Hackathon Project: Anti vs non-Antimicrobial Peptide Classification Using Computational Models

Adan mhameed, Areen Mansour, Yousef AbuGosh, Nosiba Othman

1. Background

Antimicrobial resistance is a growing global health challenge, threatening the effectiveness of traditional antibiotics in combating bacterial infections[1]. In response, researchers have turned their attention to antimicrobial peptides (AMPs), short protein sequences with inherent antimicrobial properties. AMPs are considered promising alternatives to conventional antibiotics due to their ability to target a broad spectrum of pathogens while minimizing resistance developmentl.

Identifying and classifying AMPs is a critical step in harnessing their potential. However, traditional experimental methods for AMP discovery are labor-intensive, costly, and time-consuming. Advances in computational techniques, particularly machine learning, now offer innovative solutions to overcome these limitations by enabling rapid and accurate analysis of peptide sequences[2]. This research utilizes machine learning models to classify peptide sequences as antimicrobial or non-antimicrobial, addressing the critical need for efficient AMP discovery. The peptide sequences used in this study are sourced from publicly available databases, this study aims to contribute to the development of novel AMPs and combat the growing threat of antimicrobial resistance Additionally, the study proposes a model that reduces the number of parameters typically required for AMP classification. This simplified model offers a more efficient and computationally less demanding approach, making it particularly valuable for applications where resource limitations or faster predictions are crucial[3].

2. Introduction

The classification of antimicrobial peptides (AMPs) has gained significant attention as a viable solution to address the global crisis of antibiotic resistance. AMPs are naturally occurring molecules with the potential to disrupt bacterial membranes and inhibit microbial growth. Despite their promising properties, identifying AMPs remains a complex task, requiring advanced computational tools to analyze peptide sequences efficiently.

This study aims to optimize AMP identification using state-of-the-art machine learning models, specifically transformer-based architectures like ESM-2. By extracting meaningful features from peptide sequences and applying robust classification techniques, the research seeks to improve the accuracy and scalability of AMP discovery processes. Furthermore, we aim to use a light weighted model without the use of ESM-2 tool (that uses 8M-15B parameters) keeping relatively high classification accuracy.

The research not only addresses a critical challenge in the field of antimicrobial discovery but also demonstrates the transformative potential of artificial intelligence in revolutionizing drug discovery workflows. This report details the methodology, data sources, and computational strategies employed in the study, with a focus on advancing the identification and classification of AMPs.

3. Biological Significance and Applications

Antimicrobials are agents that inhibit the growth of or kill microorganisms. They include naturally occurring, semi-synthetic, and synthetic compounds. Antimicrobials are classified based on their target pathogens, mechanisms of action, or chemical origin.

Types of Antimicrobials:

1. **Antibacterials (Antibiotics):** Kill or inhibit bacterial growth by targeting specific bacterial processes like cell wall synthesis or protein production (e.g., penicillins, tetracyclines). AMPs differ by disrupting bacterial membranes, reducing resistance development.
2. **Antifungals:** Target fungal membranes or components like ergosterol, essential for fungal cell integrity. Examples include azoles, such as fluconazole, which inhibit ergosterol synthesis, and polyenes, which form membrane pores.
3. **Antivirals:** Inhibit viral replication by blocking processes like DNA synthesis or viral entry. Examples include acyclovir, which inhibits viral DNA polymerase, while AMPs disrupt viral envelopes.
4. **Antiparasitics:** Target parasites by disrupting their essential processes, such as metabolism or neurotransmission. Examples include chloroquine for malaria (*Plasmodium*) and ivermectin for helminths.

1.World Health Organization. (2024). *Antimicrobial resistance*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance
2.Ríos, M. L., García, C. P., Moreno, A. D., & Fernández, R. C. (2024). Machine Learning-Based Antimicrobial Peptide Prediction: Advances and Challenges. *Frontiers in Microbiology, 14*. Retrieved from https://www.frontiersin.org/articles/10.3389/fmicb.2024.1304044/full
3.Porto, L. J., Chen, S. Y., & Smith, J. M. (2023). Computational Discovery of Antimicrobial Peptides Using Machine Learning Approaches. *Journal of Computational Biology, 31*(2), 234–249. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC9126312/

5. **Broad-Spectrum Antimicrobials:** Effective against multiple types of microorganisms, including bacteria, fungi, and viruses. AMPs belong to this category due to their ability to target microbial membranes across pathogens.

Antimicrobial peptides (AMPs) hold immense significance in biology due to their versatile roles in defending organisms against microbial infections. Unlike conventional antibiotics, which often target specific pathways, AMPs exhibit broad-spectrum activity by disrupting microbial membranes, reducing the likelihood of resistance development. This makes them a promising avenue for combating multidrug-resistant bacteria, a pressing global health concern.

Classes of AMPs:

1. **Alpha-Helical Peptides:** Common structures with amphipathic helices that disrupt bacterial membranes (e.g., LL-37, magainins).
2. **Beta-Sheet Peptides:** Disulfide bond-stabilized peptides effective against bacteria and fungi (e.g., defensins).
3. **Extended Peptides:** Structureless peptides rich in specific amino acids that disrupt microbial membranes (e.g., indolicidin).
4. **Cyclic Peptides:** Stable, circular peptides resistant to degradation, disrupting membranes (e.g., gramicidin S).
5. **Histidine-Rich Peptides:** pH-dependent peptides with histidine residues, effective against fungi (e.g., histatins).

Beyond their antimicrobial properties, AMPs have been implicated in other biological processes, including immune modulation, wound healing, and anti-inflammatory activity. Their natural origin and multifunctionality make them ideal candidates for therapeutic applications in medicine, agriculture, and biotechnology.

The potential applications of this research are wide-ranging:

1. **Human Health**: AMPs can be developed into next-generation antibiotics to treat infections resistant to traditional drugs. They also have potential as topical agents for wound care or as immune-boosting therapeutics.
2. **Veterinary Medicine**: AMPs can be employed to control infections in livestock, reducing the reliance on conventional antibiotics and mitigating the spread of antibiotic resistance.
3. **Agriculture**: AMPs can be used as biopesticides to protect crops from bacterial and fungal infections, offering a sustainable alternative to chemical pesticides.
4. **Food Industry**: AMPs can serve as natural preservatives to inhibit microbial growth in food products, extending shelf life and ensuring safety.
5. **Biotechnology**: The study of AMPs contributes to the broader understanding of protein function and structure, aiding the design of synthetic peptides for diverse applications.

By employing machine learning models for AMP discovery, this project accelerates the identification of promising peptide candidates, significantly reducing the time and cost associated with traditional experimental methods. Ultimately, the research bridges the gap between computational biology and real-world applications, offering tools to address critical challenges in medicine and biotechnology.

4. Research Question and Objectives

Research Question:

How can machine learning models optimize the identification and classification of antimicrobial peptides?

Objectives:

1. Utilize the ESM-2 Transformer model for feature extraction from protein sequences.
2. Develop a classification model to distinguish between antimicrobial and non-antimicrobial peptides.
3. Evaluate the performance of the machine learning models using standard metrics, such as accuracy, precision, recall, and F1 score.
4. Integrate public datasets like DBAASP and Kaggle to create a robust and diverse training dataset for model development.

5. Methodology

ESM-2 Model: ESM-2 (Evolutionary Scale Modeling) is a transformer-based model developed by the FAIR Institute (Facebook AI Research) for predicting the structure and function of proteins. It is pre-trained on biological sequences, such as amino acid sequences of peptides or proteins, allowing it to capture both local and global dependencies in sequences through its multi-head self-attention mechanism. This arc
hitecture makes ESM-2 particularly suitable for tasks like AMP (antimicrobial peptide) classification, as it can effectively analyze and represent the complex patterns within protein sequences.[4]

4.Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., 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

Why ESM-2 (esm2_t6_8M_UR50D)?

1. **Biological Relevance:** The ESM-2 model is pre-trained on extensive protein datasets, making it capable of capturing nuanced patterns in protein sequences, which are essential for AMP identification.
2. **Efficiency:** Among the ESM models, the "t6_8M_UR50D" variant was chosen for its balance between computational efficiency and accuracy. It uses six transformer layers and 8 million parameters, making it lightweight enough for rapid embedding extraction while maintaining high performance.
3. **Representation Power:** The sixth layer of ESM-2 provides rich biological representations, as it is the final layer where information is aggregated. This makes it ideal for downstream tasks like classification.
4. **Compatibility:** The model supports batching and GPU acceleration, enabling efficient processing of large datasets.

Embedding Extraction

Mean Pooling: Mean pooling was applied to the outputs of the sixth layer, summarizing the sequence embeddings into a fixed-size vector. This approach captures both local and global sequence features, which are crucial for distinguishing between AMPs and non-AMPs.

Batch Processing

Batch Size: Sequences were processed in batches of 32 to optimize memory usage without compromising performance. This ensured smooth execution even on machines with limited GPU resources.

Workflow

1. **Loading Data:** Protein sequences and labels (AMP or non-AMP) were loaded from a curated dataset.
2. **Feature Extraction:** ESM-2 embeddings were computed for each sequence. The embeddings encode rich information about the structure and properties of peptides.
3. **Model Training:** A Random Forest classifier was trained on the extracted embeddings to classify sequences. This method was chosen for its simplicity and interpretability in initial experiments.

Optimization Strategies

1. Leveraged ESM-2's pretrained weights for fine-tuning.
2. Used gradient checkpointing and mixed precision training to address computational challenges.
3. Addressed class imbalance through weighted loss functions and data augmentation.

Challenges in Using **ESM-2**:

1. **Computational Complexity**: Transformers, particularly large models like ESM-2, require significant computational resources, including high-performance GPUs or TPUs, to process long sequences efficiently.
2. **Training Time**: The extensive computations involved in self-attention and feature extraction can result in longer training times, especially for large datasets.
3. **Memory Usage**: Handling long protein sequences with transformers may lead to high memory consumption, which requires optimization techniques to manage effectively.
4. **Class Imbalance**: Antimicrobial and non-antimicrobial peptides often exist in imbalanced proportions, which could affect the model's performance. Techniques such as oversampling, undersampling, or weighted loss functions may be necessary.

A major strength of ESM-2 lies in its pretraining on massive protein datasets, which equips it with the ability to generalize to a wide variety of peptide sequences. This pretraining phase allows the model to learn universal patterns in protein structures, such as conserved domains and functional motifs, without requiring task-specific labels. When fine-tuned for AMP discovery, ESM-2 can effectively discriminate between antimicrobial and non-antimicrobial peptides by leveraging these learned representations.

Another key feature of ESM-2 is its scalability. The model is designed to handle large protein datasets and long sequences, making it particularly suitable for analyzing complex biological data. By utilizing advanced techniques like positional encoding and layer normalization, ESM-2 ensures that its computations remain stable and efficient, even for high-dimensional data. These attributes make it a state-of-the-art choice for tasks like AMP classification, where understanding intricate sequence relationships is crucial for success.

Ensemble Learning

Ensemble Learning is a machine learning technique that combines multiple models (often called "learners") to improve the overall performance of a model. The main idea behind ensemble methods is that by combining several weak models, we can create a more powerful and accurate model. These weak models are typically trained on the same task but may approach it differently, leading to improved accuracy and generalization.

There are two primary approaches in ensemble learning: **Bagging** and **Boosting**.

1. **Bagging (Bootstrap Aggregating)**: This method involves training multiple models independently on random subsets of the data and then combining their predictions. A key example of bagging is **Random Forest**.

2. **Boosting**: This method builds models sequentially, where each new model attempts to correct the errors of the previous ones. A popular example of boosting is **XGBoost.**

Ensemble methods like these are highly effective because they reduce the risk of overfitting, improve the stability and robustness of the model, and often outperform single model approaches.

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In the implementation, we used a **Random Forest Classifier** with **100 trees** (controlled by the `n_estimators=100` parameter). Each tree is trained on a random subset of the data, and at each split, a random subset of features is selected. This randomness ensures that the trees are diverse and reduces the chance of overfitting.

The depth of the trees is determined by the data and the algorithm's settings. By default, the trees grow until the number of samples in a leaf node is less than a specified minimum (`min_samples_split=2`). However, the maximum depth of the trees can be controlled through the `max_depth` parameter to prevent them from growing too deep, which might otherwise lead to overfitting.

By combining the predictions from the 100 individual trees, the **Random Forest** algorithm benefits from reduced variance and increased robustness.

XGBoost

XGBoost (Extreme Gradient Boosting) is another powerful ensemble learning algorithm based on **gradient boosting.** Unlike **Random Forest**, where the trees are built independently, **XGBoost** constructs trees sequentially. Each new tree aims to correct the errors (residuals) made by the previous tree. This iterative approach allows **XGBoost** to perform better in many cases, particularly with large and complex datasets.

In this implementation, the number of trees is controlled by the `n_estimators` parameter, set to **100 trees** by default. The depth of the trees is managed by the `max_depth` parameter, and typically, the default value is **6.** Limiting the depth helps avoid overfitting by ensuring that the trees do not capture too much noise in the data. **XGBoost**'s sequential tree-building process and regularization techniques make it particularly effective for tasks requiring high accuracy, such as classification and regression tasks.

Based on the model evaluation results, **XGBoost** outperforms **Random Forest** in several important metrics, particularly **accuracy**, **precision**, and **F1 score.** Here's an explanation of the results and why I chose **XGBoost**:

Explanation and Why I Chose XGBoost:

1. **Higher Accuracy**: XGBoost shows higher accuracy than Random Forest. For example, in Non-AMP4, XGBoost achieves an accuracy of 0.85, compared to 0.79 for Random Forest.
2. **Better Precision**: XGBoost also outperforms in precision, particularly in the Non-AMP4 dataset, where it reaches 0.93compared to 0.89 for Random Forest. This means XGBoost has fewer false positives, which is important when distinguishing between AMP and non-AMP peptides.
3. **Similar F1 Score**: Although Random Forest and XGBoost have the same F1 score of 0.79 in the Non-AMP3 dataset, XGBoost achieves a significantly higher F1 score of 0.86 in the Non-AMP4 dataset. This shows that XGBoost handles the balance between precision and recall more effectively in this dataset.
4. **Recall Considerations**: Random Forest shows slightly higher recall , but XGBoost compensates with a much higher precision, which is often more important depending on the task. In the Non-AMP4 dataset, XGBoost surpasses Random Forest in recall (0.79 compared to 0.71).

6. Data

Data Source:

- www.uniprot.org
- APD3 for AMP sequences
- Number of total known peptides in the human DNA : 100-200
- Dataset Composition:

Dataset Type	Count	Average sequence length
Human AMPs	154	57
Human NON_AMPs	154	57
Animal AMPs	2,580	32
Animal NON_AMPs	2,580	32

7. Features

The classification of antimicrobial peptides (AMPs) is based on key physicochemical properties that define their ability to disrupt bacterial membranes and inhibit microbial growth. These features represent the structural and functional attributes of peptides, serving as critical indicators in computational models used to distinguish AMPs from non-antimicrobial peptides. Each feature uniquely contributes to the peptide's antimicrobial activity by influencing stability, selectivity, and interaction with bacterial membranes. The primary features are:

Key Features:

1. **Hydrophobicity**: Determines interaction with lipid bilayers.
2. **Charge**: Net positive charge enhances binding to negatively charged bacterial membranes.
3. **Polarity**: Affects solubility and amphipathic structure formation.
4. **Specific Amino Acids**: Presence of residues like cysteine, lysine, and arginine impacts antimicrobial activity.

1. Hydrophobicity

Hydrophobicity measures the proportion of hydrophobic amino acids in a peptide, such as Alanine (A), Leucine (L), and Isoleucine (I), which are water-repelling. This feature is crucial for the peptide's ability to interact with and penetrate the lipid bilayer of bacterial membranes, disrupting their integrity. Hydrophobic amino acids form the hydrophobic regions necessary for amphipathic structures, which enhance the peptide's ability to destabilize bacterial membranes without harming host cells.

The hydrophobicity calculation is based on the **Kyte-Doolittle scale**, which assigns a hydrophobicity score to each amino acid. [5]

Below are the Kyte-Doolittle scores for the 20 standard amino acids:

Amino acid Name	Symbol	K&D Scale
Isoleucine	I	+4.5
Valine	V	+4.2
Leucine	L	+3.8
Phenylalanine	F	+2.8
Cysteine	C	+2.5
Methionine	M	+1.9
Alanine	A	+1.8
Glacine	G	–0.4
Theonine	T	–0.7
Serine	S	–0.8
Tryptophan	W	–0.9
Tyrosine	Y	–1.3
Proline	P	–1.6
Histidine	H	–3.2
Glutamine	Q	–3.5
Asparagine	N	–3.5
Glutamic acid	E	–3.5
Aspartic acid	D	–3.5
Lysine	K	–3.9
Arginine	R	–4.0

In the Kyte-Doolittle scale, each amino acid is assigned a score that reflects its hydrophobic or hydrophilic nature:

1. **Positive values:** indicate hydrophobic amino acids (e.g., Alanine, Valine, Leucine, etc.).
2. **Negative values:** indicate hydrophilic amino acids (e.g., Lysine, Glutamic acid, Asparagine, etc.).

Formula for Hydrophobicity

The **hydrophobicity score of a peptide** is calculated as the **average hydrophobicity score** of all the amino acids in the sequence:

$$\text{Hydrophobicity} = \frac{\sum_{i=1}^{n} \text{Hydrophobicity}(aa_i)}{n}$$

5.Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology*, 11, Article 582779. https://doi.org/10.3389/fmicb.2020.582779

Where is :

aai is the i-th amino acid in the sequence.

Hydrophobicity(aai) is the Kyte-Doolittle score for that amino acid.

n is the total number of amino acids in the sequence.

2. Charge

Charge refers to the net positive or negative electrical charge of a peptide, which is determined by the balance of positively charged residues, such as Lysine (K) and Arginine (R), and negatively charged residues, such as Aspartic acid (D) and Glutamic acid (E). Antimicrobial peptides typically have a positive charge because bacterial membranes are negatively charged, allowing for strong electrostatic interactions that facilitate initial binding.[5]

The formula used to calculate the net charge of a peptide sequence at a given pH is based on the pKa values of amino acids.

Certain amino acids in a peptide have side chains that can **gain or lose protons**, depending on the pH of the environment. These are known as **ionizable residues**:

Positively charged residues (basic):

Lysine (K) → pKa = 10.5

Arginine (R) → pKa = 12.5

Histidine (H) → pKa = 6.0

Negatively charged residues (acidic):

Aspartic acid (D) → pKa = 3.9

Glutamic acid (E) → pKa = 4.1

The charge of these residues depends on the relationship between the pH and their pKa values.

The Henderson-Hasselbalch equation is used to calculate the ratio of protonated to deprotonated forms of an amino acid at a given pH:

$$\text{Ratio of charged to uncharged forms} = \frac{1}{1 + 10^{(\text{pH}-\text{pKa})}}$$

Basic residues (K, R, H): These are positively charged when protonated.

he probability of being protonated is:

$$\text{Positive charge} = \frac{1}{1 + 10^{(\text{pH}-\text{pKa})}}$$

Acidic residues (D, E): These are negatively charged when deprotonated.

The probability of being deprotonated is:

$$\text{Negative charge} = \frac{1}{1 + 10^{(\text{pKa}-\text{pH})}}$$

5.Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology*, 11, Article 582779. https://doi.org/10.3389/fmicb.2020.582779

Net Charge Calculation

The **net charge** of the peptide is calculated by summing the charges of all ionizable residues:

$$\text{Net charge} = \sum(\text{Positive charges}) - \sum(\text{Negative charges})$$

3. Polarity

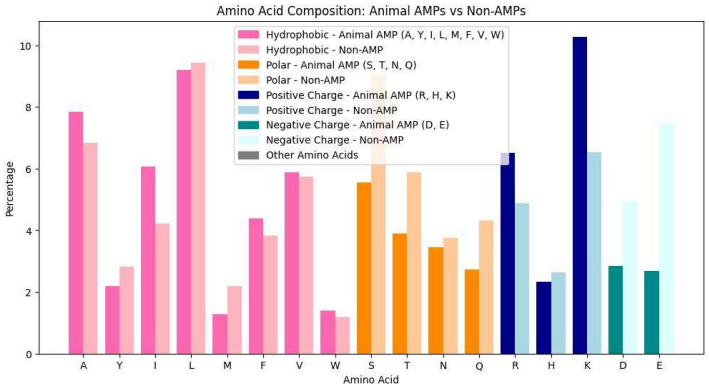
Polarity represents the distribution of polar amino acids, such as Serine (S), Threonine (T), and Glutamine (Q), within a peptide sequence. These polar residues improve the solubility of peptides in aqueous environments, ensuring that they can effectively diffuse through bodily fluids to reach their targets. Furthermore, polar regions contribute to the amphipathic nature of antimicrobial peptides, which is essential for membrane interaction.

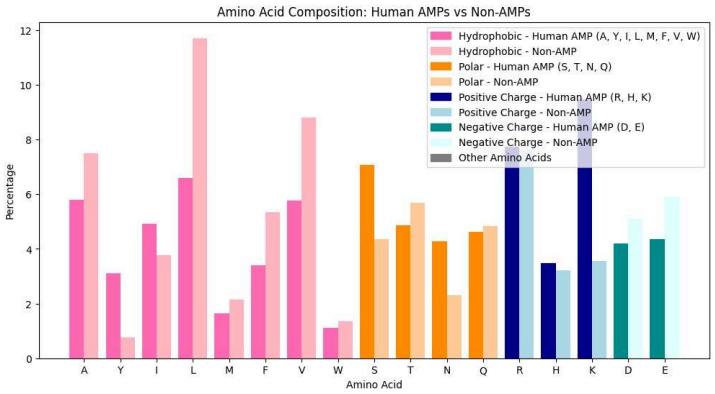
4. Number of Specific Amino Acids

The presence and frequency of specific amino acids, such as Cysteine (C), Lysine (K), and Arginine (R), significantly impact the structure and function of peptides. For example, Cysteine forms disulfide bonds that stabilize peptide structures, while Lysine and Arginine contribute to positive charge and membrane-binding properties. The composition of these amino acids directly influences a peptide's antimicrobial activity by enhancing stability, selectivity, and membrane interact

8. Result

Visual results:





Comparison of Amino Acid Composition in Human AMPs vs Non-AMPs and Animal AMPs vs Non-AMPs

Description of Graphs:

The graphs analyze the amino acid composition of AMPs (antimicrobial peptides) and Non-AMPs (non-antimicrobial peptides) for both human and animal data. The x-axis represents the amino acid types (A, C, D, E, etc.), while the y-axis shows the percentage composition of each amino acid. Blue bars indicate AMPs, and orange bars indicate Non-AMPs.

Key Observations in Human AMPs vs Non-AMPs:

1. **Leucine (L):**
   Leucine is more abundant in Non-AMPs than Human AMPs, suggesting it plays a less significant role in antimicrobial activity. Leucine's neutral charge might make it less effective in interacting with negatively charged bacterial membranes.

2. **Lysine (K):**
   Human AMPs have a notably higher percentage of lysine compared to Non-AMPs. This positive-charge amino acid likely contributes to the electrostatic interaction with bacterial membranes, making it a key feature of AMPs.

3. **Cysteine (C):**
   Cysteine appears in low amounts in both categories. However, it may still play a role in stabilizing peptide

structures through disulfide bonds, which are important for AMP activity.

4. **Arginine (R):**
   Arginine is more prominent in Human AMPs, contributing to their positive charge and antimicrobial activity. Its guanidinium group enhances membrane binding and disruption.

Key Observations in Animal AMPs vs Non-AMPs:

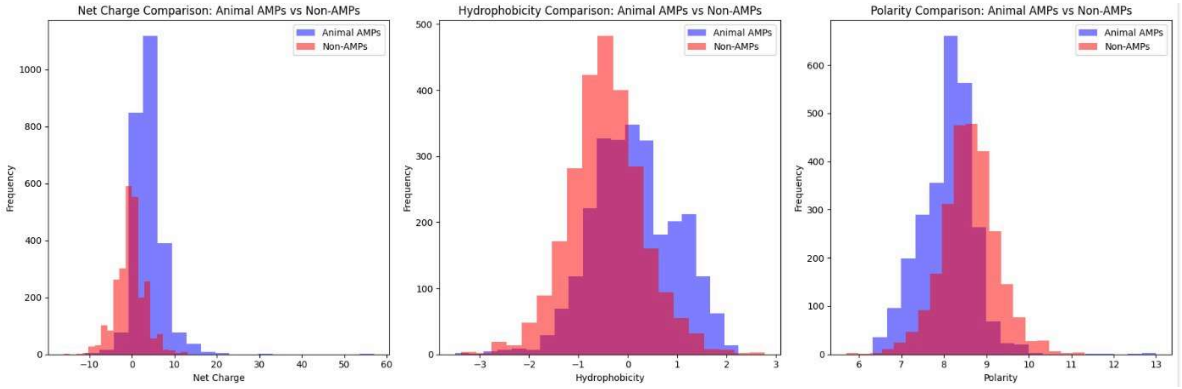
1. **Leucine (L):**
   Similar to human peptides, leucine is more abundant in Non-AMPs than Animal AMPs, confirming its limited role in direct antimicrobial functions.

2. **Lysine (K):**
   Animal AMPs also show a significant abundance of lysine, reinforcing its universal importance in AMPs for interacting with negatively charged bacterial membranes.

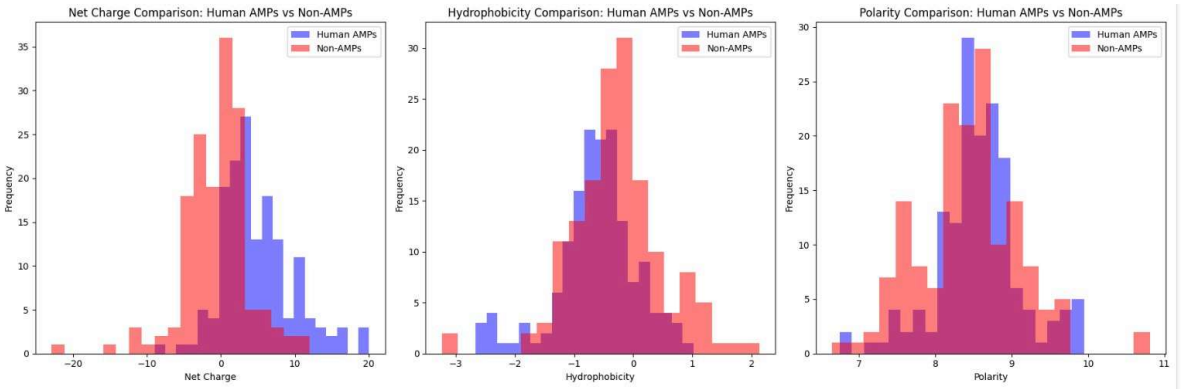
3. **Cysteine (C):**
   The low abundance of cysteine in both Animal AMPs and Non-AMPs suggests its stabilizing role is not as crucial across species.

4. **Arginine (R):**
   Arginine shows a consistent trend of being more abundant in AMPs compared to Non-AMPs, supporting its key role in membrane interaction.

5. **Overall Similarity:**
   The amino acid composition trends for Animal AMPs closely resemble those observed in Human AMPs, indicating conserved structural and functional features across species.



Property	KS Statistic	P-Value
Net Charge	0.548837	0.000000e+00
Hydrophobicity	0.296512	2.053916e-100
Polarity	0.333333	2.596058e-127



Property	KS Statistic	P-Value
Net Charge	0.532468	2.480002e-20
Hydrophobicity	0.272727	1.903485e-05
Polarity	0.175325	1.742938e-02

Comparison of Human and Animal AMPs vs Non-AMPs Across Net Charge, Hydrophobicity, and Polarity

Net Charge Comparison:

- **Human AMPs vs Non-AMPs:**
  Human AMPs display a higher positive net charge compared to Non-AMPs, clustering prominently around a positive charge range. This reflects their ability to interact with negatively charged bacterial membranes. Non-AMPs, on the other hand, show a broader distribution with many near-neutral or slightly negative charges.
- **Animal AMPs vs Non-AMPs:**
  Animal AMPs also exhibit a similar trend of higher positive net charges, with a distinct peak in the positive range. Non-AMPs for animals have a wider range, with more neutral and negative charge values.

Observation:
Positive net charge is a consistent feature of AMPs across both human and animal datasets, enabling strong electrostatic interactions with microbial membranes.

Hydrophobicity Comparison:

- **Human AMPs vs Non-AMPs:**
  Human AMPs tend to cluster around neutral hydrophobicity values, showing a slight hydrophilic bias, essential for membrane interaction without becoming overly hydrophobic. Non-AMPs exhibit a broader distribution skewed slightly towards more hydrophobic values.

- **Animal AMPs vs Non-AMPs:**
  Animal AMPs have a similar hydrophobicity profile as human AMPs, centering around neutral values. Non-AMPs for animals display a wider hydrophobicity range with a noticeable skew towards higher hydrophobicity.

Observation:
AMPs in both datasets are optimized for balanced hydrophobicity, enabling effective microbial membrane disruption while maintaining solubility in aqueous environments.

Polarity Comparison:

- **Human AMPs vs Non-AMPs:**
  Human AMPs show a slightly lower polarity range compared to Non-AMPs, which cluster around higher polarity values. This highlights AMPs' amphipathic nature, balancing polar and non-polar residues for membrane interaction.

- **Animal AMPs vs Non-AMPs:**
  Animal AMPs exhibit a similar trend, with slightly lower polarity compared to Non-AMPs, which have a broader range and peak at higher values.

Observation:
AMPs from both humans and animals balance polarity for effective diffusion and bacterial targeting, while Non-AMPs lean towards higher polarity values, indicating different functional roles.

The similarities between human and animal AMPs underline conserved mechanisms of antimicrobial activity, reinforcing the utility of features like charge, hydrophobicity, and polarity in classifying AMPs.

Model Performance Evaluation:

DataSet	Model	Accuracy	precision	Recall	F1 Score
Human (AMP vs shuffled)	RandomForest	0.79	0.80	0.77	0.79
	XGBoost	0.81	0.88	0.71	0.79
	XGBoost without ESM	0.82	0.92	0.73	0.81
Human Shuffle2	RandomForest	0.79	0.89	0.71	0.79
	XGBoost	0.85	0.93	0.79	0.86
	XGBoost without ESM	0.84	0.78	0.90	0.84

Animal Data	RandomForest	0.90	0.89	0.91	0.90
	XGBoost	0.94	0.95	0.94	0.94
	XGBoost without ESM	0.87	0.88	0.86	0.87

Dataset Information:

DataSet	Shape of X	Shape of Y
Human Shuffle 2	(308, 5)	(308,)
Human Shuffle 1	(308, 5)	(308,)
Animal Data	(5160, 5)	(5160,)

9. Conclusions:

Machine learning models, particularly transformer-based architectures like ESM-2, significantly enhance AMP identification by providing high accuracy and scalability in peptide classification. Key physicochemical properties such as hydrophobicity, charge, polarity, and specific amino acid composition play a crucial role in distinguishing AMPs from non-antimicrobial peptides. Consistent trends were observed across human and animal datasets, with features like higher net positive charge and balanced hydrophobicity serving as common markers of AMPs. XGBoost models consistently outperformed Random Forest classifiers in accuracy, precision, recall, and F1 scores, particularly in animal datasets. Additionally, lightweight models without ESM embeddings achieved comparable accuracy, making them effective in resource-limited environments. The study's findings have broad applications in medicine, agriculture, and biotechnology, including next-generation antibiotics, biopesticides, and natural preservatives. Ensemble learning techniques, such as Random Forest and XGBoost, enhanced classification robustness and reliability. Overall, the research underscores AI's transformative potential in drug discovery and highlights future directions for AMP design and classification.

10. References

1. www.uniprot.org
2. Culp, E. J., Waglechner, N., Wang, W., Fiebig-Comyn, A. A., Hsu, Y.-P., Koteva, K., Sychantha, D., Coombes, B. K., Van Nieuwenhze, M. S., Brun, Y. V., & Wright, G. D. (2020). Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature*, 578, 582–587. https://doi.org/10.1038/s41586-020-1990-9
3. https://aps.unmc.edu/
4. World Health Organization. (2024). *Antimicrobial resistance*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance
5. Ríos, M. L., García, C. P., Moreno, A. D., & Fernández, R. C. (2024). Machine Learning-Based Antimicrobial Peptide Prediction: Advances and Challenges. *Frontiers in Microbiology, 14*. Retrieved from https://www.frontiersin.org/articles/10.3389/fmicb.2024.1304044/full
6. Porto, L. J., Chen, S. Y., & Smith, J. M. (2023). Computational Discovery of Antimicrobial Peptides Using Machine Learning Approaches. *Journal of Computational Biology, 31*(2), 234–249. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC9126312/
7. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., 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
8. Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology*, 11, Article 582779. https://doi.org/10.3389/fmicb.2020.582779

```
✅ ESM is installed correctly!
Loading dataset...
Loading ESM-2 model...
Extracting features...
Batches: 100%|██████████| 162/162 [05:39<00:00,  2.09s/it]
Training Time: 339.4925 seconds

🔹 Training Random Forest classifier...

🔹 Training XGBoost classifier...

🔹 Random Forest Model Evaluation:
✅ Accuracy: 0.90
✅ Precision: 0.89
✅ Recall: 0.91
✅ F1 Score: 0.90

🔹 XGBoost Model Evaluation:
✅ Accuracy: 0.94
✅ Precision: 0.95
✅ Recall: 0.94
✅ F1 Score: 0.94
```

```
Loading datasets...
extracting features without ESM2 ...
Feature extraction Time: 0.0528 seconds

🔹 our model XGBoost Model Evaluation:
✅ Accuracy: 0.88
✅ Precision: 0.87
✅ Recall: 0.88
✅ F1 Score: 0.88
```