nature portfolio

https://doi.org/10.1038/s41551-021-00689-x

Supplementary information

Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations

In the format provided by the authors and unedited

Table of Contents

1 Supplementary Text

1.1 Conditional Generation with Autoencoders

Since the peptide sequences are represented as text strings here, we will be limiting our discussion to literature around text generation with constraints. Controlled text sequence generation is non-trivial, as the discrete and non-differentiable nature of text samples does not allow the use of a global discriminator, which is commonly used in image generation tasks to guide generation. To tackle this issue of non-differentiability, policy learning has been suggested, which suffers from high variance during training (*1, 2*). Therefore, specialized distributions, such as the Gumbel-softmax (*3, 4*), a concrete distribution (*5*), or a soft-argmax function (*6*), have been proposed to approximate the gradient of the model from discrete samples.

Alternatively, in a semi-supervised model setting, the minimization of element-wise reconstruction error has been employed (*7*), which tends to lose the holistic view of a full sentence. Hu et al. (*8*) proposed a VAE variant that allows both controllable generation and semi-supervised learning. The working principle needs labels to be present during training and encourages latent space to represent them, so the addition of new attributes will require retraining the latent variable model itself. The framework relies on a set of new discrete binary variables

in latent space to control the attributes, an ad-hoc wake-sleep procedure. It requires learning the right balance between multiple competing and tightly interacting loss objectives, which is tricky.

Engel et al. (*9*) propose a conditional generation framework without retraining the model, similar in concept to ours, by modeling in latent space post-hoc. Their approach does not need an explicit density model in $\mathbf{z}$-space, rather relies on adversarial training of generator and attribute discriminator and focuses modifying sample reconstructions rather than generating novel samples. Recent Plug and Play Language Model (PPLM) for controllable language generation combines a pre-trained language model (LM) with one or more simple attribute classifiers that guide text generation without any further training of the LM (*10*).

1.2 Conditional Generation for Molecule Design

Following the work of Gómez-Bombarelli et al. (*11*), Bayesian Optimization (BO) in the learned latent space has been employed for molecular optimization for properties such as drug likeliness (QED) or penalized logP. The standard BO routine consists of two key steps: (i) estimating the black-box function from data through a probabilistic surrogate model; usually a Gaussian process (GP), referred to as the response surface; (ii) maximizing an acquisition function that computes a score that trades off exploration and exploitation according to uncertainty and optimality of the response surface. As the dimensionality of the input latent space increases, these two steps become challenging. In most cases, such a method is restricted to local optimization using training data points as starting points, as optimizers are likely to follow gradients into regions of the latent space that the model has not been exposed to during training. Reinforcement learning (RL) based methods provide an alternative approach for molecular optimization (*12–15*), in which RL policies are learned by incorporating the desired attribute as part of the reward. However, a large number of evaluations are typically needed for both BO and

RL-based optimizations while trading off exploration and exploitation (*16*). Semi-supervised learning has also been used for conditional generation of molecules (*17–20*), which needs labels to be available during the generative model training.

CLaSS is fundamentally different from these existing approaches, as it does not need expensive optimization over latent space, policy learning, or minimization of complex loss objectives - and therefore does not suffer from cumbersome computational complexity. Furthermore, CLaSS is not limited to local optimization around an initial starting point. Adding a new constraint in CLaSS is relatively simple, as it only requires a simple predictor training; therefore, CLaSS is easily repurposable. CLaSS is embarrassingly parallelizable as well.

2 Dataset

2.1 A Dataset for Semi-Supervised Training of AMP Generative Model

We compiled a new two-part (unlabeled and labeled) dataset for learning a meaningful representation of the peptide space and conditionally generating safe antimicrobial peptides from that space using the proposed CLaSS method. We consider discriminating for several functional attributes as well as for presence of structure in peptides. Only linear and monomeric sequences with no terminal modifications and length up to 50 amino acids were considered in curating this dataset. As a further pre-processing step, the sequences with non-natural amino acids (B, J, O, U, X, and Z) and the ones with lower case letters were eliminated.

Unlabeled Sequences: The unlabeled data is from Uniprot-SwissProt and Uniprot-Trembl database (*21*) and contains just over 1.7 M sequences, when considering sequences with length up to 50 amino acid.

Labeled Sequences: Our labeled dataset comprises sequences with different attributes curated from a number of publicly available databases (*22–26*). Below we provide details of the labeled dataset:

- Antimicrobial (8683 AMP, 6536 non-AMP);
- Toxic (3149 Toxic, 16280 non-Toxic);
- Broad-spectrum (1302 Positive, 1238 Negative);
- Structured (1170 Positive, 2136 Negative);
- Hormone (569 Positive);
- Antihypertensive (1659 Positive);
- Anticancer (504 Positive).

2.1.1 Details of Labeled Datasets

Sequences with AMP/non-AMP Annotation. AMP labeled dataset comprises sequences from two major AMP databases: satPDB (*22*) and DBAASP (*23*), as well as a dataset used in an earlier AMP classification study named as AMPEP (*25*). Sequences with an antimicrobial function annotation in satPDB and AMPEP or a MIC value against any target species less than 25 μg/ml in DBAASP were considered as AMP labeled instances. The duplicates between these three datasets were removed to generate a non-redundant AMP dataset. And, the ones with mean activity against all target species $>$ 100 μg/ml in DBAASP were considered negative instances (non-AMP). Since experimentally verified non-AMP sequences are rare to find, the non-AMP instances in AMPEP were generated from UniProt sequences after discarding sequences that were annotated as AMP, membrane, toxic, secretory, defensive, antibiotic, anticancer, antiviral, and antifungal and were used in this study as well.

Sequences with Toxic/nonToxic Annotation: Sequences with toxicity labels are curated from satPDB and DBAASP databases as well as from the ToxinPred dataset (*26*). Sequences with "Major Function" or "Sub Function" annotated as toxic in satPDB and sequences with

hemolytic/cytotoxic activities against all reported target species less than 200 μg/ml in DBAASP were considered as Toxic instances. The toxic-annotated instances from ToxinPred were added to this set after removing duplicates resulting in a total to 3149 Toxic sequences. Sequences with hemolytic/cytotoxic activities $>$ 250 μg/ml were considered as nonToxic. The nonToxic instances reported in ToxinPred (sequences from SwissProt or TrEMBL that are not found in search using keyword associated with toxins, *i.e.* keyword (NOT KW800 NOT KW20) or keyword (NOT KW800 AND KW33090), were added to the nonToxic set, totaling to 16280 non-AMP sequences.

Sequences with Broad-Spectrum Annotation Antimicrobial sequences reported in the satPDB or DBAASP database can have both Gram-positive and Gram-negative strains as target groups. We consider such sequences as broad-spectrum. Otherwise, they are treated as narrow-spectrum. Through our filtering, we found 1302 broad-spectrum and 1238 narrow-spectrum sequences.

Sequences with Structure/No-Structure Annotation: Secondary Structure assignment was performed for structures from satPDB using the STRIDE algorithm (*27*). If more than 60% of the amino acids are helix or beta-strand, we label it as structured (positive). Otherwise, they are labeled as negative. Through this filtering, we found 1170 positive sequences and 2136 negative sequences.

Peptide Dataset for Baseline Simulations: For the control simulations, three datasets were prepared. The first two were taken from the satpdb dataset, filtering sequences with length smaller than 20 amino acids. The high potency dataset contains the 51 sequences with the lowest average MIC, excluding sequences with cysteine residues. All these sequences have an average MIC of less than 10 μg / ml. The low potency dataset contains the 41 sequences with the highest MIC, excluding cysteine residues. All these have an average MIC over 300 μg / ml.

To create a dataset of inactive sequences, we queried UniProt using the following keywords:

NOT keyword:"Antimicrobial [KW-0929]" length:[1 TO 20] NOT keyword:"Toxin [KW-0800]" NOT keyword:"Disulfide bond [KW-1015]" NOT annotation:(type:ptm) NOT keyword:"Lipid-binding [KW-0446]" NOT keyword:"Membrane [KW-0472]" NOT keyword:"Cytolysis [KW-0204]" NOT keyword:"Cell wall biogenesis/degradation [KW-0961]" NOT keyword:"Amphibian defense peptide [KW-0878]" NOT keyword:"Secreted [KW-0964]" NOT keyword:"Defensin [KW-0211]" NOT keyword:"Antiviral protein [KW-0930]" AND reviewed:yes. From this, we picked 54 random sequences to act as the inactive dataset for simulation.

3 Model and Methods

3.1 Autoencoder Details

3.1.1 Autoencoder Architecture

We investigate two different types of autoencoding approaches: β-VAE (*28*) and WAE (*29*) in this study. For each of these AEs the default architecture involves bidirectional-GRU encoder and GRU decoder. For the encoder, we used a bi-directional GRU with hidden state size of 80. The latent capacity was set at $D = 100$.

For VAE, we used KL term annealing β from 0 to 0.03 by default. We also present an unmodified VAE with KL term annealing β from 0 to 1.0. For WAE, we found that the random features approximation of the gaussian kernel with kernel bandwidth $\sigma = 7$ to be performing the best. For comparison sake, we have included variations with σ values of 3 and 15 too. The inclusion of z-space noise logvar regularization, $R(logVar)$, helped avoiding collapse to a deterministic encoder. Among different regularization weights used, $1e-3$ had the most desirable behavior on the metrics (see Section 3.1.4).

3.1.2 Autoencoder Training

When training the AE model, we sub selected sequences with length $\leq$ the hyperparameter max_seq_length. Furthermore, both AMP-labeled and unlabeled data were split into train, held out, and test set. This reduces the available sequences for training; *e.g* for unlabeled set the number of available training sequences are 93k for max_seq_length=25, whereas the number of AMP-labeled sequences was 5000. The sequences with reported activities were considered as confirmed labeled data, and those with confirmed labels were up-sampled at a 1:20 ratio. Such upsampling of peptides with a specific attribute label will allow mitigation of possible domain shift due to unlabeled peptides coming from a different distribution. However, the benefit of transfer learning from unlabeled to labeled data likely outweighs the effects of domain shift.

To obtain the optimal hyperparameter setting for autoencoder training, we adopted an automated hyperparameter optimization. Specifically, we performed a grid search in the hyperparameter space and tracked an L_2 distance between the reconstructions of held-out data and training sequences, which was estimated using a weighted combination of BLEU, PPL, z-classifier accuracy, and amino acid composition-based heuristics. The best hyperparameter configuration obtained using this process was the following: learning rate = 0.001, number of iterations = 200000, minibatch size = 32, word dropout = 0.3. A beam search decoder was used with a beam size of 5.

3.1.3 Autoencoder Evaluation

Evaluation of generative models is notoriously difficult (*30*). In the variational auto-encoder family, two competing objective terms are minimized: reconstruction of the input and a form of regularization in the latent space, which form a fundamental trade-off (*31*). Since we want a meaningful and consistent latent space, models that do not compromise the reconstruction quality to achieve lower constraint loss are preferred. We propose an evaluation protocol using

four metrics to judge the quality of both heldout reconstructions and prior samples (*32*). The metrics are

(i) The objective terms, evaluated on heldout data: reconstruction log likelihood $-\log p_\theta(\mathbf{x}|\mathbf{z})$ and $D(q_\phi^h(\mathbf{z})|p(\mathbf{z}))$, where $q_\phi^h(\mathbf{z}) = \frac{1}{N_{hld}} \sum_{\mathbf{x}^i \sim \text{hld}} q_\phi(\mathbf{z}|\mathbf{x}^i)$ is the average over heldout encodings.

(ii) Encoder variance $\log(\sigma_j^2(\mathbf{x}^i))$ averaged over heldout samples, in L^2 over components j. In order to achieve a meaningful latent space, we needed to regularize the encoder variance to not becoming vanishingly small, i.e., for the encoder to become deterministic (*33*). Large negative values indicate that the encoder is collapsed to deterministic.

(iii) Reconstruction BLEU score on held-out samples.

(iv) Perplexity (PPL) evaluated by an external language model, for samples from prior $p(\mathbf{z})$ and heldout encoding $q_\phi(\mathbf{z}|\mathbf{x})$.

Note that (iii) and (iv) involve sampling the decoder (we use beam search with beam 5), which will therefore also take into account any exposure bias (*34, 35*). We propose to evaluate peptide generation using the perplexity under an independently trained language model (iv), which is a reasonable heuristic (*36*) if we assume the independent language model captures the distribution $p(\mathbf{x})$ well. Perplexity of a language model is the inverse probability of the test set, normalised by the number of words. A lower perplexity indicates a better model.

External Language Model

We trained an external language model on both labeled and unlabeled sequences to determine the perplexity of the generated sequences. Specifically, we used a character-based LSTM language model (LM) with the help of LSTM and QRNN Language Model Toolkit (*37*) trained on both AMP-labeled and unlabeled data. We trained our language model with a total of 92624

sequences, with a maximum sequence length of 25. Our best model achieves a test perplexity of 13.26.

To further validate the performance of our language model, we tested it on sequences with randomly generated synthetic amino acids from the vocabulary of lengths ranging between 10 to 25. As expected, we found it to have a high perplexity of 27.29. Also, when evaluating it for repeated amino acids (sequence consisting of a single token from vocabulary), of length ranging between 10 to 25, we found the perplexity to be very low (3.48). Upon further investigation, we observed that the training data consists of amino acids with repeated sub-sequences, which the language model by nature, fails to penalize heavily. Due to this behavior, we can conclude that the perplexity of a collapsed peptide model will be closer to 3.48 (as seen in the case of β-VAE). We have summarized these observation in Table 1.

3.1.4 Autoencoder Variants: Comparison

Architecture		PPL	BLEU	Recon	Encoder variance
Reference	Repeated Sequences	3.48			
	Random	27.29			
	AMP Labeled	5.580			
	Labeled + Unlabeled	13.26			
	Peptide LSTM, (*38*) (Labeled)	22.40			
	Peptide LSTM, (*38*) (Unlabeled)	20.26			
β-VAE (1.0)		3.820	4e-03	2.768	-2.5e-4
β-VAE (0.03)		15.34	0.475	1.075	-0.620
WAE, $\sigma = 3$, $R(logVar) = 1e-3$		13.25	0.853	0.257	-3.078
WAE, $\sigma = 15$, $R(logVar) = 1e-3$		12.98	0.909	0.224	-4.180
WAE, $\sigma = 7$, $R(logVar) = 0$		12.77	0.881	0.214	-13.81
WAE, $\sigma = 7$, $R(logVar) = 1e-2$		15.16	0.665	0.685	-0.3962
WAE, $\sigma = 7$, $R(logVar) = 1e-3$		12.87	0.892	0.216	-4.141
WAE (trained on AMP-labeled)		16.12	0.510	0.354	-4.316

Table 1: Performance of various autoencoder schemes against different baselines.

The evaluated metrics on held-out samples for different autoencoder models trained on ei-

ther labeled or full dataset are also reported in Table 1. We observed that the reconstruction of WAE is more accurate compared to β-VAE: we achieve a reconstruction error of 0.2163 and a BLEU score of 0.892 on a held-out set using WAE with σ of 7 and $R(logVar)$ of 1e-3 (values are 1.079 and 0.493 for β-VAE with β set to 0.03). The advantage of using abundant unlabeled data compared to only the labeled ones for representation learning is evident, as the language model perplexity (PPL, captures sequence diversity) for the WAE model trained on a full dataset is closer to that of the test perplexity (13.26), and the BLEU score is also higher when compared to the WAE model trained only on AMP-labeled sequences. For reference, PPL of random peptide sequences is > 25, and for repeated sequences is 3.48. We have also compared the perplexity of the generated sequences using a LSTM model used in (*38*) and have reported the perplexity in Table 1. Results show that the perplexity of the generated samples using the LSTM model is around 22.4, which is close to the random sequence perplexity. In contrast, WAE (as well as VAE with appropriate KL annealing) achieves much lower perplexity that is close to the perplexity of peptide sequences. Attribute-controlled sampling using a simple LSTM model is non-trivial.

The $\mathbf{z}$-classifier trained on the latent space of the best WAE model achieved a test accuracy of 87.4, 68.9, 77.4, 98.3, 76.3% for detecting peptides with AMP/non-AMP, toxic/non-toxic, anticancer, antihypertensive, and hormone annotation.

3.2 Post-Generation Screening

3.2.1 Sequence-Level Attribute Classifiers

For post-generation screening, we used four sets of monolithic classifiers that are trained directly on peptide sequences. Each of these binary sequence-level classifiers was aimed at capturing one of the following four properties of peptide sequences, namely,

- AMP/Non-AMP : Is the sequence an AMP or not?

- Toxicity/Non-Toxic : Is the sequence toxic or not?
- Broad/Narrow : Does the sequence show antibacterial activity on both Gram+ and Gram- strains or not?
- Structure/No-Structure : Does the sequence have secondary structure or not?

For each attribute, we trained a bidirectional LSTM-based classifier on the labeled dataset. We used a hidden layer size of 100 and a dropout of 0.3. Size of dataset used as well as accuracies are reported in the Table 2.

Attribute	**Data-Split**			**Accuracy (%)**		**Screening**
	Train	*Valid*	*Test*	*Majority Class*	*Test*	**Threshold**
{AMP , Non-AMP}	6489	811	812	68.9	88.0	7.944
{Toxic , Non-Toxic}	8153	1019	1020	82.73	93.7	-1.573
{Broad, Narrow}	2031	254	255	51.37	76.0	-7.323
{Structure , No-Structure}	2644	331	331	64.65	95.1	-5.382

Table 2: Performance of classifiers based on different attributes.

Classifiers	Reported Accuracy	YI12	FK13
AMP Classifier (ours)	88.00	AMP	AMP
AMP Scanner v2 (*39*)	91.00	Non-AMP	AMP
iAMP Pred (*40*)	66.80	AMP	AMP
DBAASP-SP (*41*)	79.00	AMP	Non-AMP
iAMP-2L (*42*)	92.23	Non-AMP	AMP
CAMP-RF (*43*)	87.57	AMP	AMP
Witten E.Coli (*44*)	94.30	AMP	Borderline [1]
Witten S.aureus (*44*)	94.30	AMP	AMP

Table 3: Reported accuracy and prediction for YI12 and FK13 using different AMP classifiers: AMP Scanner v2 (*39*), iAMP-2L (*40*), DBAASP Prediction (*41*), iAMPpred (*42*), CAMP-RF (*43*), Witten (CNN-70%) (*44*), and our LSTM-based sequence-based classifier.

[1] Witten model predicts logMIC activity values. We followed their criteria for classification: if logMIC is within

Based on the distribution of the scores (classification probabilities/logits), we determined the threshold by considering the 50^{th} percentile (median) of the scores (reported in the last column of Table 2). Similarly, we selected a PPL threshold of 16.04 that is the 25^{th} percentile of the PPL distribution of samples generated from the prior distribution of the best WAE model and also closer to the perplexity of our trained language model on test data.

3.2.2 CGMD Simulations - Contact Variance as a Metric for Classifying Membrane Binding

Given a peptide sequence as an input, PeptideBuilder (*45*) is used to prepare a PDB file of the all-atom representation of the peptide. This is prepared either as an alpha helix (with dihedral angles $\phi = -57$, $\psi = -47$) or as a random coil, with ϕ and ψ dihedral angles taking random values between -50° and 50°.

This initial structure is then passed as in input to `martinize.py` (*46*), which coarse-grains the system. The resulting files are passed into `insane.py` (*47*) to create the peptide membrane system. The solvent is a 90:10 ratio of water to antifreeze particles, with the membrane being a 3:1 mixture of POPC to POPG. The system is 15 nm x 15 nm x 30 nm, with the membrane perpendicular to the longest direction. Ions are added to neutralize the system.

This is a minimal model of a bacterial membrane that serves as a high-throughput filter. While this model does not replicate the complex physics of the peptide-membrane interactions in exact detail (e.g. membrane composition difference between Gram-positive *vs.* Gram-negative), it allows us to prioritise the peptides that show the strongest interaction with a membrane for experimental characterisation.

For the CGMD simulations, we used the Martini forcefield (*48*), as Martini is optimized for predicting the interactions between proteins and membranes while being computationally effi-

-1 to 3.5, then the peptide is active (AMP), if > 3.9 then inactive (Non-AMP), and between 3.5 to 3.9 is considered borderline. LogMIC values for YI12 (*E. coli*)=1.648, YI12 (*S. aureus*=3.607, FK13 (*E. coli*)=1.389, FK13 (*S. aureus*)=1.634.

cient, it is well suited for the task of a quick but physically-inspired filtering peptide sequences.

After building, the system is minimized for 50,000 steps using Gromacs 2019.1 (*49, 50*) and the 2.0 version of the Martini forcefield (*48*). After minimization, the production run is carried for 1 μs at a 20 fs timestep. Temperature is kept constant at 310 K using Stochastic Velocity Rescaling (*51*) applied independently to the protein, lipid, and the solvent groups. The pressure is kept constant at 1 atmosphere using a Parrinello-Rahman barostat (*52, 53*) applied independently to the same groups as the thermostat.

After 1 μs of sampling, we estimated the number of peptide-membrane contacts using TCL scripting and in-house Python scripts. The number of contacts between positive residues and the lipid membranes is defined as the number of atoms belonging to a lipid at a distance less than 7.5 Å from a positive residue.

For the control simulations, three datasets consisting of reported high-potency AMP, low-potency AMP, and non-AMP sequences were used that are discussed in 2.1.1. We performed a set of 130 control simulations. We found that the variance of the number of contacts (cutoff 7.5 Å) between positive residues and Martini beads of the membrane lipids is predictive of antimicrobial activity. Specifically, the contact variance distinguishes between high potency and non-antimicrobial sequences with a sensitivity of 88% and specificity of 63%. To screen, we used a cutoff value of 2 beads for the contact variance. We carried out a set of simulations for the 163 amp-positive and 179 amp-negative generated sequences. We further restricted to sequences that bind in less than 500 ns during the $1\,\mu$s long simulation, so that the contact variance is calculated over at least half of the total simulation time. Only sequences that formed at least 5 contacts (averaged over the duration of the simulation) were considered.

3.3 All-Atom Simulations

We used the CHARMM36m (*54*) forcefield to simulate the binding of four copies of YL12 and FK13 to a model membrane. Phi and Psi angles in the initial peptide structure were set to what was predicted using a recent deep learning model (*55*). A 3:1 DLPC:DLPG bilayer, with shorter tails, was used to speed up the simulation, alongside a smaller water box (98 Å) than the Martini simulations, to investigate the short-term effect of peptide-membrane interactions. A 160 ns long trajectory was run for the FK13 system. The length of the YI12 simulation was 200 ns. The number of peptide-membrane and peptide-peptide contacts (using a threshold of 7.5 Å and ignoring hydrogen atoms) were found to be converged in less time than the maximum simulation length for both systems.

The bilayer is prepared using CHARMM-GUI (*56*), and the peptide sequence is prepared using PeptideBuilder (*45*). Solvation and assembly is performed using VMD 1.9.3 (*57*). The system is simulated using NAMD 2.13 (*58*). Temperature is kept constant using a Langevin thermostat and a Nosé-Hoover Langevin piston barostat The Particle-Mesh Ewald method was used for long-range electrostatics. All simulations used a time step of 2 fs.

3.4 Peptide Sequence Analysis

Physicochemical properties like aromaticity, Eisenberg hydrophobicity, charge, charge density, aliphatic index, hydrophobic moment, hydrophobic ratio, isoelectric point, and instability index were estimated using the GlobalAnalysis method in modLAMP (*59*). Protparam tool from Expasy (https://web.expasy.org/protparam) was used to estimate the grand average of hydropathicity (GRAVY) score.

Pairwise sequence similarity was estimated using a global alignment method, the PAM30 matrix (*60*), a gap open penalty of -9, and a gap extension penalty of -1 using Pairwise2 function of Biopython package (*61*). Higher positive values indicate better similarity. To check the

correspondence between sequence similarity and Euclidean distance in $\mathbf{z}$-space, a random set of sequence encodings were first selected, and then sequence similarity and $\mathbf{z}$-distance with their close latent space neighbors were estimated. Sequence similarity with respect to a sequence database was estimated using "blastp-short" command from NCBI BLAST sequence similarity search tool (*62, 63*) was used to query generated short sequences by using a word size of 2, the PAM30 matrix (*60*), a gap open penalty of -9, a gap extension penalty of -1, threshold of 16, $comp\_based\_stats$ set to 0 and window size of 15. Alignment score (Bit score), Expect value (E-value), percentage of alignment coverage, percentage of identity, percentage of positive matches or similarity, percentage of alignment gap were used for analyzing sequence novelty. The E-value is a measure of the probability of the high similarity score occurring by chance when searching a database of a particular size. E-values decrease exponentially as the score of the match increases. For the search against patented sequences, we used the PATSEQ database (*64*).

3.5 Wet Lab Experiments

3.5.1 MIC Measurement

All of the peptides were amidated at their C-terminus to remove the negative charge of the C-terminal carboxyl group. Antimicrobial activity of the best AMP hits was evaluated against a broad spectrum of bacteria for minimum inhibitory concentration (MIC), which include Gram-positive *Staphylococcus aureus* (ATCC 29737), Gram-negative *Escherichia coli* (ATCC 25922), *Pseudomonas aeruginosa* (ATCC 9027) and multi-drug resistant *K. pneumoniae* (ATCC 700603), which produce beta-lactamase SHV 18. The broth microdilution method was used to measure MIC values of the AMPs, and the detailed protocol was reported previously (*65, 66*).

3.5.2 Hemolytic Activity

The selectivity of AMPs towards bacteria over mammalian cells was studied using rat red blood cells (rRBCs), which were obtained from the Animal Handling Unit of Biomedical Research

Center, Singapore. Negative control: Untreated rRBC suspension in phosphate-buffered saline (PBS); Positive control: rRBC suspension treated with 0.1% Triton X. Percentage of hemolysis of rRBCs was obtained using the following formula:

$$Hemolysis(\%) = \frac{O.D._{576nm} \textit{ of treated samples } - O.D._{576nm} \textit{ of negative control}}{O.D._{576nm} \textit{ of positive samples} - O.D._{576nm} \textit{ of negative control}} \quad (1)$$

3.5.3 Acute in Vivo Toxicity Analysis

The animal study protocols were approved by the Institutional Animal Care and Use Committee of Biological Resource Center, Agency for Science, technology, and Research (A*STAR), Singapore. LD50 values of the AMPs, the dose required to kill 50% mice, were determined using a previously reported protocol (*67*). Specifically, female Balb/c mice (8 weeks old, 18-22 g) were employed. AMPs were dissolved in saline and administered to mice by intraperitoneal (i.p.) injection at various doses. Mortality was monitored for 14 days post-AMP administration, and LD50 was estimated using the maximum likelihood method.

3.5.4 CD Spectroscopy

The peptides were dissolved at 0.5 mg/mL in either deionized water or deionized water containing 25 mM SDS surfactant. It forms anioic micelles in aqueous solution, which mimic the bacterial membrane. The CD spectra were measured using a CD spectropolarimeter from Jasco Corp. J-810 at room temperature and a quartz cuvette with 1 mm path length. The spectra were acquired by scanning from 190 to 260 nm at 10nm/min after subtraction with the spectrum of the solvent.

3.6 Mechanism study using Confocal microscopy

The mechanism study was performed on an FV3000 Olympus confocal laser scanning microscope. E. coli was grown according to the MIC measurement. Briefly, 500 μl of 10^8 CFU/ml of

bacteria was prepared and placed in a centrifuged tube and 500 μl of peptide and polymyxin B at 4xMIC concentration was added. The mixture was incubated at 37 °C and shaken at 100 rpm for 2 h. The sample was centrifuged at 4000 rpm for 10 min and supernatant was discarded before fresh 1 ml PBS was added. After washing for 3 times, the pallet was re-suspended with 500 μl of glutaraldehyde (2.5% v/v) for 30 min with occasional mixing. The sample was again washed for 3 times and re-suspended in 100 μl of PBS and applied onto a poly-L-lysine coated slide. The bacteria was left to attach for an hour before the unattached bacteria was washed away with PBS. The slides was left to air dry before imaging using the FV3000 confocal microscope.

3.7 Resistance Study

Drug resistance in the E. coli can be induced by treating repeatedly with peptides and imipenem at the sub MIC concentration (1/2 MIC). MIC experiment was performed as stated earlier using the micro dilution method. The next generation of E. coli can be initiated by growing the bacteria suspension from the sub MIC concentration and performing another MIC assay. After 18h inoculation, the bacteria grown from sub MIC concentration of the tested peptides/antibiotics was assayed for MIC. This will continue until we have reached 25 generations. The change in MIC was documented by taking the normalized value of MIC at generation X against the MIC at generation 1.

E-value	$<= 0.001$	$<= 0.01$	$<= 0.1$	$<= 1$	$<= 10$	> 10
Labeled	9.36	3.42	9.70	29.59	29.88	18.05
Unlabeled	5.16	4.05	9.07	30.97	36.65	14.08

Table 4: Percentage of CLaSS-generated AMP sequences in different categories of Expect value (E-value). E-value for the match with the highest score was considered, as obtained by performing BLAST similarity search against AMP-labeled training sequences (top row) and unlabeled training sequences (bottom row).

k	Generated AMP	AMP-labeled Training	Unlabeled Training
3	0.299 ± 0.006	0.223 ± 0.006	0.110 ± 0.005
4	0.771 ± 0.004	0.607 ± 0.002	0.777 ± 0.001
5	0.947 ± 0.002	0.706 ± 0.002	0.965 ± 0.000
6	0.978 ± 0.001	0.742 ± 0.002	0.983 ± 0.001

(a) Fraction of unique k-mer present in different datasets. k = 3-6. The mean and standard errors were estimated on three different sets, each consisting of 3000 randomly chosen samples.

Kmers	Generated AMP		AMP-labeled Training		Unlabeled Training	
	top-3	Frequency	top-3	Frequency	top-3	Frequency
3	KKK	0.011	KKK	0.006	LLL	0.002
	LKK	0.007	KKL	0.004	LAA	0.001
	KLK	0.007	GLL	0.004	RRR	0.001
4	KKKK	0.005	KKKK	0.003	PLDL	0.001
	KLKK	0.004	KKLL	0.002	LDLA	0.001
	KKLK	0.003	KLLK	0.002	NFPL	0.001

(b) Top 3 k-mers (k = 3 and 4) and their corresponding frequency in generated AMPs, training AMPs, and unlabeled training sequences. The estimated standard deviation values were close to zero.

	Generated AMP	AMP-labeled Training	Unlabeled Training
Charge	2.695 ± 0.039	3.074 ± 0.238	1.172 ± 0.218
Charge Density	0.002 ± 0.000	0.002 ± 0.000	0.001 ± 0.000
Aliphatic Index	107.814 ± 0.649	101.232 ± 1.550	82.088 ± 8.440
Aromaticity	0.095 ± 0.002	0.102 ± 0.002	0.082 ± 0.003
Hydrophobicity	0.048 ± 0.007	0.068 ± 0.002	0.052 ± 0.032
Hydrophobic Moment	0.331 ± 0.000	0.335 ± 0.019	0.222 ± 0.003
Hydrophobic Ratio	0.446 ± 0.001	0.431 ± 0.010	0.425 ± 0.013

(c) Physicochemical properties, such as charge, charge density, aliphatic index, aromaticity, hydrophobicity, hydrophobic moment, hydrophobic ratio, instability index, were estimated on unlabeled training, AMP-labeled training, and generated AMP sequences using CLaSS. Mean, and standard deviation were estimated on three different sets, each consisting of 3000 randomly chosen samples.

Figure 1: Comparison of CLaSS-generated AMPs with AMP-labeled and unlabeled peptides.

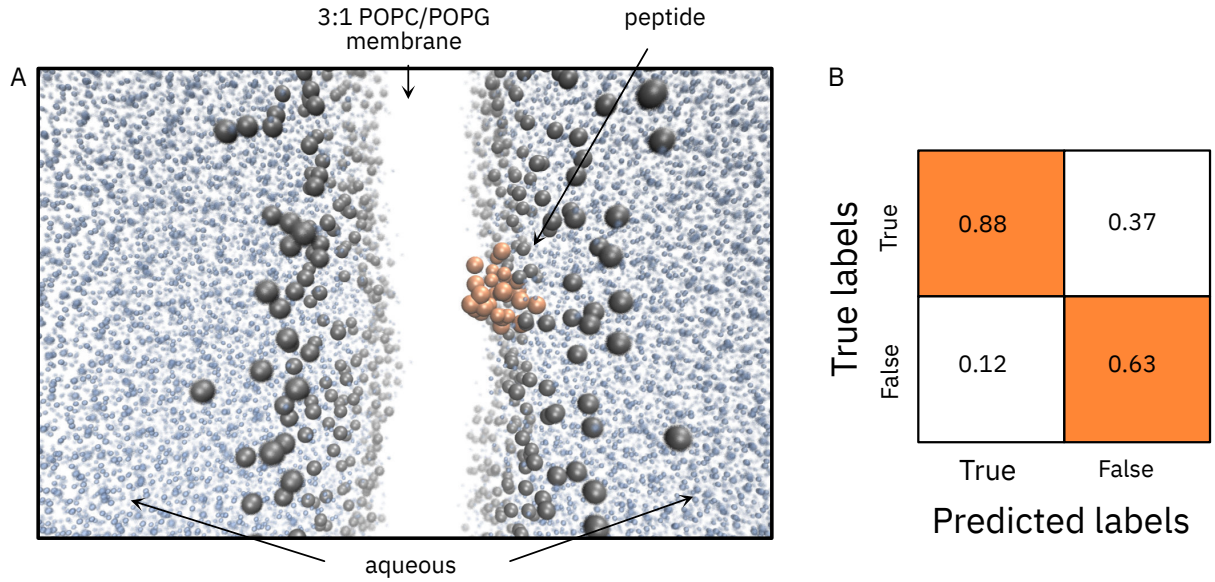
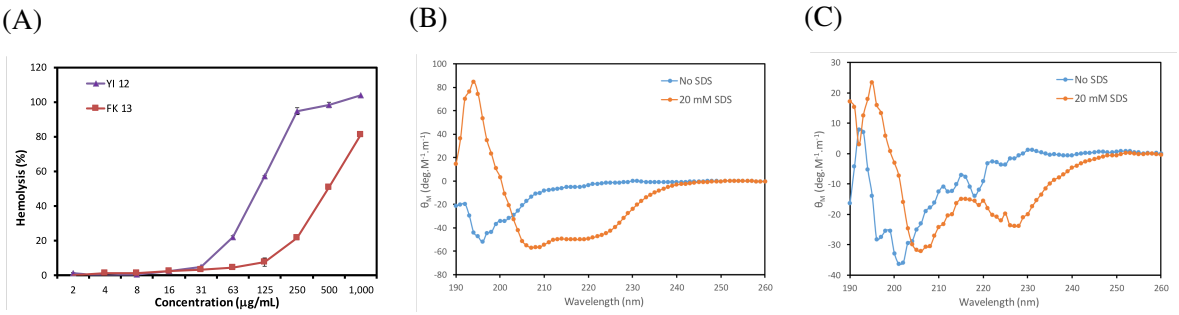


Figure 2: (A) Snapshot from a coarse-grained molecular dynamics simulation of an AMP (in orange) binding with a lipid bilayer (gray). (B) Confusion matrix of the simulation-based classifier that uses peptide-membrane contact variance as feature for detecting AMP sequences.



Sequence	Score	E-Value	% Coverage	% Identity	% Positive	% Gap
YLRLIRYMAKMI	21.88	1.20	66.67	75.00	75.00	75.00
FPLTWLKWWKWKK	33.30	4e-05	84.61	73.00	73.00	9.00

Figure 3: Percentage of hemolysis of rat red blood cells as a function of peptide concentration. (B) and (C) show CD Spectra of YI12 and FK13 peptide, respectively, at 0.5 mg/ml concentration in DI water and presence of 20 mM SDS buffer. Both YI12 and FK13 showed a random coil-like structure in the absence of SDS. When SDS was present, both sequences form α-helical structure (evident from the 208 nm and 222 nm peaks). (D) BLAST search results (alignment score, E-value, percentage of alignment coverage, percentage of identity, percentage of positive matches or similarity, percentage of alignment gap) against full training data for YI12 and FK13.

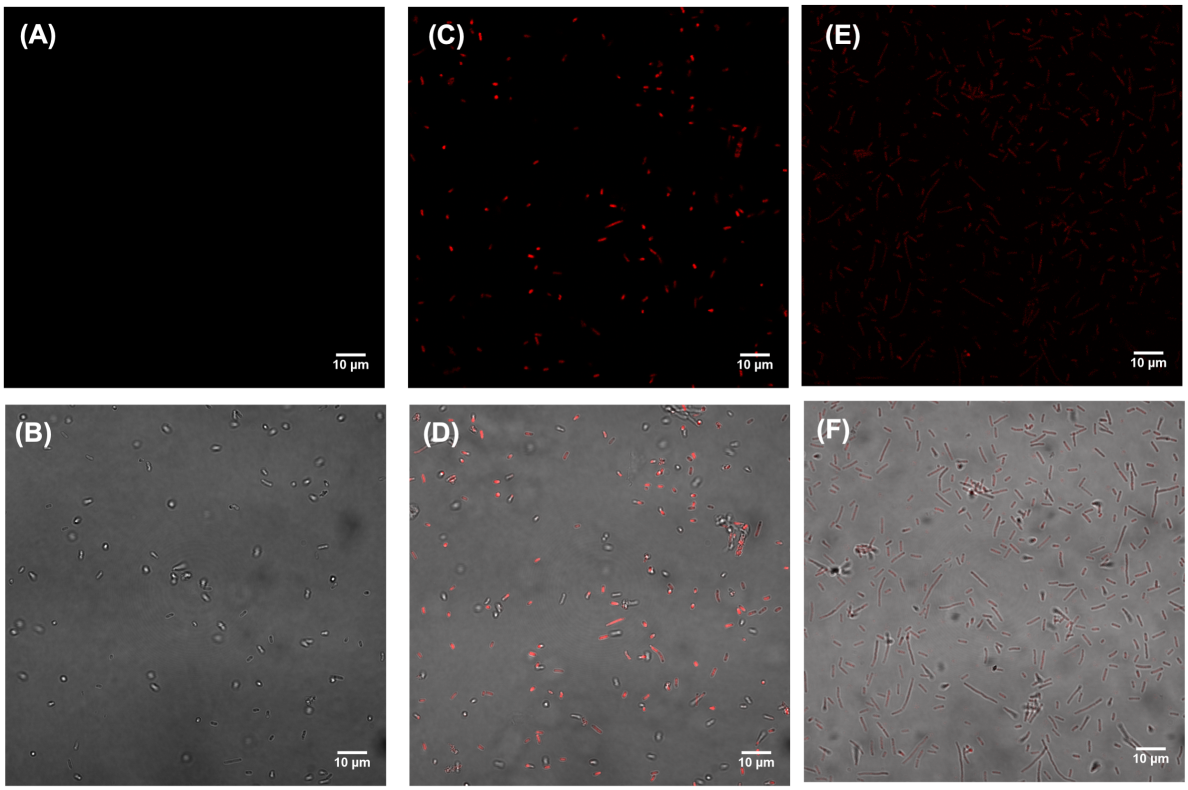


Figure 4: Confocal images of *E. coli* (A-B) without any treatment, (C-D) treated with 2xMIC FK13 and (E-F) treated with 2xMIC polymyxin B for 2 hours. Red colour denotes the propidium iodide. Images A, C and E are the dark field image and B, D and F are the bright field image. All scale bars are 10 μm.

.

Blast search results of YI12 against 223.5 M Uniprot sequences

>WP_027115309.1 EAL domain-containing protein [Lachnospiraceae bacterium P6B14]

Length=643

Score = 33.3 bits (71), Expect = 2.9 , Coverage= 11/12 (91.6%)

Identities = 9/12 (75%), Positives = 10/12 (83%), Gaps = 1/12 (8%)

```
Query 1    YLRLIRYM-AKM 11
           YLR+IRYM  KM
Sbjct 353 YLRMIRYMESKM 364
```

Blast search results of FK13 against 0.18 M training sequences

>Synthetic variant of PIN-A segment of Puroindoline-A

Length=13

Score = 33.3 bits (71), Expect = 1e-06, Coverage= 11/13 (84.6%)

Identities = 8/11 (73%), Positives = 8/11 (73%), Gaps = 1/11 (9%)

```
Query  1   FPLTWLKWWKW  11
           FPLTW  WWKW
Sbjct  1   FPLTW-RWWKW  10
```

Figure 5: BLAST search results of YI12 and FK13.

.

Sequence	Positive Residue	Binding time (ns)	Mean	Variance
YLRLIRYMAKMI	3	210	6.45	1.27
FPLTWLKWWKWKK	4	90	5.90	1.39
HILRMRIRQMMT	3	17	7.84	1.44
ILLHAILGVRKKL	3	105	7.16	1.19
YRAAMLRRQYMMT	3	19	8.79	1.25
HIRLMRIRQMMT	3	493	8.38	1.50
HIRAMRIRAQMMT	3	39	7.20	1.39
KTLAQLSAGVKRWH	3	177	7.62	1.46
HILRMRIRQGMMT	3	62	8.37	1.53
HRAIMLRIRQMMT	3	297	7.46	1.35
EYLIEVRESAKMTQ	2	150	6.65	1.79
GLITMLKVGLAKVQ	2	341	8.34	1.58
YQLLRIMRINIA	2	239	6.29	1.71
VRWIEYWREKWRT	4	125	6.41	1.28
LIQVAPLGRLLKRR	4	37	6.52	1.24
YQLRLIMKYAI	2	192	7.75	1.86
HRALMRIRQCMT	3	80	9.15	1.27
GWLPTEKWRKLC	3	227	6.11	1.63
YQLRLMRIMSRI	3	349	8.28	1.80
LRPAFKVSK	3	151	7.73	1.85

Table 5: Physics-derived features such as mean and variance of the number of contacts between positive amino acids and membrane beads (that are found to be associated with antimicrobial function in this study), as extracted from CGMD simulations of peptide membrane interactions for the top 20 AI-designed AMP sequences.

Sequence	Length	Charge	H	μH
YLRLIRYMAKMI	12	3.99	0.08	0.79
FPLTWLKWWKWKK	13	5.00	0.05	0.20
HILRMRIRQMMT	12	4.10	-0.25	0.36
ILLHAILGVRKKL	13	4.09	0.27	0.33
YRAAMLRRQYMMT	13	3.99	-0.28	0.06
HIRLMRIRQMMT	12	4.10	-0.25	0.16
HIRAMRIRAQMMT	13	4.10	-0.22	0.24
KTLAQLSAGVKRWH	14	4.09	-0.08	0.49
HILRMRIRQGMMT	13	4.10	-0.19	0.27
HRAIMLRIRQMMT	13	4.10	-0.18	0.41
EYLIEVRESAKMTQ	14	0.00	-0.16	0.26
GLITMLKVGLAKVQ	14	3.00	0.37	0.28
YQLLRIMRINIA	12	3.00	0.11	0.38
VRWIEYWREKWRT	13	3.00	-0.41	0.55
LIQVAPLGRLLKRR	14	5.00	-0.12	0.38
YQLRLIMKYAI	11	2.99	0.18	0.40
HRALMRIRQCMT	12	4.03	-0.34	0.56
GWLPTEKWRKLC	12	2.93	-0.13	0.33
YQLRLMRIMSRI	12	4.00	-0.17	0.64
LRPAFKVSK	9	4.00	-0.17	0.70

Table 6: Physico-chemical properties for the top 20 AI-designed AMP sequences.

Sequence	S. aureus (μg/mL)	E.Coli (μg/mL)
YLRLIRYMAKMI-CONH2	7.8	31.25
FPLTWLKWWKWKK-CONH2	15.6	31.25
HILRMRIRQMMT-CONH2	$>$1000	$>$1000
ILLHAILGVRKKL-CONH2	250	250
YRAAMLRRQYMMT-CONH2	$>$1000	$>$1000
HIRLMRIRQMMT-CONH2	$>$1000	$>$1000
HIRAMRIRAQMMT-CONH2	$>$1000	1000
KTLAQLSAGVKRWH-CONH2	$>$1000	$>$1000
HILRMRIRQGMMT-CONH2	$>$1000	$>$1000
HRAIMLRIRQMMT-CONH2	$>$1000	$>$1000
EYLIEVRESAKMTQ-CONH2	$>$1000	$>$1000
GLITMLKVGLAKVQ-CONH2	$>$1000	$>$1000
YQLLRIMRINIA-CONH2	$>$1000	$>$1000
VRWIEYWREKWRT-CONH2	$>$1000	$>$1000
YQLRLIMKYAI-CONH2	125	125
HRALMRIRQCMT-CONH2	1000	1000
GWLPTEKWRKLC-CONH2	1000	$>$1000
YQLRLMRIMSRI-CONH2	250	500
FFPLPAISTELKRL-CONH2	$>$1000	$>$1000
LIQVAPLGRLLKRR-CONH2	$>$1000	1000
LRPAFKVSK-CONH2	$>$1000	$>$1000

Table 7: Broad-spectrum MIC values of top 20 AI-designed AMP Sequences

Sequence	S. aureus (μg/mL)	E.Coli (μg/mL)
AMLELARIIGRR-CONH2	$>$1000	$>$1000
IPRPGPFVDPRSR-CONH2	$>$1000	$>$1000
VAKVFRAPKVPICP-CONH2	$>$1000	$>$1000
FPSFTFRLRKWKRG-CONH2	62.5	62.5
RPPFGPPFRR-CONH2	$>$1000	$>$1000
WEEMDSLRKWRIWS-CONH2	$>$1000	$>$1000
RRQAQEVRGPRH-CONH2	$>$1000	$>$1000
KKKKPLTPDFVFF-CONH2	$>$1000	$>$1000
TRGPPPTFRAFR-CONH2	$>$1000	$>$1000
LALHLEALIAGRR-CONH2	250	$>$1000

Table 8: Broad-spectrum MIC values of 11 AI-designed Non-AMP Sequences

References

1. L. Yu, W. Zhang, J. Wang, Y. Yu, *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

2. G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, *arXiv preprint arXiv:1705.10843* (2017).

3. E. Jang, S. Gu, B. Poole, *arXiv preprint arXiv:1611.01144* (2016).

4. M. J. Kusner, J. M. Hernández-Lobato, *arXiv preprint arXiv:1611.04051* (2016).

5. C. J. Maddison, A. Mnih, Y. W. Teh, *arXiv preprint arXiv:1611.00712* (2016).

6. Y. Zhang, Z. Gan, L. Carin, *NIPS workshop on Adversarial Training* (2016).

7. D. P. Kingma, S. Mohamed, D. J. Rezende, M. Welling, *Advances in Neural Information Processing Systems* (2014), pp. 3581–3589.

8. Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, E. P. Xing, *International Conference on Machine Learning* (2017), pp. 1587–1596.

9. J. Engel, M. Hoffman, A. Roberts, *arXiv preprint arXiv:1711.05772* (2017).

10. S. Dathathri, *et al.*, *arXiv preprint arXiv:1912.02164* (2019).

11. R. Gómez-Bombarelli, *et al.*, *ACS central science* **4**, 268 (2018).

12. Z. Zhou, S. Kearnes, L. Li, R. N. Zare, P. Riley, *Scientific reports* **9**, 10752 (2019).

13. J. You, B. Liu, Z. Ying, V. Pande, J. Leskovec, *Advances in Neural Information Processing Systems* (2018), pp. 6410–6421.

14. M. Popova, O. Isayev, A. Tropsha, *Science advances* **4**, eaap7885 (2018).

15. A. Zhavoronkov, *et al.*, *Nature Biotechnology* **37**, 1038 (2019).

16. K. Korovina, *et al.*, *arXiv preprint arXiv:1908.01425* (2019).

17. J. Lim, S. Ryu, J. W. Kim, W. Y. Kim, *Journal of cheminformatics* **10**, 31 (2018).

18. S. Kang, K. Cho, *Journal of chemical information and modeling* **59**, 43 (2018).

19. Y. Li, L. Zhang, Z. Liu, *Journal of cheminformatics* **10**, 33 (2018).

20. O. Méndez-Lucio, B. Baillif, D.-A. Clevert, D. Rouquié, J. Wichard, *Nature Communications* **11**, 1 (2020).

21. P. EMBL-EBI, SIB, Universal Protein Resource (UniProt), `https://www.uniprot.org` (2018). [Online; accessed August-2018].

22. S. Singh, *et al.*, *Nucleic acids research* (2015).

23. M. Pirtskhalava, *et al.*, *Nucleic acids research* **44**, D1104 (2016).

24. S. Khurana, *et al.*, *Bioinformatics* **34**, 2605 (2018).

25. P. Bhadra, J. Yan, J. Li, S. Fong, S. W. Siu, *Scientific reports* (2018).

26. S. Gupta, *et al.*, *PloS one* **8**, e73957 (2013).

27. D. Frishman, P. Argos, *Proteins* **23**, 566–579 (1995).

28. S. Bowman, *et al.*, *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (2016), pp. 10–21.

29. I. Tolstikhin, O. Bousquet, S. Gelly, B. Schölkopf, *Stat* **1050**, 12 (2018).

30. L. Theis, A. v. d. Oord, M. Bethge, *ICLR* (2016).

31. A. A. Alemi, *et al.*, *arXiv preprint arXiv:1711.00464* (2017).

32. T. Sercu, *et al.*, *ICLR Workshop on Deep Generative Models for Highly Structured Data* (2019).

33. P. K. Rubenstein, B. Schoelkopf, I. Tolstikhin, *arXiv preprint arXiv:1802.03761* (2018).

34. M. Ranzato, S. Chopra, M. Auli, W. Zaremba, *arXiv preprint arXiv:1511.06732* (2015).

35. S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, *Advances in Neural Information Processing Systems* (2015), pp. 1171–1179.

36. J. J. Zhao, Y. Kim, K. Zhang, A. Rush, Y. LeCun, *arXiv preprint arXiv:1706.04223* (2017).

37. S. Merity, N. S. Keskar, R. Socher, *arXiv preprint arXiv:1708.02182* (2017).

38. A. T. Mueller, J. A. Hiss, G. Schneider, *Journal of Chemical Information and Modeling* (2018).

39. D. Veltri, U. Kamath, A. Shehu, *Bioinformatics* **34**, 2740 (2018).

40. P. K. Meher, T. K. Sahu, V. Saini, A. R. Rao, *Scientific reports* **7**, 42362 (2017).

41. B. Vishnepolsky, *et al.*, *Journal of chemical information and modeling* **58**, 1141 (2018).

42. X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, K.-C. Chou, *Analytical biochemistry* **436**, 168 (2013).

43. S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, S. Idicula-Thomas, *Nucleic acids research* **38**, D774 (2010).

44. J. Witten, Z. Witten, *bioRxiv* (2019).

45. M. Z. Tien, D. K. Sydykova, A. G. Meyer, C. O. Wilke, *PeerJ* **1**, e80 (2013).

46. D. H. de Jong, *et al.*, *Journal of Chemical Theory and Computation* **9**, 687 (2012).

47. T. A. Wassenaar, H. I. Ingólfsson, R. A. Böckmann, D. P. Tieleman, S. J. Marrink, *Journal of Chemical Theory and Computation* **11**, 2144 (2015).

48. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, *The Journal of Physical Chemistry B* **111**, 7812 (2007).

49. H. Berendsen, D. van der Spoel, R. van Drunen, *Computer Physics Communications* **91**, 43 (1995).

50. M. J. Abraham, *et al.*, *SoftwareX* **1-2**, 19 (2015).

51. G. Bussi, D. Donadio, M. Parrinello, *The Journal of Chemical Physics* **126**, 014101 (2007).

52. M. Parrinello, A. Rahman, *Journal of Applied Physics* **52**, 7182 (1981).

53. S. Nosé, M. Klein, *Molecular Physics* **50**, 1055 (1983).

54. J. Huang, *et al.*, *Nature Methods* **14**, 71 (2016).

55. Z. Qin, *et al.*, *Extreme Mechanics Letters* p. 100652 (2020).

56. S. Jo, J. B. Lim, J. B. Klauda, W. Im, *Biophysical Journal* **97**, 50 (2009).

57. W. Humphrey, A. Dalke, K. Schulten, *Journal of Molecular Graphics* **14**, 33 (1996).

58. J. C. Phillips, *et al.*, *Journal of Computational Chemistry* **26**, 1781 (2005).

59. A. T. Müller, G. Gabernet, J. A. Hiss, G. Schneider, *Bioinformatics* **33**, 2753 (2017).

60. Y.-K. Yu, J. C. Wootton, S. F. Altschul, *Proceedings of the National Academy of Sciences* **100**, 15688 (2003).

61. P. J. Cock, *et al.*, *Bioinformatics* **25**, 1422 (2009).

62. T. Madden, *The NCBI Handbook [Internet]. 2nd edition* (National Center for Biotechnology Information (US), 2013).

63. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Journal of molecular biology* **215**, 403 (1990).

64. PATSEQ, `https://www.lens.org/lens/bio/`. [Online].

65. W. Chin, *et al.*, *Nature communications* **9**, 1 (2018).

66. V. W. L. Ng, X. Ke, A. L. Lee, J. L. Hedrick, Y. Y. Yang, *Advanced Materials* **25**, 6730 (2013).

67. S. Liu, *et al.*, *Biomaterials* **127**, 36 (2017).