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

Ernest Y. Lee\textsuperscript{a}, Gerard C.L. Wong\textsuperscript{a,b,*}, Andrew L. Ferguson\textsuperscript{c,d,*}

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

\textbf{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

\textbf{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. \textit{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.

\textcopyright{} 2017 Elsevier Ltd. All rights reserved.

\textbf{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.\textsuperscript{1} The present mini-review foregrounds the computational aspects of this work; another recent invited mini-review takes a more experimental vantage.\textsuperscript{2} We first discuss the new insights furnished by the
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 mechanisms4,12-15 – 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 action1-8,26 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 (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 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 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 holdout 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/sexLinks.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 Giguer 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 preprocessing step prior to ANN classification.34 In our own recent work, we developed a SVM classifier to predict x-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 physicochemical signatures. 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.\(^1\,\)\(^2\)\(^3\) 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 \(x_i\) in the \(m\)-dimensional feature space can be represented by an \(m\)-element vector \(x_i\). A \((m - 1)\)-dimensional hyperplane in the feature space is defined by its surface normal vector \(w\) and offset from the origin \(b / ||w||\). This plane comprises the locus of points \(x\) satisfying the relation \(w \cdot x = b\). A data point \(i\) is classified by which side of the hyperplane upon which it falls, as a “hit” if \(w \cdot x_i > b\) and a “miss” if \(w \cdot x_i < b\). The margins around the hyperplane are defined by the parallel hyperplanes \(w \cdot x = b \pm 1\) containing the closest “hit(s)” and \(w \cdot x = b - 1\) containing the closest “miss(es)”. The size of the gap is defined by the projection of the vector \((x_i - x)\) onto the unit vector \(w = w / ||w||\).

\[
\begin{align*}
\text{w} \cdot (x_i - x) & = \frac{1}{||w||} (w \cdot x_i - w \cdot x) \\
& = \frac{1}{||w||} (w \cdot x_i - b) - (w \cdot x_i - b) \\
& = \frac{1}{||w||} [1 - (-1)] = \frac{2}{||w||}.
\end{align*}
\]

Fig. 1. Schematic illustration of a support vector machine (SVM) classifier operating in a feature space of dimensionality \(m = 2\). Data points \(x\) are represented in this space by their \(m = 2\)-dimensional feature vectors. 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 \(x\) satisfying \(w \cdot x - b = 0\), where \(w\) is the (possibly non-unit) surface normal, \(w \cdot x - |w|| = \text{the corresponding unit vector, and} \ w \cdot x - b = |w|| \text{is the offset of the hyperplane from the origin. The margin is defined by the two parallel hyperplanes satisfying } w \cdot x = b + 1 \text{ and } w \cdot x = b - 1, \text{ and the SVM is trained by maximizing the width of the margin } w \cdot (x - x) = \frac{2}{||w||} \text{ where } x, \text{ is the closest “hit” to the hyperplane separator and } x, \text{ 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.}

\[
\begin{align*}
\arg \min_{w,b} \frac{1}{2} ||w|| & \quad \text{subject to } y_i(w \cdot x_i - b) \geq 1, \quad i = 1 \ldots n. \quad (2)
\end{align*}
\]

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,

\[
\begin{align*}
\arg \min_{w,b} \frac{1}{2} ||w|| + \frac{1}{n} \sum_{i=1}^{n} L_1(y_i, x_i).
\end{align*}
\]

\[
L_1(y_i, x_i) = \max (0, 1 - y_i(w \cdot x_i - b)), \quad (3)
\]

where \(L_1(y_i, x_i)\) defines a hinge loss function that penalizes points that fall into or beyond the margin (i.e., those with \(y_i(w \cdot x_i - b) < 1\)) in linear proportion to how far across the margin they reside, and \(\lambda\) is a hyperparameter controlling the relative weighting between maximizing the margin and incorrect predictions that is typically tuned by cross-validation.\(^2\)\(^3\)\(^7\) The minimization in Eq. (3) is known as the primal problem, and training of the SVM classifier amounts to determining the values of \(w\) 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.\(^7\)

SVMs are an inherently linear classification technique that discriminates 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.\(^7\) 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.\textsuperscript{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,\textsuperscript{9,40} which may be beneficial in sub-categorizing AMPs as demonstrated by Xiao et al.\textsuperscript{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 (\url{http://aps.ummc.edu/AB})\textsuperscript{3} 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 (\url{http://pdbtm.enzim.hu}).\textsuperscript{41,44} Both datasets comprised peptides of length 8–60 residues and \( \alpha \)-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 for use in SVM classifier training and cross-validation. For classification problems, one must also pay attention to the proportionality of data sets originating from the 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.\textsuperscript{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.\textsuperscript{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,\textsuperscript{46,47} great care must be taken in training over imbalanced datasets.\textsuperscript{7,48} 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.\textsuperscript{47,48} In general, the training and testing partitions should maintain the same class balance.\textsuperscript{46}

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.\textsuperscript{24} 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{52,53} Large numbers of these derived descriptors may be generated, and feature selection conducted to systematically identify an optimal subset of descriptors.\textsuperscript{24,55} 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.\textsuperscript{54,55}

In this work, we employed the propy Python package\textsuperscript{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.\textsuperscript{46} 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.\textsuperscript{56} This approach replaces the soft margin minimization defined in Eq. (3) by,

\[
\arg \min_{w, b} \left[ \frac{1}{n} \sum_{i=1}^{n} L_2(y_i, x_i) + C \sum_{j=1}^{m} \lambda_j \| w_j \|_1 \right],
\]

where \( \| w \|_1 \) is the \( \ell_1 \)-norm of \( w \), and \( L_2(y_i, x_i) \) is a squared hinge loss. Similar to the LASSO method,\textsuperscript{57} the \( \ell_1 \)-norm serves as a numerically efficient proxy for the \( \ell_0 \)-norm that enforces sparsity in the \( w \) vector. Geometrically, the classification hyperplane is perpendicular to those descriptors with zero-value elements in \( w \), 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.\textsuperscript{46} We solve this minimization over the \( n = 486 \) training points using the scikit-learn Python machine learning library,\textsuperscript{58} and determine the optimal 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.\textsuperscript{56,23} The particular descriptors selected vary between the sparse models, and we stabilize feature selection by bootstrap aggregation (bagging)\textsuperscript{59} 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 \( w \) and \( b \). Training was conducted using the scikit-learn Python machine learning library,\textsuperscript{58} 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
predicted and observed classifications. The classifier was trained cient – of 0.837, revealing a strong correlation between the well along these metrics, with 90.9% of negative predictions experimentally synthesized. Nevertheless, our classifier also performs the size of the accessible sequence space is so large that failure to cor-tive rate) are typically less of a priority for virtual screening. The High sensitivity and negative predictive value (i.e., low false nega-

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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>Description</th>
<th>Weight (w_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>netCharge</td>
<td>Net peptide charge</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>c_i^c_j</td>
<td>Length-normalized sequence order coupling number measuring physicochemical correlations between residues separated by two positions (i, j+2) measured by the Grantham chemical distance matrix</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>p_(\beta_{0})</td>
<td>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</td>
<td>0.36</td>
</tr>
<tr>
<td>4</td>
<td>SolventAccessD1025</td>
<td>Fraction of the peptide length containing 25% of the buried amino acid residues AL,FC,GL,W</td>
<td>-0.24</td>
</tr>
<tr>
<td>5</td>
<td>pc(M,K)</td>
<td>Relative fraction of M residues to K residues</td>
<td>-0.21</td>
</tr>
<tr>
<td>6</td>
<td>p_(\beta_{k})</td>
<td>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</td>
<td>0.20</td>
</tr>
<tr>
<td>7</td>
<td>AE</td>
<td>Fraction of contiguous AE residue pairs</td>
<td>0.18</td>
</tr>
<tr>
<td>8</td>
<td>c_i^c_j</td>
<td>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</td>
<td>-0.17</td>
</tr>
<tr>
<td>9</td>
<td>LW</td>
<td>Fraction of contiguous LW residue pairs</td>
<td>0.17</td>
</tr>
<tr>
<td>10</td>
<td>NK</td>
<td>Fraction of contiguous NK residue pairs</td>
<td>0.13</td>
</tr>
<tr>
<td>11</td>
<td>DP</td>
<td>Fraction of contiguous DP residue pairs</td>
<td>-0.12</td>
</tr>
<tr>
<td>12</td>
<td>FC</td>
<td>Fraction of contiguous FC residue pairs</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

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:

- accuracy = \(\frac{TP + TN}{TP + FP + TN + FN}\) = 91.9%
- specificity = \(\frac{TN}{TP + FN}\) = 93.0%
- sensitivity = \(\frac{TP}{TP + FN}\) = 90.7%
- positive predictive value (PPV) = \(\frac{TP}{TP + FP}\) = 92.9%
- negative predictive value (NPV) = \(\frac{TN}{TN + FN}\) = 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 experimen-tally synthesized. Nevertheless, our classifier also performs well along these metrics, with 90.8% 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. 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.

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. 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 \(\sigma\) 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”, \(\sigma \geq 0\)) or not (“miss”, \(\sigma < 0\)). It is also possible to convert \(\sigma\) 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 \(\sigma\) and \(P(1)\). We initially hypothesized that distance from the separating hyperplane would be correlated with antimicrobial activity, defining an inverse relationship between \(\sigma\) 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 microbialic activity far from the margin on the “miss” side, and peptides with weak microbialic 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.antistaphylolytic.com) and ran them through our SVM classifier to compute the predicted \(\sigma\) values. The Spearman correlation coefficient between \(\sigma\) 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 microbial 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. Due to these confounding factors it is perhaps not surprising in retrospect that we should see no correlation between \( \sigma \) 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). Induction of NGC has also been observed for other classes of membrane active peptides, including cell-penetrating peptides and viral fusion peptides, and is a necessary requirement of a number of membrane destabilization mechanisms including portion, blebbing, and budding. Geometrically, NGC possesses a clear mathematical and geometric interpretation. Geometrically, NGC possesses a clear mathematical and geometric interpretation. At each point on a surface we can define the curvature \( \kappa \) along an arbitrary tangent vector as the reciprocal of the mean curvature \( H \) 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 \( K_1 \) and \( K_2 \) are respectively defined by the kissing circles that give maximum and minimum values of the curvature. The mean curvature \( H \) of the surface is given by the arithmetic mean of the principal curvatures \( H = (K_1 + K_2) / 2 \), and the Gaussian curvature is given by their product \( K = K_1 K_2 \). 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 \( \alpha \)-helical peptides with varying degrees of homology to known AMPs that were positively classified by the SVM \( (\sigma > 0) \) with a range of \( \sigma \) 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 \( l(q) \) 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 \( \chi \) 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 \( \langle K \rangle = 2\pi l/A_0 a^2 \). Correlating \( \langle K \rangle \) with \( \sigma \), we find a strong and statistically significant positive correlation \( r = 0.65 \) \( p = 0.006 \), 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., \( \langle K \rangle \leq 0 \) 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 \( \sigma \) predicted by the trained SVM classifier. (a) A scatterplot of \( \sigma \) against in vitro minimum inhibitory concentration (MIC) for 478 AMPs active against Staphylococcus aureus reveals no significant correlation \( (r = -0.06, p = 0.19) \). (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 Pn3m and Im3m cubic phase space groups that are rich in NGC. (d) A scatterplot of \( \sigma \) against average NGC \( (K) \) 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 \( (r = 0.65, p = 0.006) \). Panels a, c, and d are adapted from Lee et al. Proc. Natl. Acad. Sci. USA 113 48 13588–13593 (2016).]
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 \( \mathbf{w} \) vector defining the surface normal of the separating hyperplane. Geometrically, large magnitude elements of \( \mathbf{w} \) 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 \( \mathbf{w} \) vector elements associated with each of the \( m = 12 \) descriptors ranked-order by magnitude is presented in Table 1.

The predictive performance of a machine learning classifier is, of course, independent of any \textit{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.  

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.

\textbf{Charge.} The top-ranked feature possessing the most discriminatory weight is the peptide net charge, possessing a \( \mathbf{w} \) 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.

\textbf{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.4,7,9,12–15 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 \( \alpha \)-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 \( \alpha \)-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.

\textbf{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 \( \mathbf{w} \) 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 \( \eta \) 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 \( \eta \), \( P_{\text{miss}}(\eta) = 1 - P_{\text{hit}}(\eta) \), 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 (\( p = 4.5 \times 10^{-6}, 1.2 \times 10^{-6}, 8.9 \times 10^{-7} \), 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.

\textbf{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. Hydrophobic residues can induce positive membrane curvature by steric displacement of phospholipids in the membrane. Cationic peptides can generically induce negative mean membrane curvature due to electrostatic attraction and wrapping of the membrane around the peptide. By doubly coordinating phosphate head groups, arginine can by itself induce NGC, whereas singly-coordinating lysine can only generate negative mean curvature. 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, 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 collected 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,

\[ P_{\text{accept}} = \min \{ 1, \exp \left( \frac{(\sigma_{\text{trial}} - \sigma_{\text{current}})}{T} \right) \} \]

where \( \sigma_{\text{current}} \) is the distance from the hyperplane assigned by the SVM classifier to the current peptide sequence in the MCMC chain, \( \sigma_{\text{trial}} \) is that for the trial peptide, and \( T = 0.8 \) is a fictitious “temperature” that controls the acceptance ratio. We solved this optimization problem to identify those Pareto optimal sequences for which no one criterion can be improved without diminishing another. 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. We present in Fig. 4a a scatter-plot 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. 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 \( 2^N \) possible...
sequences for an N-residue protein.\textsuperscript{35} Computationally efficient Q SAR 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.\textsuperscript{26–28,32,33} Nevertheless, for peptides longer than 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 $\sigma$ 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,\textsuperscript{92,93} and a subset may then be hierarchically ranked along any number of metrics 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,\textsuperscript{92,93} and a subset may then be hierarchically ranked along any number of metrics highly by the classifier. The computationally screened candidates may then be hierarchically ranked along any number of metrics.

Here we remark on an alternative application of Q SAR 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)\textsuperscript{101} to search for membrane activity in a variety of peptide families with established primary function.\textsuperscript{3} 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 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.\textsuperscript{102,103} This analysis illustrates the utility of trained and validated Q SAR 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.\textsuperscript{1} Our introduction to support vector machines exposes the elegant simplicity of their mathematical foundations that can provide interpretable Q SAR 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 predictions 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

