SPECIES AND TOPIC-AWARE REPRESENTATION LEARNING FOR ANTIMICROBIAL PEPTIDE ACTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Antimicrobial peptides (AMPs) play a crucial role in addressing the growing threat of antimicrobial resistance. However, their activity varies markedly across microbial targets, indicating that primary peptide sequence data alone cannot fully explain antimicrobial function. To overcome this limitation, we present STAMP: a Species- and Topic-aware representation-learning framework that models AMP activity by integrating sequence, motif, and biological context. STAMP unifies contextual protein language model representations, latent motif composition, and microbial species information within a single model, addressing data sparsity by enabling cross-species training. We evaluated STAMP on four benchmark datasets curated from DBAASP, dbAMP, and DRAMP. As a single cross-species model, STAMP matches methods that rely on tens of independently trained species-specific models and outperforms the most recently published benchmark, reaching a maximum Pearson correlation coefficient (PCC) of 0.837 and an R^2 of 0.698. To demonstrate its utility for candidate prioritization, we applied STAMP to score a large library of computationally generated peptides. Among the 18 candidates selected for experimental validation, 16 exhibited experimentally measured MIC values below 10 μM against *E. coli*, while 17 exhibited MIC values below 10 μM against *S. epidermidis*, providing experimental support for the predicted antimicrobial activity. Additionally, experiments with Tilapia Piscidin 4 (TP4)-derived peptides and residue-level attribution analysis revealed that STAMP captures target-dependent activity shifts and pinpoints sequence regions crucial for species-conditioned predictions. Collectively, these results demonstrate the value of integrating molecular and microbial context to model AMP activity, supporting computationally guided and efficient AMP discovery.

1 INTRODUCTION

Antibiotics are essential for treating life-threatening bacterial infections (CDC Report, 2019). However, their widespread use across healthcare and agriculture has accelerated antimicrobial resistance (AR), significantly reducing the effectiveness of existing treatments (CDC Report, 2024; Ahmed et al., 2024). The COVID-19 pandemic further intensified this global burden, highlighting the urgent need for alternative therapeutic strategies Murray et al. (2022); Prestinaci et al. (2015); Hunt & Kates (2024). Antimicrobial peptides (AMPs) have emerged as a promising solution; these naturally occurring molecules directly disrupt bacterial membranes, a mechanism that makes it inherently difficult for resistance to develop Zhang & Gallo (2016); Huan et al. (2020).

Despite their potential, the discovery and design of novel AMPs is hindered by the challenge of exploring a vast chemical search space and the limitations of traditional experimental methods. AMP potency is typically quantified by the minimum inhibitory concentration (MIC)—the lowest concentration required to inhibit growth—where lower values indicate higher efficacy Elbehiry et al. (2025); Gerada et al. (2024); Salam et al. (2023); Kowalska-Krochmal & Dudek-Wicher (2021). Critically, MIC is species-dependent[1], meaning that the activity of a single peptide can vary substantially across different microbial targets Kim et al. (2022). Conventional growth-based assays, such as broth microdilution, are often too time-consuming, costly, and resource-intensive for large-scale screening, making the validation of designed AMPs via MIC testing inefficient Elbehiry et al. (2025); Gerada et al. (2024); Salam et al. (2023). Consequently, computational models have become a vital culture-free strategy to rapidly estimate activity and prioritize candidates for validation Gerada et al. (2024); Wang et al. (2022); Nguyen et al. (2019).

The success of computational models depends on how effectively they represent the given peptide. Early approaches relied on handcrafted physicochemical descriptors such as charge and hydrophobicity, which, while interpretable, offer a limited view of sequence information Bajiya et al. (2025); Chung et al. (2024). Modern Protein Language Models (PLMs), such as ESM-2 Lin et al. (2023), provide much richer contextual representations by learning from

[1]The terms "species" and "target" both refer to bacterial pathogens and are used interchangeably.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

large protein sequence corpora. However, PLM embeddings primarily capture distributed sequence information and may not explicitly represent recurring functional motifs or the biological context in which a peptide acts, particularly the microbial target. Unlike prior approaches that train separate models per species Bae et al. (2025); Chung et al. (2024) or transform sequences into image representations Chiu et al. (2026), STAMP integrates sequence, motif, and species context within a single cross-species architecture. These limitations motivate a unified representation-learning approach that can capture both shared sequence-level patterns and target-specific activity information across microbial species. A comprehensive review of related work is provided in Appendix A.1.

To address this need, we propose STAMP, a Species- and Topic-aware representation-learning framework for AMP activity analysis. STAMP integrates contextual sequence representations from ESM-2, latent motif composition derived via Latent Dirichlet Allocation (LDA), and microbial species context within a unified cross-species model. This joint formulation models peptide sequences, motif compositions, and target contexts simultaneously, enabling the framework to learn shared sequence–function relationships across microbial targets while preserving species-dependent activity nuances.

Our primary contributions are as follows:

- We introduce STAMP, a unified cross-species framework that integrates protein language model embeddings, latent motif composition, and species-level context to model AMP activity as a context-dependent property of the peptide–target pair, replacing the conventional approach of training separate models for each species.
- We rigorously evaluate STAMP across four independent datasets spanning 3–35 bacterial species, different source databases, and predefined splits, demonstrating that a single model matches methods requiring up to 35 independently trained species-specific models and outperforms the most recently published benchmark.
- We validate STAMP's predictive reliability by screening a large peptide library and synthesizing the top 18 candidates; 16/18 exhibit MIC < 10 μM against E. coli and 17/18 against S. epidermidis, demonstrating effective in silico prioritization of potent antimicrobial peptides.
- We demonstrate STAMP's interpretability through species-conditioned analysis of TP4-derived peptides and residue-level attribution via Gradient × Input, revealing that the model captures target-dependent activity shifts and identifies biologically meaningful residue contributions rather than operating as a black box.

2 STAMP: Species- and Topic-Aware Representation Learning

STAMP is a multimodal representation framework for AMPs that integrates complementary sequence-, motif-, and species-level information. As shown in Figure 1, it combines contextual ESM-2 sequence representations, LDA-derived topic representations of latent peptide motifs, and one-hot encoding of the target bacterial species into a unified representation. We evaluate this representation through continuous MIC prediction across multiple bacterial species; AMP/non-AMP classification is subsequently derived from predicted MIC values using thresholds established through ROC and precision-recall analyses. The problem formulation and individual components are described below.

Given a dataset $\mathcal{D} = \{(s_i, b_i, m_i)\}_{i=1}^{N}$, where s_i denotes a peptide sequence, b_i its target microbial species, and m_i the measured MIC, we formulate antimicrobial activity as a context-dependent sequence-function relationship. Since the same peptide can exhibit different activities against different microbial targets, the log-transformed activity $y_i = \log_{10}(m_i)$ is modeled as

$$\hat{y}_i = f_\Theta(s_i, b_i), \tag{1}$$

where the target species is assumed to be known, as in conventional MIC evaluation. This formulation emphasizes that MIC is not determined by peptide sequence alone, but depends on the interaction between the peptide and its microbial target. STAMP realizes this relationship using a shared cross-species representation-learning framework in which peptide sequence is represented at complementary contextual and motif levels, while species information provides the biological context for activity prediction.

2.1 Protein Language Modeling for Sequence Semantics

To capture fundamental biological syntax, we use ESM-2 (Evolutionary Scale Modeling 2) Lin et al. (2023), a transformer-based protein language model pretrained on large-scale protein sequence data using masked language modeling. Given a protein sequence, ESM-2 produces a contextualized representation for each residue, such that the representation of an amino acid depends on its surrounding sequence context. This enables the pretrained encoder to capture local and long-range sequence dependencies learned from protein sequence data.

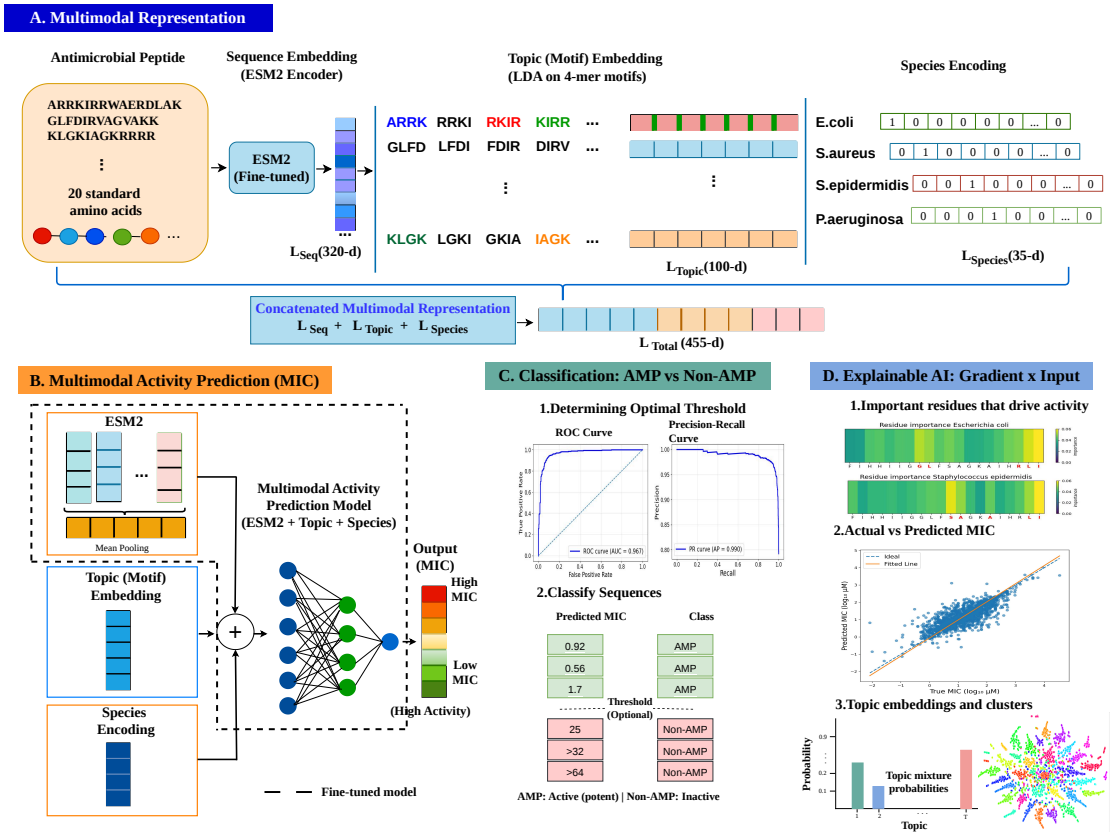

Figure 1: A framework of Species and Topic-aware representation learning for AMP activity analysis (STAMP). (A) Antimicrobial peptide (AMP) sequences are encoded using ESM2 embeddings. Topic embeddings are derived from an LDA topic model that captures latent antimicrobial motifs, while species one-hot encodings represent the target bacteria. (B) The concatenated representations are fed into a multi-layer perceptron (MLP) regression head to predict MIC values. The ESM2 model is fine-tuned for MIC prediction with fixed topic and species encodings. (C) The classification of sequences into AMP and non-AMP categories is achieved using predicted MIC values, with an optimal threshold determined through precision-recall and receiver operating characteristic analyses. (D) An explainable AI component employs the Gradient × Input method to identify key residues that influence antimicrobial activity. Accompanying scatter plots illustrate the correlation between actual and predicted values, and topic embeddings are visualized, showcasing clusters extracted for each sequence.

For a peptide s_i of length L_i, let $\mathbf{H}_i = [\mathbf{h}_{i1}, \ldots, \mathbf{h}_{iL_i}]$ denote the residue-level hidden states produced by ESM-2, where $\mathbf{h}_{ij} \in \mathbb{R}^{d_{\text{ESM}}}$ is the contextualized representation of residue j. We obtain a fixed-dimensional peptide-level representation using masked mean pooling:

$$\mathbf{h}_i^{\text{ESM}} = \frac{1}{L_i} \sum_{j=1}^{L_i} \mathbf{h}_{ij} \in \mathbb{R}^{d_{\text{ESM}}}, \tag{2}$$

where padding and special tokens are excluded from the pooling operation. The resulting vector $\mathbf{h}_i^{\text{ESM}}$ serves as the contextual sequence representation of peptide i in STAMP. Rather than using ESM-2 as a fixed feature extractor, its parameters are fine-tuned jointly with the MIC prediction network, allowing the pretrained sequence representation to adapt to AMP activity prediction task.

2.2 Probabilistic Modeling of Latent Structural Motifs

Complementing the distributed semantics of ESM-2, STAMP explicitly represents recurring local sequence patterns through probabilistic topic modeling. Latent Dirichlet Allocation (LDA) Blei et al. (2003) is a statistical model that represents documents as mixtures of latent topics defined by distributions over words. Recent topic-model-based

analysis of AMPs has demonstrated that LDA-derived latent motif groups capture recurring sequence patterns associated with physicochemical properties and antimicrobial activity Padi et al. (2025). Unlike k-mer frequency vectors Padi et al. (2025), which treat sequence patterns independently, or attention-based motifs Lee et al. (2023), which are coupled to a specific downstream objective, LDA captures contextual co-occurrence relationships among sequence patterns in a continuous, task-independent representation. For peptide sequences, we treat each peptide as a document and its overlapping amino-acid 4-mers as words. LDA represents each peptide through a distribution over K latent topics, while each topic defines a probability distribution over the 4-mer vocabulary:

$$\boldsymbol{\theta}_i \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}), \qquad \boldsymbol{\phi}_k \sim \mathrm{Dirichlet}(\boldsymbol{\eta}), \qquad P(w \mid z = k) = \phi_{kw}, \tag{3}$$

where $\boldsymbol{\phi}_k$ characterizes topic k through its distribution over 4-mers, such that high-probability 4-mers identify recurring sequence patterns associated with that latent group. After inference, peptide s_i is represented by its posterior topic-mixture vector

$$\boldsymbol{\theta} i = [P(z_1 \mid s_i), \ldots, P(z_K \mid s_i)], \qquad \sum_{k=1}^{K} \theta_{ik} = 1 \tag{4}$$

Unlike a hard motif assignment, this mixed-membership representation allows multiple latent motif groups to contribute to the same peptide, providing an explicit representation of its motif composition that is complementary to the ESM-2 representation. LDA is an unsupervised model trained on AMPs independently of MIC labels and remains fixed during STAMP training, preserving an unsupervised motif representation based on sequence co-occurrence rather than adapting the discovered patterns to the downstream MIC objective. The LDA model is fitted exclusively on the training sequences to learn latent topic distributions over the sequence features. The fitted LDA model is then used to transform both the training and test sequences into the same learned topic space, ensuring consistent representation across data splits. Further details are provided in Section A.3 in the Appendix.

2.3 Multimodal Integration of Sequence, Motif and Biological Context

Antimicrobial function depends not only on peptide sequence but also on the microbial target. Rather than training independent models on species-specific subsets, STAMP learns from the pooled dataset $\mathcal{D} = \bigcup_{b=1}^{B} \mathcal{D}_b$, enabling sequence-activity relationships to be shared across microbial targets. The target species is represented by a one-hot vector $\mathbf{e}_b \in \{0,1\}^B$, without introducing a separate learnable species embedding. The complete STAMP representation is

$$\mathbf{z}_i = \left[\mathbf{h}_i^{\mathrm{ESM}} \| \boldsymbol{\theta}_i \| \mathbf{e}_b\right], \tag{5}$$

where $\|$ denotes vector concatenation, $\mathbf{h}_i^{\mathrm{ESM}}$ is the ESM2 sequence representation, $\boldsymbol{\theta}_i$ is the LDA topic-mixture representation, and $\mathbf{e}_b$ is the one-hot encoding of the target bacterial species. Thus, motif and species context can alter the hidden activation pattern through which the ESM-2 representation contributes to MIC prediction. Although Eq. 5 is implemented through concatenation, these components interact through the nonlinear prediction network. Partitioning the first MLP layer according to the three representation components gives

$$\mathbf{a}_i = W_h \mathbf{z}_i + \mathbf{b}_h, \qquad \mathbf{r}_i = \mathrm{ReLU}(\mathbf{a}_i), \tag{6}$$

The log-transformed MIC and training objective are

$$\hat{y}_i = \mathbf{w}_o^\top \mathbf{r}_i + b_o, \qquad \mathcal{L}_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{7}$$

2.4 Species-Conditioned Learning Dynamics

A critical feature of STAMP is that the fixed topic mixture $\boldsymbol{\theta}_i$ and the species encoding $\mathbf{e}_b$ actively influence how the shared ESM-2 representation is optimized. Let $D_i = \mathrm{diag}(1[\mathbf{a}_i > 0])$ denote the ReLU activation mask. For the sample loss $\mathcal{L}_i = (\hat{y}_i - y_i)^2$, the gradient reaching the ESM-2 representation is:

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{h}_i^{\mathrm{ESM}}} = 2(\hat{y}_i - y_i) \left(W_{h_i}^{\mathrm{ESM}}\right)^\top D_i \mathbf{w}_o \tag{8}$$

Equation 8 demonstrates that context influences learning through two mechanisms: first, $\boldsymbol{\theta}_i$ and $\mathbf{e}_b$ alter the ReLU activation mask D_i, changing the gradient pathways for the sequence features; and second, they influence the prediction $\hat{y}_i$, which scales the residual learning signal. Consequently, these contextual variables condition the learning signals propagated into the shared sequence encoder. This mechanism is particularly relevant when the same

peptide is evaluated against different species. For targets t and t', the sequence representation remains unchanged, but the pre-activations differ by:

$$\mathbf{a}_{i,t} - \mathbf{a}_{i,t'} = W_{h_i}^{\text{species}}(\mathbf{e}_t - \mathbf{e}_{t'}) \tag{9}$$

This context-dependent shift allows STAMP to maintain a single shared ESM-2 encoder while learning through species-conditioned gradient signals, capturing universal sequence–function relationships that remain sensitive to biological context.

3 RESULTS

3.1 DATASETS

To evaluate the performance of the STAMP model, we consider four AMP activity datasets comprising measurements across multiple bacterial targets. The first three benchmarks were introduced by Bae et al. (2025), Chung et al. (2024) and Chiu et al. (2026), respectively, while the fourth one is a curated DBAASP-v3 dataset developed in this work. The benchmarks provide complementary evaluation settings, ranging from measurements across 35 bacterial species to activity measurements against three priority pathogens and a newly curated set spanning four target species. The first three benchmark datasets provide predefined training, validation, and test splits for model training and evaluation, as detailed in Appendix A.2. For our curated dataset, detailed descriptions of MIC standardization, handling of duplicate and conflicting measurements, and activity-stratified data partitioning are provided in Appendix A.2. An anonymized GitLab repository containing the STAMP implementation, model configurations, and dataset partitions is available at `https://gitlab.com/anonymous-iclr/stamp.git`.

3.2 CROSS-SPECIES AMP ACTIVITY PREDICTION

We evaluated STAMP under a unified cross-species setting across the datasets described above and compared it with existing MIC prediction approaches as summarized in Table 1. On Dataset A, STAMP matches LLAMP using a single cross-species model, whereas LLAMP requires 35 independent species-specific models. On Dataset B, STAMP achieves comparable accuracy to the Ensemble method, which constructs and selects from 24 separate models (eight models per species), further demonstrating that a unified architecture can rival multi-model strategies without per-species training and tuning. On the ANIA dataset (Dataset C), STAMP outperforms the Attention-based model across all metrics, achieving a PCC of 0.798 compared to 0.775 and an R^2 of 0.634 compared to 0.585, while also yielding a lower MSE of 0.221 versus 0.246. In contrast to all these approaches, STAMP conditions a shared model on species identity, enabling shared peptide-activity patterns to be learned across species while retaining species-specific context. This unified architecture avoids separate model construction and tuning for each target and allows data-sparse species to benefit from shared sequence-activity patterns.

Notably, STAMP achieved its highest performance on our curated dataset, with a PCC of 0.837, $R^2 = 0.698$, MAE of 0.285, and MSE of 0.171, demonstrating strong agreement between predicted and experimentally measured MIC values. Per-species analysis on the curated dataset reveals consistently high accuracy, with *S. epidermidis* achieving the highest PCC of 0.856 and R^2 of 0.731, followed by *P. aeruginosa* (PCC = 0.845, $R^2 = 0.711$), while *E. coli* and *S. aureus* also maintain strong correlations above 0.80. Detailed per-species metrics for all datasets are provided in the Appendix A.4. The predicted-versus-measured MIC plots in Figures 3-6 (Appendix) further show that STAMP predictions closely follow the identity line across all four datasets, and an analysis of prediction error direction and dispersion confirms no systematic bias toward over- or underestimation of MIC values (Appendix A.6).

3.3 CANDIDATE PRIORITIZATION

To evaluate STAMP as a screening tool for candidate prioritization, we applied the trained model to a large library of 70,000 computationally generated peptide sequences and filtered the candidates based on their predicted MIC values, from which 18 candidate peptides were selected for experimental validation.[2]. As shown in Table 2, for 16 of the 18 peptides tested against *E. coli*, both STAMP-predicted and experimentally measured MIC values are below 10 μM. Similarly, for 17 of the 18 peptides tested against *S. epidermidis*, the predicted and measured values are below 10 μM, demonstrating that the model successfully prioritized candidates with strong experimentally confirmed antimicrobial activity. One peptide (P3) showed a substantial quantitative discrepancy, with experimentally measured MICs of 64 μM against both species. P3 differs from accurately predicted P1 and P2 by a single substitution of $I \Rightarrow W$ at position 7, suggesting that this mutation disrupts antimicrobial activity through structural or physicochemical effects not fully captured by the sequence-level representation.

[2]Work describing the peptide generation procedure and experimental characterization is omitted here to preserve anonymity.

Table 1: The performance of our proposed model, STAMP, was compared against several state-of-the-art methods for MIC prediction analysis. While the LLAMP model builds separate models for each of the 35 species and reports only the average score, and the Ensemble method trains eight models per species to select the top two for averaging scores across three species, our STAMP model utilizes a multimodal framework trained as a single model across all species. Note: Dataset A is a benchmark dataset published by Bae et al. (2025), Dataset B: published by Chung et al. (2024), Dataset C: published by Chiu et al. (2026)

Dataset	Method	Models	Species	MSE↓	MAE↓	R2↑	PCC↑
Dataset A	LLAMP	35	Avg. across 35 species	0.272	0.379	0.536	0.735
	STAMP	1	across-species	**0.275**	**0.385**	**0.531**	**0.732**
Dataset B	Ensemble model	24	*Avg. across 3 species*	0.233	–	0.603	0.779
	STAMP	1	across species	**0.260**	**0.363**	**0.566**	**0.758**
Dataset C	Attention-model	3	*Avg. across 3 species*	0.246	–	0.585	0.775
	STAMP	1	across species	**0.221**	**0.330**	**0.634**	**0.798**
Our Dataset	**STAMP**	1	across species	**0.171**	**0.285**	**0.698**	**0.837**

Table 2: Experimental validation of STAMP-prioritized peptide candidates. 18 peptides were selected from a computationally generated library based on low STAMP-predicted MIC values ($< 10\mu$M). Measured MIC values confirmed that 16 of the 18 candidates are active (MIC < 10 μM) against *E. coli*, while 17 of the 18 candidates are active against *S. epidermidis*. M: measured MIC; P: STAMP predicted MIC values (model trained on Dataset A is used to predict MIC).

Sequence ID	Sequence	E. coli		S. epidermidis	
		M	P	M	P
P1	FIHHIIGGLFSAGKAIHRLIRRRRR	4	4.54	1	1.24
P2	FIHHIIGGLFSAGRAIHRLIRRRRR	4	5.86	1	1.42
P3	FIHHIIWILFSAGRAIHRLIRRRRR	64	15.12	64	4.34
P4	RWKGVIQAGKAIHRIIRRRR	8	5.59	1	6.24
P5	VWRHFFRGIRHVGKTIHRLVTG	2	2.69	1	2.00
P6	KWRSFIRGVTHAAHTALKHYFRRKRSG	1	1.29	0.5	1.49
P7	VFWAKFKVAFRGVRHIGRTILRHVLR	8	3.34	1	2.31
P8	VWGFIKHIVRPILRRLTRGLNKL	8	6.91	1	5.06
P9	VWQHIFRGIVHVGRTIHRLITG	2	2.08	1	2.77
P10	VTWRHIFRGIVHVGKTIHRLVTG	2	1.67	0.5	1.93
P11	VTWRAIFRGIIHVGKTIHRLVTG	4	4.79	0.5	2.67
P12	DWRFIRRIWKVGSHIYRAIKKLL	16	2.78	2	1.28
P13	LWGHIFRGIKHIGKTIHRLVTG	1	1.91	1	1.56
P14	VKSWIRVAGRAIGKLIRRK	2	4.47	0.5	5.05
P15	LWHIFKRIHRGIVHVGKTIHRLVTG	4	3.04	0.5	2.29
P16	NWRNAWRRFKGAFRHVIHTAKRF	8	2.83	0.5	3.72
P17	IWHHIVRGALHVGKLIHRLVTG	2	3.19	1	2.20
P18	RWRSILKHGLKHIGKTIHRLVTG	1	1.28	0.5	1.79

3.4 Species-Conditioned Validation

To evaluate whether STAMP captures species-dependent antimicrobial activity, we analyzed four previously characterized peptides against Gram-negative *E. coli* and Gram-positive *S. epidermidis* pathogens. Table 3 reports the peptide sequences, experimentally measured MICs, and the STAMP predictions for both bacterial targets by changing only the species context. Arginine-rich TP4 and NP-R5 are predicted to be highly active against both species, consistent with their experimentally measured low MICs, while peptides lacking extended C-terminal arginine segments exhibited more pronounced target-dependent differences. For example, STAMP correctly predicts that TP4-noR5 is more active against *S. epidermidis* than *E. coli*, preserving the experimentally observed direction of the activity shift despite quantitative differences. These results confirm that STAMP's cross-species formulation captures target-dependent activity from a single shared model.

3.5 Context-Aware Residue Attribution

The species-conditioned validation in Section 3.4 shows quantitatively that the same peptide can yield different MIC predictions when conditioned on different targets. We next examine this behavior qualitatively at the residue level to identify which amino acids contribute the most to the predicted activity. Specifically, we apply Gradient $\times$

Table 3: **Comparison of STAMP-predicted and experimentally measured MICs for TP4-derived peptides.** STAMP correctly captures the activity of peptides against *E. coli* and *S. epidermidis*, with TP4-noR5 predicted as inactive against *E. coli*. Peptides containing C-terminal arginine clusters (TP4 and NP-R5) show consistently low predicted MICs across both species, in agreement with the experimentally measured MIC values. The differences between measured and predicted MICs for these peptides are also comparable across the two microbial targets. Note: M: measured MIC; P: STAMP predicted MIC values.

Peptide	Sequence	E. coli		S. epidermidis	
		M	P	M	P
TP4	FIHHIIGGLFSAGKAIHRLIRRRRR	4	4.54	1	1.24
TP4-noR5	FIHHIIGGLFSAGKAIHRLI	16	41	2	16
NP	QLAQALAAALAALAQGW	64	33.81	64	31.21
NP-R5	QLAQALAAALAALAQGWRRRRR	8	10	2	10.37

Input Ancona et al. (2018) saliency attribution to the ESM-2 residue representations and compare attribution profiles for the same peptide conditioned on *E. coli* and *S. epidermidis*. As shown in Figure 2, the resulting profiles differ across the two species, highlighting target-dependent changes in the relative contributions of individual residues to the prediction of MIC. Because the peptide sequence is fixed and only the species context is changed, these differences provide qualitative evidence that STAMP uses the microbial context when interpreting residue-level sequence information.

Across the analyzed peptides, positively charged residues, particularly arginine (R), frequently receive high attribution scores, with C-terminal poly-arginine regions prominently emphasized in several TP4-derived sequences. Hydrophobic residues, including leucine (L), isoleucine (I), alanine (A), and tryptophan (W), also contribute strongly, consistent with established AMP properties associated with electrostatic membrane interactions and hydrophobic membrane insertion. Importantly, the magnitude and distribution of these residue contributions vary between microbial targets. Thus, STAMP does not assign a fixed importance profile to a peptide sequence; instead, it learns context-dependent sequence-activity relationships in which the relevance of individual residues is conditioned by the target species.

3.6 Ablation Study

We conducted ablation experiments to examine the contributions of the representation components used in our proposed STAMP model. We compare four configurations: **Sequence only**, using only the ESM-2 representation for a given sequence; **Sequence + Motif**, combining ESM-2 embeddings with the LDA-derived topic-mixture representation; **Sequence + Species**, which combines the ESM-2 representation and the species context; and the complete **STAMP** model, which integrates sequence, motif, and species context information. **Sequence only** and **Sequence + Motif** are trained as separate species-specific models, with performance averaged across species, whereas **Sequence + Species** and **STAMP** use a single cross-species model jointly trained across all species. The complete training and evaluation setup is summarized in Appendix A.3.

As shown in Table 4, the **Sequence only** model provides a strong baseline across all four datasets, confirming that ESM-2 embeddings capture meaningful sequence-activity relationships. Adding motif information (**Sequence + Motif**) consistently improves per-species performance, while species conditioning (**Sequence + Species**) enables a single cross-species predictor with competitive accuracy. The relative contribution of each component varies by dataset: on Dataset A, which spans 35 species, species conditioning alone underperforms the per-species baselines (R^2: 0.334 vs. 0.445), whereas on Dataset C it surpasses the motif-only variant (R^2: 0.476 vs. 0.450). Despite this variation, the complete **STAMP** model achieves the best performance on every dataset. These results demonstrate that motif information and species context play complementary roles: neither component alone matches their combined contribution within STAMP's unified cross-species prediction framework.

The consistent improvement from motif information across all datasets reflects the complementary nature of the ESM-2 and LDA representations. While ESM-2 captures global contextual sequence semantics, the LDA topic-mixture vector encodes the latent motif composition of each peptide, such that sequences sharing similar functional motifs are mapped to similar topic-mixture representations. This grouping effect enables the downstream predictor to associate recurring motif patterns with species-specific activity profiles; for example, motif groups enriched in cationic residues may correlate with low MIC values, while others may characterize hydrophobic or inactive peptide subpopulations. By combining these complementary views with species context, STAMP can leverage motif-level similarities across peptides to improve species-conditioned predictions, particularly when individual species have limited training data.

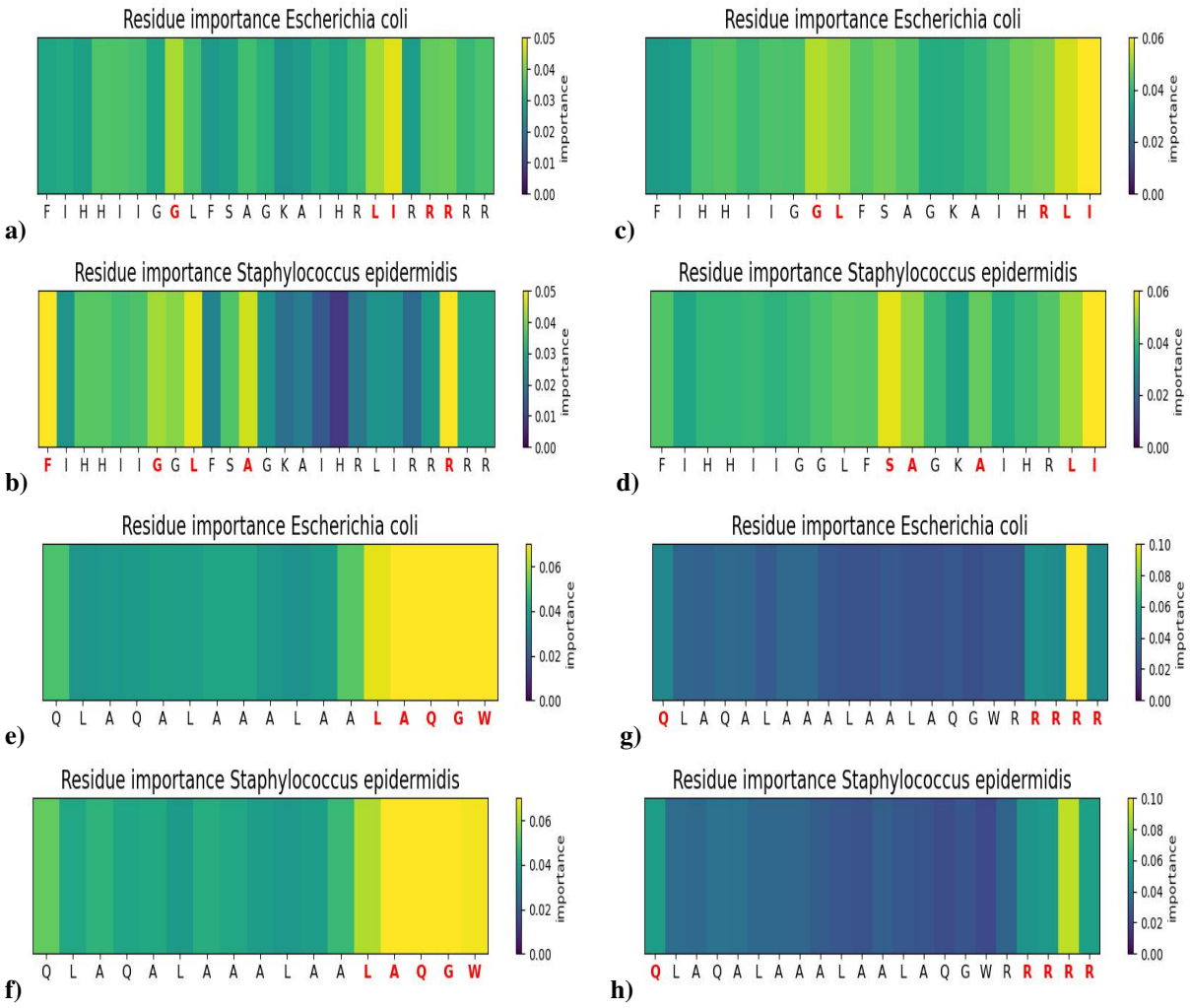

Figure 2: Residue importance analysis shows enrichment of positively charged residues, particularly arginine(R) amino acid, among the most influential positions for MIC prediction. Hydrophobic residues such as leucine (L), isoleucine (I), and tryptophan (W) are also frequently highlighted, suggesting that the model captures amphipathic sequence features associated with AMP membrane disruption. Differences in residue importance between *E. coli* and *S. epidermidis* targets further indicate species-dependent sequence feature recognition by the model. Note: Panels (a) & (b), (c) & (d), (e) & (f), and (g) & (h) represent the same peptide sequences evaluated against different microbial targets, with (a), (c), (e), and (g) corresponding to *E. coli* and (b), (d), (f), and (h) corresponding to *S. epidermidis*, highlighting species-dependent differences in residue contributions to predicted antimicrobial activity.

4 Discussion

The results across four benchmark datasets demonstrate that STAMP achieves comparable or superior performance compared to existing methods while using a single unified model. On Dataset A, STAMP matches LLAMP, which trains separate models for 35 species; on Dataset B, it matches the Ensemble method, which trains eight models per species; and on Dataset C, it outperforms the Attention-based model. The strongest performance is observed on our curated dataset, likely reflecting the benefits of standardized preprocessing. The cross-species formulation provides the greatest advantage when individual species have limited training data, as demonstrated by the largest performance gap between per-species baselines and STAMP on Dataset A (35 species). The species-conditioned attribution analysis further reveals that this is not merely a data-pooling effect: the model learns to modulate residue-level importance based on microbial context, generating biologically interpretable differences consistent with known Gram-negative and Gram-positive membrane properties. Of the 18 experimentally evaluated candidates, 16 exhibited measured MIC values below 10 μM against *E. coli* and 17 below 10 μM against *S. epidermidis*, supporting the use of STAMP as a practical screening filter for candidate prioritization. For species with limited training representation

Table 4: **STAMP representation ablation and training strategies.** Sequence and Sequence+Motif use separate per-species models, with performance averaged across species, whereas Sequence+Species and STAMP use a single cross-species model jointly trained across all species. Best performance for each dataset is shown in bold.

Model	Sequence	Motif	Species	Training	MSE↓	MAE ↓	$\mathbf{R^2}$ ↑	PCC↑
Dataset A								
Sequence only	✓	–	–	Per species	0.301	0.416	0.445	0.670
Sequence + Motif	✓	✓	–	Per species	0.291	0.40	0.481	0.697
Sequence + Species	✓	–	✓	Cross-species	0.39	0.49	0.334	0.581
STAMP	✓	✓	✓	Cross-species	**0.275**	**0.385**	**0.531**	**0.732**
Dataset B								
Sequence only	✓	–	–	Per species	0.360	0.451	0.391	0.635
Sequence + Motif	✓	✓	–	Per species	0.340	0.440	0.423	0.661
Sequence + Species	✓	–	✓	Cross-species	0.347	0.446	0.421	0.654
STAMP	✓	✓	✓	Cross-species	**0.26**	**0.363**	**0.566**	**0.758**
Dataset C								
Sequence only	✓	–	–	Per species	0.336	0.443	0.436	0.661
Sequence + Motif	✓	✓	–	Per species	0.326	0.436	0.450	0.676
Sequence + Species	✓	–	✓	Cross-species	0.316	0.430	0.476	0.692
STAMP	✓	✓	✓	Cross-species	**0.221**	**0.330**	**0.634**	**0.798**
Our Curated Dataset								
Sequence only	✓	–	–	Per species	0.29	0.429	0.471	0.687
Sequence + Motif	✓	✓	–	Per species	0.263	0.391	0.520	0.725
Sequence + Species	✓	–	✓	Cross-species	0.264	0.399	0.532	0.732
STAMP	✓	✓	✓	Cross-species	**0.171**	**0.285**	**0.698**	**0.837**

(Dataset A), STAMP yields negative R^2 values, indicating that the cross-species model does not consistently outperform a mean predictor when fewer than approximately 100 species-specific training samples are available. These cases highlight a lower bound on the per-species data required for reliable predictions and motivate future work on incorporating genomic similarity to improve generalization for underrepresented species. We use the 8M-parameter ESM-2 model, which is well suited to the scale of available AMP datasets and the short length of peptide sequences (5–50 residues), enabling stable fine-tuning with limited training samples. A comprehensive discussion of per-species analysis, modeling insights, and limitations is provided in Appendix A.7.

5 Conclusion and Future Work

The central finding of this work is that antimicrobial activity can be modeled as a context-dependent property of the peptide-target pair rather than solely as an intrinsic property of the peptide sequence. STAMP operationalizes this formulation through a single cross-species model, achieving superior or comparable MIC prediction performance across four benchmark datasets without constructing separate models for each target. By enabling rapid, species-aware prioritization of AMP candidates from a single model, STAMP offers a computationally efficient complement to experimental screening pipelines, supporting the broader effort to accelerate AMP discovery in response to the growing threat of antimicrobial resistance.

Three lines of evidence support the utility of this formulation. First, the ablation study shows that sequence, motif, and species components provide complementary information, with the cross-species formulation offering greater benefit for species with limited training data. Second, candidate prioritization experiments demonstrate the utility of STAMP for enriching active peptides from a large computationally generated library. Third, species-conditioned analysis of TP4-derived peptides and residue-level attribution reveals target-dependent differences in predicted activity and residue contributions, rather than assigning a fixed importance profile to each sequence.

The current formulation represents microbial context primarily through species identity, whereas MIC measurements can also depend on strain, assay protocol, and other experimental factors that are not explicitly modeled. Moreover, prospective experimental evaluation currently covers a limited number of peptides and microbial targets. Future work will therefore investigate richer biological context, including genomic, phylogenetic, strain-level, and physicochemical information, together with broader prospective experimental validation. Such extensions could enable more comprehensive context-aware AMP representations and improve generalization to microbial targets not represented during training.

AI USE STATEMENT

Generative AI tools were used to assist in editing, rephrasing, and condensing the manuscript for clarity and compliance with the ICLR format and page limit. All AI-assisted revisions were reviewed and verified by the authors, who take responsibility for the final content.

REFERENCES

Sirwan Khalid Ahmed, Safin Hussein, Karzan Qurbani, Radhwan Hussein Ibrahim, Abdulmalik Fareeq, Kochr Ali Mahmood, and Mona Gamal Mohamed. Antimicrobial resistance: Impacts, challenges, and future prospects. *Journal of Medicine, Surgery, and Public Health*, 2:100081, 2024. ISSN 2949-916X. doi: https://doi.org/10.1016/j.glmedi.2024.100081. URL https://www.sciencedirect.com/science/article/pii/S2949916X24000343.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Daehun Bae, Minsang Kim, Jiwon Seo, and Hojung Nam. Ai-guided discovery and optimization of antimicrobial peptides through species-aware language model. *Briefings in Bioinformatics*, 26(4):bbaf343, 07 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf343.

Nisha Bajiya, Nishant Kumar, and Gajendra PS Raghava. Prediction of inhibitory peptides against e. coli with desired mic value. *Scientific Reports*, 15(1):1–15, 2025. doi: 10.1038/s41598-025-86638-z.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. ISSN 1532-4435.

CDC Report. 2019 antibiotic resistance threats report, 2019. URL https://www.cdc.gov/antimicrobial-resistance/data-research/threats/index.html.

CDC Report. Antibiotic use and antimicrobial resistance facts, 2024. URL https://www.cdc.gov/antibiotic-use/data-research/facts-stats/index.html.

Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022. doi: 10.1109/TMI.2020.3021387.

Yen-Peng Chiu, Lantian Yao, Yun Tang, Chia-Ru Chung, Yuxuan Pang, Ying-Chih Chiang, and Tzong-Yi Lee. Ania: an inception-attention network for predicting minimum inhibitory concentration of antimicrobial peptides. *Briefings in Bioinformatics*, 27(1):bbag023, 01 2026. ISSN 1477-4054. doi: 10.1093/bib/bbag023. URL https://doi.org/10.1093/bib/bbag023.

Chia-Ru Chung, Chung-Yu Chien, Yun Tang, Li-Ching Wu, Justin Bo-Kai Hsu, Jang-Jih Lu, Tzong-Yi Lee, Chen Bai, and Jorng-Tzong Horng. An ensemble deep learning model for predicting minimum inhibitory concentrations of antimicrobial peptides against pathogenic bacteria. *iScience*, 27(9):110718, 2024. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2024.110718.

Ayman Elbehiry, Eman Marzouk, Adil Abalkhail, Moustafa H Abdelsalam, Mohamed EA Mostafa, Mazen Alasiri, Mai Ibrahem, Abousree T Ellethy, Abdulaziz Almuzaini, Sahar N Aljarallah, et al. Detection of antimicrobial resistance via state-of-the-art technologies versus conventional methods. *Frontiers in Microbiology*, 16:1549044, 2025.

Alessandro Gerada, Nicholas Harper, Alex Howard, Nada Reza, and William Hope. Determination of minimum inhibitory concentrations using machine-learning-assisted agar dilution. *Microbiology Spectrum*, 12(5):e04209–23, 2024.

LI Haimeng, WANG Lufan, ZHU Congmin, and YANG Yuqing. Esm2amp: An esm-2-based deep learning framework for antimicrobial peptide identification and mic prediction from bohai sea metagenomes. *Journal of Ocean University of China*, 25(5):1733–1740, 2026. ISSN 1672-5182. doi: 10.1007/s11802-026-6240-9.

Yuchen Huan, Qing Kong, Haijin Mou, and Huaxi Yi. Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology*, Volume 11 - 2020, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.582779. URL https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2020.582779.

Devin Hunt and Olivia S Kates. A brief history of antimicrobial resistance. *AMA Journal of Ethics*, 26(5):408–417, 2024.

Jee In Kim, Finlay Maguire, Kara K Tsang, Theodore Gouliouris, Sharon J Peacock, Tim A McAllister, Andrew G McArthur, and Robert G Beiko. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clinical microbiology reviews*, 35(3):e00179–21, 2022.

Beata Kowalska-Krochmal and Ruth Dudek-Wicher. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens*, 10, 2021. URL 10.3390/pathogens10020165.

Hojung Lee, Sohyun Lee, Ingoo Lee, and Hojung Nam. Amp-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein Science*, 32(1):e4529, 2023. doi: 10.1002/pro.4529.

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. doi: 10.1126/science.ade2574.

Christopher J L Murray, Kevin Shunji Ikuta, and Fablina Sharara. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(21)02724-0. URL https://www.sciencedirect.com/science/article/pii/S0140673621027240.

Marcus Nguyen, S. Wesley Long, Patrick F. McDermott, Randall J. Olsen, Robert Olson, Rick L. Stevens, Gregory H. Tyson, Shaohua Zhao, and James J. Davis. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *Journal of Clinical Microbiology*, 57(2), 2019. doi: 10.1128/jcm.01260-18.

Sarala Padi, Kinjal Mondal, David P Hoogerheide, Frank Heinrich, Mihaela Mihailescu, Jeffery B Klauda, and Antonio Cardone. Ai-driven antimicrobial peptide characterization unveils novel motifs for drug design. *Scientific Reports*, 2025.

Francesca Prestinaci, Patrizio Pezzotti, and Annalisa Pantosti. Antimicrobial resistance: a global multifaceted phenomenon. *Pathogens and Global Health*, 109(7):309–318, 2015. doi: 10.1179/2047773215Y.0000000030.

Md. Abdus Salam, Md. Yusuf Al-Amin, Jogendra Singh Pawar, Naseem Akhter, and Irine Banu Lucy. Conventional methods and future trends in antimicrobial susceptibility testing. *Saudi Journal of Biological Sciences*, 30(3):103582, 2023. ISSN 1319-562X. doi: https://doi.org/10.1016/j.sjbs.2023.103582.

Guangshun Wang, Iosif I Vaisman, and Monique L Van Hoek. Machine learning prediction of antimicrobial peptides. In *Computational peptide science: Methods and protocols*, pp. 1–37. Springer, 2022.

Lei Wang, Xudong Li, Han Zhang, Jinyi Wang, Dingkang Jiang, Zhidong Xue, and Yan Wang. A comprehensive review of protein language models, 2025. URL https://arxiv.org/abs/2502.06881.

Jielu Yan, Pratiti Bhadra, Ang Li, Pooja Sethiya, Longguang Qin, Hio Kuan Tai, Koon Ho Wong, and Shirley W.I. Siu. Deep-ampep30: Improve short antimicrobial peptides prediction with deep learning. *Molecular Therapy - Nucleic Acids*, 20:882–894, 2020. ISSN 2162-2531. doi: https://doi.org/10.1016/j.omtn.2020.05.006. URL https://www.sciencedirect.com/science/article/pii/S2162253120301323.

Jielu Yan, Bob Zhang, Mingliang Zhou, François-Xavier Campbell-Valois, and Shirley W. I. Siu. A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against escherichia coli using multi-branch-cnn and attention. *mSystems*, 8(4):e00345–23, 2023. doi: 10.1128/msystems.00345-23.

Zhang and Richard L. Gallo. Antimicrobial peptides. *Current Biology*, 26(1):R14–R19, 2016. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2015.11.017. URL https://www.sciencedirect.com/science/article/pii/S0960982215014098.

Qian Zhang and Yifan Liu. Hmd-amp: Protein language model-based hierarchical multi-label deep forest for antimicrobial peptide annotation. *Bioinformatics*, 2022.

Zonghan Zhang, Ramyasri Veerapaneni, Moses Ayoola, Athish Das, Zhiqian Chen, Bindu Nanduri, and Mahalingam Ramkumar. Leveraging graph neural networks for mic prediction in antimicrobial resistance studies. volume 2024, pp. 1–4, 07 2024. doi: 10.1109/EMBC53108.2024.10782350.

A Appendix

A.1 Related Work

Computational approaches for predicting antimicrobial peptide (AMP) activity have evolved from handcrafted molecular descriptors toward learned sequence representations. Traditional machine-learning methods represent peptides using physicochemical and compositional properties and employ models such as Random Forests Bajiya et al. (2025) and ensemble models for activity prediction Chung et al. (2024). Deep-learning approaches, including convolutional and attention-based architectures, learn activity-associated patterns directly from amino-acid sequences Yan et al. (2020; 2023). Graph-based approaches provide an alternative representation by explicitly modeling relational structures, including residue-level connectivity, structural interactions, or heterogeneous biological entities, and propagating information over these relations using graph neural networks Zhang et al. (2024). These approaches are particularly suitable when relationships among molecular or biological entities constitute an important component of the representation. In contrast, STAMP does not explicitly construct a molecular or heterogeneous graph; instead, it integrates complementary sequence-, motif-, and target-level information within a unified context-conditioned framework.

Protein language models (PLMs) have further expanded sequence-based AMP modeling by providing contextualized amino-acid representations learned through large-scale protein sequence pretraining. PLM-based approaches have increasingly been applied to AMP classification and activity prediction Zhang & Liu (2022); Wang et al. (2025). Complementary methods have incorporated microbial information to account for the target-dependent nature of antimicrobial activity Bae et al. (2025), while topic-model-based analyses have demonstrated that latent motif groups can capture recurring sequence patterns associated with physicochemical properties and antimicrobial activity Padi et al. (2025). These approaches capture different aspects of AMP activity: PLMs provide distributed contextual representations of amino-acid sequences, topic models provide an explicit representation of recurring motif composition, and microbial information captures the biological context in which activity is measured.

Several recent deep-learning frameworks provide additional context for biological representation and prediction. Pathomic Fusion Chen et al. (2022) is a multimodal framework developed for cancer prognosis and classification that integrates histopathology images with genomic information. ESM2AMP Haimeng et al. (2026) uses ESM-2 protein language model representations for AMP identification and continuous MIC prediction, employing strain-specific models across a collection of bacterial and fungal strains. ANIA Chiu et al. (2026) is directly related to continuous MIC prediction and transforms one-dimensional peptide sequences into two-dimensional Frequency Chaos Game Representation (FCGR) images, followed by Inception modules and a Transformer encoder for MIC prediction. LLAMP Bae et al. (2025) and ensemble-based approaches Chung et al. (2024) also incorporate microbial target information but employ target-specific prediction strategies. Collectively, these approaches demonstrate the value of learned sequence representations and microbial context for AMP activity prediction, while differing in their sequence representations, use of target information, and modeling strategies.

STAMP integrates these complementary directions within a unified cross-species MIC prediction framework. Specifically, it directly represents peptide sequences using the ESM-2 protein language model, incorporates LDA-derived topic representations to capture recurring sequence motifs, and conditions predictions on the target bacterial species. Unlike approaches that transform sequences into image representations or construct separate predictors for individual pathogens, STAMP uses a single shared architecture to jointly model peptide-species pairs across multiple bacterial targets. This design enables sequence context, latent motif composition, and microbial target identity to be represented jointly for continuous MIC prediction.

A.2 Dataset Construction and Preprocessing

We evaluated STAMP using four AMP activity datasets. Three datasets were previously introduced for AMP activity and MIC prediction, while the fourth one was independently curated from DBAASP-v3 in this work. The datasets differ in their species coverage, target organisms, and preprocessing procedures, providing complementary evaluation settings for assessing sequence-, motif-, and species-aware representations.

A.2.1 Previously Published Benchmark Datasets

The first benchmark was introduced by Bae et al. (2025) for species-aware MIC prediction. The dataset was curated from DBAASP-v3 and contains 9,181 unique AMP sequences and 36,818 peptide–species measurements spanning 35 bacterial species. Each peptide–species pair is associated with an experimentally measured MIC value in μM. The original curation retained species with at least 60 associated peptides and peptide sequences of length 5–50 amino acids. We use the predefined training, validation, and test partitions provided by the dataset authors.

The second benchmark was introduced by Chung et al. (2024) for AMP activity prediction against three priority bacterial strains: *S. aureus* ATCC 25923, *E. coli* ATCC 25922, and *P. aeruginosa* ATCC 27853. The dataset was assembled from multiple AMP repositories, including DBAASP, dbAMP, and DRAMP, and contains 8,920 AMP-strain measurements after preprocessing and duplicate removal. Peptide sequences were restricted to lengths of 6-40 amino acids. We use the predefined training, validation, and test partitions reported by the dataset authors.

The third benchmark was curated by Chiu et al. (2026) using AMP sequences with experimentally measured MIC values collected from three public databases: DBAASP, dbAMP, and DRAMP. As shown in Table 5, the dataset contains MIC measurements for three bacterial pathogens and provides predefined training and test splits, which we adopted for our experiments to enable a direct and consistent comparison with the reported benchmark results.

A.2.2 Our Curated DBAASP-v3 Dataset

To provide an additional evaluation benchmark based on a consistently processed and more recent collection of DBAASP-v3 measurements, we constructed a curated AMP activity dataset using a standardized preprocessing protocol. Existing DBAASP-derived datasets may differ in their treatment of duplicate sequences, conflicting measurements, and non-numeric MIC annotations. Our curation was designed to establish a consistent numeric MIC representation and reduce ambiguity arising from heterogeneous activity annotations.

AMP repositories contain measurements collected across different studies, experimental conditions, assay protocols, units, and reporting conventions. MIC values may therefore be reported as single values, ranges, inequalities, or measurements with associated variation. In particular, range-reported MIC values may appear in forms such as x–y or minimum–maximum values, and the ordering of the two values is not necessarily consistent across source datasets. Some previously curated datasets use the second reported value as the MIC value, implicitly treating it as the upper bound, which can result in inconsistent interpretation when the range is recorded in the opposite order.

To address these inconsistencies, we applied a standardized preprocessing protocol to the DBAASP-v3 records. All MIC measurements were converted to μM. For range-reported measurements, the larger of the reported values was retained as the MIC value, irrespective of their order in the original record. For measurements reported as $x \pm y$, the central value x was retained. For inequality-based measurements, the comparison symbol was removed and the corresponding numeric value was extracted according to the preprocessing protocol. This procedure provides a consistent numeric representation of MIC measurements for downstream regression analysis.

DBAASP contains repeated peptide sequences associated with different MIC measurements because of differences in experimental conditions or assay protocols. To reduce label inconsistency and prevent sequence overlap across data partitions, duplicate and conflicting sequence measurements were handled according to the dataset curation protocol. Starting from 43,477 collected records, the resulting dataset contains 18,831 peptide sequences associated with measurements for four target bacterial species. Additional dataset statistics are provided in Table 5.

A.2.3 Activity-Stratified Data Partitioning

Because MIC values exhibit a strongly non-uniform distribution, purely random partitioning can produce differences in activity distributions across the training, validation, and test sets. We therefore constructed quantile-based MIC bins before partitioning the curated dataset. The number of bins was selected according to the number of unique MIC values to avoid excessive fragmentation of smaller subsets.

The MIC bins were used only for stratification and were not provided as input features to STAMP. Samples were partitioned into training, validation, and test sets using an 80:10:10 ratio while approximately preserving the relative representation of low-, intermediate-, and high-MIC observations across the three partitions. This procedure provides more comparable activity distributions across the data splits for model development and evaluation.

For the two previously published benchmarks, we retained the predefined training, validation, and test partitions provided by the respective dataset authors. Detailed dataset statistics, including the numbers of sequences, peptide–species measurements, target species, and partition sizes, are summarized in Table 5.

Table 5: Datasets, species, and numbers of sequences used for model training, validation, and testing. Datasets A, B and C use the predefined train, validation, and test partitions provided with the respective benchmarks. For the curated dataset, we construct train, validation, and test partitions using quantile-based binning to preserve the MIC distribution across splits. Dataset A contains MIC measurements for 35 bacterial species; for compact presentation, only four representative species are shown, although all 35 species are included in model training and evaluation.

Database	**Species**	**Sequences**		
		Train	**Test**	**Validation**
Dataset A	*E. coli.*	6095	749	762
	S.aureus.	5518	695	695
	P.aeruginosa.	3978	509	507
	S. epidermidis.	1458	190	178
	Total	29394	3741	3683
Dataset B	*E. coli.*	2443	764	611
	S.aureus.	1692	529	423
	P.aeruginosa.	1572	492	394
	Total	5707	1785	1428
Dataset C	*E. coli.*	6626	1657	–
	S.aureus.	6065	1517	–
	P.aeruginosa.	4496	1125	–
	Total	17187	4299	–
Our Dataset	*E. coli.*	5296	654	589
	S.aureus.	4557	563	507
	P.aeruginosa.	3752	464	417
	S. epidermidis.	1646	204	182
	Total	15251	1885	1695

A.3 Experimental Setup and Configuration

For latent motif representation, LDA is applied to overlapping amino-acid 4-mers extracted with a stride of one, using $K = 100$ latent topics and symmetric Dirichlet priors $\alpha = 0.01$ and $\eta = 0.01$. For sequence representation, STAMP uses the 8M-parameter ESM-2 model, with peptide sequences truncated or padded to a maximum length of 50 residues. The ESM-2 representation, 100-dimensional LDA topic-mixture vector, and one-hot target-species encoding are jointly provided to an MLP with a 256-dimensional hidden layer and ReLU activation for MIC prediction. ESM-2 and the prediction network are jointly fine-tuned using AdamW with a learning rate of 1×10^{-5}, a batch size of 8 and a maximum of 1,000 training epochs with early stopping criteria, while the LDA topic-mixture representations and species encoding remain fixed during model training. For LDA-based representation learning, the LDA model was fitted exclusively on the training sequences to learn a set of latent topic distributions over the sequence features. The fitted LDA model was then used to transform both the training and test sequences into their corresponding topic representations. Thus, the same learned topic space was used consistently across the two splits, with the test sequences serving only as inputs to the transformation step. These LDA representations were subsequently used as input features for the downstream prediction experiments.

We model MIC using the scale $\log_{10}(\mu\mathrm{M})$ because MIC values span several orders of magnitude, and the log transformation reduces skewness and stabilizes the regression task. The metrics used to evaluate STAMP's MIC prediction performance are mean squared error (MSE), mean absolute error (MAE), coefficient of determination (R^2), and Pearson correlation coefficient (PCC). For the downstream AMP classification task, we report accuracy and weighted F1 score, together with receiver operating characteristic (ROC) and precision-recall (PR) curves.

A.4 Per-Species Metrics Across Datasets

Tables 6–8 report per-species regression metrics across all four datasets. Dataset A, the largest and most taxonomically diverse collection with 35 species, exhibits considerable performance variation: R^2 ranges from -0.965 (*Listeria innocua*) to 0.792 (*Streptococcus agalactiae*), reflecting the difficulty of generalizing across underrepresented species with limited samples. Dataset B, restricted to three well-studied pathogens, yields more uniform results (PCC ≈ 0.75, $R^2 \approx 0.55$). Our curated dataset achieves the strongest overall performance, with all four species exceeding a PCC of 0.80 and R^2 values between 0.65 and 0.73, while also recording the lowest MSE (0.171) and MAE (0.285) among all datasets. The Dataset C, combining data across sources, maintains competitive accuracy (overall PCC $= 0.798$,

$R^2 = 0.634$) despite its larger scale. Across all datasets, *E. coli*, *P. aeruginosa*, and *S. aureus* consistently appear as well-predicted species, suggesting that sufficient training representation and structural diversity are key drivers of predictive reliability.

Table 6: Per-species AMP activity metrics for Dataset A.

Species	**MSE**	**MAE**	R^2	**PCC**
Acinetobacter baumannii	0.237	0.376	0.553	0.755
Aeromonas hydrophila	0.417	0.473	−0.929	0.423
Aeromonas salmonicida	0.073	0.244	0.692	0.862
Bacillus cereus	0.496	0.496	−0.004	0.371
Bacillus megaterium	0.451	0.472	0.208	0.531
Bacillus subtilis	0.303	0.377	0.504	0.718
Enterobacter cloacae	0.201	0.347	−0.268	0.523
Enterococcus faecalis	0.331	0.421	0.333	0.612
Enterococcus faecium	0.310	0.432	0.338	0.620
E. coli	0.243	0.357	0.588	0.771
Klebsiella aerogenes	0.258	0.406	0.131	0.567
Klebsiella oxytoca	0.214	0.427	0.215	0.498
Klebsiella pneumoniae	0.236	0.362	0.567	0.769
Lactococcus lactis	0.339	0.349	−0.489	0.037
Listeria innocua	0.341	0.452	−0.965	−0.258
Listeria monocytogenes	0.489	0.502	0.371	0.621
Micrococcus luteus	0.491	0.488	0.406	0.664
Pasteurella multocida	0.114	0.271	0.715	0.882
Pectobacterium carotovorum	0.318	0.405	−0.311	0.393
Proteus mirabilis	0.252	0.400	0.492	0.751
Proteus vulgaris	0.145	0.318	0.620	0.796
P. aeruginosa	0.231	0.349	0.557	0.754
Pseudomonas syringae	0.130	0.279	0.473	0.738
Salmonella enterica	0.265	0.379	0.549	0.752
Serratia marcescens	0.603	0.594	0.260	0.561
Shigella dysenteriae	0.498	0.504	0.144	0.539
S. aureus	0.290	0.395	0.483	0.699
S. epidermidis	0.244	0.368	0.506	0.722
Staphylococcus haemolyticus	0.315	0.429	0.428	0.685
Stenotrophomonas maltophilia	0.447	0.538	0.047	0.439
Streptococcus agalactiae	0.111	0.266	0.792	0.900
Streptococcus mutans	0.237	0.353	0.313	0.677
Streptococcus pneumoniae	0.294	0.371	0.480	0.747
Streptococcus pyogenes	0.251	0.374	0.615	0.795
Vibrio parahaemolyticus	0.279	0.365	0.515	0.770
All Species	**0.275**	**0.385**	**0.531**	**0.732**

Table 7: Per-species AMP activity metrics for Dataset B.

Species	**MSE**	**MAE**	R^2	**PCC**
E. coli	0.253	0.359	0.551	0.757
P. aeruginosa	0.264	0.343	0.534	0.752
S. aureus	0.285	0.384	0.552	0.749
All Species	**0.260**	**0.363**	**0.566**	**0.758**

Table 8: Per-species AMP activity metrics for the Dataset C.

Species	MSE	MAE	R^2	PCC
E. coli	0.234	0.334	0.641	0.803
P. aeruginosa	0.188	0.310	0.659	0.814
S. aureus	0.232	0.341	0.599	0.776
All Species	**0.221**	**0.330**	**0.634**	**0.798**

Table 9: Per-species AMP activity metrics for our curated dataset.

Species	MSE	MAE	R^2	PCC
E. coli	0.194	0.302	0.685	0.828
P. aeruginosa	0.139	0.270	0.711	0.845
S. aureus	0.184	0.305	0.653	0.809
S. epidermidis	0.154	0.290	0.731	0.856
All Species	**0.171**	**0.285**	**0.698**	**0.837**

A.5 MIC-Based AMP Classification

To assess whether quantitative MIC predictions remain useful for peptide screening, predicted MIC values were mapped to AMP and non-AMP categories using the activity thresholds employed in the literature Bae et al. (2025): peptides with MIC < 25 μM were considered as active (AMP), whereas those with MIC > 100 μM were considered as inactive (non-AMP), while peptides with intermediate MIC values between 25 and 100 μM are excluded from the binary classification comparison. This convention is used only when comparing classification performance with existing methods. Separately, to characterize the relationship between the continuous MIC predictions and AMP activity across the full MIC range, we generated ROC and precision-recall (PR) curves by evaluating thresholds over all available samples, without excluding peptides with intermediate MIC values.

As shown in Table 10 and Figure 3, STAMP achieves classification accuracies greater than 80% across all four datasets, reaching approximately 89% on the curated dataset. The average precision and AUROC exceed 0.9, indicating that the predicted MIC values provide strong separation between active and inactive peptides and therefore can support the prioritization of low-MIC candidates for experimental evaluation. Prediction errors occur more frequently in the high-MIC regime, where fewer training observations are available, highlighting a limited representation of weakly active peptides as an important source of uncertainty. Figure 3 presents the detailed regression and classification results for the curated dataset, while the corresponding graphs for Datasets A and B are provided in Figures 4, 5 and 6.

Table 10: AMP classification based on STAMP-predicted MIC values for computational screening, using predefined MIC thresholds to distinguish active and inactive peptides.

Dataset	Classification	
	Accuracy↑	WF1↑
Dataset A	0.87	0.88
Dataset B	0.88	0.89
Dataset C	0.90	0.91
Our curated dataset	0.89	0.89

A.6 Prediction Error Bias Analysis

Additionally, to assess whether STAMP exhibits systematic prediction bias that may not be captured by aggregate regression metrics, we further analyze the direction and distribution of prediction errors (Table 11). The over and under-prediction errors are approximately balanced, indicating that there is no dominant directional bias in the MIC estimates. Although Datasets A, B and C exhibit greater error variability, the error distributions remain centered around zero. These results suggest that STAMP does not systematically over or underestimate MIC, providing additional evidence that its predictive performance is not driven by a consistent bias toward higher or lower AMP activity.

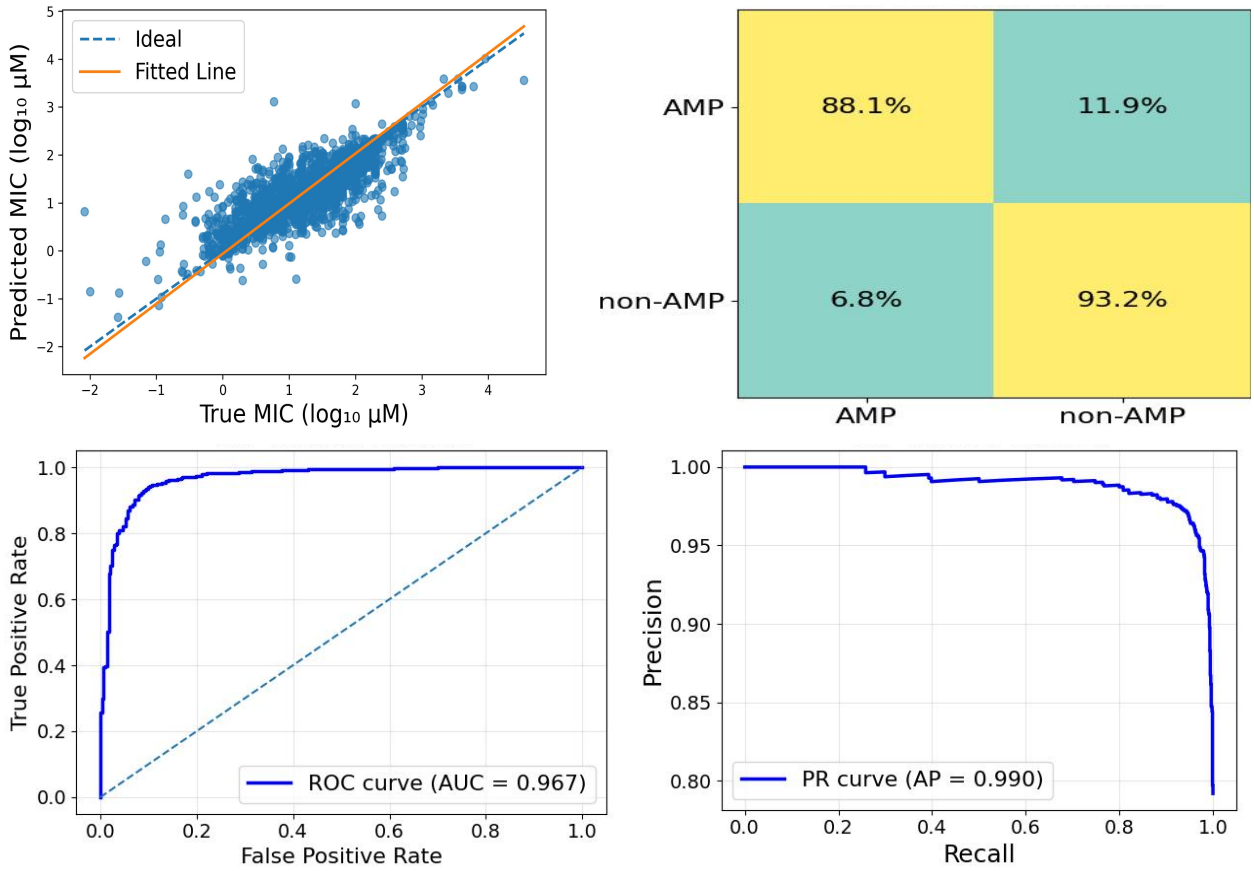

Figure 3: Performance evaluation of MIC prediction model and derived AMP classification on our curated dataset. The scatter plot demonstrates strong agreement between predicted and actual MIC values, indicating accurate regression performance. For downstream screening, predicted MIC values were categorized into AMP ($<$25 μM) and non-AMP ($>$100 μM) categories. The confusion matrix (%) summarizes classification performance. ROC and precision–recall curves show excellent discriminative capability (AUC$>$ 0.9, AP $>$ 0.9), highlighting the model's robustness and effectiveness in identifying AMPs for drug design analysis.

Table 11: Prediction error bias and dispersion across datasets. Errors are reported in $\log_{10}$(MIC) units. The mean over-prediction errors are comparable across datasets, while under-prediction errors show modest variation in magnitude. Overall, the errors remain centered around relatively small magnitudes, with no directional tendency toward systematic over- or underestimation of MIC values.

Dataset	**Over Predictions** ($\mu \pm \sigma$)	**Under Predictions** ($\mu \pm \sigma$)
Dataset A	0.396 ± 0.385	0.370 ± 0.340
Dataset B	0.373 ± 0.377	0.351 ± 0.356
Dataset C	0.355 ± 0.356	0.310 ± 0.315
Our Dataset	0.310 ± 0.312	0.279 ± 0.278

A.7 Discussions

The ablation study shows that sequence, motif, and species components provide non-redundant information, with the full STAMP model benefiting from their joint representation. On Dataset C, STAMP maintains robust predictive accuracy with an overall PCC of 0.798 and R^2 of 0.634, with *P. aeruginosa* achieving the highest per-species correlation (PCC = 0.814, R^2 = 0.659), followed by *E. coli* (PCC = 0.803, R^2 = 0.641) and *S. aureus* (PCC = 0.776, R^2 = 0.599). Despite being the largest dataset with 4299 samples for test dataset aggregated from multiple sources, the consistently high per-species correlations demonstrate that STAMP generalizes well even when data heterogeneity increases. The cross-species formulation enables data-sparse species to benefit from shared

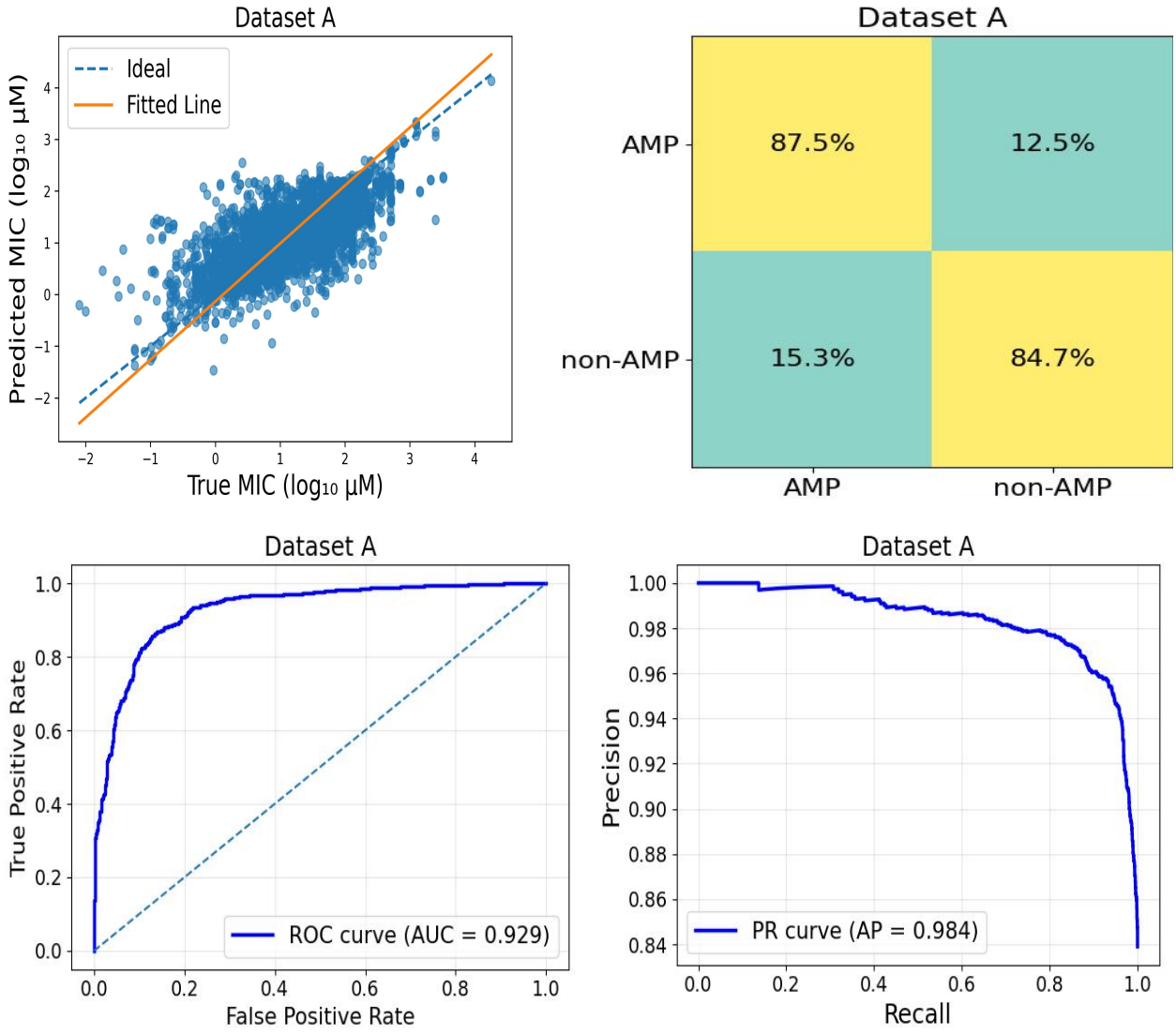

Figure 4: Performance evaluation of MIC prediction model and derived AMP classification on Dataset **A**. The scatter plot demonstrates strong agreement between predicted and actual MIC values, indicating accurate regression performance. For downstream screening, predicted MIC values were categorized into AMP ($<$25 μM) and non-AMP ($>$100 μM) categories. The confusion matrix (%) summarizes classification performance. ROC and precision–recall curves show excellent discriminative capability (AUC$>$ 0.9, AP $>$ 0.9), highlighting the model's robustness and effectiveness in identifying AMPs under class imbalance.

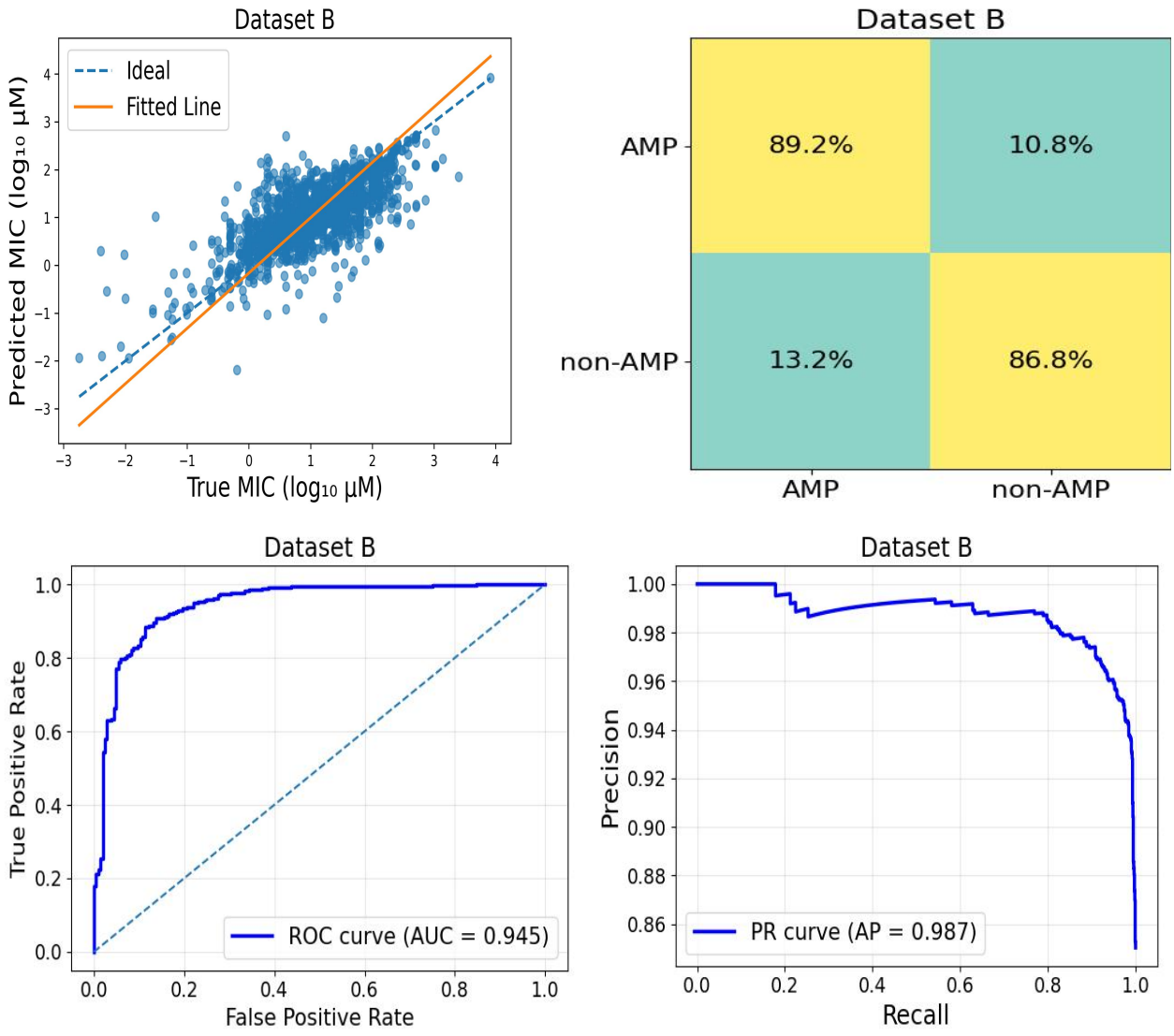

Figure 5: Performance evaluation of MIC prediction model and derived AMP classification on Dataset **B**. The scatter plot demonstrates strong agreement between predicted and actual MIC values, indicating accurate regression performance. For downstream screening, predicted MIC values were categorized into AMP ($<$25 μM) and non-AMP ($>$100 μM) categories. The confusion matrix (%) summarizes classification performance. ROC and precision–recall curves show excellent discriminative capability (AUC $>$ 0.9, AP $>$ 0.9), highlighting the model's robustness and effectiveness in identifying AMPs under class imbalance

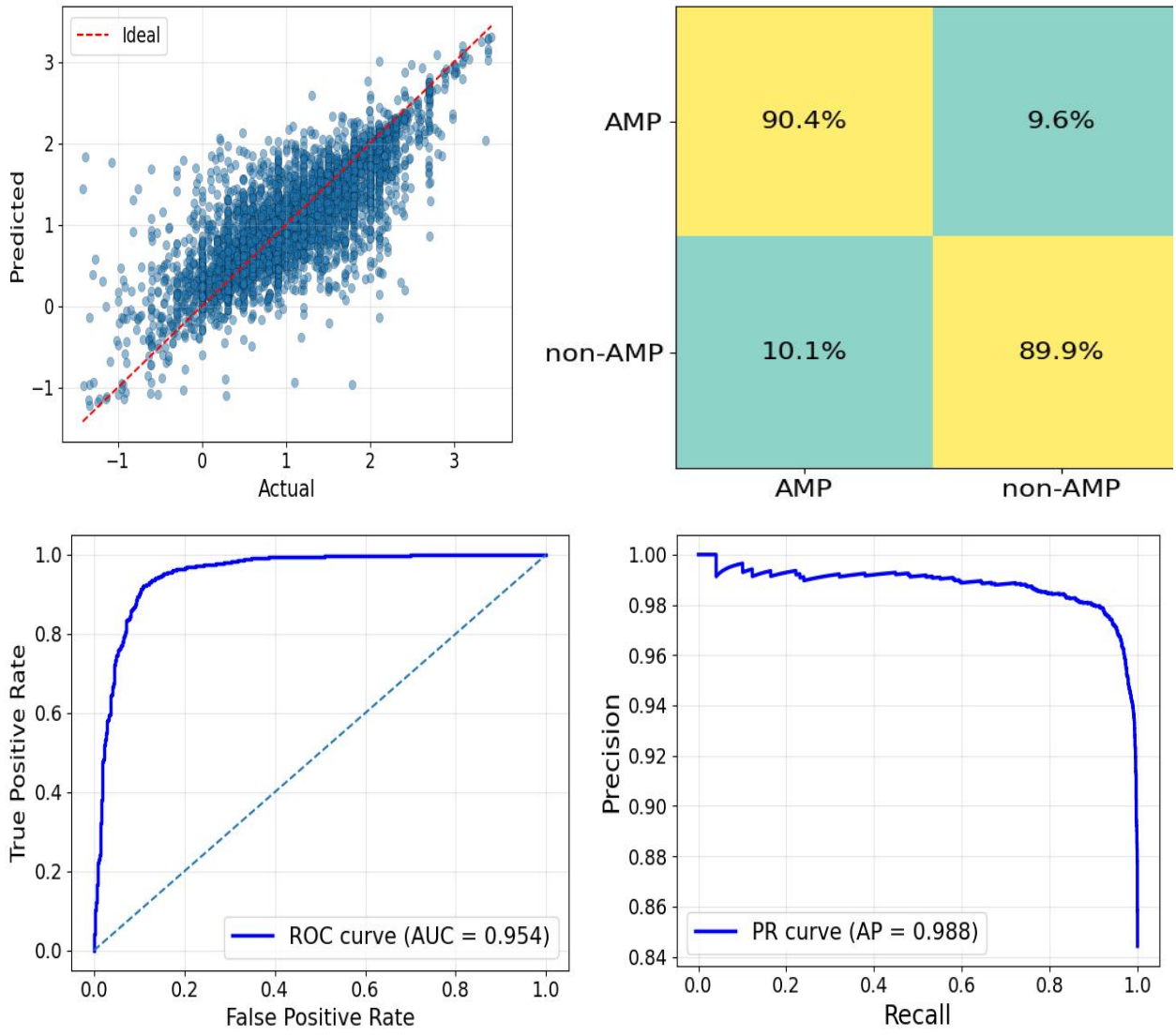

Figure 6: Performance evaluation of MIC prediction model and derived AMP classification on **Dataset C**. The scatter plot demonstrates strong agreement between predicted and actual MIC values, indicating accurate regression performance. For downstream screening, predicted MIC values were categorized into AMP ($<$25 μM) and non-AMP ($>$100 μM) categories. The confusion matrix (%) summarizes classification performance. ROC and precision–recall curves show excellent discriminative capability (AUC $>$ 0.9, AP $>$ 0.9), highlighting the model's robustness and effectiveness in identifying AMPs under class imbalance

sequence-activity relationships learned across the dataset, particularly in Dataset A spanning 35 bacterial species, while avoiding the need to construct and maintain separate models for individual targets.

Species-conditioned residue attribution provides additional insight into target-dependent activity differences. In TP4-derived peptides, differences in the contributions of C-terminal arginine clusters and hydrophobic core residues are consistent with known differences in interactions with Gram-negative and Gram-positive bacterial membranes. For example, TP4-noR5, which lacks the C-terminal poly-arginine tail, shows a greater reduction in activity against *E. coli* than against *S. epidermidis*. These results suggest that species context modulates the sequence regions used by the model for activity prediction, providing residue-level interpretability of cross-species activity differences. Because attribution identifies model associations rather than causal mechanisms, these observations should be interpreted as biologically informed hypotheses.

The P3 outlier, with a measured MIC of 64 μM versus a predicted value of approximately 15 μM against *E. coli*, illustrates the limitations of sequence-level representations in capturing effects such as structural or aggregation-related changes caused by individual substitutions. Such cases motivate future incorporation of structural and higher-order physicochemical information.

Several limitations define the scope of the current formulation. STAMP represents microbial context primarily through species identity, whereas MIC measurements can also depend on strain, assay protocol, and growth conditions. The fixed species vocabulary also prevents direct prediction for unseen pathogens without retraining. In addition, inequality-based MIC measurements are treated as exact continuous targets, which may introduce bias at the activity extremes, and the LDA representation remains fixed during model training. Future work can address these limitations by incorporating richer strain-level, genomic, phylogenetic, structural, and physicochemical context, jointly optimizing motif representations, and conducting broader prospective experimental validation, with the longer-term goal of improving generalization to microbial targets not represented during training.