7821 ± 0.0435	**0.9628 ± 0.0223**	0.7762 ± 0.0496	0.76 ± 0.03
ACP-EPC (429)	**0.9232 ± 0.0192**	0.8133 ± 0.0501	0.9594 ± 0.0211	**0.7920 ± 0.0515**	0.79 ± 0.06

the best overall performance, ACC = 0.9232 ± 0.0192, MCC = 0.7920 ± 0.0515, while preserving a balanced Sn/Sp profile (Sn = 0.8133 ± 0.0501, Sp = 0.9594 ± 0.0211).

These findings highlight the complementary contributions of the descriptor modules. DPC encodes local sequence context, AAC captures global compositional bias that primarily enhances Sn, and PCP encodes interpretable physicochemical constraints (e.g., net charge, hydrophobicity and secondary-structure propensities) that predominantly improve precision. Their combination establishes a more reliable decision boundary, as reflected by the highest MCC. Efficiency losses are minimal: DPC alone requires 0.74 ± 0.05 s per epoch, ACP-EPC incurs only a 7% increase in computational time (0.79 ± 0.06 s/epoch), and pairwise fusions fall within 0.76–0.81 s/epoch.

Table 3 summarizes the contribution of each architectural component—ESM-2 contextual embeddings, traditional descriptors (AAC + DPC + PCP), Cross-Attention fusion, and the Transformer encoder—by comparing four configurations. Using only traditional descriptors yields the weakest performance (ACC = 0.9085 ± 0.0188, Sn = 0.7562 ± 0.0381, MCC = 0.7454 ± 0.0571), indicating that handcrafted features alone are insufficient for robust discrimination despite low runtime. Using only ESM-2 provides a strong baseline (ACC = 0.9222 ± 0.0176, MCC = 0.7867 ± 0.0440) with high specificity (Sp = 0.9626 ± 0.0173) but lower sensitivity (Sn = 0.7975 ± 0.0542), suggesting that contextual sequence signals alone bias the model toward conservative, high-Sp decisions. Adding Cross-Attention fusion without the Transformer ('No_Transformer') modestly improves performance (ACC = 0.9232 ± 0.0123, MCC = 0.7902 ± 0.0287) and yields the best Sp (Sp = 0.9657 ± 0.0226), reflecting the role of physicochemical priors in reducing false positives. The full model (ACP-EPC), which refines ESM-2 features with a Transformer encoder before fusion, matches the highest accuracy (ACC = 0.9232 ± 0.0192) and achieves the best MCC (0.7920 ± 0.0515) while delivering the highest sensitivity (Sn = 0.8133 ± 0.0501) with only a minor decrease in Sp (Sp = 0.9594 ± 0.0211). This pattern indicates that the encoder enhances recall by sharpening context around motif-bearing positions, and the fusion with interpretable descriptors stabilizes the decision boundary, yielding the most balanced error profile. Although ACP-EPC requires the longest epoch time (0.79 ± 0.06 s) relative to ESM-2-only (0.49 ± 0.04 s) and fusion without the encoder (0.55 ± 0.03 s), this modest overhead is warranted by its superior MCC and Sn—metrics that are critical for high-recall screening of candidate ACPs. ACP-EPC was therefore adopted as the default configuration in subsequent experiments.

Interpretability analysis

In this section, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE) [58] to project and visualize the representations generated by the model's core module. We further computed SHAP values for the top-40 handcrafted features obtained from traditional feature extraction methods to quantify their individual contributions. Finally, since our architecture performs token-level fusion via Cross-Attention, we analyze position-specific attention patterns across all positive sequences in the training set by comparing attention score as a function of residue position, thereby revealing conserved, high-salience regions emphasized by the model.

Figure 3 (A) visualizes four stages of the pipeline using t-SNE, (I) traditional handcrafted features, (II) mean-pooled ESM-2 embeddings, (III) token-level Cross-Attention output, and (IV) the pre-classifier layer, on the Training set and two datasets (ACP135, ACP99). Across all datasets, the embeddings progress from partially overlapping clusters in (I) to markedly separated ACP vs. non-ACP manifolds in (III)–(IV). The tightening of within-class structure and widening inter-class separation after Cross-Attention and the final encoder indicate that token-level fusion yields more discriminative and transferable representations, rather than merely overfitting geometric structure.

Figure 3 (B) presents SHAP-based feature rankings according to their contributions to ACP prediction. Positive SHAP values, which is shown in red, shift predictions toward the ACP class, whereas negative values, which is shown in blue, shift them toward the non-ACP class. In the feature importance analysis, four global descriptors, pI, net charge, α-helix propensity, and polarity emerged as the most influential. These were followed by secondary contributors including molecular weight, β-sheet propensity, bulkiness, hydrophobicity, and average flexibility. Below these, residue/dipeptide features exert smaller, context-dependent effects. This organization indicates that the classifier relies primarily on global electrostatic and conformational cues and only secondarily on local sequence motifs. A strong, positive

Table 3 Architectural ablation of ESM-2 embeddings, traditional descriptors, Cross-Attention fusion, and the Transformer encoder

	ACC	Sn	Sp	MCC	Time/Epoch
ESM-2 Only	0.9222 ± 0.0176	0.7975 ± 0.0542	0.9626 ± 0.0173	0.7867 ± 0.0440	0.49 ± 0.04
No_Transformer	0.9232 ± 0.0123	0.7954 ± 0.0611	**0.9657 ± 0.0226**	0.7902 ± 0.0287	0.55 ± 0.03
Traditional Only	0.9085 ± 0.0188	0.7562 ± 0.0381	0.9593 ± 0.0184	0.7454 ± 0.0571	**0.16 ± 0.09**
ACP-EPC	**0.9232 ± 0.0192**	**0.8133 ± 0.0501**	0.9594 ± 0.0211	**0.7920 ± 0.0515**	0.79 ± 0.06

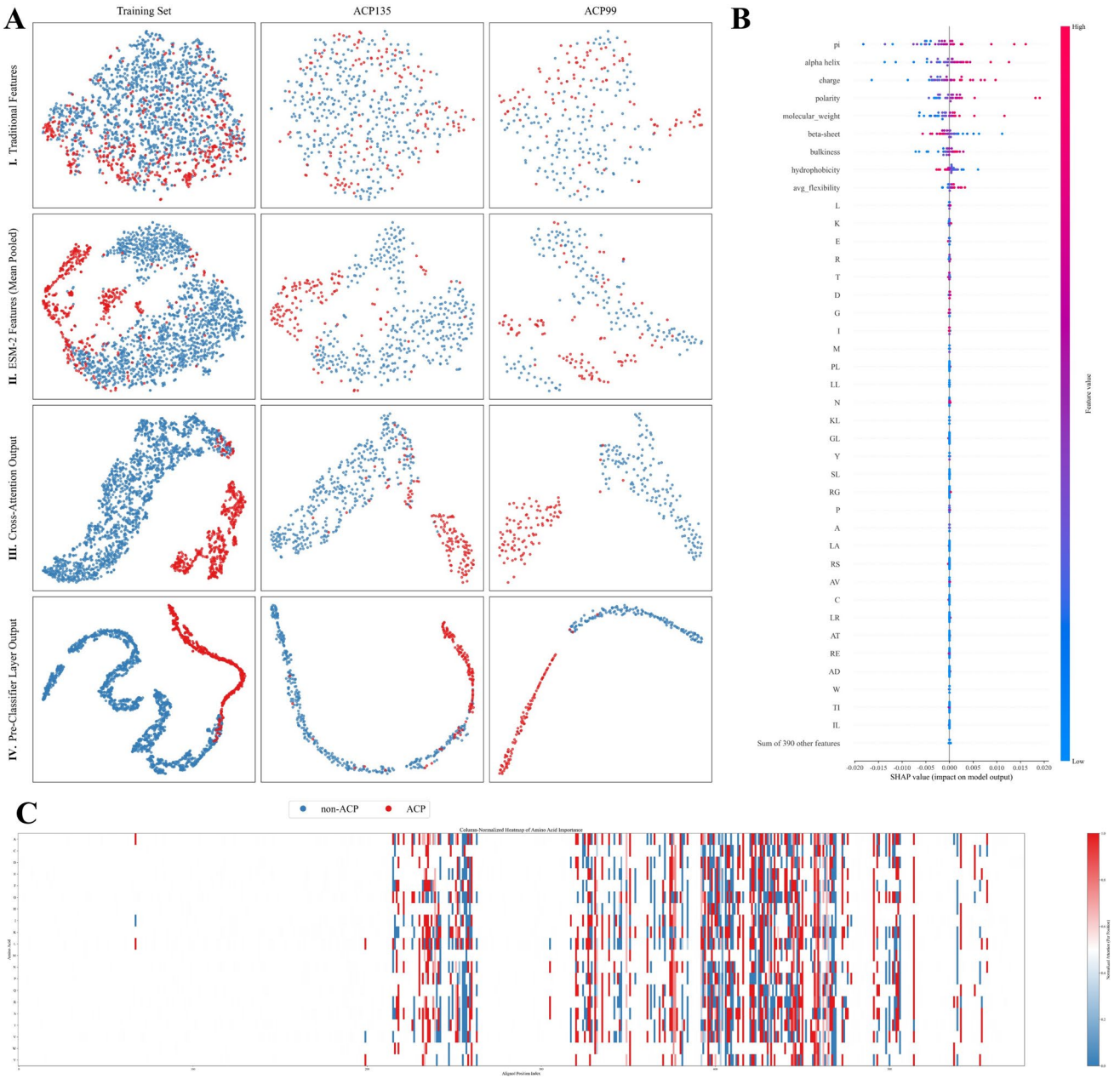

Fig. 3 Representation analysis and interpretability of ACP prediction. **A** t-SNE visualizations of sample representations at four stages of the pipeline. Rows (I–IV) show I) traditional handcrafted features; II) ESM-2 sequence embeddings (mean-pooled); III) outputs after token-level Cross-Attention fusion; IV) the pre-classifier layer outputs. Columns correspond to the Training set, ACP135 and ACP99. Points are colored by class (red: ACP; blue: non-ACP). **B** SHAP summary for the top-40 handcrafted features. Each point represents a sample's SHAP value for a given feature; horizontal position indicates impact on the model output (logit), and color encodes the original feature value (high to low). Features are ordered by mean SHAP. **C** Token-level Cross-Attention saliency across all positive sequences in the Training set. Rows denote sequences (aligned by residue index), columns denote positions, and the color scale shows normalized attention weights. Recurrent high-salience bands indicate positions consistently emphasized by the model

attribution for pI and net positive charge is mechanistically consistent with cancer cell membranes, which present a more anionic outer leaflet in part due to externalized phosphatidylserine. Cationic peptides are electrostatically captured by such surfaces, enhancing selective binding to tumor cells over healthy cells. The SHAP pattern, high pI/charge values shifting predictions rightward, accords with this electrostatic capture model and with reports that ACP activity tracks the interplay of net charge and membrane composition [43, 59]. The prominent, positive influence of α-helix propensity likewise mirrors biophysics: many ACPs are amphipathic helices that are disordered in aqueous solution

but fold upon contacting lipid bilayers, where their segregated hydrophobic and cationic faces enable insertion, pore formation, or carpet-like disruption. The model's emphasis on helical readiness therefore reads as a proxy for membrane active architecture rather than overfitting to sequence idiosyncrasies [60, 61]. Hydrophobicity shows a moderate yet bidirectional contribution, sometimes pushing predictions toward ACPs, sometimes toward non-ACPs, indicating that the model captures the need for balance rather than monotonic increases. Sufficient hydrophobic surface is required for bilayer partitioning and translocation, but excessive hydrophobicity can erode selectivity and raise nonspecific cytotoxicity; SHAP's spread around zero is congruent with this trade-off [62]. At the amino-acid level, features enriched in Lys/Arg tend to nudge predictions toward ACPs, in line with the role of guanidinium (A) and ε-amino (L) groups in mediating strong electrostatic and hydrogen bonding interactions with anionic phospholipids, while also contributing to snorkeling and amphipathic packing on a helical face. Acidic residues (Asp/Glu) more often pull the prediction leftward, opposing cationic character. These residue-level attributions are expected from sequence statistics and physicochemical studies of cationic, helix-forming ACPs [48]. Finally, the comparatively smaller and sometimes negative impact of β-sheet propensity relative to helical propensity reflects the empirical predominance of helical scaffolds among membrane-disruptive peptides, while acknowledging that β-sheet ACPs exist but are less frequent in datasets of membrane active sequences. The model's weighting thus aligns with known structure function trends rather than arbitrary feature preference [63]. Overall, the SHAP analysis recapitulates mechanism-grounded determinants of ACP activity in ACP135, high pI, and positive charge for electrostatic targeting of PS-rich cancer membranes, strong α-helix propensity for amphipathic insertion, and disruption and a tuned hydrophobic profile that balances membrane engagement with selectivity. This concordance between model explanations and biophysics supports the reliability of our classifier and suggests practical design levers Lys/Arg rich, helix-forming sequences with moderated hydrophobicity.

Figure 3 (C) depicts residue and position-specific saliency scores learned by the classifier from aligned anticancer peptides. All positive peptides from the training set were first aligned using Clustal Omega [64] to place homologous residues within a common multiple sequence alignment (MSA) frame, providing a biologically meaningful coordinate system for comparing attention across peptides. From the trained model, we extracted Cross-Attention weights where physicochemical descriptors served as Q and ESM-2 embeddings as K/V. For each aligned column, weights were aggregated across peptides and then min–max normalized within that column to emphasize intracolumn differences. The heatmap plots amino-acid types (y-axis) against aligned positions (x-axis); color encodes normalized attention (red, high; white, intermediate; blue, low), while blank cells denote alignment-induced gaps.

The saliency pattern is heterogeneous across positions and tends to concentrate in columns with high residue occupancy. Although gaps are introduced by the MSA, the model assigns negligible attention to these regions; instead, positions that are alignment-stable and consistently populated accumulate stronger, more coherent attention. Recurrent vertical bands of high attention mark positions repeatedly emphasized by the model, consistent with residues previously implicated in ACP activity. In particular, cationic residues (Lys, Arg), aromatics (Trp, Phe, Tyr), and hydrophobics (Leu, Ile, Val) receive higher attention, whereas columns dominated by polar uncharged residues (Ser, Thr, Asn, Gln) or helix-disrupting residues (Gly, Pro) exhibit lower attention [63, 65, 66].

Several segments display alternating red–blue patterns with a periodicity of about 3–4 residues, echoing the amphipathic cadence characteristic of α-helical surfaces [67]. This suggests that the model captures not only local residue identity but also higher-order physicochemical motifs relevant to membrane interaction and selectivity. Taken together, the visualization indicates that the classifier prioritizes evolutionarily conserved, mechanistically meaningful features over incidental sequence variation, thereby improving the biological interpretability of its predictions.

Performance analysis

To better evaluate the performance of our model on the ACP135 and ACP99 datasets, we utilized Kernel Density Estimation (KDE) [68] to illustrate the predicted probabilities for each sample. Additionally, we analyzed the Receiver Operating Characteristic (ROC) curves for both datasets. Given the class imbalance, we further included Precision–Recall (PR) curves to assess performance across thresholds. Compared to ROC curves, PR curves offer more insightful evaluations of the model's ability to predict the minority class in imbalanced datasets. The results are summarized in Fig. 4.

For the ACP135, shown in Fig. 4(A), the class-conditional score densities are clearly bimodal: non-ACPs concentrate near the lower tail (0–0.2), whereas ACPs cluster near the upper tail (0.9–1.0). A modest overlap appears in the mid-range, reflecting a small cohort of borderline sequences with indistinct physicochemical signatures or conflicting contextual cues. The ACP mode is broader than the non-ACP mode, suggesting slightly greater heterogeneity among positives, consistent with diverse anticancer mechanisms. From a decision-theoretic perspective, the Bayes-optimal threshold would lie near the intersection of the densities, enabling high specificity with minimal sensitivity loss.

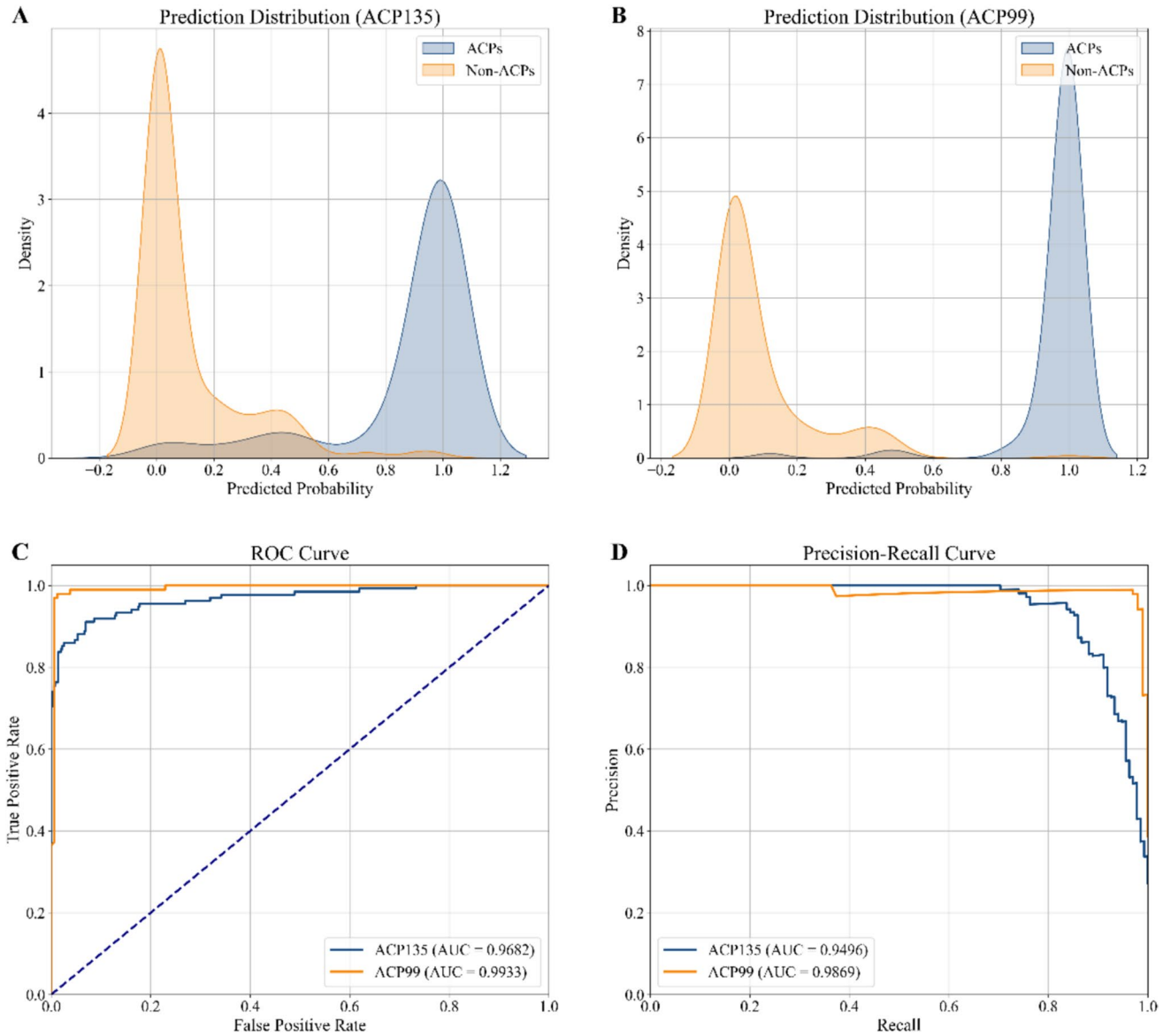

Fig. 4 Performance of ACP-EPC on ACP135 and ACP99. **A** KDE of predicted probability $p(y = ACP)$ on ACP135. Blue means ACPs; orange means non-ACPs. ACPs cluster near 1.0 and non-ACPs near 0.0 with limited overlap, indicating strong separability. **B** KDE of $p(y = ACP)$ on ACP99. **C** ROC for ACP135 (blue) and ACP99 (orange); axes: TPR vs FPR. The dashed diagonal denotes random. Reported AUCs: 0.9682 (ACP135) and 0.9933 (ACP99). **D** PR for the positive class using the same predictions; axes: precision vs recall. Reported AUC(PR): 0.9496 (ACP135) and 0.9869 (ACP99), showing high precision sustained across broad recall, especially on ACP99

Compared with ACP135, ACP99, shown in Fig. 4(B), shows tighter, more peaked modes at the extremes and markedly reduced overlap, implying a larger margin between classes and higher confidence in individual predictions. The narrow dispersion of ACP scores near 1.0 and non-ACP scores near 0 suggests low aleatoric uncertainty and well-aligned feature representations. Practically, a wide interval of thresholds would yield similar operating characteristics, making deployment less sensitive to threshold tuning. Score distributions indicate robust separability overall—near-linear separability on ACP99 and strong, though not perfect, separability on ACP135 with ambiguity confined to a narrow probability band.

The ROC curves, shown in Fig. 4(C), approach the upper-left envelope on both datasets, with AUC = 0.9682 (ACP135) and 0.9933 (ACP99), evidencing excellent threshold-independent ranking. The steep initial ascent shows that high true-positive rates are attainable at negligible false-positive rates—especially on ACP99, consistent with its sharper class-conditional densities. Operating at the Youden-optimal

point [69, 70] would yield high sensitivity with only a minor specificity trade-off, aligning with the observed margins.

Considering the inherent class imbalance in the datasets, PR curves, shown in Fig. 4(D), remain near the upper envelope, with PR-AUC = 0.9496 (ACP135) and 0.9869 (ACP99). Precision is sustained over a wide recall range, indicating that high-scoring predictions are rarely spurious; only at extreme recall on ACP135 does precision taper, mirroring the ambiguous mid-probability region. These profiles imply strong early enrichment (high precision at moderate recall) and competitive F1 across plausible thresholds, favorable for screening pipelines where false discoveries are costly. The PR results demonstrate resilience to class imbalance, confirming that the model maintains a low false-discovery rate while recovering most true positives—approaching saturation performance on ACP99 and showing robust, stable behavior on ACP135, with limitations mainly due to a small subset of borderline cases.

Performance comparison with previous models

In this section, we compared the performance of our model with previous approaches on the ACP135 and ACP99 datasets. On ACP135, which is shown in Table 4, our model achieved ACC of 0.935, Sn of 0.859, Sp of 0.964, and MCC of 0.835, representing improvements of 2.6% in ACC and 6.5% in MCC over the previous best results. On ACP99, which is shown in Table 5, the model obtained ACC of 0.984, Sn of 0.980, Sp of 0.987, and MCC of 0.993, surpassing the prior model by 5.8% in ACC, 4.3% in Sn, 6.9% in Sp, and 0.15 in MCC. Considering that our model was trained with just one training set, it achieved high performance on two datasets, compared with some existing methods, with similar values for ACC and Sp.

The marked performance gains arise from three design choices. First, we leverage a state-of-the-art protein pretrained language model (ESM-2) to obtain rich, contextualized token embeddings that capture long-range, structure-aware dependencies. Second, we complement these embeddings with interpretable, sequence-level physicochemical descriptors derived via traditional feature engineering. Third, we introduce a Cross-Attention module that injects the global physicochemical profile into every token representation, ensuring that each token encodes not only local sequence context, via positional encoding and self-attention, but also global physicochemical signals provided by the handcrafted features. This fusion yields more expressive and balanced representations, underpinning the observed improvements.

Table 4 Performance comparison of ACP- EPC with existing methods on ACP135

Model	ACC	Sn	Sp	MCC
ACPred-BMF	0.669	0.838	0.429	0.296
ACP-MHCNN	0.590	0.783	0.338	0.136
ACPred	0.857	0.914	0.719	0.648
iDACP	0.879	0.861	**0.975**	0.687
mACPpred	0.871	0.879	0.838	0.659
ACP-ML	0.909	**0.939**	0.830	0.770
ACP-EPC	**0.935**	0.859	0.964	**0.835**

Table 5 Performance comparison of ACP-EPC with existing methods on ACP999

Model	ACC	Sn	Sp	MCC
mACPpred	0.895	0.882	0.918	0.777
ACPred	0.864	0.896	0.814	0.714
ACP-MHCNN	0.699	0.738	0.625	0.354
ACPred-BMF	0.758	0.920	0.629	0.561
ACP-ML	0.926	0.937	0.908	0.843
ACP-EPC	**0.984**	**0.980**	**0.987**	**0.993**

Discussion

In this study, we present ACP-EPC, a deep learning model designed for the rapid and accurate identification of ACPs. The model employs a Cross-Attention mechanism to integrate features extracted using multiple feature extraction methods. During the feature extraction stage, we utilized four traditional feature extraction methods to capture physicochemical properties related to the functionality of ACPs. Additionally, we leveraged the 650 M-parameter version of ESM-2 to extract sequence features. The two types of features were then fused using the Cross-Attention mechanism. The resulting features passed into a MLP for classification. We evaluated our model on two datasets, ACP135 and ACP99, which are characterized by the removal of homologous sequences and significant differences in sequence length and amino-acid distribution. On these datasets, our model demonstrated significant improvements over existing models in terms of ACC, Sp, and MCC. Specifically, on the ACP135 dataset, the model achieved an ACC of 0.935 and MCC of 0.835, while on the ACP99 dataset, it achieved an ACC of 0.984, Sn of 0.980, Sp of 0.987, and MCC of 0.993. These results highlight the high accuracy and excellent generalization ability of our model in the domain of ACP identification, meaning that researchers can use our model to quickly perform preliminary screening of thousands of candidate peptides. In the actual drug development process, researchers do not

need to conduct individual studies and analyses on each peptide. Instead, they can efficiently identify peptides with anticancer potential through our model, thereby significantly reducing the number of samples required for drug experimentation. This not only reduces the cost and time required for the development of anticancer peptide drugs but also enhances the efficiency of the research and development process. We attribute the superior performance of ACP-EPC to its deliberately layered feature engineering strategy. First, the PLM (ESM-2, 650 M parameters) captures fine-grained positional dependencies among residues, implicitly encoding structural and evolutionary contexts of each peptide. Second, we manually encode a comprehensive suite of physicochemical descriptors that summarize physicochemical attributes critical to anticancer activity. By integrating these two complementary representations through a Cross-Attention module, the model learns a synergistic, information dense feature space that simultaneously preserves high-level contextual relationships and low-level physicochemical signals. This enriched representation underpins the marked gains in accuracy, specificity, and MCC observed across diverse benchmark datasets.

Despite these promising results, the model remains limited in certain respects. The ESM-2 embeddings were used without fine-tuning due to the small dataset size, and the physicochemical property descriptors may not comprehensively capture ACP functionality. Addressing these gaps will require collaboration with experimental research teams to better understand the mechanistic basis of ACP activity.

To address these challenges, the future work will focus on expanding ACP datasets and refining physicochemical descriptors through experimental studies under varying environmental condition. These efforts are expected to enhance the robustness, recognition performance, and generalization ability of ACP-EPC, thereby advancing its applicability in anticancer peptide discovery.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-025-11352-x.

Author contribution J.L. contributed toward project administration, resources, software, writing—original draft, and writing—review & editing. K.L contributed toward validation, visualization, and writing—review & editing. Y.W. contributed toward data curation, formal analysis, validation, and writing—review & editing. J.X. contributed toward software, validation, and visualization. Y.M. contributed toward investigation, methodology, software, validation, and visualization. F.C. contributed toward investigation, methodology, project administration, software, and writing—review & editing. L.W. contributed toward investigation, methodology, project administration, and resources. Q.Z. contributed toward investigation, methodology, and software. Z.Z. contributed toward conceptualization, data curation, formal analysis, methodology, project administration, and writing—review & editing.

Funding The work was supported by the National Natural Science Foundation of China (No. 62262015), Science and Technology special fund of Hainan Province (ZDYF2024GXJS018), and the Science and Technology Development Fund of Macau (No. 0177/2023/RIA3).

Data availability The work and source code are available to researchers and developers at https://github.com/EuclidLv/ACP-EPC. We also established a freely available online web server at http://www.bioai-lab.com/ACP-EPC.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval The research presented in this manuscript does not need an Ethics statement. No experimentation with animals or human subjects is contained.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the author(s) used ChatGPT in order to polishing the manuscript to enhance its clarity, coherence, and overall expressiveness. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Siegel RL, Giaquinto AN, Jemal A (2024) Cancer statistics, 2024. CA Cancer J Clin. https://doi.org/10.3322/caac.21820
2. Debela DT, Muzazu SG, Heraro KD, Ndalama MT, Mesele BW, Haile DC, Kitui SK, Manyazewal T (2021) New approaches and procedures for cancer treatment: current perspectives. SAGE Open Med 9:20503121211034370. https://doi.org/10.1177/20503121211034366
3. Saranya S, Bharathi M, Kumar NS, Chellapandi P (2024) Design and characterization of anticancer peptides derived from snake venom metalloproteinase library. Int J Pept Res Ther 30(3):24. https://doi.org/10.1007/s10989-024-10602-0
4. Norouzi P, Mirmohammadi M, Tehrani MHH (2022) Anticancer peptides mechanisms, simple and complex. Chem Biol Interact 368:110194. https://doi.org/10.1093/nar/gkae241
5. Sah BNP, Vasiljevic T, McKechnie S, Donkor O (2015) Identification of anticancer peptides from bovine milk proteins and their potential roles in management of cancer: a critical review. Compr Rev Food Sci Food Saf 14(2):123–138. https://doi.org/10.1111/1541-4337.12126
6. Yan C, Geng A, Pan Z, Zhang Z, Cui F (2024) MultiFeatVotPIP: a voting-based ensemble learning framework for predicting pro-inflammatory peptides. Brief Bioinform. https://doi.org/10.1093/bib/bbae505
7. Arif M, Musleh S, Fida H, Alam T (2024) Plmacpred prediction of anticancer peptides based on protein language model and wavelet denoising transformation. Sci Rep 14(1):16992. https://doi.org/10.1038/s41598-024-67433-8
8. Amjad A, Ahmed S, Kabir M, Arif M, Alam T (2024) A novel deep learning identifier for promoters and their strength using heterogeneous features. Methods 230:119–128. https://doi.org/10.1016/j.ymeth.2024.08.005
9. Ali F, Ibrahim N, Alsini R, Masmoudi A, Alghamdi W, Alkhalifah T, Alturise F (2025) Comprehensive analysis of computational models for prediction of anticancer peptides using machine learning and deep learning. Arch Comput Method E. https://doi.org/10.1007/s11831-025-10237-4
10. Sultan MF, Karim T, Shaon MSH, Ali MM, Ibrahim SM, Akter MS, Ahmed K, Bui FM, Moni MA (2025) Bitteren: a novel

ensemble model for the identification of bitter peptide. Comput Biol Med 195:110528. https://doi.org/10.3390/pr9060992
11. Sultan MF, Shaon MSH, Karim T, Ali MM, Hasan MZ, Ahmed K, Bui FM, Chen L, Dhasarathan V, Moni MA (2024) MLAFP-XN: Leveraging neural network model for development of antifungal peptide identification tool. Heliyon. https://doi.org/10.1016/j.heliyon.2024.e37820
12. Ao C, Jiao S, Wang Y, Yu L, Zou Q (2022) Biological sequence classification: a review on data and general methods. Research. https://doi.org/10.34133/research.0011
13. Geng A, Luo Z, Li A, Zhang Z, Zou Q, Wei L, Cui F (2025) ACP-CLB: an anticancer peptide prediction model based on multichannel discriminative processing and integration of large pretrained protein language models. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.4c02072
14. Gao S, Xia Y, Li X, Cui F, Zhang Q, Zou Q, Zhang Z (2025) ACP-esm2: enhancing anticancer peptide prediction with pre-trained protein language models. IEEE Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBBIO.2025.3547952
15. Wei L, Zhou C, Chen H, Song J, Su R (2018) ACpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 34(23):4007–4016. https://doi.org/10.1093/bioinformatics/bty451
16. Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W (2019) ACpred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 24(10):1973. https://doi.org/10.3390/molecules24101973
17. Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang D-C (2019) MAcpped: a support vector machine-based meta-predictor for identification of anticancer peptides. Int J Mol Sci 20(8):1964. https://doi.org/10.3390/ijms20081964
18. Wei L, Zhou C, Su R, Zou Q (2019) PEPred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics 35(21):4272–4280. https://doi.org/10.1093/bioinformatics/btz246
19. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3(02):185–205. https://doi.org/10.1142/S0219720005001004
20. Yi H-C, You Z-H, Zhou X, Cheng L, Li X, Jiang T-H, Chen Z-H (2019) ACP-dl: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Ther-Nucleic Acids 17:1–9. https://doi.org/10.1016/j.omtn.2019.04.025
21. Ge R, Feng G, Jing X, Zhang R, Wang P, Wu Q (2020) Enacp: an ensemble learning model for identification of anticancer peptides. Front Genet 11:760. https://doi.org/10.3389/fgene.2020.00760
22. Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GP (2021) AntiCP 2.0: an updated model for predicting anticancer peptides. Brief Bioinform 22(3):bbaa153. https://doi.org/10.1093/bib/bbaa153
23. Chen X-g, Zhang W, Yang X, Li C, Chen H (2021) Acp-da: improving the prediction of anticancer peptides using data augmentation. Front Genet 12:698477. https://doi.org/10.3389/fgene.2021.698477
24. Huang K-Y, Tseng Y-J, Kao H-J, Chen C-H, Yang H-H, Weng S-L (2021) Identification of subtypes of anticancer peptides based on sequential features and physicochemical properties. Sci Rep 11(1):13594. https://doi.org/10.1038/s41598-021-93124-9
25. Ahmed S, Muhammod R, Khan ZH, Adilina S, Sharma A, Shatabda S, Dehzangi A (2021) ACP-mhcnn: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. Sci Rep 11(1):23676. https://doi.org/10.1038/s41598-021-02703-3
26. Han B, Zhao N, Zeng C, Mu Z, Gong X (2022) Acpred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. Sci Rep 12(1):21915. https://doi.org/10.1038/s41598-022-24404-1
27. Antonini AS, Tanzola J, Asiain L, Ferracutti GR, Castro SM, Bjerg EA, Ganuza ML (2024) Machine learning model interpretability using SHAP values: application to Igneous Rock Classification task. Appl Comput Geosci. https://doi.org/10.1016/j.acags.2024.100178
28. Wu X, Zeng W, Lin F, Xu P, Li X (2022) Anticancer peptide prediction via multi-kernel CNN and attention model. Front Genet 13:887894. https://doi.org/10.3389/fgene.2022.887894
29. Park HW, Pitti T, Madhavan T, Jeon Y-J, Manavalan B (2022) MLACP 2.0: an updated machine learning tool for anticancer peptide prediction. Comput Struct Biotechnol J 20:4473–4480. https://doi.org/10.1016/j.csbj.2022.07.043
30. Bian J, Liu X, Dong G, Hou C, Huang S, Zhang D (2024) ACP-ML: a sequence-based method for anticancer peptide prediction. Comput Biol Med 170:108063. https://doi.org/10.1016/j.compbiomed.2024.108063
31. Fang H, Tang P, Si H (2020) Feature selections using minimal redundancy maximal relevance algorithm for human activity recognition in smart home environments. J Healthcare Eng 2020(1):8876782. https://doi.org/10.1155/2020/8876782
32. Lazrek G, Chetioui K, Balboul Y, Mazer S (2024) An RFE/Ridge-ML/DL based anomaly intrusion detection approach for securing IoMT system. Results Eng 23:102659. https://doi.org/10.1016/j.rineng.2024.102659
33. Xiao C, Zhou Z, She J, Yin J, Cui F, Zhang Z (2024) PEL-pvp: application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset. Int J Biol Macromol 277:134317. https://doi.org/10.1016/j.ijbiomac.2024.134317
34. Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, Joshi A, Singh S, Gautam A, Raghava GP (2015) CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res 43(D1):D837–D843. https://doi.org/10.1093/nar/gku892
35. Wang G, Li X, Wang Z (2016) APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 44(D1):D1087–D1093. https://doi.org/10.1093/nar/gkv1278
36. Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, Tuknait A, Agrawal P, Mathur D, Raghava GP (2016) SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res 44(D1):D1119–D1126. https://doi.org/10.1093/nar/gkv1114
37. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158
38. Shen W, Le S, Li Y, Hu F (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 11(10):e0163962. https://doi.org/10.1371/journal.pone.0163962
39. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202. https://doi.org/10.1006/jmbi.1999.3091
40. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389
41. Chen Y, Guarnieri MT, Vasil AI, Vasil ML, Mant CT, Hodges RS (2007) Role of peptide hydrophobicity in the mechanism of action of α-helical antimicrobial peptides. Antimicrob Agents Chemother 51(4):1398–1406. https://doi.org/10.1128/AAC.00925-06
42. Fernández-Vidal M, Jayasinghe S, Ladokhin AS, White SH (2007) Folding amphipathic helices into membranes: amphiphilicity trumps hydrophobicity. J Mol Biol 370(3):459–470. https://doi.org/10.1016/j.jmb.2007.05.016

43. Xie M, Liu D, Yang Y (2020) Anti-cancer peptides: classification, mechanism of action, reconstruction and modification. Open Biol 10(7):200004. https://doi.org/10.1098/rsob.200004
44. Galdiero S, Falanga A, Cantisani M, Vitiello M, Morelli G, Galdiero M (2013) Peptide-lipid interactions: experiments and applications. Int J Mol Sci 14(9):18758–18789. https://doi.org/10.3390/ijms140918758
45. Gaspar D, Veiga AS, Castanho MA (2013) From antimicrobial to anticancer peptides. A review. Front Microbiol 4:294. https://doi.org/10.3389/fmicb.2013.00294
46. Amos S-BT, Vermeer LS, Ferguson PM, Kozlowska J, Davy M, Bui TT, Drake AF, Lorenz CD, Mason AJ (2016) Antimicrobial peptide potency is facilitated by greater conformational flexibility when binding to gram-negative bacterial inner membranes. Sci Rep 6(1):37639. https://doi.org/10.1038/srep37639
47. Hadianamrei R, Tomeh MA, Brown S, Wang J, Zhao X (2022) Rationally designed short cationic α-helical peptides with selective anticancer activity. J Colloid Interface Sci 607:488–501. https://doi.org/10.1016/j.jcis.2021.08.200
48. Zhang Y, Wang C, Zhang W, Li X (2023) Bioactive peptides for anticancer therapies. Biomater Transl 4(1):5. https://doi.org/10.1201/9781003052777-21
49. Khemaissa S, Sagan S, Walrant A (2021) Tryptophan, an amino-acid endowed with unique properties and its many roles in membrane proteins. Crystals 11(9):1032. https://doi.org/10.3390/cryst11091032
50. Varela-Quitián YF, Mendez-Rivera FE, Bernal-Estévez DA (2025) Cationic antimicrobial peptides: potential templates for anticancer agents. Front Med 12:1548603. https://doi.org/10.1002/9783527652853.ch2
51. Cordoves-Delgado G, García-Jacas CR (2024) Predicting antimicrobial peptides using ESMFold-predicted structures and ESM-2-based amino acid features with graph deep learning. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.3c02061
52. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23(10):1282–1288. https://doi.org/10.1093/bioinformatics/btm098
53. Lv J, Geng A, Pan Z, Wei L, Zou Q, Zhang Z, Cui F (2025) iBitter-GRE: a novel stacked Bitter peptide predictor with ESM-2 and multi-view features. J Mol Biol 437(8):169005. https://doi.org/10.1016/j.jmb.2025.169005
54. Yan C, Zhang Z, Xu J, Meng Y, Yan S, Wei L, Zou Q, Zhang Q, Cui F (2025) CasPro-ESM2: accurate identification of Cas proteins integrating pre-trained protein language model and multi-scale convolutional neural network. Int J Biol Macromol 308:142309. https://doi.org/10.1016/j.ijbiomac.2025.142309
55. Wang Y, Zhai Y, Ding Y, Zou Q (2024) SBSM-Pro: support biosequence machine for proteins. Sci China Inf Sci 67(11):212106. https://doi.org/10.1007/s11432-024-4171-9
56. Alsini R, Almuhaimeed A, Ali F, Khalid M, Farrash M, Masmoudi A (2024) Deep-VEGF: deep stacked ensemble model for prediction of vascular endothelial growth factor by concatenating gated recurrent unit with two-dimensional convolutional neural network. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2024.2323144
57. Almusallam N, Ali F, Masmoudi A, Ghazalah SA, Alsini R, Yafoz A (2024) An omics-driven computational model for angiogenic protein prediction: advancing therapeutic strategies with Ens-deep-AGP. Int J Biol Macromol 282:136475. https://doi.org/10.1016/j.ijbiomac.2024.136475
58. Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH, Hartline DK (2020) T-distributed stochastic neighbor embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. Mar Genomics 51:100723. https://doi.org/10.1016/j.margen.2019.100723
59. Herrera-León C, Ramos-Martín F, Antonietti V, Sonnet P, D'amelio N (2022) The impact of phosphatidylserine exposure on cancer cell membranes on the activity of the anticancer peptide HB43. FEBS J 289(7):1984–2003. https://doi.org/10.1111/febs.16276
60. Kabelka I, Vácha R (2021) Advances in molecular understanding of α-helical membrane-active peptides. Acc Chem Res 54(9):2196–2204. https://doi.org/10.1021/acs.accounts.1c00047
61. Huang Y, Feng Q, Yan Q, Hao X, Chen Y (2015) Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs. Mini Rev Med Chem 15(1):73–81. https://doi.org/10.2174/1389557514666141107120954
62. Zare-Zardini H, Saberian E, Jenča A, Ghanipour-Meybodi R, Petrášová A, Jenčová J (2024) From defense to offense: antimicrobial peptides as promising therapeutics for cancer. Front Oncol 14:1463088. https://doi.org/10.3389/fonc.2024.1463088
63. Dathe M, Wieprecht T (1999) Structural features of helical antimicrobial peptides: their potential to modulate activity on model membranes and biological cells. Biochimica et Biophysica Acta (BBA) 1462(1–2):71–87. https://doi.org/10.1016/S0005-2736(99)00201-1
64. Madeira F, Madhusoodanan N, Lee J, Eusebi A, Niewielska A, Tivey AR, Lopez R, Butcher S (2024) The EMBL-EBI job dispatcher sequence analysis tools framework in 2024. Nucleic Acids Res 52(W1):W521–W525. https://doi.org/10.1093/nar/gkae241
65. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75(1):422–427. https://doi.org/10.1016/S0006-3495(98)77529-0
66. Yau W-M, Wimley WC, Gawrisch K, White SH (1998) The preference of tryptophan for membrane interfaces. Biochemistry 37(42):14713–14718. https://doi.org/10.1074/jbc.M802074200
67. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W: **Hydrophobic moments and protein structure**. In: *Faraday Symposia of the Chemical Society: 1982*. Royal Society of Chemistry: 109–120.
68. Węglarczyk S: **Kernel density estimation and its application**. In: *ITM web of conferences: 2018*. EDP Sciences: 00037.
69. Unal I (2017) Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput Math Methods Med 2017(1):3762651. https://doi.org/10.1155/2017/3762651
70. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF (2008) Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biometrical J 50(3):419–430. https://doi.org/10.1002/bimj.200710415