JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

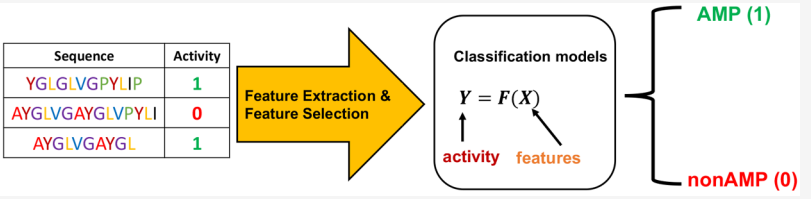
pubs.acs.org/jcim Article

Bacteria-Specific Feature Selection for Enhanced Antimicrobial Peptide Activity Predictions Using Machine-Learning Methods

Hamid Teimouri, Angela Medvedeva, and Anatoly B. Kolomeisky*

Cite This: J. Chem. Inf. Model. 2023, 63, 1723−1733 Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: There are several classes of short peptide molecules, known as antimicrobial peptides (AMPs), which are produced during the immune responses of living organisms against various infections. In recent years, substantial progress has been achieved in applying machine-learning methods to predict the activities of AMPs against bacteria. In most investigated cases, however, the outcome is not bacterium-specific since the specific features of bacteria, such as chemical composition and structure of membranes, are not considered. To overcome this problem, we developed a new computational approach that allowed us to train several supervised machine-learning models using a specific set of data associated with peptides targeting *E. coli* bacteria. LASSO regression and Support Vector Machine techniques have been utilized to select, among more than 1500 physicochemical descriptors, the most important features that can be used to classify a peptide as antimicrobial or ineffective against *E. coli*. We then performed the classification of active versus inactive AMPs using the Support Vector classifiers, Logistic Regression, and Random Forest methods. This computational study allows us to make recommendations of how to design more efficient antibacterial drug therapies.

INTRODUCTION

Antimicrobial peptides (AMPs), which can be produced by both eukaryotic and prokaryotic organisms, play an important role in the immune systems of mammals and plants.[1−3] It is well-known that AMPs have a net positive charge and are in general amphipathic, i.e., they have both hydrophobic and hydrophilic spatially separated segments. These characteristics allow them to attach to anionic (negatively charged) bacterial membranes and exhibit their antibacterial activity. Thus, the antimicrobial functioning is primarily dependent on the specific interactions between AMPs, particularly the N-terminus of these molecules,[4] and bacterial membranes,[5] such that a certain peptide can disrupt the membrane of a specific bacterium while it might not be active against other bacteria.[6] This is confirmed by the observations that larger fractions of anionic lipids in bacterial membranes result in increased membrane disruption and permeabilization by cationic AMPs.[7]

Machine-learning methods have been widely employed in studying AMPs, and they primarily aim at predicting the antimicrobial activity of an arbitrary peptide from its amino-acid sequence.[8,9] The performance and reliability of such approaches are mainly dependent on the training data, so many prediction tools have been developed in conjunction with AMP databases.[10−14] These tools can in turn be used to classify newly discovered peptides as antimicrobial or active against another target such as cancer or fungi,[15] but the predictions are not bacterium-specific. Previous machine-learning models were mainly trained based on a data set composed of AMP activity data targeting mixed species of bacteria.[8] Recently, a new machine-learning pipeline approach was developed that predicts the antimicrobial activity of peptides targeting separately Gram-positive and Gram-negative bacteria.[16] Since each group of bacteria (Gram-negative and Gram-positive) have different membrane architectures, choosing two different data sets for each group would partially take the role of the bacterial membrane into account. However, within the broad categories of Gram-positive or Gram-negative bacterial species, there are significant differences in physiological and biochemical properties. In the Gram-negative bacteria, for example, greater genetic plasticity in non-fermenting Gram-negative bacilli can lead to greater resistance

Received: December 12, 2022
Published: March 13, 2023

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

to antimicrobial agents compared to another Gram-negative bacterium, Enterobacteriaceae.[17]

The efficacy of an AMP in inhibiting a specific bacterium depends on the unique features of interactions between the AMP molecules and the bacterium. As predicted in ref 8 from a machine learning study of AMP features contributing to antimicrobial activity, the most important characteristic of the peptide that makes it effective against bacteria is the molecular net charge. However, recent studies have also shown that some other features of AMPs are important in their antimicrobial activity and that these features may be specific to the bacterial target.[18,19] Thus, a machine learning analysis of AMP characteristics associated with bacterium-specific efficacy could reveal new features of AMPs that are important against specific bacterium, creating opportunities for a better understanding of microscopic mechanisms of AMP function as well as for developing of new antimicrobial drugs.

Although the probability of development of resistance against AMPs is typically low, it is still possible.[20] Given the possibility of AMP resistance, it is crucial to identify the features that make bacteria susceptible to AMPs.

The aim of the current study is to apply a bacterium-specific machine-learning approach using a recently developed feature selection method.[21] This should contrast with previous machine-learning models that included different species of bacteria in the same data set.[16,21] In this study, we focus on investigating the properties of AMPs targeting the specific bacterium *E. coli*, and similar analysis was also performed for other species including *Acinetobacter baumannii* and *Pseudomonas aeruginosa*. Using machine-learning prediction algorithms, specifically Logistic Regression and Support Vector Machine (SVM), we demonstrate that a small proportion of the AMPs' physicochemical features is sufficient to predict whether an unknown peptide will be effective against *E. coli*. We identify the most important features using LASSO- and SVM-based feature selection methods and show that some features have a positive effect on antimicrobial activity against *E. coli* while others have a negative effect. Accordingly, we argue that the selected features can be incorporated in the rational design of more effective AMPs targeting *E. coli*. It is also argued that this method can be applied toward the design of AMPs for other bacterial targets.

METHODS

Data Set and Data Preprocessing. The schematic overview of our procedures is presented in Figure 1. We considered *E. coli* bacterium because it is listed as an urgent threat in the World Health Organization (WHO) priority list of the most dangerous health issues[22] and it is a common model organism for investigations of antimicrobial activities.[23] This is also one of the most studied bacterial species from biochemical and pharmacological points of view, providing us with the abundant data needed for our analysis.

A DBAASP database[24] has been utilized to generate a list of peptide sequences (at least 11 amino acids long and without N- or C-terminal modifications), that were tested on the bacterium of interest, *E. coli*. It also included minimum inhibitory concentration (MIC) values, which define the minimum concentration of an antimicrobial agent required to completely inhibit the bacterial growth[25] (all MICs in our study are considered in units of μg/mL). To reduce the influence of experimental error on our results, AMPs were not included if there were multiple conflicting MIC values in relation to the same bacterium. Our data set for *E. coli* includes 183 AMP (active against bacteria) and 214 non-AMP (ineffective against bacteria) peptides. For *A. baumannii*, our data set includes 35 non-AMP (ineffective) and 87 AMP (active) peptides. We selected MIC values that were collected under the same experimental conditions: same solutions in which bacteria were cultured and at standard pH, ionic conditions, buffers, and other physicochemical properties.

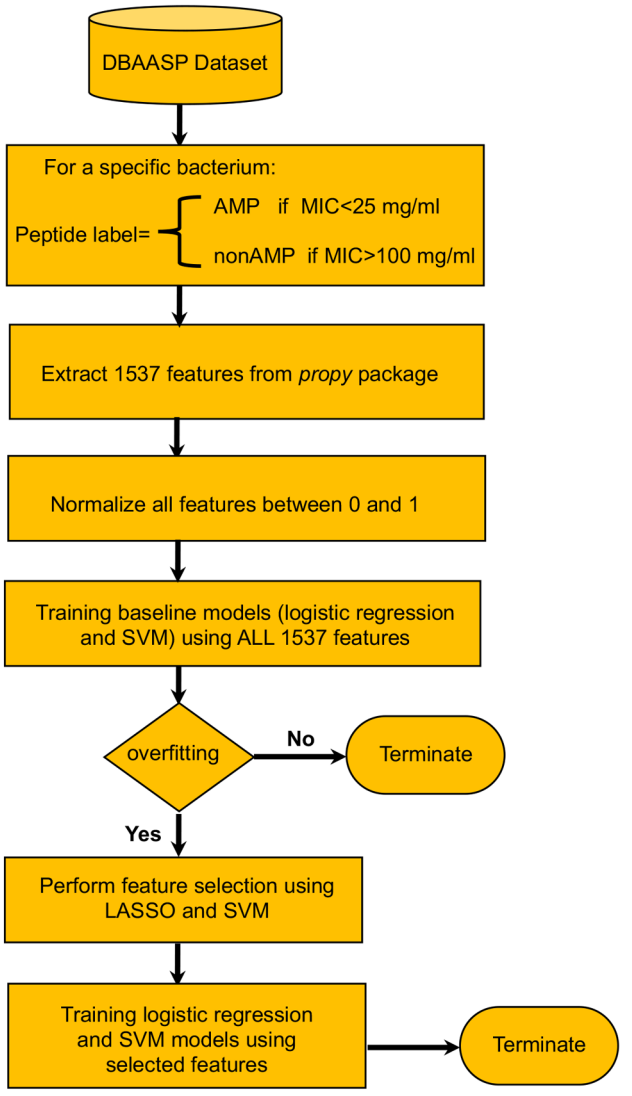


Figure 1. An overview of the bacteria-specific feature selection method for antimicrobial peptide activity prediction.

Generation of AMP Physicochemical Descriptors. From the amino-acid sequence, one can extract the physicochemical descriptors (net charge, hydrophobicity, etc.) and amino-acid composition patterns. We utilized a *propy* package to generate full descriptors for each peptide.[26] There were 1537 descriptors that broadly could be classified as having different basic character (e.g., charge), residue compositions (e.g., dipeptide composition), autocorrelations, chemical compositions, and sequence order features. Since the *propy* package only identifies natural amino acids, we did not include peptides with non-natural amino acids in our data set.

Supervised Machine Learning Algorithms. Predicting AMP activity from its sequence is a supervised learning

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

problem. Supervised learning algorithms, which use training data sets to train themselves to predict the desired output, can be divided into two types, namely, regression and classification algorithms. In our data set, a given peptide i is characterized by two entries. First, the peptide activity index, y_i, is 1 if the peptide acts as an AMP and 0 if it does not act as an AMP. Second is the feature vector $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n}\}$, whose elements describe different properties of the peptide i. Therefore, our response variable (the peptide activity index) is a categorical variable. Following standard cutoff points for antimicrobial activity,[16] the AMPs were labeled as inactive against *E. coli* if the MIC was greater than 100 μg/mL and as active if the MIC was less than or equal to 25 μg/mL. We considered predicting the MIC values instead of classifying activity, but no significant correlations were found between any *propy* descriptors and the MIC, in line with the finding from ref 21.

Predicting a categorical response for a peptide is a classification problem. Thus, we focus on the classification problem, which uses an algorithm to assign test data to certain classes. It is important to note that logistic regression is indeed a classification technique. There are different classification methods, including Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression, to predict active vs inactive AMPs. In the following, we briefly discuss these techniques to understand their advantages and disadvantages. Let us start with the simplest supervised machine-learning method, which is the Linear Regression approach.

Linear Regression. The simplest technique for supervised learning is a Linear Regression, which is a useful method for predicting the relationship between a target variable y_i and a set of independent variables $x_{i,1}, x_{i,2}, ..., x_{i,n}$,

$$y_i = b + \mathbf{w}\mathbf{x}_i = b + w_1 x_{i,1} + w_2 x_{i,2} + ... + w_n x_{i,n} \quad (1)$$

The unknown parameters are vectors $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ and parameters b which represent the slopes and intercepts, respectively, of the hyperplane defined by eq 1 with the axis of the multidimensional space. They are chosen in such a way so that they minimize the residual sum of squares (RSS), also known as the least-squares function,

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{l} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{l} (y_i - b - w_1 x_{i,1} - w_2 x_{i,2} - ... - w_n x_{i,n})^2 \end{aligned} \quad (2)$$

where l is the number of training data.

Logistic Regression. Logistic regression methods consider the problem of predicting a binary categorical response from the multiple features. The main relation of this approach is formulated as

$$\begin{aligned} y_i &= \frac{\exp(b + \mathbf{w}\mathbf{x}_i)}{1 + \exp(b + \mathbf{w}\mathbf{x}_i)} \\ &= \frac{\exp(b + w_1 x_{i,1} + ... + w_n x_{i,n})}{1 + \exp(b + w_1 x_{i,1} + ... + w_n x_{i,n})} \end{aligned} \quad (3)$$

This is a more convenient method for predicting antimicrobial activity of a peptide.

Support Vector Machine (SVM). In the Support Vector Machine (SVM), each data point is projected into an n-dimensional feature space in which each coordinate is associated with a physicochemical feature. We can think also of an $(n - 1)$-dimensional hyperplane that separates the data into two distinct volume spaces, corresponding to different classes of AMPs. The hyperplane equation is given by

$$\mathbf{w}\mathbf{x}_i + b = 0 \quad (4)$$

It is important to note that positive values of w_i correspond to the features that increase the AMP activity, while negative values indicate that these features negatively affect the antimicrobial activity. The SVM uses training data to learn the parameters $w_1, w_2, ..., w_n$, which, in turn, specify a classification rule for the data,[8]

$$y_i = F(\mathbf{x}_i) = \mathrm{sign}(\mathbf{w}\mathbf{x}_i + b) \quad (5)$$

where the sign function is defined as

$$\mathrm{sign}(a) = \begin{cases} 1, & a \geq 0 \\ -1, & a < 0 \end{cases} \quad (6)$$

For linearly separable data, one can select two parallel hyperplanes that distinguish two classes of data, so that the distance between them is as large as possible. The SVM method only uses several data points (support vectors) to determine the normal vector of the separating hyperplane. The two hyperplanes are defined as $\mathbf{w}\mathbf{x}_i - b = 1$ and $\mathbf{w}\mathbf{x}_i - b = -1$. The distance between these parallel hyperplanes is $\frac{2}{\|\mathbf{w}\|}$. Here $\|\mathbf{w}\|$ is defined as $\|\mathbf{w}\| = \sum_{i=1}^{n} w_i$. To define the most robust separating hyperplane, we must maximize the distance $\frac{2}{\|\mathbf{w}\|}$. It is important to note that, for mathematical convenience, we minimize $\frac{1}{2}\|\mathbf{w}\|_2$ instead of $\|\mathbf{w}\|$, where $\|\mathbf{w}\|_2 = (\sum_{i=1}^{n} w_i^2)^{1/2}$. Thus, one needs to evaluate

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\|\mathbf{w}\|_2\right) \quad (7)$$

subject to a condition $y_i(\mathbf{w}\mathbf{x}_i + b) - 1 \geq 0$ for $i = 1, ..., l$. The corresponding Lagrangian for specific minimization calculations reads as

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|_2 - \sum_{i=1}^{l} \alpha_i [y_i(\mathbf{w}\mathbf{x}_i + b) - 1] \quad (8)$$

where constants α_i are Lagrange multipliers. All the parameters $\mathbf{w}$, b, and α can be estimated by minimizing the Lagrangian.

However, there are situations when a single hyperplane (defined in eq 4) is not able to properly separate two classes of data. To circumvent this problem, it is convenient to define a soft margin formulation that almost separates the two classes by introducing the variable ξ_i, as explained in ref 27. The corresponding optimization problem now reads as

$$\min_{\mathbf{w},b,\xi}\left(\frac{1}{2}\|\mathbf{w}\|_2 + C\sum_{i=1}^{l}\xi_i\right) \quad (9)$$

subject to a condition $y_i(\mathbf{w}\mathbf{x}_i + b) - 1 \geq 1 - \xi_i$ for $i = 1, ..., l$. The sum over ξ_i, which measures the total amount of misclassifications of training data, is controlled by a hyperparameter C. This optimization problem can be solved using the corresponding Lagrangian,

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|_2 - \sum_{i=1}^{l} \alpha_i(y_i(\mathbf{w}\mathbf{x}_i + b) - 1 + \xi_i) + C\sum_{i=1}^{l}\xi_i \quad (10)$$

Ensemble Learning. An alternative way to build a more accurate method is to combine two or more independent models. This technique is known as an Ensemble Learning, and it can be implemented using two methods: bootstrap aggregation (bagging) and boosting.

Bagging involves segmentation of the data into bootstrap samples and applying a prediction model to each of them. The final predictions are selected by taking the majority vote for each predicted class across models (e.g., 1 if most models predicted 1) if the goal is classification or the averaged predicted values across all models if the goal is regression.[28]

Boosting is an adaptation of bagging that calculates a weighted average such that incorrect predictions are weighted higher and the procedure is repeated until accuracy is highest with the appropriate corresponding weights.[29] Boosting is used for prediction models that are too simple to perform accurately, called weak learners, such as the Decision Tree method with only two nodes.[28]

For feature classification and predictions based on the feature-classifications method, we implemented bagging with the data split into 15 stratified shuffled samples (*E. coli*), as was also done in ref 21. For our final set of features, we selected the features that were consistent across the samples,[21] and for classification we averaged the model accuracy across the samples for a robust estimate of each model's performance. We also used 15 stratified shuffled samples to tune hyperparameters with grid search, a method that tests and compares accuracy for one value at a time from a specified "grid", or vector of values. We selected the hyperparameter values that led to the highest model accuracy when averaged across the 15 samples.

Feature Selection Methods. When we deal with a high-dimensional features space, as for the AMP descriptors, it is extremely useful to find a small subset of the most predictive features. Mathematically, this corresponds to assigning zero weights to the most irrelevant or redundant features in the regression and the SVM methods [see eqs 1, 3, and 4]. Two common techniques for shrinkage and feature selection are LASSO (The Least Absolute Shrinkage and Selection Operator) regression and the Support Vector Machine.[30] LASSO regression is similar to the least-squares approach, except that one has to minimize the square error subject to the constraint $\sum_{j=1}^{n}|w_j| \le t$. Thus, the corresponding error function takes the form[31]

$$\sum_{i=1}^{l}\left(y_i - w_0 - \sum_{j=1}^{n} w_j x_{i,j}\right)^2 + \lambda\sum_{j=1}^{n}|w_j| \quad (11)$$

where $\lambda \ge 0$ is a tuning parameter, and a constant t is a function of λ. For the least-squares approach (when $\lambda = 0$), we have $t_0 = \sum_{j=1}^{n}|w_j^{(\mathrm{ls})}|$. When λ is nonzero but $t > t_0$, the coefficients remain the same as in the linear regression. However, for $t = ft_0$ with $0 \le f \le 1$, the least-squares coefficients are shrunk by about 100f%. Since both λ and t are interrelated tuning parameters, when λ is known, t must be adaptively estimated to minimize the expected prediction error [eq 11]. As shown in ref 32, the estimated weights from the LASSO method are given by

$$\hat{w}_j^{(\mathrm{lasso})}(\lambda) = (|\hat{w}_j^{(\mathrm{ls})}| - \lambda)\mathrm{sign}(\hat{w}_j^{(\mathrm{ls})}) = \begin{cases} -|\hat{w}_j^{(\mathrm{ls})}| + \lambda, & \hat{w}_j^{(\mathrm{ls})} < 0 \\ 0, & \hat{w}_j^{(\mathrm{ls})} = 0 \\ |\hat{w}_j^{(\mathrm{ls})}| - \lambda, & \hat{w}_j^{(\mathrm{ls})} > 0 \end{cases} \quad (12)$$

where $\hat{w}_j^{(\mathrm{lasso})}$ and $\hat{w}_j^{(\mathrm{ls})}$ are estimated weights from the Lasso and the least-squares methods, respectively. Equation 12 determines whether for a feature j, the corresponding coefficient w_j does satisfy the constraint $\sum_{j=1}^{n}|w_j| \le t$, and if so, it shrinks it by λ. Thus, the LASSO method shrinks the coefficients toward zero, removing the irrelevant and redundant descriptors.

Alternatively, one can use the Support Vector Machine for feature selection. The main idea here is to choose a proper norm of $\mathbf{w}$ in the optimization process, such that the estimated $\mathbf{w}$ becomes more sparse (most of the components of $\mathbf{w}$ shrink to 0). A simple way is to consider the first norm of $\mathbf{w}$ ($\|\mathbf{w}\| = \sum_{i=1}^{n} w_i$). The corresponding optimization problem reads, then,[33]

$$\min_{\mathbf{w},b,\xi}\left(\|\mathbf{w}\| + C\sum_{i=1}^{n}\xi_i\right) \quad (13)$$

with $z_i \ge w_i$, $z_i \ge -w_i$ for $i = 1, ..., l$. However, because the absolute values are not differentiable, it is convenient to rewrite the norm in the following form:

$$\min_{\mathbf{w},z,b,\xi}\left(\sum_{i=1}^{n} z_i + C\sum_{i=1}^{n}\xi_i\right) \quad (14)$$

subject to $y_i(\mathbf{w}\mathbf{x} + b) \ge 1 - \xi_i$, $\xi_i \ge 0$, and $z_i \ge w_i$, $z_i \ge -w_i$ for $i = 1, ..., l$. The solution of this optimization problem gives a set of estimated parameters (w^{*T}, b^*, ξ^*, z^*), among which we select the elements of w^{*T}, which satisfy eq 14.

Evaluating Models. The performance of any machine-learning algorithm can be characterized by the ratio of the correctly predicted samples, true positives (TP) and true negatives (TN), to the total number of samples,

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (15)$$

where FP and FN represent the number of false positives and false negatives, respectively. This ratio was utilized as a measure of accuracy in our theoretical analysis, and the average across stratified shuffled cross-validation sets was calculated, providing the score for each model's performance. Alternatively the model performance can be evaluated using a recall function, which is defined as[34]

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (16)$$

However, since our data set for *A. baumannii* is imbalanced (the number of AMPs is greater than the number of nonAMPs), it is also advantageous to use a Matthews's correlation coefficient (MCC), which is a more reliable statistical tool for more complex situations:[35,36]

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

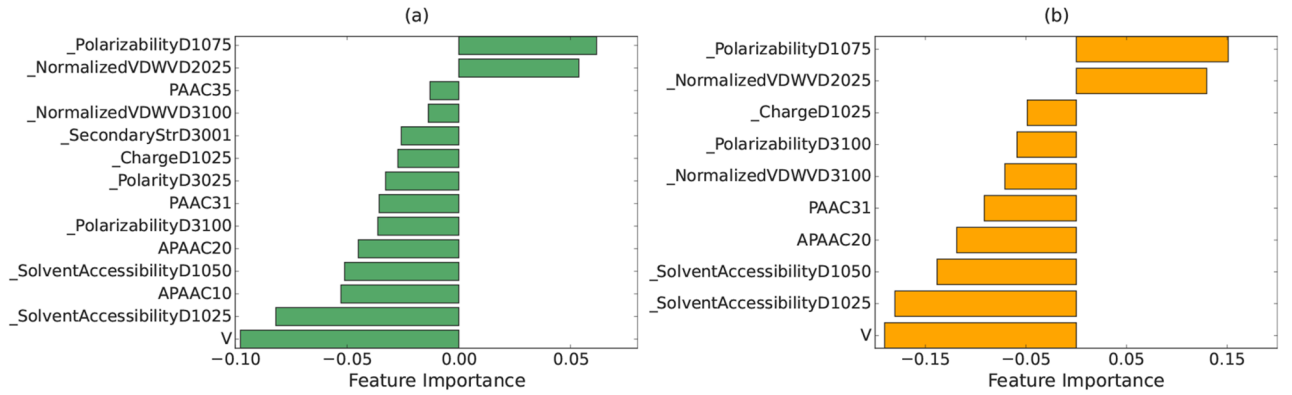


Figure 2. Relative importance of different physicochemical features in determining the antimicrobial activity against *E. coli* using (a) LASSO regression method and (b) the support vector machine. In computations, we utilized the following values for the hyperparameters: For LASSO, the Lagrange multiplier in eq 11 was set to be $\lambda = 0.01$. For SVM, the hyperparameter C (in eq 13), which is calculated via the grid search optimization, is equal to $C = 0.1$. In both methods, the number of stratified shuffled cross-validation sets is $n = 15$.

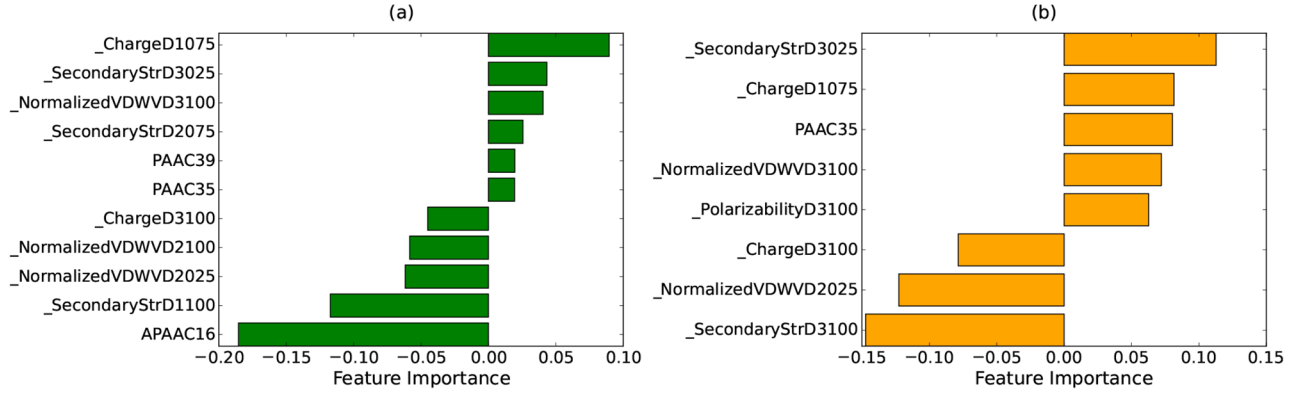


Figure 3. Relative importance of different physicochemical features in determining the antimicrobial activity against *A. baumannii* using (a) LASSO regression method and (b) the support vector machine. In computations, we utilized the following values for the hyperparameters: For LASSO, the Lagrange multiplier in eq 11 was set to be $\lambda = 0.01$. For SVM, the hyperparameter C (in eq 13), which is calculated by the grid search optimization, is equal to $C = 0.1$. In both methods, the number of stratified shuffled cross validation sets was $n = 5$.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (17)$$

RESULTS AND DISCUSSION

Results of feature selection using LASSO regression and linear SVM methods for bacteria *E. coli* and *A. baumannii* are presented in Figure 2 and Figure 3, respectively. The acronyms and their descriptions are defined in Figure 4. Features with negative scores have a negative effect on antimicrobial activity, while features with positive scores support the antimicrobial activity of a peptide.

Feature Importance Ranking for *E. coli*. As one can see in Figure 2, the composition frequency of amino acid valine (V) is selected by both methods as the most important feature with a negative impact on antimicrobial activity against *E. coli*. It is important to note that valine has very high hydrophobicity. Moreover, in the data set for *E. coli*, 87% of the peptides contain at least one valine and 22% of the peptides contain at least one dipeptide VV. Thus, the abundance of amino acid valine in the peptides targeting *E. coli* increases the hydrophobicity of those peptides. As was shown before,[37] a very high hydrophobicity can lead to self-association of peptides, and consequently it can significantly slow down the breaking of the bacterial cell membranes and the translocation through the membranes. In addition, one can see that both pseudo-amino acid composition (PAAC)[38] and amphiphilic pseudo-amino acid composition (APAAC),[39] which describe sequence order or specific pattern of amino acids, have negative effect on antimicrobial activity against *E. coli*.

Other selected features are related to the distribution of particular physicochemical properties along the peptide backbone. The corresponding features are hydrophobicity, van der Waals volume, polarity, polarizability, charge, secondary structure, and solvent accessibility.[40] As shown in Table 2 in ref 40, all 20 amino acids are categorized into three groups based on these features. The distribution descriptor (Di) for each group (i) has five components, which are defined in the following way. The fractional segment of the peptide sequence that accommodates the first residue, 25%, 50%, 75%, and 100% of the residues belonging to the group i ($i = 1, 2, 3$) are denoted as $Di001$, $Di025$, $Di050$, $Di075$, and $Di100$, respectively. Thus, these values lie between 0 and 100.

As one can see from Figure 2, among all physicochemical features, the solvent accessibility pattern of a peptide is predicted to have a strong negative effect on the antimicrobial activity against *E. coli*. Solvent accessibility of a peptide is simply defined as the total surface area of the peptide

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

Feature Acronym	Feature Description
V	Frequency of amino acid Valine
C	Frequency of amino acid Cysteine
I	Frequency of amino acid Isoleucine
_PolarizabilityD1075	The fraction of the entire sequence, where 75% the residues of group 1 (with polarizability value of 0-1) are contained.
_SolventAccessibilityD1025	The fraction of the entire sequence, where 25% the residues of group 1 (buried by solvent) are contained.
_NormalizedVDWVD2025	The fraction of the entire sequence, where 25% the residues of group 2 (high VW volume) are contained.
_SecondaryStrD3001	The fraction of the entire sequence, where the first residue of group 3 (coil) are contained.
_HydrophobicityD3025	The fraction of the entire sequence, where 25% of the residues of group 3 (hydrophobic) are contained.
_ChargeD3001	The fraction of the entire sequence, where the first residue of group 3 (negative charge) are contained.
_PolarityD1050	The fraction of the entire sequence, where 50% the residues of group 1 (with polarity value of 4.9-6.2) are contained.
APAAC10	Amphiphilic Pseudo Amino Acid Composition
PAAC35	Pseudo Amino Acid Composition

Figure 4. Description of the selected features in Figures 2 and 3

accessible to water molecules. Because the net (positive) charge of a peptide is essential for binding to the bacterial membrane, peptides that have a higher surface area are shielded with water molecules that leads to an effectively lower positive charge. Both methods predict that *SolventAccessibilityD*1025 and *SolventAccessibilityD*1050, which quantify the distribution of the amino acids of group 1 (buried by the solvent), negatively affect the antimicrobial activity. Thus, the uniform distribution of two or four (out of eight) buried residues along the peptide sequence can lessen the antimicrobial activity. However, the localized presence of buried amino acids in a compact segment of the peptide would alleviate their impact on antimicrobial activity.

Moreover, the two methods predict that *PolarizabilityD*1075 and *NormalizedVDWVD*2025 have positive effects on the antimicrobial activity. This suggests that the uniform distribution of a small number of residues with high van der Waals volume along the peptide sequence can enhance interaction of the peptide with the cell membrane. The same description applies to polarizability, which is proportional to the van der Waals volume. Likewise, the uniform distribution of large number of residues with intermediate van der Waals volume along peptide sequence (*NormalizedVDWVD*3100) reduces the antimicrobial activity.

Last but not least, one can see that the uniform distribution of a small number of residues with positive charge (*Charge*1025) along the peptide sequence does not support antimicrobial activity. In other words, concentration of positive charge in one part of the sequence can better sustain the association of peptides to the membrane.

Feature Importance Ranking for *A. baumannii*. For *A. baumannii*, both methods predict that *ChargeD*1075, which quantifies the distribution of 75% amino acids with positive charge, increases the antimicrobial activity (see Figure 3). On the other hand, *Charge*3100, which quantifies the distribution of all amino acids with negative charge (group 3) in the entire sequence, reduces the peptide association with membranes. However, it can be seen that other selected features have different effects on antimicrobial activity. For example, *NormalizedVDWVD*2025,in contrast to *E. coli*, has a negative effect on the antimicrobial activity against *A. baumannii*. Notably, *A. baumanii*, in contrast to *E. coli*, has more multidrug-resistant strains.[41,42] The differences in selected features between the two species may clarify the molecular picture of how AMPs interact differently with the membranes of *A. baumanii* species, suggesting the possibility to bypass *A. baumannii*'s resistance mechanisms to traditional antibiotics.

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

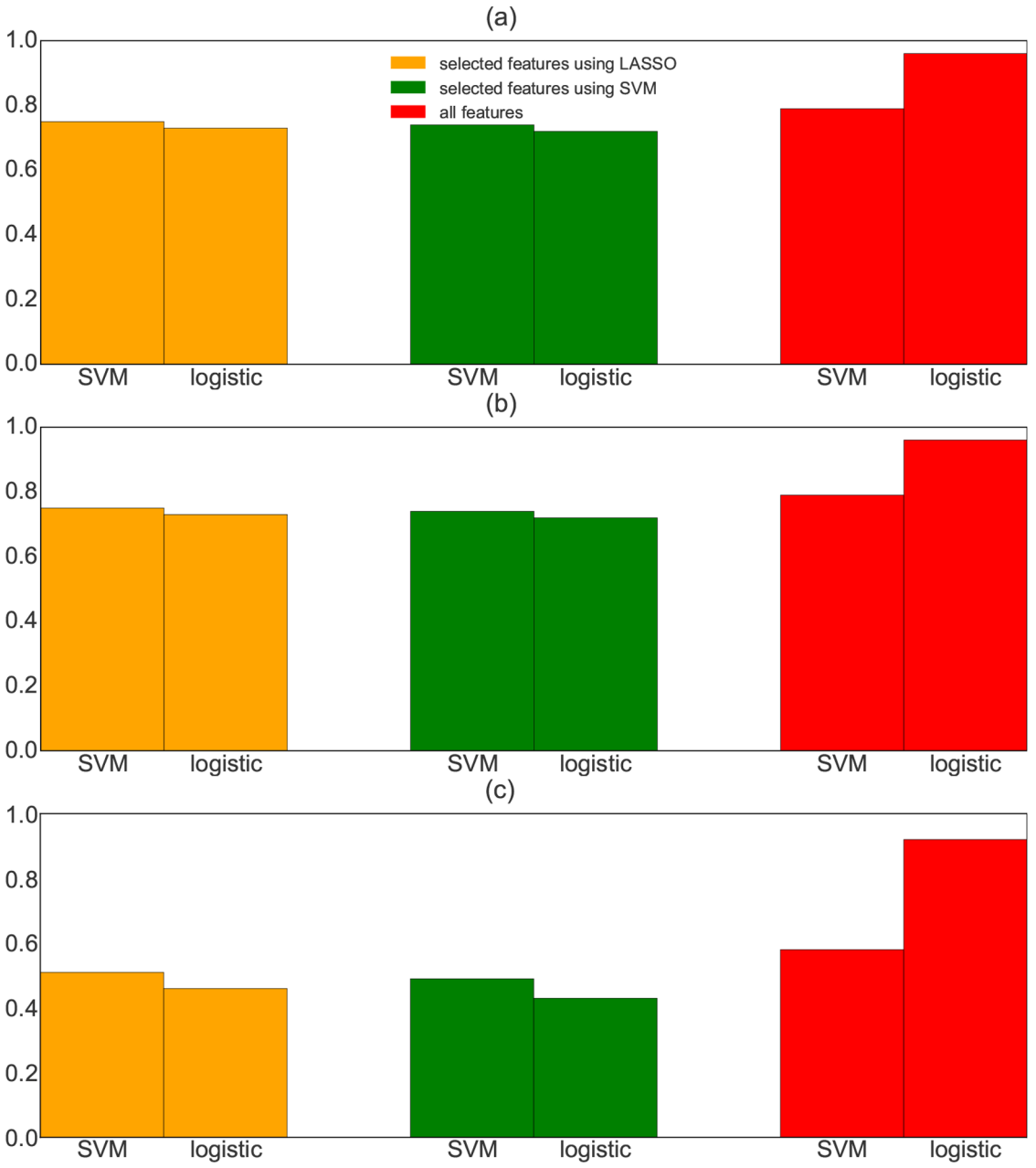


Figure 5. Results of bacteria-specific feature selection for *E. coli*. (a) Comparison of the accuracy for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) and LASSO (orange bars). (b) Comparison of the recall for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) and LASSO (orange bars). (c) Comparison of the Matthews's correlation coefficient for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) LASSO (orange bars). Each metric reflects the average value among 15 test cross-fold validation sets. A standard splitting of 80/20 (training/test) was applied for each fold.

Comparison of Different Machine-Learning Algorithms. Having extracted important physicochemical features, we now aim at comparing different machine-learning algorithms in successfully predicting whether an AMP was active against a specific bacterium. Using all features or only the features selected from the LASSO regression or only those selected from the SVM classification as the input, the accuracy, recall, and MCC of the SVM and the Logistic Regression in correctly classifying AMPs as active or inactive against bacteria *E. coli* and *A. baumannii* have been quantitatively compared (see Figures 5 and 6). Since our data for *E. coli* is slightly imbalanced, accuracy and recall are very close (Figure 5). For *A. baumannii*, however, the data are actually quite imbalanced (number of non-AMPs is less than half of AMPs) and thus recall is a bit different from accuracy (see Figure 6). As one can see in Figures 5 and 6, when all features are included, the MCC is higher for logistic regression in contrast to SVM. However, when only selected features are included, for *E. coli* the MCC is similar for SVM and logistic regression (Figure 5), while for *A. baumannii* the MCC is much higher for SVM in contrast to logistic regression (Figure 6). In ref 43, a systematic comparison of various metrics between SVM and logistic

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

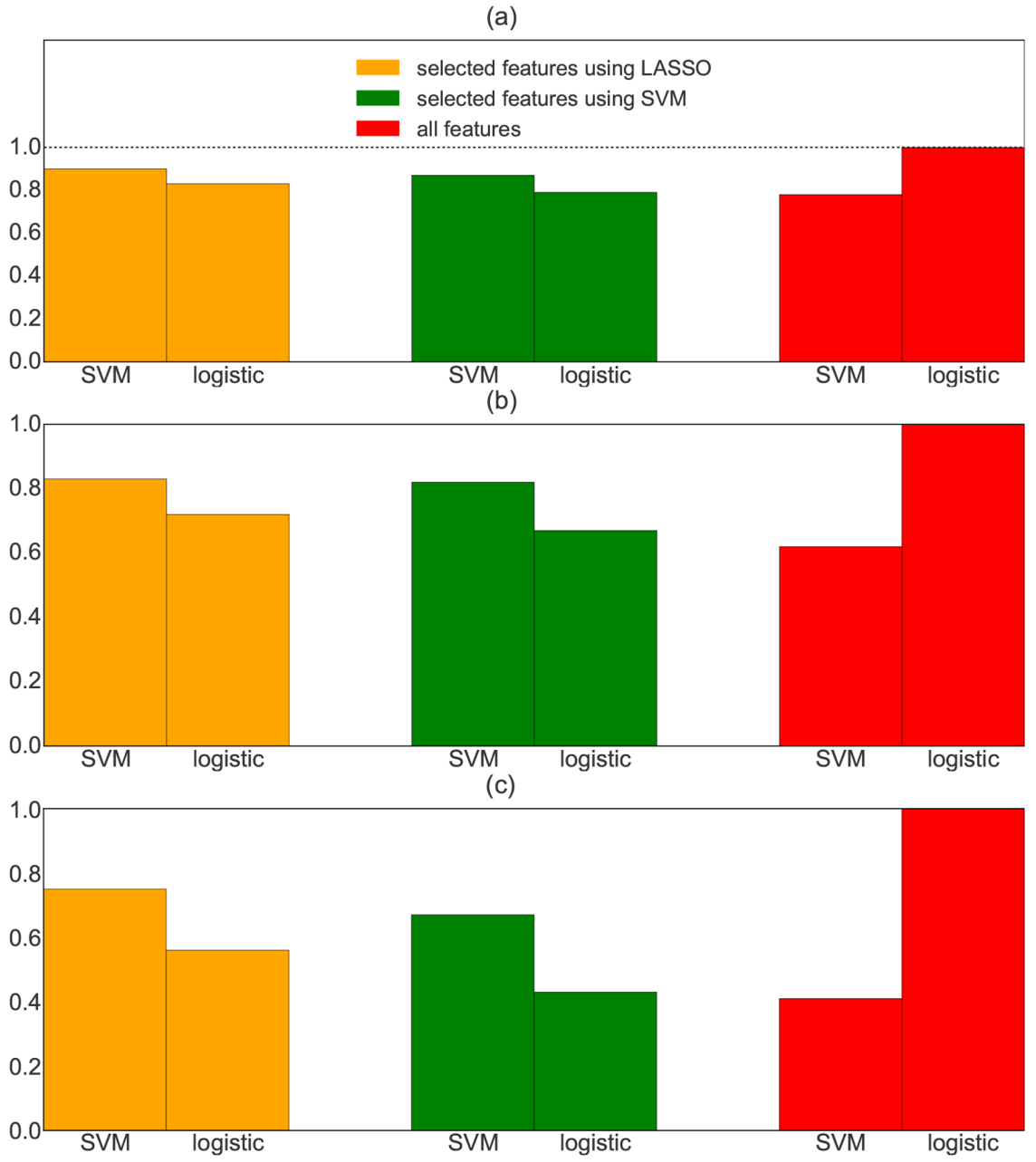


Figure 6. Results of bacteria-specific feature selection for *A. baumannii*. (a) Comparison of the accuracy for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) and LASSO (orange bars). (b) Comparison of the recall for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) and LASSO (orange bars). (c) Comparison of the Matthews's correlation coefficient for the trained baseline models (SVM and logistic regression) using all features (red bars) with the models trained using selected features from SVM (green bars) and LASSO (orange bars). Each metric reflects the average value among 5 test cross-fold validation sets. A standard splitting of 80/20 (training/test) was applied for each fold.

regression suggested that in comparison to logistic regression, SVM may be a better method to evaluate unbalanced data sets similar to *A. baumannii*.

SUMMARY AND CONCLUSIONS

In this study, we utilized supervised machine-learning methods to predict the activity of antimicrobial peptides against specific bacterial species. The baseline models were logistic regression and support vector classifiers for all physicochemical features of AMPs that have been used as input. We aimed to test the same models with the features that emerged from the feature-selection analysis. Specifically, we employed the LASSO Regression and the Support Vector techniques to select the most important set of physicochemical features, which have a positive or negative effect on antimicrobial activity against specific targets, *E. coli* and *A. baumannii*. The results indicate that for each bacterium there is a distinct set of physicochemical features related to antimicrobial activity, and it is important to consider each physicochemical feature in the context of other relevant properties of the system. For example, the results suggest that for more than one bacterium, solvent accessibility and the secondary structures are important

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

physicochemical characteristics, so when evaluating the relationship between antimicrobial activity of a peptide and specific secondary structure (e.g., α helix) for these bacteria, one must consider the degree of the peptide's solvent accessibility (buried or exposed).

It is also critical to compare our feature selections analysis with current investigations on this subject. In comparison to previous studies, the prediction accuracy for the SVM and logistic regression models that we explored was slightly lower, but this could be the result of our definition of accuracy that might differ from other studies. Our correlation matrix showed a greater number of correctly classified AMPs relative to a lower number of correctly classified non-AMPs because some AMPs were classified as non-AMPs. The results suggest that the model can accurately identify most but not all peptides with antimicrobial activity against *E. coli*, so that the cost of high accuracy is a conservative classification. Such conservative classification could be due to the non-AMP class showing antimicrobial activity against bacteria other than *E. coli*. Therefore, while these non-AMPs were not active against *E. coli*, they were still antimicrobial against other bacterial targets and, accordingly, had similarities to the AMP class active against *E. coli*. The non-AMP class in ref 21 consisted of peptides that were not active against any bacterial targets, so these peptides were likely more dissimilar to the AMP class. In other studies, the data sets were larger, so that the differences between AMP and non-AMP distributions of features would be clearer because the data were more likely to be normally distributed.[16,44] Additionally, the data are subject to experimental errors. Previous studies could have averaged across a large range of reported MIC values, which affected the classification of the peptide as AMP or non-AMP in the training data set. Finally, tuning of the hyperparameters in the feature-selection models could be improved, and it is suggested that future studies could use alternative hyperparameter tuning methods. Proper tuning of hyperparameters may increase the prediction accuracy. Finally, our approach was simple enough to enable comparison with the original implementation of the feature-selection method; it increased interpretability and decreased the likelihood of overfitting. As was shown before,[45] a combination of models can be explored to include the integrated random forest and SVM scores to linear regression to increase prediction accuracy. In ref 46, 30 baseline models were combined, including the logistic regression, random forest, and SVM. It is still notable that in our study a small proportion of features was able to satisfactorily predict the class of antimicrobial activity, and the prediction accuracy is expected to improve with additional data and greater reliability of the overall data set. Future studies can build on our approach with multiple combined models to increase the prediction accuracy.

It is important to note that the SVM and the LASSO feature selection methods are designed to identify the optimal *set* of features for prediction, while other methods such as Mann−Whitney U or z-test identify features that can independently discriminate between classes. Simply selecting all the features that independently discriminate between classes may not lead to a successful prediction algorithm, as indicated by some recent studies,[47] because it would not account for correlations between the features. In contrast, the SVM and LASSO methods will not include features that independently predict AMP vs non-AMP because their correlations with other features will not improve the predictions of the overall feature subset.

Our feature selection methods might not have captured all relevant correlations between the features, even if the independent contributions of the correlating features were high. However, it provides important microscopic insights on the relationship between peptide physicochemical properties and antimicrobial activity. Specifically, it was predicted that the abundance of amino acid valine makes a peptide *too* hydrophobic and thus reduces its activity against *E. coli*. Moreover, our feature selection methods suggest that not only the magnitude of physicochemical features but also the *distribution* of these quantities along the peptide chain can influence the antimicrobial activity. For example, the net charge was found to be the most important feature in ref 8, but here for a specific bacterium, *E. coli*, we found that the distribution of residues with positive charge in a small localized segment of the peptide backbone was among the selected features. Accordingly, our results show that it is important to consider the role of bacterial species in AMP interactions with bacteria membranes.

A major drawback in previous machine-learning predictions of antimicrobial peptide activity is that those models are not bacteria specific. A recent study tried to circumvent this problem by predicting the antimicrobial activity against Gram-negative and Gram-positive bacteria and utilized data sets for peptide activity against three Gram-negative bacteria, *E. coli* and *A. baumannii*, and *P. aeruginosa*.[16] Although these species share some common membrane architecture, there are key differences between them that might affect antimicrobial activity against them.[48] Vishnepolsky et al. were able to design an accurate clustering model trained on various Gram-negative bacteria, and the inputs to the model were nine physicochemical features of AMPs.[44] In contrast, our feature-selection approach (with inputs of over 1500 features) requires bacteria-specific training because different features of AMPs might be relevant to different bacteria. Accordingly, the features identified as important for *E. coli* might be different from the features found in ref 21 because they utilized peptides that target different species of bacteria. Future studies should investigate further how the AMP features are related to successful membrane interactions in specific multidrug-resistant bacteria strains compared to other bacteria strains. It will help to elucidate why certain species of bacteria are more likely to evolve resistant strains than others. Moreover, similar feature extraction methods can be used for antibiotics to perform feature selection, and the same approach can be used to extract features for antibiotics and AMPs to compare which features of each are associated with high antimicrobial activity.

The rational design of AMP-based therapies requires simultaneous tackling of multiple factors that include toxicity, stability, and bacterial resistance.[49] For this purpose, it is critical to identify the most important characteristics that make a peptide effective against different species of bacteria. This study, to the best of our knowledge, provides a *bacteria-specific* feature selection that can be utilized in the rational design of antimicrobial peptides targeting specific bacteria. However, the major pitfall of the current study is insufficient data. It will be important to test the predictions of this study with larger data sets, combining across databases,[50] and more advanced machine-learning and deep-learning techniques. Specifically, future studies could explore predicting other important

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733

features of AMPs like cytotoxicity[51] based on the feature selection, extending the current approach with ensemble methods to combine models for better prediction,[52] and developing models to predict specific MIC values[50] using the regression-based feature selection.

■ ASSOCIATED CONTENT

Data Availability Statement

The data (machine learning predictions) obtained in this work and the in-house scripts are available on figshare at the following URL: https://figshare.com/articles/software/A_bacteria-specific_machine_learning_study_of_individual_antimicrobial_peptide_activity/22129547.

■ AUTHOR INFORMATION

Corresponding Author

Anatoly B. Kolomeisky − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Department of Chemical and Biomolecular Engineering, and Department of Physics and Astronomy, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-5677-6690; Email: tolya@rice.edu

Authors

Hamid Teimouri − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Angela Medvedeva − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01551

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work was supported by the Welch Foundation (C-1559), by the NIH (R01 HL157714-02), by the NSF (CHE-1953453 and MCB-1941106), and by the Center for Theoretical Biological Physics sponsored by the NSF (PHY-2019745).

■ REFERENCES

(1) Brogden, K. A. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature reviews microbiology* **2005**, *3*, 238−250.

(2) Shai, Y. From innate immunity to de-novo designed antimicrobial peptides. *Current pharmaceutical design* **2002**, *8*, 715−725.

(3) Zasloff, M. Antimicrobial peptides of multicellular organisms. *nature* **2002**, *415*, 389−395.

(4) Otvos Jr, L. Antibacterial peptides and proteins with multiple cellular targets. *Journal of peptide science: an official publication of the European Peptide Society* **2005**, *11*, 697−706.

(5) Nguyen, T. N.; Teimouri, H.; Medvedeva, A.; Kolomeisky, A. B. Cooperativity in Bacterial Membrane Association Controls the Synergistic Activities of Antimicrobial Peptides. *J. Phys. Chem. B* **2022**, *126*, 7365−7372.

(6) Papo, N.; Shai, Y. Can we predict biological activity of antimicrobial peptides from their interactions with model phospholipid membranes? *Peptides* **2003**, *24*, 1693−1703.

(7) Teixeira, V.; Feio, M. J.; Bastos, M. Role of lipids in the interaction of antimicrobial peptides with membranes. *Progress in lipid research* **2012**, *51*, 149−177.

(8) Lee, E. Y.; Fulan, B. M.; Wong, G. C.; Ferguson, A. L. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 13588−13593.

(9) Lee, E. Y.; Lee, M. W.; Fulan, B. M.; Ferguson, A. L.; Wong, G. C. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface focus* **2017**, *7*, 20160153.

(10) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic acids research* **2016**, *44*, D1094−D1097.

(11) Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895.

(12) Agrawal, P.; Bhalla, S.; Chaudhary, K.; Kumar, R.; Sharma, M.; Raghava, G. P. In silico approach for prediction of antifungal peptides. *Frontiers in microbiology* **2018**, *9*, 323.

(13) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic acids research* **2016**, *44*, D1087−D1093.

(14) Aronica, P. G.; Reid, L. M.; Desai, N.; Li, J.; Fox, S. J.; Yadahalli, S.; Essex, J. W.; Verma, C. S. Computational methods and tools in antimicrobial peptide research. *J. Chem. Inf. Model.* **2021**, *61*, 3172−3196.

(15) Moretta, A.; Salvia, R.; Scieuzo, C.; Di Somma, A.; Vogel, H.; Pucci, P.; Sgambato, A.; Wolff, M.; Falabella, P. A bioinformatic study of antimicrobial peptides identified in the Black Soldier Fly (BSF) Hermetia illucens (Diptera: Stratiomyidae). *Sci. Rep.* **2020**, *10*, 16875.

(16) Söylemez, Ü. G.; Yousef, M.; Kesmen, Z.; Büyükkiraz, M. E.; Bakir-Gungor, B. Prediction of Linear Cationic Antimicrobial Peptides Active against Gram-Negative and Gram-Positive Bacteria Based on Machine Learning Models. *Applied Sciences* **2022**, *12*, 3631.

(17) Oliveira, J.; Reygaert, W. C. *Gram Negative Bacteria*; StatPearls Publishing LLC, Treasure Island, FL, USA, 2019.

(18) Strandberg, E.; Zerweck, J.; Horn, D.; Pritz, G.; Berditsch, M.; Bürck, J.; Wadhwani, P.; Ulrich, A. S. Influence of hydrophobic residues on the activity of the antimicrobial peptide magainin 2 and its synergy with PGLa. *Journal of Peptide Science* **2015**, *21*, 436−445.

(19) Liu, Z.; Brady, A.; Young, A.; Rasimick, B.; Chen, K.; Zhou, C.; Kallenbach, N. R. Length effects in antimicrobial peptides of the (RW) n series. *Antimicrob. Agents Chemother.* **2007**, *51*, 597−603.

(20) Joo, H.-S.; Fu, C.-I.; Otto, M. Bacterial strategies of resistance to antimicrobial peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2016**, *371*, 20150292.

(21) Lee, E. Y.; Wong, G. C.; Ferguson, A. L. Machine learning-enabled discovery and design of membrane-active peptides. *Bioorganic & medicinal chemistry* **2018**, *26*, 2708−2718.

(22) Shrivastava, S. R.; Shrivastava, P. S.; Ramasamy, J. World health organization releases global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *Journal of Medical Society* **2018**, *32*, 76.

(23) Blount, Z. D. The natural history of model organisms: The unexhausted potential of E. coli. *Elife* **2015**, *4*, e05826.

(24) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research* **2021**, *49*, D288−D297.

(25) Wiegand, I.; Hilpert, K.; Hancock, R. E. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **2008**, *3*, 163−175.

(26) Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960−962.

(27) Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273−297.
(28) Sutton, C. D. Classification and regression trees, bagging, and boosting. *Handbook of statistics* **2005**, *24*, 303−329.
(29) Schapire, R. E. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification* **2003**, *171*, 149−171.
(30) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Springer, New York, 2013; Vol. *112*.
(31) Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267−288.
(32) Mehta, P.; Bukov, M.; Wang, C.-H.; Day, A. G.; Richardson, C.; Fisher, C. K.; Schwab, D. J. A high-bias, low-variance introduction to machine learning for physicists. *Physics reports* **2019**, *810*, 1−124.
(33) Deng, N.; Tian, Y.; Zhang, C. *Support vector machines: optimization based theory, algorithms, and extensions*; CRC press, Boca Raton, 2012.
(34) Kaur, H.; Pannu, H. S.; Malhi, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)* **2020**, *52*, 79.
(35) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* **2020**, *21*, 6.
(36) Zhu, Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters* **2020**, *136*, 71−80.
(37) Chen, Y.; Guarnieri, M. T.; Vasil, A. I.; Vasil, M. L.; Mant, C. T.; Hodges, R. S. Role of peptide hydrophobicity in the mechanism of action of α-helical antimicrobial peptides. *Antimicrob. Agents Chemother.* **2007**, *51*, 1398−1406.
(38) Chou, K.-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics* **2009**, *6*, 262−274.
(39) Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10−19.
(40) Li, Z.-R.; Lin, H. H.; Han, L.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic acids research* **2006**, *34*, W32−W37.
(41) Pormohammad, A.; Nasiri, M. J.; Azimi, T. Prevalence of antibiotic resistance in Escherichia coli strains simultaneously isolated from humans, animals, food, and the environment: a systematic review and meta-analysis. *Infection and drug resistance* **2019**, *12*, 1181−1197.
(42) Jahangiri, A.; Neshani, A.; Mirhosseini, S. A.; Ghazvini, K.; Zare, H.; Sedighian, H. Synergistic effect of two antimicrobial peptides, Nisin and P10 with conventional antibiotics against extensively drug-resistant Acinetobacter baumannii and colistin-resistant Pseudomonas aeruginosa isolates. *Microbial Pathogenesis* **2021**, *150*, 104700.
(43) Musa, A. B. Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics* **2013**, *4*, 13−24.
(44) Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M.; Managadze, G.; Grigolava, M.; Makhatadze, G. I.; Pirtskhalava, M. Predictive model of linear antimicrobial peptides active against gram-negative bacteria. *J. Chem. Inf. Model.* **2018**, *58*, 1141−1151.
(45) Khatun, S.; Hasan, M.; Kurata, H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. *FEBS letters* **2019**, *593*, 3029−3039.
(46) Hasan, M. M.; Basith, S.; Khatun, M. S.; Lee, G.; Manavalan, B.; Kurata, H. Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Briefings in Bioinformatics* **2021**, *22*, bbaa202.
(47) Christin, C.; Hoefsloot, H. C.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics* **2013**, *12*, 263−276.
(48) Giacomucci, S.; Cros, C. D.-N.; Perron, X.; Mathieu-Denoncourt, A.; Duperthuy, M. Flagella-dependent inhibition of biofilm formation by sub-inhibitory concentration of polymyxin B in Vibrio cholerae. *PLoS One* **2019**, *14*, e0221431.
(49) Tornesello, A. L.; Borrelli, A.; Buonaguro, L.; Buonaguro, F. M.; Tornesello, M. L. Antimicrobial peptides as anticancer agents: Functional properties and biological activities. *Molecules* **2020**, *25*, 2850.
(50) Witten, J.; Witten, Z. Deep learning regression model for antimicrobial peptide design. *BioRxiv* **2019**, 692681.
(51) Salem, M.; Keshavarzi Arshadi, A.; Yuan, J. S. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. *BMC bioinformatics* **2022**, *23*, 389.
(52) Lv, H.; Yan, K.; Guo, Y.; Zou, Q.; Hesham, A. E.-L.; Liu, B. AMPpred-EL: An effective antimicrobial peptide prediction model based on ensemble learning. *Computers in Biology and Medicine* **2022**, *146*, 105577.

https://doi.org/10.1021/acs.jcim.2c01551
J. Chem. Inf. Model. 2023, 63, 1723−1733