frontiers in Microbiology

ORIGINAL RESEARCH
published: 30 September 2021
doi: 10.3389/fmicb.2021.725727

# PepVAE: Variational Autoencoder Framework for Antimicrobial Peptide Generation and Activity Prediction

*Scott N. Dean[1]*, Jerome Anthony E. Alvarez[2], Dan Zabetakis[1], Scott A. Walper[1] and Anthony P. Malanoski[1]**

*[1]US Naval Research Laboratory, Center for Bio/Molecular Science and Engineering, Washington, DC, United States, [2]STEM Student Employment Program, US Naval Research Laboratory, Center for Bio/Molecular Science and Engineering, Washington, DC, United States*

New methods for antimicrobial design are critical for combating pathogenic bacteria in the post-antibiotic era. Fortunately, competition within complex communities has led to the natural evolution of antimicrobial peptide (AMP) sequences that have promising bactericidal properties. Unfortunately, the identification, characterization, and production of AMPs can prove complex and time consuming. Here, we report a peptide generation framework, PepVAE, based around variational autoencoder (VAE) and antimicrobial activity prediction models for designing novel AMPs using only sequences and experimental minimum inhibitory concentration (MIC) data as input. Sampling from distinct regions of the learned latent space allows for controllable generation of new AMP sequences with minimal input parameters. Extensive analysis of the PepVAE-generated sequences paired with antimicrobial activity prediction models supports this modular design framework as a promising system for development of novel AMPs, demonstrating controlled production of AMPs with experimental validation of predicted antimicrobial activity.

**Keywords: antimicrobial peptides, minimum inhibitory concentration, generative deep learning, activity prediction, variational autoencoder**

OPEN ACCESS

***Edited by:***
*Jianhua Wang,*
*Gene Engineering Laboratory, Feed Research Institute, Chinese Academy of Agricultural Sciences (CAS), China*

***Reviewed by:***
*Cesar de la Fuente-Nunez, University of Pennsylvania, United States*
*Qingfeng Guan, Hainan University, China*

****Correspondence:***
*Scott N. Dean*
*scott.dean@nrl.navy.mil*
*Anthony P. Malanoski*
*anthony.malanoski@nrl.navy.mil*

***Specialty section:***
*This article was submitted to Antimicrobials, Resistance and Chemotherapy, a section of the journal Frontiers in Microbiology*

***Received:*** *15 June 2021*
***Accepted:*** *30 August 2021*
***Published:*** *30 September 2021*

***Citation:***
*Dean SN, Alvarez JAE, Zabetakis D, Walper SA and Malanoski AP (2021) PepVAE: Variational Autoencoder Framework for Antimicrobial Peptide Generation and Activity Prediction. Front. Microbiol. 12:725727. doi: 10.3389/fmicb.2021.725727*

## INTRODUCTION

Many pathogenic bacteria are resistant to the majority of, if not all, antibiotics that are currently being isolated using traditional methods. Because of this, generation of new antimicrobials is critical for survival in the post-antibiotic era (Brown and Wright, 2016). At the current rate, annual global death due to antibiotic resistance is projected to exceed 10 million by 2050, costing 100 trillion USD (O'Neill, 2014); however, investment has not kept up with the task with very few antibiotics currently in clinical development and far fewer likely to be approved for treatment of patients (Trusts, 2019). To lower the burden required for antimicrobial development, various schemes for their discovery and refinement have been proposed, including improved methods for cultivation of antibiotic-producing organisms (Ling et al., 2015) and repurposing of FDA-approved drugs as antimicrobials (Farha and Brown, 2019). Recent work has shown that generative deep learning techniques can be applied to this problem: using a generative model, Stokes et al. (2020) showed that by training on a starting database of compounds from ZINC15 (Sterling and Irwin, 2015) new small molecule antibiotics can

be identified, such as halicin, which displays activity against *Acinetobacter baumannii* in a murine model of infection (Stokes et al., 2020) and is currently in clinical trials.

Unlike many small molecule antibiotics, 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 (Lazzaro et al., 2020). For this reason, among others, peptide antibiotics have been regularly deemed "drugs of last resort" for their ability to kill multidrug resistant bacteria, an increasingly important classification due to resistance formation toward conventional antibiotics (Lewies et al., 2019). Generally acting through mechanisms associated with membrane disruption, as well as other routes of incapacitation (Chung et al., 2015), the relative immutability of bacterial membranes and other essential AMP targets make the development of resistance to AMPs rare, but possible (Kubicek-Sutherland et al., 2016), thus increasing the importance of their reliable, continued discovery, to grow to the antimicrobial stockpile (Lazzaro et al., 2020).

Attempts at both generating new AMPs and improving their activity have been carried out with varying degrees of success (Mahlapuu et al., 2020). Many of these new or enhanced AMPs have been generated using low-throughput design methods, including rational design and specific amino acid substitution, *de novo* peptide design of alpha helices such as $(LKKL)_3$, use of templates or motifs, or otherwise high throughput techniques such as rational library design, each of which requiring expert knowledge (Huan et al., 2020). 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 (Porto et al., 2018).

In order to increase the rate of discovery of AMPs, newer high-throughput and low expertise design approaches are needed. In a similar vein to Stokes et al. (2020), 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 (Müller et al., 2018; Nagarajan et al., 2018), Generative Adversarial Networks (GANs; Tucs et al., 2020), and Variational Autoencoders (VAEs; Das et al., 2018; Dean and Walper, 2020). These works have been facilitated by thousands of AMP sequences housed in various databases, including Antimicrobial Peptide Database 3 (Wang et al., 2016) and Data Repository of Antimicrobial Peptides (Kang et al., 2019), which pair peptide sequences with experimentally determined activity against Gram-negative and Gram-positive bacteria, fungi, HIV, and cancer cells. To date several of these datasets have been formatted, so the sequential amino acid residues of AMPs can be represented in the form of a string of characters enabling machine/deep learning training and analysis of these datasets to identify novel AMP sequences (Müller et al., 2018; Nagarajan et al., 2019; Witten and Witten, 2019).

Although, recent results using generative deep learning for producing new sequences has shown promise, including a handful of experimental demonstrations of their activity (Nagarajan et al., 2018; Dean and Walper, 2020; Tucs et al., 2020), 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 (Gabere and Noble, 2017), 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 (Nagarajan et al., 2018; Porto et al., 2018). 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.

# MATERIALS AND METHODS

## Dataset and Analysis of Sequence Characteristics

The dataset used in this study was based on Giant Repository of AMP Activity (GRAMPA) as described by Witten and Witten (2019), with some modifications. Witten and Witten (2019) scraped all data from APD (Wang et al., 2016), DADP (Novković et al., 2012), DBAASP (Pirtskhalava et al., 2016), DRAMP (Fan et al., 2016), and YADAMP (Piotto et al., 2012), each accessed in Spring 2018, resulting in a combined 6,760 unique AMP sequences and 51,345 MIC values, and is publicly available on GitHub: https://github.com/zswitten/Antimicrobial-Peptides. A small percentage of MIC values were independently spot-checked and confirmed; however, methods vary widely between publications, and therefore, MIC values herein should be interpreted as approximations of activity. Since MICs determined against *E. coli* were the most-commonly available (**Supplementary Figure S1A**), these were used for the study. To avoid issues with synthesis, the dataset was further modified by excluding peptides with cysteine or any recorded modifications with the exception of C-terminal amidation. For ease of synthesis and to keep costs low, sequences ≥40 amino acids in length (representing 3.1% of sequences) were excluded; see **Supplementary Figures S1B,C** for the length distribution before and after this step. 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 (Witten and Witten, 2019). The sequences were tokenized, <end> token appended to each, and were 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 3,280, 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 2,974 and 1,968 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; Wright, 2016). PredictHEC makes use of the GOR IV algorithm (Garnier et al., 1996). This method is one of the best-performing predictors that uses only the primary sequence and does not require input of other sequences. Welch's *t*-test *via* NumPy (Walt et al., 2011) was used throughout for sample comparison with a significance threshold of 0.01 unless otherwise noted.

## Variational Autoencoder

The architecture of the VAE was implemented as described by Bowman et al. (2015), as described by Dean and Walper (2020) with minor modifications. The VAE model was trained on sequences in the *E. coli* dataset. 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 (Chollet, 2015) library with a TensorFlow (Abadi et al., 2016) backend, and used the Adam optimizer. The number of neurons for the LSTM layers found in both encoder and decoder were both set to 1,024. The number of latent dimensions was set to 50. All models were trained on an Ubuntu workstation with an Nvidia GeForce GTX1070 GPU. The LSTM network used in the decoder-encoder is 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.

## New Sequence Generation

Cosine similarity was used to compare latent space vectors generated by the VAE. To carry out cosine similarity calculations for each vector encoded, NumPy dot and norm functions were used which follow the notation:

$$\text{cosine similarity}(\text{a,b}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} = \frac{ab}{\|a\|\|b\|}$$

Where a and b are vectors; $a$ and $b$ are Euclidean ($L^2$) norms of vectors $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n$; Han et al., 2012), 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 (Norberg et al., 2011).

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:

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

Where $v_i$ ith vector of Group A or Group B.

$\mathcal{X}_{ij}$ random $1 \times 50$ vector sampled from a normal distribution $\mathcal{N}$.

$\mathcal{N}$ normal distribution function (with mean $\mu$ and SD $\sigma$).

$std\left(v_i\right)$ SD of ith vector of Group A or Group B.

The $\mu$ was set to 0 and $\sigma$ equal to the SD of the vectors from Group A and Group B [$std(v)$]. The resulting random $1 \times 50$ 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, 38 remained (sequences and other information available in **Table 1**).

## Latent Space Visualization

For dimensionality reduction principal component analysis (PCA), T-distributed stochastic neighbor embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) – each with two components – were used. PCA and t-SNE used were imported from Scikit-learn, while UMAP was from McInnes et al. (2018). For t-SNE, perplexity was set to 30, and learning rate was 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 was set to 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 similarity 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.

## Regression Models

Eight different regression models for predicting AMP MIC values: convolution neural network (CNN) as implemented by Witten and Witten (2019), Elastic Net (ENet), Gradient Boosting (GB), Kernel Ridge (KR), Lasso, and Random Forest (RF) models used were from the Python Scikit-learn library (Pedregosa et al., 2011), while Light Gradient Boosting Machine (LGBM) and EXtreme Gradient Boosting (XGB) used lightgbm

**TABLE 1** | PepVAE-generated peptides.

| Peptide name | Cosine similarity[a] | Group Id[b] | Sequence | Parent peptide ID[c] | MIC prediction (log μM) |
|---|---|---|---|---|---|
| p1 | 1.000 | A | VLNANLLR | 3,088 | 3.6 |
| p2 | 1.000 | A | VLIKTRLFIKRK | 3,088 | 1.2 |
| p3 | 1.000 | A | LNWKAILKHIIK | 3,088 | 1.2 |
| p4 | 1.000 | A | VLPKVMAHMK | 3,088 | 2.0 |
| p5 | 1.000 | A | LNWGAVLKHVVK | 3,088 | 1.9 |
| p6 | 1.000 | A | LILKRKRKRKRILI | 3,088 | 1.8 |
| p7 | 0.994 | A | LNWGAIKKHIIK | 3,085 | 2.0 |
| p8 | 0.994 | A | VLNENLLA | 3,085 | 3.8 |
| p9 | 0.994 | A | LNWGAFLKHFFK | 3,085 | 1.3 |
| p10 | 0.994 | A | VLNENLLH | 3,085 | 3.9 |
| p11 | 0.993 | A | VLNENAAR | 3,090 | 3.9 |
| p12 | 0.993 | A | VLNENLRR | 3,090 | 3.7 |
| p13 | 0.993 | A | VLNENLLR | 3,090 | 3.7 |
| p14 | 0.993 | A | VDLKNLLK | 3,090 | 3.1 |
| p15 | 0.993 | A | VALNENLLR | 3,090 | 3.9 |
| p16 | 0.984 | A | LRRLRLRLLRLLRRLLRLL | 3,087 | 0.9 |
| p17 | 0.984 | A | VLNNLLR | 3,087 | 3.4 |
| p18 | 0.984 | A | VLNENLAA | 3,087 | 3.9 |
| p19 | 0.984 | A | VLNEALLR | 3,087 | 3.5 |
| p20 | 0.981 | A | LNWGAWLKHWWK | 3,089 | 1.3 |
| p21 | 0.981 | A | LVKRVKKVL | 3,089 | 1.3 |
| p22 | 0.981 | A | VNLKNLLR | 3,089 | 3.6 |
| p23 | −0.952 | B | KWKLWKKIEKWGQGIGAVLKWLTTWL | 2,220 | −0.3 |
| p24 | −0.952 | B | KWKSFLKTFKSPVKTVFYTALKPISS | 2,220 | 0.4 |
| p25 | −0.952 | B | KWKSFIKKLTSVLKKVVTTAKPLISS | 2,220 | 0.2 |
| p26 | −0.952 | B | KWKSFIKKLTSAAKKVVTTAKPLISS | 2,220 | 0.2 |
| p27 | −0.952 | B | KWKSFLKTFKSPARTVLHTALKPISS | 2,220 | 0.5 |
| p28 | −0.958 | B | KWKSFIKKLTSAAKKVLTTGLPALIS | 2,227 | 0.0 |
| p29 | −0.958 | B | KWKSFLKKLTSAAKKVLTTALKPISS | 2,227 | 0.0 |
| p30 | −0.963 | B | KWKSFLKTFKSAVKTVLHTALKAISS | 2,228 | 0.0 |
| p31 | −0.963 | B | FIGGLRRLFATVVGTVVGAINKLGGG | 2,228 | 1.1 |
| p32 | −0.965 | B | KFFKKLKKAVKKGFKKFAKV | 1,802 | 1.1 |
| p33 | −0.965 | B | FFFHIIKGLFHAGRMIHGLV | 1,802 | 1.1 |
| p34 | −0.965 | B | FFFKLLPKAIGALKKI | 1,802 | 1.1 |
| p35 | −0.981 | B | FKIKASKKLLKKVGKGALGAVAKALAQQA | 1,809 | 0.8 |
| p36 | −0.981 | B | KWKKFIKKLTSAAKKVLTTGLPALIS | 1,809 | 0.0 |
| p37 | −0.981 | B | KWKKFLKKLTSAAKKVLTTALKPISS | 1809 | 0.0 |
| p38 | −0.981 | B | FFKKFIGGVAKIAGKAAPHGVGQLIPHVTP | 1,809 | 0.6 |

[a]*Cosine similarity: cosine similarity calculated using VLNENLLA as a reference.*

[b]*Group ID: A and B groupings.*

[c]*Parent peptide ID: index of the parental peptide sampled nearby.*

(Ke et al., 2017) and xgboost (Chen and Guestrin, 2016) libraries, respectively. The model parameters for each are provided in **Supplementary 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 Root mean square error (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 Assays

Minimum inhibitory concentration measurements values were measured using broth dilution method for AMPs (Wiegand et al., 2008). Peptides synthesized for use in this study are listed in **Table 1**. Peptides were synthesized by Genscript, Inc. (Piscataway, NJ, United States) 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, *S. aureus* ATCC 12600, and *P. aeruginosa* 27853 were grown in Mueller Hinton II Broth (BD, San Jose, CA, United States) at 37°C. Cultures were diluted to a final concentration of approximately $5 \times 10^5$ 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, United States) and 10 μl of the peptide, which was diluted in series, so that final peptide concentrations examined ranged from 128 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, United States) to identify the lowest concentration of peptide which inhibited growth. Statistical analysis of predicted and experimental MIC data was performed using the Fisher's Exact Test from stats package in R (R Core Team, 2013).

## Circular Dichroism

Circular dichroism (CD) spectra of the generated AMPs were obtained using a Jasco J-815 spectropolarimeter. Samples were allowed to equilibrate to 20°C prior to data collection in a 0.1 cm path length cuvette, with a chamber temperature of 20°C throughout each scan. Spectra were collected from 195 to 260 nm in 0.1-nm intervals. Data shown is an average of three scans, performed for each sample. All AMPs were analyzed at 25 μM concentration in either 10 mM sodium phosphate (pH 7) or 60 mM sodium dodecyl sulfate (SDS) in 10 mM sodium phosphate, both from Sigma (St. Louis, Missouri, United States). Baselines obtained prior to the experiment with peptide-free buffers were subtracted from each scan. Mean residue molar ellipticity (MRME) was calculated as follows:

$$MRME = \theta cl / R$$

Where $\theta$ is ellipticity (mdeg), $c$ is peptide concentration (mol/L), $l$ is cell path length (cm), and $R$ is the length of the peptide. MRME is presented multiplied by 1,000 to improve clarity.

# RESULTS

## Dataset Characterization and Framework Design

This study makes use of the Witten and Witten (2019) GRAMPA dataset as the starting point. 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 9,150 different entries (**Supplementary Figure S1A**). 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 (**Supplementary Figure S1B**), and median of 16 amino acids following removal of long sequences (**Supplementary Figure S1C**). Of the remaining peptides ($N = 3{,}280$), the median log μM MIC value was found to be 1.19 with a net charge of +4 (**Supplementary Figures S1C,D**). 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 in **Supplementary Figure S1E** 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 2,974 and 1,968, 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* (see **Supplementary Figures S2, S3**, respectively).

Using the above defined dataset for *E. coli*, we designed a VAE AMP generation pipeline (**Figure 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 2 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.

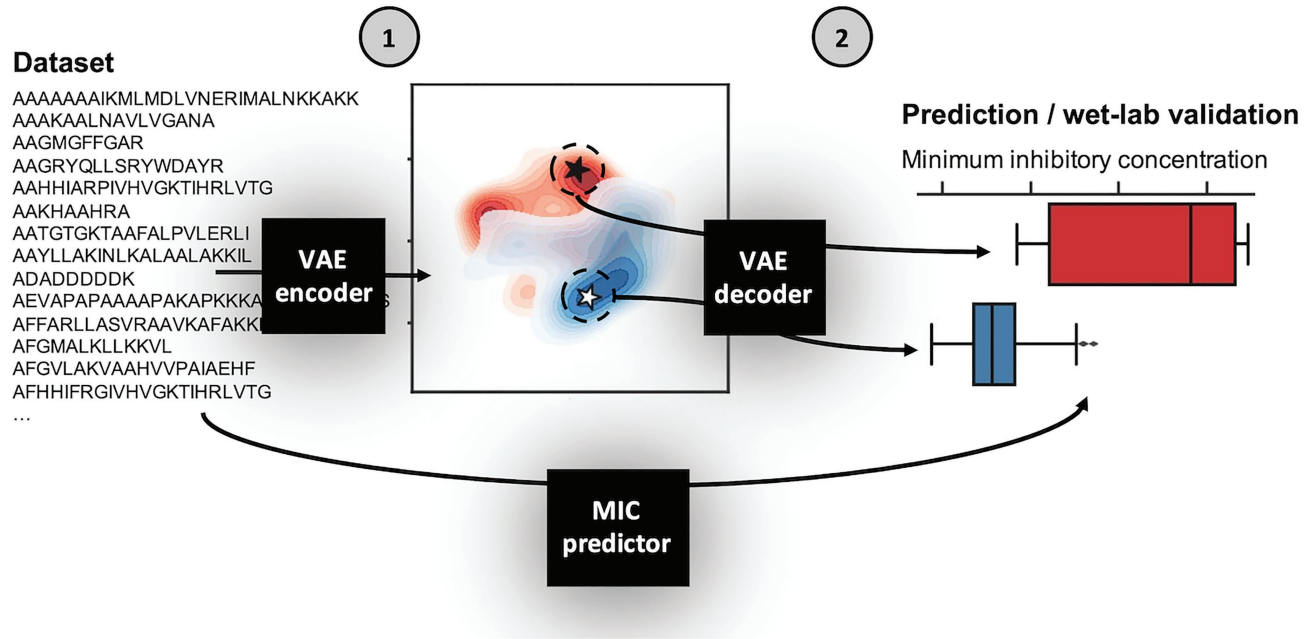


**FIGURE 1 |** Schematic of variational autoencoder (VAE) antimicrobial peptide (AMP) generation and design process. The VAE AMP generation and design process occurs in two stages: (1) training the VAE for the development of the latent space and a regression model for activity prediction, and (2) sampling 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 (Dean and Walper, 2020), making use of VAE described by Bowman et al. (2015), which was reported for use in generating new sentences. Here, the number of intermediate dimensions was set to 1,024 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 Methods section.

A VAE implemented as previously described (Dean and Walper, 2020), making use of VAE described by Bowman et al. (2015), was trained on the *E.coli* dataset as described above. The number of intermediate dimensions was set to 1,024 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 and characteristics dependent on that sequence were the components "learned" by the VAE (Dean and Walper, 2020). 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 AMP sequences using the generated decoder (see diagram in **Figure 1**).

## 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: PCA, t-SNE, and UMAP (**Supplementary Figure S4A**). 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 (**Supplementary Figure S4B**). 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 **Figures 2A,B**.

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

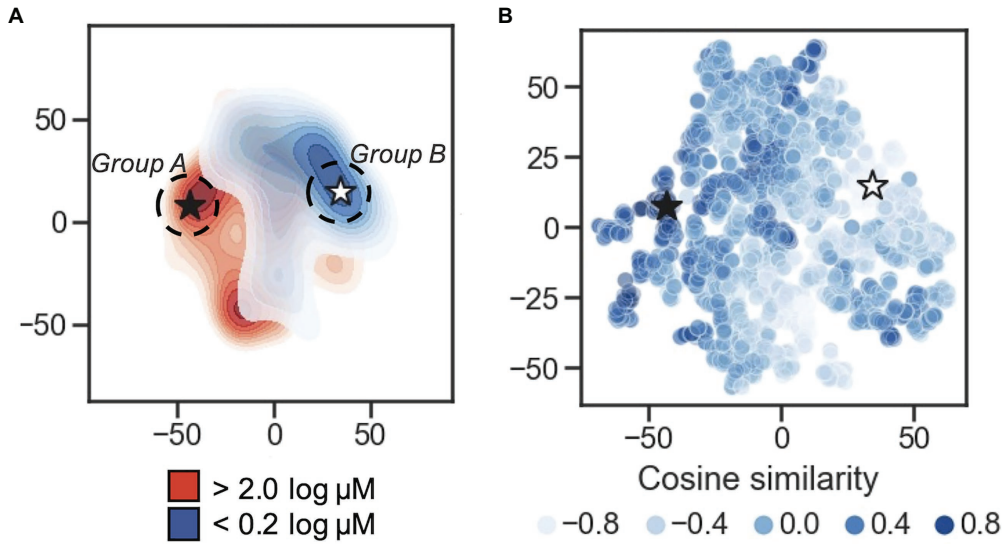


**FIGURE 2 |** Latent space characterization. Dimensionality reduction for visualization of the 50-dimension latent space. **(A)** 2D contour plot of t-distributed stochastic neighbor embedding (t-SNE) with two components performed on the encoded peptides. The MIC thresholds for coloring were: <0.2 log 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 similarity 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.

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 1,250 μM against *E. coli* NCIMB 11843 (Norberg et al., 2011). 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 **Figures 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 (Martins et al., 2006). In **Figure 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 **Figure 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. (2011). And parent Group B (most dissimilar to caseicin B-B1) was: KWKLWKKIEKWGQGIGAVLKWLTTWL, KWKSFIKK LTSAAKKVVTTAKPLISS, KWKSFIKKLTSVLKKVVTTAKP LISS, KFFKKLKKAVKKGFKKFAKV, and KFGKIVGKVLKQ LKKVSAVAKVAMKKG. The average MIC against *E. coli* for parent Group A group was 2,500 μM, and for parent Group B: 1 μM. Nearby these 10 AMPs (two groups of five) a total of 100 peptides were generated by the decoder. Following removal of duplicates, 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 found in **Supplementary Table S1** – the dataset used for training the model. 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.

## AMP Generation, Characterization, and MIC Results

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 (**Figure 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; **Supplementary Figure S5A**). 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; **Figure 3B**). 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 (see **Supplementary Figures S1F, S5B**). 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 CD in phosphate buffer with SDS micelles as a membrane-mimicking agent (Tulumello and Deber, 2009). To account for concentration and difference in peptide length, MRME was plotted to visualize the relative proportion of secondary structure for each group A and B (**Figure 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 (**Supplementary Figure S6A**) 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,[1] secondary structure was estimated from the CD data (converted to $\Delta\varepsilon$ ).

[1]http://bestsel.elte.hu/

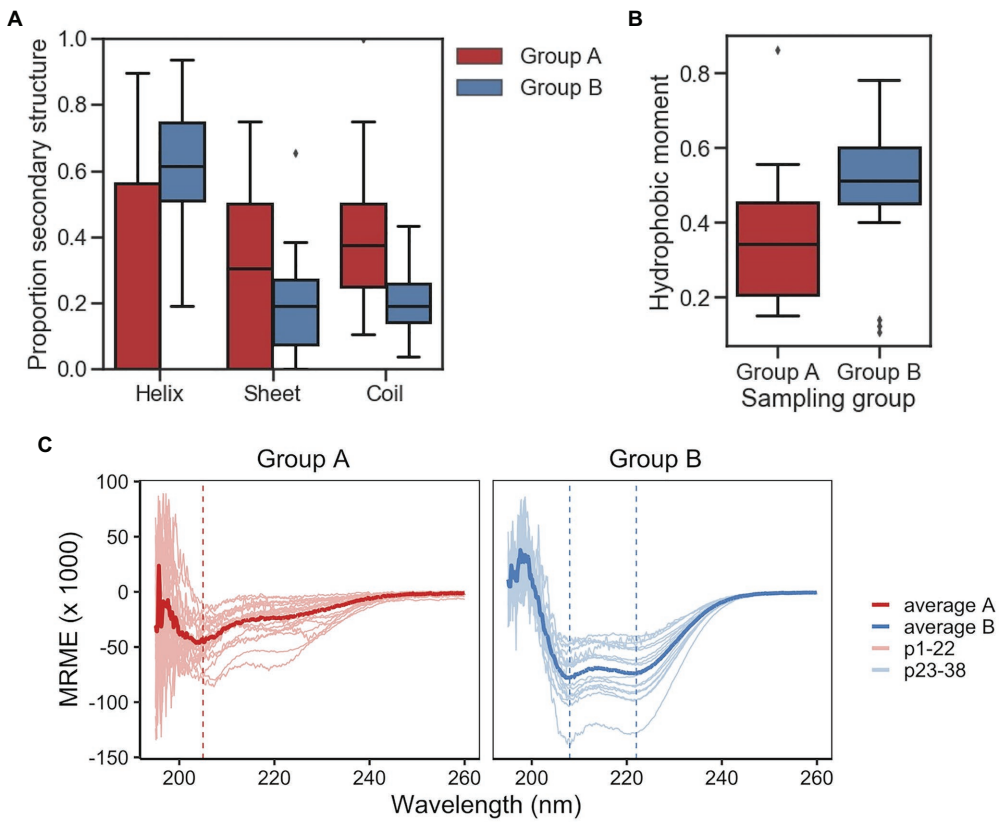


**FIGURE 3 |** Generated peptide characterization. **(A)** Boxplot of the predicted helix, sheet, and coil percentages calculated from Group A and Group B sampling groups into the GOR IV algorithm. Group A and Group B are significantly different ($p < 0.01$, Welch's two-sided $t$-test) for helix, sheet, 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 sodium dodecyl sulfate (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.

Following summation of antiparallel, parallel, and turn into "sheet," the results were plotted for both Groups A and B (**Supplementary Figure S6B**). 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.

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 and Witten (2019) and others have reported use of regression models, including a CNN. 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 (**Supplementary Figure S6**). 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, GB, kernel ridge, lasso, and RF, each from Scikit-learn, as well as LGBM and an 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 (**Supplementary Figure S8A**). This underperformance was also visible in the actual-predicted difference histograms (**Supplementary Figure S8B**), 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 and Witten (2019) 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 **Figure 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; **Figures 4B,C**) 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* (**Supplementary Figure S9**) and *P. aeruginosa* (**Supplementary Figure S10**). For *S. aureus* XGB was found to be the best predictor after cross validation (**Supplementary Figure S9**), while for *P. aeruginosa*, although, the RSME and $R^2$ measures disagreed, the RF model was selected (**Supplementary Figure S10**). Although, several regression models for MIC prediction of AMPs on *E. coli* have been reported with varying degrees of accuracy (Wu et al., 2014; Xiao and You, 2015; Nagarajan et al., 2018; Gull, 2020), none have publicly accessible models in order to directly compare with our results. Nevertheless, given the modular nature of the described PepVAE framework, any superior MIC prediction system could be used in place of the described models.

Minimum inhibitory concentrations predicted for Group A and Group B peptides are provided in **Figure 5** and **Table 2**. For *E. coli*, the median predicted MIC of the Group A group was 1,809 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 (**Supplementary Figure S5C**). 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 (**Supplementary Figure S11**). Each group was found to be significantly different from the other ($p < 0.01$). These results suggest intermediates 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* (**Supplementary Figure S1A**). For *S. aureus*, the median predicted of Group A was 181 and 11 μM for Group B (**Figure 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 and 5 μM for Group B (**Figure 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 and 14 μM, respectively (**Supplementary Figure S5C**).

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 and 4 μM, respectively, which may be in line with the median predicted values of 1,809 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 notably the unique p38, which displayed high activity against each bacteria. In addition, we found that both generated AMPs p31 and p33 have net charge ≤3, which in similar generative studies including Nagarajan et al. (2018) and others would have been

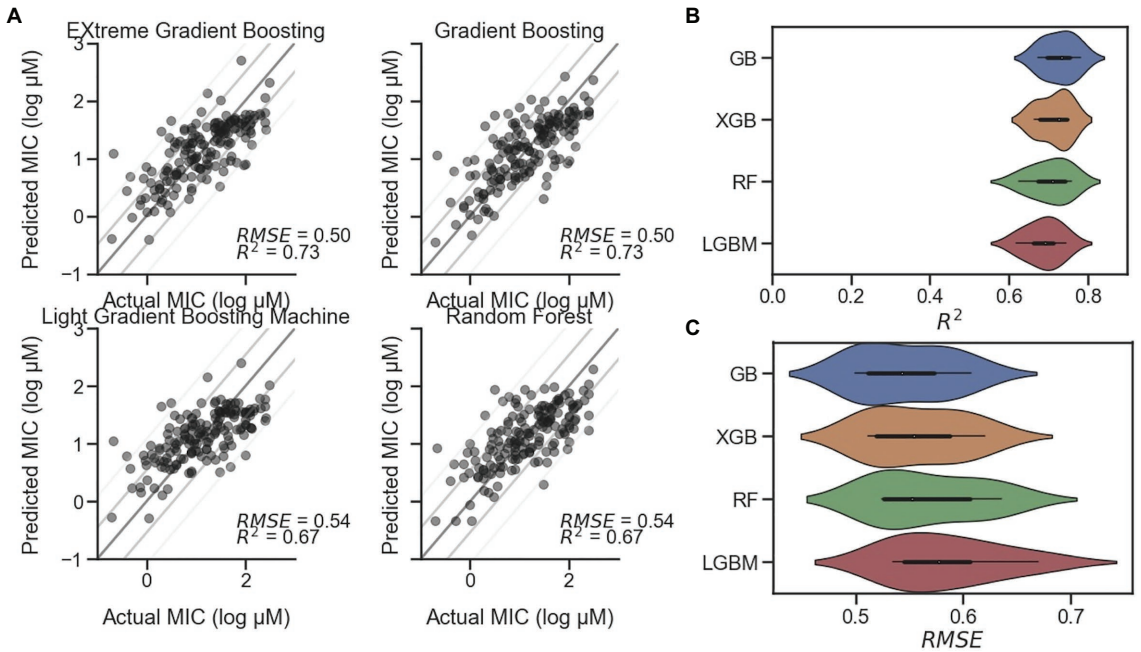


**FIGURE 4 |** Comparison of MIC prediction models. Four regression models for predicting AMP MIC values against *E. coli*. **(A)** Representative scatterplots of Predicted vs. Actual MIC (log μM) of EXtreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and 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 SDs 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.

placed below their threshold for proceeding to experimental testing of activity. 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 novel p16 peptide was also particularly interesting due to its incorrect predicted MIC (≤11 μM for each bacteria), but experimental MICs that were in-line with the rest of Group A.

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 and 32 μM, respectively, while the median predicted values were 181 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 operating characteristic analysis. 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 (**Supplementary Figure S12**). For *P. aeruginosa*, the median MICs for Group A and Group B were experimentally determined to be 32 and 4 μM, respectively, compared to median predicted values of 78 and 5 μM. Consistent with the predicted MICs, the experimentally determined MICs for *E. coli* displayed a larger separation between Groups A and B for than for the corresponding comparison for *S. aureus* and *P. aeruginosa* (**Figure 5A**; **Table 2**). When predicted and experimental MIC determinations found in **Table 2** are plotted relative to one another (**Figure 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.

## DISCUSSION

This study reports the use of a peptide generation framework, PepVAE, for discovery and design of new AMP sequences. Using the learned latent space, we demonstrate the identification of new active AMPs using reference peptides as input. Paired with antimicrobial activity prediction models, this modular framework shows the ability to produce AMPs with both predicted and experimentally validated activity against the targeted bacteria. Previous work on generative deep learning models for the design of AMPs has demonstrated the ability of VAEs to form a well-organized, usable latent space representation from which novel peptide sequences

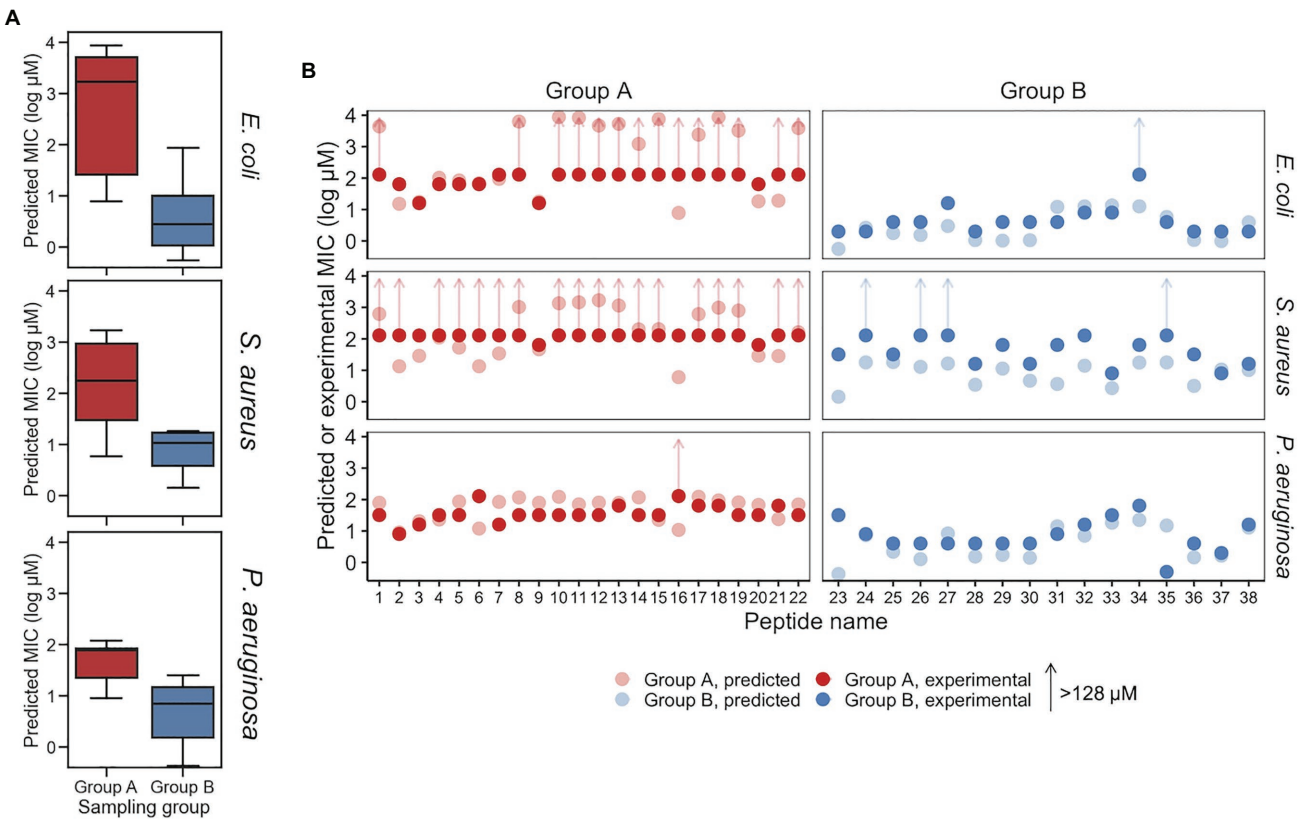


**FIGURE 5 |** Predicted and experimental MICs against *E. coli*, *S. aureus*, and *P. aeruginosa*. **(A)** Predicted MICs using the top-performing models for the Group A and Group B sampling groups are show in a boxplot, for *E. coli*, *S. aureus*, and *P. aeruginosa*, respectively. Group A and Group 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**.

can be generated (Das et al., 2018; Dean and Walper, 2020). These reports when paired with similar promising work on the discovery of small molecules using VAEs and other generative deep learning models broadly establishes these techniques as valuable new methods for molecular and material design (Sanchez-Lengeling and Aspuru-Guzik, 2018). Although, generative deep learning models for AMP generation have previously shown their ability to produce distributions of characteristics that closely match the databases of sequences on which they were trained, unconditional generation alone does not readily solve the problem of discovering new AMPs that are potent against target bacteria. In particular, previously described systems would benefit from certain functions to improve automated discovery of new potent AMPs: testable activity prediction, a mechanism for AMP generation *via* a reference peptide for controlled sequence generation, and an expanded sequence space from which to sample.

Many previously described systems readily generate sequences that are both predicted and experimentally determined to be inactive, by nature of the models used when training sets include inactive or low-activity sequences. In the literature, while classifiers for AMP activity are widely available and meta-analyses of their performance has been reported (Gabere and Noble, 2017), regression models for activity prediction – MIC, half maximal effective concentration, or other metrics – are less common. Several studies make use of the CAMPR3 predictor (Waghu et al., 2016) or other similar models that output abstract activity predictions as a probability [*probability(active AMP)*], not readily relatable to MIC or another antimicrobial activity metrics (Müller et al., 2018; Nagarajan et al., 2018; Dean and Walper, 2020). Beyond the lack of testability of these predictions, it likely that the increasing hydrophobic moment yielded higher probabilities in the CAMPR3 predictor and other models suggesting that simply alternating groups of positively ionizable and hydrophobic amino acids will score highly, highlighting the importance of experimentally verifying the antimicrobial activity of generated AMPs. To address these issues, our study implemented ML models trained on sequence data and adjoining experimental MIC values to predict MICs of new peptide sequences. Critically, we found that the MIC predictions for *E. coli* were not statistically different from the experimental MICs and can, therefore, be used for assessing whether a potential AMP is likely to be active against *E. coli* prior to peptide synthesis. While the MIC predictions and for *S. aureus* and *P. aeruginosa* were similarly found to be statistically similar to the experimental outcomes, the

**TABLE 2 |** Minimum inhibitory concentration (MIC) assay results.

| Peptide name | Group Id | Sequence | *Escherichia coli* predicted MIC (μM) | *Escherichia coli* experimental MIC (μM) | *Staphylococcus aureus* predicted MIC (μM) | *Staphylococcus aureus* experimental MIC (μM) | *Pseudomonas aeruginosa* predicted MIC (μM) | *Pseudomonas aeruginosa* experimental MIC (μM) |
|---|---|---|---|---|---|---|---|---|
| p1 | A | VLNANLLR | 4,406 | >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 | 6,281 | >128 | 1,031 | >128 | 117 | 32 |
| p9 | A | LNWGAFLKHFFK | 18 | 16 | 46 | 64 | 79 | 32 |
| p10 | A | VLNENLLH | 8,761 | >128 | 1,353 | >128 | 122 | 32 |
| p11 | A | VLNENAAR | 8,333 | >128 | 1,450 | >128 | 70 | 32 |
| p12 | A | VLNENLRR | 4,691 | >128 | 1,693 | >128 | 80 | 32 |
| p13 | A | VLNENLLR | 5,307 | >128 | 1,147 | >128 | 78 | 64 |
| p14 | A | VDLKNLLK | 1,221 | >128 | 200 | >128 | 117 | 32 |
| p15 | A | VALNENLLR | 7,546 | >128 | 203 | >128 | 22 | 32 |
| p16 | A | LRRLRLRLLRLLRRLLRLL | 8 | >128 | 6 | 128 | 11 | >128 |
| p17 | A | VLNNLLR | 2,398 | >128 | 612 | >128 | 123 | 64 |
| p18 | A | VLNENLAA | 8,661 | >128 | 981 | >128 | 94 | 64 |
| p19 | A | VLNEALLR | 3,233 | >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 | 3,894 | >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 |

inaccuracy of the predictions for MICs against these bacteria relative to *E. coli* was not unexpected due to the smaller training sets and lower accuracy of the MIC predictions on the test set. Increases in both data quantity and quality, with respect to the specific strain used, may significantly improve MIC predictions, especially for non-*E. coli* bacteria. Pairing with PepVAE, further compilation of data for would likely allow for improved generation of AMPs targeting species of interest.

Generative models for AMP design can be bifurcated into those that use a starting sequence and those that do not (i.e., *de novo* design). To address the development of a mechanism for selecting a reference AMP for controlled sequence output, as opposed to unconditional generation, which would improve existing AMP generation frameworks, 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. We show that this method can produce peptides with similar MICs as the input reference peptides, but with novel sequences not found in the training set. One possible limitation of this system is the likely requirement of colocalization with nearby similar AMPs with similar activity to raise confidence in sampling in the region. In our testing, we generated AMPs in reference to caseicin B-B, encoded nearby a series of single and double mutants with similar activity. It is possible that an isolated, single reference peptide is not sufficient, and that a small panel of AMPs may be required for assurance – sampling from a sparsely populated region will likely produce results similar to that of the random sample group in this study, who is predicted MICs span a wide range.

Another issue our study improved upon is the expanded sequence space from which to sample. Other groups have shown the benefits of tying a continuous activity prediction to output from an AMP-generating models rather than simple classification, following up with experimental validation. These methods, however, used predicted amphipathicity, MIC, and charge filtering steps in order to obtain target active sequences (Nagarajan et al., 2018; Porto et al., 2018), restricting sequence output to a specific set of positively charged, alpha helical peptides. Our results showing 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. However, we did perform pre-training biasing due to cost and synthesis limitations on both length and cysteine presence, and we expect the structural diversity of generated AMPs to have been greater had these limitations not been a factor, and we plan to avoid more of these biasing factors in future studies.

In addition to investigating AMPs generated using models trained on data without length or cysteine constraints, given sufficient funds for synthesis costs, future studies will examine our results that suggest sampling from nearer regions (i.e., not highest or lowest code similarity but interpolations in between) would on average result in generated AMPs with middling activity in a range between the polar Groups A and B. This interpolation may look similar in outcome to our proof-of-concept report on *de novo* generation of short ≤12-mer peptides using a VAE (Dean and Walper, 2020) but would further validate that the VAE is producing a smooth, well-organized latent space. Another interesting feature to investigate in the future, relating to the ability to sample nearby a particular selected AMP, are the characteristics best retained by the newly generated AMPs. The results 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, but it is unclear whether other measurable features will be retained. For example, if an AMP with highly specific activity particular species or Gram classification is used as the reference, will the generated AMPs also selectively kill? The above-mentioned possibility of pairing with PepVAE with further compiled non-*E. coli* data for likely allowing for improved generation of AMPs targeting species of interest leads us to speculate whether this framework can result in a species-level granularity in terms of activity or whether other techniques would be better suited for this goal. In addition, we highlight that this study utilized a vanilla LSTM VAE that was used previously, not a more sophisticated variant, such as a conditional VAE (CVAE) for AMP generation (Das et al., 2018). It is unclear whether the results would be improved by conditioning the latent space on MIC information without an explicit comparison paired with experimental validation; in future studies, we plan to investigate this, comparing the output of CVAE to PepVAE, as well as other generative models to determine which performs the best for controlled AMP sequence generation.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

## AUTHOR CONTRIBUTIONS

SD initiated the study, performed the experiments, generated the models, and performed data analysis. JA and DZ performed supporting data analysis. SW and AM oversaw and edited the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

We acknowledge funding support through the Jerome and Isabella Karle Distinguished Scholar Fellowship provided by the Naval Research Laboratory, base funds of the Naval Research Laboratory (WU# 1V33), and funds from the Defense Threat Reduction Agency (HDTRA1033536).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.725727/full#supplementary-material

## REFERENCES


Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning." in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* November 2, 2016.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. arXiv [Preprint].

Brown, E. D., and Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature* 529, 336–343. doi: 10.1038/nature17042

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system." in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, August 13, 2016.

Chollet, F. (2015). Keras. Available at: https://github.com/fchollet/keras (Accessed: January 14, 2021).

Chung, M.-C., Dean, S. N., and van Hoek, M. L. (2015). Acyl carrier protein is a bacterial cytoplasmic target of cationic antimicrobial peptide LL-37. *Biochem. J.* 470, 243–253. doi: 10.1042/BJ20150432

Das, P., Wadhawan, K., Chang, O., Sercu, T., Santos, C. D., Riemer, M., et al. (2018). Pepcvae: semi-supervised targeted design of antimicrobial peptide sequences. arXiv [Preprint].

Dean, S. N., and Walper, S. A. (2020). Variational autoencoder for generation of antimicrobial peptides. *ACS Omega* 5, 20746–20754. doi: 10.1021/acsomega.0c00442

Fan, L., Sun, J., Zhou, M., Zhou, J., Lao, X., Zheng, H., et al. (2016). DRAMP: a comprehensive data repository of antimicrobial peptides. *Sci. Rep.* 6:24482. doi: 10.1038/srep24482

Farha, M. A., and Brown, E. D. (2019). Drug repurposing for antimicrobial discovery. *Nat. Microbiol.* 4, 565–577. doi: 10.1038/s41564-019-0357-1

Gabere, M. N., and Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* 33, 1921–1929. doi: 10.1093/bioinformatics/btx081

Garnier, J., Gibrat, J.-F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266, 540–553. doi: 10.1016/s0076-6879(96)66034-0

Gull, S. (2020). Amp0: species-specific prediction of anti-microbial peptides using zero and few shot learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.2999399 [Epub ahead of print]

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining Concepts and Techniques. 3rd Edn. Vol. 3*. Massachusetts, USA: Elsevier.

Huan, Y., Kong, Q., Mou, H., and Yi, H. (2020). Antimicrobial peptides: classification, design, application and research progress in multiple fields. *Front. Microbiol.* 11:582779. doi: 10.3389/fmicb.2020.582779

Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., et al. (2019). DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. data* 6, 1–10. doi: 10.1038/s41597-019-0154-y

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "Lightgbm: a highly efficient gradient boosting decision tree." in *Advances in Neural Information Processing Systems*, December 4, 2017.

Kubicek-Sutherland, J. Z., Lofton, H., Vestergaard, M., Hjort, K., Ingmer, H., and Andersson, D. I. (2016). Antimicrobial peptide exposure selects for *Staphylococcus aureus* resistance to human defence peptides. *J. Antimicrob. Chemother.* 72, 115–127. doi: 10.1093/jac/dkw381

Lazzaro, B. P., Zasloff, M., and Rolff, J. (2020). Antimicrobial peptides: application informed by evolution. *Science* 368:eaau5480. doi: 10.1126/science.aau5480

Lewies, A., Du Plessis, L. H., and Wentzel, J. F. (2019). Antimicrobial peptides: the Achilles' heel of antibiotic resistance? *Probiotics Antimicrob. Proteins* 11, 370–381. doi: 10.1007/s12602-018-9465-0

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., et al. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature* 517, 455–459. doi: 10.1038/nature14098

Mahlapuu, M., Bjorn, C., and Ekblom, J. (2020). Antimicrobial peptides as therapeutic agents: opportunities and challenges. *Crit. Rev. Biotechnol.* 40, 978–992. doi: 10.1080/07388551.2020.1796576

Martins, R. M., Sforça, M. L., Amino, R., Juliano, M. A., Oyama, S., Juliano, L., et al. (2006). Lytic activity and structural differences of amphipathic peptides derived from trialysin. *Biochemistry* 45, 1765–1774. doi: 10.1021/bi0514515

McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. arXiv [Preprint].

Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model.* 58, 472–479. doi: 10.1021/acs.jcim.7b00414

Nagarajan, D., Nagarajan, T., Nanajkar, N., and Chandra, N. (2019). A uniform in vitro efficacy dataset to guide antimicrobial peptide design. *Database* 4:27. doi: 10.3390/data4010027

Nagarajan, D., Nagarajan, T., Roy, N., Kulkarni, O., Ravichandran, S., Mishra, M., et al. (2018). Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J. Biol. Chem.* 293, 3492–3509. doi: 10.1074/jbc.M117.805499

Norberg, S., O'Connor, P. M., Stanton, C., Ross, R. P., Hill, C., Fitzgerald, G. F., et al. (2011). Altering the composition of caseicins A and B as a means of determining the contribution of specific residues to antimicrobial activity. *Appl. Environ. Microbiol.* 77, 2496–2501. doi: 10.1128/AEM.02450-10

Novković, M., Simunić, J., Bojović, V., Tossi, A., and Juretić, D. (2012). DADP: the database of anuran defense peptides. *Bioinformatics* 28, 1406–1407. doi: 10.1093/bioinformatics/bts141

O'Neill, J. (2014). "Antimicrobial resistance: tackling a crisis for the health and wealth of nations," in *The Review on Antimicrobial Resistance.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Piotto, S. P., Sessa, L., Concilio, S., and Iannelli, P. (2012). YADAMP: yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents* 39, 346–351. doi: 10.1016/j.ijantimicag.2011.12.003

Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H. L., Squires, R. B., Hurt, D. E., et al. (2016). DBAASP v. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* 44, D1104–D1112. doi: 10.1093/nar/gkv1174

Porto, W. F., Irazazabal, L., Alves, E. S., Ribeiro, S. M., Matos, C. O., Pires, Á. S., et al. (2018). In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun.* 9:1490. doi: 10.1038/s41467-018-03746-3

R Core Team (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365. doi: 10.1126/science.aat2663

Sterling, T., and Irwin, J. J. (2015). ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell* 180, 688.e13–702.e13. doi: 10.1016/j.cell.2020.01.021

Trusts, P. C. (2019). Antibiotics Currently in Global Clinical Development. Available at: https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2014/antibiotics-currently-in-clinical-development (Accessed: January 14, 2021).

Tucs, A., Tran, D. P., Yumoto, A., Ito, Y., Uzawa, T., and Tsuda, K. (2020). Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks. *ACS Omega* 5, 22847–22851. doi: 10.1021/acsomega.0c02088

Tulumello, D. V., and Deber, C. M. (2009). SDS micelles as a membrane-mimetic environment for transmembrane segments. *Biochemistry* 48, 12096–12103. doi: 10.1021/bi9013819

Waghu, F. H., Barai, R. S., Gurung, P., and Idicula-Thomas, S. (2016). CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* 44, D1094–D1097. doi: 10.1093/nar/gkv1051

Walt, S. V. D., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37

Wang, G., Li, X., and Wang, Z. (2016). APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* 44, D1087–D1093. doi: 10.1093/nar/gkv1278

Wiegand, I., Hilpert, K., and Hancock, R. E. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* 3:163. doi: 10.1038/nprot.2007.521

Witten, J., and Witten, Z. (2019). Deep learning regression model for antimicrobial peptide design. bioRxiv [Preprint]. doi: 10.1101/692681v1

Wright, E. S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. *R J.* 8, 352–359. doi: 10.32614/RJ-2016-025

Wu, X., Wang, Z., Li, X., Fan, Y., He, G., Wan, Y., et al. (2014). In vitro and in vivo activities of antimicrobial peptides developed using an amino acid-based activity prediction method. *Antimicrob. Agents Chemother.* 58, 5342–5349. doi: 10.1128/AAC.02823-14

Xiao, X., and You, Z.-B. (2015). "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)*, October 14, 2015, IEEE.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.