CIS 5200 Fall 2025, Project Report

Investigating a Multi-modal Approach for Antimicrobial Peptide Screening

Group Members:

- Kelly Fay
- Kevin Shen
- Pedram Bayat

Abstract

Addressing the critical global health crisis posed by antimicrobial resistance (AMR), this project utilized a deep learning approach to mine peptide proteomes, focusing on the novel application of molecular de-extinction to identify potential antimicrobial peptides (AMPs) from ancient organisms. We developed and benchmarked an APEX-based Sequence Encoder, which achieved robust in-distribution performance (Spearman rho = 0.79), confirming its utility for sequence-based prediction. However, testing the model on an out-of-distribution (OOD) dataset of extinct peptides resulted in a significant performance collapse (rho = 0.36), confirming the challenge of evolutionary generalization. To stabilize OOD prediction, we investigated the hypothesis that 3D context is required, integrating five explicit AlphaFold structural features via transfer learning. Feature importance analysis revealed that the model rejected the structural data (contributing ¡1% feature weight), providing a crucial scientific insight: the low-value of static structural predictions (e.g., disordered state) for dynamic, membrane-targeting AMPs. This work successfully validates a high-performance sequence baseline but ultimately establishes that the primary bottleneck in next-generation AMP discovery lies in the current inability to model dynamic, bio active structural states.

1 Motivation

Antimicrobial resistance (AMR) refers to the evolution of bacteria, viruses, fungi, and parasites against traditional antimicrobial compounds. The World Health Organization predicts that AMR will be associated with over 10 million annual deaths by 2050 [5].

Traditional experimental antibiotic discovery cannot keep pace with the rate at which bacteria develop resistance mechanisms due to the high costs of development, the depletion of easily accessible natural reservoirs, and complex regulatory hurdles. As a result, computational and artificial intelligence approaches have recently emerged as powerful tools for high-throughput antibiotic discovery [7].

Specifically, novel antibiotics can be discovered through proteome mining, a data-driven exploration of the complete set of proteins in an organism to infer structural, functional, and evolutionary properties. In previous work, this approach has identified antimicrobial peptide (AMP) candidates, which are short protein sequences serving as an innate defense mechanism across a wide range of species [2]. These computational efforts have the potential to substantially accelerate early-stage discovery of antimicrobial compounds.

Critically, this computational approach enables molecular de-extinction, the novel application of AI to mine the proteomes of extinct organisms. By analyzing ancient protein sequences (such as those from the woolly mammoth), researchers can identify evolutionarily distant AMP candidates that current organisms no longer possess. This process represents a unique and powerful avenue for finding novel antibiotics untouched by modern resistance mechanisms.

Altogether, the aim of this project is twofold.

Aim 1: To train a deep learning model that learns high-dimensional latent representations of antimicrobial peptides directly from raw sequence data and to rigorously compare its accuracy with traditional baseline models.

Aim 2: To test the generalization of this model on an out-of-distribution dataset of extinct peptides.

2 Related Work

2.1 Deep Learning for Antimicrobial Peptide Discovery

Deep learning models have been employed to mine proteomes for the identification of antimicrobial peptides. A prominent example, which heavily inspired this project, is the APEX (Antibiotic Peptide de-Extinction) model [7]. APEX uses a multitask deep learning architecture consisting of a peptide-sequence encoder combining recurrent neural networks (RNNs) with attention mechanisms to extract hidden features from peptide sequences. Specifically, hidden features are extracted using a Gated Recurrent Unit (GRU), followed by attention layers which are fed into fully connected neural networks (FCNNs) to perform two distinct tasks:

1. Predict species-specific antimicrobial activity against pathogenic bacteria.
2. Predict binary antimicrobial activity using public data as a form of data augmentation.

Antimicrobial activity is quantified by the minimum inhibitory concentration (MIC), which is the lowest concentration of a peptide sequence necessary to prevent the growth of a pathogen in the laboratory. MIC can then be binarized to classify whether a certain peptide sequence is an antimicrobial peptide. For example, APEX defined an inactive peptide as any peptide sequence with an MIC greater than 30 μmol^{-1}.

3 Data Set

3.1 Large-Scale Training Data

The publicly available Database of Antimicrobial Activity and Structure of Peptides (DBAASP) [6] was used to train the encoder and the baseline models for Aim 1. A total of 23944 unique sequences, their corresponding features (e.g. peptide complexity, synthesis type, N terminal), and their peptide-target interactions were mined. The dataset also includes information about bacterial targets, particularly target group, target object, and minimum inhibitory concentration (MIC) values for each of the 7879 targets in the dataset.

3.2 Feature Engineering and Data Preprocessing

Raw peptide entries from DBAASP were transformed into a structured peptide-target MIC feature matrix to enable machine-learning-based prediction. Because MIC data were extremely sparse across the 7,879 microbial targets (Fig. 1B), a filtering step was applied in which all targets with fewer than 200 non-zero MIC values were removed. This procedure reduced the set of microbial targets to 50 high-coverage targets (Fig. 1D).

Following target filtering, the dataset was converted into a long-format representation, in which each row corresponded to a unique (peptide, target) pair with an associated MIC value. This restructuring substantially reduced sparsity and produced a dense, learning-ready dataset (Fig. 1C).

For the input feature set, physicochemical peptide properties were calculated by modlAMP (version 4.3.0) [4], including sequence length, molecular weight, sequence charge, charge density, isoelectric point, instability index, aromaticity, aliphatic index, Boman index, and hydrophobic density. Missing numerical values were imputed using the median of the cleaned dataset to preserve distribution characteristics. Additionally, irrelevant columns were dropped and categorical features such as peptide synthesis type and bacterial target groups were label encoded.

Upon completion of preprocessing, the dataset included 25,306 unique peptide-target interactions, with 6,233 total unique sequences and 50 bacterial targets, along with important physicochemical information for each peptide (Fig. 1A).

3.2.1 Class Distribution

The dataset includes a variable `is_AMP` indicating if a peptide is an active AMP or an inactive peptide, based on the earlier defined MIC threshold (Fig. 1e). Based on the number of peptide-target interactions classified as having AMP activity, a class imbalance was observed, thus emphasizing the need for data stratification in train-test splits.

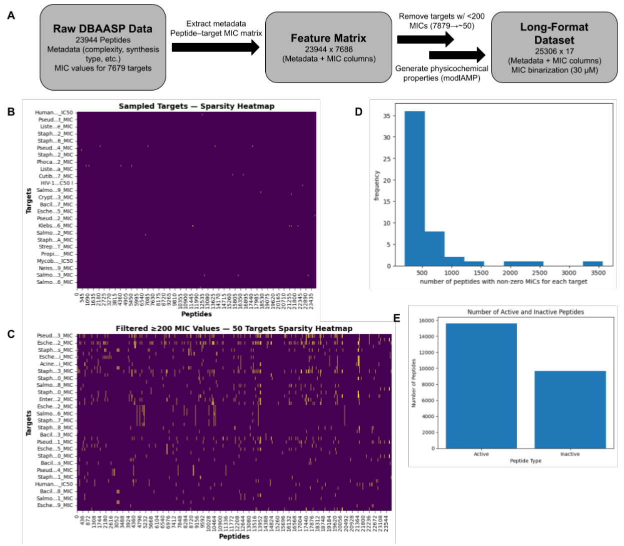

Figure 1: **A)** Preprocessing pipeline **B)** Sparsity heatmap showing the distribution of non-zero MIC values across a randomly sampled subset of targets. **C)** Sparsity heatmap for the filtered set of targets with ≥ 200 non-zero MIC measurements, illustrating the substantially denser target-peptide matrix. **D)** Histogram showing the distribution of targets that exceed the 200-peptide threshold and number of peptides with non-zero MIC values. **E)** Bar graph showing binarized peptide-target interactions.

3.3 Experimentally Validated AMPs

A set of 69 extinct peptides experimentally assessed in the APEX study was incorporated as an external evaluation cohort. Of these peptides, 41 demonstrated antimicrobial activity against at least one bacterial strain. Because these sequences were previously identified computationally as high-probability AMPs, they provide a high-quality benchmark for evaluating model generalization. The objective of this work is to incorporate additional structural and feature-based information to improve discrimination of the 41 validated AMPs from the broader set of 69 peptides.

3.4 Curation of Structural Data with AlphaFold

Augmentation of sequence-based features with structural information from AlphaFold (AF) [1] was explored for Aim 2, with the goal of creating a more accurate, multi-modal discovery pipeline. The less GPU-intensive ColabFold [3] implementation was employed to generate AlphaFold representations. However, this process was impeded by significant technical and resource limitations.

4 Problem Formulation

We formulate the prediction of log2-transformed MIC as a supervised regression task. This approach preserves the continuous nature of peptide potency, prioritizing the accurate modeling of extreme values critical for drug ranking.

4.1 Feature Engineering & Representation

Unlike prior studies that averaged MIC values, we employ a long-format representation where each row corresponds to a unique peptide-target pair. This allows the model to treat the bacterial strain as an explicit predictive feature, preserving specificity.

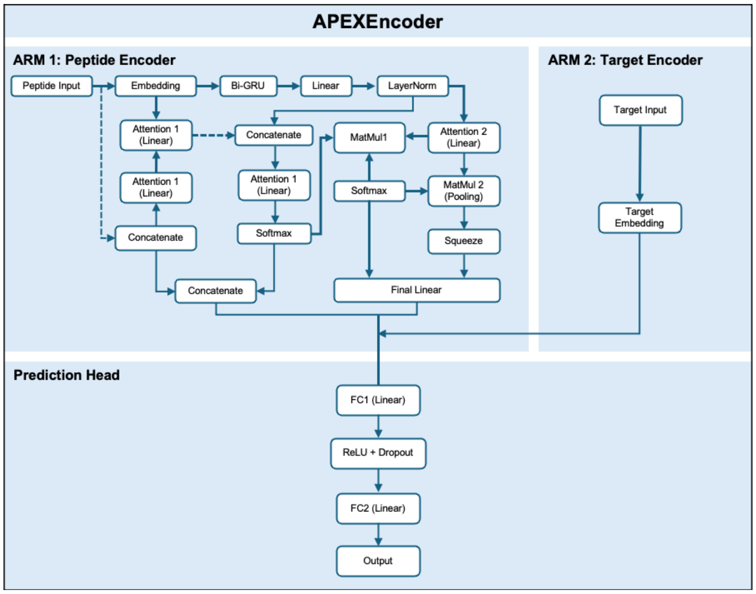

Figure 2: **APEXEncoder Architecture**. The model utilizes a dual-arm framework to predict log2(MIC). The Peptide Encoder (ARM 1) processes sequence embeddings with a bidirectional Gated Recurrent Unit (GRU), using a two-stage attention mechanism aggregating both hidden states and original input embeddings. The Target Encoder (ARM 2) maps target IDs to a dense vector space. Finally, the resulting feature representations are concatenated and processed through a Multi-Layer Perceptron (MLP) prediction head).

4.2 Model & AlphaFold Extension

To accommodate this multi-modal input, we implemented a Two-Arm Encoder (Fig. 2) processing peptide and target features separately. For Aim 2, we adopted a symmetric formulation, augmenting sequence features with condensed AlphaFold structural statistics to evaluate if explicit 3D context enhances prediction stability.

5 Methods

5.1 Initial Architecture: APEX-based Two-Arm Encoder

To predict Antimicrobial Peptide (AMP) activity, we implemented APEXEncoder, a custom Two-Arm Encoder Neural Network adapted from the APEX [7] architecture (Fig. 2). This architecture was chosen to effectively model the interaction between two distinct modalities: the peptide sequence and the target bacterial strain.

1. **Sequence Branch:** Peptide sequences are tokenized and processed through a Bi-directional Gated Recurrent Unit (Bi-GRU). This allows the model to capture sequential dependencies and contextual information from both N-terminal and C-terminal directions. An attention mechanism is applied to the GRU outputs to weight the importance of specific amino acid residues before pooling.

2. **Target Branch:** Bacterial strains are represented via learnable embeddings, allowing the model to capture latent similarities between different target organisms (e.g., Gram-positive vs. Gram-negative characteristics).

3. **Training and Optimization:** The model predicts log2-transformed Minimum Inhibitory Concentration (MIC) values. We trained across 5 random seeds with randomized 80/20 train/test splits to ensure robustness. Optimization was performed using Adam, with early stopping (patience=5) monitoring validation loss to prevent overfitting.

5.2 Baseline Benchmarking

To validate the necessity of a deep learning approach, we benchmarked APEXEncoder against traditional machine learning algorithms.

1. **Linear Models:** We evaluated Linear SVM and Nyström-SGD. These models were selected to confirm that the sequence-activity relationship is non-linear.

2. **Tree-Based Models:** We implemented Random Forest and XGBoost regressors. Unlike the linear baselines, these models captured significant signals, serving as a competitive baseline. Performance was evaluated using both R^2 and Spearman's Rank Correlation, the latter being critical for prioritizing drug candidates. All tree-based models were supplied with one-hot encoded bacterial targets to ensure fairness of comparison against the NN's Target Branch.

5.3 Hybrid Architecture

We hypothesized that explicit domain knowledge could guide the model in data-scarce regimes. We expanded the architecture into a Three-Arm "Hybrid" Encoder. This was accomplished by supplementing the peptide encoder and the target encoder with a physicochemical feature branch, where the 12 features calculated for the baseline models (e.g., net charge, hydrophobicity, isoelectric point, instability index) were normalized and processed through a dedicated dense layer before being concatenated with the sequence and target embeddings.

5.4 Ensemble Strategy

We implemented a Deep Ensemble by aggregating predictions from 25 independently trained models (5 independent training runs for 5 random seeds). While this reduced prediction variance and offered marginal stability improvements, the performance gain was insufficient to justify the 5-fold increase in computational inference cost.

5.5 Out-of-Distribution Generalization (Extinct Peptides)

To evaluate the model's utility for de novo drug discovery, we tested it on a strictly out-of-distribution (OOD) dataset of extinct peptide, sequences from ancient organisms (e.g., Woolly Mammoth, Mylodon) that share little homology with modern peptides. This generalization test mimics the real world scenario of mining novel biological sources for undiscovered AMPs.

5.6 Investigating Structural Contribution via AlphaFold Integration

In an attempt to improve performance on OOD extinct peptides, we initiated a technical investigation to determine whether explicit biophysical context could stabilize predictions where sequence homology fails. Particularly, we investigated the ability of protein structure prediction model AlphaFold in stabilizing predictions where sequence homology fails.

1. **Structural Feature Extraction:** Static structural models were generated using AlphaFold for 460 peptides in the novel datasets (limited due to the compute time of AlphaFold). Five quantitative features were engineered from the PDB files, including Mean pLDDT, Fractional Helix/Sheet, Backbone Rigidity, and Average Degree.

2. **Feature Fusion Test Design:** We employed a Transfer Learning approach where the pre-trained, 128-dimensional sequence embedding (trained on the full dataset) was extracted and concatenated with the five AlphaFold structural features. This composite vector was then fed into a simple Random Forest Regressor. This design ensures a direct, quantitative comparison of the predictive power contributed by sequence knowledge vs. structural context.

6 Experiments and Results

6.1 Baseline Performance Analysis

APEXEncoder demonstrated strong predictive capabilities, successfully converging and learning complex sequence-activity relationships. The model achieved an average R^2 of 0.619 ± 0.0139 and a Spearman-Rank Correlation (ρ) of 0.788, outperforming the metrics reported in the reference study, which achieved an R^2 of 0.37 and ρ of 0.55-0.62 for single models [7] (Fig. 3A). However, it is important to acknowledge that our use of a randomized 80/20 split likely simplified the generalization task compared to the cluster-based cross-validation or leave-one-species-out validation likely used in the reference paper.

6.1.1 Baseline Error Analysis

Parity and residual plots indicate the model "plays it safe", systematically over-predicting low MIC values (predicting them as less potent) and under-predicting high MIC values (predicting them as more potent). However, the high Spearman correlation (0.79) confirms the model effectively preserves rank order, which is the primary requirement for prioritizing lead candidates in drug discovery (Fig. 3B-C).

6.2 APEXEncoder Benchmarking

We benchmarked the Deep Learning approach against standard ML regressors to establish a performance hierarchy (Table 1).

Table 1: Benchmarking Results

Model	**Result**	
	R^2	ρ
APEXEncoder	0.62	0.79
Random Forest	0.61	0.78
XGBoost	0.61	0.78
Linear Baselines	$< .20$	–

This hierarchy validates that non-linear, high-capacity models are required to capture the complex grammar of antimicrobial peptides.

6.3 Hybrid Model Results (Injection of Physicochemical Features)

We hypothesized that injecting 12 explicit physicochemical features (e.g., charge, hydrophobicity) would improve performance. The hybrid model achieved an R^2 of 0.630 ± 0.0239, a statistically negligible improvement over the baseline. Feature importance analysis revealed that the model assigned $> 99\%$ importance to the learned sequence embeddings and $< 1\%$ to the explicit features(Fig. 3D). This suggests the deep sequence encoder had already implicitly learned these physical properties, rendering the explicit injection redundant.

6.4 Ensemble Performance

We evaluated a Deep Ensemble of 25 models to determine if aggregating predictions could reduce variance. The ensemble achieved an R^2 of 0.626 ± 0.0095 and an MSE of 0.37, offering only marginal gains over the single-best model. However, the variance did decrease (Fig. 3E). While ensembling provided a slight improvement in stability, the performance gain was insufficient to justify the 25x increase in training and inference cost. Given the high performance of our single models, we determined that model capacity was not the limiting factor.

6.5 Generalization to Extinct Peptides (OOD Testing)

To stress-test the model, we evaluated it on a dataset of extinct peptide sequences from the reference paper, a strictly out-of-distribution task involving ancient sequences. All models suffered a significant performance regression (Table 2), characteristic of OOD generalization issues (Fig. 3F).

Table 2: OOD Testing

Model	ρ
APEXEncoder	0.36
Random Forest	0.33
XGBoost	0.33

While our model retained a slight edge over the baselines, the drop in ρ highlights the difficulty of generalizing to evolutionary distant sequences based on sequence alone. The features were weak, and the model regressed to the mean.

6.6 3D Structure Integration (AlphaFold)

We integrated 3D structural features derived from AlphaFold predictions for approximately 460 peptides. However, the this yielded no improvement in predictions. Analysis of AlphaFold data revealed that $> 75\%$ of the peptides were predicted to have no secondary structures (Fig. 3G). This aligns with the fact that AMPs are intrinsically disordered in solution and only fold upon membrane contact. AlphaFold likely failed to capture the bioactive conformation, rendering the features non-informative for this specific predictive task.

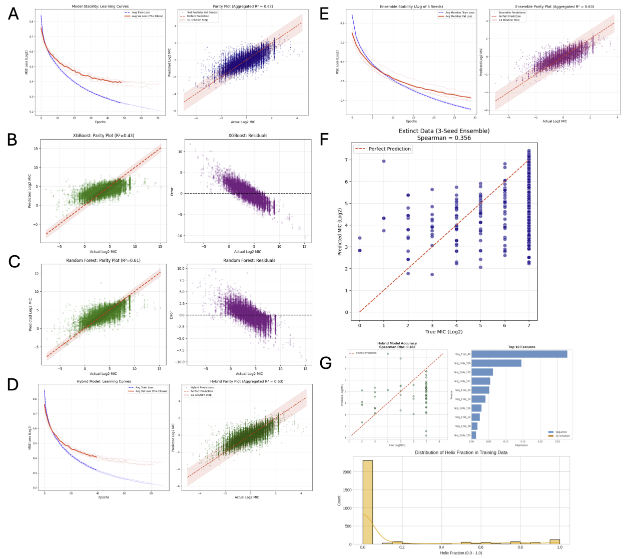

Figure 3: **A) XGBoost Baseline Performance:** Parity Plot and Residual Plot for the XGBoost regressor, illustrating a goodness of fit of $\mathbf{R^2 = 0.43}$ on the in-distribution data and confirming the model's non-linear signal capture. **B) and C) Baseline Model Performance:** Parity Plots (R^2) and Residual Plots for XGBoost (B) and Random Forest (C) regressors on the ID training data. **D) Deep Learning Model Diagnostics:** Learning Curves (and Aggregated Parity Plot for the Hybrid Neural Network, illustrating model stability and goodness of fit (R^2). **E) Ensemble Model Diagnostics:** Learning Curves and Aggregated Parity Plot for the Deep Ensemble. **F) Out-of-Distribution (OOD) Generalization:** Parity Plot for the Ensemble Model tested on the Extinct Peptide Data (OOD set). **G) Structural Feature Analysis:** Top: Hybrid Model Accuracy and Feature Importance, structural features (orange), sequence embeddings (blue). Bottom: Histogram showing the Distribution of Helix Fraction in the training peptides.

6.7 Reproducibility, Initialization, and Performance

To ensure reproducibility and robustness, all models were evaluated using a randomized 80/20 train/test split, with performance metrics reported as the average and standard deviation across five independent random seeds. The APEXEncoder was optimized using the Adam algorithm with early stopping (patience of 5 epochs) monitoring validation loss. Baseline tree-based models were initialized with fixed hyperparameters-`n_estimators=100` for Random Forest and `n_estimators=200` with `learning_rate=0.05` for XGBoost-to maintain consistency across comparisons.

6.8 Performance

Two primary metrics were used to asses model performance.

- **Coefficient of Determination(R^2):** Quantifies the proportion of the variance in the dependent variable (MIC) that is predictable from the independent variables.
- **Spearman Rank Correlation (ρ):** Used as the primary measure of the model's ranking capability. The Spearman coefficient validates the model's utility for prioritizing drug discovery hits.

7 Conclusion and Discussion

7.1 Summary of Findings and Model Assignment

The project successfully established a high-performance deep learning baseline for predicting Minimum Inhibitory Concentration (MIC), but critical limitations were revealed in the advanced feature-engineering efforts when evaluated on Out-of-Distribution (OOD) data. Baseline performance reached an $R^2 = 0.62$ and a Spearman correlation of $\rho = 0.79$, demonstrating stable in-distribution interpolation capability.

However, attempts to introduce explicit domain knowledge were found to be ineffective. Feature-redundancy analyses indicated that the inclusion of physicochemical descriptors and aggregated 3D structural features produced less than a 1% performance gain, suggesting that these external features were computationally redundant because the sequence encoder had implicitly learned the same information.

The most consequential observation arose from the generalization assessment. A critical failure point was identified when the model was evaluated on OOD "extinct" peptides: the Spearman correlation collapsed to $\rho = 0.36$, indicating that reliance on sequence information alone resulted in brittle performance for evolutionarily distant peptides.

Finally, ensemble modeling was deemed inefficient. The marginal improvement in ρ was not sufficient to justify the 25× increase in computational cost, indicating that ensemble complexity did not yield proportionate gains in predictive accuracy.

7.2 AlphaFold Exploration

While structural data is a novel feature that holds immense potential for enhancing our task, its effective implementation was hindered by significant contextual difficulties.

7.2.1 Computational Constraints and Data Representation Challenges

Generation of AF predictions was found to be highly demanding, running slowly and requiring more computational resources than were readily available. The primary bottleneck was memory consumption, which became problematic for longer sequences. Due to these resource constraints, after many days of running, AlphaFold representations were successfully generated for only 756 peptides of length ≤ 10 residues for training, out of a total of 6,233 unique sequences. Representations were also generated for 57 of the 69 extinct peptides used for testing.

A major challenge in data representation for subsequent machine learning model input was posed by the structural outputs from AlphaFold. Outputs from the most confident AF model included detailed per-residue features, such as phi and psi torsional angles, contact maps, and confidence scores for the positioning of each residue. Peptide sequences are inherently of variable length, resulting in feature vectors of variable dimensions. Consequently, individual AF representations could not be represented as a constant-dimensional (N-dimensional) array, which is required for standard machine learning models.

To address the invariant length challenge and make use of the partial structural data that had been generated, summary statistics were computed from the detailed per-residue AF outputs. This approach ensured that each peptide sequence was represented by a single, fixed-length vector in a master CSV, suitable for subsequent model training.

7.2.2 Future Dataset Generation

The most critical opportunity for future research lies in fully realizing the multi-modal goal of Aim 2. This requires overcoming the variable-length limitation to integrate rich structural information. With more available compute power, we would preferably run the full AlphaFold 2 system rather than ColabFold, as it offers higher computational efficiency and greater control over the generation process. This increased capacity is essential for generating structural representations for all 6,233 unique peptide sequences in the dataset.

Implementing a Graph Neural Network (GNN) has strong potential as a method to fully realize the multi-modal goal of Aim 2 and utilize the rich structural information from AlphaFold (AF). Future work must transition from fixed-input architectures to methods capable of processing graph-structured data. Implementing a Graph Neural Network (GNN) would allow the model to directly use the detailed, variable-sized AF outputs such as contact maps or predicted residue coordinates. This would fully test the hypothesis that structural knowledge noticeably improves MIC prediction without losing information in summary statistics.

Acknowledgments

We would like to extend our sincere gratitude to the teaching staff of CIS 5200 for their guidance feedback during the project proposal and checkpoint phases. We specifically thank Yicheng Li, Alok Shah, and Lyle Ungar for their support throughout the entire semester. Additionally, we acknowledge the authors of the APEX study [7] and the maintainers of the DBAASP database [6] for making their datasets and methodologies publicly available, which served as the foundation for this work. Finally, we thank the developers of ColabFold [3] and the modlAMP python package [4] for providing the open-source tools necessary for our feature engineering and structural analysis.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[2] Yue Ma, Zhengyan Guo, Binbin Xia, Yuwei Zhang, Xiaolin Liu, Ying Yu, Na Tang, Xiaomei Tong, Min Wang, Xin Ye, Jie Feng, Yihua Chen, and Jun Wang. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nature Biotechnology*, 40(6):921–931, 2022.

[3] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nat Methods*, 19(6):679–682, May 2022.

[4] Alex T Müller, Gisela Gabernet, Jan A Hiss, and Gisbert Schneider. modlAMP: Python for antimicrobial peptides. *Bioinformatics*, 33(17):2753–2755, September 2017.

[5] World Health Organization. New report calls for urgent action to avert antimicrobial resistance crisis. 2019.

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

[7] Fangping Wan, Marcelo D T Torres, Jacqueline Peng, and Cesar de la Fuente-Nunez. Deep-learning-enabled antibiotic discovery through molecular de-extinction. *Nature Biomedical Engineering*, 8(7):854–871, July 2024.