AMPLE: A COST-AWARE ADAPTIVE MULTI-EXPERT FRAMEWORK FOR ANTIMICROBIAL PEPTIDE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Antimicrobial peptides (AMPs) are key components of innate immune defense and promising candidates for combating antimicrobial resistance, making accurate prediction of their diverse antibacterial, antifungal, antiviral, and antiparasitic activities important for identifying broad-spectrum anti-infective peptides. However, existing AMP prediction methods often rely on limited information sources or fixed strategies for integrating heterogeneous representations, limiting their ability to capture the distinct predictive characteristics of different types of biological information. To address this issue, we propose AMPLE, a cost-aware adaptive multi-expert learning framework for antimicrobial peptide activity prediction that models heterogeneous biological representations through specialized experts and selectively integrates their predictive capabilities within a unified architecture. AMPLE achieves state-of-the-art performance on the evaluated benchmark while substantially reducing unnecessary computation from expensive experts, demonstrating the effectiveness of adaptive expert acquisition for multilabel AMP activity prediction. Its design enables flexible performance–cost trade-offs across different computational budgets, providing a general framework for studying and modeling heterogeneous information in antimicrobial peptide prediction.

1 INTRODUCTION

Antimicrobial resistance (AMR) is progressively undermining the effectiveness of conventional antibiotics, creating an urgent need for new anti-infective strategies (MacNair et al., 2024). Antimicrobial peptides (AMPs), a diverse class of naturally occurring or engineered bioactive peptides, have emerged as promising candidates because of their broad functional activities and diverse mechanisms of action (Oliveira Júnior et al., 2025; Mookherjee et al., 2020). However, experimental discovery and functional characterization of AMPs remain costly and time-consuming, spurring the development of computational approaches for efficient candidate screening and activity prediction (Wan et al., 2024). AMP function can be characterized through complementary biological representations, including sequence information, physicochemical properties, protein language model representations, and structural characteristics. This heterogeneity motivates modeling complementary representations while accounting for their distinct predictive utility and computational cost.

Existing AMP predictors typically exploit either a single representation or a predefined subset of complementary information sources. Representative multimodal approaches combine sequence with structural information or pretrained protein representations with physicochemical descriptors, while more recent models incorporate richer combinations of structural, geometric, and physicochemical features (Ojeda et al., 2025; Liu et al., 2026; Sun et al., 2025). Although these studies demonstrate the benefits of integrating heterogeneous biological information, the representation set is generally fixed at model-design time, and the same representations are computed and fused for every peptide. Consequently, two questions remain insufficiently explored: how does the utility of different biological representations vary across samples and computational regimes, and does every peptide require the same expensive representations? These questions motivate moving beyond fixed pipelines toward sample-adaptive representation acquisition that can better balance predictive performance and computational cost.

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.

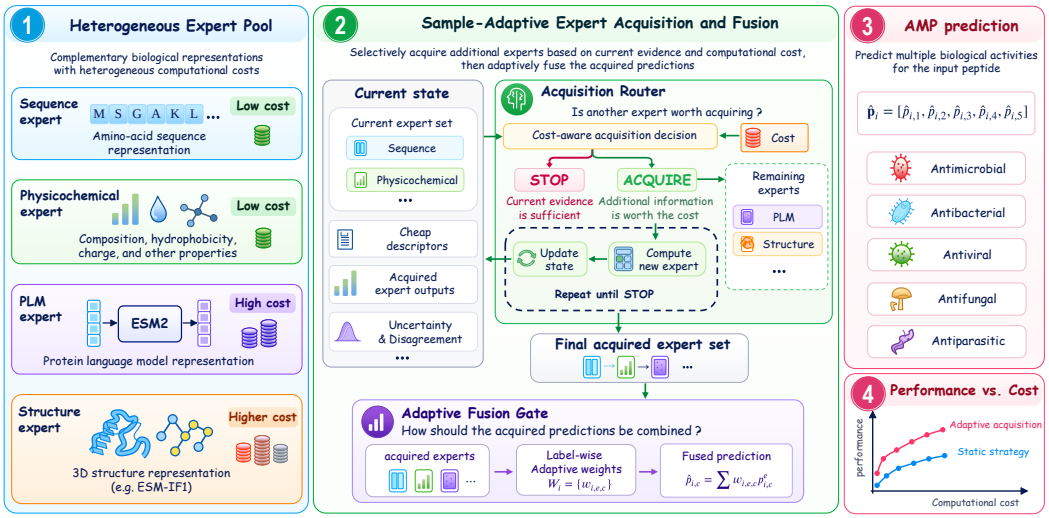

Figure 1: Overview of AMPLE. (1) Heterogeneous experts capture sequence, physicochemical, protein language model, and structural representations with different computational costs. (2) A sample-adaptive router selectively acquires additional experts, followed by label-wise adaptive fusion. (3) The fused prediction covers five AMP activities. (4) Adaptive acquisition enables different performance–cost operating points.

To address these questions, we propose AMPLE, a cost-aware adaptive multi-expert framework for antimicrobial peptide prediction (Fig. 1). AMPLE organizes sequence, physicochemical, protein language model, and structural representations as heterogeneous experts with distinct predictive capabilities and computational costs. Starting from lightweight sequence and physicochemical evidence, it selectively acquires expensive experts on a per-sample basis and adaptively fuses the acquired predictions through a decoupled acquisition–fusion design, enabling sample-specific computation under different deployment budgets.

Our experiments reveal substantial heterogeneity in expert utility across both samples and computational regimes. Expensive representations provide limited incremental benefit for many peptides, while the router preferentially allocates computation to samples with higher realized utility. Across different deployment budgets, AMPLE achieves a favorable empirical performance–cost frontier by reducing unnecessary expert computation while maintaining strong predictive performance.

Our main contributions are:

- **Heterogeneous expert analysis.** We systematically study sequence, physicochemical, protein language model, and structural experts within a unified framework, revealing substantial variation in their predictive utility across samples and computational regimes.
- **Cost-aware adaptive multi-expert framework.** We introduce AMPLE, which combines selective expert acquisition with adaptive prediction fusion in a decoupled framework, allowing different peptides to follow different computational paths while accounting for heterogeneous acquisition costs.
- **Performance–cost evaluation.** Across different deployment budgets, AMPLE achieves favorable empirical performance–cost trade-offs, substantially reducing expensive expert computation while preserving strong predictive performance.

2 Related Work

2.1 AMP Prediction with Heterogeneous Biological Representations

Computational antimicrobial peptide (AMP) prediction has progressively moved from handcrafted sequence and physicochemical descriptors toward learned representations that capture richer bio-

logical information. Early methods mainly relied on amino-acid composition, sequence-derived features, and physicochemical properties, while subsequent deep learning approaches began to learn task-specific representations directly from peptide sequences (Bhadra et al., 2018; Li et al., 2022; Lee et al., 2023). Protein language models further extended this paradigm by learning contextual and evolutionary information from large-scale protein sequence corpora, substantially enriching protein sequence representations (Rives et al., 2021; Lin et al., 2023). In parallel, structural information has received increasing attention as another important source of biological representation for AMP prediction (Sun et al., 2025).

Recent studies have increasingly combined multiple sources of biological information for AMP prediction. ESCAPE integrates sequence and structural representations for multilabel AMP classification, MAPLE jointly models protein language model representations derived from protein language models and physicochemical properties, and SSFGM further combines sequence information with predicted structural, geometric, and surface features (Ojeda et al., 2025; Liu et al., 2026; Sun et al., 2025). These studies demonstrate the potential benefits of combining heterogeneous biological representations and reflect a broader transition in AMP prediction from single-source modeling toward multimodal learning.

Despite this progress, most existing multimodal AMP methods determine the set of biological representations at model design time and apply the same representation configuration to all peptides (Ojeda et al., 2025; Liu et al., 2026; Sun et al., 2025). Consequently, prior work has mainly focused on how a predefined collection of representations should be encoded and fused, while paying less attention to whether every representation is equally necessary for every sample. This issue becomes particularly important when different biological information sources require substantially different amounts of computation. These observations motivate the study of more selective use of heterogeneous biological representations, rather than uniformly computing all representations for every peptide.

2.2 Adaptive Information Acquisition and Expert Routing

Adaptive computation provides a related perspective in which different samples need not follow the same computational path. Active feature acquisition (AFA) methods sequentially acquire informative features while balancing predictive quality against acquisition cost (Valancius et al., 2024; Guney et al., 2025). Dynamic modality selection further extends this principle to multimodal settings, allowing different input modalities to be selectively evaluated according to the current sample and computational constraints (Hu et al., 2024). These studies establish information acquisition itself as a sample-dependent decision rather than a fixed preprocessing procedure.

Mixture-of-experts (MoE) and conditional-computation methods address a complementary problem by dynamically activating or weighting subsets of experts for different inputs (Shazeer et al., 2017). Sparse routing can control computational cost through selective expert execution, while multimodal expert modeling provides a methodological basis for dynamically combining different information sources. However, much of this literature focuses on internal network modules, relatively homogeneous experts, or conventional input features and modalities. In AMP prediction, heterogeneous biological representations may instead arise from substantially different computational pipelines, including lightweight sequence models and physicochemical descriptors, large pretrained protein language models, and structure-aware pipelines involving protein structure prediction models such as AlphaFold and ESMFold (Rives et al., 2021; Lin et al., 2023; Jumper et al., 2021). Meanwhile, existing multimodal AMP methods still predominantly employ a fixed representation configuration across samples. These observations motivate the study of adaptive and cost-aware information acquisition for heterogeneous biological representations in AMP prediction.

3 Method

3.1 Problem Formulation and Framework Overview

We consider multilabel antimicrobial peptide prediction, where each peptide x_i is associated with a label vector $\mathbf{y}_i \in \{0,1\}^C$ over C target activities. We assume access to a heterogeneous expert pool $\mathcal{E} = \{E_1, \ldots, E_K\}$, where each expert captures a distinct source of biological information and produces an expert-specific output $\mathbf{z}_i^e = E_e(x_i)$.

Conventional multimodal prediction typically employs a predefined set of information sources and combines their outputs through a fixed fusion pipeline. We instead allow the set of evaluated experts to vary across individual peptides. Let $\mathcal{A}_i \subseteq \mathcal{E}$ denote the experts acquired for sample i. The prediction process can then be written abstractly as

$$\hat{\mathbf{p}}_i = F\left(\{\mathbf{z}_i^e\}_{e\in\mathcal{A}_i}\right), \tag{1}$$

where both the acquired expert set and the resulting computational path may depend on the sample.

AMPLE models this adaptive computation through two complementary components: an *acquisition router* determines whether additional expert information should be computed from the information currently available, while a *post-acquisition fusion module* determines how the acquired expert outputs should contribute to the final prediction once they become observable. To avoid additional representation alignment across heterogeneous modalities and reduce confounding effects introduced by complex feature-fusion mechanisms, we instantiate AMPLE at the prediction level by setting $\mathbf{z}_i^e = \mathbf{p}_i^e \in [0,1]^C$.

The acquisition router therefore governs which expert predictions are worth computing, whereas the fusion module governs how the observed predictions are combined, enabling sample-adaptive multimodal inference under heterogeneous computational costs.

3.2 Heterogeneous Expert Pool

In this work, we instantiate the heterogeneous expert pool as $\mathcal{E} = \{E_{\text{seq}}, E_{\text{phys}}, E_{\text{esm}}, E_{\text{struct}}\}$, covering task-specific sequence, physicochemical, pretrained evolutionary, and structural information. Although all four experts produce predictions in the same C-dimensional task space, they differ substantially in their information sources, representation mechanisms, and computational costs.

Sequence expert. E_{seq} provides a lightweight sequence-based prediction pathway. It directly takes the amino-acid sequence as input and employs a four-layer Transformer to model residue-level contextual dependencies. The resulting sequence representation is mapped to multilabel activity predictions $\mathbf{p}_i^{\text{seq}}$.Trained directly on the task-specific short-peptide data, it provides a lightweight and dataset-adapted sequence view.

Physicochemical expert. E_{phys} captures complementary information derived from physicochemical characteristics of the peptide, including charge, hydrophobicity, and amphipathicity, which are closely associated with antimicrobial activity. It models descriptors related to peptide composition and physicochemical behavior and maps the resulting representation to predictions $\mathbf{p}_i^{\text{phys}}$. These descriptors are inexpensive to obtain and provide complementary information to learned sequence representations.

Evolutionary expert. E_{esm} employs the pretrained ESM2-650M (Lin et al., 2023) protein language model to obtain residue-level representations. The ESM2 backbone is kept frozen, and its outputs are aggregated through attention pooling to construct a peptide-level representation, which is subsequently mapped to multilabel predictions $\mathbf{p}_i^{\text{esm}}$. This expert provides substantially richer pretrained protein representations at a higher computational cost than the lightweight experts.

Structural expert. E_{struct} captures structure-aware information using ESM-IF1 (Hsu et al., 2022). Residue-level representations extracted from the structural input are aggregated through attention pooling and mapped to predictions $\mathbf{p}_i^{\text{struct}}$. The structural pathway provides information complementary to sequence-based representations while introducing additional structure-related computation.

Together, the four experts provide complementary biological information with distinct predictive capabilities and computational costs. Their shared prediction space enables prediction-level integration, while these differences make uniform evaluation potentially inefficient and motivate selective expert acquisition.

3.3 ADAPTIVE EXPERT ROUTING

Conventional mixture-of-experts (MoE) models often use the same routing scores for expert activation and output weighting. In our setting, however, expert execution corresponds to acquiring a heterogeneous and potentially expensive source of biological information. The acquisition decision must therefore estimate the expected benefit of an unobserved expert relative to its computational cost, whereas post-acquisition fusion determines the contribution of an expert after its prediction has been observed. This distinction is particularly important for our heterogeneous expert pool, whose components are independently trained, partly frozen, and may exhibit sample-dependent reliability due to representation-specific noise or distribution mismatch. We therefore separate the acquisition router R_θ from the fusion gate G_ϕ: the former decides whether additional information is worth acquiring, while the latter combines the acquired predictions based on their realized predictive evidence.

Expert acquisition. For sample i, let $\mathcal{A}_i^{(t)} \subseteq \mathcal{E}$ denote the set of experts acquired up to step t. We represent the information available to the acquisition router as

$$S_i^{(t)} = \left(\mathcal{A}_i^{(t)}, P_i^{(t)}, U_i^{(t)}, D_i\right), \tag{2}$$

where $P_i^{(t)} = \{\mathbf{p}_i^e\}_{e\in\mathcal{A}_i^{(t)}}$ contains the predictions of the acquired experts, $U_i^{(t)}$ denotes uncertainty or disagreement statistics derived from the currently available predictions, and D_i contains inexpensive sample-level descriptors such as sequence length, charge, and cysteine content. Importantly, the state contains only information available before the next acquisition decision and never includes predictions from unacquired experts.

Given the current state, the acquisition router determines the additional experts to be evaluated:

$$\Delta\mathcal{A}_i^{(t)} = R_\theta\left(S_i^{(t)}\right), \qquad \Delta\mathcal{A}_i^{(t)} \subseteq \mathcal{E} \setminus \mathcal{A}_i^{(t)}. \tag{3}$$

The acquired expert set is then updated as

$$\mathcal{A}_i^{(t+1)} = \mathcal{A}_i^{(t)} \cup \Delta\mathcal{A}_i^{(t)}. \tag{4}$$

The acquisition process terminates when $\Delta\mathcal{A}_i^{(t)} = \varnothing$.

In our instantiation, the lightweight Sequence and Physicochemical experts form the initial expert set, $\mathcal{A}_i^{(0)} = \{E_{\text{seq}}, E_{\text{phys}}\}$. This design is motivated by their favorable cost–utility profile. As shown in Tables 1 and 2, the Sequence expert already provides strong standalone predictive performance at low latency, while the Physicochemical expert introduces only a small additional computational overhead compared with the substantially more expensive ESM2 and Structure experts. Using these two lightweight experts as the common starting point also provides richer initial evidence while reducing the branching complexity of the subsequent acquisition problem.

We further adopt sequential acquisition, restricting each routing step to at most one additional expert, $|\Delta\mathcal{A}_i^{(t)}| \leq 1$. This choice is supported by both computational and decision-making considerations. On a single RTX 3090 GPU, concurrent execution of multiple expensive experts provides little latency reduction because of resource contention, with the measured latency of ESM2+Structure and all-expert execution remaining close to the sum of their individual costs (Table 1). More importantly, sequential acquisition allows each newly observed expert prediction to update the router state before the next decision, enabling subsequent acquisition to be conditioned on richer evidence and allowing inference to terminate early when the current information is sufficient.

Post-acquisition fusion. After acquisition terminates, let $\mathcal{A}_i$ denote the final set of evaluated experts. The fusion gate G_ϕ takes the final routing state and produces label-wise normalized weights $w_{i,e,c}$ over the acquired experts. The final prediction is

$$\hat{p}_{i,c} = \sum_{e\in\mathcal{A}_i} w_{i,e,c} p_{i,e,c}. \tag{5}$$

Table 1: Inference latency (ms/sample, batch size 1) on a single NVIDIA RTX 3090. Multi-expert configurations are executed in parallel.

Expert configuration	$L = 10$	$L = 25$	$L = 50$	$L = 75$	$L = 100$	$L = 150$	$L = 200$	$L = 250$
Sequence	2.46	2.37	2.40	2.36	2.37	2.39	2.36	2.39
Physicochemical	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.77
ESM2	20.34	19.64	19.43	19.27	19.53	22.60	27.22	29.19
Structure (ESM-IF)	18.95	24.17	41.07	56.29	71.24	100.17	131.32	161.76
ESM2 + Structure (parallel)	37.44	43.72	59.48	74.21	89.55	121.41	154.61	186.57
All 4 experts (parallel)	40.31	46.67	62.42	77.38	92.70	124.73	157.87	189.38

3.4 Training and Cost-Aware Optimization

AMPLE is optimized in stages. The experts are first trained independently for multilabel activity prediction. For expert E_e, we denote its task-specific parameters by ψ_e and optimize

$$\mathcal{L}_e(\psi_e) = \mathbb{E}_{(x_i,y_i)} \left[\ell_{\mathrm{BCE}} \left(y_i, \mathbf{p}_i^e \right) \right]. \tag{6}$$

For experts built on frozen pretrained backbones, only the task-specific prediction components are optimized. After expert training, all expert predictors are kept fixed during subsequent fusion and routing optimization.

After expert training, the expert predictors are kept fixed and the adaptive acquisition and fusion components are optimized using cross-fitted out-of-sample expert predictions. The fusion gate G_ϕ is optimized for predictive quality,

$$\mathcal{L}_{\mathrm{fusion}}(\phi) = \mathbb{E} \left[\ell_{\mathrm{BCE}} \left(\mathbf{y}_i, \hat{\mathbf{p}}_i \right) \right], \tag{7}$$

while the acquisition router R_θ is optimized to balance predictive utility against the cost of acquiring additional information.

For the router objective, let $\pi_\theta(a \mid S_i^{(t)})$ denote the router distribution over feasible actions, and let $\hat{\mathbf{p}}_i^{(a)}$ and $c_i(a)$ denote the resulting prediction and incremental computational cost of action a, respectively. We optimize

$$\mathcal{L}_{\mathrm{route}}(\theta) = \mathbb{E}_{i,t} \left[\sum_a \pi_\theta(a \mid S_i^{(t)}) \left(\ell_{\mathrm{BCE}} \left(y_i, \hat{\mathbf{p}}_i^{(a)} \right) + \lambda c_i(a) \right) \right], \tag{8}$$

where λ is a fixed regularization coefficient that encourages cost-aware acquisition during training. In our experiments, $c_i(a)$ is instantiated using measured length-dependent inference latency.

At inference time, soft routing outputs are converted into hard acquisition decisions using expert-specific thresholds. Let $z_{i,a}^{(t)}$ denote the router score for action a. The routing decision is

$$a_i^{(t)} = \arg \max_{a \in \varnothing \cup (\mathcal{E} \setminus \mathcal{A}_i^{(t)})} \left(z_{i,a}^{(t)} - \tau_a \right), \qquad \tau_\varnothing = 0, \tag{9}$$

Increasing τ_e raises the evidence required to acquire expert e, making its acquisition more conservative and reducing its invocation rate. The training-time coefficient λ shapes the learned cost-aware acquisition policy, whereas the thresholds control the deployment operating point of the trained router.

We independently vary the acquisition thresholds for ESM2 and Structure to obtain different expert invocation rates and computational operating points. This enables a single trained AMPLE model to support different deployment budgets without retraining the experts, fusion module, or router. The resulting configurations are evaluated in terms of predictive performance and computational cost, from which we construct the empirical performance–cost Pareto frontier.

4 Experiments

4.1 Experimental Setup

Dataset and evaluation protocol. We evaluate AMPLE on the ESCAPE benchmark (Ojeda et al., 2025), a standardized dataset for multilabel antimicrobial peptide classification containing 21,409

Table 2: Predictive performance and computational efficiency on ESCAPE. *w/o Adaptive Fusion* uses global fusion with the same router. * For published methods, mAP and F1 are taken from ESCAPE (Ojeda et al., 2025), while latency and FLOPs are measured by us. † amPEPpy is evaluated using a GPU-backed implementation; the released sklearn CPU pipeline requires 58.9 ms/sample. ‡ FLOPs are not reported for amPEPpy because its random-forest inference is not directly comparable with neural-network FLOPs. § This configuration corresponds to AMPLE-Best w/o Router: all four experts are evaluated for every sample with sample-adaptive fusion.

Method	Predictive Performance		Computational Cost	
	mAP (%) ↑	F1 (%) ↑	Latency (ms/sample) ↓	FLOPs (G) ↓
Published methods*				
AMPs-Net	54.6 ± 0.86	57.7 ± 0.70	26.28	43.93
TransImbAMP	64.9 ± 1.11	62.0 ± 0.70	9.12	14.19
AMP-BERT	66.9 ± 1.17	64.7 ± 0.64	20.04	155.91
PEP-Net	68.4 ± 0.53	65.5 ± 0.61	29.06	209.16
amPEPpy	68.5 ± 0.48	66.5 ± 0.37	3.14†	N/A‡
AVP-IFT	68.8 ± 0.50	66.5 ± 0.59	<u>2.66</u>	<u>0.17</u>
AMPlify	70.3 ± 0.87	68.5 ± 0.77	4.10	3.13
ESCAPE Baseline	72.1 ± 0.60	69.8 ± 0.43	5.83	7.18
Expert configuration				
Seq expert	70.9 ± 0.20	68.7 ± 0.45	2.90	0.57
Phys expert	64.0 ± 0.14	61.2 ± 0.14	**1.01**	**0.04**
ESM2 expert	74.0 ± 0.23	70.6 ± 0.13	26.34	113.70
Struct expert	59.3 ± 0.46	56.5 ± 0.33	65.11	153.87
All 4 experts (global fusion)	75.1 ± 0.15	72.1 ± 0.35	95.37	268.18
All 4 experts (sample-adaptive fusion)§	$\mathbf{76.1 \pm 0.06}$	$\mathbf{72.6 \pm 0.17}$	95.37	268.18
AMPLE (Ours)				
AMPLE-Efficient	74.8 ± 0.06	71.6 ± 0.17	7.83	5.76
⌞ w/o Adaptive Fusion	74.4 ± 0.06	70.9 ± 0.07	7.83	5.76
AMPLE-Best	<u>76.1 ± 0.19</u>	<u>72.3 ± 0.24</u>	23.89	44.03
⌞ w/o Adaptive Fusion	75.6 ± 0.16	71.7 ± 0.11	23.89	44.03

antimicrobial peptides and 60,950 non-AMP sequences with five activity labels: antibacterial, antiviral, antifungal, antiparasitic, and antimicrobial. Following the official protocol, we use the predefined two-fold split for model development and evaluate all methods on the independent test set. Predictions from the two fold-specific models are averaged at test time. To avoid information leakage during adaptive learning, the acquisition router R_θ and adaptive fusion module G_ϕ are trained using cross-fitted out-of-sample expert predictions.

Baselines and metrics. We compare AMPLE with published ESCAPE methods, including AMPs-Net, TransImbAMP, AMP-BERT, amPEPpy, PEP-Net, AVP-IFT, AMPlify, and the ESCAPE baseline, as well as individual experts, fixed expert combinations, random acquisition, a coupled-router baseline, and full expert computation. We use mean average precision (mAP) as the primary predictive metric and additionally report F1 score. Computational efficiency is evaluated using batch-1 FP32 inference latency on a single NVIDIA RTX 3090 GPU and analytical FLOPs. Table 1 summarizes the computational costs of the expert pathways across representative peptide lengths.

4.2 Overall Performance and Computational Efficiency

Table 2 compares AMPLE with published methods, individual experts, and full-computation multi-expert configurations. AMPLE-Best achieves 76.1% mAP and 72.3% F1, improving over the ESCAPE Baseline by 4.0 and 2.5 percentage points, respectively. Importantly, this improvement does not rely on uniformly applying expensive representations to every peptide. AMPLE-Efficient reaches 74.8% mAP with only 5.76 GFLOPs, outperforming the standalone ESM2 expert by 0.8 mAP points while reducing latency from 26.34 to 7.83 ms and computation from 113.70 to 5.76

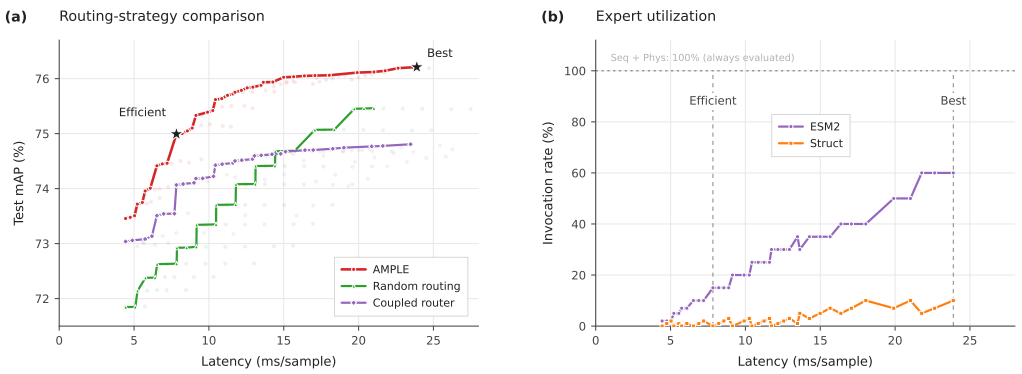

Figure 2: Cost-aware adaptive expert acquisition. (a) Performance–cost frontiers of AMPLE, random routing, and the coupled-router baseline; faint markers denote evaluated configurations. (b) ESM2 and Struct invocation rates along the AMPLE frontier; Seq and Phys are always evaluated.

GFLOPs. These results show that selectively acquiring an expensive representation can be substantially more efficient than evaluating it for every sample.

The router removes redundant computation. To isolate the effect of R_θ, we compare AMPLE-Best with a full-computation counterpart in which all four experts are always evaluated, with sample-adaptive fusion used in both settings. Both achieve 76.1% mAP, while AMPLE-Best reduces latency from 95.37 to 23.89 ms/sample and FLOPs from 268.18 to 44.03 G, corresponding to approximately 75% lower latency and 84% fewer FLOPs. This shows that sample-adaptive acquisition removes substantial redundant computation with little loss in predictive performance.

Adaptive fusion improves the use of acquired experts. We next isolate the effect of the fusion module G_ϕ while keeping the routing decisions unchanged. Replacing sample-adaptive fusion with global fusion decreases mAP from 74.8% to 74.4% at the Efficient operating point and from 76.1% to 75.6% at the Best operating point, with identical computational cost in each case. A similar effect is observed under full computation, where sample-adaptive fusion improves mAP from 75.1% to 76.1% while all four experts are evaluated in both cases. Together with the router ablation above, these results validate the effectiveness of both adaptive components: R_θ substantially reduces computation with little loss in predictive performance, while G_ϕ consistently improves prediction quality at fixed computational cost.

4.3 Adaptive Acquisition and Expert Utilization

We next examine how AMPLE allocates computation across samples. Figure 2a compares AMPLE with random acquisition and a coupled-router baseline, in which a single module jointly controls expert acquisition and prediction fusion. Because the ESM2 and Struct acquisition thresholds are varied independently, different expert-allocation policies can yield similar average latency, resulting in multiple configurations at the same or nearby cost. Across a broad range of computational budgets, AMPLE achieves higher predictive performance at comparable latency. The random baseline shows that the gain comes from allocating additional computation to appropriate samples rather than merely reducing expert invocation, while the coupled-router comparison demonstrates the benefit of decoupling acquisition from fusion.

Figure 2b shows how AMPLE allocates optional computation across the expensive experts. By design, Seq and Phys are always evaluated as the low-cost base experts, while ESM2 and Struct are selectively acquired by the router. At the Efficient operating point, ESM2 is invoked for approximately 15% of samples while Struct is effectively unused. At the Best operating point, the ESM2 invocation rate increases to approximately 60%, whereas Struct remains highly selective at around 10%. Thus, additional budget is not distributed uniformly: most of it is allocated to ESM2, while the more expensive structural pathway is reserved for a small subset of peptides.

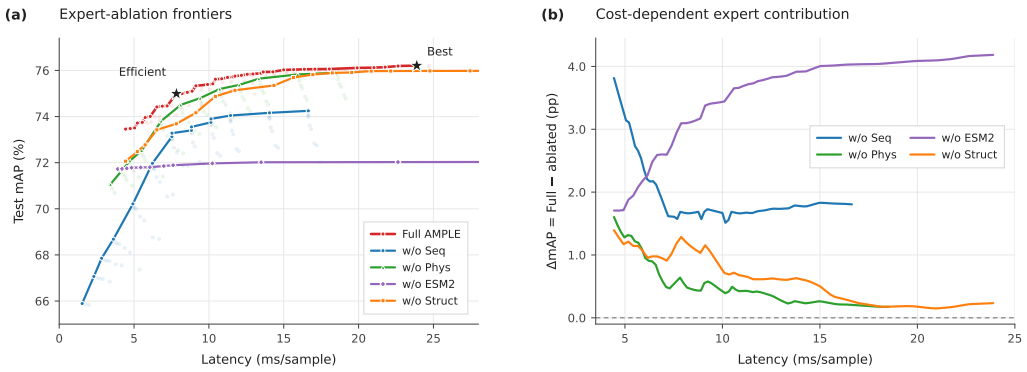

Figure 3: Expert ablation across cost regimes. (a) Performance–cost frontiers of AMPLE and leave-one-expert-out variants. (b) Cost-dependent expert contributions over overlapping cost ranges.

To assess whether selective routing targets samples that actually benefit from additional information, we define per-sample expert utility as the reduction in BCE between the fused predictions immediately before and after expert acquisition. For ESM2, the mean utility is 0.0160 on acquired samples versus 0.0014 on non-acquired samples, an 11.4× difference. For Struct, the corresponding values are 0.0383 and 0.0058, yielding a 6.6× difference. These results show that the router preferentially allocates expensive computation to samples with higher realized expert utility.

4.4 Heterogeneous Expert Utility across Computational Regimes

The heterogeneous experts differ not only in standalone predictive strength but also in how their contributions change with computational budget. We evaluate this effect using leave-one-expert-out variants, retraining the adaptive components over the remaining expert pool. Although Seq and Phys are always evaluated in the full AMPLE model, we remove them in these ablations to quantify their contributions and assess their suitability as the low-cost base experts. Figure 3 shows that removing different experts changes the performance–cost frontier to different degrees, indicating that the value of each representation depends on the operating regime rather than being uniformly additive.

Figure 3b compares expert contributions at matched computational cost. Seq and Phys contribute most in low-cost regimes, whereas ESM2 becomes increasingly important and dominates at higher budgets. Struct shows a smaller overall contribution but provides selective complementary gains. These shifts show that no fixed expert subset is uniformly effective across computational regimes, motivating AMPLE's adaptive acquisition and fusion.

5 Conclusion

This work presents AMPLE, a cost-aware multi-expert framework for multilabel antimicrobial peptide prediction that adaptively acquires heterogeneous biological representations on a per-sample basis. Our analyses show that these representations differ substantially in predictive utility and computational cost, with their value varying across samples and computational regimes. AMPLE addresses this heterogeneity with a decoupled framework that combines selective expert acquisition with adaptive prediction fusion. On ESCAPE, AMPLE achieves strong predictive performance while substantially reducing unnecessary expert computation, and adaptive fusion further improves prediction quality at fixed routing cost. These results demonstrate that efficient AMP prediction benefits from allocating information selectively rather than computing representations uniformly.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and to draft selected sections of the paper. All AI-assisted text was carefully reviewed, revised, and verified by the authors for accuracy, clarity, and consistency with the underlying methods and experimental results. We did not use generative AI tools for the other purposes listed in the submission form. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is based on the publicly available ESCAPE benchmark and does not involve human subjects or private data. AMPLE is developed for computational prediction of antimicrobial peptide activities and is not intended for direct therapeutic design or clinical decision-making. Predictions made by the model should be interpreted with appropriate caution and require experimental validation before practical use. We encourage the use of this method in accordance with established ethical, biosafety, and regulatory standards.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the expert architectures, adaptive routing and fusion mechanisms, training objectives, evaluation protocol, and computational-cost measurements in the main paper and appendix. The ESCAPE benchmark and its predefined data splits are publicly available, and all reported adaptive components are trained using cross-fitted out-of-sample expert predictions to avoid information leakage. Additional end-to-end cost analyses, including ESMFold-based structure acquisition, are provided in the appendix. Code and configuration files will be released upon publication.

REFERENCES

Pratiti Bhadra, Jielu Yan, Jinyan Li, Simon Fong, and Shirley WI Siu. Ampep: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific reports*, 8(1):1697, 2018.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B Kolachalama. Active feature acquisition via explainability-driven ranking. In *Forty-second International Conference on Machine Learning*, 2025.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 8946–8970, 2022.

Bodun Hu, Le Xu, Jeongyoon Moon, Neeraja J. Yadwadkar, and Aditya Akella. Mosel: Inference serving using dynamic modality selection. In *EMNLP*, pp. 8872–8886, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

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.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Hao Liu, Yi Shi, Feiyu Guo, Jinyi Wang, Jiaqian Li, Guangji Wang, De-Chuan Zhan, Haiping Hao, and Guo Yu. Maple: interpretable deep learning identifies selective antimicrobial peptides using joint evolutionary–physicochemical analysis. *Briefings in Bioinformatics*, 27(3):bbag318, 2026.

Craig R MacNair, Steven T Rutherford, and Man-Wah Tan. Alternative therapeutic strategies to treat antibiotic-resistant pathogens. *Nature Reviews Microbiology*, 22(5):262–275, 2024.

Neeloffer Mookherjee, Marilyn A Anderson, Henk P Haagsman, and Donald J Davidson. Antimicrobial host defence peptides: functions and clinical potential. *Nature reviews Drug discovery*, 19(5):311–332, 2020.

Sebastian Ojeda, Rafael Velasquez, Nicolás Aparicio, Juanita Puentes, Paula Cárdenas, Nicolás Andrade, Gabriel González, Sergio Rincón, Carolina Muñoz Camargo, and Pablo Arbelaez. A standardized benchmark for multilabel antimicrobial peptide classification. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, 2025. doi: 10.52202/085713-4303.

Nelson G Oliveira Júnior, Camila M Souza, Danieli F Buccini, Marlon H Cardoso, and Octávio L Franco. Antimicrobial peptides: structure, functions and translational applications. *Nature Reviews Microbiology*, 23(11):687–700, 2025.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Zehua Sun, Jing Xu, Yumeng Zhang, Yiwen Zhang, Zhikang Wang, Xiaoyu Wang, Shanshan Li, Yuming Guo, Hsin Hui Shen, and Jiangning Song. Multimodal geometric learning for antimicrobial peptide identification by leveraging alphafold2-predicted structures and surface features. *Briefings in Bioinformatics*, 26(3):bbaf261, 2025.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In *Forty-first International Conference on Machine Learning*, 2024.

Fangping Wan, Felix Wong, James J Collins, and Cesar de la Fuente-Nunez. Machine learning for antimicrobial peptide identification and design. *Nature Reviews Bioengineering*, 2(5):392–407, 2024.

A End-to-End Cost Analysis Including Structure Prediction

The main analysis reports the inference cost of each expert conditional on its required input representation being available. This choice reflects the lack of a single uniform and directly comparable structure-acquisition pipeline across the ESCAPE benchmark. Structural inputs may come from experimentally determined structures when available, whereas peptides without such structures require computational prediction. Moreover, different structure-prediction systems, such as AlphaFold or ESMFold, can incur substantially different computational costs. Including a particular upstream predictor in the main analysis would therefore make the reported structural cost depend on an additional implementation choice that is not intrinsic to the Struct expert itself. We consequently isolate the inference cost of the structure-aware expert in the main analysis and treat structure acquisition separately.

We further ask whether the main performance–cost conclusions remain valid when the cost of obtaining structural inputs is explicitly included. To isolate the effect of this stricter cost accounting, we keep the learned routing policy fixed and uniformly add ESMFold structure-acquisition cost whenever Struct is invoked. Specifically, we run ESMFold for all samples to measure its length-dependent latency and FLOPs, and use these measurements to recompute the end-to-end cost of each routing configuration. Keeping the policy fixed is intentional: retraining the router with ESMFold-inclusive costs would introduce a different optimization objective and confound the effect of cost accounting with changes in the learned routing policy.

Under this analysis, expert outputs, routing decisions, invocation rates, mAP, and F1 remain unchanged; only the computational cost assigned to the Struct pathway is updated. The resulting comparison therefore directly tests the sensitivity of AMPLE's performance–cost conclusions to whether upstream structure-prediction cost is included. As shown below, the absolute cost scale changes substantially for configurations that acquire Struct, while the qualitative conclusions regarding selective acquisition and the relative performance–cost advantage of AMPLE remain unchanged.

Table 3 shows that upstream structure prediction dominates the end-to-end cost of the structural pathway, with ESMFold latency ranging from approximately 0.8 to 4.3 seconds over the evaluated peptide lengths. Consequently, configurations that require structural information for every sample become substantially more expensive under this stricter cost accounting.

Table 4 shows how this revised cost accounting affects the main operating points while keeping predictions and routing decisions fixed. AMPLE-Efficient is unaffected because it never acquires Struct, remaining at 74.8% mAP with 7.83 ms latency and 5.76 GFLOPs. AMPLE-Best increases from 23.89 to 150.53 ms/sample because Struct is acquired for approximately 10% of samples, but remains far below full four-expert evaluation at 1361.77 ms/sample. At the reported precision, AMPLE-Best therefore retains the same 76.1% mAP as full-computation sample-adaptive fusion while requiring only about one ninth of its end-to-end latency. Under the same uniform ESMFold-inclusive accounting, the ESCAPE baseline reaches 1272.23 ms/sample, further illustrating the cost of uniformly requiring structural information.

Figure 4 replots the same learned routing policy under end-to-end structural cost accounting. Including ESMFold shifts configurations that acquire Struct substantially toward higher latency, while predictions, routing decisions, and invocation rates remain unchanged. Over the overlapping cost ranges, AMPLE still achieves higher predictive performance than random routing and the coupled-router baseline. ESM2 remains the primary additional expert, reaching approximately 60% invocation, whereas Struct is acquired for only about 10% of samples at the Best operating point. The large end-to-end cost associated with this small Struct invocation rate further emphasizes the value of selective structural acquisition.

Figure 5 shows that the qualitative pattern of expert utility is also preserved. Seq and Phys remain most important in low-cost regimes, whereas ESM2 becomes increasingly important as more computation becomes available. Struct provides more localized complementary gains at substantially higher end-to-end cost. Thus, including structure-prediction cost changes the absolute computational scale but not the qualitative finding that expert utility and acquisition cost are highly heterogeneous across computational regimes.

Overall, this fixed-policy sensitivity analysis shows that the main performance–cost conclusions are robust to whether upstream structure-prediction cost is included. AMPLE-Efficient remains unaf-

Table 3: End-to-end inference latency (ms/sample, batch size 1) when ESMFold structure prediction is included in the Struct pathway. Multi-expert configurations report measured parallel execution latency.

Expert configuration	$L = 10$	$L = 25$	$L = 50$	$L = 75$	$L = 100$	$L = 150$	$L = 200$	$L = 250$
Sequence	2.46	2.37	2.40	2.36	2.37	2.39	2.36	2.39
Physicochemical	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.77
ESM2	20.34	19.64	19.43	19.27	19.53	22.60	27.22	29.16
Structure (ESM-IF + ESMFold)	842.91	1043.38	920.54	954.20	968.44	1765.18	2861.66	4449.16
ESM2 + Structure (parallel)	861.40	1062.93	938.94	972.12	986.75	1786.42	2884.95	4473.96
All 4 experts (parallel)	864.27	1065.88	941.89	975.29	989.90	1789.75	2888.21	4476.78

Table 4: Effect of including ESMFold on configurations involving the Struct pathway. Latency and FLOPs are shown as main-analysis cost → end-to-end cost; configurations without structural acquisition are unchanged.

Method	**mAP (%)** ↑	**F1 (%)** ↑	**Latency (ms)** ↓	**FLOPs (G)** ↓
ESCAPE Baseline	72.1 ± 0.60	69.8 ± 0.43	5.83 → 1272.23	7.18 → 2513.18
Struct expert	59.3 ± 0.46	56.5 ± 0.33	65.11 → 1331.51	153.87 → 2659.87
All 4 experts (global fusion)	75.1 ± 0.15	72.1 ± 0.35	95.37 → 1361.77	268.18 → 2774.18
All 4 experts (sample-adaptive fusion)	76.1 ± 0.06	72.6 ± 0.17	95.37 → 1361.77	268.18 → 2774.18
AMPLE-Efficient (ESM2 15% / Struct 0%)	74.8 ± 0.06	71.6 ± 0.17	7.83 → 7.83	5.76 → 5.76
⌞ w/o Adaptive Fusion	74.4 ± 0.06	70.9 ± 0.07	7.83 → 7.83	5.76 → 5.76
AMPLE-Best (ESM2 60% / Struct 10%)	76.1 ± 0.19	72.3 ± 0.24	23.89 → 150.53	44.03 → 294.63
⌞ w/o Adaptive Fusion	75.6 ± 0.16	71.7 ± 0.11	23.89 → 150.53	44.03 → 294.63

fected because it avoids structural acquisition entirely, while AMPLE-Best pays the high structural cost for only a small fraction of samples. Under uniform end-to-end accounting, both operating points remain substantially less expensive than configurations that require structural information for every peptide, including the ESCAPE baseline and full four-expert evaluation. The absolute latency scale therefore changes markedly, but the relative advantage of selective acquisition remains.

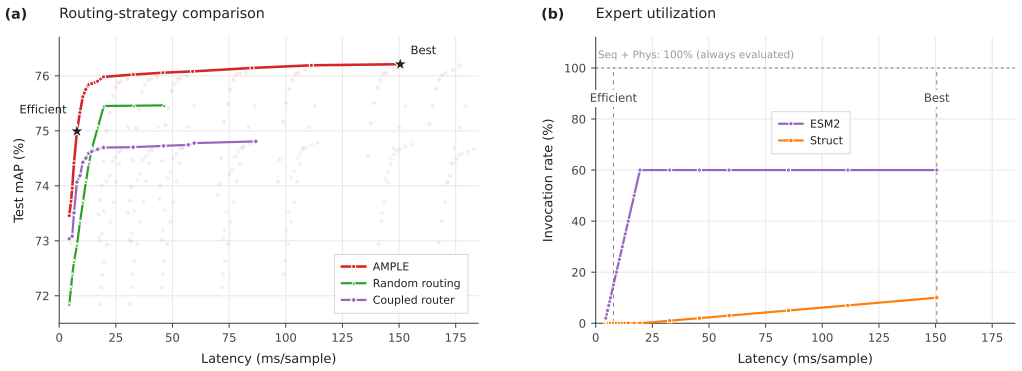

Figure 4: End-to-end adaptive acquisition including ESMFold structure prediction. (a) Performance–cost frontiers of AMPLE, random routing, and the coupled-router baseline. (b) ESM2 and Struct invocation rates along the AMPLE frontier; Seq and Phys are always evaluated.

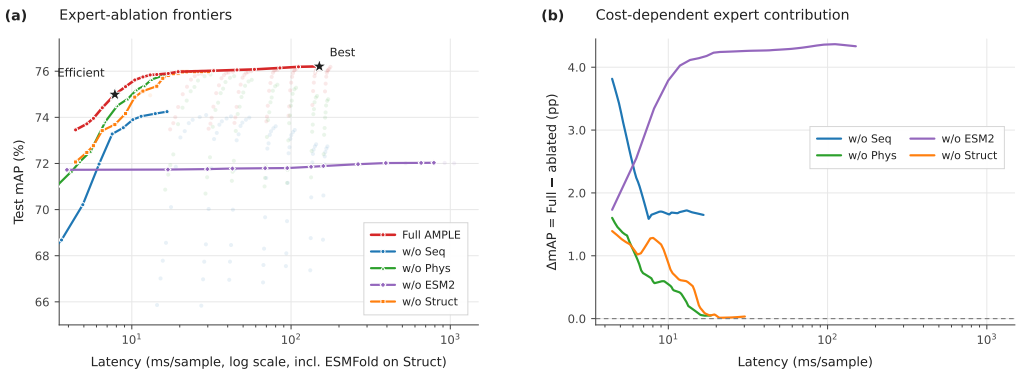

Figure 5: Expert ablation under end-to-end structural acquisition cost. (a) Performance–cost frontiers of AMPLE and leave-one-expert-out variants. (b) Cost-dependent expert contributions over overlapping cost ranges.