A Generative Approach Toward Precision Antimicrobial Peptide Design

Jonathon B. Ferrell,[1#] Jacob M. Remington,[1#] Colin M. Van Oort,[2#] Mona Sharafi,[1] Reem Aboushousha,[3] Yvonne Janssen-Heininger,[3] Severin T. Schneebeli,[1] Matthew J. Wargo,[4] Safwan Wshah,[2] Jianing Li [1*]

[1]Department of Chemistry, University of Vermont, Burlington, Vermont 05405

[2]Department of Computer Science, University of Vermont, Burlington, Vermont 05405

[3]Department of Pathology, University of Vermont, Burlington, Vermont 05405

[4]Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405

Corresponding Author: Jianing Li (jianing.li@uvm.edu)

# These authors contributed equally to this work.

Abstract:

Antimicrobial peptides (AMPs) are peptides with promising applications for healthcare, veterinary, and agriculture industries. Despite prior success in AMP design using physics- or knowledge-based approaches, there is still a critical need to create new methodologies to design peptides with a low false positive rate and high AMP activity and selectivity. Toward this goal, we invented a cost-effective approach which utilizes a generative model to produce AMP-like sequences and molecular simulations to select peptides based on their structures and interactions. For a proof of concept, we curated a dataset that comprises 500,000 non-AMP peptide sequences and 8,000 labeled AMP sequences to train the generative model, which generated novel and diverse AMP candidates to potentially target a wide variety of microbes. Following a screening process to select peptides that are cationic and likely helical, we assessed 12 candidates by simulating their membrane-binding tendency to a lipid bilayer model. With the umbrella sampling technique, we determined the free energy change during transfer from the solution to the membrane environments for each peptide. Accordingly, we selected the six peptides with the best membrane-binding tendency, synthesized them, and characterized through spectroscopies and biological assays. Three novel peptides were validated with activity to inhibit bacterial growth. In aggregate, the combination of AMP generator and molecular simulations afford an enhanced accuracy in AMP design. Towards future precision AMP design, our methodology and results demonstrate the viability to design novel AMP-like peptides to target selected pathogens and mechanisms.

Key words:
Generative Neural Network, Machine Learning, Molecular Simulation, Helicity, Toxicity

With advantages over traditional small-molecule antibiotics (e.g. broad-spectrum activity, rapid onset of killing, low levels of induced resistance, etc.), antimicrobial peptides (AMPs) show promise for treating infectious diseases (1,2) caused by bacteria, viruses, or fungi, which are detrimental to our society and the healthcare, veterinary, and agriculture industries. Unlike traditional small-molecule antibiotics, AMPs are biological polymers predominantly containing 4 to 40 natural amino acids. So far, over 8000 AMPs (3) have been found in animals and plants as a first-line defense of the host immune systems. However, the known AMPs represent only a small fraction of the vast chemical space — e.g. 20^N possible chemical compositions for a N-residue peptide composed of 20 natural amino acids. Recent methods have focused on quantitative structure activity relationship (QSAR) and machine learning (ML) predictions in order to seek new AMPs (see recent reviews (2,4)). Such approaches often rely on sequences to be generated randomly or evolved from known sequences (5), which can be computationally demanding to afford reasonable sampling in the vast peptide sequence space. This, combined with the fact that QSAR models often include a large number of chemical descriptors (which are chosen from a priori knowledge), makes is inefficient to expand into unknown chemical space that may not be well described by those descriptors. Thus implying, while QSAR and similar models can be powerful for finding novel sequences with similar properties, their poor sampling outside of their descriptors requires entirely new models to be created in order to explore new chemical space; leading to an overall inadequately robust method of exploring uncharacterized sequence space. Instead it is more efficient to utilize a generative model such as a variational autoencoder (VAE) or a generative-adversarial network (GAN), which generates new sequences from the underlying distribution of possible AMPs. However, prior studies based on the VAE model (6) suggest that a low false positive rate and high AMP activity still remains to be achieved. Using a fundamentally distinct GAN approach (7), in this work we generated plausible AMP sequences and selected six potent candidates based on molecular modeling for experimental validation, a first for such methodology. Three of the six AMPs were capable of inhibiting bacterial growth, indicating the potential of our methodology to be applied for AMP discovery.

The conditional generative adversarial network (CGAN) (8) represents a powerful approach for creating generative models, as it gives instructions to two networks that are pitted against each other in a zero-sum game. We utilized this framework and a set of only four fundamental conditions (or labels) to generate new AMP sequences. The discriminator was trained to distinguish between authentic and generated sequences, and the generator was tasked with generating sequences that fool the discriminator. In order to trick the discriminator successfully, the generator must learn the underlying AMP distribution. By generating sequences from the underlying AMP distribution and providing conditioning variables that allow the generation process to be directed, we obtained control over the properties of generated sequences. Furthermore, the conditional vectors provided a distinct capability to target particular pathogens, mechanisms, and even sequence lengths. This directed generation approach moves machine learned AMP discovery away from a rejection sampling methodology and towards precision design, which affords a high accuracy to discover AMP with desirable activity, selectivity, and safety profile.

While prior research has been mainly concerned with pre-applying labels using QSAR models to generate sequences, we validated our approach of creating a sequence generator then applying physiochemical criteria after the fact to further select AMP candidates from the GAN.

We then carried out molecular dynamics (MD) simulations to verify the stability of the secondary structures and interactions of the AMP candidates with a lipid bilayer membrane. Because membrane disruption is well accepted as a major mechanism of antibacterial AMPs, the membrane-binding propensity was estimated for the AMP candidates using free energy simulations. Half of the peptides from the free-energy rankings were confirmed with antibacterial activity in subsequent bacterial growth experiments. In aggregate, our efficient approach enables a comprehensive consideration of both AMP sequences and structures, which will serve as a valuable platform for future AMP discovery.

Results and Discussion

1. AMP-GAN generated peptide candidates with high sequence diversity.

We constructed a conditional GAN model (AMP-GAN) to produce a large number of candidate AMP sequences. AMP-GAN was trained over a training set comprised initially of 8,005 known AMPs from available databases as well as nearly 500,000 non-AMP peptides. Four conditions were chosen based on their fundamental nature and data abundancy: sequence length, microbial target, target mechanism, and MIC50. Distinct from prior research, we designed two novel labels regarding the microbial target (Gram-positive/negative bacteria, fungi, viruses, or others) and target mechanism (e.g. disrupting the lipid membrane, vital protein inhibition, or interfering with DNA/RNA), which allowed us to train our network on the broadest set of AMPs and the potential connections between them. As a result, each generated peptide sequence was associated with specific microbial targets and potential mechanisms, which greatly facilitated selection and experimental validation. As a matter of practice, we focused on AMPs within 32 amino acids in length due to the reliability of structural prediction and affordable synthetic cost. Thus, we applied a maximal 32-residue length constraint to both the training sequences and generated sequences (Fig. 1A). Moreover, while helicity or secondary structures were commonly included in the descriptors in several QSAR models (5,9-11), there was no structural constraint in our AMP-GAN. Thus, AMP-GAN can produce novel peptide sequences that can fold into different three-dimensional (3D) structures.

The diversity of generated peptide sequences is an important indicator for a good AMP generative model. It is a major challenge for many models trained on relatively scarce and sparsely labeled data to generate new sequences that are distinct, both from known sequences and from each other in the data set. In this work we generated 50,000 sequences with AMP-GAN using condition vectors drawn from the training data. We compared the training data (known AMPs) and generated data (AMP candidates) in terms of length, formal charge (FC), and helicity penalty (HP) (Fig. 1). HP was determined by the sum of energy scores (12) associated with the helix propensity of each residue, e.g. no penalty for alanine as 0 kcal/mol and high penalty for glycine 1 kcal/mol and proline 3.6 kcal/mol. Overall, the 50,000 AMP candidates displayed normal distributions in terms of length, FC and HP. These distributions were mostly consistent with the expectation based on known AMPs, except that slightly more anionic peptides were found in our generated data (Fig. 1A). To assess the sequence diversity in the AMP candidates, we calculated the pairwise similarity using the ratio from SequenceMatcher in the difflib python package (https://docs.python.org/3/library/difflib.html), which was defined as the fraction of contiguous matching subsequences between two given peptides (Fig. 1B). With the 3829 20-residue peptides as an example, the majority of peptides are distinct from each other, as indicated by the low sequence matcher ratio overall (Fig. 1B), although our analysis showed some peptides are relatively similar. In addition, we also found low similarity of our generated sequences to known proteins and peptides using the Basic Local Alignment Search Tool (BLAST) to search the UniprotKB (13) database. For sequences below 10 residues in length, we found 76% of our AMP candidates with accidental homology to regions of larger proteins (E-value 0.1-10), while the rest were likely novel sequences unrelated to known proteins (E-value >> 10). For longer AMP candidates, we had a much higher fraction of sequences with high E-values and low bit scores (low homology), which indicates the capacity of AMP-GAN to generate new sequences (E-value > 1, Tables S1 and 2) 40% more than prior studies using VAE (6) and CNN (14). Therefore, AMP-GAN is able to generate diverse, novel AMP-like sequences and explore new areas in the vast AMP chemical space.

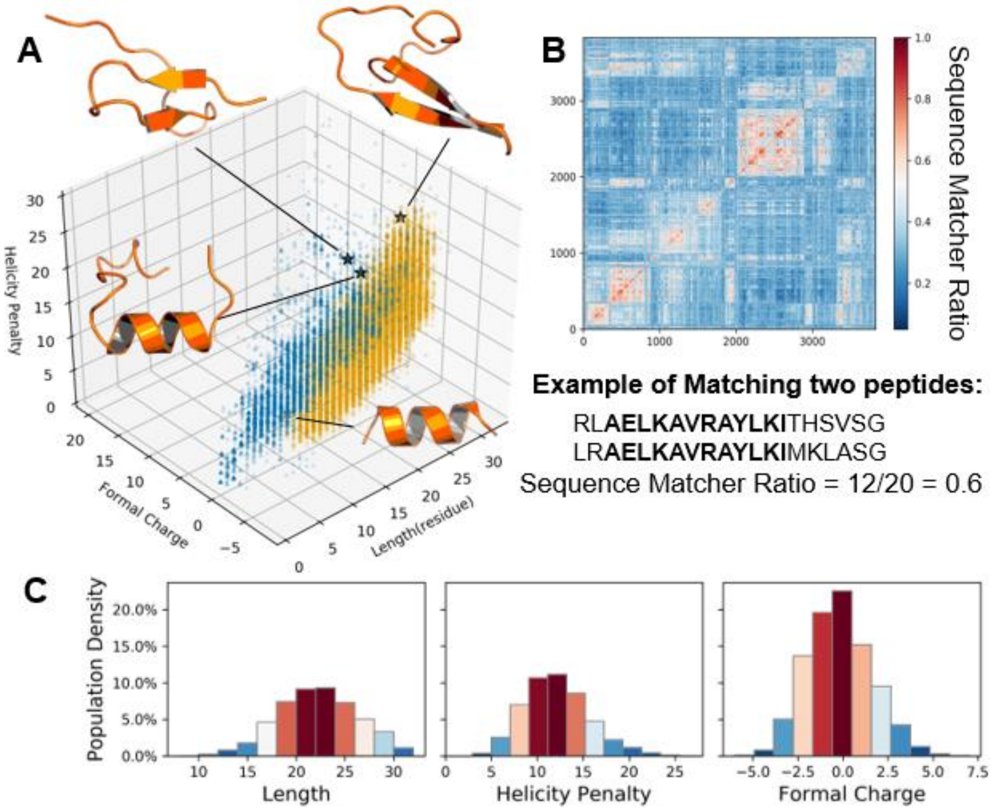

Figure 1. Diversity of AMP-GAN-generated peptide sequences and structures. **(A)** Comparison of the training set (blue dots) and the generated data set (orange dots) in terms of peptide length (in residue), formal charge (in electron charge unit), and helicity penalty (in kcal/mol). **(B)** Heatmap to show the sequence diversity among the generated peptides of 20 residues in length (3829 peptides in total). X-Y axes represent numbering of the 20-residue peptides in a sorted list. The long contiguous matching subsequences were identified between two given peptides (see the example). A ratio of 1 shows exact identical sequences (the diagonal in the heatmap), while the sequence matcher ratio remains low among most of our generated data — an indicator of sequence diversity within our data set. **(C)** Histograms to show the distribution of peptide length (in residue), formal charge (in electron charge unit), and helicity penalty (in kcal/mol) in our generated data set.

2. AMP-GAN generated peptide candidates with high structural diversity.

The sequence diversity from AMP-GAN serves as a firm foundation for the generation of structurally diverse AMP candidates. Notably, the structural diversity represents an even harder challenge besides the sequence diversity. While the vast structural diversity of AMPs has been known for some time, currently the ~1200 known structures of AMPs (collected in APD3 (15) until June 2020) show that only 1/3 of the AMP structures are classified as helical. However, most prior studies of AMP design were restricted to alpha helices or simply ignored the peptide structures. The generator of our AMP-GAN operates on the primary structure of generated peptides, and is not explicitly provided information about secondary or tertiary structure. This, in combination with our usage of conditioning variables that do not explicitly enforce aspects of secondary or tertiary structure, allows AMP-GAN to generate a broad range of peptides with diverse structures that extend beyond alpha helices.

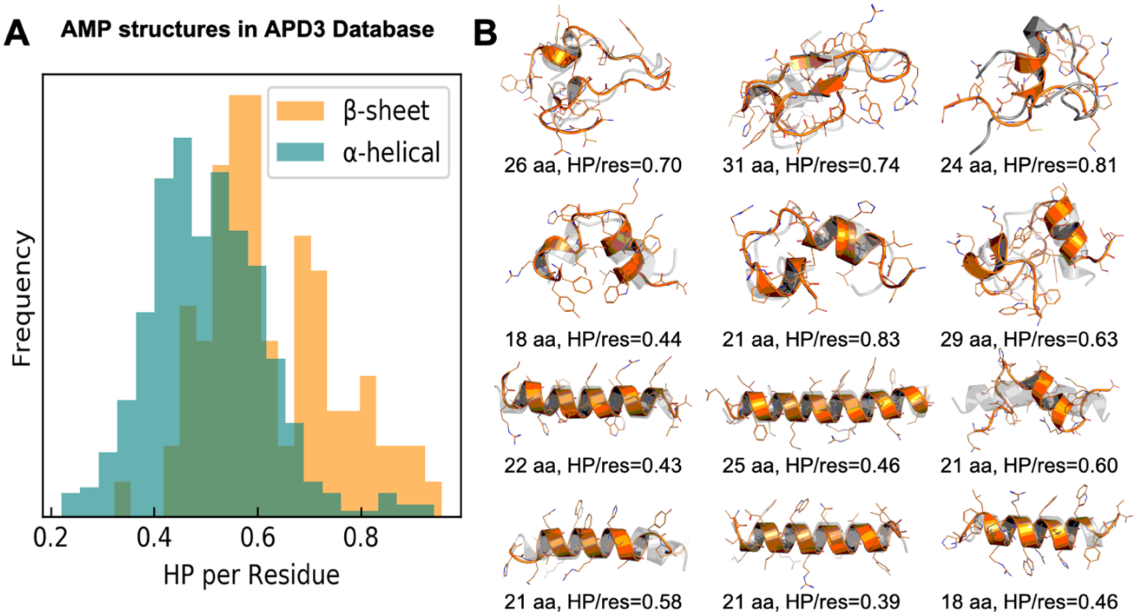

Figure 2. (A) Distribution of known AMPs that are classified as alpha or beta structures in terms of HP per residue, which included ~400 alpha helical and ~80 beta hairpin AMPs. **(B)** Predicted structures of 12 cationic AMP candidates (formal charge > +3). Each structure in the orange cartoon is from the final snapshot of a 10-ns MD simulation. The structure in grey is the initial model that is either predicted by Pep-Fold (16) or restricted to be helical. The peptides with low HP/res are more likely to adopt stable helical structures while the peptides with high HP/res may fold into beta hairpin structures (without disulfide bonds) or other motifs.

Using peptide structure prediction (16), we showed that the generated sequences likely adopt various folded structures such as alpha helices, beta hairpins, and other prototypical motifs. Despite focusing on peptides shorter than 32 residues, the structural diversity observed in selected cases (Figs. 1 and 2) was in agreement with what was expected from known AMP structures. For example, several sequences with high HP (HP/res > 0.7) were more likely to adopt stable hairpin structures, which were confirmed by 10-ns MD simulations. In contrast, almost all the peptides with low HP (HP/res < 0.5) formed helical structures. Hairpin structures of known AMPs typically contain 1-3 disulfide linkages (17). As the redox information was not included in the current design of AMP-GAN, ad hoc consideration would be needed to model the disulfide bonds in our AMP candidates. However, our MD simulations clearly indicate that the beta hairpin motif was stable for several AMP candidates even in the absence of disulfide bonds. In general, while our results confirm the diversity of 3D structures folded from the AMP-GAN-generated sequences, it is noteworthy that HP provides a simple estimation of the AMP secondary structures (Fig. 2).

3. Free-energy simulations identified helical peptides with antibacterial activities.

There are five times as many alpha structures amongst known AMPs as there are beta structures (Fig. 2). Thus, we focused on the alpha helical AMP candidates potentially against Gram-positive/negative bacteria to gain proof of principle in this work. First, we eliminated all sequences which had targets against mammalian, cancer, or fungal groups in their conditioning vector. We further excluded all AMPs that above the lowest band of MIC50 activity (~5.76 µg/ml) from our conditioning vectors. Then two selection criteria were applied, aimed to identify chemically relevant candidates for antibacterial tests. (i) *The structure rules*: a total formal charge greater than +2 or a calculated helical penalty < 5 kcal/mol (12); (ii) *The chemical rules*: elimination of sequences with multiple adjacent Ser or Gln or ones which had selenocysteine or pyrrolysine and sequences which did not contain adjacent Arg and Trp (known as the RW pattern commonly seen in many AMPs). The structural rules were used to select sequences likely to be charged or helical which would be easier for us to confidently simulate, while the chemical rules helped identify chemically relevant antibacterial peptides that avoided unstable and difficult to synthesize peptides. From the 50,000 initially generated sequences, 13 sequences passed all of these filters.

To further select the most-likely membrane-active peptides from the 13 candidates, free-energy calculations based on all-atom simulations were performed to screen their membrane-binding propensity. For each peptide, an estimate of the free energy change ΔG (from embedding in a model *E. coli* bacterial inner membrane) was obtained via a quick set of MD simulations with umbrella sampling (US). US restrained the helical peptides to different heights above and below the membrane surface, and calculated potential of mean force (PMF) to approximate ΔG. The determined ΔG for each peptide was used to obtain a score of membrane-binding propensity, given by Equation 1, which compares a candidate sequence to a known membrane-active peptide, magainin 2 (ΔG_+) and a known non-AMP sequence (ΔG_-, sequence in the SI).

$$Score = (\Delta G_+ - \Delta G)/(\Delta G_- - \Delta G_+) \qquad \text{(Eq. 1)}$$

In Equation 1, a positive (or negative) score represents an AMP sequence with a higher (or lower) propensity to bind the membrane than magainin 2. This difference is then scaled by the range of the positive and negative controls for relative comparison. US was repeated in three replicas for the control sequences and showed they have significantly different ΔG values with ΔG_+ 7 kcal/mol lower than ΔG_- and a 5-13% relative error (calculated ΔG values in Table S3), suggesting this methodology can distinguish poor from good membrane binding peptides. It is noted that the significance of our free-energy calculation is in the relative values, which is confirmed by the control sequences following in the expected trend that a known AMP would have a more negative change in energy upon membrane binding than a non-AMP (i.e. known AMPs have a more favorable change in energy upon binding to anionic membranes). Given the membrane-binding propensity score, we ranked the 13 helical peptides and further reduced to six for synthesis and experimental validation (Table 1).

Table 1. Properties of six selected AMP candidates. The E-value and sequence identity were obtained from BLAST, while MIC50 values were measured by serial dilution growth inhibition assays. All the peptides are amidated in the synthesis.

	Sequence	MW (kD)	Actual Length (residues)	Formal Charge	Helicity Penalty (kcal/mol)	Score	E-value	Identity (%)	*E. coli* MIC50(μg/ml)
Pep1	SGRIASHFTQLWRWLRGYYKLM	2.77	22	+4	9.4	0.23	0.5	90	64
Pep2	QSGIFMHLKQLCRWLRGYMQWAGIG	2.98	25	+3	11.6	-0.25	0.5	62	>256
Pep3	QSNVFLSHFMQPCRWLRGKMG	2.52	21	+3	12.6	-0.52	0.4	52	>256
Pep4	MHKTQLRWFRCHLSQYPGAGL	2.53	21	+3	12.1	-0.65	6.8	89	>256
Pep5	CSKWELRWRRYQGKVSYQLAL	2.67	21	+4	8.3	-1.23	4.8	65	256
Pep6	WHLRRTRWRFEHLWSYGV	2.48	18	+3	8.2	-1.23	0.1	83	256

Note: It is common for short peptides to show high sequence identity to large proteins collected in the UniproKB database. However, the high expectation value (>0.1) denied the biological significance of the sequence similarity. Thus, these six peptides are novel AMPs.

Of the simulated helical AMP sequences, the relative orientation of the charged and non-polar sidechains correlated with their predicted membrane binding score. This molecular origin is demonstrated in the membrane bound windows from the MD simulations of the highest scoring (Pep1) and lowest scoring (Pep6) peptides (Table 1 and Fig. 3). For Pep6 bound to the membrane, the hydrophobic sidechains in the helical segment were exposed to the polar solvent (Fig. 3A). On the other hand, the non-polar sidechains on Pep 1 were embedded in the non-polar environment of the membrane core (Fig. 3B). For the high scoring Pep 1, this orientation coincided with charged sidechains interacting with the charged phosphate groups on the membrane surface, establishing a combination of strong interactions on all sides of the helical segment. While Pep 6 also established strong interactions between its charged sidechains and the phosphate groups of the membrane, due to the orientation of the helix, these interactions actually blocked entry of the Pep6 deeper into the membrane. These observations are in agreement with known amphiphilic and helical AMPs. In addition to the membrane-binding score, other factors (such as conformational changes and aggregation) may affect the actual antibacterial activity. For example, Pep2 and Pep4 have low solubility (< 5 mg/ml in water) and likely aggregate in aqueous solution, while Pep3 may adopt U-shape conformations rather than a straight helix. In general, our results suggest that AMP-GAN has the capability to generate sequences with patterns of hydrophobic and hydrophilic sidechains that can enhance the membrane binding affinity.

While conventional MD simulations of membrane disruption or poration induced by AMPs (18,19) are often challenging due to the high computational cost of all-atom MD simulations, our simulation approach is valuable to provide a scoring method, which is complementary to the sequence-based AMP-GAN. Notably, our methodology is fundamentally distinct from previous approaches which employed both sequence- and structure-dependent descriptors or labeling, especially in regards to the following major features. Firstly, the number of labels in the GAN model is minimized, which enables focus on a greater diversity of peptide sequences. Furthermore, we can also obtain a higher accuracy in selection of AMP candidates with simulations based on molecular structures. Last but not least, the sequence-based generator and structure-based scoring are well integrated in our methodology. This latter advantage may allow this methodology to capture the diversity of AMPs which are known to act on multiple targets (membrane, vital proteins, and DNA) and via various mechanisms of actions. Because the AMP-GAN can be used to generate sequences specifically targeted at the aforementioned targets, this should allow for the construction of precise molecular models to verify the effectiveness of AMPs upon certain targets in subsequent simulations.

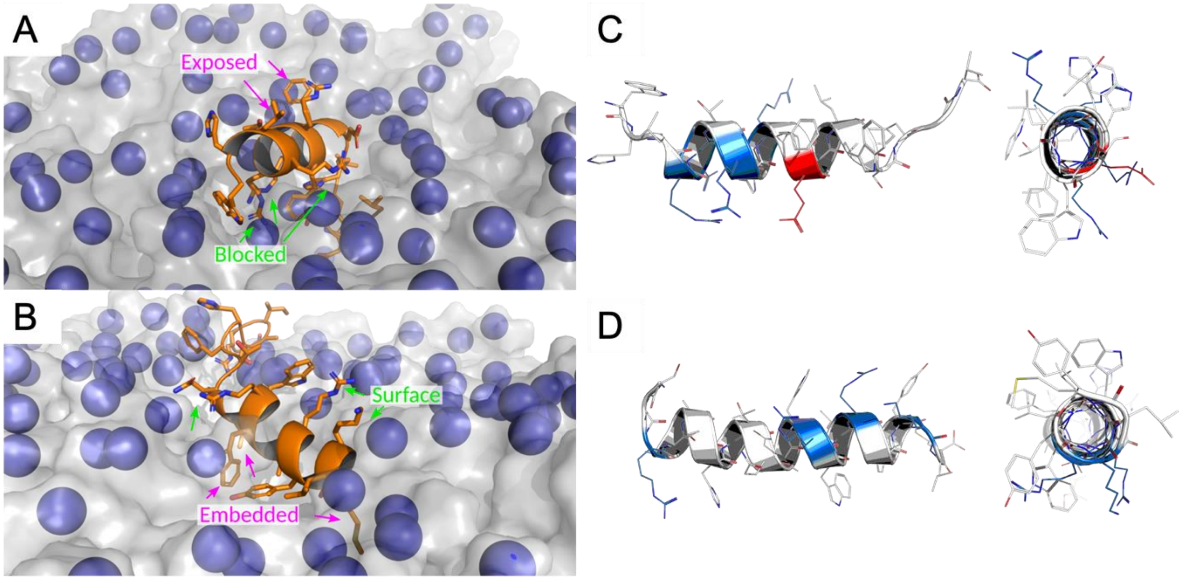

Figure 3. Membrane binding motifs and solution structures of Pep6 (**A & C**) and Pep1 (**B & D**). The peptides embedded in the membrane are shown as orange cartoons. The membrane surface is shown as a transparent surface and phosphorous atoms as blue spheres. Charged interactions between sidechains and the phosphorous atoms that either block membrane penetration (as in A) or stabilize membrane penetration (as in B) are emphasized in green. Non-polar interactions of the solvent exposed (A) or membrane embedded (B) sidechains are highlighted in green. For the peptide models in solution (relaxed in 10-ns MD simulations) were shown with white cartoon and charged residues in color (blue for Lys/Arg and red for Asp/Glu). The charged surface of these peptides is shown in contrast to the hydrophobic surface with many aromatic side chains.

4. <u>Inhibition of Bacterial Growth</u>

From the 50,000 AMP candidates generated by AMP-GAN, we selected 13 plausible ones that were predicted to be cationic, helical, and active against Gram-positive and -negative bacteria, which were further scored by simulated PMFs as above-mentioned. The best six peptides with the most preferred membrane-binding propensity were purchase from Genscript (> 90% HPLC purity) and tested against four microbes, *E. coli*, *S. aureus*, *P. aeruginosa*, and *K. pneumoniae*., in order to validate our sequence-based generation and structure-based scoring methodology. We used a serial dilution method from 256 μg/mL to 0.01 μg/mL in an attempt to find the value for which approximately 50% of the bacterial growth was inhibited (MIC50). Most of the peptides show some growth-reduction activity against Gram-negative *E. coli*, *P. aeruginosa* and *K. pneumoniae*, and Gram-positive *S. aureus*. Their antibacterial activity increased with increasing concentration, but increases in light scattering due to peptide aggregation also increased the apparent bacterial growth at high concentrations (Fig. 4). While this was partially controlled for by subtracting peptide-alone optical density from corresponding growth measurements, this effect does reduce the accuracy of the assays at higher peptide concentrations. For *E. coli*, Pep1 at 64 ug/ml and Pep5 and Pep6 at 256 ug/ml inhibited over 50% of bacterial growth, while the others achieved less than 40% inhibition at the concentration below 256 ug/ml. For *S. aureus*, four of the six peptides inhibited ~50% bacterial growth at the concentration of 256 ug/ml, except Pep3 and Pep4. In addition, most of our peptides inhibited *P. aeruginosa* and *K. pneumoniae* growth by 20-30% at a concentration below 256 ug/ml (Fig. S4). AMP specificity was achieved with Pep3 which targeted *S. aureus* but not *E. coli* (Fig. 4), while more broad-spectrum activity was observed for other peptides. In general, the AMP-GAN model produced novel and active peptides against the variety of bacteria tested. Two of the most interesting peptides, Pep1 and Pep6, were further tested for toxicity. Pep1 displayed notable toxicity at our highest measured value of 300μg/mL, with less than 5% of the cells still viable compared to our control. In all other concentrations both Pep1 and Pep6 had a greater than 90% viability compared to our control (Table S4).

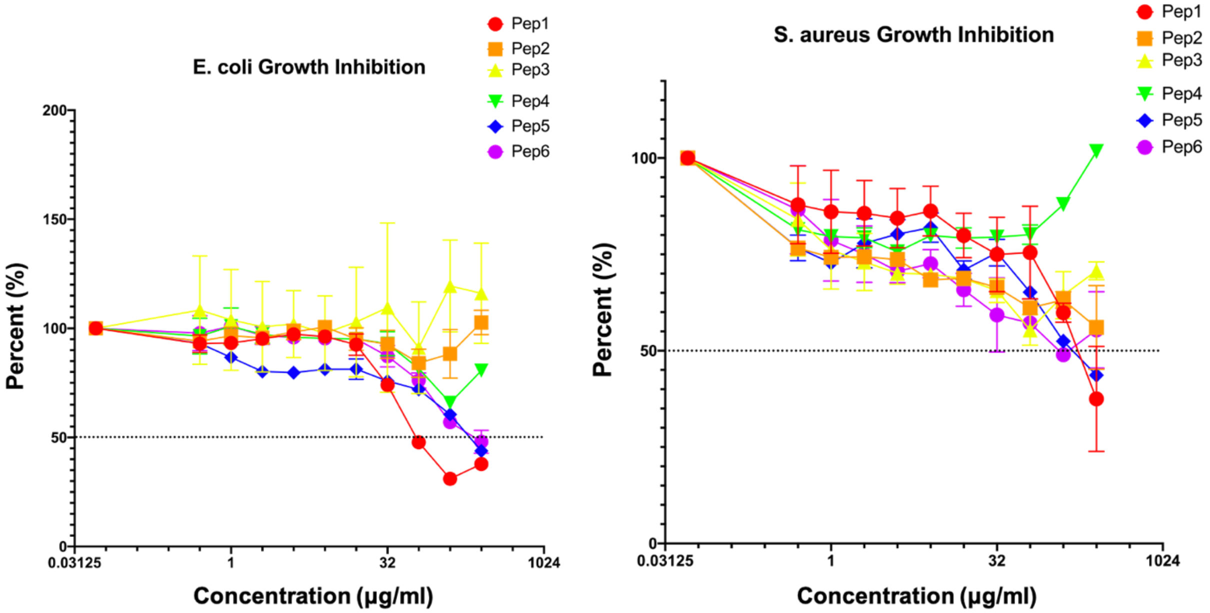

Figure 4. Bacterial growth assays for the six synthesized peptides against *E. coli* (left) and *S. aureus* (Right). The growth of the bacteria was measured by optical density at 600 nm (OD_{600}) and normalized as a percentage to bacterial growth with no peptide. These normalized values are plotted against AMP concentration. Dashed lines to judge MIC50 concentrations are shown. Error bars are from three independent experiments.

Considering the experimental results, we think there are several major factors that impacted the accuracy of our AMP selection in the free-energy simulations. Firstly, we assumed that all the peptides to be helical and we restricted them to fully helical conformations during the molecular simulations. However, further examination of circular dichroism (CD) spectra revealed that all 6 peptide candidates were no more than 32% helical in solution without SDS (Table 2, Fig. S5). With the presence of SDS, it is suggested Pep1 dramatically increases its helicity in the SDS micelles, in good agreement with the mechanisms of typical helical AMPs like melittin (20). However, Pep6 remains largely disordered regardless of SDS, which well explains why we were not able to properly evaluate its membrane-binding propensity. Secondly, we only considered a single peptide in solution and membrane during, but AMP aggregation is common at high concentrations. As a result, the effective peptide concentration may be reduced, which likely is the reason why Pep2, Pep3, and Pep4 showed lower activity typically at concentrations above 32 μg/ml. In addition to peptide folding/unfolding and aggregation in solution and on the membrane, there are other transitions like peptide tilting that have not been fully captured in current free-energy simulation. Finally, in the MD simulations it was assumed that the peptides attack the membrane due to that being the prevailing commonality of the peptides from our training set. However, the labels used to generate the candidate sequences were not exclusively membrane active and therefore other mechanisms could also be involved. While this work serves as a proof of principle, it is possible to develop more rigorous, more accurate simulation approaches to improve AMP selection among different targets and mechanisms.

Table 2. Peptide helicity (%) measured from CD spectra. The secondary structure analysis was carried out with the program Dichromweb developed by Wallace et al. (21). Raw data of the CD spectra are provided in Fig. S5.

% of helicity	Pep1	Pep2	Pep3	Pep4	Pep5	Pep6
w/o SDS	7	32	17	4	22	3
w SDS	73	80	58	15	6	5

Conclusion Remarks

In summary, we have invented and demonstrated an efficient and accurate AMP design methodology. By curating a dataset that comprises 500,000 of non-AMP peptide sequences and 8000 labeled AMP sequences to train the AMP-GAN model for generating new AMP sequences, we created a general generator that can be used to target more than just one type of microbe or one mechanism. We then demonstrated a proof of concept method for evaluating peptide candidates by evaluating membrane binding tendency toward a model *E. coli* inner membrane. This technique can be extended to different membrane compositions and therefore different microbes, thus retaining the generality of our generator in our screening as well. This synergy of sequence-based generation and structure-based ranking represents a new methodology toward flexible, precision AMP design. Extensive analysis of the generated antimicrobial sequences reveals that the proposed framework, beyond being general, is indeed capable of learning and generating from a richer representation than trained upon and yields AMPs that are both diverse in sequence and structure. Our methodology as an entity will be valuable in the ever-necessary rational design of AMPs with the potential to be expanded to even broader classes of biologics.

Materials and Methods

1. Computational Methods

We utilized a GAN approach that incorporates several elements from recent research. The base of our model is a Wasserstein GAN with Gradient Penalty (WGAN-GP), which claims to mitigate many of the pathological learning dynamics that can occur when training GANs, including dissociation and mode collapse. Next, we add conditioning information following the structure outlined by Conditional GANs (CGAN), allowing human designers to influence the features of the generated peptide sequences. Additionally, we include an encoder component following work on Bidirectional GANs (BiGAN), which greatly simplifies the implementation of latent space interpolations and allows for more effective exploration of the learned latent space through the use of known AMPs. Fig. 5 provides a flow diagram that depicts the high-level organization of our AMP-GAN, which is similar to what was investigated by Perarnau *et. al.* (22), excluding the label/conditioning vector encoder. Architecture details for the generator, discriminator, and encoder networks are shown in Figs. S1, S2, and S3 respectively.

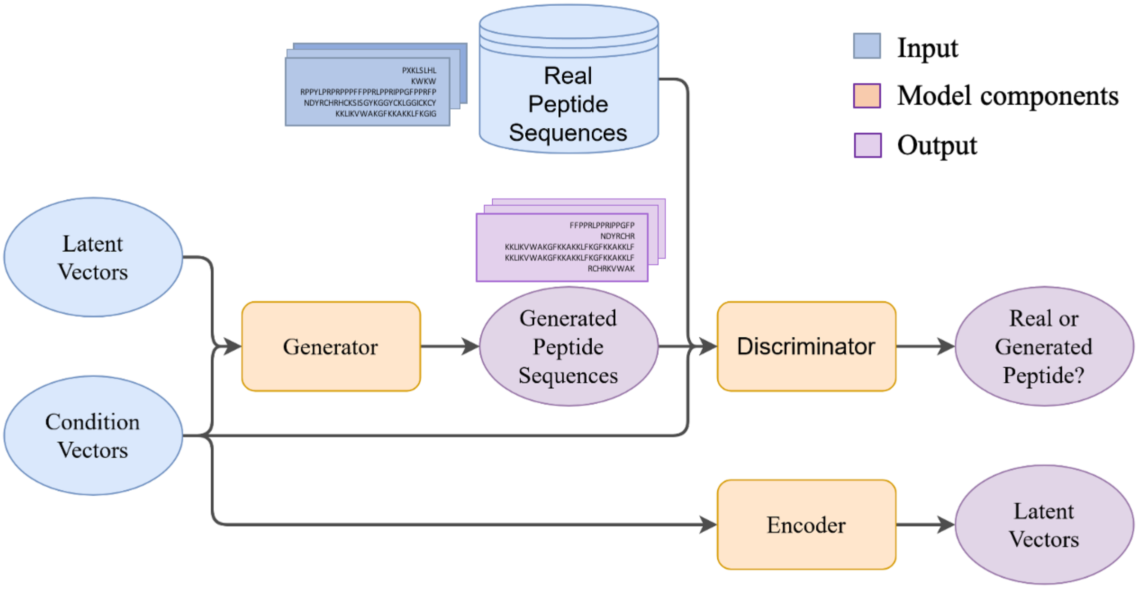

Figure 5. Illustration of the organization of AMP-GAN.

For our training data we utilized the Database of Antimicrobial Activity and Structure of Peptides (DBAASP) as our known good AMPs (3). For those AMPs with multiple MIC50 readings we converted all readings into μg/mL and averaged them these were then binned into 10 deciles and those deciles which were provided to the AMP-GAN. We also combined categories of different microbes into a reduced set of Gram-positive, Gram-negative, Viral, Fungal, Mammalian, and Cancer. Similarly, we collapsed targets into, Lipid Bilayer, RNA/DNA, Cytoplasmic Protein, Membrane Protein, and Virus Entry. We also removed any sequences of length greater than 32 amino acids. For any sequences with a wildcard FASTA symbol (X, B, Z, or J) the symbol was replaced with a randomly selected concrete symbol during training. For our null data set we selected sequences from the Uniprot database, filtering any sequences that were present in DBAASP or were longer than 32 residues. The conditioning vectors for our null database were constructed to indicate no target microbes, no target mechanisms, maximal MIC50, and the appropriate sequence length. We trained AMP-GAN on 100,000 batches with 128 samples per batch. Each batch was composed of 64 randomly sampled AMPs and 64 randomly sampled non-AMP peptides. This translates to approximately 1,000 epochs of training over the positive dataset and 13 epochs over the negative dataset. Implementation details for the data processing, model creation, and model training elements can be found on GitLab.

For all atom molecular dynamics simulations, initial peptide structures were modeled as a helix and translated 25 Å above a 3:1 mass-ratio POPE:POPG membrane. The CHARMM-36m (23) forcefield was then applied using CHARMM-GUI (24,25) and the systems solvated with TIP3P water model and 120 mM NaCl. All simulations were performed using the AMBER18 simulation package (26). The following US protocol was developed to enable rapid screening and estimation of the relative free energy change upon peptide binding to the model membrane. Each peptide was put through the multi-stage minimization and equilibration protocol. Short, 500 ps, steered molecular dynamics (SMD) trajectories were run with a very stiff spring constant of 500 kcal/molÅ applied to a custom collective variable representing the z-component of the center of mass position of the peptide, over a total distance of 80 Å into the membrane. While this distance was larger than the membrane thickness, it was required to embed the peptide into the membrane over the short period due to translation of the membrane during SMD. Following SMD, frames most closely representing eight umbrella sampling windows linearly spaced between 20 angstroms below the phosphorous atoms in the top leaflet of the membrane and the initial position above the membrane were selected. To allow the membrane atoms and peptide sidechains to relax from the SMD simulations each window was equilibrated for 1 ns using a stiff spring constant of 1 kcal/mol Å. Following this period, 20 ns of production umbrella sampling simulations were performed with a weaker spring constant of 0.1 kcal/molÅ that ensured proper overlap of the umbrella windows. The final data was analyzed using the weighted histogram analysis method, and the final difference between the first and last window were used to estimate the free energy difference for embedding the peptide in the membrane.

2. Experimental Methods

These peptides were purchased from Genscript and characterized for solubility and helicity. Helicity was determined by circular dichroism (CD) spectra with a peptide concentration of 20μM in 1mL Milli-Q Water. The Jasco J-815 spectropolarimeter was set to analyze the peptides at a wavelength range of 260-190nm, a cell length of 2mm, and 3 scans. The resulting spectra was then fed to Dichromweb (21) and the helicity was then calculated using the CDSSTR method and a reference dataset.

Bacteria were stored as 15% glycerol stocks at -80 ^{o}C and routinely propagated on LB agar or broth (Lysogeny broth, Lennox formulation). The specific species and strains used for these assays were *Escherichia coli* K12, *Pseudomonas aeruginosa* PAO1, *Klebsiella pneumoniae var. pneumoniae* KPPR1, and *Staphylococcus aureus var. aureus* ATCC 12600. For antimicrobial testing, bacteria were streaked to an LB agar plate and incubated for 24 h at 37 ^{o}C. Colonies from the resulting plates were used to start 3 ml LB broth cultures that were grown at 37 ^{o}C for 18 h overnight. To the 'no peptide' wells, 75 ul of sterile deionized water were added into the wells of tissue-culture treated 96-well polystyrene plates, and the remaining wells were set up as two-fold serial dilutions of each peptide starting at 256 ug/ml. Optical density of the bacterial cultures was measured at 600 nm (OD600) and cells were collected by centrifugation and normalized to an OD600 of 0.02 in LB and 75 ul of this suspension were aliquoted into the 75 ul of water or peptide solution. Thus, the final well contained 150 ul of liquid that was ½ strength LB and bacteria at a starting OD600 of 0.01. Identical plates were also generated with no bacteria added (only LB broth) to assess peptide aggregation to subtract from bacterial growth measurements. These 96-well plates were incubated with horizontal shaking at 120 rpm at 37 °C for 24 h and then OD600 was measured in a Biotek Synergy 2 plate reader.

The AMP cytotoxicity was tested with human A549 cells. The A549 cells was plated in 24-well plates at 300,000 cells/well. When cells were ~85% confluent, we added three different concentrations (30, 100, and 300 ug/ml) of the tested peptides for 24 hours. The cell viability was measured using a fluorescent dye calcein AM.

Reference:

1. Fjell, C.D., Hiss, J.A., Hancock, R.E.W. and Schneider, G. (2012) Designing antimicrobial peptides: form follows function. *Nature Reviews Drug Discovery*, **11**, 37-51.

2. Magana, M., Pushpanathan, M., Santos, A.L., Leanse, L., Fernandez, M., Ioannidis, A., Giulianotti, M.A., Apidianakis, Y., Bradfute, S., Ferguson, A.L. *et al.* (2020) The value of antimicrobial peptides in the age of resistance. *The Lancet Infectious Diseases*.

3. Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H.L., Squires, R.B., Hurt, D.E., Grigolava, M., Chubinidze, M., Gogoladze, G., Vishnepolsky, B. *et al.* (2015) DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Research*, **44**, D1104-D1112.

4. Lee, E.Y., Wong, G.C.L. and Ferguson, A.L. (2018) Machine learning-enabled discovery and design of membrane-active peptides. *Bioorg Med Chem*, **26**, 2708-2718.

5. Lee, E.Y., Fulan, B.M., Wong, G.C.L. and Ferguson, A.L. (2016) Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences*, **113**, 13588-13593.

6. Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Dos Santos, C., Chen, P.-Y. *et al.* (2020) *Accelerating Antimicrobial Discovery with Controllable Deep Generative Models and Molecular Dynamics*.

7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Networks.

8. Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets.

9. Torrent, M., Andreu, D., Nogués, V.M. and Boix, E. (2011) Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLOS ONE*, **6**, e16968.

10. Maccari, G., Di Luca, M., Nifosí, R., Cardarelli, F., Signore, G., Boccardi, C. and Bifone, A. (2013) Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization. *PLoS Comp. Biol.*, **9**, e1003212.

11. Nagarajan, D., Nagarajan, T., Roy, N., Kulkarni, O., Ravichandran, S., Mishra, M., Chakravortty, D. and Chandra, N. (2018) Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *The Journal of biological chemistry*, **293**, 3492-3509.

12. Nick Pace, C. and Martin Scholtz, J. (1998) A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical Journal*, **75**, 422-427.

13. Consortium, T.U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506-D515.

14. Witten, J. and Witten, Z. (2019) Deep learning regression model for antimicrobial peptide design. *bioRxiv*, 692681.

15. Wang, G., Li, X. and Wang, Z. (2015) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, **44**, D1087-D1093.

16. Thévenet, P., Shen, Y., Maupetit, J., Guyon, F., Derreumaux, P. and Tufféry, P. (2012) PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Research*, **40**, W288-W293.

17. Panteleev, P.V., Bolosov, I.A., Balandin, S.V. and Ovchinnikova, T.V. (2015) Structure and Biological Functions of β-Hairpin Antimicrobial Peptides. *Acta Naturae*, **7**, 37-47.

18. Lipkin, R., Pino-Angeles, A. and Lazaridis, T. (2017) Transmembrane Pore Structures of β-Hairpin Antimicrobial Peptides by All-Atom Simulations. *The Journal of Physical Chemistry B*, **121**, 9126-9140.

19. Zhang, Z., Subramaniam, S., Kale, J., Liao, C., Huang, B., Brahmbhatt, H., Condon, S.G., Lapolla, S.M., Hays, F.A., Ding, J. *et al.* (2016) BH3 ‐ in ‐ groove dimerization initiates and helix 9 dimerization expands Bax pore assembly in membranes. *The EMBO Journal*, **35**, 208-236.

20. Liao, C., Esai Selvan, M., Zhao, J., Slimovitch, J.L., Schneebeli, S.T., Shelley, M., Shelley, J.C. and Li, J. (2015) Melittin aggregation in aqueous solutions: insight from molecular dynamics simulations. *J. Phys. Chem. B*, **119**, 10390-10398.

21. Whitmore, L. and Wallace, B.A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers*, **89**, 392-400.

22. Perarnau, G., van de Weijer, J., Raducanu, B. and Álvarez, J.M. (2016) Invertible Conditional GANs for image editing.

23. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B.L., Grubmüller, H. and MacKerell, A.D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, **14**, 71-73.

24. Jo, S., Kim, T., Iyer, V.G. and Im, W. (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, **29**, 1859-1865.

25. Wu, E.L., Cheng, X., Jo, S., Rui, H., Song, K.C., Dávila-Contreras, E.M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R.M. *et al.* (2014) CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J. Comput. Chem.*, **35**, 1997-2004.

26. Lee, T.-S., Cerutti, D.S., Mermelstein, D., Lin, C., LeGrand, S., Giese, T.J., Roitberg, A., Case, D.A., Walker, R.C. and York, D.M. (2018) GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J. Chem. Inf. Mod.*, **58**, 2043-2050.