Bioorganic & Medicinal Chemistry 26 (2018) 2708–2718

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Machine learning-enabled discovery and design of membrane-active peptides

Ernest Y. Lee [a], Gerard C.L. Wong [a,b,*], Andrew L. Ferguson [c,d,*]

[a] Department of Bioengineering, University of California, Los Angeles, CA 90095, United States
[b] California NanoSystems Institute, Los Angeles, CA 90095, United States
[c] Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States
[d] Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

ARTICLE INFO

Article history:
Received 17 May 2017
Revised 29 June 2017
Accepted 6 July 2017
Available online 8 July 2017

Keywords:
Machine learning
Quantitative structure activity relationship models
Antimicrobial peptides
Cell-penetrating peptides
Membrane-active peptides

ABSTRACT

Antimicrobial peptides are a class of membrane-active peptides that form a critical component of innate host immunity and possess a diversity of sequence and structure. Machine learning approaches have been profitably employed to efficiently screen sequence space and guide experiment towards promising candidates with high putative activity. In this mini-review, we provide an introduction to antimicrobial peptides and summarize recent advances in machine learning-enabled antimicrobial peptide discovery and design with a focus on a recent work Lee et al. *Proc. Natl. Acad. Sci. USA* 2016;113 (48):13588–13593. This study reports the development of a support vector machine classifier to aid in the design of membrane active peptides. We use this model to discover membrane activity as a multiplexed function in diverse peptide families and provide interpretable understanding of the physicochemical properties and mechanisms governing membrane activity. Experimental validation of the classifier reveals it to have learned membrane activity as a unifying signature of antimicrobial peptides with diverse modes of action. Some of the discriminating rules by which it performs classification are in line with existing "human learned" understanding, but it also unveils new previously unknown determinants and multidimensional couplings governing membrane activity. Integrating machine learning with targeted experimentation can guide both antimicrobial peptide discovery and design and new understanding of the properties and mechanisms underpinning their modes of action.

1. Structure of this mini-review

The tandem expansion of experimental databases of antimicrobial peptides (AMPs) and maturation of robust machine learning algorithms has led to profitable synergies in which computational models trained on large and high-quality data sets can perform high-throughput "virtual screening" to guide the discovery and design of novel AMPs. Predictive computational models can serve as fast and inexpensive pre-screening tools to efficiently traverse the combinatorially vast sequence space and direct time, labor, and cost-intensive experimentation towards promising candidates with high putative activity. Beyond computational hit finding, machine learning models can also furnish new understanding of the underlying peptide properties underpinning antimicrobial activity and inform experiments to validate and calibrate these predictions. In this mini-review we provide an introduction to antimicrobial peptides and recent advances in machine learning-enabled AMP design, with a particular focus on a recent publication reporting the development of machine learning classifiers designed not only to aid in peptide discovery but also provide new understanding of the common physicochemical determinants underpinning the activity of this diverse group of peptides.[1] The present mini-review foregrounds the computational aspects of this work; another recent invited mini-review takes a more experimental vantage.[2] We first discuss the new insights furnished by the

Abbreviations: AMP, antimicrobial peptide; ANN, artificial neural network; AUROC, area under the receiver operating characteristic; HMM, hidden Markov model; k-NN, k-nearest neighbor; MCC, Matthews correlation coefficient; MCMC, Markov Chain Monte Carlo; MIC, minimum inhibitory concentration; NGC, negative Gaussian curvature; NPV, negative predictive value; PPV, positive predictive value; QM, quantitative matrix; QSAR, quantitative structure–activity relationship; RF, random forest; SAXS, small angle X-ray scattering; SOM, self-organizing map; SVM, support vector machine.

* Corresponding authors at: Department of Bioengineering, University of California, Los Angeles, CA 90095, United States (G.C.L. Wong). Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States (A.L. Ferguson).

E-mail addresses: gclwong@seas.ucla.edu (G.C.L. Wong), alf@illinois.edu (A.L. Ferguson).

http://dx.doi.org/10.1016/j.bmc.2017.07.012
0968-0896/

machine learning model, some consistent with existing "human learned" understanding and some entirely new. We then demonstrate its utility in efficiently screening the combinatorially vast sequence space to design new non-natural membrane-active peptides and discover membrane activity in diverse families of peptides with established primary functions. We close with an outlook and perspective for this rapidly evolving field.

2. Antimicrobial peptide structure and function

Antimicrobial peptides are a class of short peptides with the capacity to disrupt and/or penetrate microbial membranes and induce cell death.[3–8] In excess of 2,000 naturally occurring and synthetic AMPs have been experimentally defined, with a large fraction forming part of the innate immune response.[3–11] AMPs tend to be short (<50 amino acids), positively charged (+2 to +9), and facially amphipathic.[3–8] There is evidence for a variety of modes of action by which AMPs disrupt the membrane – including the barrel stave, toroidal pore, and carpet mechanisms[4,12–15] – and effect their microbicidal activity – including membrane depolarization, leakage of cell contents, disruption of intracellular function, and immunomodulation.[4,8,16–22] Regardless of the precise mode of action, membrane activity is a critical prerequisite to antimicrobial activity. In general, AMPs tend to bind to prokaryotic cell membranes due to strong Coulombic attractions between cationic peptide residues and anionic lipid head groups. Their amphipathic character and relatively small size permits them to embed into the lipid bilayer with the hydrophobic face favorably interacting with the lipid tails and hydrophilic face with the head groups. Ultimately, this leads to membrane disruption, membrane permeation, and/or peptide translocation resulting in cell death.[13] However, the enormous diversity in sequence, secondary structure,[3,8] and modes of action[16–22] makes it challenging to define more precise determinants of antimicrobial activity to serve as actionable precepts for AMP design. Machine learning models trained to recognize antimicrobial activity can be of great value in advancing understanding and accelerating AMP discovery.

3. Prior applications of machine learning to antimicrobial peptide discovery

Machine learning models of AMP activity come in many forms and employ diverse mathematical approaches. All techniques have the same fundamental goal of predicting antimicrobial activity (y) based on properties of the peptide ($\mathbf{x}$). As such, they fall under the umbrella of quantitative structure–activity relationship (QSAR) models. Regression models seek to predict the strength of the antimicrobial activity – measured by, for example, the minimum inhibitory concentration (MIC) – whereas classification models seek to distinguish candidates as either hits or misses – based, for example, on some threshold in the MIC. In contrast to physical models, machine learning models are data-driven in the sense that they seek to infer a relationship between $\mathbf{x}$ and y by statistical learning over characterized experimental databases. Training models over datasets in this manner in which inputs and outputs are known is known as *supervised learning*.[23] To be useful as a predictive model, four criteria must be satisfied. First, a relationship between the peptide properties $\mathbf{x}$ – the *features* or *descriptors* in machine learning parlance – and the antimicrobial activity y – the *response* variable – must exist and must be detectable by the training algorithm. In general, care must be taken not to neither underfit the data, as this results in overly simplistic models with poor predictive capacity, nor overfit, as this results in overly complex models with poor generalizability. This can be framed as a bias-variance tradeoff. Underfitted models contain high bias and low variance, whereas overfitted models possess low bias and high variance.[23] In general, one employs some from of cross-validation to tune the model parameters to minimize the error over a hold-out data subset to optimize this tradeoff.[23] Second, sufficiently large databases containing candidates representative of the distribution in the population must exist over which to conduct model training and validation. While it is difficult to predict *a priori* how big is big enough, *post hoc* testing of the model against a hold out set can typically inform expected model performance. Third, the trained machine learning model must provide useful predictions, where "useful" is dependent on the context and intended application of the model. In some machine learning applications high accuracy is required (e.g., self-driving cars), whereas in others it may be sufficient to simply do better than average to provide a competitive edge (e.g., financial predictions). Alternatively, high *sensitivity* is required where the intended goal requires a low frequency of false negatives (type II errors). High sensitivity tests allow the user to confidently rule out the occurrence of an event upon showing a negative result (e.g., a pregnancy test). Conversely, high *specificity* is the priority where the application requires a low frequency of false positives (type I errors). High specificity tests allow the operator to confidently rule in the occurrence of an event upon showing a positive result (e.g., confirming the presence of disease). Fourth, the descriptors should be cheaper and faster to compute and/or measure than the response variable itself, or otherwise the QSAR model is typically rendered redundant.

Enabled by the advent of robust machine learning algorithms and large AMP databases (www.camp.bicnirrh.res.in/exLinks.php),[24,9–11] a body of work emerged beginning in the mid-2000s reporting the development of high-performance QSAR models to predict AMP activity. The preponderance of these studies have focused on the development of predictive models to perform efficient *in silico* screening of large ensembles of peptide sequences to identify candidates with high putative activity. A number of machine learning algorithms have been deployed for this purpose, although there is no clear consensus of the superiority of any one algorithm. Lata et al. reported an approach based on an artificial neural network (ANN), support vector machine (SVM), and quantitative matrix (QM) model to classify AMPs based on C- and N-terminal residues.[25] Fjell et al. employed a hidden Markov model (HMM) to discover a novel bovine AMP.[26] Cherkasov et al. and Fjell et al. trained ANNs to perform *in silico* screening of 100,000 peptide candidates to discover two peptides with higher potency against multi-drug resistant "superbugs" than existing antibiotic therapies and AMPs in clinical trials.[27,28] Wang et al. integrated BLASTP sequence alignment with amino acid composition descriptors to develop an AMP classifier with ~80% prediction accuracy,[29] and Torrent et al. developed an 8-descriptor SVM classifier with up to 90% accuracy.[30] Xiao et al. developed a fuzzy k-nearest neighbor (k-NN) classifier to identify AMPs and bin them into one or more of ten sub-categories: antibacterial, anticancer, antifungal, anti-HIV, antiviral, antiparasital, anti-protist, chemotactic, insecticidal, and spermicidal.[31] Maccari et al. trained a random forest (RF) to design and experimentally validate two *de novo* AMPs, enhance the activity of an existing AMP and engineer a novel AMP containing non-natural amino acids.[32] Giguere et al. developed a graph-based approach to identify and test four peptides with high *in vitro* activity.[33] Schneider et al. reported a two-step process wherein a self-organizing map (SOM) was used to perform nonlinear dimensionality reduction over 147 peptide descriptors as a pre-processing step prior to ANN classification.[34]

In our own recent work, we developed a SVM classifier to predict α-helical AMP activity with ~90% accuracy.[1] In a departure from the typical goals of QSAR model development, our intent was to develop a machine learning classifier that not only had high predictive accuracy and specificity (i.e., low false positive rate) but

to also provide interpretable insight into the underlying physicochemical determinants of antimicrobial activity. To this end, we purposefully selected a linear SVM classifier as our QSAR model for its transparent interpretability in linking the descriptors with the predicted response. We performed rigorous filtering and embedded feature selection over an initial candidate set of 1588 physicochemical descriptors to distill those 12 most predictive of antimicrobial activity. Some of these features are in line with previously "human learned" determinants of AMP activity, while others provide new understanding. We compared our classifier predictions against small angle X-ray scattering (SAXS) experiments to both validate its predictions and connect the identified physicochemical peptide properties with induced structural changes in the target membranes and thus the AMP mode of action. Unexpectedly, these guided experiments exposed our classifier not to have learned to predict antimicrobial activity based on physicochemical signatures of "antimicrobial-ness", but rather membrane activity as a unifying property of AMPs with diverse modes of action. This result provides a clear illustration that the rules learned during QSAR training may be adjacent to those initially intended, and that calibrating experiments are essential in both validating classifier performance and revealing its mechanistic basis. We subsequently exploited our model in light of this new understanding to identify novel membrane active peptides and discover membrane activity in diverse peptide families with other putative primary functions. In the remainder of this mini-review, we offer a fuller examination and appraisal of this work.

4. Training a support vector machine to distinguish antimicrobial activity

4.1. Introduction to support vector machines

A support vector machine (SVM) classifier is a machine learning algorithm designed to perform deterministic classification of data into one of two distinct categories.[35,36] The essence of the approach is to project each data point into a m-dimensional Euclidean space in which its location is defined by a list of m ordered features. The SVM is trained to define a $(m-1)$-dimensional hyperplane that optimally partitions the data into the two distinct classes by maximizing the distance (i.e., margin) from the closest point in each class to the separating hyperplane. This optimal hyperplane is known as the *maximum-margin hyperplane* and is determined by supervised training of the SVM classifier over a labeled set of training data in which the features and classifications of the data are known. An schematic illustration of a linear SVM classifier for $m = 2$ is presented in Fig. 1.

Mathematically, a data point i in the m-dimensional feature space can be represented by an m-element vector $\mathbf{x}_i$. A $(m-1)$-dimensional hyperplane in the feature space is defined by its surface normal vector $\mathbf{w}$ and offset from the origin $b/\,||\,\mathbf{w}\,||$. This plane comprises the locus of points $\mathbf{x}$ satisfying the relation $\mathbf{w}.\mathbf{x} - b = 0$. A data point i is classified by which side of the hyperplane upon which it falls, defined as a "hit" if $\mathbf{w}.\mathbf{x_i} - b > 0$ and a "miss" if $\mathbf{w}.\mathbf{x_i} - b < 0$. The margins around the hyperplane are defined by the parallel hyperplanes $\mathbf{w}.\mathbf{x}_+ - b = +1$ containing the closest "hit(s)" and $\mathbf{w}.\mathbf{x}_- - b = -1$ containing the closest "miss(es)". The size of the gap is defined by the projection of the vector $(\mathbf{x}_+ - \mathbf{x}_-)$ onto the unit vector $\hat{\mathbf{w}} = \mathbf{w}/\,||\,\mathbf{w}\,||$,

$$\begin{aligned}\hat{\mathbf{w}}.(\mathbf{x}_+ - \mathbf{x}_-) &= \frac{1}{||\,\mathbf{w}\,||}(\mathbf{w}.\mathbf{x}_+ - \mathbf{w}.\mathbf{x}_-)\\ &= \frac{1}{||\,\mathbf{w}\,||}[(\mathbf{w}.\mathbf{x}_+ - b) - (\mathbf{w}.\mathbf{x}_- - b)]\\ &= \frac{1}{||\,\mathbf{w}\,||}[1-(-1)] = \frac{2}{||\,\mathbf{w}\,||}. \end{aligned} \quad (1)$$

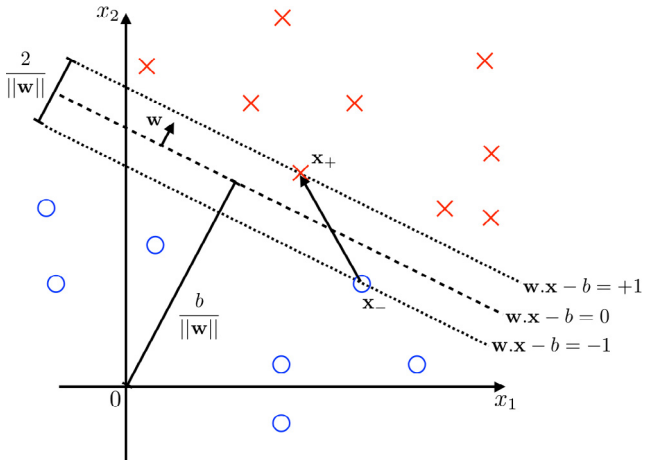

Fig. 1. Schematic illustration of a support vector machine (SVM) classifier operating in a feature space of dimensionality $m = 2$. Data points $\mathbf{x}$ are represented in this space by their $m = 2$-dimensional feature vetors. The maximum margin hyperplane defines a $(m-1)$-dimensional surface that maximally separates the two classes "hits" (red crosses) and "misses" (blue circles). Mathematically, the hyperplane is defined by the locus of points $\mathbf{x}$ satisfying $\mathbf{w}.\mathbf{x} - b = 0$, where $\mathbf{w}$ is the (possibly non-unit) surface normal, $\hat{\mathbf{w}} = \mathbf{w}/\,||\,\mathbf{w}\,||$ is the corresponding unit vector, and $\hat{\mathbf{w}}.\mathbf{x} = b/\,||\,\mathbf{w}\,||$ is the offset of the hyperplane from the origin. The margin is defined by the two parallel hyperplanes satisfying $\mathbf{w}.\mathbf{x} - b = +1$ and $\mathbf{w}.\mathbf{x} - b = -1$, and the SVM is trained by maximizing the width of the margin $\hat{\mathbf{w}}.(\mathbf{x}_+ - \mathbf{x}_-) = \frac{2}{||\mathbf{w}||}$, where $\mathbf{x}_+$ is the closest "hit" to the hyperplane separator and $\mathbf{x}_-$ the closest "miss". For data that are not linearly separable, the hard margin formulation (Eq. (2)) is supplanted by the soft margin (Eq. (3)) version that allows for classification errors.

The maximum margin hyperplane maximizes this gap and the corresponding maximum-margin SVM classifier is defined by the following optimization,[35,23,37]

$$\arg\min_{\mathbf{w},b} \frac{1}{2}||\,\mathbf{w}\,|| \quad \text{subject to } y_i(\mathbf{w}.\mathbf{x}_i - b) \geqslant 1, \quad i = 1\ldots n. \quad (2)$$

This formulation assumes that the data are linearly separable such that each point may be correctly classified, and is known as a *hard margin* SVM. In general, the data may not admit perfect linear separation and we appeal to the *soft margin* formulation that allows for classification errors,[35,23,37]

$$\begin{aligned}&\arg\min_{\mathbf{w},b}\left[\frac{1}{2}||\,\mathbf{w}\,|| + \lambda\frac{1}{n}\sum_{i=1}^{n} L_1(y_i,\mathbf{x}_i)\right],\\ &L_1(y_i,\mathbf{x}_i) = \max(0, 1 - y_i(\mathbf{w}.\mathbf{x}_i - b)),\end{aligned} \quad (3)$$

where $L_1(y_i.\mathbf{x}_i)$ defines a *hinge loss* function that penalizes points that fall into or beyond the margin (i.e., those with $y_i(\mathbf{w}.\mathbf{x}_i - b) < 1$) in linear proportion to how far across the margin they reside, and λ is a hyperparameter controlling the relative weighting between maximizing the margin and incorrect predictions that is typically tuned by cross-valuation.[23,37] The minimization in Eq. (3) is known as the *primal problem*, and training of the SVM classifier amounts to determining the values of $\mathbf{w}$ and b that minimize the objective function over all training points $i = 1\ldots n$. In practice, it is convenient to reformulate this problem using Lagrangian multipliers to obtain an equivalent *dual problem* that admits efficient solutions by quadratic programming.[23]

SVMs are an inherently linear classification technique that discriminate class membership based on a hyperplane separator. This linearity is advantageous in providing intuitive understanding of the classification rationale and the relative importance and relationship between the features of the data. We exploited this attractive feature to aid in our interpretation of the underlying physicochemical determinants discovered by the classifier.[1] Nevertheless, SVMs can perform poorly on data sets that are not linearly

separable. Nonlinear generalizations exist in which the data are projected by the so-called "kernel trick" into a high-dimensional (or even infinite-dimensional) space and a linear hyperplane constructed in this transformed feature space.[36,38,23] If the transformation is nonlinear, then the separating hyperplane is a nonlinear surface in the original space that can provide greater flexibility in separating the data. Finally, we note that generalizations of SVMs exist to handle multi-class data, typically through nested binary classifications,[39,40] which may be beneficial in sub-categorizing AMPs as demonstrated by Xiao et al.[31]

4.2. Curation of training data

The positive data set for training of our linear SVM comprised 286 AMPs downloaded from the Antimicrobial Peptide Database (http://aps.unmc.edu/AP)[9–11] for which microbicidal activity had been experimentally confirmed by plate killing or broth microdilution. The negative data set comprised 286 membrane active proteins reported to possess no antimicrobial activity that were downloaded from the Protein Data Bank of Transmembrane Proteins (http://pdbtm.enzim.hu).[41–43] Both datasets comprised peptides of length 8–60 residues and α-helical secondary structure. We randomly selected 85% of each of the positive and negative datasets to define a balanced training set comprising 243 AMPs and 243 decoys for use in SVM classifier training and cross-validation. The remaining 15% of each dataset defined a balanced blind test set of 43 AMPs and 43 decoys that was used only to evaluate the performance of the final classifier and not used at any stage of classifier training, tuning, or feature selection.

There are no hard-and-fast rules for the optimal split between training and test data, since this can depend strongly on the total number of observations, the complexity of the models to be fit, and the level of noise in the samples.[23] A good rule of thumb is to reserve 40–80% of the data for training, erring towards larger fractions for larger datasets and lower signal-to-noise ratios.[23,44,45] For classification problems, one must also pay attention to the proportions of the two (or more) classes within the data. While one might be inclined to assemble training sets with a class balance matching that of the population at large,[46] great care must be taken in training over imbalanced datasets.[47,48] In particular, classification accuracy can be a misleading training objective for highly imbalanced datasets since high performance can just reflect the underlying distribution. It is advisable to balance the data by resampling, collecting additional samples, or generating synthetic samples, and to consider the use of alternative performance measures such as the sensitivity or positive predictive value.[47,48] In general, the training and testing partitions should maintain the same class balance.[45]

4.3. Descriptor generation and embedded feature selection

The goal of our machine learning classifier is to predict whether a candidate peptide is an AMP from its amino acid sequence. In principle, one could compute the similarity of the candidate sequence space to those in the training data according to some proximity metric (e.g., Hamming, Jukes-Cantor) and make an assignment based on the classification of its nearest neighbors using a k-nearest neighbor (k-NN) classification protocol.[23] In practice, this approach will tend to perform poorly for query peptides possessing little sequence similarity with those in the training data, and is particularly ill-suited for AMP classification due to the large diversity of peptide lengths and sequences.[25,49] An alternative approach instead makes classifications based on a set of features derived from the peptide sequence that are potentially relevant determinants of antimicrobial activity.[50,51] In the case of peptides, these features typically comprise physicochemical descriptors (e.g., charge, hydrophobicity) and patterns in amino acid composition (e.g., prevalence of contiguous residue pairs, correlated distributions of residues with similar properties) that may be readily calculated directly from the peptide sequence without additional experimental knowledge.[52,53] Large numbers of these derived descriptors may be generated, and *feature selection* conducted to systematically identify an optimal subset of descriptors.[54,55] By excluding irrelevant or confounding descriptors, feature selection is beneficial in improving model performance, reducing the time and cost of descriptor generation, improving model generalization, and enhancing interpretability of the model.[54,55]

In this work, we employed the propy Python package[52,53] to generate 1588 descriptors for each peptide in the training set and then conducted two rounds of feature selection. First, we filtered the features to eliminate two irrelevant descriptors that were invariant over the training data, and 257 redundant descriptors that were highly correlated with one or more other descriptors. The remaining 1329 descriptors were then Z-scored by subtracting out their mean and dividing by their standard deviation over the training data. This linear transformation is a standard pre-processing step to place all features on an even mathematical footing by rendering each descriptor dimensionless and standardized to zero mean and unit variance.[54] Second, we performed a form of embedded feature selection based on an elegant approach developed by Bi et al. that simultaneously performs on-the-fly feature selection and model training by modifying of the standard SVM classifier objective function.[56] This approach replaces the soft margin minimization defined in Eq. (3) by,

$$\arg\min_{\mathbf{w},b}\left[\frac{1}{2}\,||\,\mathbf{w}\,||_1 + C\frac{1}{n}\sum_{i=1}^{n}L_2(y_i,\mathbf{x}_i)\right],$$
$$L_2(y_i,\mathbf{x}_i) = [\max(0, 1 - y_i(\mathbf{w}.\mathbf{x}_i - b))]^2, \quad (4)$$

where $||\,\mathbf{w}\,||_1$ is the ℓ_1-norm of $\mathbf{w}$, and $L_2(y_i,\mathbf{x}_i)$ is a squared hinge loss. Similar to the LASSO method,[57] the ℓ_1-norm serves as a numerically efficient proxy for the ℓ_0-norm that enforces sparsity in the $\mathbf{w}$ vector. Geometrically, the classification hyperplane is perpendicular to those descriptors with zero-value elements in $\mathbf{w}$, meaning that these descriptors do not play a role in classification and may be discarded. The parameter C controls the tradeoff between sparsity and classification errors penalized according to the square of their distance from the margin.[56] We solve this minimization over the $n = 486$ training points using the scikit-learn Python machine learning library,[37] and determine the optimum value of $C = 0.0995$ using $k = 15$ rounds of stratified shuffled cross validation to maximize the prediction accuracy of the sparse model trained over 80% of the training data on a 20% validation partition.[56,23] The particular descriptors selected vary between the sparse models, and we stabilize feature selection by bootstrap aggregation (bagging)[58] to identify those $m = 12$ descriptors retained in all $k = 15$ rounds of cross validation. These descriptors comprise the terminal ensemble identified by our feature selection procedure and are listed in Table 1.

4.4. Classifier training, validation, and performance

The $m = 12$ descriptors identified by our feature selection procedure (Table 1) were then used to train a linear SVM classifier by minimizing the objective function in Eq. (3) over the $n = 486$ peptides in the training set to determine the optimal values of $\mathbf{w}$ and b. Training was conducted using the scikit-learn Python machine learning library,[37] and the optimal value of $\lambda = 0.0127$ determined by maximizing the accuracy of the classifier trained over 80% of the training data on a 20% validation partition over $k = 15$ independent rounds of stratified shuffled cross validation. The elements of the

Table 1
The 12 descriptors identified by feature selection protocol. We provide a brief physical interpretation of each descriptor; full details are provided in Ref. 1. The descriptors are rank ordered according to their weights in the trained SVM classifier (i.e., the value of their corresponding element in the **w** vector). Positive weights indicate a positive association with antimicrobial activity, and negative weights a negative association.

Rank	Feature	Description	Weight ($\mathbf{w}_i$)
1	netCharge	Net peptide charge	0.80
2	τ_2^G	Length-normalized sequence order coupling number measuring physicochemical correlations between residues separated by two positions (i, i+2) measured by the Grantham chemical distance matrix[49,59]	0.48
3	p_{29}^G	Pseudo amino acid composition generalization at tier k = 9 measuring pairwise correlations of the physicochemical properties of residues separated by nine positions (i, i+9) measured by the Grantham chemical distance matrix[49,59]	0.36
4	SolventAccessD1025	Fraction of the peptide length containing 25% of the buried amino acid residues A,L,F,C,G,I,V,W	−0.24
5	pc(M,K)	Relative fraction of M residues to K residues	−0.21
6	p_{50}^G	Pseudo amino acid composition generalization at tier k = 30 measuring pairwise correlations of the physicochemical properties of residues separated by 30 positions (i, i +30) measured by the Grantham chemical distance matrix[49,59]	0.20
7	AE	Fraction of contiguous AE residue pairs	0.18
8	τ_4^G	Length-normalized sequence order coupling number measuring physicochemical correlations between residues separated by four positions (i, i+4) measured by the Grantham chemical distance matrix[49,59]	−0.17
9	LW	Fraction of contiguous LW residue pairs	0.17
10	NK	Fraction of contiguous NK residue pairs	0.13
11	DP	Fraction of contiguous DP residue pairs	−0.12
12	FC	Fraction of contiguous FC residue pairs	−0.04

optimal **w** vector corresponding to each descriptor are provided in Table 1.

We evaluated the performance of the trained classifier over the blind test set of 43 AMPs and 43 decoy peptides according to the following five metrics, where *TP* is the number of true positives, *TN* the number of true negatives, *FP* the number of false positives, and *FN* the number of false negatives:[50,60]

- accuracy = $(TP + TN)/(TP + FP + TN + FN) = 91.9\%$
- specificity = $TN/(TN + FP) = 93.0\%$
- sensitivity = $TP/(TP + FN) = 90.7\%$
- positive predictive value (PPV) = $TP/(TP + FP) = 92.9\%$
- negative predictive value (NPV) = $TN/(TN + FN) = 90.9\%$

The classifier exhibits excellent performance in excess of 90% along all metrics. For the purposes of high throughput *in silico* screening of peptide sequence space, typically one wishes to prioritize specificity and positive predictive accuracy such that false positive rate is very low and positive classifications trusted with high confidence. Our classifier exhibits this desirable performance with 92.9% of positive classifications expected to be true positives. High sensitivity and negative predictive value (i.e., low false negative rate) are typically less of a priority for virtual screening. The size of the accessible sequence space is so large that failure to correctly identify all positive candidates is not critical, since we are typically able to generate far more putative hits than can be experimentally synthesized. Nevertheless, our classifier also performs well along these metrics, with 90.9% of negative predictions expected to be true negatives. We also computed a strong Matthews correlation coefficient (MCC) – also known as the phi coefficient – of 0.837, revealing a strong correlation between the predicted and observed classifications.[61] The classifier was trained for predictive accuracy in both positive and negative predictions to attain 91.9% accuracy. However, the offset of the hyperplane from the origin (i.e., the *b* value) can be increased (decreased) to increase (decrease) the threshold for a positive prediction. This allows the classifier to be tuned to meet the requirements of a particular application by enhancing the specificity and PPV (sensitivity and NPV) at the expense of overall accuracy. The computed area under the receiver operating characteristic (AUROC) of 0.981 indicates that our classifier possesses excellent sensitivity–specificity trade-off.[60]

We compared the performance of this 12-descriptor linear SVM classifier against a 1329-descriptor linear classifier employing all of the original 1588 descriptors excluding those 259 determined to be irrelevant or redundant, and against 12-descriptor and 1329-descriptor nonlinear SVM classifiers employing optimized polynomial and radial basis function kernels.[36,23,1] In all cases we found the performance of the 12-descriptor linear SVM to be as good or superior to the more complicated variants with the added benefits of clearer mechanistic interpretability and faster computation times. This illustrates the value and success of rigorous feature selection in developing robust and high performance classifiers. Computationally, calculation of the 12 descriptors and classification by the trained SVM requires only 0.14 s per peptide on a 2.13-GHz Intel Core 2 Duo processor. Classification of large peptide libraries can be performed by deploying multiple independent copies of the classifier in parallel, facilitating accurate and efficient *in silico* screening of peptide sequence space.

5. Calibration and interpretation of classifier predictions

5.1. The classifier predicts membrane activity, not antimicrobial activity

The distance σ of a candidate peptide in the 12-dimensional feature space from the separating hyperplane of the trained classifier possesses a clear geometric interpretation and serves as the discriminating metric to classify a peptide as an AMP ("hit", $\sigma \geqslant 0$) or not ("miss", $\sigma < 0$). It is also possible to convert σ into a probability of membership in the "hit" class $P(+1)$ by performing logistic regression over the training data to define a monotonic mapping between σ and $P(+1)$. [37] We initially hypothesized that distance from the separating hyperplane would be correlated with antimicrobial activity, defining an inverse relationship between σ and the *in vitro* minimum inhibitory concentration (MIC). Accordingly, we expected potent antimicrobial peptides to be located far from the margin on the "hit" side of the hyperplane, peptides with no microbicidal activity far from the margin on the "miss" side, and peptides with weak microbicidal activity close to the hyperplane. To test this hypothesis, we collated a list of 478 AMPs active against *Staphylococcus aureus* for which standardized MIC values are known (http://www.antistaphybase.com)[62] and ran them through our SVM classifier to compute the predicted σ values. The Spearman correlation coefficient between σ and MIC revealed no significant correlation, with the 95% confidence intervals spanning zero and an insignificant *p*-value ($\rho_{\text{Spearman}} = -0.06[-0.15,$

0.03]; $p = 0.19$, two-tailed bootstrap significance test with $n = 10{,}000$ trials) (Fig. 2a). This result does not support the hypothesis that the classifier has learned to distinguish AMPs based on antimicrobial potency. This negative result may be understood by the diverse mechanisms of microbicidal activity in addition to membrane activity. For example, there are peptides within the set of 478 AMPs considered that mediate their antimicrobial activity through inhibition of DNA synthesis, inhibition of macromolecular synthesis, and immunomodulation.[2] Due to these confounding factors it is perhaps not surprising in retrospect that we should see no correlation between σ and MIC. What, then, is the mechanistic basis by which the SVM classifier has learned to make its AMP/non-AMP classification predictions?.

We developed a new hypothesis that the SVM did not learn to distinguish AMPs based on *antimicrobial activity*, but rather *membrane activity* as a uniting property of AMPs that possess diverse modes of action. In prior work, we have shown AMPs that effect their microbicidal activity through membrane permeation to generate a specific type of membrane curvature known as negative Gaussian curvature (NGC).[63–67] Induction of NGC has also been observed for other classes of membrane active peptides, including cell-penetrating peptides[68] and viral fusion peptides,[69,70] and is a necessary requirement of a number of membrane destabilization mechanisms including portion, blebbing, and budding.[12,71–75,1] Geometrically, NGC possesses a clear mathematical and geometric interpretation.[76,2] At each point on a surface we can define the curvature κ along an arbitrary tangent vector as the reciprocal of the signed radius R of the corresponding kissing circle. The radius is defined to be positive if the vector connecting the point on the surface to the center of the kissing circle points in the same direction as the surface normal, and negative otherwise. The *principal curvatures* $\kappa_1 = 1/R_1$ and $\kappa_2 = R_2$ are respectively defined by the kissing circles that give maximum and minimum values of the curvature. The *mean curvature* of the surface is given by the arithmetic mean of the principal curvatures $H = (\kappa_1 + \kappa_2)/2$, and the *Gaussian curvature* is given by their product $K = \kappa_1\kappa_2$. Positive Gaussian curvature ($K > 0$) is indicative of a dome-like shape (either convex or concave), whereas negative Gaussian curvature ($K < 0$) is indicative of a saddle-like topography (Fig. 2b).

We tested our hypothesis by synthesizing 16 α-helical peptides with varying degrees of homology to known AMPs that were positively classified by the SVM ($\sigma > 0$) with a range of σ values. We characterized their capacity to induce NGC by incubating the peptides with small unilamellar vesicles as artificial mimics of bacterial cell membranes, and interrogated the membrane structure using synchrotron small angle X-ray scattering (SAXS). The peak positions in the integrated scattering intensity as a function of scattering vector $I(q)$ revealed that 14 of the 16 peptides reorganized the membranes into Pn3m or Im3m cubic phases replete with NGC (Fig. 2c). To quantify the degree of induced NGC, we inferred the best-fit cubic lattice parameter a for each cubic phase and combined it with the Euler characteristic χ and surface area per unit cell A_0 ($\chi = -4$ and $A_0 = 2.345$ for Im3m, $\chi = -2$ and $A_0 = 1.919$ for Pn3m) to compute the average NGC in the phase $|\langle K\rangle| = 2\pi\chi/A_0a^2$. [1,69,77,78] Correlating $|\langle K\rangle|$ with σ, we find a strong and statistically significant positive correlation ($\rho_{\text{Spearman}} = 0.65[0.23,\ 0.89]$; $p = 0.006$, two-tailed bootstrap hypothesis test with $n = 10{,}000$ trials) (Fig. 2d). As a negative control, we synthesized three additional peptides negatively classified by the SVM ($\sigma < 0$), and found all three unable to generate NGC (i.e., $|\langle K\rangle| = 0\ \text{Å}^{-2}$). These results provide strong support for our hypothesis, unveiling the mechanistic basis for the predictions of the SVM classifier as the capacity of the peptides to generate NGC in bacterial membranes. It somewhat remarkable that a SVM classifier trained over only 12 physicochemical descriptors generated from sequence information alone can learn a rule based on an intrinsically geometric and topological mechanism of action. Our use of calibrated experimentation informed by the machine learning model to define this relationship presents a compelling illustration of the synergies between these two modes of investigation.

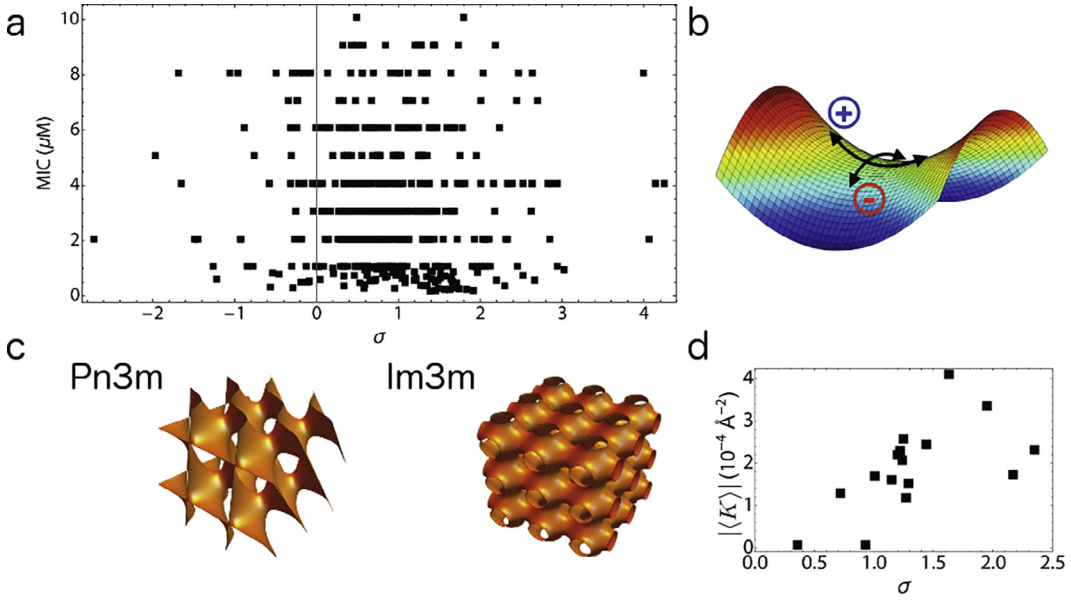

Fig. 2. Physical interpretation of the distance to hyperplane σ predicted by the trained SVM classifier. (a) A scatterplot of σ against *in vitro* minimum inhibitory concentration (MIC) for 478 AMPs active against *Staphylococcus aureus* reveals no significant correlation ($\rho_{\text{Spearman}} = -0.06$ [-0.15, 0.03], $p = 0.19$). (b) Schematic illustration of a surface possessing negative Gaussian curvature (NGC) wherein principal curvatures of opposing sign give rise to a saddle-shaped topography. (c) Illustration of the Pm3m and Im3m cubic phase space groups that are rich in NGC. (d) A scatterplot of σ against average NGC $|\langle K\rangle|$ induced in artificial mimics of bacterial cell membranes for 16 peptides selected for synthesis and experimental characterization reveals a strong and statistically significant positive correlation ($\rho_{\text{Spearman}} = 0.65$ [0.23, 0.89], $p = 0.006$). Panels a, c, and d are adapted from Lee et al. *Proc. Natl. Acad. Sci. USA* **113** 48 13588–13593 (2016).[1]

5.2. Interpretation of selected features and comparison with human learning

By defining a separating hyperplane in the m-dimensional feature space spanned by the input descriptors (cf. Fig. 1), linear SVM classifiers define a relatively transparent relationship between the input features and classification prediction. The relationship is encoded in the elements of the **w** vector defining the surface normal of the separating hyperplane. Geometrically, large magnitude elements of **w** indicate that the surface normal of the hyperplane possesses a large component oriented along the corresponding feature axis, and that this feature is an important determinant in classification. Positive values indicate a positive association between the corresponding descriptor and membrane activity, and negative values a negative association. The **w** vector elements associated with each of the $m = 12$ descriptors rank-ordered by magnitude is presented in Table 1.

The predictive performance of a machine learning classifier is, of course, independent of any *post hoc* analysis of its mathematical structure, and the multidimensional relationships it embodies cannot always be straightforwardly translated into mechanistic understanding. In particular, care must be taken in resolving univariate trends with respect to single descriptors without accounting for the full multidimensionality of the classification function.[23] Nevertheless, the relative interpretability of the SVM classifier, together with targeted experimentation informed by its predictions, reveal that the classifier has learned (at least) four peptide properties with a comprehensible mechanistic link to the determinants of membrane activity.

Charge. The top-ranked feature possessing the most discriminatory weight is the peptide net charge, possessing a **w** vector element of +0.80, which is >65% larger in magnitude than that for any of the remaining 11 features. The positive sign of the weight indicates that positive charge is positively associated with positive classification. This is consistent with the well-known cationic nature of AMPs that mediates an attractive Coulombic attraction with positively-charged lipid head groups in bacterial cell membranes.[6] Without providing any physical model of AMP action, the classifier identified positive charge as a principal determinant of membrane activity.

Amphipathicity. Another hallmark of membrane active peptides is a facially amphipathic nature that can mediate membrane disruption through a number of proposed mechanisms including the barrel stave, toroidal pore, and carpet models.[4,79,12–15] Our SVM classifier also learns facial amphipathicity as a defining signature of membrane activity, with fully one third (4 of 12) of the descriptors identified in our feature selection procedure associated with this property. Specifically, the 2nd, 3rd, and 6th-ranked descriptors possessing positive weights of 0.48, 0.36, and 0.20 favor positive classification of a peptide as an AMP if residues separated by 2, 9, and 30 positions tend to have opposing physicochemical character. The 8th ranked descriptor possessing a negative weight of −0.17 favors positive classification if residues separated by 4 positions tend to have similar physicochemical character. Trained over peptides with α-helical secondary structure, the classifier has identified and exploited physicochemical periodicity within the 3.6-residue period of the helix as a discriminating classification rule (Fig. 3). Specifically, facially amphipathic α-helical candidates are scored highly by our classifier along these four features due to the physicochemical similarity of the residues residing on each of the hydrophilic and hydrophobic faces.

Dipeptide incidence. The feature selection procedure identified five contiguous residue pairs as important discriminants of membrane activity. The contiguous pairs AE, LW, and NK possess positive **w** weights (0.18, 0.17, 0.13), indicating that an elevated prevalence of these pairs favors positive classification, and DP and FC negative weights (−0.12, −0.04) such that a reduced prevalence favors positive classification. To investigate this trend, we conducted univariate logistic regression over the length-normalized prevalence of each of these five residue pairs over our curated library of 286 AMP and 286 decoy peptides (Section 4.2). We constructed a least-squares fit of the function $\log\left(\frac{P_{\text{hit}}(\eta)}{P_{\text{miss}}(\eta)}\right) = \beta_0 + \beta_1\eta$, where η is the per residue incidence rate of a particular contiguous dipeptide pair, $P_{\text{hit}}(\eta)$ is the probability of a peptide being an AMP given a particular value of η, $P_{\text{miss}}(\eta) = 1 - P_{\text{hit}}(\eta)$, and $\frac{P_{\text{hit}}(\eta)}{P_{\text{miss}}(\eta)}$ is the odds. Performing the logistic regressions reveals AE, LW, and NK to have positive values of the regression coefficient on the length-normalized incidence ($\beta_1 = 23.7,\ 22.2,\ 35.7$) that reach statistical significance ($p = 4.5 \times 10^{-9}$, 1.2×10^{-9}, 8.9×10^{-7}, Wald test). Conversely, DP and FC possess negative coefficients ($\beta_1 = -15.2,\ -2.45$) that do not reach significance ($p = 0.48$, 0.89, Wald test). Accordingly, the classifier has indeed identified that the pairs of amino acids AE, LW, and NK tend to appear contiguously at a statistically significant higher prevalence in AMPs. Classic analyses of individual amino acid incidences in AMPs vs. non-AMPs show that cationic amino acids like lysine (K) and hydrophobic amino acids like leucine (L) and tryptophan (W) appear more commonly in AMPs. However, to the best of our knowledge, this discovery of important dipeptide motifs in AMPs is a novel finding for which the physicochemical basis remains unclear. This result illustrates the value of interpretable machine learning models in spurring and informing new inquiry, and suggests an avenue for future studies to resolve the physicochemical root of these elevated dipeptide prevalences. This interplay of statistical learning and experimentation can assist in mapping the "linguistic" tendencies of AMPs.

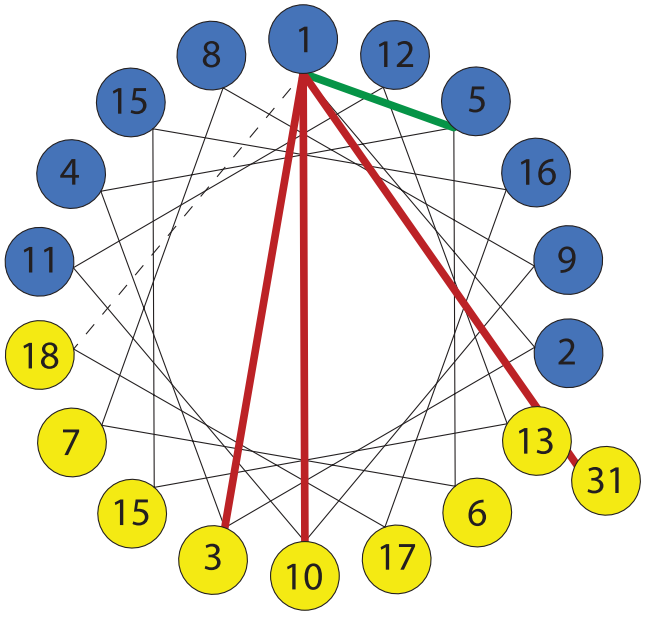

Fig. 3. Helical wheel plot showing the relative residue locations along the α-helical backbone. Our classifier favors positive classification of peptides in which residues separated by 2, 9, and 30 positions are of opposing physicochemical character (i.e., those located on opposing faces of the helix; red bold lines) and those separated by 4 positions tend to have similar character (i.e., those located on the same face; green bold line). Facially amphipathic peptides are therefore scored highly by the classifier over these four features, and it has learned these patterns as a discriminating rule with which to distinguish membrane activity. Image adapted from Lee et al. *Proc. Natl. Acad. Sci. USA* **113** 48 13588–13593 (2016).[1]

Saddle splay selection rule. The combination of cationic charge, hydrophobicity, and amphipathicity is central to the recognizability and mechanism of action of AMPs.[80,81] In prior work, we

established the so-called "saddle splay selection rule" that codified a trade-off between mean peptide hydrophobicity and the relative proportions of the positively charged arginine and lysine residues.[82,65,83,84] Hydrophobic residues can induce positive membrane curvature by steric displacement of phospholipids in the membrane.[85] Cationic peptides can generically induce negative mean membrane curvature due to electrostatic attraction and wrapping of the membrane around the peptide.[65,68] By doubly coordinating phosphate head groups, arginine can by itself induce NGC, whereas singly-coordinating lysine can only generate negative mean curvature.[65] Accordingly, arginine-rich AMPs (and cell-penetrating peptides) can rely on their arginine content alone to simultaneously induce both positive and negative curvature and produce the conditions required for NGC. In contrast, AMPs whose cationic content is largely provided by lysine residues typically also contain a number of highly hydrophobic residues to cooperatively generate NGC. We have previously shown the peptides within the AMP database to obey the saddle splay selection criterion,[82,9] and wished to determine if this rule was also learned by our SVM classifier.

To test this hypothesis, we generated for consideration by our classifier an ensemble of 242,110 peptide candidates of length 20–25 residues by a directed traversal of sequence space. We collated the 76 AMPs in the training and test data within the size range of interest, and added to these the 33,079 sequences formed by making all single point mutations. We then supplemented these with 208,955 sequences generated by a Markov Chain Monte Carlo (MCMC) procedure that biased sampling towards high-σ candidates. We initialized 10 MCMC runs with a randomly selected AMP from the database and performed 25,000 rounds of random point mutation, insertion, and deletion. Proposed transitions were accepted or rejected according to a Metropolis acceptance criterion,

$$P_{\text{accept}} = \min\{1, \exp((\sigma_{\text{trial}} - \sigma_{\text{current}})/T)\}, \quad (5)$$

where σ_{current} is the distance from the hyperplane assigned by the SVM classifier to the current peptide sequence in the MCMC chain, σ_{trial} is that for the trial peptide, and $T = 0.8$ is a fictitious "temperature" that controls the acceptance ratio.[86–88] Within this ensemble of 242,110 candidates, we identified the set of optimal candidates against which to test the saddle splay selection rule. To do so, we defined a multi-objective optimization to simultaneously maximize (i) distance from the hyperplane σ, (ii) degree of α-helical structure assessed using the PSIPRED secondary structure prediction algorithm[89,90] implemented in the PROTEUS2 program,[91] and (iii) minimum sequence homology to any known AMP measured by the Jukes–Cantor distance. In doing so we select those candidates that are not only predicted by our classifier to possess a high probability of possessing membrane activity, but also possess the correct secondary structure, and share sequence similarity with existing AMPs. We solve this optimization problem to identify those *Pareto optimal* sequences for which no one criterion can be improved without diminishing another.[92,93] To guard against unwarranted extrapolations of our classifier into regions of feature space not observed during classifier training we also restrict our analysis to those candidates among the 242,110 peptides for which the 12 descriptors used by the classifier lie no more than 10% outside the range spanned by the training peptides.[27] We present in Fig. 4a a scatterplot of the 242,110 candidates considered by our search procedure within which we have highlighted the 13 Pareto optimal candidates. Fig. 4b illustrates that these 13 peptides obey precisely the hydrophobicity versus proportion of arginine and lysine tradeoff defined by the saddle splay selection rule as the AMPs present in the AMP database.[82,9] It is somewhat remarkable that this relatively complex determinant of membrane activity was discovered by the classifier precisely in line with existing understanding.

6. Discovery of multiplexed membrane activity in peptide families with other primary functions

To employ a recently-coined neologism to express the frequently enormous size of genomics data sets, peptide sequence space is "genomically" large in that it comprises 20^N possible

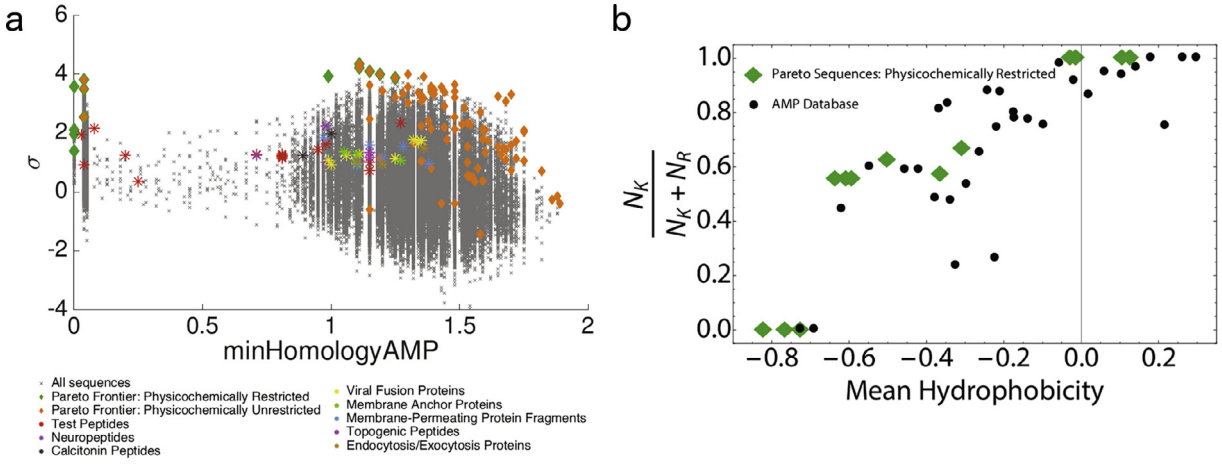

Fig. 4. Directed search of sequence space and adherence of Pareto optimal candidates to the saddle splay selection rule. (a) Projection of the 242,110 peptide candidates considered in our directed traversal of sequence space into the minimum sequence homology to any known AMP measured by the Jukes-Cantor distance and classifier distance to hyperplane σ. Highlighted are the 85 Pareto optimal candidates within the three dimensional search space of [σ, degree of α-helical structure, sequence homology to a known AMP] (orange diamonds), and the 13 Pareto optimal candidates subject to the additional condition that the 12 descriptors employed by the classifier (Table 1) lie no more than 10% outside the range of the training data (green diamonds). Peptides from families with other putative primary functions are situated close to the Pareto frontier and are positively classified by the SVM ($\sigma > 0$) suggesting that they possess membrane activity as part of a multiplexed functionality (colored stars). (b) Optimal peptide candidates identified in a guided traversal of sequence space by the SVM classifier obey the previously identified saddle splay selection rule governing a trade-off between peptide hydrophobicity and proportion of arginine and lysine residues. N_K and N_R respectively denote the number of lysine and arginine residues in the peptide. The mean hydrophobicity is defined as the mean value of the Eisenberg consensus hydrophobicity averaged over all residues in the peptide.[94] The 13 physicochemical restricted Pareto optimal peptides identified within our directed search of 242,110 peptide candidates (green diamonds) fall precisely on the saddle splay selection rule trend defined by the α-helical AMPs extracted from the AMP database (black circles).[9] We plot all 299 α-helical peptides harvested from the database clustered into 31 bins according to mean hydrophobicity in order to smooth the distribution and improve visual clarity. Panels a and b are adapted from Lee et al. *Proc. Natl. Acad. Sci. USA* **113** 48 13588–13593 (2016).[1]

sequences for an *N*-residue protein.[95] Computationally efficient QSAR models can perform high-throughput *in silico* screening of sequence space to sieve through orders of magnitude larger numbers of candidates than would be possible by experiment.[26–28,32,33] Nevertheless, for peptides longer then a few amino acids the sequence space remains too large to screen exhaustively. The Monte-Carlo search procedure described in Section 5.2 presents a means to perform a guided traversal of space that biases sampling towards promising candidates while simultaneously preventing trapping of the search procedure in local optima by adjusting the fictitious "temperature" T in the Metropolis acceptance criterion (Eq. (5)). Tuning $T \to 0$ results in the acceptance only of those moves that improve σ but risks trapping in local maxima and poor sampling of sequence space. Conversely, tuning $T \to \infty$ results in acceptance of all trial moves to produce a random walk through sequence space that disregards the location of the SVM hyperplane. In practice, we find tuning $T \approx 0.8$ provides a good balance between sampling and bias towards candidate sequences ranked highly by the classifier. The computationally screened candidates may then be hierarchically ranked along any number of metrics by computing a series of nested Pareto frontiers,[92,93] and a subset of the top-ranked candidates – along with low-ranked controls – put forward for experimental testing. We observe that more sophisticated temperature control approaches to direct sampling may be implemented by appealing to the vast literature in simulated annealing,[96] simulated tempering,[97] parallel tempering,[98] expanded ensembles,[99] and J-walking.[100]

Here we remark on an alternative application of QSAR models not to discover novel highly functional candidates, but rather detect multiplexed functionality in existing peptide families. We employed our SVM classifier as an experimentally-validated predictor of membrane activity to analyze a diversity of proteins from the Protein Data Bank (www.rcsb.org)[101] to search for membrane activity in a variety of peptide families with established primary function.[1] Despite sharing very little homology with any known AMP, our classifier situates a number of peptides close to the Pareto frontier of the 242, 110 candidates generated in our *in silico* screen (Fig. 4a). Specifically, we predict membrane activity in proteins belonging to diverse functional families, including endocytosis/exocytosis peptides (brown stars), membrane anchor proteins (green stars), membrane-permeating protein fragments (blue stars), and topogenic peptides (pink stars). More interestingly, we identify membrane activity within neuropeptides, suggesting potential intracellular regulation targets (purple stars), viral fusion proteins, indicative of a role for membrane deformations within the viral life cycle (yellow stars), and calcitonin, as a hormone involved in calcium regulation but also part of the amyloid family of which other members have been reported to aggregate on and permeabilize lipid membranes.[102,103] This analysis illustrates the utility of trained and validated QSAR models in helping to understand modes of peptide action and in guiding new experimental inquiry.

7. Conclusions and outlook

In this mini-review, we have examined a recent piece of work in which a machine learning classifier was developed to aid in the design of membrane active peptides, discover membrane activity in diverse peptide families, and provide interpretable understanding of the underlying mechanisms by which membrane activity is effected.[1] Our introduction to support vector machines exposes the elegant simplicity of their mathematical foundations that can provide interpretable QSAR models that can inform mechanistic understanding. Frequently one must trade off model simplicity with predictive accuracy, but we demonstrated that judicious and rigorous variable selection procedures coupled with sufficiently large and high quality training data can produce simple, interpretable, low-dimensional models with performance equal to or exceeding their more complex cousins. The trained model illuminated a number of machine-learned discriminatory rules precisely in line with human understanding, but also other features – and multidimensional couplings between them – that were not previously known. Furthermore, this work illustrates the imperative importance of experiment to furnish high-quality training and test data, validate the classifier, and calibrate its predications to ascertain what rules have actually been learned. In this case, these experiments revealed – to our initial consternation – that the classifier had not learned a rule to discriminate antimicrobial peptides based on antimicrobial activity, but rather membrane activity as a unifying prerequisite of AMPs possessing diverse modes of action. We subsequently exploited our classifier to perform a high throughput directed computational search of sequence space, and identify membrane activity as a multiplexed function within diverse peptide families.

Looking to the future, we see an increasing role for modes of investigation comprising tightly coupled and mutually reinforcing experimentation and machine learning. Human-directed experimental trial-and-error searches of the "genomically" large peptide sequence space can frequently lead to highly inefficient deployment of resources. Integrating machine learning with targeted experimentation to guide experimentation and provide new data to improve model performance establishes a mutually beneficial cycle providing savings in money, time, and labor to massively accelerate peptide discovery and design. Furthermore, we see great potential in developing and combining multiple classifiers to engineer multiplexed peptide activity. The machine learning models described in this work are extremely generic and extensible, permitting them to be straightforwardly translated to the classification of peptides and proteins – not just membrane proteins – with arbitrary functions, including intra-cellular transport, immunomodulation, signaling, vesicle fusion, or toxicity. For example, combining our membrane activity classifier with another designed to screen for immunomodulatory peptides could be used to design peptides with both functions encoded either modularly within distinct domains or combined within multifunctional sequences. Finally, we emphasize the value of "white box" or "grey box" machine learning models (e.g., linear support vector machines, decision trees) wherein the mathematical underpinnings are sufficiently interpretable to not only deliver high predictive performance, but also expose and inform mechanistic understanding.

Acknowledgements

E.Y.L. acknowledges support from the T32 Systems and Integrative Biology Training Grant at University of California, Los Angeles (UCLA) (T32GM008185) and the T32 Medical Scientist Training Program at UCLA (T32GM008042). G.C.L.W. acknowledges support from NIH Grant 1R21AI122212. X-ray research was conducted at Stanford Synchrotron Radiation Lightsource, SLAC National Laboratory, supported by the US DOE Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

References

1. Lee EY, Fulan BM, Wong GC, Ferguson AL. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences*. 2016;113(48):13588–13593.
2. Lee EY, Lee MW, Fulan BM, Ferguson AL, Wong GCL. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface Focus*. 2017 (in press).
3. Zasloff M. Antimicrobial peptides of multicellular organisms. *Nature*. 2002;415 (6870):389–395. http://dx.doi.org/10.1038/415389a. URL: http://www.nature.com/doifinder/10.1038/415389a.

4. Shai Y. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by α-helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochim Biophys Acta Biomembr*. 1999;1462 (1–2):55–70. http://dx.doi.org/10.1016/S0005-2736(99)00200-X. URL: http://linkinghub.elsevier.com/retrieve/pii/S000527369900200X.
5. Brogden KA. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature Rev Microbiol*. 2005;3(3):238–250. http://dx.doi.org/10.1038/nrmicro1098. URL: http://www.nature.com/doifinder/10.1038/nrmicro1098.
6. Hancock REW, Lehrer R. Cationic peptides: a new source of antibiotics. *Trends Biotechnol*. 1998;16(2):82–88. http://dx.doi.org/10.1016/S0167-7799(97)01156-6. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167779997011566.
7. Hancock REW, Sahl H-G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnol*. 2006;24(12):1551–1557. http://dx.doi.org/10.1038/nbt1267. URL: http://www.nature.com/doifinder/10.1038/nbt1267.
8. Yeaman MR, Yount NY. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol Rev*. 2003;55(1):27–55. http://dx.doi.org/10.1124/pr.55.1.2. URL: http://pharmrev.aspetjournals.org/content/55/1/27.full.
9. Wang Z, Wang G. Apd: The antimicrobial peptide database. *Nucleic Acids Res*. 2004;32(suppl 1):D590–D592.
10. Wang G, Li X, Wang Z. Apd2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res*. 2009;37(suppl 1): D933–D937.
11. 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. http://dx.doi.org/10.1093/nar/gkv1278. URL http://nar.oxfordjournals.org/content/44/D1/D1087.full.
12. Yang L, Harroun TA, Weiss TM, Ding L, Huang HW. Barrel-Stave model or toroidal model? A case study on Melittin Pores. *Biophys J*. 2001;81 (3):1475–1485. http://dx.doi.org/10.1016/S0006-3495(01)75802-X. URL: http://linkinghub.elsevier.com/retrieve/pii/S000634950175802X.
13. Bechinger B, Kim Y, Chirlian LE, et al. Orientations of amphipathic helical peptides in membrane bilayers determined by solid-state NMR spectroscopy. *J Biomol NMR*. 1991;1(2):167–173. http://dx.doi.org/10.1007/BF01877228. URL: http://link.springer.com/article/10.1007/BF01877228.
14. Pouny Y, Rapaport D, Mor A, Nicolas P, Shai Y. Interaction of antimicrobial dermaseptin and its fluorescently labeled analogs with phospholipid membranes. *Biochemistry*. 1992;31(49):12416–12423. http://dx.doi.org/10.1021/bi00164a017. URL: http://pubs.acs.org/doi/abs/10.1021/bi00164a017.
15. K. Matsuzaki, O. Murase, N. Fujii, K. Miyajima, An Antimicrobial Peptide, Magainin 2, Induced Rapid Flip-Flop of Phospholipids Coupled with Pore Formation and Peptide Translocation, vol. 35, American Chemical Society, 1996. http://dx.doi.org/10.1021/bi960016v. URL: http://pubs.acs.org/doi/abs/10.1021/bi960016v.
16. Brötz H, Bierbaum G, Leopold K, Reynolds PE, Sahl HG. The lantibiotic mersacidin inhibits peptidoglycan synthesis by targeting lipid II. *Antimicrob Agents Chemother*. 1998;42(1):154–160. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=9449277&retmode=ref&cmd=prlinks.
17. Park CB, Kim HS, Kim SC. Mechanism of action of the antimicrobial peptide buforin II: Buforin II kills microorganisms by penetrating the cell membrane and inhibiting cellular functions. *Biochem. Biophys. Res. Commun.*. 1998;244 (1):253–257. http://dx.doi.org/10.1006/bbrc.1998.8159. URL http://linkinghub.elsevier.com/retrieve/pii/S0006291X98981591.
18. Yonezawa A, Kuwahara J, Fujii N, Sugiura Y. Binding of tachyplesin I to DNA revealed by footprinting analysis: significant contribution of secondary structure to DNA binding and implication for biological action. *Biochemistry*. 2002;31(11):2998–3004. http://dx.doi.org/10.1021/bi00126a022.
19. Patrzykat A, Friedrich CL, Zhang L, Mendoza V, Hancock REW. Sublethal concentrations of pleurocidin-derived antimicrobial peptides inhibit macromolecular synthesis in Escherichia coli. *Antimicrob Agents Chemother*. 2002;46(3):605–614. http://dx.doi.org/10.1128/AAC.46.3.605-614.2002. URL: http://aac.asm.org/content/46/3/605.full.
20. Otvos L, O I, Rogers ME, et al. Interaction between heat shock proteins and antimicrobial peptides. *Biochemistry*. 2000;39(46):14150–14159. http://dx.doi.org/10.1021/bi0012843.
21. Bowdish DME, Davidson DJ, Hancock REW. A re-evaluation of the role of host defence peptides in mammalian immunity. *Curr.Protein Peptide Sci.*. 2005;6 (1):35–51. http://dx.doi.org/10.2174/1389203053027494. URL: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2037&volume=6&issue=1&spage=35.
22. Gilliet M, Lande R. Antimicrobial peptides and self-dna in autoimmune skin inflammation. *Curr Opin Immunol*. 2008;20(4):401–407.
23. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer Science & Business Media; 2009.
24. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S. CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res*. 2009;38 (10). http://dx.doi.org/10.1093/nar/gkp1021. gkp1021–D780.
25. Lata S, Sharma BK, Raghava G. Analysis and prediction of antibacterial peptides. *BMC Bioinf*. 2007;8(1):1. http://dx.doi.org/10.1186/1471-2105-8-263.
26. Fjell CD, Jenssen H, Fries P, et al. Identification of novel host defense peptides and the absence of α-defensins in the bovine genome, Proteins: Structure. *Funct Bioinformat*. 2008;73(2):420–430. http://dx.doi.org/10.1002/prot.22059.
27. Fjell CD, Jenssen H, Hilpert K, et al. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J Med Chem*. 2009;52 (7):2006–2015. http://dx.doi.org/10.1021/jm8015365.
28. Cherkasov A, Hilpert K, Jenssen H, et al. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chem Biol*. 2008;4(1):65–74. http://dx.doi.org/10.1021/cb800240j.
29. Wang P, Hu L, Liu G, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS ONE*. 2011;6(4): e18476. http://dx.doi.org/10.1371/journal.pone.0018476.
30. Torrent M, Andreu D, Nogués VM, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS ONE*. 2011;6 (2):e16968. http://dx.doi.org/10.1371/journal.pone.0016968. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21347392&retmode=ref&cmd=prlinks.
31. 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):168–177. http://dx.doi.org/10.1016/j.ab.2013.01.019. URL: http://linkinghub.elsevier.com/retrieve/pii/S0003269713000390.
32. Maccari G, Di Luca M, Nifosí R, et al. Antimicrobial peptides design by evolutionary multiobjective optimization. *PLOS Comput Biol*. 2013;9(9): e1003212. http://dx.doi.org/10.1371/journal.pcbi.1003212.
33. Giguère S, Laviolette F, Marchand M, et al. Machine learning assisted design of highly active peptides for drug discovery. *PLOS Comput Biol*. 2015;11(4): e1004074. http://dx.doi.org/10.1371/journal.pcbi.1004074. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25849257&retmode=ref&cmd=prlinks.
34. Schneider P, Müller AT, Gabernet G, et al. Hybrid network model for deep learning of chemical data: application to antimicrobial peptides. *Mol Inf*. 2017;36(1–2):1600011. http://dx.doi.org/10.1002/minf.201600011.
35. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20 (3):273–297. http://dx.doi.org/10.1007/BF00994018.
36. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers*Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM Press; 1992:144–152.
37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830. URL: http://dl.acm.org/citation.cfm?id=1953048.2078195.
38. Aizerman A, Braverman EM, Rozoner LI. Theoretical foundations of the potential function method in pattern recognition learning. *Automat Remote Control*. 1964;25:821–837. http://dx.doi.org/10.1234/12345678. URL: http://www.citeulike.org/group/664/article/431797.
39. Duan K-B, Keerthi SS. Which is the best multiclass svm method? an empirical study*International Workshop on Multiple Classifier Systems*. Springer; 2005:278–285.
40. Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. *IEEE Trans.Neural Netw.*. 2002;13(2):415–425.
41. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.*. 2012;40(Database issue):D370–6. http://dx.doi.org/10.1093/nar/gkr703.
42. Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Research*. 2012;41(D1). http://dx.doi.org/10.1093/nar/gks1169. gks1169–D529.
43. Tusnády GE, Dosztányi Z, Simon I. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*. 2004;20 (17):2964–2972. http://dx.doi.org/10.1093/bioinformatics/bth340.
44. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genom*. 2011;4(1):31.
45. Lever J, Krzywinski M, Altman N. Points of significance: Model selection and overfitting. *Nature Methods*. 2016;13(9):703–704.
46. Oommen T, Baise LG, Vogel RM. Sampling bias and class imbalance in maximum-likelihood logistic regression. *Math Geosci*. 2011;43(1):99–120.
47. Chawla NV. Data mining for imbalanced datasets: An overview. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*. US, Boston, MA: Springer; 2010:875–886. http://dx.doi.org/10.1007/978-0-387-09823-445.
48. He H, Ma Y. *Imbalanced Learning: Foundations, algorithms, and applications*. John Wiley & Sons; 2013.
49. Chou K-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteom*. 2009;6 (4):262–274. http://dx.doi.org/10.2174/157016409789973707. URL: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1570-1646&volume=6&issue=4&spage=262.
50. Porto WF, Pires ÁS, Franco OL. CS-AMPPred: an updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLoS ONE*. 2012;7(12):e51444. http://dx.doi.org/10.1371/journal.pone.0051444.
51. Porto WF, Fernandes FC, Franco OL. An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs*Advances in Bioinformatics and Computational Biology*. Berlin Heidelberg: Springer; 2010:59–62. http://dx.doi.org/10.1007/978-3-642-15060-96.
52. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*. 2006;34(Web Server issue): W32–7. http://dx.doi.org/10.1093/nar/gkl305.

53. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*. 2013;29(7):960–962. http://dx.doi.org/10.1093/bioinformatics/btt072.
54. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182. URL: http://dl.acm.org/citation.cfm?id=944919.944968.
55. Kittler J. Feature selection and extraction. *Handbook of Pattern Recognition and Image Processing*. URL: http://personal.ee.surrey.ac.uk/Personal/J.Kittler/lecturenotes/biometrics/EEM.asp2.grey.pdf.
56. Bi J, Bennett K, Embrechts M, Breneman C, Song M. Dimensionality reduction via sparse support vector machines. *J Mach Learn Res*. 2003;3:1229–1243. URL: http://dl.acm.org/citation.cfm?id=944919.944971.
57. Tibshirani R. Regression selection and shrinkage via the lasso. *J Royal Statist Soc Ser B*. 1996;58(1):267–288.
58. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–140. http://dx.doi.org/10.1007/BF00058655.
59. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185(4154):862–864. http://dx.doi.org/10.1126/science.185.4154.862.
60. Gorunescu F. *Data Mining: Concepts, Models and Techniques*. Springer Science & Business Media; 2011.
61. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Protein Struct*. 1975;405(2):442–451. http://dx.doi.org/10.1016/0005-2795(75)90109-9. URL: http://linkinghub.elsevier.com/retrieve/pii/0005279575901099.
62. Zouhir A, Taieb M, Lamine MA, et al. Antistaphybase: database of antimicrobial peptides (amps) and essential oils (eos) against methicillin-resistant staphylococcus aureus (mrsa) and staphylococcus aureus. *Arch Microbiol*. 2016;1–8.
63. Yang L, Gordon VD, Trinkle DR, et al. Mechanism of a prototypical synthetic membrane-active antimicrobial: efficient hole-punching via interaction with negative intrinsic curvature lipids. *Proc Natl Acad Sci*. 2008;105(52):20595–20600.
64. Schmidt NW, Wong GC. Antimicrobial peptides and induced membrane curvature: geometry, coordination chemistry, and molecular engineering. *Curr Opin Solid State Mater Sci*. 2013;17(4):151–163.
65. Schmidt NW, Lis M, Zhao K, Lai GH, Alexandrova A, Tew GN, Wong GC. Molecular basis for nanoscopic membrane curvature generation from quantum mechanical models and synthetic transporter sequences. *J Am Chem Soc*. 2012;134(46):19207.
66. Lee MW, Chakraborty S, Schmidt NW, Murgai R, Gellman SH, Wong GC. Two interdependent mechanisms of antimicrobial activity allow for efficient killing in nylon-3-based polymeric mimics of innate immunity peptides. *BiochimBiophys Acta Biomembr*. 2014;1838(9):2269–2279.
67. Xiong M, Lee MW, Mansbach RA, et al. Helical antimicrobial polypeptides with radial amphiphilicity. *Proc Natl Acad Sci*. 2015;112(43):13155–13160. http://dx.doi.org/10.1073/pnas.1507893112. arXiv:http://www.pnas.org/content/112/43/13155.full.pdf. URL: http://www.pnas.org/content/112/43/13155.abstract.
68. Schmidt N, Mishra A, Lai GH, Wong GC. Arginine-rich cell-penetrating peptides. *FEBS Lett*. 2010;584(9):1806–1813.
69. Schmidt NW, Mishra A, Wang J, DeGrado WF, Wong GC. Influenza virus a m2 protein generates negative gaussian membrane curvature necessary for budding and scission. *J Am Chem Soc*. 2013;135(37):13710.
70. Yao H, Lee MW, Waring AJ, Wong GC, Hong M. Viral fusion protein transmembrane domain adopts β-strand structure to facilitate membrane topological changes for virus–cell fusion. *Proc Natl Acad Sci*. 2015;112(35):10926–10931.
71. Yang L, Weiss TM, Lehrer RI, Huang HW. Crystallization of antimicrobial pores in membranes: magainin and protegrin. *Biophys J*. 2000;79(4):2002–2009. http://dx.doi.org/10.1016/S0006-3495(00)76448-4. URL: http://linkinghub.elsevier.com/retrieve/pii/S0006349500764484.
72. Ludtke SJ, He K, Heller WT, Harroun TA, Yang L, Huang HW. Membrane Pores Induced by Magainin. *Membrane Pores Induced by Magainin*, Vol. 35. American Chemical Society; 1996. http://dx.doi.org/10.1021/bi9620621.
73. Saiman L, Tabibi S, Starner TD, et al. Cathelicidin peptides inhibit multiply antibiotic-resistant pathogens from patients with cystic fibrosis. *Antimicrob Agents Chemother*. 2001;45(10):2838–2844. http://dx.doi.org/10.1128/AAC.45.10.2838-2844.2001.
74. Kalfa VC, Jia HP, Kunkle RA, McCray PB, Tack BF, Brogden KA. Congeners of SMAP29 kill ovine pathogens and induce ultrastructural damage in bacterial cells. *Antimicrob Agents Chemother*. 2001;45(11):3256–3261. http://dx.doi.org/10.1128/AAC.45.11.3256-3261.2001.
75. Yu Y, Vroman JA, Bae SC, Granick S. Vesicle budding induced by a pore-forming peptide. *J Am Chem Soc*. 2010;132(1):195–201. http://dx.doi.org/10.1021/ja9059014.
76. Kreyszig E. *Differential geometry*. New York: Dover Publications; 1991. URL: http://lccn.loc.gov/91014321.
77. Harper P, Gruner S. Electron density modeling and reconstruction of infinite periodic minimal surfaces (ipms) based phases in lipid-water systems. I. Modeling ipms-based phases. *Eur Phys J E*. 2000;2(3):217–228. http://dx.doi.org/10.1007/PL00013660.
78. Anderson D, Wennerstrom H, Olsson U. Isotropic bicontinuous solutions in surfactant-solvent systems: the l3 phase. *J Phys Chem*. 1989;93(10):4243–4253.
79. Oren Z, Shai Y. Mode of action of linear amphipathic α-helical antimicrobial peptides. *Peptide Sci*. 1998;47(6):451–463. 10.1002/(SICI)1097-0282(1998)47:6451::AID-BIP43.0.CO;2-F. URL: http://doi.wiley.com/10.1002/%28SICI%291097-0282%281998%2947%3A6%3C451%3A%3AAID-BIP4%3E3.0.CO%3B2-F.
80. Epand RM, Shai Y, Segrest JP, Anantharamiah GM. Mechanisms for the modulation of membrane bilayer properties by amphipathic helical peptides. *Biopolymers*. 1995;37(5):319–338. http://dx.doi.org/10.1002/bip.360370504.
81. Segrest JP, De Loof H, Dohlman JG, et al. Amphipathic helix motif: classes and properties. *Proteins Struct Funct Bioinf*. 1990;8(2):103–117. http://dx.doi.org/10.1002/prot.340080202.
82. Schmidt NW, Mishra A, Lai GH, et al. Criterion for amino acid composition of defensins and antimicrobial peptides based on geometry of membrane destabilization. *J Am Chem Soc*. 2011;133(17):6720.
83. Wu Z, Cui Q, Yethiraj A. Why do arginine and lysine organize lipids differently? Insights from coarse-grained and atomistic simulations. *J Phys Chem B*. 2013;117(40):12145–12156. http://dx.doi.org/10.1021/jp4068729.
84. Cui Q, Zhang L, Wu Z, Yethiraj A. Generation and sensing of membrane curvature: where materials science and biophysics meet. *Curr Opin Solid State Mater Sci*. 2013;17(4):164–174. http://dx.doi.org/10.1016/j.cossms.2013.06.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S1359028613000387.
85. McMahon HT, Gallop JL. Membrane curvature and mechanisms of dynamic cell membrane remodelling. *Nature*. 2005;438(7068):590–596.
86. Gilks WR, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. CRC Press; 1995. URL: http://books.google.com/books?id=TRXrMWYi2IC&printsec=frontcover&dq=Markov+Chain+Monte+Carlo+in+Practice&hl=&cd=1&source=gbsapi.
87. Gilks WR. *Markov Chain Monte Carlo*. Chichester, UK: John Wiley & Sons Ltd; 2005. http://dx.doi.org/10.1002/0470011815.b2a14021.
88. Geyer CJ. Practical markov chain Monte Carlo. *Statist Sci*. 1992;7(4):473–483. http://dx.doi.org/10.2307/2246094.
89. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics*. 2000;16(4):404–405. http://dx.doi.org/10.1093/bioinformatics/16.4.404.
90. Zhang H, Zhang T, Chen K, et al. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings Bioinf*.. 2011;12(6):672–688. http://dx.doi.org/10.1093/bib/bbq088.
91. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinf*. 2006;7(1):1. http://dx.doi.org/10.1186/1471-2105-7-301.
92. Arora JS. *Introduction to Optimum Design*. Academic Press; 2011. URL: http://books.google.com/books?id=Qhpdh9xDsa4C&printsec=frontcover&dq=Introduction+to+Optimum+Design+Third+Edition&hl=&cd=1&source=gbsapi.
93. Shoval O, Sheftel H, Shinar G, et al. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science*. 2012;336(6085):1157–1160. http://dx.doi.org/10.1126/science.1217405.
94. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. Hydrophobic moments and protein structure. *Faraday Symp Chem Soc*. 1982;17:109–120. http://dx.doi.org/10.1039/fs9821700109.
95. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? *PLoS Biol*. 2015;13(7):e1002195.
96. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *science*. 1983;220(4598):671–680.
97. Marinari E, Parisi G. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*. 1992;19(6):451.
98. Swendsen RH, Wang J-S. Replica monte carlo simulation of spin-glasses. *Phys Rev Lett*. 1986;57(21):2607.
99. Lyubartsev A, Martsinovski A, Shevkunov S, Vorontsov-Velyaminov P. New approach to monte carlo calculation of the free energy: method of expanded ensembles. *J Chem Phys*. 1992;96(3):1776–1783.
100. Frantz D, Freeman DL, Doll J. Reducing quasi-ergodic behavior in monte carlo simulations by j-walking: applications to atomic clusters. *J Chem Phys*. 1990;93(4):2769–2784.
101. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235–242.
102. Friedman R, Pellarin R, Caflisch A. Amyloid aggregation on lipid bilayers and its impact on membrane permeability. *J Mol Biol*. 2009;387(2):407–415.
103. Caillon L, Killian JA, Lequin O, Khemtémourian L. Biophysical investigation of the membrane-disrupting mechanism of the antimicrobial and amyloid-like peptide dermaseptin s9. *PloS One*. 2013;8(10):e75528.