


RESEARCH ARTICLE

De novo synthetic antimicrobial peptide design with a recurrent neural network

Chenkai Li^{1,2}  | Darcy Sutherland^{1,3,4} | Amelia Richter^{1,3} | Lauren Coombe¹ | Anat Yanai^{1,3} | René L. Warren¹ | Monica Kotkoff¹ | Fraser Hof⁵ | Linda M. N. Hoang^{3,4} | Caren C. Helbing⁶ | Inanc Birol^{1,3,4,7}

¹Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada

²Bioinformatics Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada

³Public Health Laboratory, British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada

⁴Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada

⁵Department of Chemistry and the Centre for Advanced Materials and Related Technology, University of Victoria, Victoria, British Columbia, Canada

⁶Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada

⁷Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence

Inanc Birol, Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada.
Email: ibirol@bcgsc.ca

Funding information

Genome BC, Grant/Award Number: 291PEP; Genome Canada, Grant/Award Number: 291PEP; Investment Agriculture Foundation of BC, Grant/Award Number: INV106

Abstract

Antibiotic resistance is recognized as an imminent and growing global health threat. New antimicrobial drugs are urgently needed due to the decreasing effectiveness of conventional small-molecule antibiotics. Antimicrobial peptides (AMPs), a class of host defense peptides, are emerging as promising candidates to address this need. The potential sequence space of amino acids is combinatorially vast, making it possible to extend the current arsenal of antimicrobial agents with a practically infinite number of new peptide-based candidates. However, mining naturally occurring AMPs, whether directly by wet lab screening methods or aided by bioinformatics prediction tools, has its theoretical limit regarding the number of samples or genomic/transcriptomic resources researchers have access to. Further, manually designing novel synthetic AMPs requires prior field knowledge, restricting its throughput. *In silico* sequence generation methods are gaining interest as a high-throughput solution to the problem. Here, we introduce AMPd-Up, a recurrent neural network based tool for *de novo* AMP design, and demonstrate its utility over existing methods. Validation of candidates designed by AMPd-Up through antimicrobial susceptibility testing revealed that 40 of the 58 generated sequences possessed antimicrobial activity against *Escherichia coli* and/or *Staphylococcus aureus*. These results illustrate that AMPd-Up can be used to design novel synthetic AMPs with potent activities.

KEYWORDS

antibiotic resistance, antimicrobial peptide, *de novo* peptide design, recurrent neural network

Reviewing Editor: Nir Ben-Tal

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

1 | INTRODUCTION

The worldwide overuse of antibiotics has created an alarming number of bacteria that possess antibiotic resistance, resulting in conventional antibiotics being less effective (Reardon, 2014). It is estimated that 1.27 million people died due to antibiotic resistance in 2019 (Antimicrobial Resistance Collaborators, 2022), and the speed of bacterial evolution, resulting in antibiotic resistance, is expected to greatly increase this death toll in the next few decades (Laxminarayan et al., 2013; O'Neill, 2014). Moreover, the sluggish pace of discovery and development of new therapeutics is exacerbating this public health crisis (Koo and Seo, 2019). As a result, novel effective substitutes for conventional antibiotics are urgently needed as weapons to fight against multidrug-resistant bacteria, also referred to as “superbugs”.

Antimicrobial peptides (AMPs), a diverse class of short and often cationic peptides, are considered a viable alternative to conventional antibiotics (van der Does et al., 2019). Naturally occurring AMPs are observed among all forms of life (Zhang and Gallo, 2016). In higher eukaryotic organisms, AMPs have co-evolved with environmental microbes as part of the host innate immune system (Zhang and Gallo, 2016). Microbes can also produce AMPs for inter-competition purposes against the growth of other microbes (Zhang and Gallo, 2016). Most of the known AMPs reported in public databases are antibacterial, with some AMPs active or additionally active against other types of microbes (e.g., fungi, viruses) (Wang et al., 2016). Unlike most conventional antibiotics, which have specific functional or structural targets, most AMPs act directly on bacterial membranes or cell walls leading to non-enzymatic disruption, with some eukaryotic AMPs performing additional modulation of the host immune system (Nguyen et al., 2011; Zhang and Gallo, 2016). As a result, it may be more difficult for bacteria to develop resistance to AMPs compared with conventional antibiotics (Boman, 2003). However, resistance to AMPs can still be observed if bacteria are exposed to AMPs for sufficient periods of time (Boman, 2003), indicating that antibiotic resistance is an enduring phenomenon. Thus, high-throughput methods for the rapid discovery and design of novel AMPs would be instrumental in our fight against superbugs (Lin et al., 2022).

Recently, a number of *in silico* AMP prediction tools have been developed to reduce the labor and costs associated with large-scale wet lab screening for AMP discovery (Jukić and Bren, 2022; Li et al., 2022; Meher et al., 2017; Veltri et al., 2018; Xiao et al., 2013). State-of-the-art AMP prediction tools include AMPlify (Li et al., 2022), AMP Scanner Vr.2 (Veltri et al., 2018), iAMPpred (Meher

et al., 2017), and iAMP-2L (Xiao et al., 2013). Each of these tools utilizes machine learning methods, with AMPlify outperforming the latter three tools by adapting a deep learning model with attention mechanisms (Li et al., 2022; Vaswani et al., 2017; Yang et al., 2016). These *in silico* tools have successfully been applied in identifying novel, naturally occurring AMPs from genomic or transcriptomic resources (Li et al., 2022; Lin et al., 2022; Richter et al., 2022). Nevertheless, the discovery of these AMPs is limited by the availability of organism sources, such as tissue samples for direct wet lab screening or sequencing data for *in silico* mining. Even though *in silico* mining methods are high-throughput, they require massive amounts of upstream work for careful data preparation (Li et al., 2022; Lin et al., 2022; Richter et al., 2022), which further limits the pace of development and the number of novel AMPs that can be discovered.

The potential sequence space of amino acids is combinatorially large, allowing for the design of peptide sequences that may not exist in nature but still possess desirable antimicrobial properties. Traditional approaches for AMP design include (1) modification of known AMP sequences to generate their congeners, fragments, or hybrids; (2) minimalist approaches by which AMPs are designed *de novo* purely based on structural requirements (e.g., amphipathic alpha-helical structures) but with limited types (e.g., physicochemical properties) of residues used; (3) sequence-template-guided approaches that create sequence templates by comparing structurally homologous fragments from known AMPs for conserved patterns in terms of residue types; and (4) utilizing combinatorial peptide libraries (Huan et al., 2020; Tossi, 2011). However, these methods require prior expertise in AMPs' research for more accurate designs, which restricts the throughput.

Recently, a series of machine learning models based on neural networks have been proposed for the automatic *de novo* design of AMP sequences (Das et al., 2021; Dean et al., 2021; Gupta and Zou, 2019; Nagarajan et al., 2018; Szymczak et al., 2022; Tucs et al., 2020; Van Oort et al., 2021). They make it possible for users to sample novel AMP sequences directly from the models without any artificial design. Common sequence generation models include recurrent neural network (RNN) language models (Mikolov et al., 2010), variational autoencoders (VAEs) (Kingma and Welling, 2014), and generative adversarial networks (GANs) (Goodfellow et al., 2014). Nagarajan et al. developed a long short-term memory (LSTM) RNN language model (Hochreiter and Schmidhuber, 1997; Mikolov et al., 2010), and embedded it into a framework with multiple filtering steps for the generation of novel AMPs with strong antibacterial activity (Nagarajan et al., 2018). Dean et al. proposed a

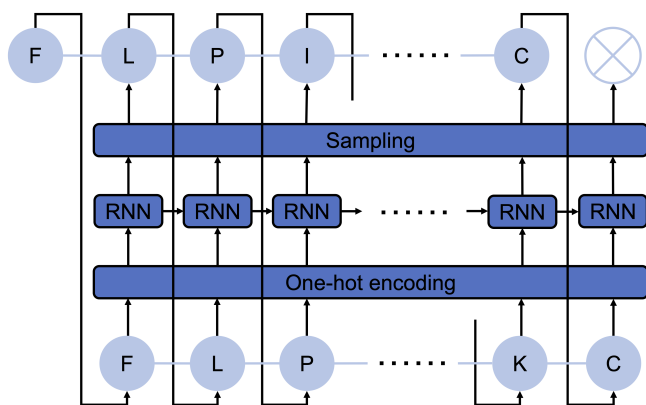


FIGURE 1 Architecture of the recurrent neural network (RNN) language model. Given a starting amino acid, the RNN language model predicts the next amino acids residue by residue until reaching the end-of-sequence (EOS) signal (represented as a cross marker). Amino acids, including the EOS signal, are one-hot encoded. The output of RNN at each time step is a probability vector of amino acid and EOS occurrence at the next position, to which sampling strategies can be applied.

VAE-based AMP sequence generation framework, named PepVAE, for generation of highly active AMPs (Dean et al., 2021). Das et al. further adapted VAE and introduced CLaSS for controlled AMP sequence generation with attributes of interest (Das et al., 2021). HydrAMP, another VAE-based model, incorporates two pre-trained classifiers monitoring the quality of the generated peptides during training (Szymczak et al., 2022), improving upon a conditional VAE (cVAE) (Sohn et al., 2015). Gupta et al. proposed Feedback GAN for generating DNA sequences that encode proteins with optimized properties, and applied it to AMP sequence generation as an example (Gupta and Zou, 2019). Tucs et al. adapted an activity-aware LeakGAN (Guo et al., 2018) to generate highly active AMPs (Tucs et al., 2020), while Van Oort et al. introduced AMPGAN v2 based on a bidirectional conditional GAN (BiCGAN) (Donahue et al., 2017; Dumoulin et al., 2017) to generate AMP sequences of different types and properties (Van Oort et al., 2021). The flurry of activities represented by these methods illustrate a strong interest in the field for *de novo* AMP design and explore expertise-free approaches. Nonetheless, there is still room for improvement in generating AMP designs with desirable properties and high potency.

In the presented work, we introduce AMPd-Up, a novel AMP sequence generation tool that implements a standard RNN language model (Mikolov et al., 2010) (Figure 1). The tool focuses on generating short AMP sequences ≤ 50 amino acids (aa) in length, with potential antibacterial activity. AMPd-Up samples candidate AMP sequences from multiple model instances trained with

different random initializations. For *de novo* AMP sequence generation, our RNN language model learns the “grammar”—the arrangement of the amino acids—of the training AMP sequences and estimates the probabilities of amino acid occurrence at each position recurrently starting from the N-terminus. Thus, the model generates a putative AMP sequence, residue by residue, based on the probability distribution estimated at each residue position (or each time step of the process), until reaching the end-of-sequence (EOS) signal. We expect different model instances to capture the complicated underlying features of AMP sequences from slightly different aspects, thus exploring various localities in the state space represented by a rich repertoire of natural AMPs. With this approach, we generated 40 novel AMPs that have not been reported in public databases but were proven to be active against laboratory strains of *Escherichia coli* and/or *Staphylococcus aureus*. Our results illustrate the power of AMPd-Up in contributing to our expanding arsenal of synthetic antimicrobial agents.

2 | RESULTS

2.1 | Performance comparison with state-of-the-art methods

We measured the performance of AMPd-Up by assessing the generated sequences using three state-of-the-art AMP prediction tools: AMPlify (Li et al., 2022), AMP Scanner Vr.2 (Veltri et al., 2018), and iAMPpred (Meher et al., 2017). The estimated sequence generation accuracy values, expressed as the percentages of sequences predicted as AMPs by each AMP prediction tool, are reported in Table 1. The results of three other AMP sequence generation methods: the LSTM language model (Nagarajan et al., 2018), AMPGAN v2 (Van Oort et al., 2021), and HydrAMP (Szymczak et al., 2022), are listed in Table 1 for comparison. Although none of the *in silico* prediction tools are perfect in identifying AMPs, their reported performance (Li et al., 2022; Meher et al., 2017; Veltri et al., 2018) would be suitable for evaluating the AMP sequence generation methods. Details of how we calculated the estimated accuracy values can be found in Section 4.

As measured by AMPlify, AMPd-Up obtains the highest estimated accuracy with 95.50% of the generated sequences predicted as AMPs on average, which outperforms the best comparator AMPGAN v2 by 4.60%, followed by HydrAMP (by 8.00%) and then the LSTM language model (by 10.65%). When evaluated using AMP Scanner Vr.2 and iAMPpred, AMPd-Up generates AMP sequences with estimated accuracies of 100.00%

TABLE 1 Performance comparison of different AMP sequence generation methods.

AMP sequence generation method	Estimated accuracy evaluated by AMP prediction tools (%)		
	By AMPLify	By AMP Scanner Vr.2	By iAMPpred
AMPd-Up	95.50 ± 0.35	100.00 ± 0.00	99.30 ± 0.37
LSTM ^a	84.85 ± 0.75	84.20 ± 1.04	82.80 ± 0.97
AMPGAN v2 ^b	90.90 ± 2.10	87.55 ± 1.29	94.85 ± 1.29
HydrAMP ^c	87.50 ± 1.15	94.60 ± 0.46	97.70 ± 0.64

Note: Different methods were evaluated using three *in silico* AMP prediction tools: AMPLify (Li et al., 2022), AMP Scanner Vr.2 (Veltri et al., 2018), and iAMPpred (Meher et al., 2017), based on sequences generated by each of the methods. The estimated AMP sequence generation accuracy measured by a selected prediction tool was defined as the percentage of peptide sequences predicted as AMPs among a generated sequence set. For each sequence generation method, five sets of sequences were generated, with 400 in each set. For each AMP sequence generation method, an average estimated accuracy value of the five generated sets was reported when measured by a specific AMP prediction tool, along with the corresponding standard deviation value. One-sided Welch's *t*-tests indicate that the superior performance of AMPd-Up over its comparators is statistically significant ($p < 0.05$).

Abbreviations: AMP, antimicrobial peptide; LSTM, long short-term memory.

^aSequences sampled from the generated sequence set provided by the authors (Nagarajan et al., 2018).

^bAntibacterial peptides were selected for a fairer comparison with other methods (Van Oort et al., 2021).

^cSequences generated using online server on November 7, 2022 (Szymczak et al., 2022).

and 99.30%, surpassing the best comparator HydrAMP by 5.40% and 1.60%, respectively. Although the rankings of the AMP sequence generation methods evaluated by the three AMP prediction tools are slightly different from each other, AMPd-Up always performs the best compared with its comparators.

2.2 | De novo generated sequences

Besides using the outputs of *in silico* AMP prediction tools as a proxy for performance, we also analyzed the generated sequences based on their amino acid compositions, length and net charge distributions, as well as their sequence similarity levels to the training set and all known AMP sequences. Details of how we analyzed the sequences generated by AMPd-Up can be found in Section 4.

Figure S1 in Data S1 summarizes the amino acid compositions of the generated sequences. The sequences generated by AMPd-Up were substantially rich in lysine (K) and leucine (L) residues, with average proportions of 29.87% and 24.51% per peptide sequence, respectively. In comparison, the sequences in our training set were rich in leucine (L), glycine (G), and lysine (K) residues, with average proportions of 11.50%, 10.94%, and 10.67%, respectively. Figure S1 in Data S1 additionally provides the amino acid composition information of the putative AMP sequences generated by three other methods. Two of the other methods (i.e., the LSTM language model and AMPGAN v2) highlighted lysine (K) and leucine (L) as predominant amino acid residues in their generated sequences, similar to the pattern observed in AMPd-Up.

Short lengths and net positive charges are common characteristics for most previously discovered AMPs (Zhang and Gallo, 2016), therefore many AMP studies investigate these key properties (Gagnon et al., 2017). Shorter peptides are also cheaper to synthesize (Lin et al., 2022), making translating shorter sequences for clinical application potentially more cost-effective. Further, the net positive charges of cationic AMPs are responsible for the electrostatic interaction with the negatively charged bacterial membranes or cell walls (Zhang and Gallo, 2016), with studies illustrating that the antimicrobial activity of some AMPs can be improved by increasing their net charges (Zelezetsky and Tossi, 2006). The top section of Figure 2 compares the length distributions of the sequences generated by AMPd-Up with those constituting the training set. We note that the model may fail to reach the EOS signals when generating some sequences (referred to as “incomplete sequences”; see Section 4 for details); we thus additionally compared the generated sequence set with those incomplete sequences removed. The average generated sequence length was 28.90 aa, but was reduced to 21.56 aa after incomplete sequences were removed. The incomplete sequences are 50 aa by default. The complete sequences were 4.65 aa shorter than the training sequences on average. The bottom section of Figure 2 shows a similar comparison for net charge distributions. The average generated sequence net charge was 9.08, but was reduced to 6.45 after incomplete sequence removal. However, the net charge of the complete sequences was still 3.15 greater than the training sequences on average.

The sequence similarity of each AMPd-Up-generated sequence to the training set, composed of antibacterial peptides, was calculated for analysis (details in Section 4).

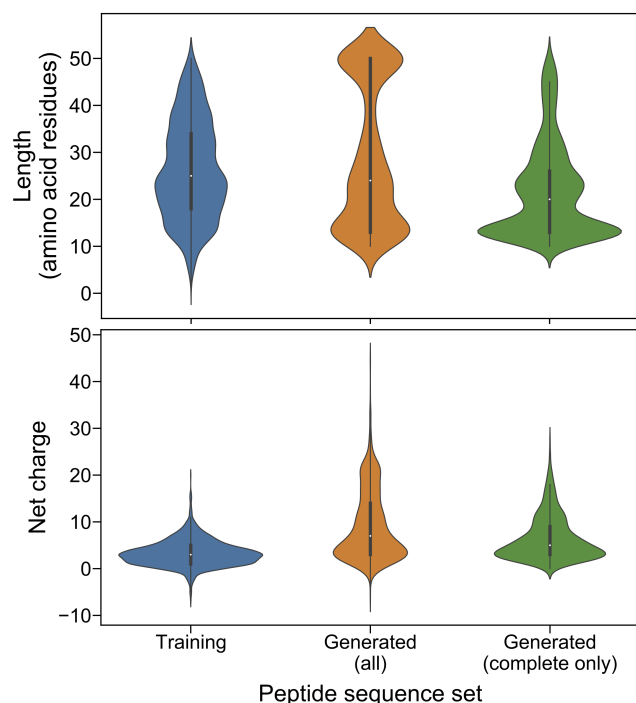


FIGURE 2 Length and net charge distributions of the sequences generated by AMPd-Up. Length and net charge distributions were calculated based on 2000 sequences generated by AMPd-Up, along with training sequences for comparison; 1484 of the 2000 generated sequences in “complete” status were chosen for an additional comparison. Mean (μ) and standard deviation (σ) of each distribution are as follows: training sequences (length: $\mu = 26.21$ aa, $\sigma = 10.34$ aa; net charge: $\mu = 3.30$, $\sigma = 2.74$), all generated sequences (length: $\mu = 28.90$ aa, $\sigma = 15.07$ aa; net charge: $\mu = 9.08$, $\sigma = 7.33$), and complete generated sequences (length: $\mu = 21.56$ aa, $\sigma = 9.87$ aa; net charge: $\mu = 6.45$, $\sigma = 4.73$).

Figure 3 shows the sequence similarity distribution of the AMPd-Up-generated sequences to the training set, with a peak between 50.00% and 55.00%. The generated sequences possess a similarity level of 49.97% compared with the training sequences on average, indicating that AMPd-Up generates novel AMP sequences different from the training sequences. This implies that AMPd-Up may be capturing high-level features of AMPs, rather than only memorizing sequence-level information during training. An additional test on the sequence similarity of each AMPd-Up-generated sequence to all available known AMPs from Antimicrobial Peptide Database (APD3, <https://aps.unmc.edu>) (Wang et al., 2016) and Database of Anuran Defense Peptides (DADP, <http://split4.pmfst.hr/dadp>) (Novković et al., 2012) was done (details in Section 4), with an average sequence similarity level of 51.03%, indicating the novelty of our generated sequences as compared with known AMPs (Figure S2 in Data S1). To supplement the sequence similarity analysis, we also visualized the pairwise sequence similarities

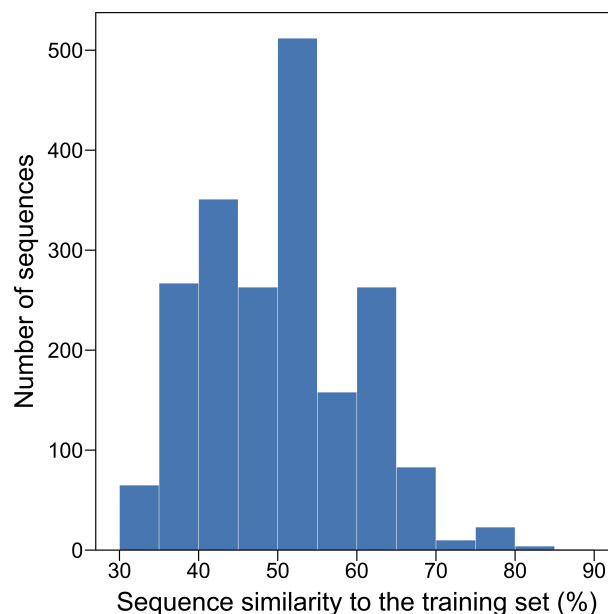


FIGURE 3 Sequence similarity distribution of the AMPd-Up-generated sequences to the training set. The sequence similarity distribution, with a mean of 49.97% and a standard deviation of 9.83%, was calculated based on the 2000 sequences generated by AMPd-Up. The sequence similarity of each generated sequence to the training set was considered as the similarity of that sequence to its most similar sequence in the training set, based on which the distribution was plotted.

between different sequence sets (Figure S3 in Data S1). A lower generated sequence similarity level between different model instances of AMPd-Up (33.56%) than within the same model instance (39.14%) indicates that different model instances tend to capture features of AMPs from slightly different aspects. We expect the novelty of generated sequences by our tool to add diversity to the current AMP sequence databases.

2.3 | *In vitro* validation results

We selected 58 peptide sequences, generated by 1000 AMPd-Up model instances, for *in vitro* validation and bioactivity assessment. We organized our candidates into three lists: List A (DeNo1001 to DeNo1038) and List B (DeNo1039 to DeNo1042) were sampled through AMPd-Up scores, and 16 more sequences that appeared with high frequencies of ≥ 40 in the generated set were selected to make List C (DeNo1043 to DeNo1058). AMPd-Up score ranges from 0 to 1 and is a measure of the confidence level of the model when generating the sequence (see Section 4 for detailed definition). Table 2 summarizes the sequence specifications of our 58 selected putative AMPs. All sequences in Lists A and C were

TABLE 2 Putative AMP sequences generated by AMPd-Up that have been prioritized for synthesis.

List name	Peptide name	Sequence	# aa	Net charge ^a	Molecular weight (Da)	AMPd-Up score ^b	AMPLify score ^c	Count ^d	Sequence similarity to known ^e (%)
A	DeNo1001	DLISGLGKAARKVAKTVLKNLLKC	24	5	2512.12	0.2362	80.00	1	45.83
	DeNo1002	NLLDTLKNLAKKLLAKKLLKLLKLL	25	8	2889.71	0.2626	80.00	1	51.61
	DeNo1003	NLLSTLLDAAKKAAGAAKSAAKKLAKKLL	33	9	3364.14	0.2679	80.00	1	48.48
	DeNo1004	HLISGLLSAAKKAAKKAALKKLLKLLKLL	33	12	3553.57	0.2891	80.00	1	45.45
	DeNo1005	GLFSLKLLKLLKLLKLLKLLKLLKLL	29	12	3431.61	0.2942	80.00	1	58.62
	DeNo1006	NLLDTLKKKAKVAKVAKVLLKLLKLLKLL	29	12	3373.40	0.3150	80.00	1	48.48
	DeNo1007	FLPSIHKGAARKLPIFKILKKC	24	7	2688.49	0.3313	80.00	1	75.00
	DeNo1008	GLLSLLKLLKLLKLLKLLKLL	21	8	2432.27	0.2949	80.00	11	66.67
	DeNo1009	DLKTLGKAARKAARTALKAALGLLKLLAKKL	33	10	3446.37	0.2467	69.24	1	45.45
	DeNo1010	VLGGLLKLLKLLKLL	17	6	1905.55	0.2520	69.24	1	58.82
	DeNo1011	HLISLLKKAARKLKLKLLKLLAKKL	25	10	2868.78	0.3049	69.24	1	48.00
	DeNo1012	CLLDTLKCVAKGVAGTLLDTLCKKITGKC	29	3	3009.73	0.2732	67.48	1	70.00
	DeNo1013	KIFGKILKLLKLLKLLKLL	22	10	2635.55	0.2995	64.47	1	54.55
	DeNo1014	ALPSLLKLLAKKLLAKKLLKLLKLLKLLKLL	33	14	3794.08	0.3100	64.47	1	48.48
	DeNo1015	NLLDTLKNVAKNVAKNVLDTLCKKITCK	29	4	3204.89	0.3346	64.47	1	65.52
	DeNo1016	FLPIAGLAARKFLPKIFCKITKCK	24	5	2664.42	0.3579	64.47	1	87.50
	DeNo1017	FLPIAGLAARKLPLKIFCKITKCK	24	5	2630.41	0.3519	60.79	1	87.50
	DeNo1018	WLPKIAGKIAGKLLKLLKLLKLLKLLKLL	24	10	2744.60	0.2573	60.21	1	50.00
	DeNo1019	FLPKIAGKAARKLPIFKITKCK	24	8	2675.45	0.3314	58.82	2	83.33
	DeNo1020	TLPDVAKNVAKNVAKTVLDTLCKKITGKC	29	4	3072.70	0.3259	58.45	1	75.86
	DeNo1021	KLFGKLLKLLKLLKLLKLLKLLKLL	25	13	2977.99	0.2309	57.62	1	44.83
	DeNo1022	GLLSLLKLLKLLKLLKLLKLLKLL	17	5	1822.38	0.2415	55.72	1	58.82
	DeNo1023	DLKTLKKAARKLKLKLLKLLKLLKLLKLL	29	11	3415.52	0.2838	53.67	1	48.28
	DeNo1024	KLFGKILGKIARKLGLKLLKLLKLLKLL	30	7	3149.05	0.2150	53.22	1	46.67
	DeNo1025	DLISCLKLLKLLKLLKLLKLLKLL	17	4	1903.41	0.1885	49.48	1	47.06
	DeNo1026	RLPSLFLKLLKLLKLLKLLKLLKLLKLL	27	11	3124.05	0.2203	48.34	1	45.71
	DeNo1027	RLPSIPIAGKLLGGLLGGLLKGL	24	3	2313.88	0.2044	43.84	1	54.17

TABLE 2 (Continued)

Peptide List name	Sequence	# aa	Net charge ^a	Molecular weight (Da)	AMPd-Up score ^b	AMPLify score ^c	Count ^d	Sequence similarity to known ^e (%)	
DeNo1028	CLPSLLPSLFKKL	13	2	1458.86	0.1713	34.88	3	53.85	
DeNo1029	SLPSILSGIAGKL	13	1	1255.51	0.1991	32.75	1	61.54	
DeNo1030	RLPRIFRGIRGKL	13	5	1581.96	0.1674	31.93	1	46.15	
DeNo1031	PLPPIPGIAGKLLGGLLGLLKKL	24	3	2392.07	0.2256	31.81	1	54.17	
DeNo1032	YLPVLPVSLKPL	13	1	1425.76	0.1779	31.78	1	53.85	
DeNo1033	PLPPIPGIAGKLLGGLLGLLKKL	18	0	1718.12	0.1938	24.72	1	61.11	
DeNo1034	KLPSIIKAAAKALPKLF	17	4	1809.30	0.2097	20.81	1	52.94	
DeNo1035	QLPRIAGKIAKKL	13	4	1435.81	0.1833	17.41	1	61.54	
DeNo1036	QLPSVLPAlAKAL	13	1	1320.63	0.1609	9.02	1	53.85	
DeNo1037	CLPSILC	7	0	747.97	0.1462	7.88	1	41.67	
DeNo1038	MLPSIAGAAAKGLPKLFCKITKKC	24	5	2490.16	0.2784	3.88	1	79.17	
B	DeNo1039	MLPKIFGKIFKKILKKILKKILKKILKKLKKL	33	14	3976.37	0.3091	1.32	1	42.42
DeNo1040	MLPSILGALLKLL	13	1	1381.82	0.2000	0.83	3	61.54	
DeNo1041	MLPKIAGKIAKKL	13	4	1410.86	0.2210	0.80	62	61.54	
DeNo1042	MLPKIAGAIKKL	13	2	1338.75	0.2080	0.77	8	61.54	
C ^f	DeNo1043	WLPKIAGKIAGKL	13	3	1394.75	0.2202	19.83	95	61.54
DeNo1044	CLPSILCKITKKC	13	3	1449.90	0.2147	35.64	90	53.85	
DeNo1045	FLPKIFKIIAKKL	13	5	1574.06	0.2659	25.78	77	61.54	
DeNo1046	VLGSLKGLLKKL	13	3	1381.80	0.2240	80.00	65	61.54	
DeNo1047	ALPSIIKGLLKKL	13	3	1393.81	0.2104	53.55	63	53.85	
DeNo1048	LLPSLLKGLLKKL	13	3	1435.89	0.2282	56.01	58	61.54	
DeNo1049	ALLSLLKLLKKL	13	4	1480.97	0.2426	69.24	56	61.54	
DeNo1050	FLPKIAGKIAGKL	13	3	1355.72	0.2400	10.81	55	69.23	
DeNo1051	ALPSLLKLLKKL	13	4	1464.93	0.2259	54.39	50	53.85	
DeNo1052	YLPVLPKGLLKKL	13	3	1471.88	0.2032	46.71	48	53.85	
DeNo1053	LLPSLLKGLAKKL	13	3	1393.81	0.2267	52.80	47	61.54	
DeNo1054	QLPKIAGKIAKKL	13	4	1407.79	0.2159	12.97	47	61.54	

(Continues)

TABLE 2 (Continued)

Peptide List name	Sequence	# aa	Net charge ^a	Molecular weight (Da)	AMPd-Up score ^b	AMPLify score ^c	Count ^d	Sequence similarity to known ^e (%)
DeNo1055	FLPKFKKIAKKI	13	5	1574.06	0.2514	39.72	46	55.56
DeNo1056	GLLSLLKLLKLL	13	4	1466.94	0.2490	80.00	43	69.23
DeNo1057	ILGKLLKLLKLL	13	5	1508.04	0.2325	51.38	40	61.54
DeNo1058	FLPKIAGKIAKKL	13	4	1426.84	0.2474	19.60	40	69.23

Note: Lists A, B, and C include 38, 4, and 16 sequences, respectively. All sequences in Lists A and C were predicted as AMPs by AMPLify (Li et al., 2022), while those in List B were predicted as non-AMPs. Sequences were sampled from all candidate peptide sequences generated by 1000 model instances, with incomplete sequences removed. Sequences in Lists A and B were sampled through AMPd-Up scores, while List C comprises a set of sequences that appeared with high frequencies (≥ 40 in sequence counts) in all candidate peptides. The numbering of peptide names for Lists A and B was by AMPLify score, while List C was by sequence count, both in descending order.

Abbreviation: AMP, antimicrobial peptide.

^aNet charge at pH = 7.

^bAMPd-Up scores range from 0 to 1; average AMPd-Up score was reported for the same sequence generated by multiple model instances.

^cAMPLify scores range from 0 to 80; sequences with AMPLify scores > 3.01 (i.e., AMPLify probability scores > 0.5) are predicted as AMPs.

^dFrequency of the sequence appearing in the generated set.

^eSequence similarity to the most similar known AMP sequence from Antimicrobial Peptide Database (APD3) (Wang et al., 2016) and Database of Anuran Defense Peptides (DADP) (Novković et al., 2012).

^fSequences with shorter lengths (≤ 20 aa) but higher net charges ($\geq +3$) were prioritized for List C.

predicted as AMPs by AMPLify, while all sequences in List B were predicted as non-AMPs.

Our 58 candidate peptides were tested against two bacterial isolates: the Gram-negative *E. coli* ATCC 25922 and the Gram-positive *S. aureus* ATCC 29213. Porcine red blood cells (RBCs) were used to assess the hemolytic activity of the peptides. Out of the 58 peptides selected for *in vitro* validation, 40 peptides displayed antimicrobial activity against at least one bacterial strain tested. All 15 peptides that were active against *S. aureus* ATCC 29213 also showed antimicrobial activity against *E. coli* ATCC 25922. Figure 4a visualizes the antimicrobial and hemolytic activities of the 40 peptides, in minimum inhibitory concentration (MIC) and concentration that lyses 50% of the RBCs (HC_{50}), respectively. The entire *in vitro* validation results of the 58 peptides are shown in Table S1 in Data S1. For a better interpretation of the results, we split the activity of the tested peptides into four levels according to the MIC/ HC_{50} ranges: high (≤ 4 $\mu\text{g/mL}$), moderate (8–16 $\mu\text{g/mL}$), low (32–128 $\mu\text{g/mL}$), and without observable activity (> 128 $\mu\text{g/mL}$).

Among the 38 List A peptides tested, 28 peptides displayed antimicrobial activity, 12 of which were active against both strains tested (Figure 4a). Nine of the List A peptides were highly active against *E. coli* ATCC 25922, with DeNo1018 being the most active with an MIC of 1–2 $\mu\text{g/mL}$. These same nine peptides were also active against *S. aureus* ATCC 29213. Four of the nine peptides were highly active against *S. aureus* ATCC 29213 (MIC = 2–4 $\mu\text{g/mL}$ for DeNo1016 and DeNo1017; MIC = 4 $\mu\text{g/mL}$ for DeNo1007 and DeNo1022), with one (DeNo1018) moderately active (MIC = 8 $\mu\text{g/mL}$). Six peptides from List A were moderately active against *E. coli* ATCC 25922, and another two showed low to moderate activity against the strain. Three of these eight peptides displayed some antimicrobial activity against *S. aureus* ATCC 29213, one of which (DeNo1031) was moderately active (MIC = 16 $\mu\text{g/mL}$) with the other two (DeNo1021 and DeNo1026) showed low (MIC = 32–64 $\mu\text{g/mL}$) and minimal activity (MIC ≥ 128 $\mu\text{g/mL}$), respectively. Among all 28 List A peptides with proven antimicrobial activity, three were minimally hemolytic ($HC_{50} \geq 128$ $\mu\text{g/mL}$) and 17 did not show any hemolytic activity ($HC_{50} > 128$ $\mu\text{g/mL}$) in our tests. DeNo1007 was the only AMP with high antimicrobial activity against both bacterial strains tested (MIC = 4 $\mu\text{g/mL}$) and without observable hemolytic activity ($HC_{50} > 128$ $\mu\text{g/mL}$).

Among the four peptides from List B tested, only DeNo1040 displayed some low-level activity against the two bacterial strains tested (Figure 4a). Specifically, this peptide inhibited the growth of *E. coli* ATCC 25922 and *S. aureus* ATCC 29213 providing MICs of 64–128 $\mu\text{g/mL}$

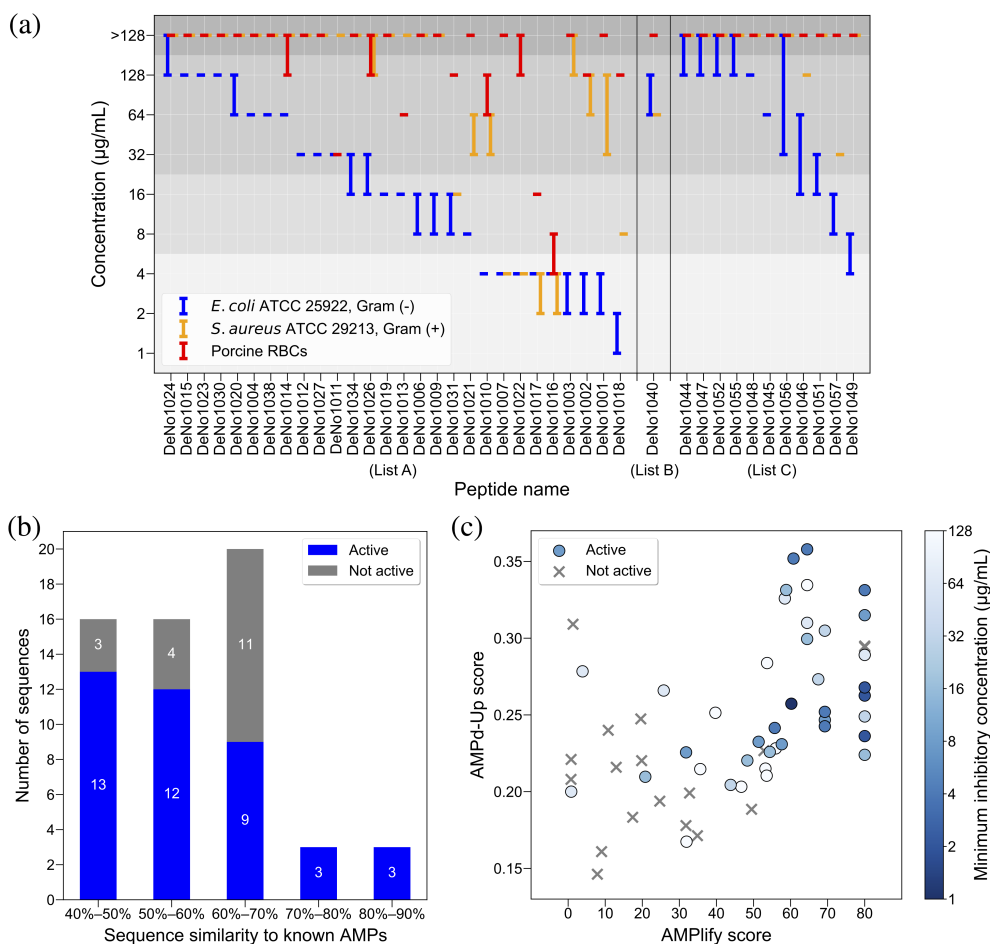


FIGURE 4 *In vitro* validation results of the 58 selected putative AMPs. (a) Antimicrobial and hemolytic activities of the 40 peptides that were active against at least one bacterial strain of *Escherichia coli* ATCC 25922 and *Staphylococcus aureus* ATCC 29213. Antimicrobial and hemolytic activities were measured by minimum inhibitory concentration (MIC) and concentration that lyses 50% (HC₅₀) of the red blood cells (RBCs), respectively. HC₅₀ was determined using porcine RBCs. Data are presented as the lowest effective peptide concentration range (μg/mL) observed in three independent experiments performed in duplicate, with one maximum data point and one minimum data point dropped for each measurement. The three sections from left to right correspond to peptides with observable antimicrobial activity from List A ($n = 28$), List B ($n = 1$), and List C ($n = 11$), respectively. Activity of the peptides was split into four levels: high (≤ 4 μg/mL), moderate (8–16 μg/mL), low (32–128 μg/mL), and without observable activity (>128 μg/mL), as separated by different background colors in the plot. (b) Stacked bar chart showing proportions of peptides that displayed antimicrobial activity with different sequence similarity levels to known AMPs from Antimicrobial Peptide Database (APD3) (Wang et al., 2016) and Database of Anuran Defense Peptides (DADP) (Novković et al., 2012). All similarity ranges are left-open and right-closed, and the sequence similarity of each candidate peptide to known AMPs was considered as the sequence similarity of that sequence to its most similar known AMP sequence. (c) Visualization of antimicrobial activity of the 58 tested peptides with respect to AMPlify (x -axis) and AMPd-Up (y -axis) scores. AMPd-Up scores of the same peptide sequences generated by multiple model instances were averaged. Peptides without any observable antimicrobial activity are presented as gray crosses, and the active peptides are presented as blue dots. Dots with darker colors indicate stronger antimicrobial activity against *Escherichia coli* ATCC 25922, determined by the lowest MIC value of each peptide against the strain. AMP, antimicrobial peptide; RBC, red blood cell.

and 64 μg/mL, respectively. DeNo1040 also did not show any hemolytic activity in our tests (HC₅₀ > 128 μg/mL). We note again that peptides in List B were predicted as non-AMPs by AMPlify.

Among the 16 List C peptides tested, a total of 11 peptides showed antimicrobial activity against *E. coli* ATCC 25922, with two of them additionally active against *S. aureus* ATCC 29213 (Figure 4a). DeNo1049 displayed

moderate to high activity against *E. coli* ATCC 25922 (MIC = 4–8 μg/mL), which was the strongest in List C. DeNo1057 was moderately antibacterial against *E. coli* ATCC 25922 (MIC = 8–16 μg/mL), followed by DeNo1051 (MIC = 16–32 μg/mL) and DeNo1046 (MIC = 16–64 μg/mL). DeNo1057 and DeNo1046 were the only two List C peptides with antibacterial activity against *S. aureus* ATCC 29213, though with low

activity (MIC = 32 $\mu\text{g}/\text{mL}$ and 128 $\mu\text{g}/\text{mL}$, respectively). None of the List C peptides displayed hemolytic activity ($\text{HC}_{50} > 128 \mu\text{g}/\text{mL}$).

Among the peptides that did not show any antimicrobial activity against the bacterial strains tested, most of them were also not hemolytic to the porcine RBCs except DeNo1008 ($\text{HC}_{50} = 16\text{--}32 \mu\text{g}/\text{mL}$) and DeNo1039 ($\text{HC}_{50} = 32\text{--}64 \mu\text{g}/\text{mL}$) as shown in Table S1 in Data S1.

In summary, List A has the largest proportion (73.68%) of putative AMPs observed with antimicrobial activity in our tests, followed by List C (68.75%) and List B (25.00%). Figure 4b presents the proportions of peptides that were active against at least one of the bacterial strains tested under different sequence similarity levels to the known AMPs from APD3 (Wang et al., 2016) and DADP (Novković et al., 2012). All six peptides between sequence similarities of 70.00% and 90.00% to known AMPs showed antimicrobial activity in our tests. The largest proportion of the tested peptides fall between sequence similarities of 60.00% and 70.00% to known AMPs, with nine out of 20 sequences displaying antimicrobial activity. Interestingly, lower similarity intervals of 50.00%–60.00% and 40.00%–50.00% possess relatively high proportions of antimicrobially active peptides with rates of 75.00% (12/16) and 81.25% (13/16), respectively. More than half (62.50%) of the peptides with antimicrobial activity from our experiments fall into these intervals, implying there is much to be explored in the sequence space for novel AMPs. Figure 4c visualizes the distribution of the 58 tested putative AMPs with regard to AMPlify scores and AMPd-Up scores. AMPlify score, ranging from 0 to 80, is a prediction score reported by AMPlify, which is a log transformation of the AMPlify probability score p_{AMPlify} as $-10\log_{10}(1 - p_{\text{AMPlify}})$. Considering the fact that multiple model instances may generate the same sequence but with different AMPd-Up scores, the average was taken in the visualization for a more comprehensive analysis. As evident in Figure 4c, most of the peptides without any observable antimicrobial activity in our tests are located at the bottom left of the figure, suggesting that it is a viable strategy to prioritize generative sequences with both high AMPlify and AMPd-Up scores for *in vitro* validation assays.

3 | DISCUSSION

In the presented work, we introduce AMPd-Up, a tool for *de novo* AMP sequence generation. AMPd-Up adopts an RNN language model, sampling from multiple model instances trained with different random initializations. AMPd-Up is available online as an open-source tool at <https://github.com/bcgsc/AMPd-Up>. Although the

architecture of our model is relatively simple compared with existing methods, we show that simple models like AMPd-Up can work well if properly trained. The simplicity of our model architecture also brings with it lower computational costs. Moreover, the sequences generated by AMPd-Up are of high novelty compared with existing AMP sequences in public databases, demonstrating the ability of our model to learn high-level AMP features.

While AMPd-Up shows great promise and favorable performance, the size of its training set is still relatively small (2253 sequences) compared with that of many traditional machine learning tasks for broader sequence data analysis, such as sentiment analysis or machine translation, which typically use hundreds of thousands to millions of data points for training available through public databases (Khurana et al., 2023). Furthermore, AMPd-Up does not take the strength of antimicrobial activities (i.e., MIC values) into consideration during training. The MIC values of an AMP against the same bacterial strain may vary due to the differences in protocols utilized across different laboratories (Schuurmans et al., 2009), thereby diminishing the comparability of those values within existing public AMP databases. We expect these limitations to be gradually resolved as the ongoing discovery and validation of AMPs is bringing more high-quality and well-organized data, leading to further improvement in *de novo* AMP sequence generation tools like AMPd-Up.

Although the AMPd-Up-generated putative AMPs have a considerable level of sequence diversity (Figure S3 in Data S1), we still noticed some patterns at the sequence level. Analyzing a set of 20,000 generated sequences, we observed that “LLKK” and “LKKL” were the two most frequently occurring 4-mer motifs, appearing in 44.09% and 40.73% of the generated sequences, respectively. Previous studies have shown that synthetic amphipathic alpha-helical peptides made up of repeat units $[\text{LLKK}]_n$ or $[\text{LKKL}]_n$ have antimicrobial properties (Khara et al., 2017; Wiradharma et al., 2011), which can explain these findings to some extent. In fact, it is suggested that repeats of 4-mer units such as these are responsible for the formation of cationic amphipathic alpha-helical structures, a key initiating step to the bioactivity and membrane-disrupting properties of many AMPs (Khara et al., 2017; Wiradharma et al., 2011).

Among the 58 novel putative AMP sequences generated by AMPd-Up, 40 showed antimicrobial activity, 15 of which were broadly antibacterial against both Gram-positive and Gram-negative isolates. Promisingly, one of the most active peptides, DeNo1007, not only possessed high antimicrobial activity against the two bacterial strains tested, but was also without observable hemolytic activity. We expect the AMP candidates

generated by AMPd-Up to increase the diversity of known peptide-derived antibiotics, currently populated by mostly naturally occurring sequences, and to augment the candidate set of potential alternatives to conventional antibiotics. Although some of our putative AMPs did not show any antimicrobial activity against the two bacterial strains tested *in vitro*, they may still be active against other bacterial species and/or possess unexplored modes of action. Also, the structures of some AMPs may vary based on their microenvironment (Cândido et al., 2019). Further experimentation could be done to test candidate sequences on a wider panel of bacterial species, to investigate the variances in their antimicrobial mechanisms against bacteria with different membrane and cell wall structures (e.g., Gram-positive vs. Gram-negative bacteria), or to interrogate *in vivo* biological interactions.

Results from work like ours also have broader potential impact. Resistance to last-line peptide-based therapeutics, such as colistin and other polymyxins, is increasingly being reported (Aghapour et al., 2019). Concerningly, this is sometimes presented with cross-resistance to multiple AMPs (Fleitas and Franco, 2016), highlighting the need for multiple and diverse classes of peptide-based antimicrobials. *De novo* AMP sequence generation provides a rational solution to this problem, as one would theoretically expect that pathogens would be naïve to many of the diverse *de novo* generated AMPs. Even though there may be natural AMPs similar to some of the *de novo* generated ones, the vast sequence space of amino acids (e.g., 10^{20} or one hundred quintillion for a 10-residue peptide sequence) virtually ensures that there would be a practically infinite number of them out there that are “new” to most common pathogens. Thus, we expect high-throughput *in silico* AMP sequence design tools like AMPd-Up to play a vital role in the fight against antibiotic resistance and the imminent rise of antibiotic-resistant bacteria.

4 | MATERIALS AND METHODS

4.1 | Training set

To get our RNN language model well trained, a curated set of known AMP sequences are required to comprise the training set. Our work primarily focused on AMPs with direct antibacterial activity, a major function of most known AMPs. We also limited the generated AMP sequences to include only standard amino acids with a maximum length of 50 aa, reflecting the fact that most documented AMPs are relatively short (Zhang and Gallo, 2016).

All antibacterial peptide sequences were downloaded from APD3 (Wang et al., 2016) on March 20, 2019, a manually curated and annotated database for AMPs. This set of sequences contained 2571 AMP records with antibacterial activity, 2276 of which were ≤ 50 aa long. After removing duplicates and sequences with non-standard amino acids, we ended up with a non-redundant set of 2253 antibacterial sequences ≤ 50 aa in length, forming the training set for our RNN language model.

4.2 | Model architecture and implementation

The implementation of the RNN language model was adapted from the PyTorch online tutorial by Sean Robertson (Robertson, 2017), with PyTorch library 1.7.1 (Paszke et al., 2019) in Python 3.6.7. During the training process, cross-entropy was used as the loss function, and stochastic gradient descent (Robbins and Monro, 1951) was applied to optimize the model weights. We also adopted dropout technique (Srivastava et al., 2014) to prevent overfitting. The hyperparameters, which cannot be learned directly from training, were tuned through stratified five-fold cross-validation on the training set. The set of hyperparameters for model architecture and training settings with the lowest average cross-validation loss was determined to be the optimal one to train the final model.

Figure 1 shows the architecture of the RNN language model, represented as a chain of repeating RNN cells. Given the first N-terminal amino acid, the RNN language model generates a peptide sequence residue by residue until reaching the EOS signal. In this specific task of AMP sequence generation, we set the maximum length to be 50 and only the 20 standard amino acids are considered. Amino acids, together with the EOS signal, are encoded as 21 distinct one-hot vectors, with $\mathbf{x}_t \in \mathbb{R}^{21}$ representing the t -th residue of a generated sequence. In this task, a time step t is defined as the process of an RNN cell predicting the $(t+1)$ -th residue \mathbf{x}_{t+1} of a sequence. At each time step t of the generation process, the RNN cell takes the hidden state \mathbf{h}_{t-1} from the previous time step and the predicted amino acid for the t -th residue \mathbf{x}_t as input, and outputs a set of probabilities \mathbf{p}_t of amino acid and EOS occurrence at the next position, from which \mathbf{x}_{t+1} can be sampled. The hidden state $\mathbf{h}_t \in \mathbb{R}^{d_h}$ and probability vector $\mathbf{p}_t \in \mathbb{R}^{21}$ at each time step are calculated as:

$$\mathbf{h}_t = W_h \begin{bmatrix} \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{bmatrix} + \mathbf{b}_h,$$

$$\mathbf{p}_t = \text{softmax} \left(W_p \left[W_o \begin{bmatrix} \mathbf{h}_t \\ \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{bmatrix} + \mathbf{b}_o \right] + \mathbf{b}_p \right),$$

where $W_h \in \mathbb{R}^{d_h \times (d_h+21)}$, $W_o \in \mathbb{R}^{21 \times (d_h+21)}$, and $W_p \in \mathbb{R}^{21 \times (d_h+21)}$ are weight matrices, and $\mathbf{h}_h \in \mathbb{R}^{d_h}$, $\mathbf{b}_o \in \mathbb{R}^{21}$, and $\mathbf{b}_p \in \mathbb{R}^{21}$ are bias vectors. Here, $\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$ denotes the concatenation of two vectors \mathbf{v}_1 and \mathbf{v}_2 , and the softmax function ensures that the probabilities sum up to 1. The initial hidden state \mathbf{h}_0 is set to be a zero vector. We found the best tuned d_h to be 128. A dropout rate of 0.1 was applied before the softmax function during training, and the training process was conducted with 100,000 iterations and a learning rate of 0.0005.

Predictions can be made by sampling from the output probabilities of the RNN cells. The sequence generation process stops if an EOS signal is predicted or if the maximum length is reached without EOS signal predicted. Sequences generated in the former case are annotated as “complete”, while those in the latter case as “incomplete”. AMPd-Up computes a confidence score when generating each sequence. The score is calculated as the geometric mean of probabilities of all predicted symbols in a sequence, including the EOS signal if the sequence is complete. We refer to this score as the “AMPd-Up score”, and we use it as a measure of confidence of the RNN language model in generating a sequence. In AMPd-Up, the model is trained multiple times with different random initializations, yielding multiple model instances.

Given one of the 20 possible starting amino acids, the symbol with the highest probability estimated at each time step is taken as the next amino acid prediction (including the EOS signal), resulting in a maximum of 20 candidate AMP sequences generated by a single model instance. In a practical use case, the model will be trained k times and the users would get a candidate AMP list of up to $20k$ sequences. Assuming we have a non-convex loss function like most neural network based tasks, different initializations may result in different trained models (Fort et al., 2019), allowing different model instances of AMPd-Up to capture slightly different aspects of the complex but unknown features of AMPs.

4.3 | Model evaluation

In order to measure the performance of AMPd-Up in an efficient and cost-effective way, we used the predictions from three state-of-the-art *in silico* AMP prediction tools: AMPlify (Li et al., 2022), AMP Scanner Vr.2 (Veltri

et al., 2018), and iAMPpred (Meher et al., 2017), as a proxy for AMP sequence generation accuracy. These AMP prediction tools determine whether an input peptide sequence is an AMP or not. Here, the estimated AMP sequence generation accuracy measured by a selected prediction tool was calculated based on the percentage of peptide sequences predicted as AMPs among a generated sequence set. A default setting of balanced model was chosen for AMPlify (v1.1.0) as described in a data note (Li et al., 2023), while the “original production model” was chosen for AMP Scanner Vr.2 on its online server (Veltri et al., 2018). Predictions by iAMPpred were obtained through its online server with its trained model as described in the publication (Meher et al., 2017).

We compared AMPd-Up with three other AMP sequence generation methods with publicly available models or generated sequences: the LSTM language model (Nagarajan et al., 2018), AMPGAN v2 (Van Oort et al., 2021), and HydrAMP (Szymczak et al., 2022). For each method, a total of 2000 sequences were generated for comparison in five batches. This resulted in five generated sequence sets of 400 sequences for each method. Sequences for the LSTM language model were sampled from the dataset the authors provided (Nagarajan et al., 2018), while those for HydrAMP were obtained through their online server (Szymczak et al., 2022) on November 7, 2022. While all other methods focus on the generation of antibacterial peptides, AMPGAN v2 additionally allows for generating AMP sequences of other function types (e.g., antifungal, antiviral) and the generated sequences are annotated with their predicted functions in the results (Van Oort et al., 2021). For a fairer comparison, only AMPs targeting bacteria were selected for AMPGAN v2. For each AMP sequence generation method measured by each AMP prediction tool, the average estimated accuracy value of the five generated sets was reported, along with the corresponding standard deviation value.

In addition to the estimated sequence generation accuracy, we evaluated the sequences generated by AMPd-Up based on their amino acid compositions, physicochemical properties, as well as their sequence similarities to the training set and all publicly available known AMPs. The same 2000 sequences generated by AMPd-Up for performance comparison were used in these analyses.

The properties that cause a peptide sequence to have antimicrobial activity are complex and the mechanisms are still not well understood (Teimouri et al., 2021). Considering the fact that most known AMPs share common characteristics of short lengths and net positive charges (Zhang and Gallo, 2016), we focused on these two important and easy-to-calculate physicochemical properties in addition to an amino acid composition analysis.

Moreover, sequence similarities of the AMPd-Up-generated sequences to the training set were calculated to evaluate whether the model instances capture high-level features of AMPs rather than only generating the same or highly similar sequences to the training set. A similar comparison between the AMPd-Up-generated sequences and all publicly available known AMPs was done to evaluate the novelty of the generated sequences compared with those known AMP sequences. We note that the training AMPs are antibacterial, while the known AMP sequence set additionally includes those targeting microbes other than bacteria. The known AMP sequence set comprises 4538 distinct sequences that were downloaded from APD3 (Wang et al., 2016) and DADP (Novković et al., 2012) on July 11, 2022 and December 6, 2018, respectively. The similarity between two sequences was calculated as $\left(1 - \frac{d_{ij}}{\max(l_i, l_j)}\right) \times 100\%$, where d_{ij} is the edit distance and l_i, l_j are lengths of the sequences regarding the numbers of amino acid residues. The similarity of a sequence to a set of sequences was defined as the maximum of all similarity values calculated between that sequence and the sequences in the target set for comparison (i.e., the similarity of that sequence to its most similar sequence in the target set).

4.4 | Selecting putative AMPs for validation

To demonstrate the utility of our tool, we trained the model 1000 times, yielding 1000 model instances and 20,000 sequences, 14,188 of which were complete, and 8737 of the complete sequences were distinct. The trained models applied to generate the sequences for validation can be accessed at <https://doi.org/10.5281/zenodo.7905591> (Li and Birol, 2023). We define the “count” of a sequence as the number of times it appears in the entire generated set. We further filtered for short sequences with lengths ≤ 35 aa and obtained 7434 peptide sequences, since shorter peptides are more cost-effective for synthesis (Lin et al., 2022). We selected 58 of these peptides using different strategies (forming Lists A, B, and C), and validated their bioactivity through *in vitro* experiments (Table 2).

The peptides comprising Lists A and B were chosen following a strategy that stratifies the AMPd-Up score range of 7434 sequences into same-length score intervals. For n intervals, each interval can be written as a range from $a + \frac{(k-1)(b-a)}{n}$ to $a + \frac{k(b-a)}{n}$, with $k = 1, 2, \dots, n$ and a, b being the minimum and maximum AMPd-Up scores investigated in the generated sequence set. In our case,

$a = 0.1462$ and $b = 0.3579$. All intervals are left-open and right-closed, except the first one ($k = 1$) that is closed. If multiple model instances generated the same sequence, the AMPd-Up score from the first model that generated this sequence was used for stratification. Peptides for List A were sampled by splitting the AMPd-Up score range of $[0.1462, 0.3579]$ into 40 intervals, and the sequence with top AMPd-Up score within each interval was chosen. List B peptides were chosen by splitting the same AMPd-Up score range into five intervals, and then selecting one predicted non-AMP (as assessed by AMPLify) in each interval with the highest count, or with top AMPd-Up score if all sequences have the same count in the interval. We note that some intervals did not have any sequences, resulting in 38 sequences in List A and 4 sequences in List B. Additionally, 16 more peptide sequences that appeared with high frequencies (≥ 40 in sequence counts) in the generated set were selected as List C. All sequences in Lists A and C were predicted as AMPs by AMPLify. In Table 2, we also present the sequence similarity of each sequence to the known AMPs, showing the novelty of those sequences compared with the known AMP sequences.

4.5 | Antimicrobial susceptibility testing

The antimicrobial activity of our selected peptides was measured in the laboratory by broth microdilution assays to determine the minimum inhibitory and minimum bactericidal concentrations (MICs and MBCs, respectively) as outlined by the Clinical and Laboratory Standards Institute (CLSI) (Clinical and Laboratory Standards Institute, 2015) with some adaptations for testing cationic AMPs as described previously (Wiegand et al., 2008). Laboratory isolates of *E. coli* 25922 and *S. aureus* 29213 were purchased from the American Type Culture Collection (ATCC; Manassas, VA, USA) and were used to test the 58 selected putative AMPs. Bacteria from frozen stocks were streaked onto non-selective Columbia blood agar with 5% sheep blood (Oxoid) and incubated for 18–24 h at 37°C. The following day, 2–4 colonies were streaked onto a new agar plate and incubated for 18–24 h at 37°C to ensure uniform colony health prior to the assay. A standardized bacterial inoculum was prepared by suspending isolated colonies in Mueller-Hinton Broth (MHB; Sigma-Aldrich, St. Louis, MO, USA). The suspension was adjusted to an optical density of 0.08–0.1 at 600 nm, equivalent to a 0.5 McFarland standard of approximately $1-2 \times 10^8$ CFU/mL (CFU: colony forming units). The inoculum was then diluted 1:250 to achieve a final concentration of $5 \pm 3 \times 10^5$ CFU/mL. The target bacterial density was

confirmed by examining the total viability counts from the final inoculum.

Candidate AMPs were purchased from and synthesized by GenScript (Piscataway, NJ, USA). These were received in lyophilized format and stored at -20°C , and were suspended in sterile ultrapure water prior to testing. A two-fold serial dilution of 1280 down to $2.5\ \mu\text{g}/\text{mL}$ was prepared in sterile 96-well polypropylene microtiter plates (Greiner Bio-One #650261, Kremsmünster, Austria) before the addition of $100\ \mu\text{L}$ of the standardized bacterial inoculum, providing a final AMP testing range of 128 down to $0.25\ \mu\text{g}/\text{mL}$. The MIC values were reported as the lowest peptide concentration where no visible bacterial growth was observed following a 20–24 h incubation at 37°C . For determination of MBC, well contents of the MIC and the two adjacent wells containing the two- and four-fold higher peptide concentrations were plated onto non-selective nutrient agar. The concentration in which 99.9% of the inoculum were killed after an incubation for 24 h at 37°C was reported as the MBC.

A known AMP Ranatuerin-4 (Goraya et al., 1998) from the American bullfrog and an in-house peptide [TKPKG]₃ (OT15) were used as the positive and negative control peptides, respectively. We note that OT15 was truncated and derived from a negative control peptide [TKPKG]₄ (OT20), which, while not antimicrobial, shares similar characteristics with AMPs and has been used in previous studies (Horváti et al., 2017).

4.6 | Hemolysis assay

The toxicity of the selected peptides to RBCs was evaluated by hemolysis experiments. Whole blood from healthy donor pigs was purchased from Lampire Biological Laboratories (Pipersville, PA, USA). RBCs were washed and isolated by centrifugation using Roswell Park Memorial Institute (RPMI) medium (Life Technologies, Grand Island, NY, USA). All centrifugation steps were performed at $500\times g$ for 5 min in an Allegra-6R centrifuge (Beckman Coulter, CA, USA). Peptides were suspended and serially diluted from 1280 down to $10\ \mu\text{g}/\text{mL}$ using RPMI medium in a 96-well polypropylene microtiter plate, and then they were combined with $100\ \mu\text{L}$ of the 1% RBC solution. This resulted in a final AMP testing range of 128 down to $1\ \mu\text{g}/\text{mL}$. Following an incubation at 37°C for 30–45 min, plates were centrifuged and a 1/2 volume from each supernatant was transferred to a new 96-well plate. The absorbance of the wells was measured at 415 nm utilizing the Cytation 5 Cell Imaging Multimode Reader (BioTek, CA, USA); the peptide concentration that lysed 50% of the RBCs (HC_{50}) was used to

report the hemolytic activity. Absorbance readings from wells containing RBCs treated with $11\ \mu\text{L}$ of a 2% Triton-X100 solution or RPMI medium (AMP solvent-only) were used to define 100% and 0% hemolysis, respectively.

AUTHOR CONTRIBUTIONS

Chenkai Li: Conceptualization; writing – original draft; methodology; software; investigation; data curation; visualization; formal analysis; writing – review and editing; validation. **Darcy Sutherland:** Formal analysis; validation; investigation; writing – review and editing; methodology. **Amelia Richter:** Formal analysis; validation; investigation; writing – review and editing; methodology. **Lauren Coombe:** Formal analysis; writing – review and editing. **Anat Yanai:** Formal analysis; validation; investigation; writing – review and editing; methodology. **René L. Warren:** Formal analysis; investigation; writing – review and editing. **Monica Kotkoff:** Project administration; writing – review and editing. **Fraser Hof:** Conceptualization; funding acquisition; writing – review and editing; supervision. **Linda M. N. Hoang:** Conceptualization; funding acquisition; writing – review and editing; supervision. **Caren C. Helbing:** Conceptualization; funding acquisition; writing – review and editing; supervision. **Inanc Birol:** Conceptualization; funding acquisition; supervision; methodology; software; formal analysis; investigation; writing – review and editing.

ACKNOWLEDGMENTS

This work was supported by Genome BC and Genome Canada [291PEP]. The content of this paper is solely the responsibility of the authors, and does not necessarily represent the official views of our funding organizations. Additional support was provided by the Canadian Agricultural Partnership, a federal-provincial-territorial initiative, under the Canada-BC Agri-Innovation Program. The program is delivered by the Investment Agriculture Foundation of BC. Opinions expressed in this document are those of the authors and not necessarily those of the Governments of Canada and British Columbia or the Investment Agriculture Foundation of BC. The Governments of Canada and British Columbia, and the Investment Agriculture Foundation of BC, and their directors, agents, employees, or contractors will not be liable for any claims, damages, or losses of any kind whatsoever arising out of the use of, or reliance upon, this information.

CONFLICT OF INTEREST STATEMENT

IB is a co-founder of and executive at Amphorax Life Sciences Inc.

ORCID

Chenkai Li  <https://orcid.org/0000-0002-8748-0099>

REFERENCES

- Aghapour Z, Gholizadeh P, Ganbarov K, Bialvaei AZ, Mahmood SS, Tanomand A, et al. Molecular mechanisms related to colistin resistance in Enterobacteriaceae. *Infect Drug Resist.* 2019;12:965–75. <https://doi.org/10.2147/IDR.S199844>
- Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet.* 2022;399(10325):629–55. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0)
- Boman HG. Antibacterial peptides: basic facts and emerging concepts. *J Intern Med.* 2003;254(3):197–215. <https://doi.org/10.1046/j.1365-2796.2003.01228.x>
- Cândido ES, Cardoso MH, Chan LY, Torres MDT, Oshiro KGN, Porto WF, et al. Short cationic peptide derived from Archaea with dual antibacterial properties and anti-infective potential. *ACS Infect Dis.* 2019;5(7):1081–6. <https://doi.org/10.1021/acsinfectdis.9b00073>
- Clinical and Laboratory Standards Institute. *Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: approved standard.* Wayne, PA: Clinical and Laboratory Standards Institute; 2015.
- Das P, Sercu T, Wadhawan K, Padhi I, Gehrmann S, Cipcigan F, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng.* 2021;5(6):613–23. <https://doi.org/10.1038/s41551-021-00689-x>
- Dean SN, Alvarez JAE, Zabetakis D, Walper SA, Malanoski AP. PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Front Microbiol.* 2021;12:725727. <https://doi.org/10.3389/fmicb.2021.725727>
- Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. In: 5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings. 2017.
- Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, et al. Adversarially learned inference. In: 5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings. 2017.
- Fleitas O, Franco OL. Induced bacterial cross-resistance toward host antimicrobial peptides: a worrying phenomenon. *Front Microbiol.* 2016;7:381. <https://doi.org/10.3389/fmicb.2016.00381>
- Fort S, Hu H, Lakshminarayanan B. Deep ensembles: a loss landscape perspective. *arXiv.* 2019. <https://doi.org/10.48550/arXiv.1912.02757>
- Gagnon M-C, Strandberg E, Grau-Campistany A, Wadhvani P, Reichert J, Bürck J, et al. Influence of the length and charge on the activity of α -helical amphipathic antimicrobial peptides. *Biochemistry.* 2017;56(11):1680–95. <https://doi.org/10.1021/acs.biochem.6b01071>
- Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems.* 2014.
- Goraya J, Knoop FC, Conlon JM. Ranatuerins: antimicrobial peptides isolated from the skin of the American bullfrog, *Rana catesbeiana*. *Biochem Biophys Res Commun.* 1998;250(3):589–92. <https://doi.org/10.1006/bbrc.1998.9362>
- Guo J, Lu S, Cai H, Zhang W, Yu Y, Wang J. Long text generation via adversarial training with leaked information. In: *Proceedings of the AAAI Conference on Artificial Intelligence, Long Text Generation via Adversarial Training with Leaked Information.* 2018.
- Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. *Nat Mach Intell.* 2019;1(2):105–11. <https://doi.org/10.1038/s42256-019-0017-4>
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Horvati K, Bacsa B, Mlinko T, Szabo N, Hudecz F, Zsila F, et al. Comparative analysis of internalisation, haemolytic, cytotoxic and antibacterial effect of membrane-active cationic peptides: aspects of experimental setup. *Amino Acids.* 2017;49(6):1053–67. <https://doi.org/10.1007/s00726-017-2402-9>
- Huan Y, Kong Q, Mou H, Yi H. Antimicrobial peptides: classification, design, application and research progress in multiple fields. *Front Microbiol.* 2020;11:582779. <https://doi.org/10.3389/fmicb.2020.582779>
- Jukic M, Bren U. Machine learning in antibacterial drug design. *Front Pharmacol.* 2022;13:864412. <https://doi.org/10.3389/fphar.2022.864412>
- Khara JS, Obuobi S, Wang Y, Hamilton MS, Robertson BD, Newton SM, et al. Disruption of drug-resistant biofilms using de novo designed short α -helical antimicrobial peptides with idealized facial amphiphilicity. *Acta Biomater.* 2017;57:103–14. <https://doi.org/10.1016/j.actbio.2017.04.032>
- Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl.* 2023;82(3):3713–44. <https://doi.org/10.1007/s11042-022-13428-4>
- Kingma DP, Welling M. Auto-encoding variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 – Conference Track Proceedings. 2014.
- Koo HB, Seo J. Antimicrobial peptides under clinical investigation. *Pept Sci.* 2019;111(5):e24122. <https://doi.org/10.1002/pep2.24122>
- Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance—the need for global solutions. *Lancet Infect Dis.* 2013;13(12):1057–98. [https://doi.org/10.1016/S1473-3099\(13\)70318-9](https://doi.org/10.1016/S1473-3099(13)70318-9)
- Li C, Birol I. Model files of AMPd-up: a tool for antimicrobial peptide sequence generation. *Zenodo.* 2023. <https://doi.org/10.5281/zenodo.7905591>
- Li C, Sutherland D, Hammond SA, Yang C, Taho F, Bergman L, et al. AMPLify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics.* 2022;23:77. <https://doi.org/10.1186/s12864-022-08310-4>
- Li C, Warren RL, Birol I. Models and data of AMPLify: a deep learning tool for antimicrobial peptide prediction. *BMC Res Notes.* 2023;16:11. <https://doi.org/10.1186/s13104-023-06279-1>
- Lin D, Sutherland D, Aninta SI, Louie N, Nip KM, Li C, et al. Mining amphibian and insect transcriptomes for antimicrobial peptide sequences with rAMPage. *Antibiotics (Basel).* 2022;11(7):952. <https://doi.org/10.3390/antibiotics11070952>
- Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep.* 2017;7:42362. <https://doi.org/10.1038/srep42362>
- Mikolov T, Karafiat M, Burget L, ˇCernocky J, Khudanpur S. Recurrent neural network based language model. In: *Proceedings of*

- the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. p. 1045–8. 2010.
- Nagarajan D, Nagarajan T, Roy N, Kulkarni O, Ravichandran S, Mishra M, et al. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J Biol Chem*. 2018;293(10):3492–509. <https://doi.org/10.1074/jbc.M117.805499>
- Nguyen LT, Haney EF, Vogel HJ. The expanding scope of antimicrobial peptide structures and their modes of action. *Trends Biotechnol*. 2011;29(9):464–72. <https://doi.org/10.1016/j.tibtech.2011.05.001>
- Novković M, Simunić J, Bojović V, Tossi A, Juretić D. DADP: the database of anuran defense peptides. *Bioinformatics*. 2012;28(10):1406–7. <https://doi.org/10.1093/bioinformatics/bts141>
- O'Neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. The review on antimicrobial resistance [accessed 2022 Oct 31]. 2014 https://amr-review.org/sites/default/files/AMR%20Review%20Paper%20-%20Tackling%20a%20crisis%20for%20the%20health%20and%20wealth%20of%20nations_1.pdf
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*. 2019.
- Reardon S. Antibiotic resistance sweeping developing world. *Nature*. 2014;509(7499):141–2. <https://doi.org/10.1038/509141a>
- Richter A, Sutherland D, Ebrahimikondori H, Babcock A, Louie N, Li C, et al. Associating biological activity and predicted structure of antimicrobial peptides from amphibians and insects. *Antibiotics (Basel)*. 2022;11(12):1710. <https://doi.org/10.3390/antibiotics11121710>
- Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat*. 1951;22(3):400–7. <https://doi.org/10.1214/aoms/1177729586>
- Robertson S. NLP from scratch: generating names with a character-level RNN. [accessed 2021 Mar 15]. 2017 https://pytorch.org/tutorials/intermediate/char_rnn_generation_tutorial.html#nlp-from-scratch-generating-names-with-a-character-level-rnn
- Schuermans JM, Nuri Hayali AS, Koenders BB, ter Kuile BH. Variations in MIC value caused by differences in experimental protocol. *J Microbiol Methods*. 2009;79(1):44–7. <https://doi.org/10.1016/j.mimet.2009.07.017>
- Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: *Advances in Neural Information Processing Systems*. 2015.
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–58.
- Szymczak P, Możejko M, Grzegorzec T, Bauer M, Neubauer D, Michalski M, et al. HydrAMP: a deep generative model for antimicrobial peptide discovery. *bioRxiv*. 2022; <https://www.biorxiv.org/content/10.1101/2022.01.27.478054v1>
- Teimouri H, Nguyen TN, Kolomeisky AB. Single-cell stochastic modelling of the action of antimicrobial peptides on bacteria. *J R Soc Interface*. 2021;18(182):20210392. <https://doi.org/10.1098/rsif.2021.0392>
- Tossi A. Design and engineering strategies for synthetic antimicrobial peptides. In: Drider D, Rebuffat S, editors. *Prokaryotic antimicrobial peptides*. New York, NY: Springer; 2011. p. 81–98.
- Tucs A, Tran DP, Yumoto A, Ito Y, Uzawa T, Tsuda K. Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks. *ACS Omega*. 2020;5(36):22847–51. <https://doi.org/10.1021/acsomega.0c02088>
- van der Does AM, Hiemstra PS, Mookherjee N. Antimicrobial host defence peptides: immunomodulatory functions and translational prospects. In: Matsuzaki K, editor. *Antimicrobial peptides*. Advances in experimental medicine and biology. Singapore: Springer; 2019. p. 149–71.
- Van Oort CM, Ferrell JB, Remington JM, Wshah S, Li J. AMPGAN v2: machine learning-guided design of antimicrobial peptides. *J Chem Inf Model*. 2021;61(5):2198–207. <https://doi.org/10.1021/acs.jcim.0c01441>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. 2017.
- Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*. 2018;34(16):2740–7. <https://doi.org/10.1093/bioinformatics/bty179>
- Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res*. 2016;44(D1):D1087–93. <https://doi.org/10.1093/nar/gkv1278>
- Wiegand I, Hilpert K, Hancock REW. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc*. 2008;3(2):163–75. <https://doi.org/10.1038/nprot.2007.521>
- Wiradharma N, Khoe U, Hauser CAE, Seow SV, Zhang S, Yang Y-Y. Synthetic cationic amphiphilic α -helical peptides as antimicrobial agents. *Biomaterials*. 2011;32(8):2204–12. <https://doi.org/10.1016/j.biomaterials.2010.11.054>
- Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal Biochem*. 2013;436(2):168–77. <https://doi.org/10.1016/j.ab.2013.01.019>
- Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. p. 1480–9. 2016.
- Zelezetsky I, Tossi A. Alpha-helical antimicrobial peptides—using a sequence template to guide structure–activity relationship studies. *Biochim Biophys Acta Biomembr*. 2006;1758(9):1436–49. <https://doi.org/10.1016/j.bbmem.2006.03.021>
- Zhang L-J, Gallo RL. Antimicrobial peptides. *Curr Biol*. 2016;26(1):R14–9. <https://doi.org/10.1016/j.cub.2015.11.017>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Li C, Sutherland D, Richter A, Coombe L, Yanai A, Warren RL, et al. *De novo* synthetic antimicrobial peptide design with a recurrent neural network. *Protein Science*. 2024;33(8):e5088. <https://doi.org/10.1002/pro.5088>