**Naval Research Laboratory**
Washington, DC 20375-5320

**NRL/6910/MR—2021/2**

# Variational Autoencoder for the Generation of New Antimicrobial Peptides

SCOTT N. DEAN

*Laboratory for Bio/Nano Science and Technology Branch*
*Center for Bio/Molecular Science & Engineering*

April 8, 2021

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From - To)* |
|---|---|---|
| 08-04-2021 | NRL Memorandum Report | 09/03/2019 – 09/02/2020 |

**4. TITLE AND SUBTITLE**

Variational Autoencoder for the Generation of New Antimicrobial Peptides

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
NISE

**6. AUTHOR(S)**

Scott N. Dean

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**
N2U6

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

**8. PERFORMING ORGANIZATION REPORT NUMBER**

NRL/6910/MR--2021/2

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

**10. SPONSOR / MONITOR'S ACRONYM(S)**

NRL NISE

**11. SPONSOR / MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

**DISTRIBUTION STATEMENT A:** Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**

Karle Fellowship

**14. ABSTRACT**

New techniques for antimicrobial peptide (AMP) discovery are necessary for overcoming pathogenic bacteria in the post-antibiotic era. AMPs have widely arisen via natural evolution from complex communities of competing organisms, making them promising targets for antimicrobial development; unfortunately, their identification, characterization, and production of AMPs can be complex and time consuming. This report details the development of a peptide generation framework based around variational autoencoder (VAE) and antimicrobial activity prediction models for designing novel AMPs using minimal data inputs (sequences and experimental minimum inhibitory concentration (MIC)). By sampling from different, select regions of the latent space enables controlled production of new promising AMP sequences with desirable properties. Extensive analysis of the sequences and experimental validation showed this design framework as a promising system for development of novel AMPs.

**15. SUBJECT TERMS**

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | Scott N. Dean |
| UU | UU | UU | UU | 21 | **19b. TELEPHONE NUMBER** *(include area code)* (703) 587-4188 |

This page intentionally left blank.

## CONTENTS

## FIGURES



## TABLES

# EXECUTIVE SUMMARY

This report presents research conducted by Scott N. Dean for the NRL Karles Fellowship (September 3, 2019 – September 2, 2020). New techniques for antimicrobial peptide (AMP) discovery are necessary for overcoming pathogenic bacteria in the post-antibiotic era. AMPs have widely arisen via natural evolution from complex communities of competing organisms, making them promising targets for antimicrobial development; unfortunately, their identification, characterization, and production of AMPs can be complex and time consuming. This report details the development of a peptide generation framework based around variational autoencoder (VAE) and antimicrobial activity prediction models for designing novel AMPs using minimal data inputs (sequences and experimental minimum inhibitory concentration (MIC)). By sampling from different, select regions of the latent space enables controlled production of new promising AMP sequences with desirable properties. Extensive analysis of the sequences and experimental validation showed this design framework as a promising system for development of novel AMPs.

This page intentionally left blank.

# VARIATIONAL AUTOENCODER FOR THE GENERATION OF NEW ANTIMICROBIAL PEPTIDES

## 1. INTRODUCTION

### 1.1 Antimicrobial discovery and antimicrobial peptide design

Generation of new antimicrobials is critical for survival in the post-antibiotic era [1]. At the current rate, annual global death due to antibiotic resistance is projected to exceed 10 million by 2050, costing 100 trillion USD [2]. To lower the burden required for antimicrobial development, various schemes for their discovery and refinement have been proposed. Recent work has shown that generative deep learning techniques can be applied to this problem: using a generative model, Stokes et al. showed that by training on a starting database of compounds from ZINC15 [3] new small molecule antibiotics can be identified, such as halicin, which displays activity against *Acinetobacter baumannii* in a murine model of infection [4] and is currently in clinical trials.

Antimicrobial peptides (AMPs), essential components of the innate immune system of humans and other organisms, have retained effectiveness as antimicrobials despite their ancient origins and widespread and continual contact with pathogens [5], and thus have been regularly deemed "drugs of last resort" for their ability to kill multidrug resistant bacteria. The relative immutability of bacterial membranes and other essential AMP targets make the development of resistance to AMPs rare, but possible [6], thus increasing the importance of their reliable, continued discovery, to grow to the antimicrobial stockpile [5].

Attempts at both generating new AMPs and improving their activity have been carried out with varying degrees of success [7]. Many of these new or enhanced AMPs have been generated using low-throughput design methods. Certain high throughput computational techniques such as genetic algorithms have shown promise; however, in many applications starting sequences are directed toward canonical amphipathic alpha-helical peptides, restricting output to a small subset of possible structures and sequences [8]. In order to increase the rate of discovery of AMPs, newer high-throughput and low expertise design approaches are needed. Several recent preprints and publications have demonstrated the application of generative deep learning on the design of AMPs, using long short-term memory (LSTM) networks [9, 10], Generative Adversarial Networks (GANs) [11], and Variational Autoencoders (VAEs) [12, 13].

Although recent results using generative deep learning for producing new sequences has shown promise, including a handful of experimental demonstrations of their activity [10, 11, 13], some improvements are necessary. Of foremost importance, as many of these systems readily generate sequences that are both predicted and experimentally found to be inactive, machine/deep learning systems would benefit from integrated activity prediction functions. Although AMP activity prediction applications exist, many are classifiers, predicting a binary antimicrobial vs. not [14], where regression models would be of greater value. Additionally, a general reduction in filtering steps would reduce bias in predicted AMPs, since amphipathicity and helicity are explicitly selected for by certain models [8, 10]. Thus, generation of new sequences with desired predicted activity via a semi-unsupervised and streamlined AMP-generation framework with minimal input parameters and less potential for biased output would be an improvement over previous works. In this study, we report the joining of predictive models for AMP activity (minimum inhibitory concentration (MIC)) with a generative VAE for an automated framework to produce new peptides with experimentally testable predicted activity. We demonstrate an improved automated semi-supervised approach for generating promising new sequences and experimental investigation, resulting in low MIC AMPs against *Escherichia coli*, *Staphylococcus aureus*, and *Pseudomonas aeruginosa* output from a handful of input parameters.

Manuscript approved March 29, 2021.

## 2. APPROACH

### 2.1 Dataset, analysis of sequence characteristics, and variational autoencoder

The dataset used in this study was based on Giant Repository of AMP Activity (GRAMPA) as described by Witten *et al.* [15], with some modifications. Witten et al. scraped all data from APD [16], DADP [17], DBAASP [18], DRAMP [19], and YADAMP [20], each accessed in Spring 2018, resulting in a combined 6760 unique AMP sequences and 51345 MIC values, and is publicly available on GitHub. MIC values were independently spot-checked and confirmed; however, methods vary widely between publications, suggesting MIC values herein should be interpreted as approximations of activity. Since MICs determined against *E. coli* were the most-commonly available, these were used for the study. To avoid issues with synthesis, the dataset was further modified by excluding peptides with cysteine and avoiding recorded modifications. For ease of synthesis and to keep costs low, sequences ≥ 40 amino acids in length (representing 3.1% of sequences) were excluded). Since only the sequences and MIC values were needed, all other data from the modified GRAMPA dataset was removed. All MIC values were log μM transformed as done previously [15]. The sequences were tokenized, <end> token appended to each, and represented by a one-hot encoding scheme using binary vectors with length equal to the size of the amino acid vocabulary: the stopping token <end>, a, d, e, f, g, h, i, k, l, m, n, p, q, r, s, t, v, w, y, and a padding character. This resulted in a 3D data matrix of dimension 3280, 21, and 41 for the number of sequences, length of the vocabulary, and feature vector length, respectively. This process was repeated for *S. aureus* and *P. aeruginosa* identically to *E. coli.* The final 3D data matrices for *S. aureus* and *P. aeruginosa* had 2974 and 1968 sequences, respectively, with 21 and 41 length of the vocabulary and feature vector length. The *S. aureus* and *P. aeruginosa* datasets were only used for training of the MIC prediction regression models. Secondary structure of AMPs was predicted using the PredictHEC function from the DECIPHER R package within Bioconductor, providing the probability of helix, beta sheet, or coil (H, E, or C) [21]. PredictHEC makes use of the GOR IV algorithm [22]. This method is one of the best-performing that uses only the primary sequence and doesn't require input of other sequences. Welch's t-test via NumPy [23] was used throughout for sample comparison with a significance threshold of 0.01 unless otherwise noted.

The architecture of the VAE was implemented as described by Bowman *et al.* [24], as described by Dean and Walper [13] with minor modifications. The loss function was comprised of reconstruction loss and KL loss to penalize poor reconstruction of the data by the decoder and encoder output representations of z (latent space variables) that are different from a standard normal distribution. The preprocessed data was encoded into vectors using a LSTM network. The encoder LSTM was paired with a decoder LSTM in order to do sequence-to-sequence learning. The decoder results were converted from binary one-hot encoded vectors back to peptide sequences. Training stoppage criteria was met when loss values did not decrease >0.0001 for five consecutive epochs. The VAE was trained using the Keras [25] library with a TensorFlow [26] backend, and used the Adam optimizer. The number of neurons for the LSTM layers found in both encoder and decoder were both set to 1024. The number of latent dimensions was set to 50. All models were trained on an Ubuntu workstation with a Nvidia GeForce GTX1070 GPU. The LSTM network used in the decoder-encoder are stochastic – decoding from the same point in latent space may result in a different peptide being generated and is dependent on the random seed set prior to running. Models were saved to binary files and are available upon request.

### 2.2 New sequence generation and latent space visualization

Cosine similarity was used to compare latent space vectors generated by the VAE, where each vector (b) was compared to the same reference (a): the vector representation of VLNENLLA, caseicin B-B1. Selected for its relative inactivity, caseicin B-B1 was reported to show a MIC of ≥ 1.25 mM against *E. coli* NCIMB 11843 in a study of caseicin B [27].

Using the cosine similarity values, the five nearest to the caseicin B-B1 vector (Group A) and the five furthest from caseicin B-B1 (Group B) were identified. Around each of these vectors ($v$), new vectors were sampled by selecting random points from a normal distribution. In order to accommodate the relative variation of the latent codes, we denote $w_{ij}$ as a new vector with the following equation,

$$w_{ij} = \sum_{i=1}^{10} \sum_{i=1}^{10} v_i + (\mathcal{X}_{ij} \sim \mathcal{N}(\mu = 0, \sigma = std(v_i)))$$

where $v_i$ i[th] vector of Group A or Group B
$\mathcal{X}_{ij}$ random 1x50 vector sampled from a normal distribution $\mathcal{N}$
$\mathcal{N}$ normal distribution function (with mean $\mu$ and standard deviation $\sigma$)
$std(v_i)$ standard deviation of i[th] vector of Group A or Group B.

The $\mu$ was set to 0 and $\sigma$ equal to the standard deviation of the vectors from Group A and Group B ($std(v)$). The resulting random 1x50 vector ($\mathcal{X}$) was then added to the input vector, resulting in a new vector ($w$). Using this method, 10 new vectors were sampled for Group A and Group B, and all vectors were decoded to new peptide sequences. Following removal of duplicate sequences and those already present in the modified GRAMPA dataset, 38 remained (sequences and other information available in **Table 1**).

For dimensionality reduction PCA, t-SNE, UMAP – each with two components – was used. PCA and t-SNE used were imported from Scikit-learn, while UMAP was from McInnes *et al.* [28]. For t-SNE, perplexity set to 30, and learning rate set to 100. UMAP was performed using Bray-Curtis Similarity as the metric, with default settings. The MIC thresholds for coloring were: < 0.2 log μM is shown in blue, >2 log μM was set to red, and those with values ≥ 0.2 and ≤ 2 were set to light gray. To visualize the cosine similarity values of each encoded vector in latent space, the vector for each peptide was colored according to cosine similiartity value on a 2D t-SNE projection. To highlight the locations of Group A and Group B, black and white stars were placed at the points representing the peptides caseicin B-B1 and trialysin peptide P4, respectively.

### 2.3 Regression models and experimental validation

Eight different regression models for predicting AMP MIC values: convolution neural network (CNN) as implemented by [15], Elastic Net (ENet), Gradient Boosting (GB), Kernel Ridge (KR), Lasso, and Random Forest (RF) models used were from the Python Scikit-learn library [29], while Light Gradient Boosting Machine (LGBM) and EXtreme Gradient Boosting (XGB) used lightgbm [30] and xgboost [31] libraries, respectively. The model parameters for each are provided in Supplemental Material. The data used for the regression models was the same as described in the *Dataset and analysis of sequence characteristics* section above, prior to one-hot encoding. The input peptides sequences were encoded numerically to vectors, each amino acid or padding characters – which were appended to the end vector below the maximum length (40) – receiving a unique number. The data was randomly shuffled and split into training:test sets at a 90:10 ratio. Initial comparison of the eight different regression models for predicting AMP MIC values against *E. coli* was performed by calculating RMSE and $R^2$ values from actual MIC (log μM) vs. predicted on a holdout test dataset. From these the top performing four – GB, RF, LGBM, and XGB – were examined with shuffled split cross validation in each case (n=25) for predicting the MICs in the *E. coli, S. aureus,* and *P. aeruginosa* datasets. The top-performing MIC predictor for each organism was selected by lowest median RMSE.

Minimum inhibitory concentration (MIC) measurements values were measured using broth dilution method for AMPs [32]. Peptides synthesized for use in this study are listed in **Table 1**. Peptides were synthesized by Genscript, Inc. (Piscataway, NJ, USA) and each confirmed to have greater than 80% purity. Lyophilized peptides were solubilized in water, aliquoted, and stored at -20 °C. Overnight cultures of *E. coli* K-12, *Staphylococcus aureus* ATCC 12600, and *P. aeruginosa* 27853 were grown in Mueller Hinton II Broth, (BD, San Jose, CA, USA) at 37 °C. Cultures were diluted to a final concentration of approximately 5 x $10^5$ CFU/mL into fresh broth. An inoculum volume of 100 µl was added to each well of a 96-well non-treated polystyrene plate (Celltreat Scientific Products, Pepperell, MA, USA) and 10 µL of the peptide, which was diluted in series so that final peptide concentrations examined ranged from 128 µM to 0.5 µM. After incubation at 37 °C for 24 h, the MIC was determined by $OD_{600}$ measurement using a BioTek Synergy Neo2 plate reader (Winooski, VT, USA) to identify the lowest concentration of peptide which inhibit growth. Statistical analysis of predicted and experimental MIC data was performed using the Fisher's Exact Test from stats package in R [33].

Table 1 — VAE-generated peptides

| Peptide name | Cosine similarity | Group Id | Sequence | Parent peptide ID | MIC prediction (log µM) |
|---|---|---|---|---|---|
| p1 | 1.000 | A | VLNANLLR | 3088 | 3.6 |
| p2 | 1.000 | A | VLIKTRLFIKRK | 3088 | 1.2 |
| p3 | 1.000 | A | LNWKAILKHIIK | 3088 | 1.2 |
| p4 | 1.000 | A | VLPKVMAHMK | 3088 | 2.0 |
| p5 | 1.000 | A | LNWGAVLKHVVK | 3088 | 1.9 |
| p6 | 1.000 | A | LILKRKRKRKRILI | 3088 | 1.8 |
| p7 | 0.994 | A | LNWGAIKKHIIK | 3085 | 2.0 |
| p8 | 0.994 | A | VLNENLLA | 3085 | 3.8 |
| p9 | 0.994 | A | LNWGAFLKHFFK | 3085 | 1.3 |
| p10 | 0.994 | A | VLNENLLH | 3085 | 3.9 |
| p11 | 0.993 | A | VLNENAAR | 3090 | 3.9 |
| p12 | 0.993 | A | VLNENLRR | 3090 | 3.7 |
| p13 | 0.993 | A | VLNENLLR | 3090 | 3.7 |
| p14 | 0.993 | A | VDLKNLLK | 3090 | 3.1 |
| p15 | 0.993 | A | VALNENLLR | 3090 | 3.9 |
| p16 | 0.984 | A | LRRLRLRLLRLLRRLLRLL | 3087 | 0.9 |
| p17 | 0.984 | A | VLNNLLR | 3087 | 3.4 |
| p18 | 0.984 | A | VLNENLAA | 3087 | 3.9 |
| p19 | 0.984 | A | VLNEALLR | 3087 | 3.5 |
| p20 | 0.981 | A | LNWGAWLKHWWK | 3089 | 1.3 |
| p21 | 0.981 | A | LVKRVKKVL | 3089 | 1.3 |
| p22 | 0.981 | A | VNLKNLLR | 3089 | 3.6 |
| p23 | -0.952 | B | KWKLWKKIEKWGQGIGAVLKWLTTWL | 2220 | -0.3 |
| p24 | -0.952 | B | KWKSFLKTFKSPVKTVFYTALKPISS | 2220 | 0.4 |
| p25 | -0.952 | B | KWKSFIKKLTSVLKKVVTTAKPLISS | 2220 | 0.2 |
| p26 | -0.952 | B | KWKSFIKKLTSAAKKVVTTAKPLISS | 2220 | 0.2 |
| p27 | -0.952 | B | KWKSFLKTFKSPARTVLHTALKPISS | 2220 | 0.5 |
| p28 | -0.958 | B | KWKSFIKKLTSAAKKVLTTGLPALIS | 2227 | 0.0 |
| p29 | -0.958 | B | KWKSFLKKLTSAAKKVLTTALKPISS | 2227 | 0.0 |
| p30 | -0.963 | B | KWKSFLKTFKSAVKTVLHTALKAISS | 2228 | 0.0 |

| | | | | | |
|---|---|---|---|---|---|
| p31 | -0.963 | B | FIGGLRRLFATVVGTVVGAINKLGGG | 2228 | 1.1 |
| p32 | -0.965 | B | KFFKKLKKAVKKGFKKFAKV | 1802 | 1.1 |
| p33 | -0.965 | B | FFFHIIKGLFHAGRMIHGLV | 1802 | 1.1 |
| p34 | -0.965 | B | FFFKLLPKAIGALKKI | 1802 | 1.1 |
| p35 | -0.981 | B | FKIKASKKLLKKVGKGALGAVAKALAQQA | 1809 | 0.8 |
| p36 | -0.981 | B | KWKKFIKKLTSAAKKVLTTGLPALIS | 1809 | 0.0 |
| p37 | -0.981 | B | KWKKFLKKLTSAAKKVLTTALKPISS | 1809 | 0.0 |
| p38 | -0.981 | B | FFKKFIGGVAKIAGKAAPHGVGQLIPHVTP | 1809 | 0.6 |

## 3. MODELING

### 3.1 Dataset characterization and framework design

This study makes use of the Witten and Witten GRAMPA dataset as the starting point [15]. Within this dataset, amino acid sequence and MIC values for peptides targeting several common bacterial species including *E. coli*, *S. aureus*, and *P. aeruginosa* are reported with *E. coli* being the most counted at 9150 different entries. After filtering the dataset on bacteria species, peptides that were of length ≥ 40 or contain cysteine were removed in order to avoid costly and difficult synthesis of long peptides, as well as the complications cysteine-containing peptide create for their production, activity testing, and aggregation. The peptide length distribution of AMPs (without cysteine) tested against *E. coli* had a median of 17 amino acids prior to filtering on length, and median of 16 amino acids following removal of long sequences. Of the remaining peptides (N = 3280), the median log μM MIC value was found to be 1.19 with a net charge of +4. Finally, to obtain a glimpse of possible secondary structures of the AMPs in this dataset, we calculated the hydrophobic moments at different angles. The noticeably higher hydrophobic moment at 100 degrees suggests helical secondary structure likely predominates, with a relatively minor proportion of beta sheet and random coil comprising the remainder.

For the next two most common species found in the dataset following *E. coli, S. aureus* and *P. aeruginosa,* we performed the same characterization as described above. Although the overall counts were lower for the *S. aureus* and *P. aeruginosa* AMP datasets at 2974 and 1968, respectively, both the distributions and median values for length, MIC, charge, and hydrophobic moment were found to be similar to those found for *E. coli*.

Using the above defined dataset for *E. coli*, we designed a VAE AMP generation pipeline (**Fig. 1**). Broadly, the VAE AMP generation and design process occurs in two stages: (1) algorithm training and (2) AMP evaluation. In the first stage, the VAE is trained on a curated AMP dataset followed by development of a regression model for activity prediction and the subsequent development of the latent space. Stage two includes the identification of new AMP sequences from the latent space (sampling) and the subsequent production and characterization of the AMPs including determination of the MIC values. A VAE implemented as previously described [13], making use of VAE described by [24], was trained on the *E.coli* dataset as described above. The number of intermediate dimensions was set to 1024 and latent dimensions was set to 50. Training was stopped after 500 epochs or when loss decreased at a sufficiently low rate. The final state of the model was saved and used for sampling novel sequences. A more detailed description of the framework design is provided in the Methods section. As demonstrated in previous work, the implicit starting assumption was that sequence order, or "peptide grammar," and characteristics dependent on that sequence were the components "learned" by the VAE. Output of the VAE was a 50-dimensional latent space where each of the sequences is encoded to a unique location. Once generated,

coordinates can be chosen from the latent space and translated to novel AMP sequences using the generated decoder (see diagram in **Fig. 1**).

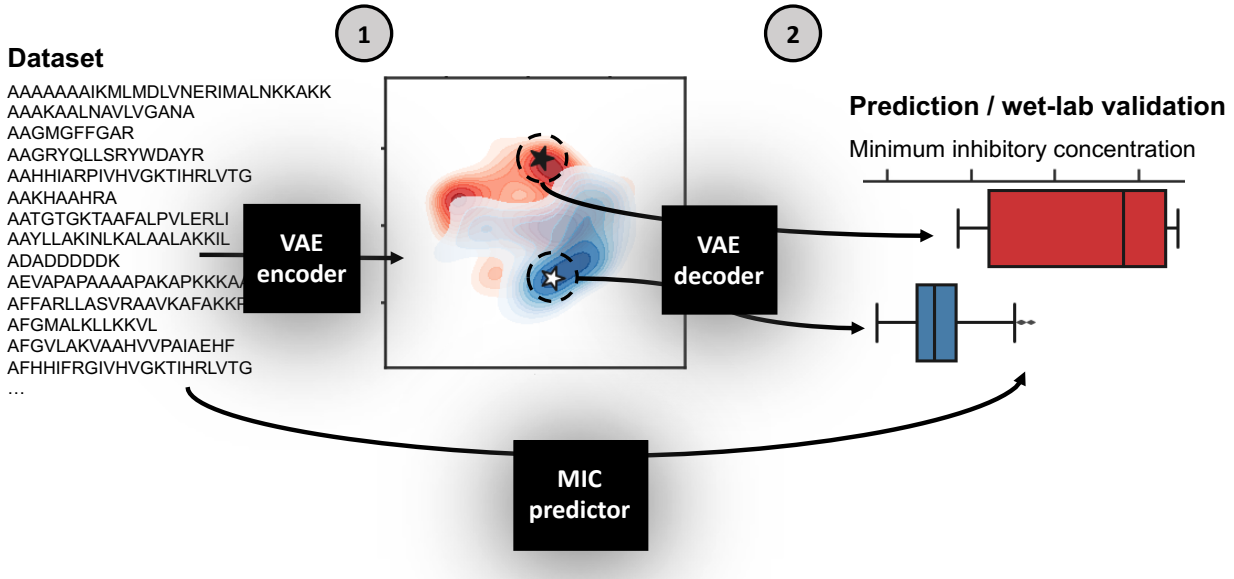


Fig. 1 – Schematic of VAE AMP generation and design process. The VAE AMP generation and design process occurs in two stages: (1) training the VAE for the development of the latent space and a regression model for activity prediction, and (2) sampling from the latent space, generation of new AMPs, and assignment of predicted MIC values. The first step of Stage 1 was to train the VAE on the *E. coli* dataset. The general design of the VAE was previously described [13], making use of VAE described by [24] which was reported for use in generating new sentences. Here, the number of intermediate dimensions was set to 1024 and latent dimensions was set to 50. Training was stopped after 500 epochs or when loss decreased at a sufficiently low rate. The final state of the model was saved; the encoder is used in Stage 1 and the decoder is used in Stage 2. The MIC prediction regression model is similarly trained on the same dataset and used in Stage 2 following sequence generation by the VAE decoder to assign MIC values for those new AMPs against *E. coli*. A more detailed description of the framework design is provided in the Approach section.

### 3.2 VAE, latent space visualization, and sampling

In order to visualize the organization of the developed 50-dimension latent space, multiple dimensionality reduction techniques were tested: principal component analysis (PCA), T-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). Upon visual inspection, t-SNE and UMAP show separation between the distant MIC thresholds of < 0.2 log μM and >2 log μM, while separation between the two groups in the first two components of the PCA is less clear; this is supported using Adjusted Rand Index (ARI) measurement and Adjusted Mutual Information (AMI) scores. Here, 2D projects (via PCA, t-SNE, and UMAP) of the latent representation was used as input to the *K*-means algorithm and measure the overlap between the resulting clustering annotations and the pre-specified subpopulations (the < 0.2 log μM and >2 log μM labels) using the Rand index and AMI scores. By ARI, t-SNE shows the highest separation measure with 0.62, with UMAP at 0.58, and PCA at 0.47. Using AMI score, t-SNE is also highest at 0.59, t-SNE at 0.56, and PCA at 0.47. These results suggests that 1) the AMPs encoded to the latent space are not randomly distributed in terms of their MIC value classification, and 2) t-SNE provides superior 2D clustering visualization in this application, relative to PCA and to a lesser extent UMAP. t-SNE projections are shown in **Fig. 2A-B.**

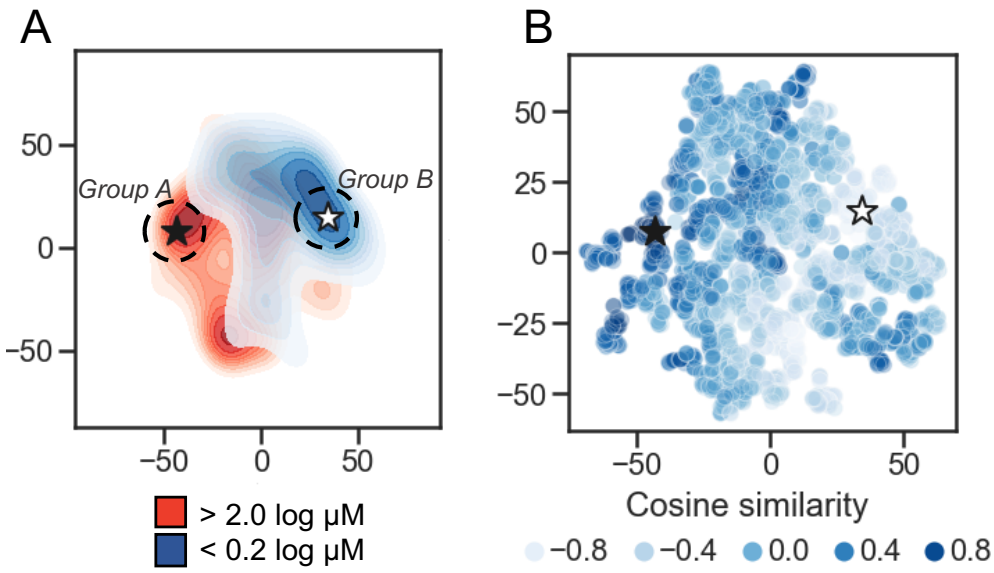


Fig. 2 — Latent space characterization. Dimensionality reduction for visualization of the 50-dimension latent space. A) 2D contour plot of t-distributed stochastic neighbor embedding (t-SNE) with two components performed on the encoded peptides. The MIC thresholds for coloring were: < 0.2 log μM is shown in blue, >2 log μM shown in red, and those with values ≥ 0.2 and ≤ 2 were set to light gray. Regions of higher density are darker. The black star is located at the peptide caseicin B-B1, and a white star is located at the embedding most distant in cosine similarity (encoding for the peptide trialysin peptide P4). Group A and Group B sampling locations are schematically shown at the dashed circle around caseicin B-B1 and trialysin peptide P4, respectively. B) Scatterplot of t-SNE projection showing encoded peptides colored by cosine similiartity calculated using the vector encoding for peptide caseicin B-B1 (black star) as reference. Higher similarity values indicate more similarity between vectors; lower values indicate more difference.

To delve into the organization of specific AMPs encoded to the latent space, we used cosine similarity as a measure of distance between AMPs using their 50-dimension vectors as input. First, the cosine similarity for all vectors was calculated relative to each vector, generating a similarity list for every AMP. For each similarity list, the associated MICs for the five most similar vectors were averaged; this process was repeated for the five least similar vectors, and the difference between the two averages was taken. From this process, the greatest difference highlighted a largely inactive AMP, VLNENLLA, called caseicin B variant B1, a variant of caseicin B which is found in milk. Regardless of mutation, caseicin B exhibited a MIC of approximately 1250 μM against *E. coli* NCIMB 11843 [27]. Since this variant of caseicin B and other mutants each showed low activity against *E. coli* (≥ 1.25 mM), we were confident in sampling from latent space near their location would likely produce new AMPs similarly inactive and hypothesized that AMPs generated in a region most distant from this reference point were likely to be highly active against *E. coli.* Caseicin B-B1 is identified in **Fig. 2A-B** at the black star; a white star is located at the embedding most distant in cosine similarity encoding for the peptide KFGKIVGKVLKQLKKVSAVAKVAMKKG, trialysin peptide P4. Trialysin peptide P4 is a potent pore-forming peptide found in the saliva of *Triatoma infestans*, the insect vector of Chagas' disease, and lytic to *E. coli* in LB broth at approximately 10 μM [34]. In **Fig. 2A**, both caseicin B-B1 and trialysin peptide P4 locate within regions of low activity and high activity AMPs as organized within latent space, respectively, and in **Fig. 2B** both encode to regions of similarly high and low cosine similarity.

Next, we took the highest and lowest five AMPs (by cosine similarity) and grouped them: parent Group A and parent Group B. Here, the parent Group A five AMPs (closest to and including caseicin B-B1) were identified as VLNENLLA, VLNENLAA, VLNENLLK, VLNENLL, and VLNENLLH, each of which are caseicin B variants reported by Norberg *et al.* [27]. And parent Group B (most dissimilar to caseicin B-B1) was: KWKLWKKIEKWGQGIGAVLKWLTTWL,

KWKSFIKKLTSAAKKVVTTAKPLISS, KWKSFIKKLTSVLKKVVTTAKPLISS, KFFKKLKKAVKKGFKKFAKV, and KFGKIVGKVLKQLKKVSAVAKVAMKKG. The average MIC against *E. coli* for parent Group A group was 2500 μM, and for parent Group B: 1 μM. Nearby these ten AMPs (two groups of five) a total of 100 peptides were generated by the decoder. Following removal of duplicates or those that are already present in the database, 38 remained and were synthesized (see sequences in **Table 1**), with 22 peptides in the Group A and the remaining 16 peptides were in the Group B. The 38 sequences were designated names p1-38 and were associated with Parent peptide IDs corresponding to those Peptide IDs. Cosine similarity listed in **Table 1** is relative to caseicin B variant B1. For purposes of comparison, in addition to the 38 peptides from Group A and Group B, 100 control sequences were decoded from random 50-dimension vectors.

## 4. EXPERIMENTS

A secondary structure prediction algorithm (GOR IV) was used to predict helix, sheet, and coil percentages of the Group A and Group B sampling groups. Group A peptides were predicted to have similar proportions sheet and coil with medians 30% sheet and 37% coil, with a median of 0% helix (**Fig. 3A**). Conversely, Group B peptides were predominately helical at 62%, with the remainder composed of approximately equal proportion sheet and coil. Group A and Group B peptides are significantly different for both predicted proportion of helix and coil ($p < 0.01$, Welch's two-sided t-test). For comparison, the group of 100 random sequences were not found to be statistically different from Group A, while Group B had significantly higher in percent helix and significantly lower in percent coil ($p < 0.01$, Welch's two-sided t-test). Given the relatively high proportion of peptides predicted to be helical in Group B, the amphipathic nature of both groups was examined via calculated hydrophobic moments at 100 degrees. Predictably, differences between Group A and Group B are significant ($p < 0.01$, Welch's two-sided t-test). As expected, the distribution of hydrophobic moment of the randomly generated group was similar to that of the sequences found in the training set when the hydrophobic moment at 100 degrees is calculated, suggesting the VAE generations aligned well with real data. Altogether these results suggest the 22 Group A peptides are predicted to be significantly less helical than the 16 Group B peptides, while randomly sampling captures a wider range of predicted structures and non-amphipathic/amphipathic AMPs. Importantly, the predictions suggest controlled sampling from distinct subpopulations of latent space generates sequences with significantly different characteristics.

Experimental investigation of secondary structure was performed using circular dichroism (CD) in phosphate buffer with sodium dodecyl sulfate (SDS) micelles as a membrane-mimicking agent [35]. To account for concentration and difference in peptide length, mean residue molar ellipticity (MRME) was plotted to visualize the relative proportion of secondary structure for each group A and B (**Fig. 3C).** Results in the presence of SDS show that the average Group A peptide presents a mixture of random coil and helical character with minima predominant at ~205 nm suggests a largely random structure, with a smaller but noticeable dip at 222 nm suggesting a minor percentage of helix. Individual scans separated out shows a mixture of structures. Group B peptides were predominantly helical, with paired minima at ~208 and ~222 nm and, unlike Group A, were more uniform in the scans of individual peptides. Using the circular dichroism analysis program Beta Structure Selection, secondary structure was estimated from the CD data (converted to $\Delta\varepsilon$). Following summation of antiparallel, parallel, and turn into "sheet", the results were plotted for both Groups A and B. Analysis showed that Group A and B with median percent helicity of 4% and 63%, respectively. The results are comparable to those estimated from sequence via GOR IV.

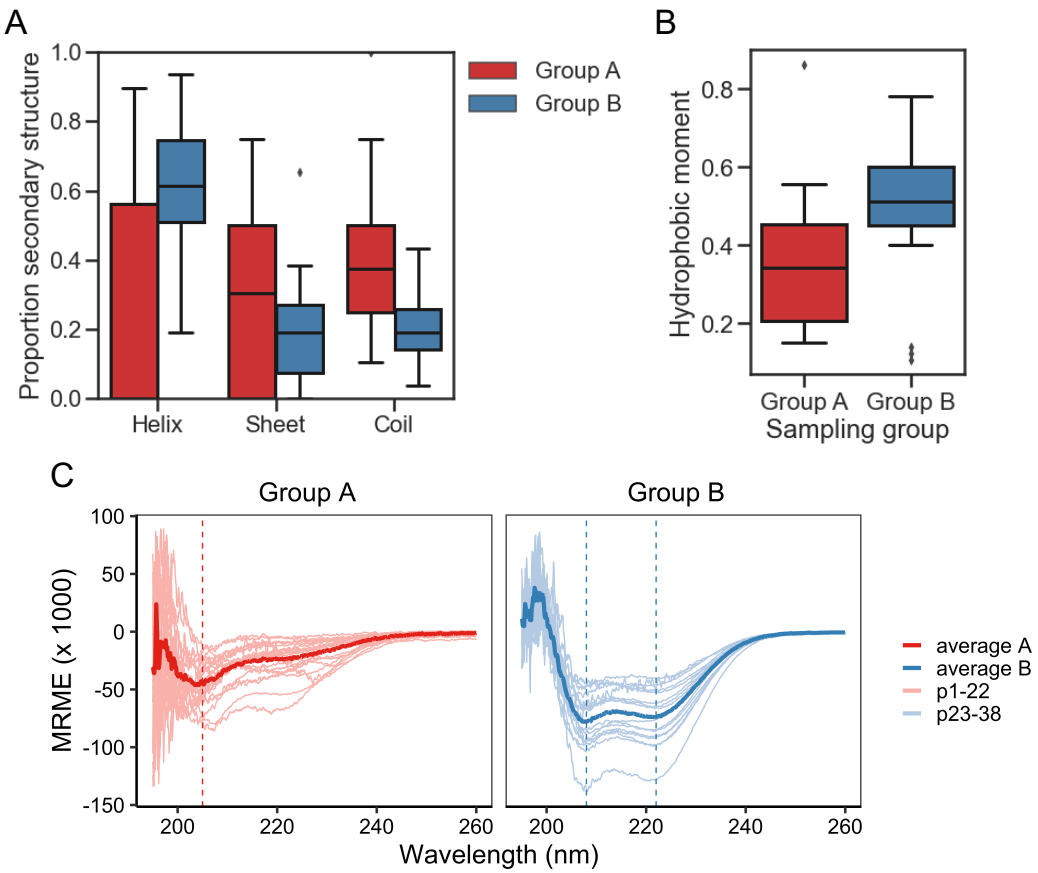


Fig. 3 — Generated peptide characterization. A) Boxplot of the predicted helix, sheet, and coil percentages calculated from Group A and Group B sampling groups into the GOR IV algorithm. Group A and Group B are significantly different ($p < 0.01$, Welch's two-sided t-test) for helix, sheet, and coil. B) Boxplot of the calculated hydrophobic moments obtained by applying the modlAMP function calculate_moment on the sequences from Group A and Group B sampling groups. The differences between Group A and Group B groups are significant ($p < 0.01$, Welch's two-sided t-test). C) Mean residue molar ellipticity (MRME) plots of peptides from group A (left) and group B (right) in the presence of membrane mimic 60 mM SDS. Each scan was averaged from three scans for each peptide with peptide-free buffer baseline scan subtracted. Lighter colored lines are individual peptide scans; darker colored lines are average scan for the group. Vertical lines highlight approximate minima of average scans: 205 nm for Group A; 208 and 222 nm for Group B.

Although, secondary structure and particular measurements such as hydrophobic moment are closely related to the antimicrobial activity of AMPs, particularly those with alpha helical character, more predictive measures are available in the form of regression models. Witten et al. and others have reported use of regression models, including a convolution neural network (CNN) [15]. In our study, we implemented the reported CNN as well as several machine learning regression models for MIC prediction against *E. coli*. Preliminary tests utilized model frameworks within the Statistics and Machine Learning Toolbox and Regression Learner app in MATLAB. However, the long training time of the best performing model identified (Rational Quadratic Gaussian Process Regression) led us to migrate to other models implemented in Python. The CNN from Keras, elastic net, gradient boosting (GB), kernel ridge, lasso, and random forest (RF), each from Scikit-learn, as well as light gradient boosting machine (LGBM) and an extreme gradient boosting (XGB) model were initially tested. Our results showed that three of the examined models significantly underperformed the others: lasso, kernel ridge, and elastic net, each with $R^2 < 0.4$ and relatively high RMSEs. This underperformance was also visible in the actual-predicted difference histograms where each of their distributions were flatter than the others. For this reason, these

three models were excluded from further study. In addition, the relatively complex CNN implemented as described by Witten *et al.* in Keras was underperforming in relationship to the length of time required to train the model and was therefore also excluded. The remaining four models – GB, RF, LGBM, and XGB – were further interrogated. Representative scatterplots are shown in **Fig. 4A**, with each showing $R^2$ higher than 0.67. Following successive train-test split-shuffling cross validation, GB was found to have both the highest median $R^2$ (0.73) and lowest RMSE (0.50) (**Fig. 4B-4C**) and was used going forward for MIC prediction of AMPs against *E. coli.* Likewise, using the *S. aureus* and *P. aeruginosa* datasets, we identified the best-performing models for predicting MIC against both *S. aureus* and *P. aeruginosa.* For *S. aureus* XGB was found to be the best predictor after cross validation, while for *P. aeruginosa,* although the RSME and $R^2$ measures disagreed, the RF model was selected. Although several regression models for MIC prediction of AMPs on *E. coli* have been reported with varying degrees of accuracy [10, 36-38], none have publicly accessible models in order to directly compare with our results. Nevertheless, given the modular nature of the described VAE framework, any superior MIC prediction system could be used in place of the described models.

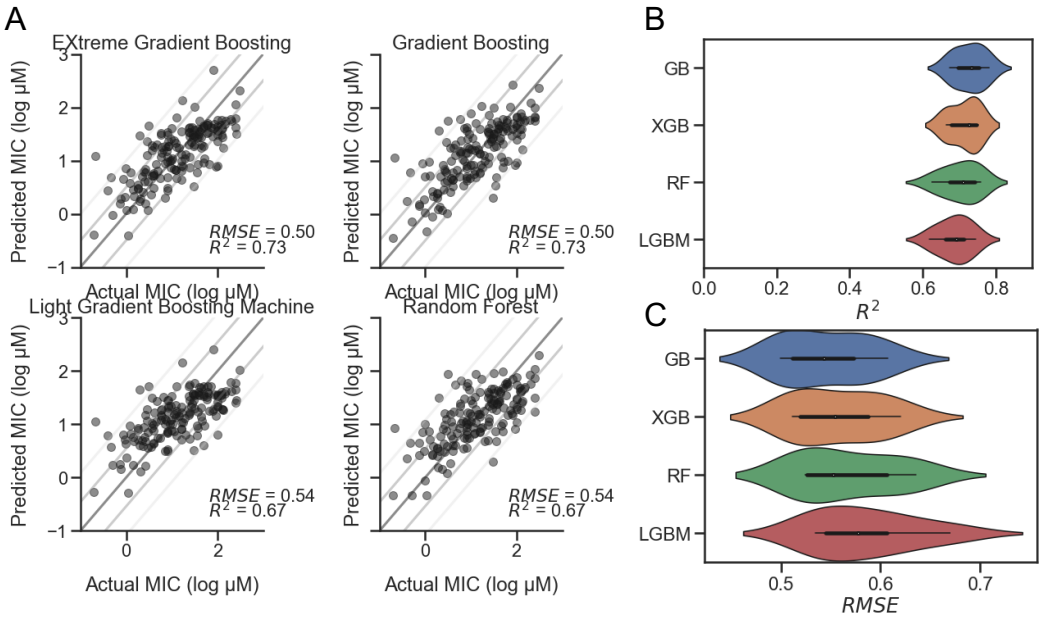


Fig. 4 — Comparison of MIC prediction models. Four regression models for predicting AMP MIC values against *E. coli*. A) Representative scatterplots of Predicted vs. Actual MIC (log μM) of EXtreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and Random Forest (RF) predictions on holdout test dataset with RMSE and $R^2$ values displayed. Lines represent standard diagonals, in addition to the diagonal +/- 1 and 2 standard deviations of the points. B) Results of cross validation using shuffled split (n=25) shown as a violin plot for each model, sorted from highest to lowest mean $R^2$ value, with GB at the top with a median of 0.73. C) Cross validation results for RMSE sorted from lowest to highest mean RMSE value, with GB at the top with a median of 0.54.

MICs predicted for Group A and Group B peptides are provided in **Fig. 5** and **Table 2**. For *E. coli*, the median predicted MIC of the Group A group was 1809 μM and 2 μM for Group B, and the sets were found to be significantly different ($p = 1 \times 10^{-8}$, Welch's two-sided t-test). As expected, the median of predicted MICs for the peptides decoded from randomly selected points in latent space was in between groups A and B: 12 μM. These predicted results are similar to the MICs of the parent AMPs used for generation of Group A and B. To investigate the predicted MICs of an intermediate location, we filtered regions of the latent space by cosine similarity relative the caseicin B variant B1 reference, randomly sampled 10 sequences, then generated new sequences (n=10) around these using the same method as

described above and predicted the MICs for each group. Each group was found to be significantly different from the other ($p < 0.01$). These results suggest intermediate locations between the polar ends of cosine similarity (and between Group A and Group B) would likely on average have corresponding intermediate MICs.

While the VAE was trained on the *E. coli* dataset, and sampling was performed with activity against *E. coli* in mind, we additionally examined the effectiveness of the generated AMPs on *S. aureus* and *P. aeruginosa,* the two most common bacteria in the modified GRAMPA dataset after *E. coli*. For *S. aureus*, the median predicted of Group A was 181 µM and 11 µM for Group B (**Fig. 5**). A Welch's two-sided t-test on Group A and B found $p = 1 \times 10^{-1}$. Meanwhile against *P. aeruginosa* the median predicted of Group A was 78 µM and 5 µM for Group B (**Fig. 5**). A Welch's two-sided t-test indicated a significant difference with $p = 3 \times 10^{-7}$. For both *S. aureus* and *P. aeruginosa,* random sampling yield median predicted MICs of 13 µM and 14 µM, respectively.

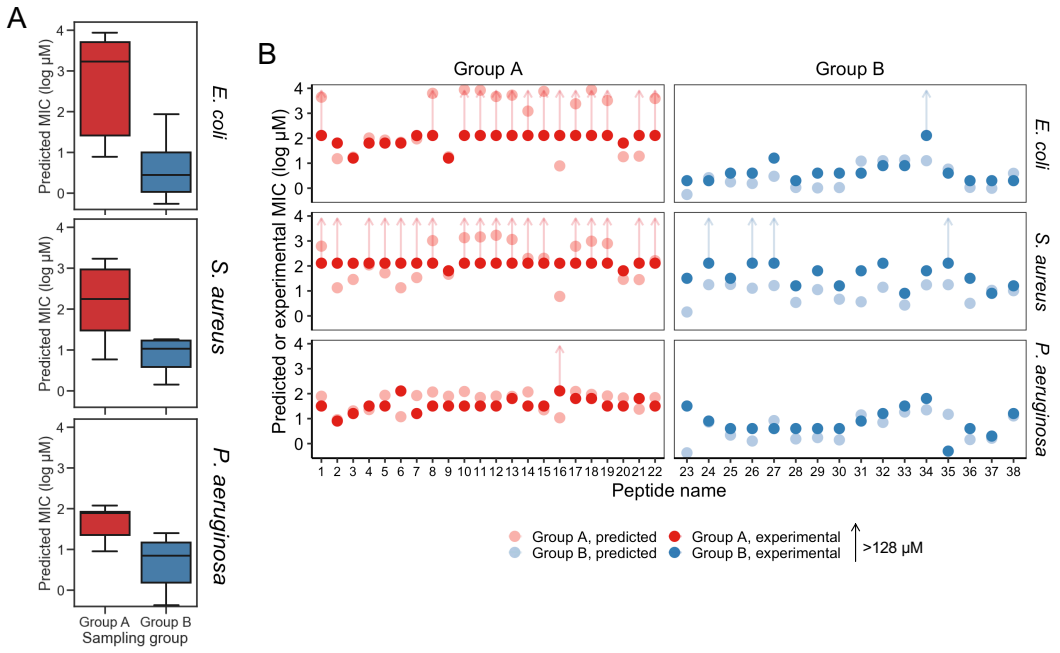


Fig. 5 — Predicted and experimental MICs against *E. coli, S. aureus*, and *P. aeruginosa*. A) Predicted MICs using the top-performing models for the Group A and Group B sampling groups are show in a boxplot, for *E. coli*, *S. aureus*, and *P. aeruginosa*, respectively. Group A and Group B are significantly different for each of the species ($p < 0.01$). B) Experimental MICs of Group A and Group B peptides against *E. coli*, *S. aureus*, and *P. aeruginosa*, with predicted MICs shown (with lighter coloring) for reference. Experimental MICs resulting in >128 µM determinations are shown with arrows extending up from 128 µM. All values are provided in Table 2.

To experimentally investigate the MIC of each of the 38 synthesized peptides against *E. coli*, the peptides were diluted in Mueller Hinton broth with a constant number of bacteria. Following incubation, we found that consistent with above predictions, the recorded MICs between the two sampling groups were significantly different. Among Group A AMPs, 63% of the MICs were found to be greater than 128 µM, while none of the 16 Group B AMPs were found to have MICs above 16 µM, other than peptide p34 (**Table 2**). After sorting for activity, the median value for Group A and Group B, was found to be >128 µM and 4 µM, respectively, which may be in line with the median predicted values of 1809 µM and 2 µM, although determining the accuracy of the values for Group A that are over 128 µM was not possible due to solubility issues. Importantly, the MIC predictions were not found to be different or independent from the experimental MICs, when categorized into >128 and ≤128 µM ($p < 0.01$, Fisher's exact test).

The MIC results paired with secondary structure estimates from circular dichroism experiments, highlight a number of active AMPs within Group B were composed of a low (< 50%) proportion of helix, including peptides p23, p24, p27, and p38. In addition, we found that both generated AMPs p31 and p33 have net charge ≤ 3, which in similar generative studies including Nagarajan et al and others would have been placed below their threshold for proceeding to experimental testing of activity [10]. Similarly, within Group A, there was a low-activity peptide with high helicity (> 50%) for p16, and relatively low-activity peptides with high net charge.

The experimentally determined MICs against *S. aureus* and *P. aeruginosa* were – unlike *E. coli* – further from their respective MIC predictions. For *S. aureus*, while 12 AMPs in Group A were predicted to have MICs >128 μM, 18 of the 22 peptides were found to have MICs at that level experimentally. The median MICs for Group A and Group B was found to be >128 μM and 32 μM, respectively, while the median predicted values were 181 μM and 11 μM. Similar to *E. coli*, the MIC predictions for *S. aureus* were similar, consistent with experimentally determined MICs when categorized into >128 and ≤128 μM ($p < 0.01$, Fisher's exact test). Although more of the predictions and experimental results were not in agreement, an insignificant difference was found using receiver. Receiver operating characteristic analysis was performed to evaluate for classification predictions for both species, which resulted in area under the curve values for *E. coli* and *S. aureus* at 0.93 and 0.9, respectively. For *P. aeruginosa*, the median MICs for Group A and Group B were experimentally determined to be 32 μM and 4 μM, respectively, compared to median predicted values of 78 μM and 5 μM. Consistent with the predicted MICs, the experimentally determined values too showed the larger separation between the median MICs for Groups A and B for *E. coli* than for *S. aureus* and *P. aeruginosa* (**Fig. 5A**; **Table 2**). When predicted and experimental MIC determinations found in **Table 2** are plotted relative to one another (**Fig. 5B**), after accounting for the >128 μM inequalities the predictions and experimental results closely align for *E. coli* and *P. aeruginosa*, with the exception of a few peptides deviating beyond a 5-fold difference in experiment vs. prediction, notably p34 tested against *E. coli* highlighted above. Peptide 16 for both *E. coli* and *P. aeruginosa* was also found be inactive while having a low predicted MIC. In contrast, for *S. aureus,* experiments notably deviate from predicted MICs for many of the peptides, where most of Group B predicted MICs significantly overestimated activity.

Table 2 — MIC assay results

| Peptide name | Group Id | Sequence | E. coli predicted MIC (μM) | E. coli experimental MIC (μM) | S. aureus predicted MIC (μM) | S. aureus experimental MIC (μM) | P. aeruginosa predicted MIC (μM) | P. aeruginosa experimental MIC (μM) |
|---|---|---|---|---|---|---|---|---|
| p1 | A | VLNANLLR | 4406 | >128 | 620 | >128 | 79 | 32 |
| p2 | A | VLIKTRLFIKRK | 15 | 64 | 13 | >128 | 9 | 8 |
| p3 | A | LNWKAILKHIIK | 18 | 16 | 29 | 128 | 20 | 16 |
| p4 | A | VLPKVMAHMK | 102 | 64 | 110 | >128 | 23 | 32 |
| p5 | A | LNWGAVLKHVVK | 84 | 64 | 53 | >128 | 86 | 32 |
| p6 | A | LILKRKRKRKRILI | 69 | 64 | 13 | >128 | 12 | 128 |
| p7 | A | LNWGAIKKHIIK | 94 | 128 | 34 | >128 | 84 | 16 |
| p8 | A | VLNENLLA | 6281 | >128 | 1031 | >128 | 117 | 32 |
| p9 | A | LNWGAFLKHFFK | 18 | 16 | 46 | 64 | 79 | 32 |
| p10 | A | VLNENLLH | 8761 | >128 | 1353 | >128 | 122 | 32 |
| p11 | A | VLNENAAR | 8333 | >128 | 1450 | >128 | 70 | 32 |
| p12 | A | VLNENLRR | 4691 | >128 | 1693 | >128 | 80 | 32 |
| p13 | A | VLNENLLR | 5307 | >128 | 1147 | >128 | 78 | 64 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| p14 | A | VDLKNLLK | 1221 | >128 | 200 | >128 | 117 | 32 |
| p15 | A | VALNENLLR | 7546 | >128 | 203 | >128 | 22 | 32 |
| p16 | A | LRRLRLRLLRLLRRLLRLL | 8 | >128 | 6 | 128 | 11 | >128 |
| p17 | A | VLNNLLR | 2398 | >128 | 612 | >128 | 123 | 64 |
| p18 | A | VLNENLAA | 8661 | >128 | 981 | >128 | 94 | 64 |
| p19 | A | VLNEALLR | 3233 | >128 | 793 | >128 | 81 | 32 |
| p20 | A | LNWGAWLKHWWK | 18 | 64 | 29 | 64 | 68 | 32 |
| p21 | A | LVKRVKKVL | 19 | >128 | 28 | >128 | 24 | 64 |
| p22 | A | VNLKNLLR | 3894 | >128 | 162 | >128 | 70 | 32 |
| p23 | B | KWKLWKKIEKWGQGIGAVLKWLTTWL | 1 | 2 | 1 | 32 | 0 | 32 |
| p24 | B | KWKSFLKTFKSPVKTVFYTALKPISS | 3 | 2 | 18 | >128 | 7 | 8 |
| p25 | B | KWKSFIKKLTSVLKKVVTTAKPLISS | 2 | 4 | 18 | 32 | 2 | 4 |
| p26 | B | KWKSFIKKLTSAAKKVVTTAKPLISS | 2 | 4 | 13 | >128 | 1 | 4 |
| p27 | B | KWKSFLKTFKSPARTVLHTALKPISS | 3 | 16 | 16 | >128 | 8 | 4 |
| p28 | B | KWKSFIKKLTSAAKKVLTTGLPALIS | 1 | 2 | 3 | 16 | 2 | 4 |
| p29 | B | KWKSFLKKLTSAAKKVLTTALKPISS | 1 | 4 | 11 | 64 | 2 | 4 |
| p30 | B | KWKSFLKTFKSAVKTVLHTALKAISS | 1 | 4 | 5 | 16 | 1 | 4 |
| p31 | B | FIGGLRRLFATVVGTVVGAINKLGGG | 12 | 4 | 4 | 64 | 14 | 8 |
| p32 | B | KFFKKLKKAVKKGFKKFAKV | 13 | 8 | 14 | 128 | 7 | 16 |
| p33 | B | FFFHIIKGLFHAGRMIHGLV | 13 | 8 | 3 | 8 | 18 | 32 |
| p34 | B | FFFKLLPKAIGALKKI | 13 | >128 | 18 | 64 | 22 | 64 |
| p35 | B | FKIKASKKLLKKVGKGALGAVAKALAQQA | 6 | 4 | 18 | >128 | 15 | 0.5 |
| p36 | B | KWKKFIKKLTSAAKKVLTTGLPALIS | 1 | 2 | 3 | 32 | 1 | 4 |
| p37 | B | KWKKFLKKLTSAAKKVLTTALKPISS | 1 | 2 | 11 | 8 | 2 | 2 |
| p38 | B | FFKKFIGGVAKIAGKAAPHGVGQLIPHVTP | 4 | 2 | 10 | 16 | 13 | 16 |

## 5. CONCLUSIONS

This study performed over the course of the Karles Fellowship (Sept. 3, 2019 – Sept. 2, 2020) demonstrates the use of a new peptide generation framework for discovery/design of new AMP sequences which, when tested experimentally, display expected and desired antimicrobial properties. This work builds upon our previously reported proof-of-concept report on *de novo* generation of short ≤ 12-mer peptides using a VAE [13], furtherer validating that the VAE is producing a smooth, well-organized latent space useful for AMP discovery. Utilizing a VAE-developed latent space paired with a series of modular antimicrobial activity predictive models, this framework shows the ability to produce AMPs with both predicted and experimentally validated activity against the targeted bacteria.

In order to address the development of a mechanism for selecting a reference AMP for controlled sequence output, we implemented a simple AMP generation-by-reference method that uses limited input parameters: reference peptide selection and the number of new sequences to generate. Using this method, we show that this method produces peptides with similar MICs as the input reference peptides, but with novel sequences not found in the training set. Specifically, in our testing, we generated AMPs in reference to caseicin B - variant B1, encoded nearby a series of single and double mutants with similar activity. Importantly, our results show that our list of newly generated active peptides includes non-canonical AMPs of low helicity and low net charge supports using the described VAE method, without imposing thresholds on peptide characteristics or otherwise biasing output post-sequence generation.

In the future, we hope to investigate the ability to sample nearby a particular selected AMP and determine which characteristics best retained by the newly generated AMPs. Results reported here suggest the average MIC of the generated AMPs from both Group A and Group B is similar to that of those AMP encoded near the location they were sampled from; however, it's unknown whether other measurable features will be retained: *e.g.,* Gram-specific or anti-biofilm activity. In addition, in future work we plan to investigate whether more sophisticated state-of-the-art generative models are more efficient for AMP discovery, comparing the output of CVAE to the described framework, as well as other generative models to determine which performs the best for controlled AMP sequence generation.

Work published during the course of the Karles Fellowship are listed below.

### Publications during Karles Fellowship

- Variational autoencoder for generation of antimicrobial peptides. **Scott N Dean**, Scott A Walper ACS Omega 1.2 (2020)
- Multi-layer epitaxial graphene on SiC: A stable working electrode for seawater samples spiked with environmental contaminants. Lisa C Shriver-Lake, Rachael L Myers-Ward, **Scott N Dean**, Jeffrey S Erickson, Scott A Trammell. Sensors (2020) p. 2392
- Lipid-Tagged Single Domain Antibodies for Improved Enzyme-Linked Immunosorbent Assays. Lisa C Shriver-Lake, Ellen Goldman, Jinny Liu, **Scott N Dean**, George P Anderson Journal of Immunological Methods (2020) p. 112790
- Francisella novicida two-component system response regulator BfpR modulates iglC gene expression, antimicrobial peptide resistance, and biofilm production. **Scott N Dean**, Morgan Milton, John Cavanagh, Monique L van Hoek. Frontiers in Cellular and Infection Microbiology 10 (2020) p. 82.
- Lactobacillus acidophilus membrane vesicles as a vehicle for bacteriocin delivery. **Scott N Dean**, Mary Ashely Rimmer, Kendrick B Turner, Dan A Phillips, Julie C Caruana, William Judson Hervey IV, Dagmar H Leary, Scott A Walper. Frontiers in Microbiology 11 (2020) p. 710.
- Quantum Dots and Gold Nanoparticles as Scaffolds for Enzymatic Enhancement: Recent Advances and the Influence of Nanoparticle Size. Gregory A Ellis, **Scott N Dean**, Scott A Walper, Igor L Medintz. Catalysts 10.1 (2020) p. 83.

**REFERENCES**


1. Brown, E.D. and G.D. Wright, *Antibacterial drug discovery in the resistance era.* Nature, 2016. **529**(7586): p. 336-343.
2. O'Neill, J., *Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations.* The Review on Antimicrobial Resistance 2014.
3. Sterling, T. and J.J. Irwin, *ZINC 15–ligand discovery for everyone.* Journal of chemical information and modeling, 2015. **55**(11): p. 2324-2337.
4. Stokes, J.M., et al., *A deep learning approach to antibiotic discovery.* Cell, 2020. **180**(4): p. 688-702. e13.
5. Lazzaro, B.P., M. Zasloff, and J. Rolff, *Antimicrobial peptides: Application informed by evolution.* Science, 2020. **368**(6490).
6. Kubicek-Sutherland, J.Z., et al., *Antimicrobial peptide exposure selects for Staphylococcus aureus resistance to human defence peptides.* Journal of Antimicrobial Chemotherapy, 2016. **72**(1): p. 115-127.
7. Mahlapuu, M., C. Bjorn, and J. Ekblom, *Antimicrobial peptides as therapeutic agents: opportunities and challenges.* Crit Rev Biotechnol, 2020. **40**(7): p. 978-992.
8. Porto, W.F., et al., *In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design.* Nature communications, 2018. **9**(1): p. 1-12.
9. Müller, A.T., J.A. Hiss, and G. Schneider, *Recurrent neural network model for constructive peptide design.* Journal of chemical information and modeling, 2018. **58**(2): p. 472-479.
10. Nagarajan, D., et al., *Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria.* Journal of Biological Chemistry, 2018. **293**(10): p. 3492-3509.
11. Tucs, A., et al., *Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks.* 2020.
12. Das, P., et al., *Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences.* arXiv preprint arXiv:1810.07743, 2018.
13. Dean, S.N. and S.A. Walper, *Variational Autoencoder for Generation of Antimicrobial Peptides.* ACS Omega, 2020.
14. Gabere, M.N. and W.S. Noble, *Empirical comparison of web-based antimicrobial peptide prediction tools.* Bioinformatics, 2017. **33**(13): p. 1921-1929.
15. Witten, J. and Z. Witten, *Deep learning regression model for antimicrobial peptide design.* bioRxiv, 2019: p. 692681.
16. Wang, G., X. Li, and Z. Wang, *APD3: the antimicrobial peptide database as a tool for research and education.* Nucleic Acids Res, 2016. **44**(D1): p. D1087-93.
17. Novković, M., et al., *DADP: the database of anuran defense peptides.* Bioinformatics, 2012. **28**(10): p. 1406-1407.
18. Pirtskhalava, M., et al., *DBAASP v. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides.* Nucleic acids research, 2016. **44**(D1): p. D1104-D1112.
19. Fan, L., et al., *DRAMP: a comprehensive data repository of antimicrobial peptides.* Scientific reports, 2016. **6**(1): p. 1-7.
20. Piotto, S.P., et al., *YADAMP: yet another database of antimicrobial peptides.* International journal of antimicrobial agents, 2012. **39**(4): p. 346-351.
21. Wright, E.S., *Using DECIPHER v2. 0 to analyze big biological sequence data in R.* R Journal, 2016. **8**(1).

22. Garnier, J., J.-F. Gibrat, and B. Robson, *GOR method for predicting protein secondary structure from amino acid sequence*, in *Methods in enzymology*. 1996, Elsevier. p. 540-553.
23. Walt, S.v.d., S.C. Colbert, and G. Varoquaux, *The NumPy array: a structure for efficient numerical computation.* Computing in science & engineering, 2011. **13**(2): p. 22-30.
24. Bowman, S.R., et al., *Generating sentences from a continuous space.* arXiv preprint arXiv:1511.06349, 2015.
25. Chollet, F. *Keras*. 2015; Available from: https://github.com/fchollet/keras.
26. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
27. Norberg, S., et al., *Altering the composition of caseicins A and B as a means of determining the contribution of specific residues to antimicrobial activity.* Applied and environmental microbiology, 2011. **77**(7): p. 2496-2501.
28. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
29. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
30. Ke, G., et al. *Lightgbm: A highly efficient gradient boosting decision tree*. in *Advances in neural information processing systems*. 2017.
31. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
32. Wiegand, I., K. Hilpert, and R.E. Hancock, *Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances.* Nature protocols, 2008. **3**(2): p. 163.
33. Team, R.C., *R: A language and environment for statistical computing*. 2013, Vienna, Austria.
34. Martins, R.M., et al., *Lytic activity and structural differences of amphipathic peptides derived from trialysin.* Biochemistry, 2006. **45**(6): p. 1765-1774.
35. Tulumello, D.V. and C.M. Deber, *SDS micelles as a membrane-mimetic environment for transmembrane segments.* Biochemistry, 2009. **48**(51): p. 12096-103.
36. Wu, X., et al., *In vitro and in vivo activities of antimicrobial peptides developed using an amino acid-based activity prediction method.* Antimicrobial agents and chemotherapy, 2014. **58**(9): p. 5342-5349.
37. Xiao, X. and Z.-B. You. *Predicting minimum inhibitory concentration of antimicrobial peptides by the pseudo-amino acid composition and Gaussian kernel regression*. in *2015 8th International Conference on Biomedical Engineering and Informatics (BMEI)*. 2015. IEEE.
38. Gull, S., *Amp0: Species-specific prediction of anti-microbial peptides using zero and few shot learning.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.