DreamWalker: Hallucination-guided Antimicrobial Peptides Generator

Jiazheng Miao
Harvard Medical School
Harvard University
Boston, MA 02115
jiazheng_miao@hms.harvard.edu

Siying Yang
Harvard Medical School
Harvard University
Boston, MA 02115
avon_yang@hms.harvard.edu

Adele Collin
Harvard Medical School
Harvard University
Boston, MA 02115
adele_collin@hms.harvard.edu

1 Abstract

In the field of antimicrobial research, the design and optimization of antimicrobial peptides (AMPs) present a significant challenge, stemming from the complex nature of biological interactions and the vast potential sequence space. Traditional methods are often labor-intensive and time-consuming, emphasizing the need for innovative approaches. The current state of the art in AMP design involves a combination of experimental methods and computational modeling, with a growing emphasis on machine learning techniques for sequence prediction and optimization.

Our approach introduces a novel approach for the design of AMPs using a groundbreaking deep learning model called DreamWalker to generate AMPs through a unique hallucination-guided mechanism. The model consists of 2 components, the Oracle and DreamWalker itself. Initially, the Oracle, trained on peptide and target bacterial species data, predicts the Minimal Inhibitory Concentration (MIC) for given peptides. Subsequently, the DreamWalker model, pre-trained on a diverse dataset of small peptides including both natural and artificial amino acids, utilizes these predictions to generate new peptides. The innovation lies in the DreamWalker's ability to create hallucinated inputs to deceive the Oracle, thereby enhancing its peptide generation capability. This iterative process is underpinned by the Oracle's feedback on the hallucinated peptides' MIC values, which serves as the loss function for DreamWalker's training. The aim is to converge towards peptides with enhanced antimicrobial properties at each iteration.

The results of this study are remarkable and promising. After several epochs of fine-tuning, the model successfully generated new AMP sequences that demonstrated significantly enhanced antimicrobial property compared to the original AMP0. Notably, the model's ability to rapidly iterate and evolve sequences far surpasses traditional methods, both in speed and in the breadth of sequence exploration. Also, the sequences generated by this model are evolutionarily distinct from the original AMP0, showcasing the model's ability to traverse uncharted sequence spaces and contribute novel designs to the field of antimicrobial peptides.

The broader impact of this research is substantial. The developed model represents a significant advancement in bioactive sequence engineering, offering a powerful tool for rapid and efficient AMP optimization. This method holds promise for accelerating the discovery of new antimicrobial agents, a critical need given the rising threat of antibiotic resistance. Furthermore, the model's versatility suggests potential applications in other areas of peptide-based therapeutics, opening new avenues for computational drug design. Ultimately, this study not only demonstrates a successful application of hallucination-guided models in biotechnology but also represents future innovations in computational peptide engineering.

2 Background

Since the invention of antimicrobials, the bacterial resistance to them has been constantly selected. Nowadays, antimicrobial resistance has been a major global health issue. Therefore, it became urgent for scientists to race with bacteria to find novel antimicrobials [6]. Among the chemical space, AMR is a group of less studied compounds, to which the bacterial resistance is still weak. Previous studies have applied a generative-predictive framework to small-molecular drug design and achieved considerable success [4]. However, the application of such generative models is still limited in industry, partially due to the difficulty in synthesizing the predicted molecules as validation [16]. Under these circumstances, designing therapeutic peptides seems to be more suitable for this framework. Compared with small-molecule compounds, therapeutic peptides not only enjoy the advantage of higher specificity and target affinity, but also could be chemically synthesized in an automated process [9]. This property allows the generated peptide sequence to be validated quickly.

Traditional approaches to antimicrobial peptide (AMP) prediction often fall short in addressing the specificity required for determining the antimicrobial characteristics of peptides. This is exemplified by discriminative models such as AMP0 [6], a targeted antimicrobial activity predictor developed through zero and few shot learning. AMP0 overcomes the limitations of conventional machine learning techniques, demonstrating superior performance in cross-validation analyses and ease of integration into experimental discovery for novel species [6]. In tandem with discriminative models like AMP0, an attentive deep learning model for AMP prediction, AMPlify has been developped [10]. AMPlify identifies four novel AMPs from the bullfrog genome, showcasing promising antibacterial activity against multidrug-resistant (MDR) World Health Organization priority pathogens. This underscores the value of attention mechanisms and a novel ensemble approach in mining genome resources for AMP discovery. It is to be noted that Gull and Minhas [6] made available their dataset of multi labelled AMP which is the one we will use for the training and testing of our model.

Complementing these discriminative models, the research community has explored optimization-based, generative methods for biological sequence design [8]. Motivated by global health challenges, this study introduces a generative active learning algorithm based on GFlowNets for sequence design. The innovative approach combines principles from Bayesian optimization and generative modeling to produce diverse and novel sets of candidates. GFlowNets, as the candidate generator, plays a crucial role in this process. This generative method serves as a foundation for designing antimicrobial peptides using a combination of generative and discriminative models. Another approach is to incorporate a reinforcement learning and a PPO discriminative model for enhanced efficiency in AMP design [1]. To refine and optimize the proposed model, considerations can be drawn from various sources. Firstly, training efficiency can be enhanced using Bayesian Optimization [15]. This approach addresses the need for more efficient training in high-dimensional search spaces. Additionally, insights from the diffusion model [5] offer valuable perspectives on further refining the proposed AMP design approach. The exploration of these refinements aims to improve the overall performance and efficacy of the computational model.

Recent breakthroughs in protein structure prediction, specifically in elucidating inter-residue distances from amino acid sequences [2], have paved the way for innovative approaches to *de novo* protein design. A notable method in this realm is deep network hallucination which involves inverting a neural network to design new proteins based on unrelated sequences, resulting in diverse proteins with unique sequences and predicted structures. Wicky et al. [17] further emphasize the potential of deep network hallucination in exploring protein structure space. By using this technique, the study generates a range of symmetric protein homo-oligomers, showcasing the design of complex components for nanomachines and biomaterials.

However, these optimization-based methods have two common drawbacks. First, these methods only optimize towards one target species (e.g., *Escherichia coli*) at a time, but in the real clinic setting, the pathogens could be diverse. Second, these methods generate new peptides through iteratively optimize a random peptide. This strategy could be time-consuming and computationally intensive. Fortunately, the deep network hallucination method gave us inspiration. Now that the output of the Oracle could be used to guide search algorithms, it is possible that the Oracle's output could also be used as the loss for training another neural network. With the guidance of the Oracle, the second model could explore the space of peptide sequences and learn the projection

between bacterial species to its optimal inhibitory peptides. In other words, a neural network could directly explore the landscape of hallucination. This approach may make iteration unnecessary and significantly reduce running time. Therefore, this study will first utilize the multi-species AMP data from [6] (AMP0) to train an Oracle model, and then train a generator to explore the hallucination of the Oracle.

3 Results and Discussion

This study uses deep network hallucination to train the DreamWalker, the peptide generator model. Specifically, we first trained an Oracle model which accepts the peptides and target bacterial species as input. The Oracle model predicts the Minimal Inhibitory Concentration (MIC) of the peptides to the target species. Next, we pre-trained a peptide generator with small peptide data. Finally, we transferred the weights of the peptide generator to the DreamWalker model. The DreamWalker takes bacteria species as input, and generates peptides for inhibiting their growth. The peptides are evaluated by the Oracle, and the predicted MIC values are returned to the DreamWalker as the loss. During this process, the DreamWalker is learning to generate fake input (hallucinations) to deceive the Oracle into believing it see a good peptide. In other words, the DreamWalker explores the landscape of the Oracle's hallucinations. This is why we name the generator as the DreamWalker. This process of optimizing the hallucination is shown in Figure 1 (A) & (B). The codes are available at https://github.com/AvonYangXX1/AMPLify-Feedback.git

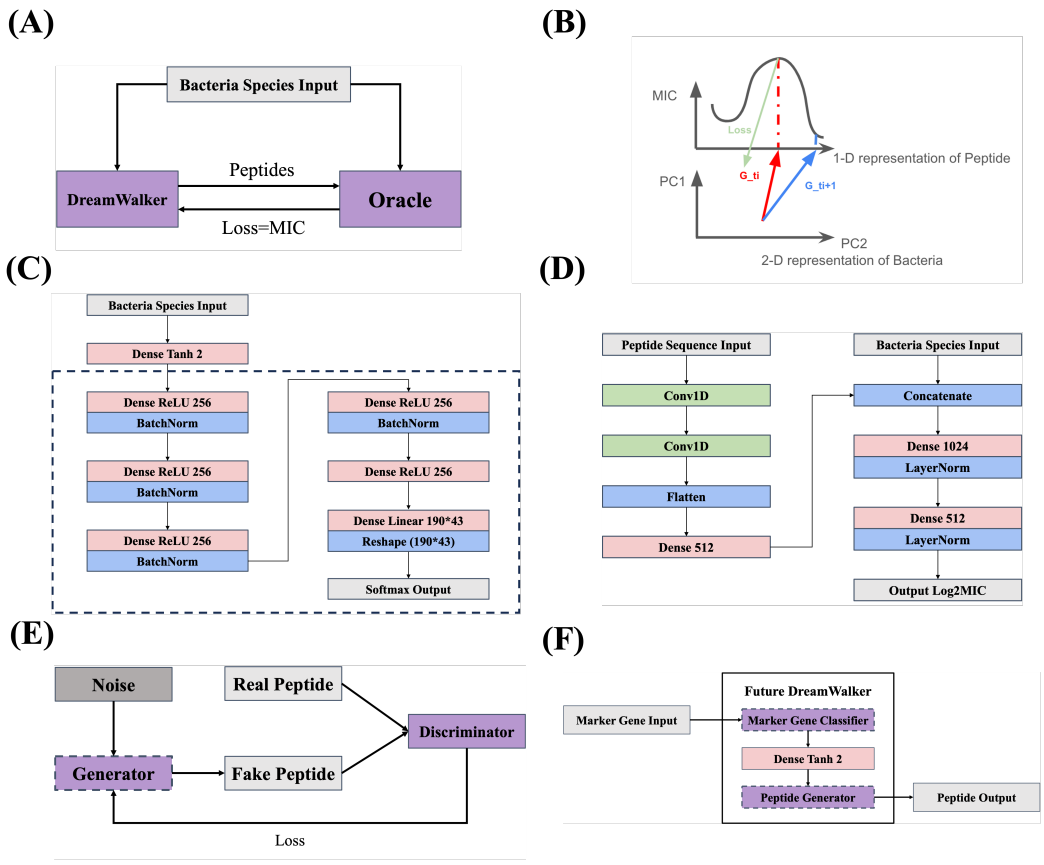

Figure 1: Framework of this study. (A) The process of fine-tuning the DreamWalker. (B) The mathematical nature of the DreamWalker: Learn to project bacterial input to peptides with low MIC. (C) The architecture of the DreamWalker model. (D) Architecture of the Oracle Model for predicting the MIC of peptides. (E) The framework of GAN for pre-training the generator. (F) Proposal of future version DreamWalker. Dashed-line boxes indicate pre-trained networks.

3.1 MIC Oracle

3.1.1 Challenges

The major challenge encountered in this part is the unavailability of pre-trained models, caused by multiple factors. First, the source of our AMP dataset, the AMP0 study [6], did not release their codes, model weights, or model architecture. This largely hindered us from reproducing the study. Also, the characteristics of the AMP0 dataset harshened the situation. The dataset contains not only natural left-handed amino acids, but also 20 artificial right-handed amino acids. Peptides containing right-handed amino acids account for 56.7% of the data. This makes it hard to transfer pre-trained protein language models (PLM) since most PLMs are only trained with natural amino acids. Specifically, we tried to fine tune PLMs such as ProteinBERT [3] and RITA small [7] with the AMP0 dataset, but the results were unsatisfactory. Meanwhile, limited by computing resources, it is also unfeasible to utilize larger PLMs.

3.1.2 Architecture

The challenges discussed previously led us to explore building the Oracle model from scratch with simpler model architectures, including GRUs, LSTMs, and CNNs. Our experiments culminated in the innovative approach of treating peptide sequences as analogous to images, allowing us to employ CNNs for feature extraction. This strategy proved to be a turning point, yielding a model that struck a balance between efficient training and acceptable performance. In the end, the Oracle model takes two input: one is peptide sequences, the other is the bacterial species that the peptide is countering. Oracle will use CNN part to extract features in peptide sequences, concatenate these information with bacterial species, and feed the information throught fully-connected layers. The output is the Minimal Inhibitory Concentration (MIC) in the log2 scale. The architecture is shown in Figure 1 (D).

3.1.3 Performance

In this study, we innovatively treat MIC prediction as a regression task. This could allow the Oracle model to give a continuous value, instead of discrete, to the Generator model as the feedback. Therefore, we used Mean Absolute Error (MAE) as the loss, and R^2 as the metric. A customized metric, Fraction of Useful Predictions, is also introduced, where a prediction within the range of [log2(True) - 1, log2(True) + 1] could be considered as clinically useful. Figure 2 (B) shows the predictive results of the Oracle model on the testing set. There is a correlation between predicted and true values, but there is also a considerable spread indicating variance in the accuracy of the predictions. The variance may be improved in the future.

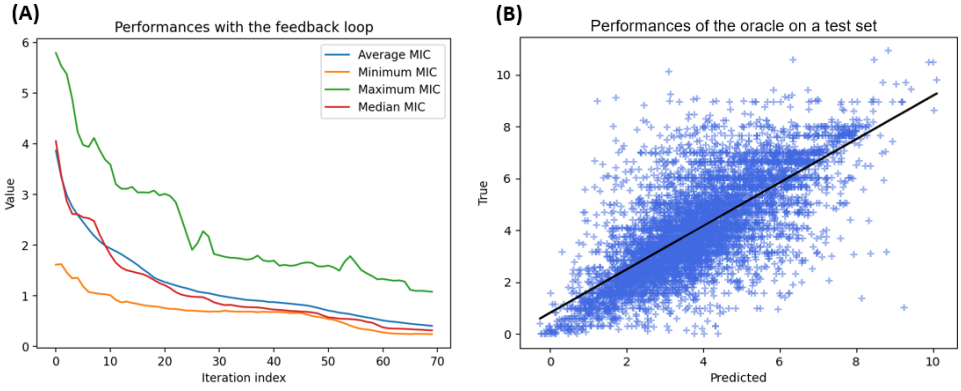

Figure 2: (A) Evolution of the MIC of the predicted sequences along the training of the DreamWalker for the specific target. The mininum, maximum, average and median are steadily decreasing over the course of the training, (B) Performances of the oracle used to compute the MIC. The accuracy of the model is about 0.65 on the validation test.

3.2 Generator Pre-training

3.2.1 Challenges

This part met similar challenge with the MIC Oracle part. Due to the right-handed amino acids, pre-trained PLMs do not work ideally. For example, in our attempt to fine-tune RITA small [7], an autoregressive generative models of 80M parameters, the accuracy of predicting the next token is highly unsatisfactory (about 11%).

3.2.2 Architecture

After multiple attempts, we found that the GAN architecture worked well in this scenario, which produces meaningful peptides and is relatively easy to train. By pretraining the Generative Adversarial Networks (GAN) to generate the new AMP sequence, the GAN architecture consists of 2 main components: the generator and the discriminator. The generator is used to create the new AMP sequences, while the discriminator evaluated whether the input sequences are real or generated. The small peptide datasets in AMP0 and AMPlify [6, 8] were combined for training the discriminator. In total, 133,584 small peptides were contained in the dataset, including 128,445 non-AMP and 5,139 AMP. To generate novel peptides, the generator would learn to project points from the input latent space to the space of real peptide.

Generator Architecture:
As shown in Fig 1 (c) above, the generator in our GAN framework is designed using a deep learning architecture, mainly composed of dense neural networks. The architecture is given a noise as the input and starts with a latent input layer that captures the diverse possibilities in AMP characteristics within a compressed latent space. This layer sets the foundation for generating peptide sequences. Subsequently, the architecture involves several dense layers, each followed by batch normalization. These layers utilize the ReLU activation function to facilitate the transformation of the latent input into a meaningful sequence format. The final output of the generator is data reshaped into the dimensions of AMP sequences, each represented by a 43-dimensional vector. This design enables the generator to learn and replicate the complex relationships inherent in AMP sequences and their corresponding features.

Discriminator Architecture:
Contrastingly, the discriminator in our GAN model is a convolutional neural network primarily focusing on differentiating between real and generated AMP sequences. It features a Conv1D layer followed by a flattening operation. The network then processes the data through two dense layers with 512 and 256 neurons, respectively, incorporating ReLU activation to maintain gradient flow and mitigate potential issues of vanishing gradients. A dropout layer with a rate of 0.3 is included to prevent overfitting. The architecture concludes with a sigmoid-activated output layer, which classifies the input as real or fake. This configuration, focusing on convolutional and fully connected layers, is adept at evaluating and distinguishing genuine AMP sequences from those generated by the model, a critical function in the discriminator's role within the GAN framework.

3.3 DreamWalker Performance

3.3.1 Challenges

The first challenge for the DreamWalker model is to efficiently integrate the oracle and the generator in a training loop. Efficiency is measured as a significant decrease of the MIC of the sequences generated estimated by the oracle and let the model converge to a few solutions. The second challenge is to develop two versions of the model, one that is target-specific and converges to a solution given a specific bacteria specie, and one that would scan all the given species to determine possible solutions for the whole bacteria space. As a matter of facts, we want our models to provide sequences both in cases where there is a need for a targeted or action or a broad-spectrum one. On the one hand, the model provides specific AMP capable of targeting a selected bacteria species even at a very low concentration. On the other hand the model can also design multi specific AMP with a broad spectrum of action against multiple bacteria sub types.

Meeting the first requirement has been challenging and we has to adapt our initial architec-

ture several times. Initially, we implemented a reinforcement loop which iteratively generates sequences and selects sequences below a MIC threshold to fine-tune the DreamWalker. However, this method proved inefficient, as evidenced by the constant MIC scores. We eventually solved this challenge by implementing a hallucination approach that uses the predicted MIC directly as the loss.

3.3.2 Generated outputs

Recognizing the limitations of the previous approaches, we shifted to a gradient descent approach, directly passing the predicted MIC as the loss for the DreamWalker. In the end, the DreamWalker loop operates as follow. First, a noise space is computed and inputed in the generator which outputs a sequence space. Then this sequence space is fed into the oracle which outputs a MIC space such as the one shown in Figure 3 (A). This MIC landscape is then used as a loss function to compute the gradients with respect to the noise input. Figure **??** displays the results of the DreamWalker model after 70 fine-tuning gradient-descent iterations. The sequence with the lowest MIC has a predicted score more than 8 times smaller than the sequence with the lowest MIC before the fine-tuning. Therefore the loop has a very positive effect on the generation of better performing sequences targeted against this specific pathogen.

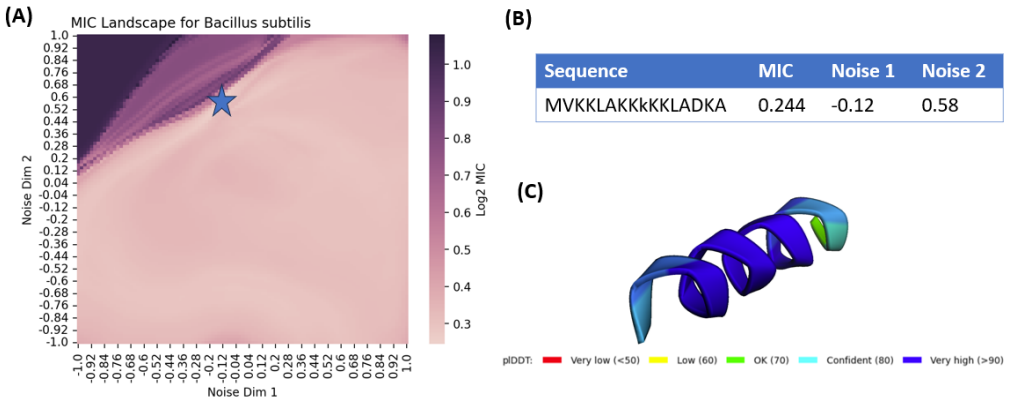

Figure 3: (A) Heatmap representing the MIC score output according to the 2D noise input after 70 fine tuning iterations for the target Bacilus subtilis. The red star represent the portion of the landscape with the lowest MIC from where the best performing sequence can be extracted. (B) Details about the sequence generated with the lowest MIC at the 70th iteration. The predicted MIC is 0.244, 8 times lower than the best performing sequence at the first iteration (C) 3D structure of the peptide generated using AlphaFold2 [11, 12, 13, 14]

After implementing gradient descent, we observed a progressive decrease in the MIC scores over subsequent iterations as shown in 2. This not only demonstrated the efficacy of the new selection method but also highlighted the dynamic and adaptive nature of our model in optimizing AMP sequences.

Finally, we extended this gradient descent with MIC approach to the computation of sequences that would be active against several species. To do so, we added one layer that projects the the bacteria into the noise space. This allows for the model to automatically search the noise space for several species simultaneously. We tested this approach with the 326 species of bacteria available in our database and get the following results for 5 of them:

These results show the ability of the loop to generate sequences for multi-species problems. Despite the variability between sequence generated, common patterns can be seen and the overall performances can be easily enhanced by a training on more iterations. It is also to be noted that the test of this multi-species loop has been done in the most unfavorable case, with more than 320 species involved. A higher convergence is to be expected for a training with fewer species and over more iterations.

Species	Peptide	Log2 MIC
Bacillus subtilis	GVlKLLKKFLKFAKI	1.792013
Staphylococcus aureus	TVyLLLKKFKKFAKDAI	2.636046
Escherichia coli	TVyLkNKKkKKFAKDAI	0.757396
Pseudomonas aeruginosa	TVKKLFIFLL	3.09396
Candida albicans	TVlKLFKFFKLFAKI	4.174263

Table 1: Multi-species AMP prediction results for 5 species among the 326 used for the test. The average MIC over this test decreased from 4.2 to 3 in 20 iterations, once again showing the importance of the loop in the optimization of the MIC of the generated sequences.

3.3.3 Measurement of Evolutionary Divergence

To evaluate the generated new sequence by the model with the original dataset, we use the BLOSUM62 matrix score to evaluate the evolutionary divergence of the peptide sequence generated by the DreamWalker model and those in the AMP0 dataset. The heatmap generated is shown in Figure 4. Notably, 87.6% of the generated sequences have a negative BLOSUM62 score, most of them range between -20 to -60. The predominated negative scores imply that the generated peptide sequences are evolutionarily distinct from those in the existing dataset, suggesting the model's capability to innovate beyond the current sequence space. The positive BLOSM62 scores would indicate the evolutionary relatedeness or the functional similarity due to conservation of amino acid pairings. The negative scores highlight that our model can generate novel and unexplored peptide sequences.

The creation of the distinct sequences from original dataset AMP0 is crucial. It definitely paves for the potential of our model in discovering the new antimicrobial peptide with unique structures and mechanisms yet have enhanced antimicrobial property. The need for the empirical validation may be the next step. It should be noted that the BLOSUM62 score is a robust tool for the sequence comparison, but it only represents one aspect of evolutionary analysis. The functional potential may not be fully captured. Phylogenetic analysis may be included in the next step to further predict the efficacy of the evolutionarily divergent peptides in antimirobial research.

4 Future Goals

Currently, the target species of the DreamWalker are limited to those in the AMP0 dataset. In the future, we hope to break this boundary of species. We would like to add a phylogenetic classifier on microbial marker genes (e.g., 16S rRNA gene) to both the Oracle and the DreamWalker, as shown in Figure 1 (E). This may allow the models to learn patterns among the phylogeny (for example, one certain peptide may be effective towards all bacteria in one family), and generalize the knowledge to species absent in the training data. Generally speaking, further evaluation of DreamWalker performances needs to be conducted to explore the full potential of this implementation.

5 Comparison with Original Proposal

Overall, the aim of this report comparing to what we proposed on the original proposal is largely congruent. Our aim is still to innovate the design of antimicrobial peptides (AMPs), and this has been achieved. However, the proposed dynamic feedback mechanism between the Generative Adversarial Network (GAN) and Convolutional Neural Network (CNN) models and the reinforcement learning method is abandoned given the bad performance. We pivoted towards a more innovative approach: the utilization of deep network hallucination to train the DreamWalker, our novel peptide generator model.

Throughout the project, we addressed most of the reviewers' concerns, specifically focusing on the robustness of our predictive model and ensuring the diversity of the generated peptide sequences to avoid overfitting and mode collapse.

Differences from the original proposal is our specific design and model architecture we use.

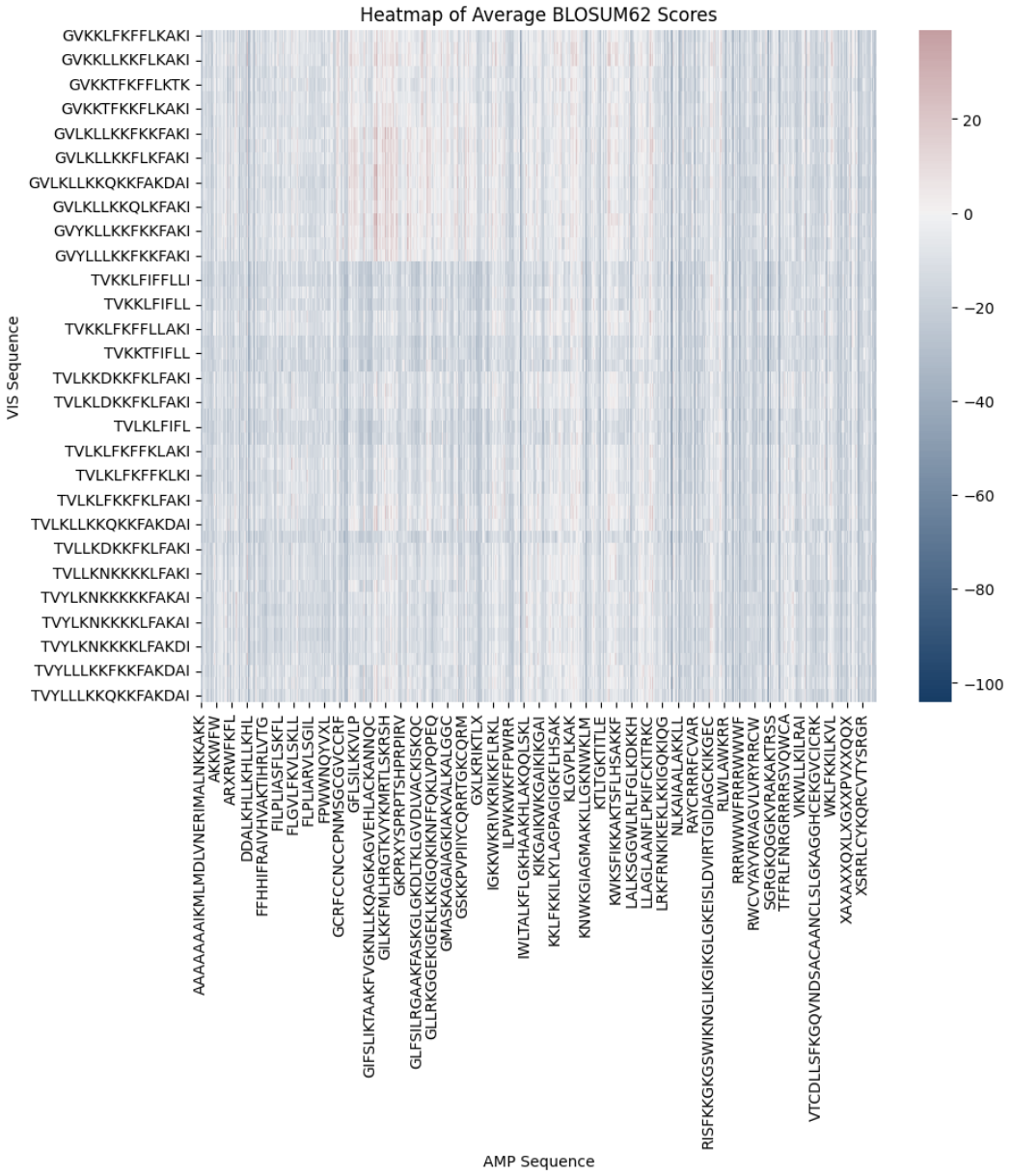

Figure 4: Heatmap of Average BLOSUM62 Scores

For instance, while we initially planned to use multi-label AMP sequence data, the project's evolution led us to focus more on the hallucination guide architecture and using loss from the MIC Oracle in refining AMP designs. We aim to break the target species boundary and develop a universal AMP that learns the pattern among the phylogeny and generalize the species outside of the original AMP0 dataset.

6 Commentary on your experience

Going back and rethinking our project, we believe that we follow the time chart for the project pretty well, and were able to achieve a good and efficient time management in both the practical and theoretical work. Despite this, we recognize that the duration allotted for experimental testing exceeded our initial estimates. To optimize our schedule in future endeavors, we could benefit from

initiating the experimental and coding phases earlier in the timeline.

In the end we are able to produce the results that align with our stated goal. The methods did change after we saw the results from the initial method do not perform very well. We ask advice from our mentor, Jason Yim and Jeremy. Both did offer us very great thoughts and innovating ideas. Moreover, we were very inspired by the talk given by Dr Anishchenko about novo hallucination design of protein structure. Our overall experience is great. We work hard, and we do get support and advice when we are struggling.

The most challenging aspect of the project was to decide which model we should select for the Generator.We definitely have a lot of choices ranging from the protein language model to GAN and GAIL. One of our main struggle during this part and the entire project has been the limited computing resource we had access to which obliged us to be especially rigorous and creative. Regarding the GAN, since we connect two complex neural networks together, it has been hard to finetune when the final results dis not go as expected. The final result is definitely very rewarding as we are able to generate the new sequences that have the antimicrobial activity. We are able to prove the sequence is different from the original dataset based on the BLOSUM62 heatmap. This achievement not only validates the efficacy of our methods but also underscores the potential of our approach in contributing valuable insights to the field of antimicrobial research.

7 Division of labor

All of us contributed to this project with best efforts and dedication. Jiazheng worked on the topic selection, the Oracle Model predictor and the evaluation. His work was fundamental in setting the direction and ensuring the project's relevance and effectiveness. Siying worked on the generator, discriminator architecture, and the evolutionary divergence. Her technical skills in these areas was critical for the development of the project's core components. Adele worked on the hallucination mapping, the model loop connecting GAN and Oracle and the performance visualization results in the end. Her ability to interconnect different parts of the project was essential for its cohesive development.

The collaboration process is very inspiring and educational. We learn from each other, give support and discuss when the ideas don't work. This collaborative environment fostered innovation and problem-solving. One challenge that arouse might have been coordinating different parts of the project, ensuring that each component aligns well with others, and maintaining consistent communication.

If starting over, the advice we would give ourselves maybe conduct more literature review in the beginning and ask more people from diverse backgrounds to give us ideas and evaluate our model. We change our direction several times, and are inspired by different guest speakers from lectures and eventually get the model performance to perform better and archieve a desirable result in the end.

References

[1] Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy J. Colwell. Model-based reinforcement learning for biological sequence design. In *International Conference on Learning Representations*, 2020.

[2] Ivan Anishchenko, Samuel J. Pellock, and Tamuka M. et al. Chidyausiku. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, Dec 2021.

[3] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 02 2022.

[4] Piotr S. Gromski, Alon B. Henson, Jarosław M. Granda, and Leroy Cronin. How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry*, 3(2):119–128, Feb 2019.

[5] Nate Gruver, Samuel Stanton, Nathan C. Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion, 2023.

[6] Sadaf Gull and Fayyaz Minhas. Amp0: Species-specific prediction of anti-microbial peptides using zero and few shot learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):275–283, 2022.

[7] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models, 2022.

[8] Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with gflownets, 2023.

[9] Christina Lamers. Overcoming the shortcomings of peptide-based therapeutics. *Future Drug Discovery*, 4(2):FDD75, 2022.

[10] Chenkai Li, Darcy Sutherland, S. Austin Hammond, Chen Yang, Figali Taho, Lauren Bergman, Simon Houston, René L. Warren, Titus Wong, Linda M. N. Hoang, Caroline E. Cameron, Caren C. Helbing, and Inanc Birol. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. *BMC Genomics*, 23(1):77, Jan 2022.

[11] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: Making Protein folding accessible to all. *Nature Methods*, 2022.

[12] Milot Mirdita, Martin Steinegger, and Johannes S"oding. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, 2019.

[13] Milot Mirdita, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes S"oding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, 45(D1):D170–D176, 2017.

[14] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Korobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D Finn. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, 2019.

[15] Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating bayesian optimization for biological sequence design with denoising autoencoders, 2022.

[16] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020.

[17] B.I.M. Wicky, L.F. Milles, and A. et al. Courbet. Hallucinating symmetric protein assemblies. *Science*, 378(6615):56–61, Oct 2022.