Title

Mapping membrane activity in undiscovered peptide sequence space using machine learning

Permalink
https://escholarship.org/uc/item/1553n90m

Journal
Proceedings of the National Academy of Sciences of the United States of America, 113(48)

ISSN
0027-8424

Authors
Lee, Ernest Y
Fulan, Benjamin M
Wong, Gerard CL
et al.

Publication Date
2016-11-29

DOI
10.1073/pnas.1609893113

Peer reviewed

Mapping membrane activity in undiscovered peptide sequence space using machine learning

Ernest Y. Lee[a], Benjamin M. Fulan[b], Gerard C. L. Wong[a,1], and Andrew L. Ferguson[c,d,1]

[a]Department of Bioengineering, University of California, Los Angeles, CA 90095; [b]Department of Mathematics, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [c]Department of Materials Science and Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and [d]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved October 11, 2016 (received for review July 4, 2016)

There are some ∼1,100 known antimicrobial peptides (AMPs), which permeabilize microbial membranes but have diverse sequences. Here, we develop a support vector machine (SVM)-based classifier to investigate α-helical AMPs and the interrelated nature of their functional commonality and sequence homology. SVM is used to search the undiscovered peptide sequence space and identify Pareto-optimal candidates that simultaneously maximize the distance σ from the SVM hyperplane (thus maximize its "antimicrobialness") and its α-helicity, but minimize mutational distance to known AMPs. By calibrating SVM machine learning results with killing assays and small-angle X-ray scattering (SAXS), we find that the SVM metric σ correlates not with a peptide's minimum inhibitory concentration (MIC), but rather its ability to generate negative Gaussian membrane curvature. This surprising result provides a topological basis for membrane activity common to AMPs. Moreover, we highlight an important distinction between the maximal recognizability of a sequence to a trained AMP classifier (its ability to generate membrane curvature) and its maximal antimicrobial efficacy. As mutational distances are increased from known AMPs, we find AMP-like sequences that are increasingly difficult for nature to discover via simple mutation. Using the sequence map as a discovery tool, we find a unexpectedly diverse taxonomy of sequences that are just as membrane-active as known AMPs, but with a broad range of primary functions distinct from AMP functions, including endogenous neuropeptides, viral fusion proteins, topogenic peptides, and amyloids. The SVM classifier is useful as a general detector of membrane activity in peptide sequences.

machine learning | membrane curvature | membrane permeation | antimicrobial peptides | cell-penetrating peptides

The ∼1,100 known antimicrobial peptides (AMPs) (1–6) are known collectively to have broad spectrum antimicrobial activity (1, 3, 5) via nonspecific interactions to target generic features in the many pathogen membranes (1, 7). Machine learning can in principle be used to help discover the "blueprint" for natural AMP sequences; however, such an enterprise presents significant structural difficulties. AMPs do not share a common core structure, but tend to be short (<50 amino acids), cationic (+2 to +9), and amphiphilic (1–6). One of the principal components of AMP activity involves the selective permeabilization of microbial membranes (1–3, 5, 8–11). However, there is increasing evidence that membrane activity is but one of several modes of antimicrobial activity: Translocated AMPs can interact with intracellular targets to inhibit cell wall synthesis, nucleic acid synthesis, protein synthesis, and enzymatic activity (12–16). Recent studies have shown that AMPs can be immunomodulatory (17, 18): In fact, LL-37 plays a role in autoimmune disorders such as lupus and psoriasis (18). These confounding factors make it difficult to implement adaptive learning for AMPs.

Prior AMP machine-learning studies have focused primarily on empirical quantitative structure activity relationship (QSAR) models to evaluate large pools of candidate sequences and identify AMP candidates with improved minimum inhibitory concentrations (MICs) (19–21). QSAR models for AMP discovery use a variety of statistical learning approaches, including multiple linear regression, linear discriminant analysis, principal component analysis, partial least-squares regression, artificial neural networks (ANN), support vector classifiers (SVC)/support vector machines (SVM), quantitative matrices (QM), hidden Markov models (HMM), and random forests (RFs) (21–24). Lata et al. developed ANN, SVC, and QM models based on an analysis of the C- and N-terminal residues in 486 antibacterial peptides (25). Porto et al. reported the development of the CS-AMPPred predictor of cysteine-stabilized AMPs based on a SVC trained over five physicochemical descriptors (19). Fjell et al. developed a 44-descriptor ANN model to screen ∼100,000 candidates and produce 18 peptides that showed high activity against drug-resistant bacteria (20). Torrent et al. trained an eight-descriptor AMP and SVC to predict AMP sequences (26). Maccari et al. used RFs to design and validate the antimicrobial activity of two natural peptides, and one peptide incorporating nonnatural amino acids (27). Fjell et al. used an HMM to screen the bovine genome for AMPs, one of which was discovered in bovine intestinal tissue (28). Cherkasov et al. iteratively synthesized thousands of nine-residue peptides and trained ANN models to discover peptides with activities against drug-resistant superbugs (29). The present work is similar in scope to these works—most notably those of Fjell et al. (20) and Cherkasov et al. (29)—in that we train QSAR models using limited experimental data to perform high-throughput virtual screening and identification of promising peptides for experimental synthesis and testing. Specifically, we train the SVM to recognize α-helical AMPs (Fig. 1*A*) (30–34), so our work is cognate with computer-assisted AMP discovery and design (22, 35, 36), which have resulted in clinical trials of synthetic AMPs (22). The goals of this work are quite different from the above. Whereas a QSAR model with good predictive performance is expected to be able to identify physical

Significance

We use machine learning on membrane-permeating α-helical host defense peptides to study the nature of their functional commonality and sequence homology. Machine learning is combined with calibrating experiments to show that the metric in our support vector machine model correlates not with antimicrobial activity but with a peptide's ability to generate the negative Gaussian membrane curvature necessary for membrane permeation. Moreover, we use the classifier reflexively to map the undiscovered sequence space of antimicrobial peptides and identify taxonomies of peptides with similar topological membrane remodeling activity, including endogenous neuropeptides, viral fusion proteins, topogenic peptides, and amyloids.

Author contributions: E.Y.L., G.C.L.W., and A.L.F. designed research; E.Y.L., B.M.F., and A.L.F. performed research; E.Y.L., B.M.F., and A.L.F. contributed new reagents/analytic tools; E.Y.L., G.C.L.W., and A.L.F. analyzed data; and E.Y.L., G.C.L.W., and A.L.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: gclwong@seas.ucla.edu or alf@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1609893113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1609893113

determinants of AMP activity (22, 35, 36), our primary aim is not to use QSAR classifiers to find AMPs with improved activity, but rather to use computational modeling in conjunction with calibrating experiments to examine the interrelated nature of AMP functional commonality, AMP sequence homologies, and general physicochemical mediators of AMP function at multiple length scales. We believe this approach will help uncover new unifying relationships between the discouragingly diverse peptide taxonomies that currently exist.

Here, we focus our study on α-helical AMPs, which have a structure common to many peptides and protein motifs. Specifically, we use our SVM to guide traversal and mapping of the undiscovered peptide sequence space. To help navigate this enormous space, we use known AMP sequences as "landmarks" and construct a "Pareto frontier" using the concept of Pareto optimality from microeconomics (37, 38) as the subset of sequences that simultaneously maximize the probability that the sequence is antimicrobial (the distance from the SVM hyperplane σ) and the degree of α-helical secondary structure, and minimize the mutational distance to known AMPs (Fig. 1*B*). Using a combination of killing assays and small-angle X-ray scattering (SAXS) experiments on synthesized test peptide sequences that are not previously known AMPs, we find a strong correlation between the SVM distance to hyperplane σ and the ability for peptides to generate negative Gaussian curvature (NGC) in model membranes. Because NGC is the type of membrane curvature topologically required for common membrane permeation mechanisms such as pore formation, blebbing, and budding (Fig. 1*C*), it provides a structural basis for this common component of AMP activity. Using the SVM sequence map as a discovery tool for membrane activity, we increase mutational distances from known AMP sequences, probing the sequence space that is increasingly difficult for nature to discover via simple mutation from existing AMP sequences. What emerges is a diverse taxonomy of sequences that are expected to be not only just as membrane-active as known AMPs, but also have a broad range of putative primary functions beyond antimicrobial activity. We highlight several families, including endogenous neuropeptides, viral fusion proteins, topogenic peptides, and amyloids. Had their primary functions been undiscovered, these peptides could have been classified as AMPs. Not only is membrane activity not coextensive with antimicrobial activity, it is surprisingly common for many classes of natural peptides as one component of multiplexed intracellular functions. Moreover, the calibrated SVM we construct is an efficient discovery tool to identify and discover membrane-active sequences.

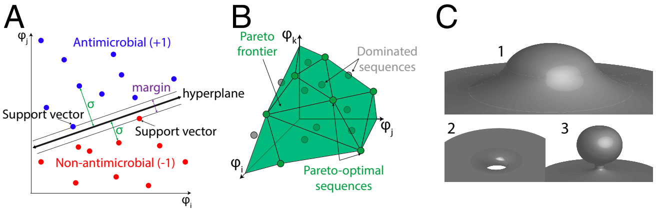

Fig. 1. SVM learning and Pareto-optimization select for antimicrobial and membrane curvature-generating peptides. (*A*) Schematic depicting the use of an SVM binary classifier to partition hypothetical antimicrobial peptide sequences (blue circles) described by the $n = 2$ descriptors $\{\phi_i,\phi_j\}$ from non-antimicrobial sequences (red circles) using an $(n - 1)$-dimensional maximum-margin linear hyperplane. The support vectors are the sequences lying on the margins. The separating hyperplane lies midway between the margins. The metric σ (green arrows) indicates the distance to hyperplane for each peptide. Positive distances denote antimicrobial sequences whereas negative distances denote nonantimicrobial sequences. (*B*) Schematic demonstrating separation of Pareto-optimal sequences (green circles) from dominated sequences (gray circles) in an arbitrary 3D subspace of descriptors. The Pareto frontier is the hypersurface containing the Pareto-optimal sequences. (*C*) Common biologically relevant manifestations of negative Gaussian curvature generation in cell membranes, including (*C, 1*) blebbing, (*C, 2*) pore formation, and (*C, 3*) scission and budding.

Results and Discussion

SVM Model Development, Monte Carlo Sampling, and Pareto Analysis. We constructed a linear SVM classifier to predict whether a candidate peptide is likely to be antimicrobial. The SVM was trained by cross-validation over a balanced training set comprising 243 α-helical AMPs and 243 α-helical decoy peptides derived from natural and synthetic sources and a variety of microbial and multicellular species, and a balanced blind test set of 43 AMP and 43 decoy α-helical peptides (*SI Appendix*, Figs. S1 and S2). Variable selection was used to identify from a library of 1,588 physicochemical descriptors (*SI Appendix*, Table S1) a subset of 12 descriptors used to make classifications (*SI Appendix*, Table S2). The classifier demonstrated excellent performance against the blind test set, with a prediction accuracy of 91.9%, specificity of 93.0%, and sensitivity of 90.7% (*SI Appendix*, Table S3).

The trained SVM enables rapid computational screening of peptides for antimicrobial activity. Generation of the 12 descriptors and classification of all 572 AMPs and decoy peptides in our training and test sets required only 80 s on a 2.13-GHz Intel Core 2 Duo processor, equal to 0.14 s per peptide. Nevertheless, comprehensive screening of all 10^{78} peptides of length 8–60 residues—the size range of training peptides—is intractable, so we instead traversed sequence space via a directed search according to the following four criteria: (*i*) Defined peptide length. The general mechanism of α-helical AMPs involves a combination of electrostatic and hydrophobic interactions with negatively charged bacterial cell membranes. Typical AMPs have a length of 20–25 amino acids, and generally generate membrane deformation by spanning the lipid bilayer and inducing opposite membrane curvature in orthogonal directions (2, 6, 39). Accordingly, we restrict our screen to 20–25-residue peptides. (*ii*) Homology with a known AMP. Classifier accuracy is expected to diminish in regions of sequence space far from the AMP sequences upon which it was trained. Accordingly, we control within our screen peptide homology to known AMPs. (*iii*) Large positive distance from the SVM hyperplane. This criterion favors sequences for which the classifier possesses high positive predictive value (the expected proportion of positive results that are true positives) and specificity at the expense of sensitivity (left side of the receiver operating characteristic curve, right side of the precision-recall curve in *SI Appendix*, Fig. S2). In driving the false-positive rate toward zero—at the expense of a high false-negative rate—we focus our screen toward the most promising candidates. (*iv*) High helical content. Trained on α-helical peptides, classifier accuracy is expected to diminish for peptides with nonhelical structure. Accordingly, we seek candidates that are predominantly helical. [We note that certain α-helical AMPs only adopt a helical structure upon interacting with the cell membrane (2, 6), and this criterion will necessarily disfavor such candidates.]

Using these criteria, we performed a guided search of sequence space to generate a subsampling of 242,110 candidates using two complementary approaches. To traverse the sequence space with close homology to existing AMPs, we generated 33,079 sequences corresponding to all one-point mutants of the 76 AMPs in the database within the size range 20–25 residues. To probe the sequence space of AMPs that are less likely for nature to discover via simple mutation, we generated 208,955 additional sequences from Monte Carlo sampling by initializing 10 independent Monte Carlo chains with a randomly selected AMP in the size range of interest, and ran 25,000 rounds of random point mutation, insertion, and deletion. The 12 bagged descriptors were generated for each new trial sequence and the distance from the classifier hyperplane σ computed using the SVM classifier. Treating $(-\sigma)$ as an effective energy to bias the search towards promising candidates, trial sequences are accepted or rejected according to the Metropolis criterion $p_{acc} = \min\{1, \exp(\Delta\sigma/T)\}$, where $\Delta\sigma = \sigma_{trial} - \sigma_{current}$ and T is an effective temperature (40–42). In this work, we found that $T = 0.8$ provides a good compromise between focusing the search toward large σ candidates while also providing good sampling. All unique candidates were saved for the computational screen. Finally, the helical content of all 242,110 candidates was evaluated using the ab initio secondary structure prediction algorithm PSIPRED (bioinf.cs.ucl.ac.uk/psipred/) (43, 44) implemented in PROTEUS2 (www.proteus2.ca/proteus2/) (45). We do not actively direct sampling toward high helical candidates, but identify them post hoc. We note that more comprehensive sampling of design space could be achieved using umbrella sampling in our Monte Carlo procedure (46).

After conducting this directed search, we wished to identify optimal sequences within our candidates. To do this, we borrowed the concept of Pareto optimality from microeconomics (37, 38) and used multiobjective optimization to determine the Pareto optimal sequences that dominate all other candidates in simultaneously maximizing the distance from the SVM hyperplane and the degree of α-helical secondary structure, and minimizing the mutational distance to a known AMP. These sequences are optimal in the sense that no other candidates exist for which any one criterion can be improved without degrading at least one other. We term this Pareto frontier "physicochemically unrestricted" because we place no restriction on the value of the 12 bagged descriptors. It is possible that this Pareto set may contain candidates with descriptor values far outside the range of the training data. Accordingly, we constructed a "physicochemically restricted" Pareto frontier constrained such that none of the 12 descriptors could lie more than 10% outside the range observed in the training set (20). Together, these two frontiers serve as guides for our exploration of the sequence space, and our subsequent interpretation of discovered sequences.

We present a 3D scatterplot of all 242,110 candidate sequences (gray crosses) plotting the distance to the SVM hyperplane σ, minimum Jukes–Cantor distance to a known AMP *minHomologyAMP*, and predicted fractional helicity (Fig. 2). Two-dimensional projections of the point cloud are presented in *SI Appendix*, Fig. S3. The candidate sequences partition into two point clouds. The cloud at low *minHomologyAMP* corresponds to candidates most homologous to known AMPs comprising the 76 AMPs in the database within the size range of interest plus all of their point mutants. The cloud at high *minHomologyAMP* corresponds to sequences generated by the directed Monte Carlo search biased toward candidates with a large distance from the SVM hyperplane σ. Importantly, the latter may contain sequences with divergent physicochemical properties from known AMPs that are nonetheless predicted by our classifier to possess antimicrobial activity. We highlight the 13 sequences residing on the Pareto frontier of physicochemically restricted sequences (green diamonds, *SI Appendix*, Table S4), and the 85 sequences residing on the Pareto frontier of physicochemically unrestricted sequences (orange diamonds, *SI Appendix*, Table S5). We also indicate 16 peptides (red stars, *SI Appendix*, Table S6) close to both Pareto frontiers that we selected to synthesize and test with SAXS and killing assays.

Distance to Hyperplane of Known AMPs Does Not Correlate with Antimicrobial Efficacy. To engage the question of whether our SVM model can predict the efficacy of known AMPs, we analyzed a standardized database of MIC values of 478 known AMPs against *Staphylococcus aureus* collated from the literature (https://www.antistaphybase.com/index.php), and calculated their distance to the SVM hyperplane σ (*SI Appendix*, Fig. S4). We plotted their reported MIC values against σ and found poor and statistically insignificant correlation ($R_{Spearman} = -0.060$ [−0.154, 0.034], $P = 0.187$). Analysis of the literature suggests that the majority of the known AMPs that we tested are compounds with multiplexed activities in addition to membrane penetration. This observation highlights a generic problem with machine-learning approaches: High classification accuracy does not necessarily imply understanding or even amenability to traditional forms of understanding. Although it is known that AMPs can have immunologically relevant activity outside of membrane activity, there is currently no general way to identify AMPs with additional functions. We propose a way to identify candidate sequences with multiplexed functions in the next section.

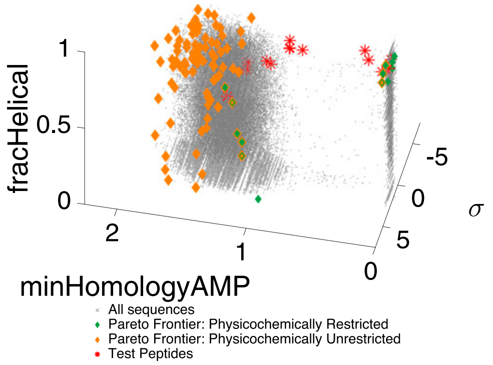

Fig. 2. Sequence atlas and Pareto frontier constructed by directed sampling of sequence space. Embedding of the 242,110 peptides generated by our directed sequence space search into the 3D space spanned by (*i*) predicted helical content, (*ii*) Jukes–Cantor distance to known AMPs, and (*iii*) distance to hyperplane (σ). Sequences with $\sigma > 0$ are predicted by the classifier to be antimicrobial or membrane-active, whereas those with $\sigma < 0$ are not. The more positive σ becomes, the higher probability of being antimicrobial P(+1). The orange diamonds pick out the 85 peptides lying on the physicochemically unrestricted Pareto frontier in which we place no restriction on the value of the descriptors generated for these candidates. Green diamonds highlight the 13 peptides on the physicochemically restricted Pareto frontier in which the descriptors are restricted to lie no more than 10% outside the range observed in the training data. Red stars are the 16 peptides proximate to the frontiers that were selected for testing.

Predicted Distance to Hyperplane of Synthesized Test Peptides Correlates with Strength of Negative Gaussian Curvature Generation. The distance from hyperplane σ possesses a clear mathematical interpretation as the distance of a candidate peptide from a hyperplane in the 12-dimensional space of the bagged descriptors (*SI Appendix*, Table S2). Nevertheless, the high dimensionality of the space, opaque nature of some descriptors, and absence of a mechanistic model linking peptide sequence to function makes it challenging to assign physical interpretability to this discriminatory metric. To glean physical understanding, we synthesized peptide sequences with defined values of σ and assayed the peptide–membrane interactions. Specifically, we selected for synthesis 16 candidate sequences identified by the directed search procedure according to the following criteria: they did not feature in the AMP database, were classified by the SVM as antimicrobial, were predicted to have >50% α-helical content by the secondary structure prediction algorithm PSIPRED (43–45), and were confirmed to possess antimicrobial activity in vitro (*SI Appendix*, Table S6). Because a common contribution to antimicrobial activity is membrane permeation, we investigated this aspect using SAXS. In our prior work, we have shown that the ability of a peptide to generate NGC in model membranes is an excellent proxy for antimicrobial activity through membrane-permeating properties (7, 11, 39). We incubated these test sequences with small unilamellar vesicles mimicking bacterial membranes {compositions 1,2-dioleoyl-sn-glycero-3-phospho-L-serine (PS)/1,2-dioleoyl-sn-glycero-3-phosphoethanolamine (PE) = 20/80 and 1,2-dioleoyl-sn-glycero-3-[phospho-rac-(1-glycerol)]

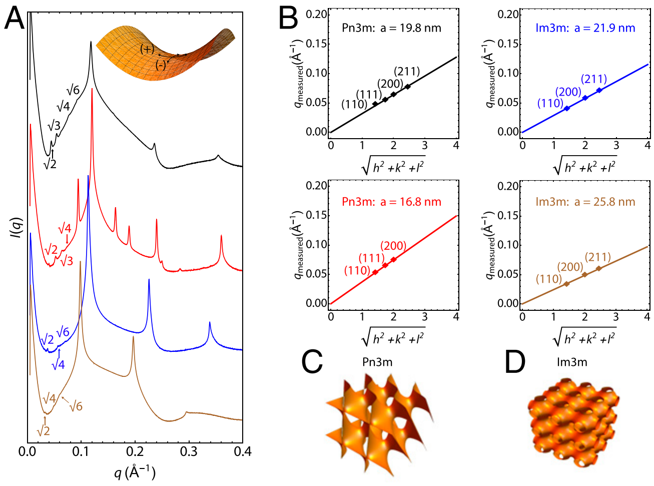

Fig. 3. Synthesized test peptides from directed Monte Carlo search and SVM screening generate negative Gaussian curvature in model membranes. (*A*) Representative SAXS data of four test peptides indicate ability to generate negative Gaussian curvature in model bacterial membranes. Peaks with cubic symmetry are labeled according to their x coordinates $\sqrt{(h^2 + k^2 + l^2)}$ in *B*. Unlabeled peaks correspond to coexisting lamellar and/or hexagonal phases induced by peptides. (*Inset*) Local topology of saddle-splay curvature. (*B*) Linear fits indicating the *q* positions of the Bragg peaks with cubic symmetry, their respective Miller indices (*hkl*), their respective space groups, and resulting lattice parameter *a*. (*C*) Contour surface representation of the Pn3m space group. (*D*) Contour surface representation of the Im3m space group.

(PG)/PE = 20/80} under physiological salt and pH (100 mM NaCl, 10 mM Hepes, pH 7.4) at specific peptide-to-lipid charge ratios, and measured the resulting peptide-induced membrane curvature quantitatively using SAXS. We find that 14 of the synthesized peptides reorganize membranes into cubic phases rich in NGC [Fig. 3 *A* (*Inset*), *C*, and *D*]. We show representative SAXS patterns for four of the test peptides, with labeled Bragg peaks corresponding to their cubic phases (Fig. 3*A*). The best-fit lines to determine the lattice parameter *a* of the cubic phases are shown, along with the Miller indices (*hkl*) of each cubic reflection (Fig. 3*B*). We find these peptides typically reorganize model membranes into either Pn3m or Im3m cubic phases (Fig. 3 *C* and *D*). The amount of induced membrane curvature in these phases can be directly compared via calculation of the NGC from the lattice parameter and the Euler characteristic (*Materials and Methods*). The same procedure was carried out for the other 10 sequences that generated cubic phases in these membranes. They also generated either Pn3m or Im3m phases (Fig. 3 *C* and *D*). The NGC values calculated from SAXS data (*SI Appendix*, Table S6) indicate that these peptides generate similar quantities of NGC as natural AMPs, which already strongly suggest that these predicted peptides permeabilize membranes. To confirm that our classifier can also recognize decoy peptides that do not generate NGC, we selected three negatively classified peptides to synthesize and test with SAXS. We find that none of them generates NGC-rich cubic phases in the same membranes tested with the 16 test peptides (*SI Appendix*, Fig. S8 and Table S8). To test the ability of our classifier to detect peptide sequences lacking antimicrobial activity, we also screened five nonactive granulysin fragment peptides from the literature (*SI Appendix*, Table S9). We find that these peptides have low or negative values of σ and are far from the Pareto frontier, which is consistent with their inability to generate NGC. Our algorithm also correctly classifies the two active granulysin fragments (47), which are known to be membrane permeating.

To test the hypothesis that the SVM classifier has learned to discriminate peptides based on their capacity to generate NGC, we computed the Spearman rank correlations of the magnitude of NGC calculated from the SAXS data for the 16 selected peptides with metrics used in the SVM classification algorithm (*SI Appendix*, Table S6). We asked whether the magnitude of NGC correlates with the homology to known AMPs and/or the distance to hyperplane σ (*SI Appendix*, Fig. S5). We find that there is no significant correlation ($R_{Spearman}$ = 0.155 [−0.425, 0.736], *P* = 0.565) between the magnitude of NGC generated by a peptide and its homology to a known AMP (Fig. 4*A*). This finding implies that sequence homology to a known AMP is not a necessary requirement to generate NGC. In other words, sequences that are far from known AMPs via simple mutation have potential to generate membrane curvature. Given this result, we ask whether there is a correlation between σ and NGC (|<k>|). In *SI Appendix*, Fig. S4, we showed that σ correlates poorly with antimicrobial efficacy. Here we observe a strong, statistically significant positive correlation between the distance to hyperplane of a peptide, and its ability to generate NGC ($R_{Spearman}$ = 0.653 [0.234, 0.891], *P* = 0.006) (Fig. 4*B*). This provides strong support for the hypothesis that our SVM classifier has learned to discriminate peptides based on their capacity to generate NGC. Because the distance to hyperplane σ also provides a measure of the confidence of the algorithm P(+1) in whether or not a sequence is antimicrobial, we conjecture that a higher confidence in classification may correlate with an enhanced ability to generate NGC. Looking at the 16 sequences that were identified by the SVM as high-probability hits [P(+1) > 0.95], we find a strong positive and statistically significant correlation between the confidence in the prediction of the SVM P(+1), and the magnitude of NGC ($R_{Spearman}$ = 0.653 [0.231, 0.896], *P* = 0.006) (Fig. 4*C*). This result makes sense because P(+1) and σ are monotonically but nonlinearly related. This result has strong implications for the utility of our SVM classifier—we can now stratify the predicted quality of sequences by using the σ-metric as a surrogate for expected NGC generating ability to efficiently explore sequence space and predict peptides with high membrane activity. Importantly, sequences less homologous to known AMPs with large values of σ that fall near, but not necessarily on, the Pareto frontier can generate the same level of membrane curvature expected of prototypical α-helical AMPs (such as Magainin, |<k>| = 2.536 × 10^{-4}·Å^{-2}), and candidate AMPs spanning a large range of helicities and homologies to known AMPs can also generate magnitudes of NGC similar to those of known AMPs (*SI Appendix*, Fig. S6 and Table S6). Accordingly, we expect to be able to use σ to predict the membrane activity of peptide families that may be very dissimilar from AMPs. We also investigated the relationship between amphiphilicity and ability to generate curvature in membranes. Interestingly, we find that 4 out of the 12 descriptors in our final SVM model enforce amphiphilicity in positively classified peptides (#2, 3, 6, 8 in *SI Appendix* Table S2, and Fig. S9*A*), suggesting that the SVM encodes amphiphilicity implicitly in its selection criteria. To quantitatively compare amphiphilicity with |<k>|, we calculate the mean hydrophobic

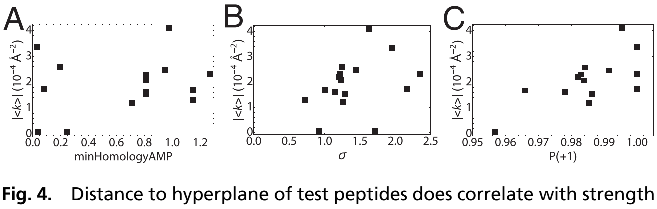

Fig. 4. Distance to hyperplane of test peptides does correlate with strength of negative Gaussian curvature. There is no significant correlation between the magnitude of NGC generation and homology of test peptides (*n* = 16) to known membrane-active peptides (*A*, $R_{Spearman}$ = 0.155 [−0.425, 0.736], *P* = 0.155), but there is a statistically significant (*B*, $R_{Spearman}$ = 0.653 [0.234, 0.891], *P* = 0.006) positive correlation between the magnitude of NGC generation and distance to hyperplane σ, as well as the probability of being antimicrobial (*C*, $R_{Spearman}$ = 0.653 [0.231, 0.896], *P* = 0.006). This validates the use of σ as a proxy for optimization of curvature generation as opposed to antimicrobial efficacy (*SI Appendix*, Fig. S4).

moment of the test peptides and find that they compare favorably to known helical AMPs (*SI Appendix*, Fig. S9*B*) despite having large mutational distances from known AMPs. We find that amphiphilicity is highly correlated with the ability to generate NGC ($R_{\mathrm{Spearman}} = 0.680$, $P = 0.0038$).

These findings strongly support the hypothesis that our SVM classifier has learned to distinguish membrane-permeating from non–membrane-permeating α-helical peptides. This result illustrates a simultaneous potentiality and deficiency of our SVM, and of QSAR approaches in general. Our classifier has discovered membrane permeation activity as a highly recognizable feature of AMPs within the training set, and used it to identify such AMPs with high accuracy and efficiency. A limitation of the classifier is that it is therefore capable of identifying and indexing membrane activity, but not necessarily antimicrobial activity. Although this result may at first blush appear to be a shortcoming of our QSAR approach, it emphasizes the transformative potential of the SVM classifier: It yields combinations of physicochemical properties that describe any peptide that can mediate membrane activity through induction of NGC. By using α-helical AMPs as a bootstrap dataset to learn about membrane curvature generating sequences in general, regardless of their primary known function, we now have a general tool for screening peptides for membrane crossing or permeating activity.

Directed Search of the Sequence Space of Physicochemically Restricted and Unrestricted Peptides Discovers Diverse Families of Membrane Curvature-Generating Peptides. To test the capacity of our SVM classifier to identify peptide sequences with membrane activity, we compiled from the Protein Data Bank 31 peptides belonging to diverse families of known and unknown function including viral peptides that attack membranes, intrinsically disordered peptides, and exocytosis/endocytosis related sequences, and used our SVM classifier to project them into our sequence map (Fig. 5 and *SI Appendix*, Fig. S3). Interestingly, a number of sequences were found to reside near the Pareto frontiers, suggesting the existence of membrane activity within these candidates. Specifically, we found several neuropeptides (purple stars), calcitonin peptide hormones (black stars), viral fusion proteins (yellow stars), membrane anchor proteins (green stars), membrane-permeating protein fragments (blue stars), and topogenic peptides (pink stars) proximate to the Pareto frontiers (Fig. 5 and *SI Appendix*, Table S7). This is exciting for several reasons. Several neuropeptides have been shown to be antimicrobial in vitro (48–50), but this is an indication that many members of the family can generate the kind of membrane curvature required for permeabilization. This alludes to possible intracellular targets of these neuropeptides and receptor-independent mechanisms of signal transduction in addition to their regular mode of activity. This observation reinforces a previously known structural tendency of AMPs, because several of them are known to have endocrine and homeostatic functions. For example, hepcidin permeates membranes but is also involved in regulation of iron (51), whereas α-melanocyte stimulating hormone has antimicrobial and antiinflammatory effects in addition to its signaling properties (52). Calcitonin is a peptide hormone involved in calcium homeostasis, but is also a known amyloid that deposits in medullary thyroid carcinoma. Other studies of amyloid proteins have demonstrated the ability to increase membrane permeability (53), and may provide a physicochemical basis for this observation (54). We have also previously described the role of the M2 proton channel in budding and scission of the influenza virus (55). Using this algorithm, we find that a variety of other viral fusion proteins likely share similar characteristics, including peptide domains from medically relevant viruses like ebolavirus, HIV, coronavirus, and hepatitis C. Other diverse discovered proteins include membrane-permeating protein fragments from enzymes, DNA-binding proteins, and prion precursors. Our analysis also identified helical membrane-active sequences from topogenic peptides, which are known membrane curvature-generating proteins. These special signal sequences present at the N-terminal portions of newly translated proteins help target and translocate large proteins across intracellular membranes (56). In general, it can be shown that the sequence content of the Pareto-optimal and peptide sequences from the newly identified taxonomies follow the same sequence trends as existing AMPs (*SI Appendix*, discussion section, and Fig. S7). This demonstrates that our algorithm can efficiently and effectively identify candidates that can reorganize bacterial membranes from a large sequence space.

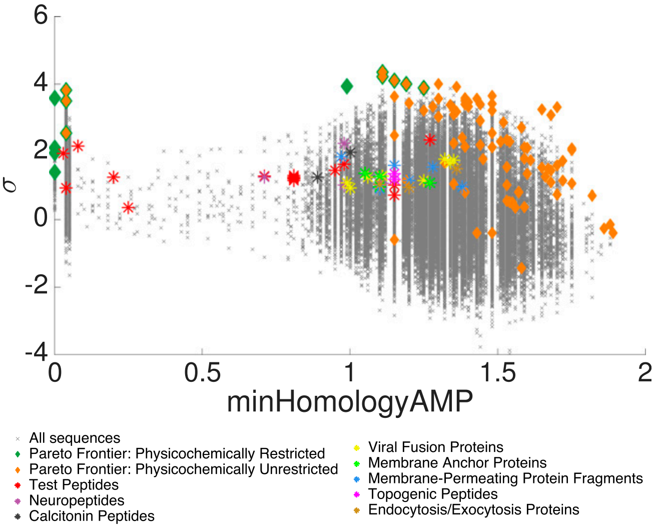

Fig. 5. Directed search of the sequence space discovers diverse families of membrane curvature-generating peptides. We visualize the 2D projection of the 242,110 candidate peptides generated by directed sampling of sequence space (Fig. 2) into distance to-hyperplane σ and Jukes–Cantor distance to known AMPs and supplemented by the 31 sequences belonging to diverse peptide families listed in *SI Appendix*, Table S7. To guide the interpretation of the discovered membrane-active sequences, we highlight the physicochemically restricted (13 peptides, green diamonds) and unrestricted Pareto frontiers (85 peptides, orange diamonds) For reference, the peptides experimentally tested are also shown (16 peptides, red stars). Screening of a variety of protein families yields sequences with predicted $\sigma > 0$ near the physicochemically unrestricted Pareto frontier. These sequences span a variety of protein families, including neuropeptides (purple stars), calcitonin peptides (black stars), viral fusion proteins (yellow stars), membrane anchor proteins (light green stars), membrane-permeating protein fragments (blue stars), and topogenic peptides (pink stars). Some of the proteins have unexpected predicted membrane activity, whereas others have confirmed experimental evidence for membrane permeation. In fact, these other classes of peptides are expected to be just as membrane-active as AMPs. This diversity demonstrates the power of the SVM-directed search framework as a tool for discovery of new membrane reorganizing protein sequences.

In summary, we have trained an SVM classifier to recognize membrane activity and experimentally calibrated the recognition metric by peptide synthesis and characterization. The results, which highlight the difference between the efficacy of an antimicrobial and its recognizability as such, are surprising. An SVM classifier trained only on physicochemical information can effectively recapitulate geometric and topological principles required for membrane permeation. We use machine learning not only to predict unknown membrane-active peptides from known ones, but also reflexively to identify peptides with multiple functions and to discover previously unknown interrelations between existing peptide classifications. Using the SVM classifier as a discovery tool to map the sequence space of AMPs, we find a diverse taxonomy of sequences that are expected to be just as membrane-active as known AMPs, but with a broad range of primary functions outside of immunity. Finally, we show how our

SVM classifier can be generalized to other fields and used as a search engine for membrane activity in peptide sequences and a detector of AMPs with multiplexed functions beyond membrane activity.

Materials and Methods

We trained and validated an SVM classifier to distinguish membrane-active sequences from non–membrane-active sequences using the Python packages propy (57) and scikit-learn (58). We used variable selection to train a classifier based on 12 physicochemical descriptors, and used this model to perform a directed search of peptide sequence space. Optimal candidates were identified using a Pareto analysis, and 16 test peptides were validated for membrane activity using SAXS and antimicrobial assays. Synthesized peptides were incubated with model membranes and the magnitude of negative Gaussian curvature was measured. Full materials and methods are found in the *SI Appendix*.

ACKNOWLEDGMENTS. We thank Nathan Schmidt and Xintong Lin for their helpful contributions to this project. 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). B.M.F. acknowledges support from National Science Foundation (NSF) Grant DMS 1345032 "MCTP: PI4: Program for Interdisciplinary and Industrial Internships at Illinois." 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.

1. Zasloff M (2002) Antimicrobial peptides of multicellular organisms. *Nature* 415(6870): 389–395.
2. Shai Y (1999) 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* 1462(1-2):55–70.
3. Brogden KA (2005) Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat Rev Microbiol* 3(3):238–250.
4. Hancock REW, Lehrer R (1998) Cationic peptides: A new source of antibiotics. *Trends Biotechnol* 16(2):82–88.
5. Hancock REW, Sahl H-G (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat Biotechnol* 24(12):1551–1557.
6. Yeaman MR, Yount NY (2003) Mechanisms of antimicrobial peptide action and resistance. *Pharmacol Rev* 55(1):27–55.
7. Yang L, et al. (2008) Mechanism of a prototypical synthetic membrane-active antimicrobial: Efficient hole-punching via interaction with negative intrinsic curvature lipids. *Proc Natl Acad Sci USA* 105(52):20595–20600.
8. Matsuzaki K (1999) Why and how are peptide–lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes. *Biochim Biophys Acta* 1462(1-2):1–10.
9. Matsuzaki K, et al. (1998) Relationship of membrane curvature to the formation of pores by magainin 2. *Biochemistry* 37(34):11856–11863.
10. Huang HW (2000) Action of antimicrobial peptides: Two-state model. *Biochemistry* 39(29):8347–8352.
11. Schmidt NW, Wong GCL (2013) Antimicrobial peptides and induced membrane curvature: Geometry, coordination chemistry, and molecular engineering. *Curr Opin Solid State Mater Sci* 17(4):151–163.
12. Brötz H, Bierbaum G, Leopold K, Reynolds PE, Sahl HG (1998) The lantibiotic mersacidin inhibits peptidoglycan synthesis by targeting lipid II. *Antimicrob Agents Chemother* 42(1): 154–160.
13. Park CB, Kim HS, Kim SC (1998) 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* 244(1):253–257.
14. Yonezawa A, Kuwahara J, Fujii N, Sugiura Y (1992) Binding of tachyplesin I to DNA revealed by footprinting analysis: Significant contribution of secondary structure to DNA binding and implication for biological action. *Biochemistry* 31(11):2998–3004.
15. Patrzykat A, Friedrich CL, Zhang L, Mendoza V, Hancock REW (2002) Sublethal concentrations of pleurocidin-derived antimicrobial peptides inhibit macromolecular synthesis in Escherichia coli. *Antimicrob Agents Chemother* 46(3):605–614.
16. Otvos L, Jr, et al. (2000) Interaction between heat shock proteins and antimicrobial peptides. *Biochemistry* 39(46):14150–14159.
17. Bowdish DME, Davidson DJ, Hancock REW (2005) A re-evaluation of the role of host defence peptides in mammalian immunity. *Curr Protein Pept Sci* 6(1):35–51.
18. Gilliet M, Lande R (2008) Antimicrobial peptides and self-DNA in autoimmune skin inflammation. *Curr Opin Immunol* 20(4):401–407.
19. Porto WF, Pires ÁS, Franco OL (2012) CS-AMPPred: An updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLoS One* 7(12):e51444.
20. Fjell CD, et al. (2009) Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J Med Chem* 52(7):2006–2015.
21. Hilpert K, Fjell CD, Cherkasov A (2008) Short linear cationic antimicrobial peptides: Screening, optimizing, and prediction. *Methods Mol Biol* 494(8):127–159.
22. Fjell CD, Hiss JA, Hancock REW, Schneider G (2011) Designing antimicrobial peptides: Form follows function. *Nat Rev Drug Discov* 11(1):37–51.
23. Mitchell JBO (2014) Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* 4(5):468–481.
24. Yee LC, Wei YC (2012) Current modeling methods used in QSAR/QSPR. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, eds Dehmer M, Varmuza K, Bonchev D (Wiley-VCH, Weinheim, Germany), Vol 2, pp 1–31.
25. Lata S, Sharma BK, Raghava G (2007) Analysis and prediction of antibacterial peptides. *BMC Bioinformatics* 8:263.
26. Torrent M, Andreu D, Nogués VM, Boix E (2011) Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS One* 6(2):e16968.
27. Maccari G, et al. (2013) Antimicrobial peptides design by evolutionary multiobjective optimization. *PLOS Comput Biol* 9(9):e1003212.
28. Fjell CD, et al. (2008) Identification of novel host defense peptides and the absence of α-defensins in the bovine genome. *Proteins* 73(2):420–430.
29. Cherkasov A, et al. (2009) Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chem Biol* 4(1):65–74.
30. Bi J, Bennett K, Embrechts M, Breneman C, Song M (2003) Dimensionality reduction via sparse support vector machines. *J Mach Learn Res* 3:1229–1243.
31. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297.
32. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.
33. Aizerman A, Braverman EM, Rozoner LI (1964) Theoretical foundations of the potential function method in pattern recognition learning. *Autom Remote Control* 25: 821–837.
34. Boser BE, Guyon IM, Vapnik VN (1992) *A Training Algorithm for Optimal Margin Classifiers* (ACM, New York), pp 144–152.
35. Engel T (2006) Basic overview of chemoinformatics. *J Chem Inf Model* 46(6): 2267–2277.
36. Schneider G, Baringhaus K-H (2008) *Molecular Design* (Wiley-VCH, Weinheim, Germany), pp 212–231.
37. Arora JS (2017) *Introduction to Optimum Design* (Academic, New York), 4th Ed, pp 771–781.
38. Shoval O, et al. (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336(6085):1157–1160.
39. Schmidt NW, et al. (2012) Molecular basis for nanoscopic membrane curvature generation from quantum mechanical models and synthetic transporter sequences. *J Am Chem Soc* 134(46):19207–19216.
40. Gilks WR (July 15, 2005) Markov Chain Monte Carlo. *Encyclopedia of Biostatistics*, eds Armitage P, Colton T, (John Wiley & Sons, Ltd., Chichester, UK), 10.1002/0470011815.b2a14021.
41. Gilks WR, Richardson S, Spiegelhalter DJ (1995) Introducing Markov Chain Monte Carlo. *Markov Chain Monte Carlo in Practice*, eds Gilks WR, Richardson S, Spiegelhalter DJ (Chapman & Hall/CRC, Boca Raton, FL), pp 1–19.
42. Geyer CJ (1992) Practical Markov Chain Monte Carlo. *Statist Sci* 7(4):473–483.
43. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405.
44. Zhang H, et al. (2011) Critical assessment of high-throughput standalone methods for secondary structure prediction. *Brief Bioinform* 12(6):672–688.
45. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* 7:301.
46. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J Comput Phys* 23(2):187–199.
47. McInturff JE, et al. (2005) Granulysin-derived peptides demonstrate antimicrobial and anti-inflammatory effects against Propionibacterium acnes. *J Invest Dermatol* 125(2): 256–263.
48. El Karim IA, Linden GJ, Orr DF, Lundy FT (2008) Antimicrobial activity of neuropeptides against a range of micro-organisms from skin, oral, respiratory and gastrointestinal tract sites. *J Neuroimmunol* 200(1-2):11–16.
49. Hansen CJ, Burnell KK, Brogden KA (2006) Antimicrobial activity of Substance P and Neuropeptide Y against laboratory strains of bacteria and oral microorganisms. *J Neuroimmunol* 177(1-2):215–218.
50. Brogden KA, Guthmiller JM, Salzet M, Zasloff M (2005) The nervous system and innate immunity: The neuropeptide connection. *Nat Immunol* 6(6):558–564.
51. Maisetta G, et al. (2013) pH-dependent disruption of Escherichia coli ATCC 25922 and model membranes by the human antimicrobial peptides hepcidin 20 and 25. *FEBS J* 280(12):2842–2854.
52. Singh M, Mukhopadhyay K (2014) Alpha-melanocyte stimulating hormone: An emerging anti-inflammatory antimicrobial peptide. *BioMed Res Int* 2014(6):874610.
53. Friedman R, Pellarin R, Caflisch A (2009) Amyloid aggregation on lipid bilayers and its impact on membrane permeability. *J Mol Biol* 387(2):407–415.
54. Caillon L, Killian JA, Lequin O, Khemtémourian L (2013) Biophysical investigation of the membrane-disrupting mechanism of the antimicrobial and amyloid-like peptide dermaseptin S9. *PLoS One* 8(10):e75528.
55. Schmidt NW, Mishra A, Wang J, DeGrado WF, Wong GCL (2013) Influenza virus A M2 protein generates negative Gaussian membrane curvature necessary for budding and scission. *J Am Chem Soc* 135(37):13710–13719.
56. von Heijne G (1986) Towards a comparative anatomy of N-terminal topogenic protein sequences. *J Mol Biol* 189(1):239–242.
57. Cao D-S, Xu Q-S, Liang Y-Z (2013) Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29(7):960–962.
58. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in python. *Journal Mach Learn Res* 12:2825–2830.