MACHINE LEARNING-DRIVEN DISCOVERY AND DESIGN OF ANTIMICROBIAL PEPTIDES: A COMPREHENSIVE REVIEW OF COMPUTATIONAL FRAMEWORKS, BENCHMARKS, AND MECHANISTIC INSIGHTS

[1*]Arokia Ahino Jessie A, [2]Raja M, [3]Jisha Hannie Selson, [4]Alex Yagoo
[1&3]Research Scholar, [2&4]Assisstant Professor

[1,2,3]*Department of Advanced Zoology and Biotechnology, Loyola College (Autonomous), Chennai, Tamil Nadu, India*

[4]*Department of Zoology, St. Xavier's College (Autonomous), Palayamkottai, Tirunelveli, Tamil Nadu 627002, India*

[*]*Corresponding author: arokia.jessie@gmail.com*

Abstract: The escalating global burden of antimicrobial resistance (AMR) has renewed interest in antimicrobial peptides (AMPs) as a structurally diverse, evolutionarily validated class of anti-infective molecules capable of circumventing many conventional resistance mechanisms. However, conventional AMP discovery remains constrained by the vast combinatorial peptide sequence space, cost and throughput limits of wet-laboratory screening, and difficulty in simultaneously optimizing potency, selectivity, and host safety. Machine learning (ML) has emerged as the principal strategy for closing this gap, spanning classical statistical classifiers, deep sequence models, structure- and graph-aware architectures, generative frameworks for de novo peptide design, and reinforcement learning-guided sequence optimization. This review synthesizes the current computational landscape of ML-driven AMP research across five dimensions: representation and embedding strategies that convert peptide sequences into model-ready features; supervised architectures for activity and toxicity prediction; generative and reinforcement-learning frameworks for de novo and multi-objective peptide design; comparative benchmarking of algorithm families, public databases (APD3, DRAMP, DBAASP), and reported performance metrics; and the mechanistic interpretability layer supplied by explainable artificial intelligence (XAI) and hybrid ML-molecular dynamics (MD) workflows. We further examine persistent bottlenecks as data scarcity and class imbalance, the experimental validation gap, host toxicity and proteolytic instability, and the translation of in silico leads into in vivo therapeutics and outline actionable strategies, including active-learning-guided design-build-test-learn cycles, physics-informed hybrid modelling, and mechanism-labelled generative frameworks, that are positioned to move the field from correlational, black-box prediction toward mechanism-aware, sample-efficient, and experimentally closed-loop AMP design.

Index Terms **- Antimicrobial Peptides, Antimicrobial Resistance, Machine Learning, Deep Learning, Generative Models, xplainable AI, Molecular Dynamic, Peptide Engineering**

1. INTRODUCTION

1.1 The Antimicrobial Resistance (AMR) Crisis and the Potential of AMPs

Antimicrobial resistance is now recognized as one of the foremost threats to global public health: bacterial antimicrobial resistance (AMR) was associated with an estimated 4.95 million deaths in 2019 alone, of which 1.27 million were directly attributable to resistant infection (Antimicrobial Resistance Collaborators, 2022). Forecasts extending this analysis to 2050 project continued increases in the associated global mortality and disease burden in the absence of effective intervention (GBD 2021 Antimicrobial Resistance Collaborators, 2024). The conventional antibiotic pipeline has not kept pace with this threat. The majority of antibiotics approved in recent decades are chemical derivatives of existing drug classes rather than agents with genuinely novel mechanisms of action, leaving a shrinking arsenal against multidrug-resistant gram-negative pathogens.

Antimicrobial peptides (AMPs) is a short, typically cationic, and amphipathic peptides produced across all domains of life as part of innate immune defense. They have attracted sustained interest as complementary or alternative therapeutic modalities. Unlike most conventional antibiotics, which act on a single, highly specific intracellular target, many AMPs primarily act through physical disruption of the microbial membrane. A mechanism that is comparatively difficult for a pathogen to circumvent through single-point mutations and confers activity against bacteria, fungi, and, in some cases, enveloped viruses and parasites, such as Plasmodium species. This broad-spectrum, multi-target mode of action, together with the rapid killing kinetics reported for many AMPs, underlies their designation as a structurally and mechanistically distinct reservoir of anti-infective chemical matter at a time when the conventional discovery pipeline has stalled.

1.2 Limitations of Conventional AMP Discovery Pipelines

Despite this promise, the translation of AMP biology into approved therapeutics has been slow, and the bottleneck lies substantially in the discovery pipeline itself. The natural and synthetic AMP sequence space is combinatorially enormous; even short peptides of 10-20 residues admit more possible sequences than can be exhaustively synthesized and assayed. Therefore, conventional discovery strategies, which rely on natural product isolation, homology-guided design, or iterative rational mutagenesis of a small number of scaffold peptides, sample only a small fraction of this space.

The wet-laboratory characterization compounds the problem. Minimum inhibitory concentration (MIC) determination, hemolysis, cytotoxicity, and proteolytic stability testing are comparatively low-throughput, reagent, and time-intensive procedures, so most discovery campaigns can only afford to synthesize and test a small top-ranked subset of candidate sequences. Because antimicrobial potency and host-cell toxicity are governed by overlapping biophysical features like cationic charge and hydrophobic amphipathicity drive both bacterial membrane disruption and, at sufficient magnitude, human erythrocyte lysis. The conventional pipelines have also struggled to optimize potency and safety simultaneously rather than sequentially, often discovering toxicity liabilities only after a promising antimicrobial candidate has already been advanced through several rounds of synthesis. The net effect is a discovery process that is slow, expensive, and structurally biased toward incremental variation around already known peptide scaffolds rather than genuine exploration of unexplored, potentially more effective regions of sequence space.

1.3 Scope and Objectives of This Review

This review focuses specifically on the computational, machine-learning-driven layer of the modern AMP discovery pipeline, examining how ML methods are reshaping each stage from sequence representation through candidate generation, activity/toxicity prediction, mechanistic interpretation, and closed-loop experimental validation. The objectives of this review are fourfold: First, to provide a structured overview of the ML frameworks currently applied to peptide engineering, organized by representation stra tegy, supervised prediction task, generative design paradigm, and reinforcement-learning-based optimization. Second, to assemble a comparative benchmark of representative algorithms, training data sources, and reported performance metrics across classical, deep-sequence, structure-aware, and generative model families, alongside the major public AMP databases that underpin them. Third, to examine the mechanistic and explainability layer, the feature-importance and attention-based interpretability methods, structure-activity relationship (SAR) mapping, and hybrid ML-molecular dynamics workflows, that connects statistical prediction back to underlying membrane-interaction biophysics. Fourth, to critically assess the field's outstanding challenges, including data bias, the experimental validation bottleneck, host toxicity and stability prediction, and the translation of computationally prioritized candidates into in vivo therapeutics, while highlighting concrete, literature-supported strategies for addressing each.

2. OVERVIEW OF MACHINE LEARNING FRAMEWORKS IN PEPTIDE ENGINEERING

2.1 Sequence-Based Representations and Embedding Strategies

The predictive and generative power of any AMP model is fundamentally limited by how the peptide is numerically represented before it reaches the algorithm. Representation strategies for AMP sequences can be organized into three tiers of increasing context awareness. The first tier comprises composition and physico-chemistry based descriptors: Amino Acid Composition (AAC) reduces a peptide to a fixed 20-dimensional vector of residue frequencies, efficiently capturing global compositional bias such as cationic- and hydrophobic-residue enrichment while discarding sequence order; dipeptide and higher-order k-mer/n-gram counts partially restore local order information; and Pseudo Amino Acid Composition (PseAAC; Chou, 2001) augments AAC with sequence-order correlation factors derived from physicochemical-property autocorrelation. In addition, explicit biophysical descriptors such as net charge, hydrophobic moment, the Boman index, isoelectric point, instability index, and grand average of hydropathy (GRAVY) ar e computed directly from the sequence and used either as direct classifier inputs or as conditioning vectors for generative models.

The second tier addresses positional and evolutionary information. One-hot encoding represents each residue as a binary indicator across the twenty canonical amino acids at every sequence position, producing an L×20 matrix consumed directly by convolution al and recurrent architectures. Transformer-based models add sinusoidal or learned positional encodings to each residue embedding so that self-attention, which is otherwise permutation-invariant, can distinguish residue identity from residue position. Position-specific scoring matrices (PSSMs), generated by aligning a query sequence against a reference database, add an evolutionary dimension by scoring the likelihood of every possible residue substitution at each position based on homologous sequences; hybrid descriptors such as PsePSSM-DCT apply pseudo-averaging and a discrete cosine transform to compress this information into a fixed-length, noise-suppressed representation.

The third and most context-rich tier draws on embeddings from large pretrained protein language models such as ESM-2 and ProtTrans/ProtT5, trained via masked-language-modelling objectives on tens to hundreds of millions of natural protein sequences. These models generate dense per-residue or pooled per-sequence vectors that encode long-range dependencies and implicit evolutionary and structural regularities without ever observing an explicit three-dimensional coordinate. In AMP-specific applications, pooled embeddings feed gradient-boosted trees or shallow classifiers, token-wise embeddings feed CNN or attention modules for finer-grained motif detection, and embeddings from multiple language models are concatenated with physicochemical descriptors for multimodal prediction. Some frameworks further continue pretraining on peptide-length sequences, better matching the short-sequence regime characteristic of AMPs. Table 1 summarizes these representation strategies alongside their typical model pairings (drawing on ESM-2, Lin et al., 2023; ProtTrans, Elnaggar et al., 2022; and the k-mer-based AmpGram framework, Burdukiewicz et al., 2020).

Table 1. Feature extraction and representation strategies for AMP sequences

Feature Type	Representation	What It Captures	Typical Model Pairing
Amino Acid Composition (AAC)	20-dimensional vector of residue frequencies	Global compositional bias (e.g., cationic/hydrophobic enrichment) No positional information	Classical ML (SVM, RF, XGBoost) as baseline input
Dipeptide / k-mer composition	400-dimensional (dipeptide) or higher-order n-gram frequency vectors	Local, adjacent-residue sequence-order patterns missed by AAC alone	RF/SVM ensembles n-gram random forest models (e.g., AmpGram)
Pseudo Amino Acid Composition (PseAAC)	AAC augmented with physicochemical-property autocorrelation factors	A fixed-length compromise between full sequence order and pure composition	Classical ML and shallow neural classifiers
Physicochemical descriptors	Net charge, hydrophobic moment, Boman index, isoelectric point, instability index, GRAVY	Biophysical properties governing membrane interaction (cationicity, amphipathicity, hydrophobicity)	Direct classical-ML input Conditioning vectors for generative models
One-hot / positional	Binary L×20 residue-identity	Explicit residue-position	CNN, RNN/LSTM,

encoding	matrix, optionally with sinusoidal/learned positional vectors	information enabling learning of position-specific motifs	Transformer input layers
Position-Specific Scoring Matrix (PSSM)	L×20 matrix from multiple-sequence-alignment/PSI-BLAST profiles	Evolutionary conservation and residue substitutability at each position	Hybrid pipelines (e.g., PsePSSM-DCT) CNN/attention hybrids
Protein language model embeddings (ESM-2, ProtTrans/ProtT5)	Dense contextual vectors (roughly 320-5120 dimensions, per-residue or pooled)	Learned evolutionary and latent structural context beyond hand-crafted descriptors Long-range dependencies	Fine-tuned classifiers/GBMs on pooled embeddings Graph and hybrid-attention predictors

2.2 Supervised Learning for Activity and Toxicity Prediction

The majority of ML applications in AMP research remain supervised classification or regression tasks: given a peptide representation, predict whether the sequence is antimicrobial (and against which target organisms) and separately, whether it is likely to be hemolytic or cytotoxic to host cells. Classical algorithms like Support Vector Machines, Random Forests, Gradient-Boosted Trees (e.g., XGBoost), and simpler baselines such as k-nearest neighbors or logistic regression operate on physicochemical or compositional feature vectors and remain competitive on small-to-medium curated datasets. Offering the additional advantage of native interpretability through feature-importance rankings, though typically at the cost of requiring classification reframing rather than native MIC regression.

Deep sequence models like convolutional neural networks (CNNs, e.g., AMPScanner; Veltri et al., 2018), recurrent architectures (RNN/LSTM), and Transformer or protein-language-model-based architectures (e.g., UniAMP, AMP-Designer; Wang et al., 2025). Instead learn representations directly from one-hot-encoded or embedded sequences, generally achieving the highest reported classification accuracies. Particularly when built on large pretrained embeddings, at the cost of reduced interpretability and a larger appetite for labelled training data. Toxicity prediction specifically remains comparatively less mature than activity prediction across all model families, since hemolytic and cytotoxic outcomes correlate only weakly with sequence descriptors alone and wet-lab toxicity confirmation is reported for only a small fraction of computationally predicted hits in most published studies. A limitation that recurs throughout the benchmarking discussion in Section 3.

2.3 Generative Frameworks for De Novo Design

Where supervised models classify or rank existing or externally proposed sequences, generative frameworks aim to synthesize novel peptide sequences directly from a learned distribution over the AMP sequence space. Variational autoencoders (VAEs), including HydrAMP (Szymczak et al., 2023), ICVAE, and conditional β-VAE variants, learn a continuous latent-space encoding of peptide sequences and generate new candidates by decoding samples from this latent space. Because there is no single accuracy or AUC figure for a generative task, success is instead judged by generation novelty and quality together with downstream classifier scores. In the case of HydrAMP, by direct in vitro and erythrocyte-toxicity validation against Gram-positive, Gram-negative, and resistant bacterial strains. Generative adversarial networks (GANs), including Feedback GAN (FBGAN; Gupta & Zou, 2019) and multi-property-optimizing variants, instead pit a generator against a discriminator trained to distinguish real from synthetic sequences. Downstream activity/toxicity classifiers are used to filter the generator's output, and a persistent risk in this family is mode collapse, and generated candidates are rarely tested experimentally at the same scale as they are generated.

A central technical challenge for all generative frameworks is that antimicrobial potency and host toxicity are not independent objectives, since the cationic charge and amphipathicity that drive bacterial membrane disruption also promote human erythrocyte lysis at sufficient magnitude. Explicit conditioning (conditional VAE/GAN) addresses this by concatenating a target property vector. A desired activity class or physicochemical/toxicity target with the latent code or noise vector before decoding or generation. So that sampling with a high potency, low toxicity target vector biases the output toward that region of sequence space. Multimodal, multitask conditioning frameworks extend this approach to cover predicted structure, species-specific activity, and mechanism of action simultaneously. Classifier-in-the-loop approaches (FBGAN and the more integrated classifier-driven GAN, cdGAN) instead use a periodically or continuously applied classifier to filter or directly shape the generator's training signal. While bidirectional conditional GAN frameworks such as AMPGAN v2

(Van Oort et al., 2021) and property-regularised VAEs disentangle the latent space so that specific dimensions correlate with specific continuous properties (e.g., a predicted-MIC axis and a separate hemolysis-score axis), enabling controlled latent-space traversal toward a target value without retraining the whole model. Regardless of the generation-time constraint mechanism, nearly all pipelines retain a final, independent post-generation screening stage in which downstream classifiers apply hard potency/toxicity threshold cut-offs before any candidate is prioritized for synthesis.

Diffusion models are a more recent addition to the generative toolkit for peptide and protein design. They learn to reverse a gradual noising process applied to sequence, structural, or joint sequence-structure representations and offer more stable training dynamics than VAEs and GANs. In addition, structure-conditioned variants can generate sequences consistent with a specified backbone geometry. Diffusion-based approaches remain comparatively less established for short antimicrobial peptides than for larger proteins, and their integration with the potency-toxicity conditioning strategies described above (explicit conditioning, classifier guidance, and multi-objective latent control) is an active extension of the constrained-generation paradigm rather than a distinct one.

2.4 Reinforcement Learning and Multi-Objective Sequence Optimisation

Reinforcement learning (RL) approaches frame AMP design as a sequential decision process: an agent proposes or edits a peptide sequence and receives a reward signal, typically derived from predicted binding affinity or activity against a target rather than from a fixed, static training corpus, which it uses to iteratively improve its policy. Monte Carlo Tree Search (MCTS)-based frameworks such as HighPlay and CYC_BUILDER exemplify this approach, reporting outcomes in terms of predicted binding affinities, rather than the accuracy/AUC/F1 metrics used to benchmark classification models. For example, micromolar-level affinities against target proteins. Because the reward function is only a proxy for true biological activity, its design is the critical determinant of RL performance, and toxicity is not automatically optimized unless it is explicitly incorporated into a multi-objective reward alongside potency. Direct in vitro validation of RL-designed AMPs remains minimal in the surveyed literature relative to the volume of computationally proposed candidates. Therefore, RL is best understood as complementary to the generative frameworks of Section 2.3, where VAEs and GANs learn to sample from a distribution resembling known AMPs, and RL explicitly optimizes toward a target property even when this pushes candidates outside the distribution of previously observed sequences, at the cost of requiring a reliable, computationally cheap reward proxy for what is ultimately a wet-lab-measured biological outcome.

3. COMPREHENSIVE BENCHMARKING OF ML MODELS FOR AMP DISCOVERY

A direct head-to-head comparison across the ML literature on AMP discovery is complicated by the fact that individual studies use varied datasets, sequence-length ranges, target organisms, and evaluation protocols (Xu et al., 2021); the performance ranges compiled here therefore reflect typical values reported across the cited model families rather than a single standardized benchmark. Table 2 consolidates this comparison across five methodological paradigms: classical (non-deep) machine learning, sequence-based deep learning, graph and structure aware models, generative models, and reinforcement learning. Spanning representative algorithms, input data type, typical training data sources and scale, reported performance metrics, and the principal limitations bearing on biological validity, including in vitro validation rate and toxicity-prediction accuracy where reported.

Table 2. Comprehensive benchmarking of ML model paradigms for AMP discovery

Model Family	Representative Algorithms	Input Data Type	Typical Training Data	Reported Performance	Key Limitations
Classical ML (non-deep)	SVM Random Forest XGBoost/Gradient-Boosted Trees k-NN / Logistic Regression	Physicochemical descriptors, amino-acid composition, engineered tabular features	APD3, DRAMP 2.0, DBAASP v3 Hundreds to ~10K+ curated peptides	Accuracy ~75-95% AUC-ROC ~0.75-0.96 depending on algorithm XGBoost typically best-in-class among classical	Low interpretability (SVM / RBF) Classification-only outputs (no native MIC regression)

				methods	Moderate toxicity-prediction accuracy (~70-85%) Prone to overfitting on small / imbalanced sets
Sequence-based deep learning	CNN (e.g., AMPScanner, Veltri et al., 2018) RNN/LSTM Transformer / protein-LM-based (UniAMP, AMP-Designer, Wang et al., 2025)	One-hot or embedding-encoded amino-acid sequences	Same large curated sources, often augmented (1K-10K+) Transformer variants pretrain on millions of general protein sequences	Accuracy 85-95%+ AUC-ROC 0.88-0.96 AMP-Designer reports ~94.4% design success against Gram-negative targets	Black-box, low interpretability Large data / compute requirements Wet-lab validation typically limited to a small candidate subset per study
Graph & structure-aware models	Graph Neural Networks (GNN) Graph Attention Networks (e.g., sAMPpred-GAT, Yan et al., 2023) 3D-CNN architectures	Peptide/molecular graphs or voxelised 3D structures (experimental or AlphaFold / trRosetta-predicted)	DBAASP v3 structural annotations PDB-derived structures Medium scale (hundreds-thousands)	AUC-ROC typically ≥0.90 sAMPpred-GAT reported to exceed nine other state-of-the-art predictors across independent test sets	Accuracy depends on structure-prediction quality (less reliable for short / flexible peptides) Computationally intensive Validation largely in-silico / benchmark-based
Generative models (VAE/GAN)	VAE / HydrAMP (Szymczak et al., 2023), ICVAE, conditional β-VAE GAN / FBGAN (Gupta & Zou, 2019), Multi-CGAN (Yu et al., 2024)	Sequence-based latent encodings Sequence- or graph-based representations for GANs	APD3, DRAMP, DBAASP-derived curated active / inactive sets	No single accuracy / AUC metric Judged by generation novelty / quality and downstream classifier or wet-lab hit rate HydrAMP experimentally validated against multiple strain types	No standard benchmark metric GANs risk mode collapse Biological validity bounded by the downstream classifier used to filter candidates
Reinforcement learning	MCTS-based frameworks (HighPlay,	Sequence plus structure-prediction	No fixed benchmark dataset	Reported via predicted binding	Reward-function design is only a

	CYC_BUILDER)	feedback loop	Reward derived from structure-prediction models	affinity (e.g., micromolar-level) No standard accuracy / AUC / F1	proxy for true activity Minimal direct in vitro validation Toxicity not optimised unless built into a multi-objective reward

3.1 Sequence-to-Activity Prediction Models

Among classical algorithms, Support Vector Machines operating on physicochemical descriptors and amino-acid composition achieve accuracies of approximately 80-90% and AUC-ROC of 0.85-0.92 on small-to-medium curated sets, although performance is sensitive to kernel choice and interpretability is limited with RBF kernels. Random Forests achieve somewhat higher accuracy (85-92%, AUC-ROC 0.88-0.94) while retaining strong, interpretable feature-importance rankings, although performance plateaus on very large or high-dimensional data and toxicity-prediction accuracy remain moderate (~70-85%). Gradient-boosted tree methods such as XGBoost achieve the highest classical-ML performance (accuracy 90-95%, AUC-ROC 0.90-0.96) on medium-to-large curated sets but are prone to overfitting on small, imbalanced AMP data without careful tuning, and MIC/toxicity is not directly modelled unless explicitly reframed as a regression task. Simpler k-NN and logistic-regression baselines (accuracy 75-85%) remain useful as comparators but are weak on high-dimensional, non-linear relationships.

Sequence-based deep learning models extend this performance ceiling further. CNN architectures such as AMPScanner (Veltri et al., 2018), trained on one-hot or embedding-encoded sequences, report accuracies of 88-95% and AUC-ROC of 0.90-0.96. Although as black-box classification-only models their outputs are not directly correlated with quantitative MIC values, and wet-lab validation is typically limited to a small candidate subset per study. RNN/LSTM architectures achieve broadly comparable accuracy (85-93%) with a particular strength in capturing long-range positional dependencies, at a higher computational cost for long sequences. Transformer and protein-language-model based architectures: including UniAMP and AMP-Designer (Wang et al., 2025) report the highest accuracies among sequence-only models (>90%), with AMP-Designer specifically reporting a 94.4% design success rate against Gram-negative targets. Notably, synthesis and testing of 18 AMPs within an 11-day cycle, illustrating high per-candidate success alongside limited absolute experimental throughput. These models carry the highest data and compute requirements and the lowest native interpretability of the sequence-based family, motivating the explainability methods discussed in Section 4.

3.2 Structure and Graph Based Predictors

Structure-aware models incorporate three-dimensional or graph-based peptide representations, typically derived from experimental structures or AlphaFold/trRosetta predictions, rather than sequence alone. General Graph Neural Networks (GNNs) operating on peptide/molecular graph nodes as residues or atoms, edges as bonds or spatial contacts, achieve AUC-ROC typically at or above 0.90 in benchmark comparisons, although accuracy is bounded by structure-prediction quality, which is less reliable for short or conformationally flexible peptides. Graph Attention Networks that combine predicted structure, sequence, and evolutionary information into an attention-weighted graph, exemplified by sAMPpred-GAT (Yan et al., 2023). It has been reported to exceed nine other state-of-the-art AMP predictors in AUC when evaluated across eight independent test datasets compiled from APD3, DRAMP, and DBAASP. 3D-CNN architectures operating on voxelized conformational data report good accuracy particularly for toxicity and conformational-stability prediction, at the cost of the highest data/compute burden among structural methods and a continued dependence on reliable 3D structures that remain challenging to obtain for short, flexible AMPs. Across this family, validation remains predominantly in-silico or benchmark-based rather than confirmed at large experimental scale.

3.3 Generative and De Novo Design Platforms

Generative platforms are evaluated on fundamentally different criteria from classifiers, since there is no single ground-truth label for a "correctly generated" novel sequence. Conditional VAE frameworks, such as HydrAMP, incorporate dual pretrained classifiers during training to monitor generation quality and are, notably, among the few generative tools with explicit in vitro and erythrocyte-toxicity

validation, having been experimentally tested against Gram-positive, Gram-negative, and resistant bacterial strains, a validation standard not typical of this paradigm. GAN-based platforms, including Feedback GAN (which iteratively re-labels or replaces training data above an activity/toxicity threshold) and classifier-driven GAN variants that integrate classifier gradients directly into the generator's loss. They are evaluated via discriminator authenticity scores and downstream classifier validation rather than direct accuracy / AUC, and remain susceptible to mode collapse. As with other generative approaches, generated candidates are rarely tested experimentally at scale, so biological validity is only as reliable as the downstream classifier used to filter the candidate pool. Multi-property and bidirectional conditional GAN variants (Multi-CGAN, Yu et al., 2024; AMPGAN v2, Van Oort et al., 2021) extend this framework to simultaneous multi-objective design. For example, target species specificity together with low toxicity in a single generation pass. Diffusion-based generative platforms, while less established specifically for short AMPs than for general protein design, represent an emerging addition to this landscape and are a natural candidate for future integration with the conditioning and classifier-guidance strategies already established for VAEs and GANs (Section 2.3).

3.4 Comparative Analysis of Datasets (APD3, DRAMP, DBAASP) and Benchmark Metrics

Nearly all ML models surveyed in this review draw their training and evaluation data predominantly from three public reposito ries like APD3, DRAMP, and DBAASP. They are supplemented, for structure-aware models, by AlphaFold or trRosetta predicted structures. Table 3 compares these resources by scope, approximate scale, and their typical role in ML pipelines. A recurring benchmarking limitation across the field is that these databases differ in curation criteria, sequence length distributions, and the balance of confirmed negative (inactive) examples. Therefore, reported performance metrics (accuracy, AUC-ROC, and F1) are not always directly comparable across studies trained on different database subsets or splits. This heterogeneity, combined with the general scarcity of standardized, community-agreed benchmark splits (in contrast to fields such as general protein structure prediction), is one of the field's outstanding methodological challenges and is revisited in Section 5.1.

Table 3. Comparative overview of the principal public AMP databases used for ML training and benchmarking

Database	Content / Scope	Approximate Scale	Distinguishing Features	Primary Use in ML Pipelines
APD3 (Antimicrobial Peptide Database; Wang et al., 2016)	Curated, experimentally characterised natural AMPs across bacteria, fungi, viruses, and higher organisms	~3,000-5,000+ entries	One of the earliest and most widely cited curated AMP repositories Strong sequence-level annotation	Primary training / benchmark source for classical ML and sequence-based deep learning classifiers
DRAMP (Data Repository of Antimicrobial Peptides), v2.0/3.0 (Kang et al., 2019)	Aggregates natural, synthetic, and patent-derived AMPs with activity, structure, and clinical-stage annotation where available	Tens of thousands of entries (general + curated subsets)	Broader scope than APD3, including synthetic / patent sequences and some clinical-development status	Large-scale pretraining / fine-tuning source Supplemented by structure-prediction pipelines for structure-aware models
DBAASP (Database of Antimicrobial Activity and Structure of Peptides), v3 (Pirtskhalava et al., 2021)	Peptides annotated with quantitative activity (MIC) against specific target strains and, for many entries, structural information	Tens of thousands of activity measurements	Fine-grained, strain-specific quantitative activity data Structural annotations used for graph / 3D-aware models	Preferred source for structure-aware (GNN / GAT / 3D-CNN) models and for regression-oriented MIC prediction

4. MECHANISTIC INSIGHTS AND EXPLAINABLE AI (XAI)

4.1 Deciphering ML Models via Feature Importance and Attention Maps

Deep learning models, particularly CNNs, RNN/LSTMs, and Transformers, routinely outperform classical machine learning on AMP classification and generation tasks; however, their internal decision process is not directly inspectable from model weights alone, which is a genuine barrier to biological insight and regulatory trust rather than a mere inconvenience: without knowing which sequence or structural features actually drove a prediction, it is difficult to distinguish a model that has learned true membrane-interaction biology from one that

has learned a dataset-specific artifact. SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) addresses this by treating each input feature as a "player" contributing to the model's prediction and computing its Shapley value, the average marginal contribution across all possible feature-subset orderings; applied to AMP classifiers, SHAP produces both per-peptide explanations (which descriptors pushed a given sequence toward an active or inactive call) and dataset-level summaries of overall feature influence, and has been used specifically to identify which physicochemical and compositional features best explain predicted antibacterial and hemolytic activity.

LIME (Local Interpretable Model-Agnostic Explanations; Ribeiro et al., 2016) offers a complementary, more computationally efficient alternative, perturbing the input around a specific prediction and fitting a locally interpretable linear surrogate model to approximate the black-box model's local behavior; however, its local-linearity assumption can be a poor approximation for highly non-linear deep models, whereas SHAP's game-theoretic formulation is more computationally expensive but generally considered more consistent. Several AMP-prediction studies apply both methods side by side across multiple classifier types specifically so that feature-importance conclusions can be cross-checked rather than trusted from a single explanation method. Where SHAP and LIME are applied post-hoc to any model type, Transformer and other attention-augmented architectures offer explanations intrinsic to their computation: visualizing self-attention weight matrices as per-residue heatmaps can highlight which positions the model is effectively focusing on, and in a reported species-aware peptide language model pipeline, this analysis identified highly attended residues that were then deliberately mutated to increase amphipathicity and hydrophobic moment before synthesis and experimental testing, a genuinely actionable use of attention interpretation for rational, targeted sequence redesign. Attention weights are nonetheless not a guaranteed measure of causal feature importance and are best treated as a hypothesis-generating tool to be cross-validated with SHAP/LIME-style attribution or structural/biophysical follow-up. Table 4 summarizes these explainability and physics-informed methods alongside their reported AMP applications.

Table 4. Explainable AI and physics-informed methods applied to ML-driven AMP design

XAI / Physics-Informed Method	How It Works	What It Reveals in AMP Models	Reported Use in the AMP Literature
SHAP (Lundberg & Lee, 2017)	Game-theoretic, model-agnostic attribution: computes each feature's average marginal contribution (Shapley value) across all possible feature orderings	Ranks physicochemical and compositional descriptors by contribution to the antibacterial / haemolytic classification decision, per-peptide or dataset-wide	Applied across classical-ML and hybrid AMP predictors to identify which properties best explain predicted antibacterial and haemolytic activity
LIME (Ribeiro et al., 2016)	Perturbs the input around one prediction and fits a simple, locally interpretable linear surrogate model	Provides a fast, per-prediction feature-importance plot Local-linearity assumption can be a poor fit for highly non-linear deep models	Used alongside SHAP as a complementary local-explanation tool across multiple classifier types (Decision Tree, RF, MLP, GBT, Naive Bayes)
Attention-map analysis (Transformer / attention-based models)	Extracts self-attention weight matrices showing how strongly each residue attends to every other position	Highlights residues the model weights most heavily In some pipelines aligns with the hydrophobic face of an amphipathic helix	Used in species-aware peptide language models to identify residues subsequently mutated to increase amphipathicity prior to synthesis
Hybrid ML + Molecular Dynamics (MD)	Couples a trained ML predictor with atomistic or coarse-grained MD simulation of the peptide in	Provides a mechanistic, physics-based check on ML-predicted candidates	Used as a post-hoc validation layer for ML-prioritised candidates and, increasingly, as a source of physics-

	a membrane-mimetic environment	(e.g., confirming membrane insertion and the qualitative disruption mechanism)	grounded labels for structure-aware models
Free Energy Perturbation (FEP)	Alchemical, all-atom MD technique computing the free-energy change of mutating one chemical/sequence state into another along a simulated pathway	Offers a rigorous, physically grounded potency estimate able to rank closely related peptide variants more reliably than data-driven scores alone	Established in peptide/protein-ligand binding workflows; increasingly proposed as an ML-guided refinement step via active-learning-accelerated sampling

4.2 Mapping Structure-Activity Relationships (SAR) and Biophysical Constraints

AMPs kill or inhibit microbes predominantly by disrupting the microbial membrane, and several distinct biophysical models describe how this disruption occurs, all beginning with an initial electrostatic attraction between the net-cationic peptide and the net-anionic phospholipid head groups characteristic of microbial (as opposed to more zwitterionic mammalian) membranes, but diverging in the geometry of the resulting interaction. In the barrel-stave model, peptides insert perpendicularly into the bilayer and self-assemble into a bundle resembling barrel staves, forming a transmembrane pore lined entirely by peptide, with hydrophilic faces lining the aqueous channel and hydrophobic faces contacting the lipid tails; this mechanism requires sufficient peptide length to span the bilay er and strong peptide-peptide self-association. The toroidal-pore model instead describes peptides inducing the lipid monolayers themselves to bend continuously through the membrane, so the pore lining comprises both peptide and lipid head groups rather than peptide alone. The carpet (detergent-like) model involves peptides accumulating parallel to the membrane surface until a critical local concentration triggers a detergent-like thinning and disintegration of the membrane into micelles, without ever forming a discrete stable pore. A fourth, less membrane-centric category covers peptides that form disordered surface aggregates or translocate across the membrane largely intact to bind intracellular targets, such as nucleic acids, ribosomes, or cell-wall synthesis enzymes, rather than relying on stable pore formation.

Incorporation of this mechanistic knowledge into ML pipelines is uneven across the four models. The clearest link exists for pore-forming, alpha-helical mechanisms (barrel-stave and, to a lesser extent, toroidal-pore): physicochemical descriptor sets explicitly compute hydrophobic moment and periodicity-matched charge/hydrophobicity alternation along a helical-wheel projection, encoding the facial amphipathicity these mechanisms require, and feature-selection studies on SVM/RF classifiers have independently confirmed that helical-periodicity descriptors rank among the most discriminative features for membrane-active peptides. Structure-aware architectures (3D-CNNs, GNNs) trained on experimental or predicted conformations can in principle learn this periodic amphipathic geometry directly from three-dimensional coordinates, representing the most promising route toward mechanism-aware prediction, although most current structure-aware AMP models are trained on isolated peptide conformations rather than membrane-embedded or MD-derived structures, capturing intrinsic amphipathic geometry without directly modelling the interaction trajectory with a lipid bilayer. The toroidal-pore and carpet mechanisms, which depend on lipid participation and concentration-dependent behavior, are captured only indirectly through charge-density and amphipathicity descriptors, since few mainstream classifiers explicitly predict which pore-forming mechanism a peptide will use, this requiring biophysical mechanism annotation (circular dichroism, MD simulation, NMR) that most AMP databases do not provide at scale. The aggregate/intracellular-target mechanism remains the least explicitly modelled, typically inferred by exclusion, an active peptide lacking the amphipathic signatures of the other three mechanisms or addressed through separate downstream molecular-docking modules rather than a dedicated sequence-based encoder. The most direct attempts to close this gap are multitask, mechanism-labelled generative frameworks that include antimicrobial mechanism as an explicit conditioning category, and the hybrid ML-molecular dynamics pipelines discussed next.

4.3 Hybrid Workflows: Coupling Machine Learning with Molecular Dynamics (MD)

XAI methods explain what a purely data-driven model has learned; however, they cannot supply physics the model was never trained on. Hybrid ML-MD workflows address this by coupling ML predictions with physics-based simulations: atomistic or coarse-grained MD simulations of a candidate peptide in a membrane-mimetic bilayer generate direct structural and energetic evidence, whether the peptide actually inserts into the membrane, how it perturbs lipid packing, and which qualitative mechanism (barrel-stave-like insertion, surface carpeting, or toroidal-pore-like bending) its trajectory resembles, that a sequence-only classifier cannot observe.

This evidence can be used either as a post-hoc validation filter on ML-prioritized candidates or, more ambitiously, as a source of physics-grounded labels for training the next generation of structure- and mechanism-aware models.

Free energy perturbation (FEP) is the most rigorous end of this spectrum, an alchemical all-atom MD technique that computes the free-

energy change of mutating one chemical/sequence state into another along a simulated pathway, yielding quantitative binding or membrane-insertion free-energy estimates that, for well-behaved systems, can approach experimental accuracy. Its principal limitation is computational cost, as a single relative free-energy calculation can require substantial simulation time, historically restricting its use to late-stage ranking of a handful of close variants rather than large-scale screening. Recent work integrating machine learning into FEP workflows, using active learning to select which alchemical calculations are most informative to run, and ML-based sampling and protocol optimization to accelerate convergence. It is beginning to close this gap, pointing toward a tiered pipeline in which fast ML classifiers perform initial large-scale triage, attention/SHAP analysis flags mechanistically plausible candidates, and FEP or full MD is reserved for final, physics-grounded confirmation of the most promising shortlist.

5. CRITICAL CHALLENGES AND FUTURE HORIZONS

5.1 Addressing Data Bias and Experimental Validation Bottlenecks

Two intertwined data-level limitations recur throughout the ML-AMP literature reviewed in Sections 2-4: data scarcity and class imbalance. Curated, activity-labelled AMP sequences number in the low thousands, even in the largest public databases, several orders of magnitude smaller than the training corpora typical of deep learning, and confirmed negative (inactive) examples to gether with balanced multi-label activity annotations (antibacterial versus antifungal versus antiviral) are especially scarce, biasing models toward majority classes. Self-supervised pretraining on large unlabelled protein corpora, followed by fine-tuning on the small labelled AMP set, decouples the bulk of representation learning from this scarce labelled data and has been shown to measurably improve downstream AMP classification accuracy for precisely this reason. Few-shot learning pipelines target the low-data regime directly — a coarse-to-fine, sequentially assembled machine-learning pipeline has been used to mine hundreds of billions of candidate hexapeptide-to-nonapeptide sequences for antimicrobial activity from only a small labelled training set (Huang et al., 2023) — and imbalance-aware loss functions such as asymmetric loss in multi-label Transformer classifiers improve balanced accuracy on imbalanced multi-functional AMP prediction without discarding minority-class data.

However, the experimental validation gap is the most persistent bottleneck: computational prediction, however accurate in cross-validation, still requires physical synthesis and assay before a candidate can be considered a genuine hit, a slow and costly one-candidate-at-a-time loop that limits most studies to wet-lab testing of only a small top-ranked subset of computationally generated candidates, meaning published in-silico accuracy figures are rarely stress-tested against real wet-lab hit rates at scale. The actionable response emerging across the recent literature is to formalize an active learning loop in place of a one-shot computation-to-bench hand-off: rather than ranking a fixed candidate pool once and testing the top N, the model iteratively nominates the next most informative peptide(s) to synthesize — typically those with the highest predictive uncertainty or expected information gain — receives the experimental result, retrains, and repeats. Early multi-round synthesise-and-test campaigns illustrate both the promise and current scale limits of this approach: an LLM-based foundation-model pipeline (AMP-Designer; Wang et al., 2025) designed 18 candidate AMPs within 11 days, achieved a 94.4% in vitro success rate, and progressed from design to two wet-lab-confirmed, murine-validated hits within a 48-day design-synthesise-test cycle, illustrating that even a high-throughput generative pipeline currently validates candidates on the order of tens per campaign rather than at the full computational scale of the sequence space explored. Table 5 consolidates these challenges alongside their proposed solutions and representative approaches.

Table 5. Key challenges in ML-driven AMP design paired with actionable solutions

Challenge	Why It Persists	Actionable Solution(s)	Representative Approach
Data scarcity	Curated, activity-labelled AMP sequences number in the low thousands even in the largest databases (APD3, DRAMP, DBAASP), several orders of	Data augmentation via sequence-level perturbation and template / docking-generated synthetic examples	Transfer learning from general-purpose protein language models improves AMP

	magnitude smaller than typical deep-learning training corpora, and wet-lab-confirmed negative examples are especially rare	Self-supervised pretraining on large unlabelled protein corpora (ESM-2, Lin et al., 2023 ProtTrans, Elnaggar et al., 2022) followed by fine-tuning on the small labelled AMP set	classification by decoupling representation learning from the scarce labelled data
Class imbalance	Positive (active) sequences are heavily over-represented relative to confirmed inactive sequences, and multi-label activity annotations (antibacterial vs antifungal vs antiviral) are imbalanced across classes	Few-shot learning pipelines combining pretraining with lightweight fine-tuning for target tasks (e.g., pathogen-specific) Imbalance-aware loss functions (asymmetric/focal loss) in multi-label classifiers	A pretrained, fine-tuned pipeline has screened hundreds of billions of candidate short peptides against a pathogen with extremely scarce prior data (Huang et al., 2023)
Experimental validation gap (synthesize-and-test bottleneck)	Even highly accurate in-silico predictors must still be chemically synthesised and assayed (MIC, haemolysis, cytotoxicity) one candidate at a time, so only a small fraction of computationally prioritised candidates are ever tested, and in-silico accuracy claims are rarely validated at scale against wet-lab hit rates	Active-learning workflows in which the model iteratively selects the next most informative peptide(s) to synthesise and retrains on the result Closed-loop design-build-test-learn pipelines coupling generative design, prediction, synthesis, and assay	AMP-Designer (Wang et al., 2025) reached two wet-lab-confirmed, murine-validated hits within a 48-day design-synthesise-test cycle at a 94.4% in vitro success rate among 18 designed candidates

5.2 Predicting Host Toxicity, Proteolytic Stability, and Resistance Evolution

Host toxicity prediction remains one of the least reliable outputs across all model families surveyed in this review: hemolytic and cytotoxic outcomes correlate only weakly with sequence descriptors alone, since the same cationic charge and amphipathicity that drive bacterial membrane selectivity. Also predispose a peptide toward mammalian membrane interaction at sufficient concentration, and independently reported toxicity-prediction accuracies (roughly 70–85% for classical descriptor-based models) lag behind antimicrobial-activity accuracies. Improving host-selectivity prediction will likely require moving beyond descriptor-based proxies toward structure and mechanism aware approaches discussed in Section 4, explicitly modelling the differential electrostatic and compositional profile of microbial versus mammalian membranes rather than relying on charge and hydrophobicity alone as generic toxicity correlates.

Proteolytic stability is a peptide's resistance to degradation by host or microbial proteases, and the longer-term risk of resistance evolution against AMP-based therapeutics are comparatively underexplored relative to potency and acute toxicity prediction within the ML frameworks reviewed here, reflecting both the scarcity of systematically labelled stability/resistance data and the multi-week, multi-passage experimental designs required to generate such labels (serum or protease-panel stability assays, and serial-passage resistance-selection experiments, respectively). Extending the mechanism-labelled and physics-informed frameworks discussed in Sections 2.3 and 4 for instance, incorporating protease-cleavage-site prediction as an additional conditioning or filtering criterion in generative pipelines, analogous to existing toxicity and activity filters. Pairing computational resistance-risk screening with serial-passage validation in a manner parallel to the active-learning loops proposed for the validation gap in Section 5.1. It represents a natural, currently underdeveloped extension of the design-build-test-learn paradigm to these longer-horizon safety and durability endpoints.

5.3 Translation from In Silico Predictions to In Vivo Therapeutics

Even computationally prioritized and in vitro-validated AMP candidates face a substantial gap before they can be considered viable therapeutics. In vivo pharmacokinetic behavior was rapid proteolytic clearance, potential immunogenicity, and, for many peptides. Poor oral bioavailability necessitating parenteral or topical administration. It is governed by physiological variables that neither sequence-based classifiers nor most current structure-aware or generative models explicitly optimize for. Since these models are trained predominantly on in vitro MIC, hemolysis, and cytotoxicity endpoints rather than on animal pharmacokinetic or efficacy data. Formulation strategies (e.g., PEGylation, cyclization, or delivery-vehicle encapsulation) that are routinely used to mitigate these liabilities for peptide therapeutics are, at present, largely applied downstream of the computational discovery pipeline described in this review rather than being integrated as an explicit design objective within it.

Closing this translational gap is likely to require extending the multi-objective conditioning strategies already established for the potency-toxicity trade-off (Section 2.3) to explicitly include pharmacokinetic and immunogenicity-relevant properties as additional generative constraints, alongside expanding the active-learning design-build-test-learn loops proposed in Section 5.1 to incorporate animal model efficacy and tolerability data as a later-stage feedback signal rather than treating in vitro validation as the pipeline's terminal endpoint. In this sense, the translation challenge is less a separate problem than a natural, currently underdeveloped extension of the same closed-loop, multi-objective, physics-and-mechanism-aware design philosophy that this review has traced across sequence representation, generative design, and explainable prediction.

6. CONCLUSION

Machine learning has fundamentally reshaped the antimicrobial peptide discovery pipeline, replacing what was once a slow, largely manual search through a vanishingly small fraction of peptide sequence space with a layered computational workflow spanning representation learning, supervised activity/toxicity prediction, generative and reinforcement-learning-guided de novo design, and increasingly, mechanistic interpretation through explainable AI and hybrid ML-molecular dynamics modelling. The comparative benchmarking assembled in this review indicates that no single model family dominates across all criteria: classical machine learning retains value for its interpretability and data efficiency; deep sequence models and protein-language-model embeddings currently achieve the highest raw predictive accuracy; structure- and graph-aware architectures offer the clearest path toward mechanism-grounded prediction; and generative and reinforcement-learning frameworks are the primary engines of genuine sequence-space exploration beyond known AMP scaffolds.

At the same time, the challenges catalogued in Section 5 emphasize the data scarcity and class imbalance, the persistent gap between in-silico prediction and wet-lab-confirmed hit rates, comparatively immature toxicity, proteolytic-stability, and resistance-evolution prediction, and the further distance between in vitro validation and genuine in vivo therapeutic viability make it clear that the field's next phase of progress will be defined less by incremental gains in classification accuracy on static benchmarks and more by the field's ability to close the loop: coupling generative design, mechanistic and physics-informed validation, and iterative, uncertainty-guided experimental feedback into a single, sample-efficient design-build-test-learn system. Active-learning campaigns, mechanism-labelled generative frameworks, and hybrid ML-MD/FEP workflows have already been reported in the literature, although they are still limited in terms of absolute experimental throughput. They represent concrete, early-stage instances of this integration and collectively chart a credible path from today's largely correlational, data-hungry, black-box predictors toward mechanism-aware, sample-efficient, and experimentally closed-loop antimicrobial peptide design systems capable of contributing meaningfully to the broader response to the global antimicrobial resistance crisis.

Author Contributions

The author conceived the review, performed the literature synthesis and wrote the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This review did not receive any specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

No new data were generated for this review. All the datasets discussed (APD3, DRAMP, and DBAASP) are publicly available at their respective repositories.

References

1. Antimicrobial Resistance Collaborators. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The Lancet*, *399*(10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

2. Burdukiewicz, M., Sidorczuk, K., Rafacz, D., Pietluch, F., Chilimoniuk, J., Rödiger, S., & Gagat, P. (2020). Proteomic screening for prediction and design of antimicrobial peptides with AmpGram. *International Journal of Molecular Sciences*, *21*(12), 4310. https://doi.org/10.3390/ijms21124310

3. Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Genetics*, *43*(3), 246–255. https://doi.org/10.1002/prot.1035

4. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M. , Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127.

5. GBD 2021 Antimicrobial Resistance Collaborators. (2024). Global burden of bacterial antimicrobial resistance 1990–2021: A systematic analysis with forecasts to 2050. *The Lancet*, *404*(10459), 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1

6. Gupta, A., & Zou, J. (2019). Feedback GAN for DNA optimizes protein functions. *Nature Machine Intelligence*, *1*(2), 105–111. https://doi.org/10.1038/s42256-019-0017-4

7. Huang, J., Xu, Y., Xue, Y., Huang, Y., Li, X., Chen, X., Xu, Y., Zhang, D., Zhang, P., Zhao, J., & Ji, J. (2023). Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. *Nature Biomedical Engineering*, *7*(6), 797–810. https://doi.org/10.1038/s41551-022-00991-2

8. Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., Li, H., Xu, H., Lao, X., & Zheng, H. (2019). DRAMP 2.0, an updated data repository of antimicrobial peptides. *Scientific Data*, *6*, 148. https://doi.org/10.1038/s41597-019-0154-y

9. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. dos S., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.

11. Pirtskhalava, M., Amstrong, A. A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D. E., & Tartakovsky, M. (2021). DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Research*, *49*(D1), D288–D297. https://doi.org/10.1093/nar/gkaa991

12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

13. Szymczak, P., Możejko, M., Grzegorzek, T., Jurczak, R., Bauer, M., Neubauer, D., Sikora, K., Michalski, M., Sroka, J., Setny, P., Kamysz, W., & Szczurek, E. (2023). Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nature Communications*, *14*, 1453. https://doi.org/10.1038/s41467-023-36994-z

14. Van Oort, C. M., Ferrell, J. B., Remington, J. M., Wshah, S., & Li, J. (2021). AMPGAN v2: Machine learning-guided design of antimicrobial peptides. *Journal of Chemical Information and Modeling*, *61*(5), 2198–2207. https://doi.org/10.1021/acs.jcim.0c01441

15. Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, *34*(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

16. Wang, G., Li, X., & Wang, Z. (2016). APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, *44*(D1), D1087–D1093. https://doi.org/10.1093/nar/gkv1278

17. Wang, J., Feng, J., Kang, Y., Pan, P., Ge, J., Wang, Y., Wang, M., Wu, Z., Zhang, X., Yu, J., Zhang, X., Wang, T., Wen, L., Y an, G., Deng, Y., Shi, H., Hsieh, C.-Y., Jiang, Z., & Hou, T. (2025). Discovery of antimicrobial peptides with notable antibacterial potency by an LLM-based foundation model. *Science Advances*, *11*(10), eads8932. https://doi.org/10.1126/sciadv.ads8932

18. Xu, J., Li, F., Leier, A., Xiang, D., Shen, H.-H., Marquez Lago, T. T., Li, J., Yu, D.-J., & Song, J. (2021). Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, *22*(5), bbab083. https://doi.org/10.1093/bib/bbab083

19. Yan, K., Lv, H., Guo, Y., Peng, W., & Liu, B. (2023). sAMPpred-GAT: Prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics*, *39*(1), btac715. https://doi.org/10.1093/bioinformatics/btac715

20. Yu, H., Wang, R., Qiao, J., & Wei, L. (2024). Multi-CGAN: Deep generative model-based multiproperty antimicrobial peptide design. *Journal of Chemical Information and Modeling*, *64*(1), 316–326. https://doi.org/10.1021/acs.jcim.3c01881

21. Zervou, M. A., Doutsi, E., Pantazis, Y., & Tsakalides, P. (2025). Classifier-driven generative adversarial networks for enhanced antimicrobial peptide design. *Briefings in Bioinformatics*, *26*(5), bbaf500. https://doi.org/10.1093/bib/bbaf500

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.