A Standardized Benchmark for Multilabel Antimicrobial Peptide Classification

Sebastian Ojeda[*], **Rafael Velasquez**, **Nicolás Aparicio**, **Juanita Puentes**, **Paula Cárdenas**, **Nicolás Andrade**, **Gabriel González**, **Sergio Rincón**, **Carolina Muñoz-Camargo**, **Pablo Arbeláez**

Universidad de los Andes, Colombia

Abstract

Antimicrobial peptides have emerged as promising molecules to combat antimicrobial resistance. However, fragmented datasets, inconsistent annotations, and the lack of standardized benchmarks hinder computational approaches and slow down the discovery of new candidates. To address these challenges, we present the Expanded Standardized Collection for Antimicrobial Peptide Evaluation (**ESCAPE**), an experimental framework integrating over $80\,000$ peptides from 27 validated repositories. Our dataset separates antimicrobial peptides from negative sequences and incorporates their functional annotations into a biologically coherent multilabel hierarchy, capturing activities across antibacterial, antifungal, antiviral, and antiparasitic classes. Building on ESCAPE, we propose a transformer-based model that leverages sequence and structural information to predict multiple functional activities of peptides. Our method achieves up to a 2.56% relative average improvement in mean Average Precision over the second-best method adapted for this task, establishing a new state-of-the-art multilabel peptide classification. ESCAPE provides a comprehensive and reproducible evaluation framework to advance AI-driven antimicrobial peptide research. [1]

1 Introduction

Antibiotics are crucial in modern medicine, enabling routine procedures and treating common infections. However, widespread misuse and overuse have led to the rise of antimicrobial resistance (AMR), where bacteria and other pathogens develop mechanisms to endure these drugs [1, 2]. This issue extends beyond bacteria, including viruses, fungi, and parasites that have rapidly evolved, hindering the treatment of infectious diseases globally [3]. The Institute of Health Metrics and Evaluation estimates that antimicrobial-resistant infections could cause over 39 million deaths between 2025 and 2050, with South Asia and Latin America facing the highest mortality rates [4]. Besides the impact of AMR on healthcare systems and population lifespan, it is also a concern for national economies across the globe [5]. A recent United Nations report warns that, without action on AMR, not only will healthcare costs increase, but also the global GDP may decrease by US$3.4 trillion and drive an additional 24 million people into extreme poverty by 2030 [6].

As a result, the scientific community has turned to alternative molecules capable of fighting infectious microorganisms without quickly triggering resistance. This process has led to the exploration of antimicrobial peptides (AMPs), which are either naturally occurring or synthetically designed proteins with a broad spectrum of antimicrobial properties [7]. Unlike traditional antibiotics, AMPs often act

[*]Corresponding author: s.ojedaa@uniandes.edu.co

[1]The ESCAPE Dataset is available at https://doi.org/10.7910/DVN/C69MCD and the ESCAPE Baseline code at https://github.com/BCV-Uniandes/ESCAPE.

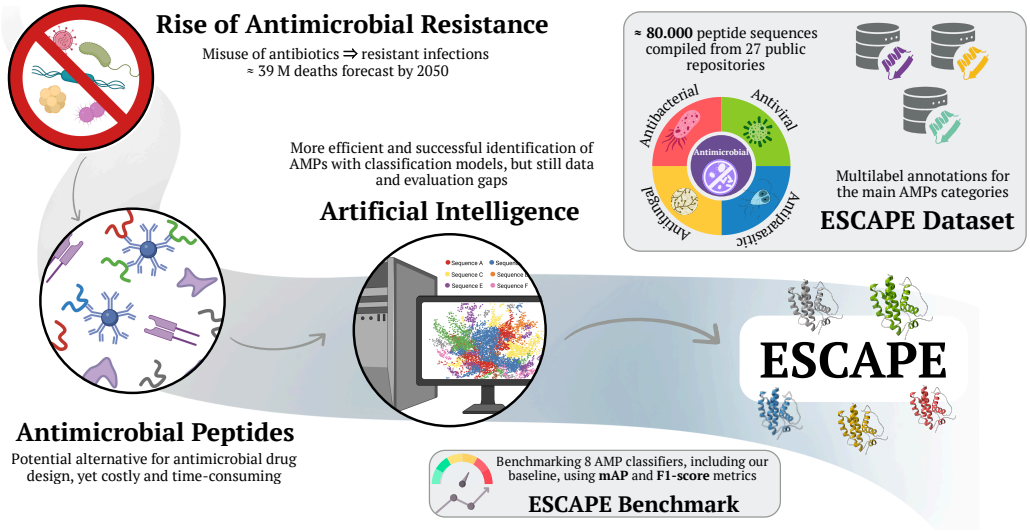

Figure 1: **Timeline of AMP Discovery and Computational Advances.** The rise of AMR underscores the urgent need for alternative therapies such as AMPs. While AI has shown promise in accelerating AMP discovery, progress is hindered by heterogeneous data and the absence of standardized evaluation protocols. We introduce ESCAPE to address these challenges and provide a robust foundation for future AI-driven methods.

through mechanisms harder for pathogens to evade, reducing the likelihood of resistance development [8]. Despite their therapeutic potential, the research and development of antimicrobial peptides, as with many pharmaceutical compounds, remains costly, time-consuming, and frequently unsuccessful. These difficulties are due to the inherent complexity of identifying effective AMPs and the extensive clinical trials and requirements novel drugs must undergo to reach the market and become profitable [9]. These challenges create a bottleneck that limits the widespread adoption of AMPs and highlights the need for more efficient and scalable discovery protocols.

To address the challenges of AMP discovery, researchers have explored using Artificial Intelligence (AI) tools to accelerate the identification of promising antimicrobial candidates [10]. Most existing models aim to predict whether a given peptide exhibits antimicrobial activity, often casting the task as a binary classification problem [11, 12, 13]. While this binary approach helps identify potentially active peptides, it disregards the proven ability of AMPs to interact with multiple types of microorganisms [14]. Despite this constraint, interest in AI-driven AMP classification tools has expanded in recent years. Some of the proposed methods include models based on Graph Neural Networks (GNNs) and other architectures inspired by advances in Natural Language Processing (NLP) [15, 16]. However, the performance of these models remains suboptimal, indicating a need for more effective modeling strategies and further exploration of architectures.

A significant limitation in the current literature is the inconsistency and insufficiency of data selected for training AI models [17]. Although the number of publicly available datasets continues to grow, most AI methods rely on individual datasets containing only a few hundred to a few thousand peptides [18]. Having such a limited number of examples from a single dataset when developing an AI model may restrict its learning potential and demonstrates the need for a more extensive set of data [19]. Furthermore, the absence of a standardized dataset or benchmarking framework makes it difficult to compare models reliably and to confidently identify which approach represents the state-of-the-art.

To overcome these limitations, we present three main contributions displayed in Fig. 1. First, we compile, curate, and standardize 27 public AMPs databases into the Expanded Standardized Collection for Antimicrobial Peptide Evaluation (ESCAPE), a comprehensive dataset containing over 80 000 peptides. We pre-process and validate all sequences to support robust AI-driven antimicrobial research. Second, we evaluate seven publicly available AMP classification methods on ESCAPE, adapting those designed initially for binary tasks to handle multilabel classification. To the best of our knowledge, this work results in the first benchmark that fairly compares existing approaches on a unified and scalable dataset for AMP multilabel classification. Finally, we introduce the ESCAPE

Baseline, a transformer-based architecture that leverages both sequence and structural information from peptides to predict not only whether it is antimicrobial, but also the types of pathogens it targets. Our baseline outperforms the state-of-the-art methods on the complete ESCAPE Dataset with a relative average improvement with respect to the second-best method of 2.56% and 1.90% in mean Average Precision (mAP) and F1-score, respectively.

2 Related Work

Given the global concern for antimicrobial resistance, there has been a noticeable rise in the development of AMP databases in recent years. Building on these resources, researchers have developed numerous AI models to analyze and identify potential AMP candidates to optimize antimicrobial drug discovery.

2.1 AMP Databases

In the search for novel AMPs, researchers have developed several databases, each containing peptides annotated with diverse biological activities [20]. These databases can be broadly categorized into general [21, 22, 23, 24] and specialized databases [25, 26, 27]. General databases span a wide range of peptide functions or classes. In contrast, specialized databases focus on narrower aspects, such as peptide origin [26] or the type of target organism [25], [27]. The number of peptide entries and the granularity of functional classes vary substantially among databases. For instance, dbAMP [21] comprises $33\,065$ peptides annotated with 58 distinct functional classes, while DRAMP [22] includes $30\,260$ entries but only 8 classes. Moreover, the hierarchical level of these classes is often inconsistent across databases. Namely, LAMP2 [23] annotates $23\,253$ peptides into 38 classes, including "anti-Gram negative" and "anti-Gram positive," which researchers can interpret as subclasses of a broader "Antibacterial" category. In contrast, SATPdb [24] assigns its $19\,192$ peptides to just 10 classes, among them a single "Antibacterial" label. Hence, currently available AMP datasets present key limitations. The class granularity and hierarchy discrepancies complicate training predictive models and comparing their performance across datasets.

Additionally, training models on datasets composed entirely of peptides obtained using a single experimental technology may introduce methodological biases, ultimately hampering the model's ability to generalize across broader peptide sources [28]. The ESCAPE Database addresses these limitations by integrating a wide range of public AMP datasets with more than $80\,000$ peptides obtained from various sources. In addition, we standardize the class annotation system by curating a concise and biologically meaningful hierarchy of antimicrobial functions, thus improving the interpretability of annotations and facilitating dataset integration.

2.2 AMP Benchmarks

Benchmarking is critical in evaluating AMP prediction models, yet standardized protocols for consistent comparison are still missing. Many studies rely on custom datasets and train-test splits without releasing exact partitions, hindering reproducibility and fair comparison against other methods [29]. Furthermore, most benchmarking efforts focus on binary classification tasks that fail to capture the functional diversity and therapeutic relevance of AMPs [30]. Although several studies have introduced multilabel classification approaches for AMPs [15, 31], the field still lacks a standardized, openly accessible benchmark designed to support rigorous evaluation in multilabel antimicrobial peptide prediction. In this context, the ESCAPE Benchmark is a significant advancement by enabling rigorous evaluation of seven state-of-the-art models on the multilabel AMP classification task, thereby facilitating fair and transparent comparisons of AI methodologies on a large-scale peptide dataset.

2.3 AMP Classification Models

There are two main categories of AMP classification methods in the current literature: those that rely exclusively on raw amino acid sequences and those that incorporate bioinformatically derived descriptors using specialized libraries. Sequence-focused methods include AMPlify [11], which employs a bidirectional LSTM with multi-head and context attention to generate a summary vector. Other methods use Large Language Models (LLMs) [32] as the backbone of the architecture. For example, TransImbAMP [31] uses a BERT model pretrained in a self-supervised manner via masked token

prediction and fine-tunes a fully connected layer attached to its outputs for the AMP classification task. AMP-BERT [12], another sequence-based model, also employs a pretrained BERT, but with an inserted class token whose embedding guides the classifier. More recently, dsAMPGAN [33] integrates CNN, Attention, and BiLSTM layers with transfer learning to perform AMP classification and function prediction, while AMPpred-DLFF [34] combines ESM-2 [35] embeddings with graph attention networks and CNN modules to capture both spatial and sequential information.

In contrast, feature-augmented approaches compute additional descriptors before modeling. For instance, amPEPpy [13] feeds CTD features (composition, transition, distribution of physicochemical amino-acid classes) to a Random Forest, and AMPs-Net [15] represents peptides as graphs enriched with physicochemical node and edge attributes to then process the peptides with a GNN. PEP-Net [36] fuses one-hot amino-acid identities, computed physicochemical properties, and high-dimensional protein language model embeddings through residual convolutional and Transformer blocks to capture local information and global contextual information. AVP-IFT [37] employs a dual-branch framework integrating contrastive learning with a transformer network enhanced with biophysical and chemical properties. Recent work has also introduced ensemble-based feature-augmented approaches, including StackAMP [38], AMP-RNNpro [39], and StackPIP [40], which combine diverse peptide descriptors with multiple machine learning models to improve classification performance. Unlike prior methods, the ESCAPE Baseline introduces a bidirectional cross-attention mechanism by integrating the peptide sequence and its 3D distance representation, enabling the combination of spatial structural information with the sequence and leading to superior performance.

3 Expanded Standardized Collection for Antimicrobial Peptide Evaluation

Current research on AMP prediction faces a critical bottleneck due to fragmented, inconsistent, and small-scale datasets that vary widely in format, annotation standards, and functional coverage [17]. These limitations hinder the development of robust predictive models and complicate fair comparisons across methods. To overcome this gap, we introduce the ESCAPE Dataset, a unified collection of antimicrobial peptides compiled from 27 public repositories. This multilabel framework standardizes functional annotations to reflect the biological taxonomy of infectious agents, resulting in five classes: four main AMP activities (antibacterial, antifungal, antiviral and antiparasitic) and a fifth category (antimicrobial) that represents the general ability to act on any microorganism. Peptides that do not exhibit antimicrobial properties, and thus do not belong to any of the five classes, are considered Non-AMPs.

3.1 Data Compilation

To build the ESCAPE Dataset, we collect experimentally validated and manually curated antimicrobial peptide entries from 27 public databases: BIOPEP-UWM Database [41], CPPsite 2.0 [42], CAMPR3 [43], TumorHoPe [44], APD3 [45], SPdb [46], ParaPep [47], CancerPPD [27], BrainPreps [48], Quorumpeps [49], YADAMP [25], LAMP2 [23], Milkampdb [50], DADP [26], AntiTbPdb [51], PeptideDB [52], NeuroPrep [53], SATPdb [24], BioDADPep [54], NeuroPedia [55], DFBP [56], dbAMP 3.0 [21], DRAMP 4.0 [22], AVPdb [57], Hemolytik [58], DBAASP v3 [59], and UniProt [60]. Each source contributes unique peptide profiles across four functional classes: antibacterial, antifungal, antiviral, and antiparasitic. This structure encapsulates key differences in mechanisms of action, such as disarrangement of bacterial membranes [61], inhibition of cell wall biosynthesis [62], and interference with viral assembly [57], among others.

For the integration of Non-AMP samples, we follow the methodology outlined in TransImbAMP [31], focusing on selecting non-antimicrobial sequences from UniProt [60]. We apply stringent exclusion criteria, removing entries associated with keywords such as “membrane,” “toxic,” “secretory,” “defensive,” “antibiotic,” “anticancer,” “antiviral,” or “antifungal”. To expand this set, we incorporate peptides from curated datasets known for their non-antimicrobial functions [54, 47]. This dual strategy of exclusion-based filtering and targeted selection constructs a high-confidence negative set that effectively distinguishes non-AMP sequences, providing a robust contrast for supervised training. Supplementary Material Section A offers specific details on the creation of the dataset, regarding handling the compiled databases and their associated licenses.

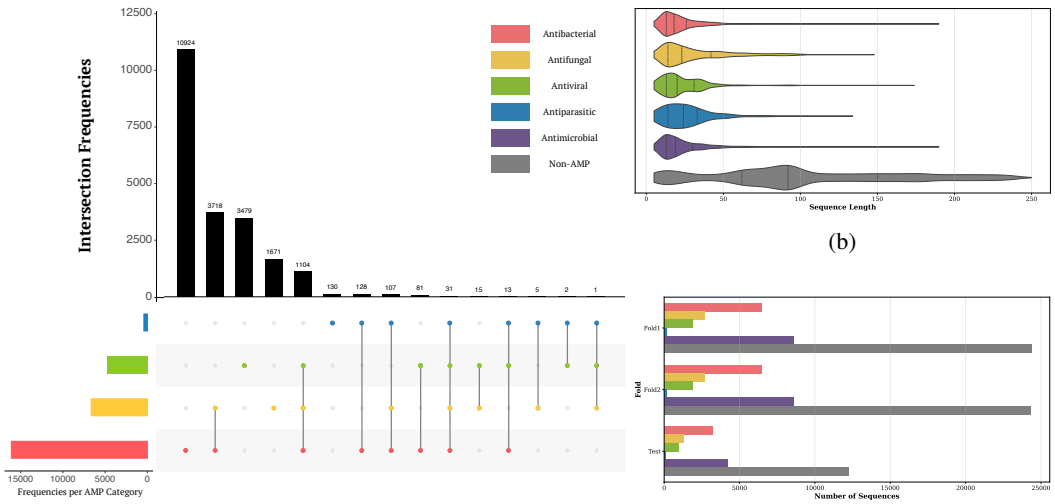

Figure 2: **Overview of ESCAPE Dataset Composition and Statistics.** (a) Multilabel distribution of AMPs across the four functional classes in ESCAPE Dataset, (b) Sequence length distribution for AMPs and non-AMPs, and (c) Distribution of AMP and non-AMP sequences in the two folds and the test set of the dataset.

3.2 Data Processing and Cleaning

We remove sequences containing synthetic residues such as pyrrolysine (O), selenocysteine (U), β-alanine (Bal), 3-naphthylalanine (Nal), or 2-aminobutanoic acid (Abu), following the pipeline that AMP-Net [15] proposes. We exclude entries with undefined amino acids (X) and retain degenerate codes J, B, and Z, treating them as biologically valid representations of leucine/isoleucine, aspartic acid/asparagine, and glutamic acid/glutamine, respectively. We also enforce length constraints, retaining only peptides with lengths between 5 and 250 residues, ensuring structural relevance and alignment with established peptide standards [15]. Finally, to mitigate redundancy, we consolidate duplicate sequences across repositories and integrate their corresponding functional annotations into a unified multilabel vector.

3.3 ESCAPE Dataset

The ESCAPE Dataset comprises 60 950 non-AMPs and 21 409 AMPs, which are functionally annotated into four major antimicrobial categories: antibacterial, antifungal, antiviral, and antiparasitic. Figure 2a shows that most AMPs (16 106) exhibit antibacterial activity, and 10 924 belong exclusively to this class. The most common combination is antibacterial and antifungal sequences (4960 peptides), while only 1671 peptides are solely antifungal. Antiviral entries total 4726 (3479 unique), and antiparasitic peptides number 417 (130 unique). Rarer multilabel groups include primarily antiparasitic sequences exhibiting antifungal or antiviral activity.

Fig. 2b presents the sequence length distribution for AMPs and non-AMPs. AMP sequences are predominantly centered around 30 amino acids, which is characteristic of this type of peptide structure [63]. In contrast, non-AMPs exhibit a broader length range, with an average of 90 amino acids, and most sequences fall between 50 and 100 residues. This distinction highlights the inherent structural differences between AMPs and non-AMP sequences.

The ESCAPE Dataset is divided into three folds: two used for cross-validation and one reserved for independent testing. This partitioning strategy ensures comprehensive model evaluation by maintaining equal label distributions across folds, as shown in Fig. 2c. This consistency supports reliable performance assessment and minimizes overfitting risks between functional classes.

3.4 ESCAPE Benchmark

To evaluate the performance of existing AMP classification methods under the standardized multilabel framework introduced by ESCAPE, we benchmark seven representative models with publicly available implementations that allow reproducibility: AMPlify [11], AMP-BERT [12], TransImbAMP [31], amPEPpy [13], AMPs-Net [15], PEP-Net [36], and AVP-IFT [37]. We adapt each model to support multilabel classification and we train it with a two-fold cross-validation scheme. In Supplementary Section B, we provide per-model implementation details, with the training hyperparameters summarized in Supplementary Table 2. To establish robust and consistent evaluation, we train each method three times with the random seeds (42, 1665, 8914) across all models. We perform the final evaluation with an ensemble that averages the probabilities from the two trained models. We assess performance using mean Average Precision (mAP) and F1-score, two widely used metrics for multilabel settings with severe class imbalance typical of AMP prediction [64]. We report the mean and standard deviation across the three runs.

4 ESCAPE Baseline

Given the limitations of existing models in addressing multiclass classification of antimicrobial peptides, we design a method that combines sequential and structural information to perform the classification task on the ESCAPE Database. We introduce the ESCAPE Baseline model, a transformer-based architecture that integrates sequence and structural modalities through bidirectional cross-attention, and we provide a detailed account of its design and implementation.

4.1 Input Peptide Representation

Sequence Representation. Let $\mathcal{S}$ denote the set of all peptide sequences. Each sequence $s \in \mathcal{S}$ is defined as an ordered list of amino acids $s = [a_1, a_2, \ldots, a_N]$ with N residues, where each $a_i \in \mathcal{A}$. The vocabulary $\mathcal{A}$ consists of 26 amino acid symbols (including rare and ambiguous codes), together with a special padding token, resulting in a vocabulary length of 27 [11]. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_{27}\}$ be a finite set of discrete tokens, with $|\mathcal{T}| = 27$, representing the token vocabulary. We define a bijective mapping $f : \mathcal{A} \to \mathcal{T}$ such that for every $a_i \in \mathcal{A}$ there exists a unique token $t_i = f(a_i) \in \mathcal{T}$. This one-to-one correspondence allows each sequence s to be equivalently represented as a token sequence $t = [t_1, t_2, \ldots, t_{\mathcal{L}}]$ over the vocabulary $\mathcal{T}$. All sequences are either truncated or zero-padded to a fixed length $\mathcal{L}$.

Structural Representation. For each peptide, structural information is collected from the UniProt [60] and Protein Data Bank (PDB) [65] repositories. When experimental structures are unavailable, we use RosettaFold [66] and AlphaFold3 [67] to predict the three-dimensional conformations. These models are state-of-the-art deep learning methods for protein structure prediction. Given a peptide with N amino acids, we compute a distance matrix $\mathcal{M} \in \mathbb{R}^{N \times N}$, based on the 3D structure. Each element $\mathcal{M}_{i,j}$ corresponds to the Euclidean distance between the Cα atoms of residues i and j: $\mathcal{M}_{i,j} = \|\mathbf{r}_i - \mathbf{r}_j\|$, where $\mathbf{r}_{i,j} \in \mathbb{R}^3$ denotes the spatial coordinates of the i^{th}, j^{th} residue. To ensure compatibility across peptides of varying lengths, $\mathcal{M}$ is resized to a fixed dimension of 224×224, enabling uniform input to the structural encoder.

4.2 Model Architecture

The ESCAPE Baseline model is built upon a dual-branch transformer architecture designed to jointly encode the sequence and structural modalities of peptides, as illustrated in Fig. 3. Each branch independently processes one modality using a specific transformer encoder, and the resulting representations are fused via a bidirectional cross-attention mechanism to enable cross-domain interaction.

Sequence Module. The sequence branch takes the tokenized peptide sequence and maps each token to a 256-dimensional vector using a learnable embedding matrix. In addition, we include a special [CLS] token to capture global sequence-level information and add positional embeddings to preserve the order of amino acids. Later, this feeds the resulting embeddings through a stack of $\mathcal{D}$ Transformer encoder layers. This setup enables the model to learn local and long-range dependencies within the

sequence. The [CLS] token at the output serves as a compact representation of the peptide's primary structure.

Structure Module. The structural branch receives as input a single-channel 224×224 distance matrix $\mathcal{M}$, where each entry indicates the Euclidean distance between a pair of $C\alpha$ atoms in the peptide's 3D structure. A 2D convolution with kernel size and stride of 16 partitions this matrix into non-overlapping 16×16 patches, producing a grid of flattened patches. The model projects each patch into a 192-dimensional embedding, creating a sequence of patch embeddings. The model adds a learnable [CLS] token at the beginning of the sequence to aggregate global structural information and appends fixed positional encodings to preserve spatial relationships. The model processes the sequence through a stack of $\mathcal{D}$ Transformer encoder layers that capture local and long-range spatial dependencies. The structure branch outputs the [CLS] token, which captures a compact representation of the peptide's 3D conformation.

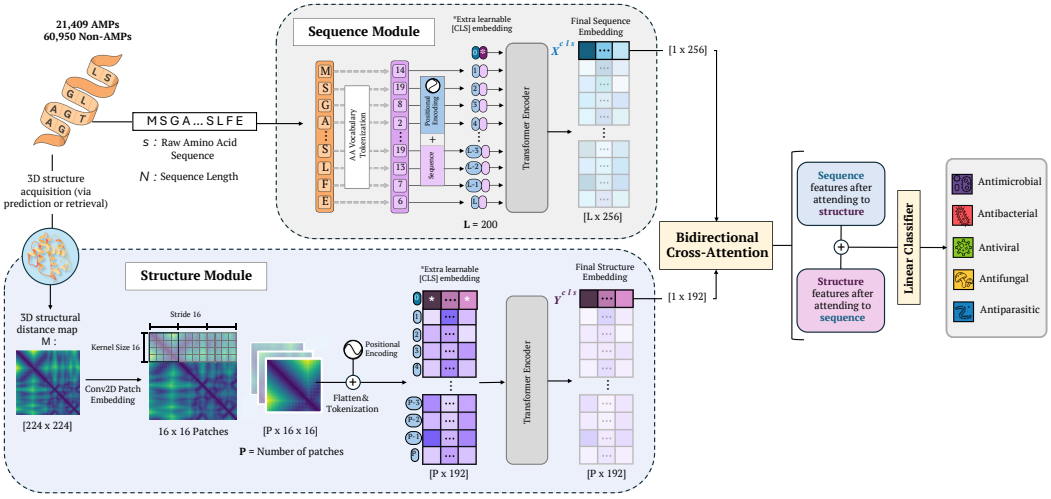

Figure 3: **ESCAPE Baseline Architecture Overview.** The model encodes each peptide using two parallel branches: the sequence module tokenizes amino acid residues. It extracts a [CLS] representation through a Transformer encoder. In contrast, the structure module processes a 224×224 distance matrix by embedding non-overlapping patches and applying a Transformer stack to produce a structural [CLS] token. A bidirectional cross-attention mechanism fuses these two representations by allowing each modality to attend to the other. The model concatenates the resulting attended CLS vectors and passes them through a linear layer to generate the final multilabel prediction vector.

Bidirectional Cross-Attention. We apply a bidirectional cross-attention mechanism to integrate information from sequence and structure modalities after independently encoding each branch. As detailed in the Sequence and Structure Modules, our model processes the amino acid sequence and the distance matrix separately through their transformer encoders. Each encoder produces a contextual embedding matrix and a corresponding [CLS] token that summarizes global information.

Let $\mathbf{X} \in \mathbb{R}^{\mathcal{L} \times 256}$ and $\mathbf{Y} \in \mathbb{R}^{\mathcal{P} \times 192}$ denote the sequence and structure embedding matrices, respectively, where $\mathcal{L}$ and $\mathcal{P}$ are the modality-specific lengths, including the [CLS] token. The corresponding [CLS] embeddings are denoted as $\mathbf{X}_{\text{cls}}$ and $\mathbf{Y}_{\text{cls}}$.

To enable cross-modal interaction, we apply attention in both directions. First, the sequence attends to the structural features where $\mathbf{Q}_x = \mathbf{X}\mathbf{W}_Q^x$ are the queries derived from the sequence, and $\mathbf{K}_y = \mathbf{Y}\mathbf{W}_K^y$, $\mathbf{V}_y = \mathbf{Y}\mathbf{W}_V^y$ are the keys and values projected from the structure branch. The attention output $\mathbf{A}_x$ is added to the original sequence embeddings via a residual connection and refined through a feedforward network. Conversely, the structural representation attends to the sequence where $\mathbf{Q}_y = \mathbf{Y}\mathbf{W}_Q^y$, and $\mathbf{K}_x = \mathbf{X}\mathbf{W}_K^x$, $\mathbf{V}_x = \mathbf{X}\mathbf{W}_V^x$ are the projections from the sequence encoder.

$$\mathbf{A}_x = \text{softmax}\left(\frac{\mathbf{Q}_x\mathbf{K}_y^\top}{\sqrt{d}}\right)\mathbf{V}_y, \qquad \mathbf{A}_y = \text{softmax}\left(\frac{\mathbf{Q}_y\mathbf{K}_x^\top}{\sqrt{d}}\right)\mathbf{V}_x,$$

The bidirectional attention mechanism enables each modality to gather contextual information from the other by focusing on the most informative regions of the complementary representation. After bidirectional cross-attention, we concatenate the updated [CLS] from both modalities and pass the resulting vector through a classification head to produce the final prediction.

4.3 Implementation Details

We train the ESCAPE Baseline on a NVIDIA GPU A100 with a batch size 64 for 100 epochs, using the AdamW optimizer and a learning rate of 1×10^{-4}. We incorporate dropout layers within each transformer encoder block to mitigate overfitting and enhance generalization. In the sequence branch, we tokenize each input peptide and pad it to a fixed length of $\mathcal{S} = 200$, embedding each amino acid into a 256-dimensional space. The sequence encoder comprises 4 transformer layers, each with 8 attention heads. In the structural branch, we represent each peptide's distance matrix as a 224×224 single-channel image and divide it into non-overlapping 16×16 patches. We flatten each patch and project it into a 192-dimensional embedding. We then process the resulting patch embeddings with a transformer encoder that shares the same configuration as the sequence branch, consisting of 4 layers and 8 attention heads, to capture spatial and geometric dependencies within the peptide structure.

5 Experiments and Discussion

5.1 Main Results

Table 1 reports the mean and standard deviations over the three random seeds of the overall and per-class F1-scores, while Table 2 presents the corresponding mAP values. The results demonstrate that our baseline consistently surpasses all seven state-of-the-art methods across both metrics, reaching relative improvements of 1.90% in overall F1-score and 2.56% in mAP compared to the second-best method.

The per-class evaluation reveals that ESCAPE Baseline achieves the most substantial gains in the least represented categories. In particular, it increases the AP for the antiparasitic class by 35.7% relative to the second-best method. Among sequence-based models, AMPlify attains the highest performance, with an average F1-score of 68.5% and mAP of 70.3%. AVP-IFT follows closely, achieving 66.5% F1 and 68.8% mAP by integrating physicochemical descriptors through a feature-augmented design. These comparisons indicate that no single modeling strategy, whether sequence-focused or feature-augmented, consistently dominates the others.

Table 1: **Overall and Per-Class F1-Scores on the ESCAPE Benchmark.** F1-scores for each model averaged over the 42, 1665, and 8914 random seeds on the 5-class multilabel classification task in the ESCAPE Benchmark (%).

Method	F1-Score	Antibacterial	Antiviral	Antifungal	Antiparasitic	Antimicrobial
AMPs-Net [15]	57.7 ± 0.70	78.9 ± 0.77	59.2 ± 0.79	61.1 ± 0.51	5.9 ± 0.71	83.5 ± 0.79
TranslmbAMP [31]	62.0 ± 0.70	87.1 ± 0.96	59.2 ± 0.50	54.7 ± 0.51	21.8 ± 0.81	87.2 ± 0.75
AMP-BERT [12]	64.7 ± 0.64	89.3 ± 0.27	63.0 ± 0.95	60.2 ± 0.26	20.6 ± 3.52	90.5 ± 0.22
PEP-Net [36]	65.5 ± 0.61	$\mathbf{89.5 \pm 0.10}$	58.1 ± 0.78	$\mathbf{65.2 \pm 0.55}$	22.8 ± 0.61	$\mathbf{91.2 \pm 0.15}$
amPEPpy [13]	66.5 ± 0.37	87.6 ± 0.07	61.6 ± 2.02	60.4 ± 1.90	34.7 ± 0.98	90.9 ± 3.78
AVP-IFT [37]	66.5 ± 0.59	89.1 ± 0.47	$\mathbf{64.8 \pm 0.06}$	60.7 ± 0.55	28.0 ± 3.84	89.9 ± 0.31
AMPlify [11]	68.5 ± 0.77	88.8 ± 0.26	60.0 ± 1.05	65.0 ± 1.57	40.9 ± 2.48	90.0 ± 0.30
ESCAPE Baseline (Ours)	$\mathbf{69.8 \pm 0.43}$	88.8 ± 0.34	64.4 ± 0.88	61.0 ± 0.75	$\mathbf{44.8 \pm 0.50}$	90.0 ± 0.32

Overall, performance trends highlight that the effectiveness of the model depends on the synergy between input representations and architectural design, rather than on the mere inclusion of additional descriptors. By combining sequence and structural modalities, ESCAPE Baseline leverages richer biological representations to improve generalization. The architecture also supports flexible operation under different configurations using only sequence information, only 3D structural data, or both, achieving its best performance when integrating the two. This versatility comes from its higher capacity and ability to align complementary modalities during training, establishing ESCAPE Baseline as both a robust and adaptable framework for multilabel AMP classification.

Furthermore, as shown in Figure 2 of the Supplementary Material, model size does not exhibit a consistent relationship with predictive performance across the evaluated methods. Notably, although

the top-performing model has nearly 9 million parameters, the second-best model overall is based on a Random Forest classifier, making it the least computationally demanding method in our benchmark. Conversely, as Table 1 and Table 2 report, BERT-based approaches do not rank among the top three performing models. These findings suggest that more complex architectures do not necessarily yield superior results for multilabel AMP classification. Additionally, our results underscore the limitations of large language models when applied to domains outside of natural language. Despite the domain adaptation efforts through BERT fine-tuning in models such as TransImbAMP and AMP-BERT, these methods fail to fully accommodate the peptide-specific data and, as a result, underperform in this task.

Table 2: **Mean and Per-Class AP Results on the ESCAPE Benchmark.** AP for each model averaged over the 42, 1665, and 8914 random seeds on the 5-class multilabel classification task in the ESCAPE Benchmark (%).

Method	mAP	Antibacterial	Antiviral	Antifungal	Antiparasitic	Antimicrobial
AMPs-Net [15]	54.6 ± 0.86	82.5 ± 0.72	51.2 ± 0.88	53.1 ± 0.84	5.3 ± 0.67	82.1 ± 0.80
TransImbAMP [31]	64.9 ± 1.11	92.5 ± 1.23	65.0 ± 1.63	56.3 ± 0.96	16.7 ± 0.86	94.0 ± 0.90
AMP-BERT [12]	66.9 ± 1.17	92.3 ± 0.59	65.9 ± 1.84	61.5 ± 2.28	21.4 ± 2.61	93.6 ± 1.25
amPEPpy [13]	68.5 ± 0.48	93.9 ± 0.24	67.7 ± 0.28	62.2 ± 0.27	23.8 ± 1.61	95.2 ± 0.05
PEP-Net [36]	68.4 ± 0.53	$\mathbf{95.2 \pm 0.21}$	61.2 ± 0.67	$\mathbf{72.6 \pm 0.78}$	16.2 ± 0.84	$\mathbf{96.7 \pm 0.26}$
AVP-IFT [37]	68.8 ± 0.50	94.3 ± 0.49	$\mathbf{71.1 \pm 0.36}$	63.3 ± 1.36	20.0 ± 4.25	95.5 ± 0.50
AMPlify [11]	70.3 ± 0.87	94.0 ± 0.19	66.1 ± 5.56	68.3 ± 4.27	27.7 ± 1.33	95.3 ± 0.31
ESCAPE Baseline (Ours)	$\mathbf{72.1 \pm 0.60}$	94.2 ± 0.21	69.8 ± 0.46	63.4 ± 0.74	$\mathbf{37.6 \pm 2.87}$	95.6 ± 0.04

Per-class analysis reveals substantial variation in predictive performance across functional categories. As the number of samples per class decreases, most models exhibit a proportional decline in both mAP and F1-score, independent of their underlying architecture or feature design. The antiparasitic and antiviral classes, which contain the fewest examples, yield the lowest scores across all evaluated methods, highlighting the intrinsic difficulty of learning under severe data scarcity. In contrast, categories with broader representation, such as antibacterial and antifungal, display more stable results and narrower variability among models. This trend underscores the impact of label imbalance as the dominant factor shaping overall performance, suggesting that future improvements should focus on better representation learning rather than on increasing model complexity alone.

5.2 Ablation Experiments

To assess the individual contribution of each peptide representation modality in our baseline, we conduct an ablation experiment using the 42 seed. We evaluate three configurations: one using only the sequence module, another using only the structural module based on distance matrices, and a third combining both through the cross-attention module. Table 3 shows that the sequence representation provides considerably more informative features than the structural one: with sequence-only input, our model achieves 21.7% higher mAP and 20.7% higher F1 score than with the distance matrix alone. This gap likely arises because the structural view captures spatial arrangement but omits explicit biochemical identities, thereby limiting the model's ability to exploit residue-level patterns critical for antimicrobial activity. These findings highlight that the biological composition of the peptide encoded in the sequence plays a decisive role in classification performance. Yet Table 3 also shows that combining both representations via cross-attention yields the best overall results: while the sequence-only variant is already strong, adding structural cues provides a complementary signal that further improves prediction quality.

Table 3: **ESCAPE Baseline Ablation Experiments.** mAP (%) and Overall F1-score (%) reported for the 42 random seed trained model.

Structure Module	Sequence Module	Cross Attention	mAP	F1
✓	·	·	47.7	46.9
·	✓	·	69.4	67.6
✓	✓	✓	**72.7**	**69.5**

As detailed in Section 4, we obtain structural information for most peptides from UniProt and PDB, and infer the remaining structures using public generative models [66, 67]. We also run a sensitivity experiment to assess the impact of using predicted structures. Supplementary Section C

(Table 5) shows that relying solely on predicted structures leads to reduced performance compared to experimental structures, with absolute drops of 1.5% in mAP and 1.9% in F1. These results suggest that artificially generated structures may introduce additional sources of error inherent to those models, potentially degrading the quality of the structural data and impairing the model's ability to accurately classify peptides.

5.3 Limitations and Broader Impact

In this work, we present a carefully curated and extensive dataset for AMP discovery. An important limitation arises from the scope of the domain, as the diversity of peptides in nature is vast and it is not feasible to capture all existing variants or ensure that our dataset fully represents the underlying distribution. Nonetheless, our benchmark provides a standardized and transparent framework for evaluating AMP classification models, helping to identify methodological gaps and guide future improvements. By promoting reproducibility and comparative analysis in this area, our work contributes to advancing computational tools for AMP discovery, which may support efforts in global health, particularly in the context of antibiotic resistance. However, further validation in real-world biological and clinical settings is required before deployment of such models.

Another limitation arises from the inherent differences in sequence length distributions. Antimicrobial peptides are naturally shorter than most non-AMPs, a characteristic tied to their biological function. In constructing ESCAPE, we aim to preserve this molecular distinction while avoiding strong correlations that could bias classification. The resulting dataset maintains realistic differences between classes without allowing sequence length to dominate predictive performance, supporting a fairer evaluation of computational models.

5.4 Ethical Considerations

From an ethical standpoint, ESCAPE is constructed entirely from publicly available, experimentally validated datasets, each distributed under its respective license. While the benchmark provides a foundation for advancing computational methods in antimicrobial peptide research, our contribution remains focused on methodological innovation rather than direct therapeutic or drug design applications.

At the same time, we acknowledge the potential risks arising from irresponsible use of this resource. Models trained on ESCAPE could, if misused, be employed to generate peptides without appropriate experimental validation, raising concerns about toxicity or biosecurity. To mitigate such risks, we encourage the research community to operate within established ethical, biosafety, and regulatory standards and to ensure that all experimental and computational findings are reported transparently and with accountability.

6 Conclusions

We introduce **ESCAPE**, the first standardized benchmark for multilabel antimicrobial peptide classification, designed to overcome key limitations of existing resources, including data fragmentation, inconsistent annotations, and limited functional scope. ESCAPE integrates over 80.000 peptides from 27 curated repositories into a biologically grounded multilabel framework encompassing antibacterial, antifungal, antiviral, antiparasitic and antimicrobial classes. It also includes a rigorously filtered set of non-antimicrobial sequences to support reliable supervised training and better reflect real-world prediction settings. Building upon this foundation, we propose a baseline using a transformer-based architecture that leverages sequence and structural information. We demonstrate that ESCAPE Benchmark enables fair and reproducible comparison across models and functional classes, setting a new standard for AI-driven AMP discovery, particularly in underrepresented categories such as antiviral and antiparasitic.

7 Acknowledgements

This research was partially funded by the Colombian Ministry of Science, Technology, and Innovation (Minciencias), under Cod. 1204-937-101846, CR 19576-2024 Call for Fundamental Research. This work was supported by Azure sponsorship credits granted by Microsoft's AI for Good Research Lab.

References

[1] World Health Organization. Global research agenda for antimicrobial resistance in human health. 2024.

[2] Amit K Mittal, Rohit Bhardwaj, Priya Mishra, and Satyendra K Rajput. Antimicrobials misuse/overuse: adverse effect, mechanism, challenges and strategies to combat resistance. *The Open Biotechnology Journal*, 14(1), 2020.

[3] Priyanka Chambial, Neelam Thakur, Prudhvi Lal Bhukya, Anbazhagan Subbaiyan, and Umesh Kumar. Frontiers in superbug management: innovating approaches to combat antimicrobial resistance. *Archives of Microbiology*, 207(3):60, 2025.

[4] Mohsen Naghavi, Stein Emil Vollset, Kevin S Ikuta, Lucien R Swetschinski, Authia P Gray, Eve E Wool, Gisela Robles Aguilar, Tomislav Mestrovic, Georgia Smith, Chieh Han, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, 404(10459):1199–1226, 2024.

[5] Mark Elsner, Grace Atkinson, and Saadia Zahidi. Global risks report 2025. Technical report, World Economic Forum, Geneva, January 2025.

[6] United Nations Environment Programme. What is fuelling the world's antimicrobial resistance crisis?, November 2023. Accessed: 2025-05-12.

[7] Jiaqi Xuan, Weiguo Feng, Jiaye Wang, Ruichen Wang, Bowen Zhang, Letao Bo, Zhe-Sheng Chen, Hui Yang, and Leming Sun. Antimicrobial peptides for combating drug-resistant bacterial infections. *Drug Resistance Updates*, 68:100954, 2023.

[8] Tarequl Islam, Noshin Tabassum Tamanna, Md Shahjalal Sagor, Randa Mohammed Zaki, Muhammad Fazle Rabbee, and Maximilian Lackner. Antimicrobial peptides: A promising solution to the rising threat of antibiotic resistance. *Pharmaceutics*, 16(12):1542, 2024.

[9] Mohamad Hamad, Farah Al-Marzooq, Gorka Orive, and Taleb H Al-Tel. Superbugs but no drugs: steps in averting a post-antibiotic era, 2019.

[10] Paulina Szymczak and Ewa Szczurek. Artificial intelligence-driven antimicrobial peptide discovery. *Current Opinion in Structural Biology*, 83:102733, 2023.

[11] Chenkai Li, Darcy Sutherland, S Austin Hammond, Chen Yang, Figali Taho, Lauren Bergman, Simon Houston, René L Warren, Titus Wong, Linda MN Hoang, et al. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. *BMC genomics*, 23(1):77, 2022.

[12] Hansol Lee, Songyeon Lee, Ingoo Lee, and Hojung Nam. Amp-bert: Prediction of antimicrobial peptide function based on a bert model. *Protein Science*, 32(1):e4529, 2023.

[13] Travis J Lawrence, Dana L Carper, Margaret K Spangler, Alyssa A Carrell, Tomás A Rush, Stephen J Minter, David J Weston, and Jessy L Labbé. ampeppy 1.0: a portable and accurate antimicrobial peptide prediction tool. *Bioinformatics*, 37(14):2058–2060, 2021.

[14] Kamila Botelho Sampaio de Oliveira, Michel Lopes Leite, Victor Albuquerque Cunha, Nicolau Brito da Cunha, and Octávio Luiz Franco. Challenges and advances in antimicrobial peptide development. *Drug Discovery Today*, 28(8):103629, 2023.

[15] Paola Ruiz Puentes, Maria C Henao, Javier Cifuentes, Carolina Muñoz-Camargo, Luis H Reyes, Juan C Cruz, and Pablo Arbeláez. Rational discovery of antimicrobial peptides by means of artificial intelligence. *Membranes*, 12(7):708, 2022.

[16] Markus Orsi and Jean-Louis Reymond. Can large language models predict antimicrobial peptide activity and toxicity? *RSC Medicinal Chemistry*, 15(6):2030–2036, 2024.

[17] Jinzhen Yan and Zhang B Wang Y Wong D F Siu S W I Cai, J. Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning. *Antibiotics*, 11(10):1451, 2022.

[18] Daniel Veltri, Uday Kamath, and Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740–2747, 03 2018.

[19] Guangshun Wang, Iosif I Vaisman, and Monique L van Hoek. Machine learning prediction of antimicrobial peptides. In *Computational Peptide Science: Methods and Protocols*, pages 1–37. Springer, 2022.

[20] Nicolas Aparicio Claros, Paula Cárdenas, Juanita Puentes, Paola Ruiz Puentes, and Pablo Arbeláez. Computational tools for handling large databases of biological relevance. In *Antimicrobial Peptides*, pages 81–96. Elsevier, 2025.

[21] Lantian Yao, Jiahui Guan, Peilin Xie, Chia-Ru Chung, Zhihao Zhao, Danhong Dong, Yilin Guo, Wenyang Zhang, Junyang Deng, Yuxuan Pang, et al. dbamp 3.0: updated resource of antimicrobial activity and structural annotation of peptides in the post-pandemic era. *Nucleic Acids Research*, 53(D1):D364–D376, 2025.

[22] Tianyue Ma, Yanchao Liu, Bingxin Yu, Xin Sun, Huiyuan Yao, Chen Hao, Jianhui Li, Maryam Nawaz, Xun Jiang, Xingzhen Lao, et al. Dramp 4.0: an open-access data repository dedicated to the clinical translation of antimicrobial peptides. *Nucleic Acids Research*, 53(D1):D403–D410, 2025.

[23] Xiaowei Zhao, Hongyu Wu, Hairong Lu, Guodong Li, and Qingshan Huang. Lamp: a database linking antimicrobial peptides. *PloS one*, 8(6):e66557, 2013.

[24] Sandeep Singh, Kumardeep Chaudhary, Sandeep Kumar Dhanda, Sherry Bhalla, Salman Sadullah Usmani, Ankur Gautam, Abhishek Tuknait, Piyush Agrawal, Deepika Mathur, and Gajendra PS Raghava. Satpdb: a database of structurally annotated therapeutic peptides. *Nucleic acids research*, 44(D1):D1119–D1126, 2016.

[25] Stefano P Piotto, Lucia Sessa, Simona Concilio, and Pio Iannelli. Yadamp: yet another database of antimicrobial peptides. *International journal of antimicrobial agents*, 39(4):346–351, 2012.

[26] Mario Novković, Juraj Simunić, Viktor Bojović, Alessandro Tossi, and Davor Juretić. Dadp: the database of anuran defense peptides. *Bioinformatics*, 28(10):1406–1407, 2012.

[27] Atul Tyagi, Abhishek Tuknait, Priya Anand, Sudheer Gupta, Minakshi Sharma, Deepika Mathur, Anshika Joshi, Sandeep Singh, Ankur Gautam, and Gajendra PS Raghava. Cancerppd: a database of anticancer peptides and proteins. *Nucleic acids research*, 43(D1):D837–D843, 2015.

[28] Ahmad M Al-Omari, Yazan H Akkam, Ala'a Zyout, Shayma'a Younis, Shefa M Tawalbeh, Khaled Al-Sawalmeh, Amjed Al Fahoum, and Jonathan Arnold. Accelerating antimicrobial peptide design: Leveraging deep learning for rapid discovery. *Plos one*, 19(12):e0315477, 2024.

[29] Katarzyna Sidorczuk, Przemysław Gagat, Filip Pietluch, Jakub Kała, Dominik Rafacz, Laura Bąkała, Jadwiga Słowik, Rafał Kolenda, Stefan Roediger, Legana CHW Fingerhut, et al. The impact of negative data sampling on antimicrobial peptide prediction. *bioRxiv*, pages 2022–05, 2022.

[30] Qi-Yu Zhang, Zhi-Bin Yan, Yue-Ming Meng, Xiang-Yu Hong, Gang Shao, Jun-Jie Ma, Xu-Rui Cheng, Jun Liu, Jian Kang, and Cai-Yun Fu. Antimicrobial peptides: mechanism of action, activity and clinical potential. *Military Medical Research*, 8:1–25, 2021.

[31] Yuxuan Pang, Lantian Yao, Jingyi Xu, Zhuo Wang, and Tzong-Yi Lee. Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities. *Bioinformatics*, 38(24):5368–5374, 2022.

[32] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

[33] Min Zhao, Yu Zhang, Maolin Wang, and Luyan Z Ma. dsamp and dsampgan: deep learning networks for antimicrobial peptides recognition and generation. *Antibiotics*, 13(10):948, 2024.

[34] Yu Chen, Xingpeng Jiang, and Weizhong Zhao. Amppred-dlff: prediction of amps based on deep learning and multi-view features fusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 891–896. IEEE, 2024.

[35] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[36] Jiyun Han, Tongxin Kong, and Juntao Liu. Pepnet: an interpretable neural network for anti-inflammatory and antimicrobial peptides prediction using a pre-trained protein language model. *Communications Biology*, 7(1):1198, 2024.

[37] Jiahui Guan, Lantian Yao, Peilin Xie, Chia-Ru Chung, Yixian Huang, Ying-Chih Chiang, and Tzong-Yi Lee. A two-stage computational framework for identifying antiviral peptides and their functional types based on contrastive learning and multi-feature fusion strategy. *Briefings in Bioinformatics*, 25(3):bbae208, 05 2024.

[38] Tasmin Karim, Md Shazzad Hossain Shaon, Md Mamun Ali, Kawsar Ahmed, Francis M Bui, and Li Chen. Stackamp: stacking-based ensemble classifier for antimicrobial peptide identification. *IEEE Transactions on Artificial Intelligence*, 5(11):5666–5675, 2024.

[39] Md Shazzad Hossain Shaon, Tasmin Karim, Md Fahim Sultan, Md Mamun Ali, Kawsar Ahmed, Md Zahid Hasan, Ahmed Moustafa, Francis M Bui, and Fahad Ahmed Al-Zahrani. Amp-rnnpro: a two-stage approach for identification of antimicrobials using probabilistic features. *Scientific Reports*, 14(1):12892, 2024.

[40] Lantian Yao, Feng Wang, Peilin Xie, Jiahui Guan, Zhihao Zhao, Xuxin He, Xingchen Liu, Ying-Chih Chiang, and Tzong-Yi Lee. Stackpip: An effective computational framework for accurate and balanced identification of proinflammatory peptides. *Journal of Chemical Information and Modeling*, 65(14):7777–7788, 2025.

[41] Piotr Minkiewicz, Anna Iwaniak, and Małgorzata Darewicz. Biopep-uwm database of bioactive peptides: Current opportunities. *International journal of molecular sciences*, 20(23):5978, 2019.

[42] Piyush Agrawal, Sherry Bhalla, Salman Sadullah Usmani, Sandeep Singh, Kumardeep Chaudhary, Gajendra PS Raghava, and Ankur Gautam. Cppsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic acids research*, 44(D1):D1098–D1103, 2016.

[43] Faiza Hanif Waghu, Ram Shankar Barai, Pratima Gurung, and Susan Idicula-Thomas. Campr3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic acids research*, 44(D1):D1094–D1097, 2016.

[44] Pallavi Kapoor, Harinder Singh, Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, and Gajendra PS Raghava. Tumorhope: a database of tumor homing peptides. *PloS one*, 7(4):e35187, 2012.

[45] Guangshun Wang, Xia Li, and Zhe Wang. Apd3: the antimicrobial peptide database as a tool for research and education. *Nucleic acids research*, 44(D1):D1087–D1093, 2016.

[46] Khar Heng Choo, Tin Wee Tan, and Shoba Ranganathan. Spdb–a signal peptide database. *BMC bioinformatics*, 6:1–8, 2005.

[47] Jette Pretzel, Franziska Mohring, Stefan Rahlfs, and Katja Becker. Antiparasitic peptides. *Yellow Biotechnology I: Insect Biotechnologie in Drug Discovery and Preclinical Research*, pages 157–192, 2013.

[48] Sylvia Van Dorpe, Antoon Bronselaer, Joachim Nielandt, Sofie Stalmans, Evelien Wynendaele, Kurt Audenaert, Christophe Van De Wiele, Christian Burvenich, Kathelijne Peremans, Hung Hsuchou, et al. Brainpeps: the blood–brain barrier peptide database. *Brain Structure and Function*, 217:687–718, 2012.

[49] Evelien Wynendaele, Antoon Bronselaer, Joachim Nielandt, Matthias D'Hondt, Sofie Stalmans, Nathalie Bracke, Frederick Verbeke, Christophe Van De Wiele, Guy De Tré, and Bart De Spiegeleer. Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic acids research*, 41(D1):D655–D659, 2013.

[50] Jérémie Théolier, Ismail Fliss, Julie Jean, and Riadh Hammami. Milkamp: a comprehensive database of antimicrobial peptides of dairy origin. *Dairy Science & Technology*, 94:181–193, 2014.

[51] Salman Sadullah Usmani, Rajesh Kumar, Vinod Kumar, Sandeep Singh, and Gajendra PS Raghava. Antitbpdb: a knowledgebase of anti-tubercular peptides. *Database*, 2018:bay025, 2018.

[52] Data Analysis & Modeling Group at Hasselt University and Functional Genomics and Proteomics Unit at K.U. Leuven. PeptideDB: Bioactive Peptide Database. `http://www.peptides.be/?p=contact`, 2022. Leuven, Belgium. Accessed on 13 June 2022.

[53] Yan Wang, Mingxia Wang, Sanwen Yin, Richard Jang, Jian Wang, Zhidong Xue, and Tao Xu. Neuropep: a comprehensive resource of neuropeptides. *Database*, 2015:bav038, 2015.

[54] Susanta Roy and Robindra Teron. Biodadpep: A bioinformatics database for anti diabetic peptides. *Bioinformation*, 15(11):780, 2019.

[55] Yoona Kim, Steven Bark, Vivian Hook, and Nuno Bandeira. Neuropedia: neuropeptide database and spectral library. *Bioinformatics*, 27(19):2772–2773, 2011.

[56] Dongya Qin, Weichen Bo, Xin Zheng, Youjin Hao, Bo Li, Jie Zheng, and Guizhao Liang. Dfbp: a comprehensive database of food-derived bioactive peptides for peptidomics research. *Bioinformatics*, 38(12):3275–3280, 2022.

[57] Abid Qureshi, Nishant Thakur, Himani Tandon, and Manoj Kumar. Avpdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic acids research*, 42(D1):D1147–D1153, 2014.

[58] Ankur Gautam, Kumardeep Chaudhary, Sandeep Singh, Anshika Joshi, Priya Anand, Abhishek Tuknait, Deepika Mathur, Grish C Varshney, and Gajendra PS Raghava. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic acids research*, 42(D1):D444–D449, 2014.

[59] Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research*, 49(D1):D288–D297, 2021.

[60] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[61] Karl A Brogden. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature Reviews Microbiology*, 3(3):238–250, 2005.

[62] Margit Mahlapuu, Jonas Håkansson, Lovisa Ringstad, and Carina Björn. Antimicrobial peptides: an emerging category of therapeutic agents. *Frontiers in Cellular and Infection Microbiology*, 6:194, 2016.

[63] Xu Ma, Qiang Wang, Kexin Ren, Tongtong Xu, Zigang Zhang, Meijuan Xu, Zhiming Rao, and Xian Zhang. A review of antimicrobial peptides: Structure, mechanism of action, and molecular optimization strategies. *Fermentation*, 10(11), 2024.

[64] Paola Ruiz Puentes, Nicolas Aparicio Claros, and Pablo Arbeláez. Artificial intelligence for the discovery of antimicrobial peptides. In *Antimicrobial Peptides*, pages 59–79. Elsevier, 2025.

[65] Protein Data Bank. Protein data bank. *Nature New Biol*, 233(223):10–1038, 1971.

[66] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[67] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately articulate the dataset and benchmark proposed in Section 3 and then the baseline model proposed in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we address the limitations of our work in Section 5.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results that require formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.3 details the procedure to create the proposed dataset with clearly defined partitions that strictly follow the experimental protocol. For complete reproducibility, the Supplementary Material provides full details. Section 4 describes the training procedure and the information needed to reproduce baseline results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:The ESCAPE Dataset is already available here, and the source code for ESCAPE Baseline is already available here.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.3 details the hyperparameters and type of optimizer for our model baseline. We include these details for the other models in the Supplementary Material. Section 3.4 describes the data splits used for all models. These sections ensure both reproducibility and interpretability of the reported results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The Supplementary Material provides the results with error bars for each fold, ensuring a complete understanding of the statistical significance. The main results in Section 5.1 include error bars as we report the ensemble from the 2 cross-folds for each model in the benchmark across 3 random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational resources used for the experiments are detailed in Section 4.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: ESCAPE uses publicly available data, follows community adopted standards for dataset curation, and does not involve human subjects or sensitive personal information. All contributions comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We address potential societal impacts in Section 5.3. The paper highlights positive impacts by advancing antimicrobial peptide discovery to support public health efforts against antibiotic resistance and notes the need for further validation. However, all results derived from this benchmark require subsequent experimental validation before any practical or societal application, substantially reducing the risk of negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve models or data with a high risk of misuse. The dataset consists of curated peptide sequences from licensed public sources, and we cite properly all datasets and comply with its usage terms.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite all employed methods from other authors and respect the licenses of the datasets that contributed to the creation of the main dataset. The Supplementary Material provides more information about these licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not include new assets. Section 3.1 shows all the datasets from which this benchmark was compiled.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any experiments with human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs as a core component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.