ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

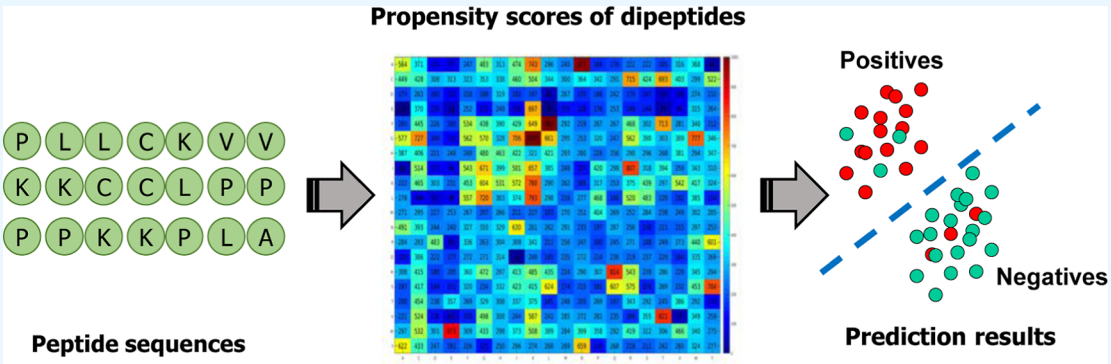
SCMRSA: a New Approach for Identifying and Analyzing Anti-MRSA Peptides Using Estimated Propensity Scores of Dipeptides

Phasit Charoenkwan, Sakawrat Kanthawong, Nalini Schaduangrat, Pietro Li', Mohammad Ali Moni, and Watshara Shoombuatong*

Cite This: *ACS Omega* 2022, 7, 32653−32664

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: *Staphylococcus aureus* is deemed to be one of the major causes of hospital and community-acquired infections, especially in methicillin-resistant *S. aureus* (MRSA) strains. Because antimicrobial peptides have captured attention as novel drug candidates due to their rapid and broad-spectrum antimicrobial activity, anti-MRSA peptides have emerged as potential therapeutics for the treatment of bacterial infections. Although experimental approaches can precisely identify anti-MRSA peptides, they are usually cost-ineffective and labor-intensive. Therefore, computational approaches that are able to identify and characterize anti-MRSA peptides by using sequence information are highly desirable. In this study, we present the first computational approach (termed SCMRSA) for identifying and characterizing anti-MRSA peptides by using sequence information without the use of 3D structural information. In SCMRSA, we employed an interpretable scoring card method (SCM) coupled with the estimated propensity scores of 400 dipeptides. Comparative experiments indicated that SCMRSA was more effective and could outperform several machine learning-based classifiers with an accuracy of 0.960 and Matthews correlation coefficient of 0.848 on the independent test data set. In addition, we employed the SCMRSA-derived propensity scores to provide a more in-depth explanation regarding the functional mechanisms of anti-MRSA peptides. Finally, in order to serve community-wide use of the proposed SCMRSA, we established a user-friendly webserver which can be accessed online at http://pmlabstack.pythonanywhere.com/SCMRSA. SCMRSA is anticipated to be an open-source and useful tool for screening and identifying novel anti-MRSA peptides for follow-up experimental studies.

INTRODUCTION

Staphylococcus aureus is a Gram-positive pathogen forming grape-like clusters of cocci that can also be found on multiple parts of healthy people, such as the skin and nose.[1] It causes a variety of infections, from skin infections, abscesses, impetigo, cellulitis, and folliculitis to life-threatening diseases, such as pneumonia, endocarditis, toxic shock syndrome, and sepsis.[2] Moreover, *S. aureus* is considered to be one of the major causes of hospital and community-acquired infections, especially methicillin-resistant *S. aureus* (MRSA) strains.[3] MRSA was initially isolated from patients hospitalized in the 1960s and rapidly became an important problem in the community and healthcare system.[4] The data from the National Antimicrobial Resistance Surveillance center (NARST) of Thailand showed that, MRSA had been found in around 50% of the total *S. aureus* isolates from clinical specimens in some tertiary hospitals of Thailand.[5] In addition, MRSA accounted for 59.5% of nosocomial *S. aureus* infections in intensive care units (ICUs) in the United States.[6] Even with the current development of new antimicrobial agents, MRSA remains a difficult-to-treat superbug with persistently high mortality.

Received: July 8, 2022
Accepted: August 22, 2022
Published: September 1, 2022

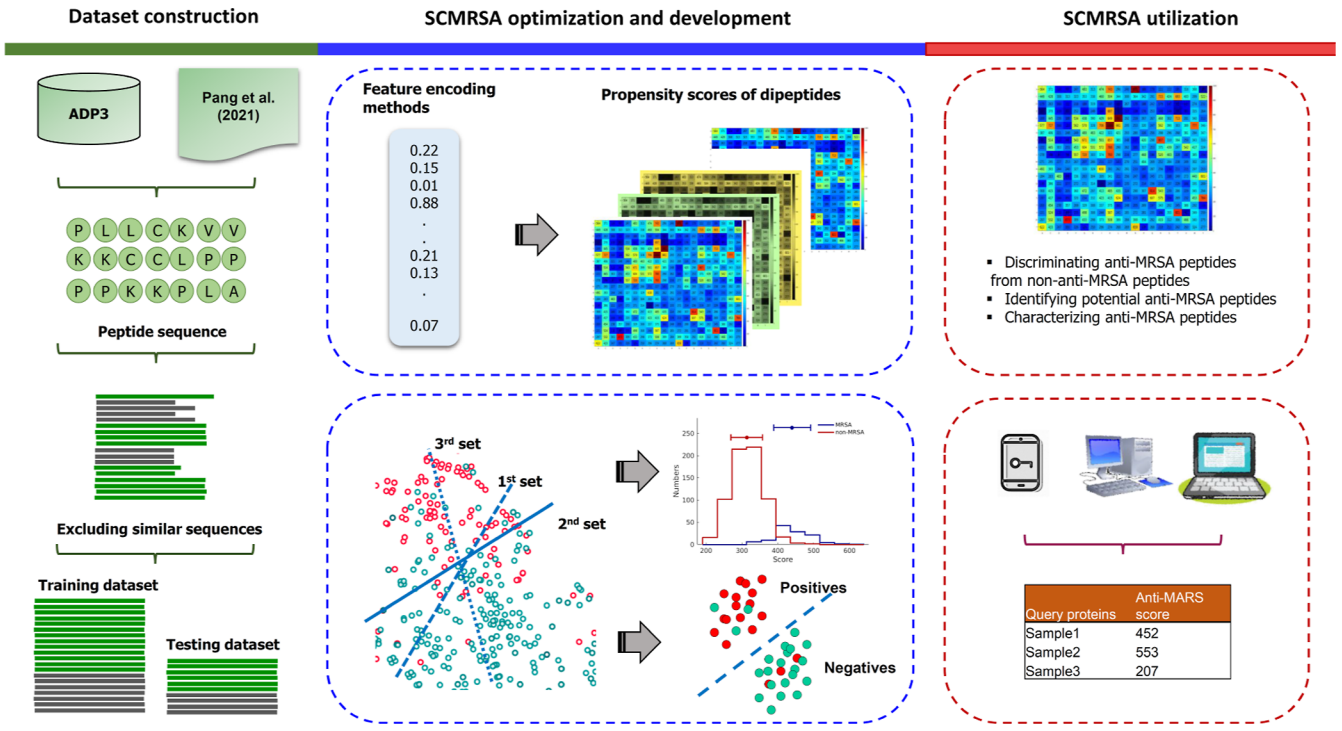


Figure 1. Schematic framework of the development of SCMRSA. The schematic framework of the development of SCMRSA contains four main steps: (i) training and independent data set preparation, (ii) SCMRSA-based propensity score generation and optimization, (iii) anti-MRSA peptide identification and characterization, and (iv) SCMRSA webserver construction.

Therefore, MRSA has emerged as one of the greatest challenges for healthcare professionals globally.

Due to the ongoing resistance toward conventional antibiotics of various pathogens, antimicrobial peptides (AMPs) have captured attention as novel drug candidates with rapidly and broad-spectrum antimicrobial activity, including antibiotic-resistant microorganisms, such as MRSA. In general, AMPs are small cationic and amphipathic molecules[7] ranging between 12 and 50 amino acids in length.[8] Owing to their intrinsic physicochemical properties, such as amino acid sequence, charge, amphipathic property, and secondary structure, AMPs adopts various mechanisms to kill bacteria.[7] Disrupting the integrity of the cell membrane has been reported to be the main mechanism of action of AMPs. However, the broad-spectrum activity of AMPs may result in the coincidental selection and proliferation of resistant bacterial strains, creating pathogens from previously harmless organisms and increasing the possibility of an ecological imbalance in the microbiota.[9] Therefore, species-specific peptides are of tremendous value. Furthermore, peptides with potent activity against MRSA have been discovered either naturally or designed synthetically.[10] Nevertheless, the conventional methods for effectively screening and designing novel AMPs, such as antimicrobial susceptibility testing (AST), are time-consuming and expensive. Therefore, sequence-based computational tools that can rapidly and accurately identify potential anti-MRSA peptides based on sequence information could serve a great purpose in their large-scale identification.

Until now, there is no computational approach in existence that has been proposed for identifying anti-MRSA peptides by using sequence information without the use of 3D structural information. With such potential of anti-MRSA peptides for the treatment of bacterial infections, in this study, we develop SCMRSA, a sequence-based computational approach for identifying and analyzing anti-MRSA peptides. In brief, the construction and development of SCMRSA involves the following steps (as summarized in Figure 1): (i) we established a benchmark data set by collecting positive samples from the AMP database version 3 (ADP3)[11] and negative samples from the article of Pang et al.;[12] (ii) we employed an interpretable scoring card method (SCM) to develop the prediction model (SCMRSA); and (iii) SCMRSA-derived propensities of 20 amino acids and 400 dipeptides were generated and optimized using the genetic algorithm (GA) based on the 10-fold cross-validation scheme. The predictive performance on the independent test data set show that SCMRSA can outperform several machine learning (ML)-based classifiers, including decision tree (DT), k-nearest neighbor (KNN), logistic regression (LR), naive Bayes (NB), and partial least squares regression (PLS) classifiers, as judged by predictive ability, cost-effectiveness, and interpretability. In addition, the SCMRSA-derived propensity scores were employed to provide insights into the functional mechanisms of anti-MRSA peptides. Finally, in order to serve a community-wide use of the proposed SCMRSA, we implemented an online webserver, which is available at http://pmlabstack.pythonanywhere.com/SCMRSA.

2. MATERIALS AND METHODS

2.1. Data Collection and Curation. In this study, we established a new data set, including 183 experimentally validated anti-MRSA peptides (positive samples) from ADP3.[11] For the negative samples, we employed 4979 peptides without anti-MRSA activity (called non-anti-MRSA peptides) from an article of Pang et al.[12] Next, a CD-HIT [40] threshold of 0.8 was applied to remove sequence redundancy in both the anti-MRSA and non-anti-MRSA samples and to avoid overestimation of the predictive performance. As a result, we

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

obtained the final non-redundant data set containing 148 positive and 847 negative samples. After obtaining the non-redundant data sets, we used 80% of the samples (118 positives and 678 negatives) as the training data set. As a result, the independent test data set contained 30 positives and 169 negatives.

2.2. Overview Framework of SCMRSA. Previously, Huang et al.[13] and Charoenkwan et al.[14,15] introduced a simple and interpretable SCM method. To date, this method has successfully overcome the limitations of existing computational black-box approaches, such as SVM[16−18] and deep learning (DL)[19] approaches, by generating the propensity scores of amino acids and dipeptides in order to provide information about the global property of general proteins and peptides. Herein, we utilized the SCM method[13] to build an interpretable model for predicting and characterizing anti-MRSA peptides. The design and development of SCMRSA based on the SCM method are summarized in Figure 1, including initial (initial-DPS) and optimal (DPS) propensity score generation, a scoring function construction, prediction of an uncharacterized peptide, and performance evaluation.

Phase 1: generating a matrix initial-DPS, as follows:

Step 1: calculating the numbers of 400 dipeptides (i = 1,2,3, ..., 20) in positive and negative classes to construct matrices (P_{ij} = ($n_{ij}|C = 1$)) and (N_{ij} = ($n_{ij}|C = 0$)), respectively, where 1 and 0 denote positive and negative classes, respectively.

Step 2: computing compositions of nP_i and nN_i by dividing them with the number of occurrences of dipeptides in each class, as follows

$$nP_i = \left(\frac{n_i}{L_{p-1}} \middle| C = 1 \right) \quad (1)$$

$$nN_i = \left(\frac{n_i}{L_{n-1}} \middle| C = 0 \right) \quad (2)$$

where L_p and L_n denote total numbers of dipeptides in positive and negative classes, respectively.

Step 3: computing the score of the ith dipeptide by subtracting nP_i from nN_i.

$$\text{init} - \text{DPS}_{ij} = nP_{ij} - nN_{ij} \quad (3)$$

Step 4: the scores of init-DPS$_{ij}$ are normalized into the range of 0−1000 for the convenience of the prediction and characterization of anti-MRSA peptides.

Phase 2: optimizing initial-DPS by using the GA algorithm in order to enhance the discriminative ability and conserve the information of anti-MRSA peptides. The fitness function used in this study is represented by

$$\text{FF(DPS)} = 0.9 \times \text{AUC} + 0.1 \times R \quad (4)$$

As can be seen, this fitness function includes AUC and R, which represent area under the receiver operating characteristics (ROC) curve (AUC) and Pearson's correlation coefficient (R value) between initial-DPS and DPS, respectively. The 10-fold cross-validation scheme was performed to control overfitting and biasness.

Phase 3: building a scoring-based predictor $S(P)$. The $S(P)$ is defined as follows

$$S(P) = \sum_{i=1}^{400} S_i \text{DPS}_i \quad (5)$$

where S_i and DPS$_i$ define the occurrence number and the ith dipeptide propensity score.

Phase 4: discriminating uncharacterized peptide P. In the meantime, the optimal threshold value was determined by maximizing the performance in terms of accuracy (ACC) based on the training data set. Particularly, P is predicted as the positive class if the score from $S(P)$ is > threshold value, otherwise P is predicted as the negative class. The propensity scores of the 20 amino acids can be generated in the same procedure. Herein, the optimized propensity scores of amino acids and dipeptides are denoted as APS and DPS, respectively, for the convenience of discussion. More detailed information for the SCM classifier construction is provided in our previous studies.[13−15,17,20]

2.3. Identification of Informative Physicochemical Properties. In this study, the SCM approach coupled with 20 amino acid propensities (or APS) were utilized for determining informative physicochemical properties (PCPs) from the amino acid index database (AAindex) so as to analyze and characterize anti-MRSA peptides. The process of identifying informative PCP approach contained multiple main steps, as follows. First, we excluded all PCPs having not applicable (NA) and then obtained 531 PCPs in this study. Each PCP is defined by a set of 20 numerical values for 20 amino acids. Second, we calculated the R value between each of the 531 PCPs and APS. Please note that PCPs affording the largest R values are deemed to be the most important properties in anti-MRSA peptides. Third, the 20 top-ranked PCPs with the largest R values were reported and used for further analysis.

2.4. ML-Based and Blast-Based Classifiers. Here, we compared the performance of SCMRSA with several ML-based and Blast-based classifiers. Specifically, we employed seven popular ML algorithms [i.e., DT, KNN, LR, NB, PLS, and SVM with linear kernel and radial basis function kernel (referred herein as SVMLN and SVMRBF, respectively)] and five conventional feature descriptors[21] [i.e., AAC, AAI, DPC, PCP, and composition−transition−distribution (CTD)] to generate a total of 35 ML classifiers. All the ML classifiers were created using Scikit-learn v0.22.0 package[22] (Supporting Information, Table S1). For the BLAST-based classifier, the training data set was considered as the BLASTP database, while the independent data set was employed as the BLASTP query sequences. Detailed information for the construction of ML-based[23−26] and Blast-based[20,27] classifiers are available in our previous studies.

2.5. Performance Evaluation. We employed six common performance measures to examine the predictive ability and effectiveness of the proposed model, including ACC, AUC, sensitivity (Sn), specificity (Sp), balanced accuracy (BACC), and Matthew's correlation coefficient (MCC).[28,29] These performance measures are described by the following equations

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (6)$$

$$\text{Sn} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (7)$$

$$\text{Sp} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (8)$$

$$\text{BACC} = (\text{Sn} + \text{Sp}) \times 0.5 \quad (9)$$

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

Table 1. Performance Comparison for the Optimal Sets of APS and DPS

cross-validation	feature	ACC	BACC	Sn	Sp	MCC	AUC
10-fold CV	APS	0.950	0.883	0.788	0.978	0.797	0.940
	DPS	0.970	0.927	0.865	0.988	0.880	0.969
independent test	APS	0.960	0.894	0.800	0.988	0.837	0.940
	DPS	0.960	0.935	0.900	0.970	0.848	0.986

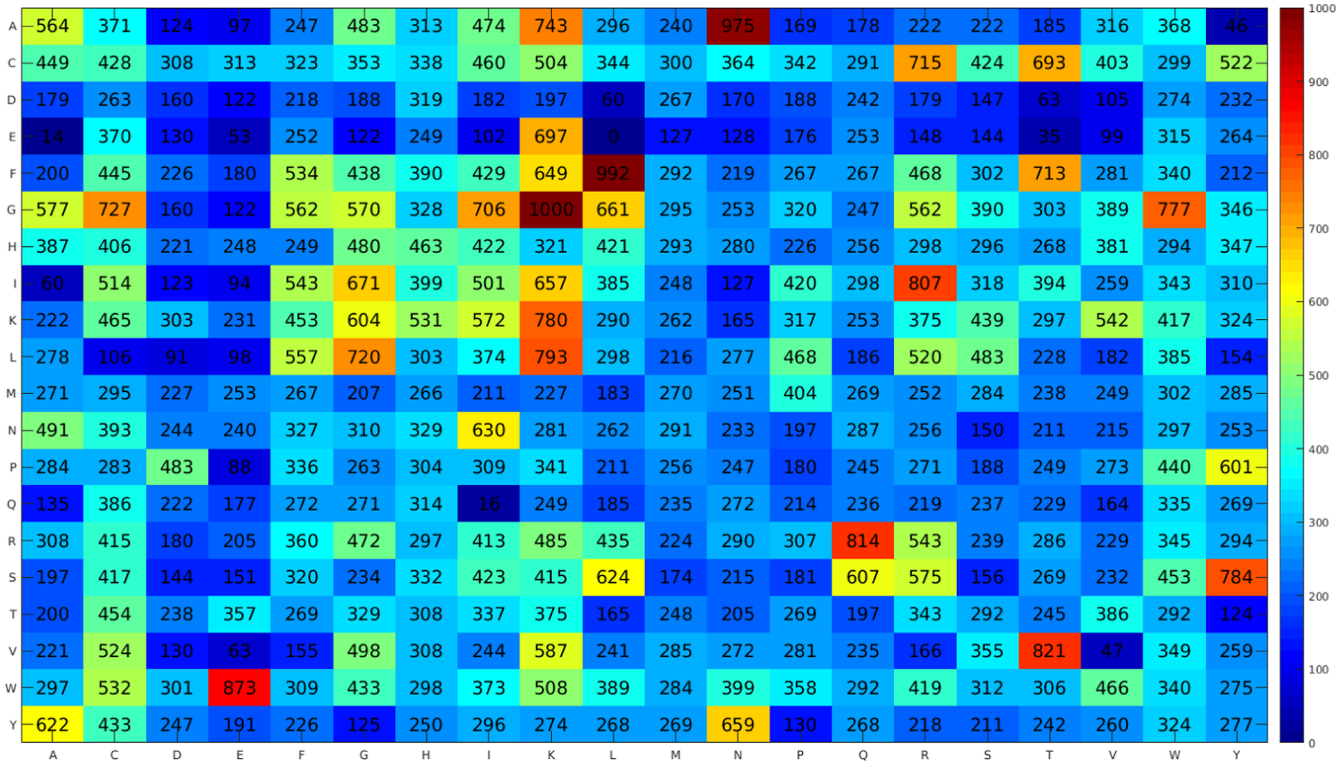


Figure 2. Propensity scores of 400 dipeptides to be anti-MRSA peptides obtained from the proposed SCMRSA.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (10)$$

where TP and TN indicate the number of true positives and true negatives, respectively, while FP and FN represent the number of false positives and false negatives, respectively.[14,20,30−32]

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation of Different Sets of Propensity Scores. In this section, we evaluated the impact of different sets of APS and DPS in anti-MRSA peptide identification. As mentioned in the Overview Framework of SCMRSA section, the GA algorithm was employed for optimization of APS and DPS based on the training data set. Because of the characteristics of the GA algorithm, we generated 10 different sets for each type of propensity scores (10 APS and 10 DPS). Each set of propensity scores was used to individually build the SCM classifier and its performance was evaluated using the 10-fold cross-validation tests. Supporting Information, tables S2 and S3 show the 10-fold cross-validation results of the different SCM classifiers trained using different sets of APS and DPS.

Supporting Information, Table S2 shows that the APS from the 4th experiment achieves the highest MCC of 0.797 with BACC of 0.883 and AUC of 0.940. On the other hand, the second best and third best sets of APS were obtained from the first (MCC of 0.791) and ninth (MCC of 0.790) experiments. In the case of DPS, the sixth experiment provided the highest MCC of 0.880 with a BACC of 0.927 and AUC of 0.969, while the DPS from the third and seventh experiments yielded the second and third highest MCC of 0.871 and 0.865, respectively (Supporting Information, Table S3). As a result, the APS from the fourth experiment and the DPS from the sixth experiment were regarded as the optimal sets of APS and DPS for identifying anti-MRSA peptides, respectively (Table 1). From Table 1, we notice that the optimal set of DPS exhibits the best of all five performance metrics. Remarkably, BACC, Sn, and MCC of the optimal set of DPS were 4.4, 7.7, and 8.3% higher than the optimal set of APS, respectively. For the performance on the independent test data set (Supporting Information, Tables S4 and S5), the optimal set of DPS still performed higher than the optimal set of APS in terms BACC, Sn, MCC, and AUC. Thus, our comparative results confirmed that the optimal set of DPS could effectively be used for identifying anti-MRSA peptides. In addition, the propensities of 400 dipeptides are provided in Figure 2.

As mentioned in the Scoring card method section, the improved performance of the SCM classifier is mainly due to

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

Table 2. Performance Comparison of Initial-DPS and DPS

cross-validation	feature	ACC	BACC	Sn	Sp	MCC	AUC
10-fold CV	initial-DPS	0.925	0.720	0.829	0.960	0.697	0.957
	DPS	0.970	0.927	0.865	0.988	0.880	0.969
independent test	initial-DPS	0.935	0.833	0.854	0.953	0.756	0.976
	DPS	0.960	0.935	0.900	0.970	0.848	0.986

the optimal set of DPS (called DPS for short), derived from the GA algorithm. Thus, to explain this evidence, the performance of DPS was compared with initial-DPS. Table 2 lists the performance of the DPS and initial-DPS on the training and independent test data set. The results indicated that the DPS significantly outperformed initial-DPS on both the training and independent data sets. To be specific, the ACC, BACC, Sn, and MCC of the DPS were 2.51, 10.19, 4.56, and 9.16% higher than the initial-DPS, respectively, on the independent test data set. In addition, Figure 3 shows that the initial-DPS (Figure 3A) has a large overlapped region between the positives and negatives as compared to DPS (Figure 3B) as analyzed on the training data set. Altogether, our analysis results confirmed that the estimated propensities of 400 dipeptides (DPS) provide more discriminative power for identifying anti-MRSA peptides than initial-DPS.

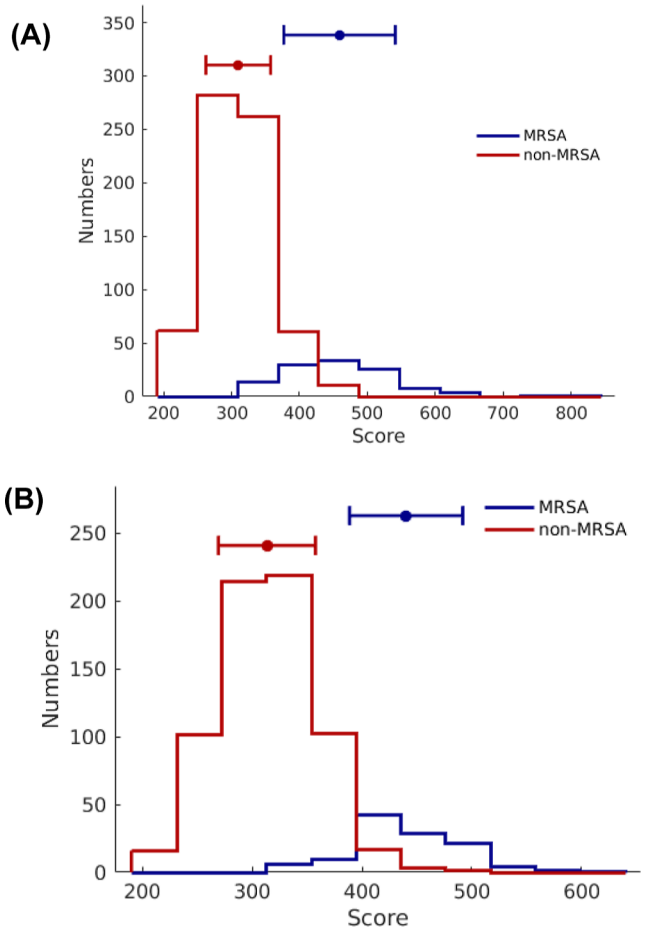


Figure 3. Histogram plot represents the scores of positives (anti-MRSA peptides) and negatives (non-anti-MRSA peptides) derived from SCMRSA by using initial-DPS (A) and DPS (B) on the training data set, where the mean and standard deviation are indicated by the bars and closed circles.

3.2. Comparison of SCMRSA with BLAST-Based Predictor and Conventional ML-Based Classifiers. To the best of the authors' knowledge, SCMRSA is the first computational model for identifying peptides with or without anti-MRSA activity. Thus, we evaluated and compared the performance of SCMRSA against the BLAST-based predictors and conventional ML-based classifiers trained with seven ML algorithms and five sequence-based feature descriptors (as described in the ML-Based and Blast-Based Classifiers section). Supporting Information, Tables S6−S8 show the performance of the BLAST-based predictor and 35 ML classifiers. In addition, we compared the performance of the proposed SCMRSA with five best-performing ML classifiers having the highest cross-validation BACC for convenience of discussion. Figures 4 and 5 and Supporting Information, Table S9 present the performance comparison of SCMRSA with the five best-performing ML classifiers, including SVMRBF-AAI, SVMRBF-AAC, SVMRBF-CTD, SVMLN-AAI, and LR-CTD.

We first compared the performance of SCMRSA with the BLAST-based predictor. Supporting Information, Table S8 shows that the BLAST-based predictor using an *E*-value cut off of 0.1 provides the highest BACC of 0.869 with an ACC of 0.940 and MCC of 0.758. However, the MCC, BACC, and Sn of SCMRSA were 8.93, 6.67, and 13.33%, respectively, higher than the BLAST-based predictor. When compared with the five best-performing ML classifiers on the training data set, SCMRSA still attained the best overall performance in terms of ACC, BACC, Sn, and MCC (Figures 4 and 5). In the meantime, SVMRBF-AAI yielded the second highest BACC of 0.899. As can be seen, the BACC, Sn, and MCC of SCMRSA were 2.76, 5.08, and 4.66%, respectively, higher than the second-best method SVMRBF-AAI. From Figures 4 and 5, the largest BACC of 0.947 is achieved by SVMRBF-AAC, while SVMRBF-AAI and SCMRSA yield the second and third largest BACC of 0.941 and 0.935, respectively.

It is well-known that SVMRBF-AAC and SVMRBF-AAI are defined as computational black-box approaches because the SVM algorithm cannot directly provide information of the relationship between each feature and the model output. As such, we were motivated to utilize SCM methods for developing a simple and highly interpretable model SCMRSA. The contribution of SCMRSA can be summarized in the following three aspects: (i) SCMRSA is able to discriminate between anti-MRSA and non-anti-MRSA using only the simple weighted-sum function, highlighting that SCMRSA will be a powerful tool for large-scale identification of anti-MRSA; (ii) SCMRSA is able to create the propensities of amino acids and dipeptides. These propensity scores could provide a good understanding of functional mechanisms of anti-MRSA peptides and (iii) SCMRSA attained a competitive performance as compared with SVM-based classifiers and also outperformed several ML-based classifiers, including DT, KNN, LR, NB, and PLS, in terms of cost-effectiveness and interpretability.

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

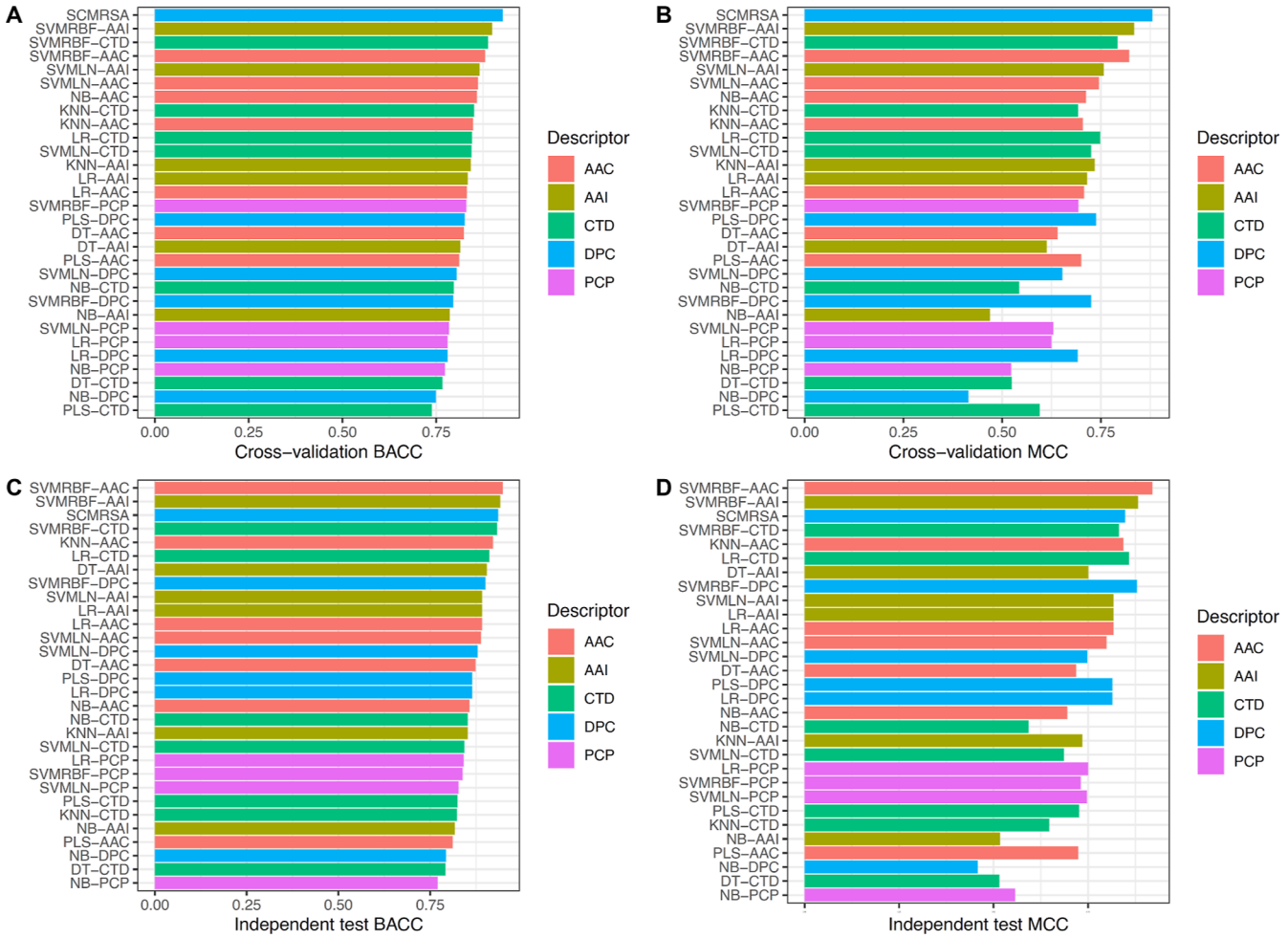


Figure 4. Performance evaluations of SCMRSA and conventional ML-based classifiers based on 10-fold cross-validation test (A,B) and independent test (C,D).

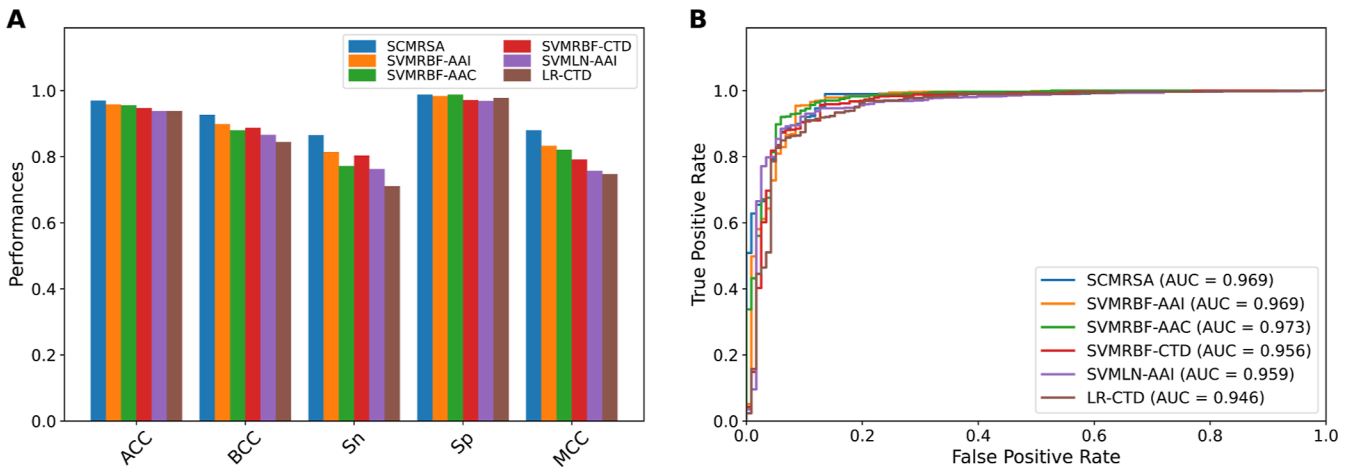


Figure 5. Performance evaluations of SCMRSA and five best-performing ML classifiers as indicated by the 10-fold cross-validation test. Prediction results of SCMRSA and five best-performing ML classifiers in terms of ACC, BACC, Sn, Sp, and MCC (A). ROC curves and AUC values of SCMRSA and five best-performing ML classifiers (B).

3.3. Analysis of Potential Anti-MRSA Peptides on the Training Data Set. In this section, we utilized our proposed approach SCMRSA to determine potential anti-MRSA peptides on the training data set. Particularly, for a given peptide sequence *P*, SCMRSA will provide us its anti-MRSA score. Note that peptides with the largest anti-MRSA score were assumed to be potential anti-MRSA peptides. Tables 3 and 4 list the top 20 potential anti-MRSA peptides having the highest and lowest anti-MRSA scores, respectively, along with their important PCPs, including hydrophobicity (PCP1), hydropathicity (PCP2), charge (PCP3), and pI (PCP4).

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

Table 3. Top 20 High-Potential Anti-MRSA Peptides Having the Highest $S(P)$ along with Their Important Physicochemical Properties[a]

#	peptide sequence	score	PCP1	PCP2	PCP3	PCP4	references
1	FLKAIKKFGKEFKKIGAKLK	598.26	−0.2	−0.33	7	10.48	33
2	FLPAALAGIGGILGKLF	560.63	0.28	1.56	1	9.11	34
3	KRIGLIRLIGKILRGLRRLG	558.37	−0.23	0.24	7	12.61	35
4	FLPLIAGLFGKIF	529.92	0.31	1.74	1	9.11	36
5	FLSAITSILGKFF	528.92	0.21	1.44	1	9.11	37
6	FLSIIAKVLGSLF	523.00	0.26	1.86	1	9.11	38
7	KRFKKFFRKIKKGFRKIFKKTKIFIGGTIPI	520.07	−0.26	−0.4	13	12.34	39
8	GLSLLLSLGLKLL	519.42	0.21	1.72	1	9.11	
9	GWKKWLRKGAKHLGQAAIKGLAS	514.27	−0.16	−0.49	6.5	11.39	40
10	GFLGSLLKTGLKVGSNLL	509.65	0.08	0.77	2	10.02	41
11	FLPLLAGLAANFLPKIFCKITRK	508.82	0.04	0.85	4	10.33	42
12	GLLSLLSLLGKLL	506.42	0.21	1.72	1	9.11	43
13	GFWGKLFKLGLHGIGLLHLHL	503.45	0.16	0.75	3.5	10.02	44
14	GRRKRKWLRRIGKGVKIIGGAALDHL	502.60	−0.31	−0.56	8.5	12.19	44
15	FLGGLIKIVPAMICAVTKKC	502.16	0.12	1.33	3	9.42	45
16	FLGAVLKVAGKLVPAAICKISKKC	501.22	0.03	1.01	5	9.91	46
17	FLQHIIGALGHLF	498.75	0.24	1.22	1	7.26	47
18	KWKSFIKKLTKKFLHSAKKF	496.84	−0.27	−0.73	8.5	10.85	48
19	GGGCGIGGGCGPIGDCGPIGGGCGPIGGGCGPVGGW	496.31	0.17	0.43	−1	3.8	
20	FLPFIAGMAAKFLPKIFCAISKK	495.05	0.1	1.04	4	10.05	42

[a]PCP1 = hydrophobicity, PCP2 = hydropathicity, PCP3 = charge, PCP4 = pI.

Table 4. Top 20 Non-Anti-MRSA Peptides Having the Lowest $S(P)$ along with Their Important Physicochemical Properties[a]

#	peptide sequence	score	PCP1	PCP2	PCP3	PCP4
1	VNVEALQKVVDES	190.00	−0.12	0.01	−2	4.14
2	RAYREDELIQLL	198.55	−0.28	−0.51	−1	4.68
3	KPLDDTLILEMA	205.64	−0.04	0.22	−2	4.03
4	PHLVIPEIEAIATQTLVEMEAEGLN	207.29	0.03	0.3	−4.5	3.98
5	DADLYTPSIHLYFNDDLTEL	216.16	−0.07	−0.3	−4.5	3.71
6	ADDLLNERYEAVG	218.33	−0.2	−0.59	−3	3.92
7	LEMNVNQLSKETSELKALAVELVEENVALQ	219.10	−0.12	−0.05	−4	4.21
8	TRLQFQALDSTQFATAQGEVPELVLVNPPRR	220.07	−0.18	−0.34	0	6.53
9	FDTVVGRDTLIEPNVV	222.87	−0.04	0.35	−2	4.03
10	EDEDEEEILDHEMREIVHIQAGQCGN	223.40	−0.26	−1.09	−8	4
11	ELVRSKNPDMDE	223.91	−0.4	−1.42	−2	4.32
12	VPGAEGQYFAYIAYDLDLFEPGSI	225.09	0.07	0.21	−4	3.44
13	QDIELCPECFSAG	225.75	−0.04	0.08	−3	3.58
14	CGVIDLAELVRNAHP	229.43	−0.04	0.4	−0.5	5.33
15	DREGTLFIEESDNNNVWTTTA	229.55	−0.22	−0.92	−4	3.84
16	IAEQVASFQEEK	230.09	−0.17	−0.55	−2	4.26
17	SSDPASSEMLSPSTQLLFYETSASFSTEV	230.89	−0.07	−0.15	−4	3.51
18	LQYEPEDPMSSNGDKLLVRSKF	230.90	−0.25	−0.88	−1	4.79
19	NLTLTLDKGTLHQEVNLV	232.88	−0.08	0.02	−0.5	5.33
20	YRVDSDATHSPMFHQVEGL	233.11	−0.17	−0.62	−1	5.22

[a]PCP1 = hydrophobicity, PCP2 = hydropathicity, PCP3 = charge, PCP4 = pI.

From Tables 3 and 4, several observations can be made: (i) the anti-MRSA scores of all top 20 high-potential anti-MRSA peptides were greater than 495, while the anti-MRSA scores of all top 20 non-anti-MRSA peptides were lower than 232, where the threshold value was set to 399; (ii) the anti-MRSA scores derived from our proposed approach SCMRSA, showed a high correlation with the hydrophobicity and charge properties, and (iii) 16 out of 20 potential anti-MRSA peptides having the highest anti-MRSA score exhibited high hydrophobicity and low positive charge (+1 to +5) (Table 3), while all the top 20 non-anti-MRSA peptides having lowest anti-MRSA scores (lower than 233) prefer to exhibit negative charge and low hydrophobicity (Table 4). Mishra and Wang.[43] studied the most critical parameters (e.g., composition, peptide hydrophobic content, and net charge) of potent anti-MRSA peptides using a database filtering technology (DFT). The researchers found that high hydrophobicity and low cationicity (a net charge molecule of +1) were two critical parameters for designing potential anti-MRSA peptides. Taken together, this analysis revealed that high hydrophobicity and low cationicity could be important properties in the design and development of novel anti-MRSA peptides.

Furthermore, Table 3 highlights that the top-five high-potential anti-MRSA peptides contained FLKAIKKFGKEFK-

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

Table 5. Propensity Scores of 20 Amino Acids to be Anti-MRSA Peptides (PS) along with Amino Acid Compositions of Anti-MRSA and Non-Anti-MRSA Peptides

amino acid	PS (rank)	anti-MRSA	non-anti-MRSA	difference (rank)	*p*-value
K-Lys	448(1)	0.111	0.058	0.054(1)	0.000
G-Gly	427(2)	0.111	0.071	0.040(2)	0.000
C-Cys	410(3)	0.046	0.014	0.032(3)	0.000
W-Trp	376(4)	0.023	0.011	0.012(7)	0.007
I-Ile	374(5)	0.082	0.060	0.022(6)	0.001
R-Arg	367(6)	0.062	0.056	0.006(9)	0.440
F-Phe	366(7)	0.064	0.038	0.026(5)	0.000
L-Leu	336(8)	0.126	0.096	0.030(4)	0.005
H-His	330(9)	0.029	0.021	0.007(8)	0.168
A-Ala	315(10)	0.079	0.082	−0.003(10)	0.683
S-Ser	312(11)	0.055	0.069	−0.014(13)	0.013
Y-Tyr	299(12)	0.016	0.029	−0.014(12)	0.000
T-Thr	298(13)	0.039	0.051	−0.012(11)	0.025
N-Asn	297(14)	0.021	0.040	−0.020(17)	0.000
V-Val	288(15)	0.054	0.072	−0.018(15)	0.001
P-Pro	282(16)	0.029	0.046	−0.017(14)	0.000
Q-Gln	264(17)	0.016	0.039	−0.023(18)	0.000
M-Met	257(18)	0.008	0.026	−0.018(16)	0.000
D-Asp	200(19)	0.012	0.053	−0.040(19)	0.000
E-Glu	196(20)	0.017	0.068	−0.051(20)	0.000
R	1.000	0.683	−0.072	0.953	

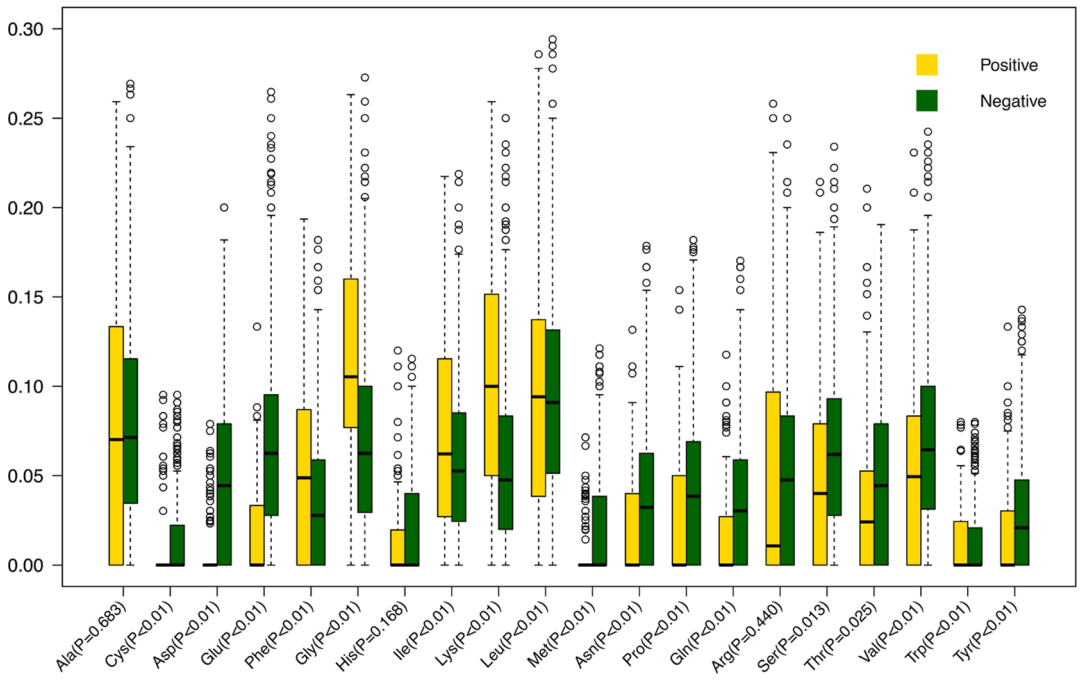


Figure 6. Boxplots of amino acid compositions of 20 amino acids for positives and negatives. *X*- and *Y*-axes represent 20 amino acids along with their *p*-value.

KIGAKLK (Omega 76 (Ω76)), FLPAALAGIGGILGKLF (temporin-SHd), KRIGLIRLIGKILRGLRRLG (Saha-CATH5), FLPLIAGLFGKIF (temporin-PF) and FLSAIT-SILGKFF (temporin-1Spa) having anti-MRSA scores higher than 528. To be specific, Ω76 was considered as the most potential anti-MRSA peptide as indicated by the anti-MRSA score. Previously, Ω76 is well recognized as a helical peptide exhibiting a 50% minimal bactericidal concentration (MBC_{50}) of 16 mg/liter for MRSA without toxicity against HeLa cells, even at the highest concentration tested (128 mg/liter). In addition, Ω76, which was designed using the Heligrapher software package, has reported antimicrobial activity against multidrug-resistant ESKAPE pathogens, namely, *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Enterobacter* spp, and *Pseudomonas aeruginosa*.[33] Interestingly, three out of the top-five high-potential anti-MRSA peptides are temporins derived from a frog-derived AMP family.[49] Several studies mentioned that temporins were able to exhibit greater potencies against Gram-positive bacteria, such as *S. aureus* as compared to Gram-negative bacteria.[35,36,50] Moreover, temporin-SHd and Tasmanian devil cathelicidin Saha-CATH5 were ranked as the second and third potential

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

anti-MRSA peptides, respectively. Temporin-SHd in the frog skin of *Pelophylax saharicus* was highly active against MRSA (ATCC 43300 and ATCC BAA-44) with a minimal inhibitory concentration (MIC) of 6.25 μM.[34] Likewise, temporin-PF identified from *Pelophylax fukienensis* displayed antimicrobial activity against MRSA (NCTC 12493), while it was ineffective against Gram-negative strains, such as *Escherichia coli* (NCTC 10418), *P. aeruginosa* (ATCC 27853), and *K. pneumoniae* (ATCC 43816), at a concentration of up to 128 μM.[36] However, Saha-CATH5 also exhibited a strong bactericidal activity against MRSA with a MIC of 32 μg/mL, while a concentration of 64 μg/mL had no effect against *P. aeruginosa* (ATCC 27853).[35]

3.4. Characterization of Anti-MRSA Peptides Using SCMRSA-Derived Propensity Scores. Here, the propensities of amino acids and dipeptides to be anti-MRSA peptides (PS) were utilized to characterize the biochemical and biophysical properties of anti-MRSA peptides. Table 5 displays the propensity scores of amino acids and the amino acid compositions (%). Amino acids having the highest propensity scores are deemed to be the most important in anti-MRSA peptides. As can be seen from Table 5, Lys, Gly, Cys, Trp, and Ile with the corresponding scores of 448, 427, 410, 376, and 374, respectively, are the top five amino acids exhibiting the largest propensities, while Glu, Asp, Met, Gln, and Pro with the corresponding scores of 196, 200, 257, 264, and 282, respectively, are the top five amino acids exhibiting the lowest propensities. The most preferred amino acid in natural AMPs that have been reported in the AMP database (APD) are Gly, Ser, and Lys.[51] Although, Gly and Lys are also dominating residues in both anti-MRSA peptides and natural AMPs. Interestingly, Ser has been found at number 11 in rank of the propensities in this study. Moreover, anti-MRSA peptides displayed a significantly difference from non-anti-MRSA peptides in terms of these 10 important amino acids at a level of $p < 0.01$ (Figure 6).

Xie et al.[50] analyzed the temporin-GHa (GHa) peptide using a combination between the template-based design coupled with the database-assisted design. Their analysis demonstrated that the GHa peptide exhibited stronger antimicrobial activity against Gram-positive bacteria, such as MRSA, as compared to Gram-negative bacteria. In addition, the results showed that Lys, Gly, Arg, and Leu were found to be abundant in AMPs, while Asp and Glu were relatively less abundant in temporins. These results are quite consistent with the SCMRSA-derived propensities of amino acids, as recorded in Table 5. As can be seen from further in Table 5, the ranks of the propensities for Lys, Gly, Arg, and Leu were 1, 2, 6, and 8, respectively, while the rank of the propensities for Asp and Glu was 19 and 20. Moreover, substitution of His with Lys at both ends of the GHa peptide to design the single-point or multi-point mutation peptides as GHaK, GHa4K, and GHa11K could enhance its antibacterial activity against MRSA (ATCC 43300) and MRSA-2 (isolate).

The same results were also found in the study of Zouhir et al.,[52] where they showed that 118 peptides with high activity against MRSA at low MIC were determined using their physicochemical data. For >80% of all anti-MRSA peptides, they consisted of Lys, Gly, Ile, and Leu, whereas Met and Asp were rarely found in Zouhir et al.'s study. Moreover, newly designed peptide S2, constructed by modifying the auto-inducing peptide of *S. aureus* and adding a disulfide bond (Cys1−Cys6), was able to improve the antimicrobial selectivity against *S. aureus*, both in vitro and in vivo.[53] This study mentioned that the disulfide bonds from Cys played an important role in the targeting activity against *S. aureus*. Although Trp is found to be less frequently deployed on average in natural AMPs, a triple Trp (WWW) motif has been reported as a critical determinant in the Trp-rich peptide, TetraF2W-RK (WWWLRKIW-amide) for their bactericidal activity against MRSA (USA300) and their disruption of bacterial biofilms.[54]

3.5. Characterization of Anti-MRSA Peptides Using Informative PCPs. Impressively, SCMRSA was able to identify informative PCPs in order to provide insights into the functional mechanisms of anti-MRSA peptides.[55−58] The 20 top-ranked informative PCPs are listed in Supporting Information, Table S10. KLEP840101, FINA910104, ZASB820101, ZIMJ680104, and ROBB760111 with the corresponding R values of 0.709, 0.644, 0.617, 0.610, and 0.557, respectively, were considered as the five top-ranked PCPs herein. Table 6 lists the selected three PCPs for analyzing anti-MRSA peptides, including KLEP840101 (R = 0.709), ZIMJ680104 (R = 0.610), and WIMW960101 (R = 0.523).

Table 6. Two Important Physicochemical Property (PCP)-Derived from SCMRSA

amino acid	PS (rank)	KLEP840101	ZIMJ680104 (rank)	WIMW960101\ (rank)
K-Lys	448(1)	1	9.74(2)	4.08(12)
G-Gly	427(2)	0	5.97(8)	4.49(6)
C-Cys	410(3)	0	5.05(18)	3.02(19)
W-Trp	376(4)	0	5.89(10)	2.23(20)
I-Ile	374(5)	0	6.02(5)	5.38(2)
R-Arg	367(6)	1	10.76(1)	4.24(8)
F-Phe	366(7)	0	5.48(16)	4.08(13)
L-Leu	336(8)	0	5.98(7)	4.52(5)
H-His	330(9)	0	7.59(3)	3.77(17)
A-Ala	315(10)	0	6(6)	4.81(4)
S-Ser	312(11)	0	5.68(12)	4.48(7)
Y-Tyr	299(12)	0	5.66(14)	3.83(15)
T-Thr	298(13)	0	5.66(13)	3.8(16)
N-Asn	297(14)	0	5.41(17)	3.67(18)
V-Val	288(15)	0	5.96(9)	3.91(14)
P-Pro	282(16)	0	6.3(4)	4.12(10)
Q-Gln	264(17)	0	5.65(15)	4.11(11)
M-Met	257(18)	0	5.74(11)	4.18(9)
D-Asp	200(19)	−1	2.77(20)	6.1(1)
E-Glu	196(20)	−1	3.22(19)	5.19(3)
R	1.000	0.709	0.610	0.523

KLEP840101, which is denoted as the "Net charge",[59] exhibited the highest positive correlation of 0.709. From Table 6, we notice that Lys and Arg are both positively charged residues, while Asp and Glu are both negatively charged residues. In addition, Lys, Arg, Asp, and Glu were ranked the 1st, 6th, 19th and 20th, respectively, important amino acids based on the propensities. The high positive R value of KLEP840101 suggested that anti-MRSA peptides favored positively charged amino acids. As previously mentioned, Zouhir et al.[52] analyzed the physicochemical data of 118 peptides derived from the APD[60] database with low MIC values against MRSA. Their results showed that the majority of anti-MRSA peptides were cationic amphipathic proteins

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

containing a high number of cationic Arg and Lys residues. Moreover, 268 AMPs with anti-G+ (Gram positive bacteria) activity in APD were analyzed using DFT. This study found that the most frequently occurring amino acids of anti-G+ peptides were Leu, Gly, and Lys, which also displayed a net charge molecule of +1 based on the positively charged amino acid, such as Lys.[43]

ZIMJ680104, which is denoted as the "Isoelectric point",[61] had a high positive correlation of 0.610. Zimmerman et al.[61] studied the side chain physical properties, such as an isoelectric point (pI) of 20 amino acids. An isoelectric point is the pH of a solution at which the net charge of a protein becomes zero. If the side chain of an amino acid is basic, the pI shows a higher pH. However, if the side chain is acidic, the pI shows a lower pH. The high positive *R* value between pI and the propensity scores of amino acids suggested that anti-MRSA peptides favored basic amino acid side chains (high pH value). There are three amino acids containing basic side chains, including Lys, Arg, and His. As could be seen in Table 6, the ranks of propensity scores (PS, pI) for Lys, Arg, and His are (1, 2), (6, 1), and (9, 3) respectively. Moreover, the pI of 19 out of the 20 high-potential anti-MRSA peptides were higher than 9. On the other hand, the pI of all 20 non-anti-MRSA peptides were lower than 6. In a previous study, an average pI derived from 118 peptides with high activity against MRSA also showed a higher pH preference (an average mean of 11.07).[52]

WIMW960101, which is described as the "interfacial values",[62] had a positive correlation of 0.523. Wimley and White.[62] determined the hydrophobicity scale for the partitioning of amino acid residues of AcWl-X-LL peptides into electrically neutral (zwitterionic) membrane interfaces from palmitoyloleoylphosphatidylcholine (POPC). The result displayed the whole-residue interfacial values (side chain plus peptide backbone) fall into three distinct classes and the aromatic amino acids (i.e., Phe, Trp, and Tyr) are highly favorable in membrane interfaces. From Table 6, it can be observed that two aromatic amino acids (Trp and Phe) are found in the 10 top-ranked amino acids with the highest propensities. To be specific, the ranks of propensities (PS and interfacial values) for Trp and Phe were (4, 1) and (7, 2), respectively. It is well recognized that all aromatic amino acids are generally hydrophobic. This result suggested that the anti-MRSA peptides favor aromatic and hydrophobic amino acids. As described earlier, Zarena et al.[54] analyzed the composition of Trp-rich peptide and TetraF2W-RK peptide with high antimicrobial activity against MRSA (USA300) and found that a Trp triplet (WWW) motif played a critical role in killing and disrupting the performance of bacterial biofilms. Furthermore, 118 peptides with low MIC values against MRSA were analyzed for their amino acid composition. It was discovered that at least 65% of all peptides contained Phe, which is an aromatic amino acid.[52]

3.6. Implementation and Utility of SCMRSA. SCMRSA is the first open-source computational tool developed for identifying and characterizing anti-MRSA peptides by employing sequence information without the use of 3D structural information. Therefore, in order to serve the scientific community, we have employed our best SCM model as an easy-to-use web server (named SCMRSA). The SCMRSA web server could be beneficial for the large-scale identification of peptides having anti-MRSA activity. It could be stated that peptides with the highest anti-MRSA score were deemed to be potential anti-MRSA peptides and could then be prioritized for experimental testing. An easy-to-use webserver of SCMRSA is freely accessible at http://pmlabstack.pythonanywhere.com/SCMRSA.

4. CONCLUSIONS

This study presents SCMRSA, an interpretable ML-based approach, which makes use of the SCM algorithm in conjunction with the propensities of amino acids and dipeptides. To the best of our knowledge, SCMRSA is the first computational tool for the identification and characterization of anti-MRSA peptides using only sequence information without the use of 3D structural information. When comparing SCMRSA with other conventional ML-based classifiers (i.e., DT, KNN, LR, NB, and PLS) on the independent test data set, SCMRSA was more effective and outperformed the compared ML-based classifiers, with an ACC of 0.960, MCC of 0.848, and AUC of 0.986. In addition, the SCMRSA-derived propensity scores were employed to provide insights into the biophysical and biochemical properties of anti-MRSA. Finally, in order to serve the community-wide use of the proposed SCMRSA, we established a user-friendly web server online at http://pmlabstack.pythonanywhere.com/SCMRSA. SCMRSA is anticipated to be an open-sourced and useful tool for facilitating the user convenience to determine potential anti-MRSA peptides for follow-up experimental validation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04305.

Performance evaluations of SCMRSA and several ML-based classifiers and 20 top-ranked informative PCPs (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Watshara Shoombuatong − *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand;* orcid.org/0000-0002-3394-8709; Email: watshara.sho@mahidol.ac.th

Authors

Phasit Charoenkwan − *Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand*

Sakawrat Kanthawong − *Department of Microbiology, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand*

Nalini Schaduangrat − *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

Pietro Li' − *Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, U.K.*

Mohammad Ali Moni − *Artificial Intelligence & Digital Health, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, Queensland 4072, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04305

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

Notes

The authors declare no competing financial interest.

The datasets used in the current study are available at http://pmlabstack.pythonanywhere.com/SCMRSA.

■ ACKNOWLEDGMENTS

This work was fully supported by College of Arts, Media, and Technology, Chiang Mai University, and partially supported by Chiang Mai University and Mahidol University. In addition, computational resources were supported by Information Technology Service Center (ITSC) of Chiang Mai University.

■ REFERENCES

(1) Gould, D.; Chamberlaine, A. Staphylococcus aureus: a review of the literature. *J. Clin. Nurs.* **1995**, *4*, 5−12.

(2) Lowy, F. D. Staphylococcus aureusInfections. *N. Engl. J. Med.* **1998**, *339*, 520−532.

(3) Gordon, R. J.; Lowy, F. D. Pathogenesis of methicillin-resistant Staphylococcus aureus infection. *Clin. Infect. Dis.* **2008**, *46*, S350−S359.

(4) Turner, N. A.; Sharma-Kuinkel, B. K.; Maskarinec, S. A.; Eichenberger, E. M.; Shah, P. P.; Carugati, M.; Holland, T. L.; Fowler, V. G. Methicillin-resistant Staphylococcus aureus: an overview of basic and clinical research. *Nat. Rev. Microbiol.* **2019**, *17*, 203−218.

(5) Jariyasethpong, T.; Tribuddharat, C.; Dejsirilert, S.; Kerdsin, A.; Tishyadhigama, P.; Rahule, S.; Sawanpanyalert, P.; Yosapol, P.; Aswapokee, N. MRSA carriage in a tertiary governmental hospital in Thailand: emphasis on prevalence and molecular epidemiology. *Eur. J. Clin. Microbiol. Infect. Dis.* **2010**, *29*, 977−985.

(6) National Nosocomial Infections Surveillance System. National Nosocomial Infections Surveillance (NNIS) system report, data summary from January 1992 through June 2004, issued October 2004. *Am. J. Infect. Control* **2004**, *32*, 470−485.

(7) Mojsoska, B.; Zuckermann, R. N.; Jenssen, H. Structure-activity relationship study of novel peptoids that mimic the structure of antimicrobial peptides. *Antimicrob. Agents Chemother.* **2015**, *59*, 4112−4120.

(8) Huang, Y.; Huang, J.; Chen, Y. Alpha-helical cationic antimicrobial peptides: relationships of structure and function. *Protein Cell* **2010**, *1*, 143−152.

(9) Qiu, X.-Q.; Wang, H.; Lu, X.-F.; Zhang, J.; Li, S.-F.; Cheng, G.; Wan, L.; Yang, L.; Zuo, J.-Y.; Zhou, Y.-Q.; Wang, H.-Y.; Cheng, X.; Zhang, S.-H.; Ou, Z.-R.; Zhong, Z.-C.; Cheng, J.-Q.; Li, Y.-P.; Wu, G. Y. An engineered multidomain bactericidal peptide as a model for targeted antibiotics against specific bacteria. *Nat. Biotechnol.* **2003**, *21*, 1480−1485.

(10) Liu, Y.; Shi, D.; Wang, J.; Chen, X.; Zhou, M.; Xi, X.; Cheng, J.; Ma, C.; Chen, T.; Shaw, C. A novel amphibian antimicrobial peptide, phylloseptin-PV1, exhibits effective anti-staphylococcal activity without inducing either hepatic or renal toxicity in mice. *Front. Microbiol.* **2020**, *11*, 565158.

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

(12) Pang, Y.; Wang, Z.; Jhong, J.-H.; Lee, T.-Y. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. *Briefings Bioinf.* **2021**, *22*, 1085−1095.

(13) Huang, H.-L.; Charoenkwan, P.; Kao, T.-F.; Lee, H.-C.; Chang, F.-L.; Huang, W.-L.; Ho, S.-J.; Shu, L.-S.; Chen, W.-L.; Ho, S.-Y. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinf.* **2012**, *13*, S3.

(14) Charoenkwan, P.; Chiangjong, W.; Lee, V. S.; Nantasenamat, C.; Hasan, M. M.; Shoombuatong, W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Sci. Rep.* **2021**, *11*, 3017.

(15) Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS One* **2013**, *8*, No. e72368.

(16) Vasylenko, T.; Liou, Y.-F.; Chiou, P.-C.; Chu, H.-W.; Lai, Y.-S.; Chou, Y.-L.; Huang, H.-L.; Ho, S.-Y. SCMBYK: prediction and characterization of bacterial tyrosine-kinases based on propensity scores of dipeptides. *BMC Bioinf.* **2016**, *17*, 514.

(17) Liou, Y.-F.; Charoenkwan, P.; Srinivasulu, Y. S.; Vasylenko, T.; Lai, S.-C.; Lee, H.-C.; Chen, Y.-H.; Huang, H.-L.; Ho, S.-Y. SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinf.* **2014**, *15*, S4.

(18) Vasylenko, T.; Liou, Y.-F.; Chen, H.-A.; Charoenkwan, P.; Huang, H.-L.; Ho, S.-Y. SCMPSP: Prediction and characterization of photosynthetic proteins based on a scoring card method. *BMC Bioinf.* **2015**, *16*, S8.

(19) Xie, R.; Li, J.; Wang, J.; Dai, W.; Leier, A.; Marquez-Lago, T. T.; Akutsu, T.; Lithgow, T.; Song, J.; Zhang, Y. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. *Briefings Bioinf.* **2021**, *22*, bbaa125.

(20) Charoenkwan, P.; Chotpatiwetchkul, W.; Lee, V. S.; Nantasenamat, C.; Shoombuatong, W. A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Sci. Rep.* **2021**, *11*, 23782.

(21) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Wang, Y.; Webb, G. I.; Smith, A. I.; Daly, R. J.; Chou, K.-C.; Song, J. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499−2502.

(22) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(23) Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Briefings Bioinf.* **2020**, *21*, 408−420.

(24) Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.* **2020**, *40*, 1276−1314.

(25) Shoombuatong, W.; Prathipati, P.; Owasirikul, W.; Worachartcheewan, A.; Simeon, S.; Anuwongcharoen, N.; Wikberg, J. E.; Nantasenamat, C., Towards the revival of interpretable QSAR models. *Advances in QSAR Modeling*; Springer, 2017; pp 3−55.

(26) Hasan, M. M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* **2020**, *36*, 3350−3356.

(27) Charoenkwan, P.; Kanthawong, S.; Nantasenamat, C.; Hasan, M. M.; Shoombuatong, W. iAMY-SCM: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics* **2021**, *113*, 689.

(28) Azadpour, M.; McKay, C. M.; Smith, R. L. Estimating confidence intervals for information transfer analysis of confusion matrices. *J. Acoust. Soc. Am.* **2014**, *135*, EL140−EL146.

(29) Zhang, D.; Xu, Z.-C.; Su, W.; Yang, Y.-H.; Lv, H.; Yang, H.; Lin, H. iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics* **2021**, *37*, 171−177.

(30) Charoenkwan, P.; Chiangjong, W.; Nantasenamat, C.; Hasan, M. M.; Manavalan, B.; Shoombuatong, W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings Bioinf.* **2021**, *22*, bbab172.

(31) Li, H.; Gong, Y.; Liu, Y.; Lin, H.; Wang, G. Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Briefings Bioinf.* **2022**, *23*, bbab533.

(32) Zulfiqar, H.; Yuan, S.-S.; Huang, Q.-L.; Sun, Z.-J.; Dao, F.-Y.; Yu, X.-L.; Lin, H. Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4123−4131.

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664

(33) Nagarajan, D.; Roy, N.; Kulkarni, O.; Nanajkar, N.; Datey, A.; Ravichandran, S.; Thakur, C.; T, I. V.; Aprameya, S. P.; Sarma, D.; Chakravortty, D.; Chandra, N. Ω76: A designed antimicrobial peptide to combat carbapenem-and tigecycline-resistant Acinetobacter baumannii. *Sci. Adv.* **2019**, *5*, No. eaax1946.
(34) Abbassi, F.; Raja, Z.; Oury, B.; Gazanion, E.; Piesse, C.; Sereno, D.; Nicolas, P.; Foulon, T.; Ladram, A. Antibacterial and leishmanicidal activities of temporin-SHd, a 17-residue long membrane-damaging peptide. *Biochimie* **2013**, *95*, 388−399.
(35) Peel, E.; Cheng, Y.; Djordjevic, J.; Fox, S.; Sorrell, T.; Belov, K. Cathelicidins in the Tasmanian devil (Sarcophilus harrisii). *Sci. Rep.* **2016**, *6*, 35019.
(36) Zai, Y.; Xi, X.; Ye, Z.; Ma, C.; Zhou, M.; Chen, X.; Siu, S. W.; Chen, T.; Wang, L.; Kwok, H. F. Aggregation and its influence on the bioactivities of a novel antimicrobial peptide, temporin-PF, and its analogues. *Int. J. Mol. Sci.* **2021**, *22*, 4509.
(37) Mishra, B.; Wang, X.; Lushnikova, T.; Zhang, Y.; Golla, R. M.; Narayana, J. L.; Wang, C.; McGuire, T. R.; Wang, G. Antibacterial, antifungal, anticancer activities and structural bioinformatics analysis of six naturally occurring temporins. *Peptides* **2018**, *106*, 9−20.
(38) Conlon, J. M.; Abraham, B.; Sonnevend, A.; Jouenne, T.; Cosette, P.; Leprince, J.; Vaudry, H.; Bevier, C. R. Purification and characterization of antimicrobial peptides from the skin secretions of the carpenter frog Rana virgatipes (Ranidae, Aquarana). *Regul. Pept.* **2005**, *131*, 38−45.
(39) de Barros, E.; Gonçalves, R. M.; Cardoso, M. H.; Santos, N. C.; Franco, O. L.; Cândido, E. S. Snake venom cathelicidins as natural antimicrobial peptides. *Front. Pharmacol.* **2019**, *10*, 1415.
(40) Gopal, R.; Lee, J. H.; Kim, Y. G.; Kim, M.-S.; Seo, C. H.; Park, Y. Anti-microbial, anti-biofilm activities and cell selectivity of the NRC-16 peptide derived from witch flounder, Glyptocephalus cynoglossus. *Mar. Drugs* **2013**, *11*, 1836−1852.
(41) Conlon, J. M.; Mechkarska, M.; Prajeep, M.; Sonnevend, A.; Coquet, L.; Leprince, J.; Jouenne, T.; Vaudry, H.; King, J. D. Host-defense peptides in skin secretions of the tetraploid frog Silurana epitropicalis with potent activity against methicillin-resistant Staphylococcus aureus (MRSA). *Peptides* **2012**, *37*, 113−119.
(42) Abraham, P.; George, S.; Kumar, K. S. Novel antibacterial peptides from the skin secretion of the Indian bicoloured frog *Clinotarsus curtipes*. *Biochimie* **2014**, *97*, 144−151.
(43) Mishra, B.; Wang, G. Ab initio design of potent anti-MRSA peptides based on database filtering technology. *J. Am. Chem. Soc.* **2012**, *134*, 12426−12429.
(44) Patrzykat, A.; Gallant, J. W.; Seo, J.-K.; Pytyck, J.; Douglas, S. E. Novel antimicrobial peptides derived from flatfish genes. *Antimicrob. Agents Chemother.* **2003**, *47*, 2464−2470.
(45) Clark, D. P.; Durell, S.; Maloy, W. L.; Zasloff, M. Ranalexin. A novel antimicrobial peptide from bullfrog (Rana catesbeiana) skin, structurally related to the bacterial antibiotic, polymyxin. *J. Biol. Chem.* **1994**, *269*, 10849−10855.
(46) Chen, Q.; Cheng, P.; Ma, C.; Xi, X.; Wang, L.; Zhou, M.; Bian, H.; Chen, T. Evaluating the bioactivity of a novel broad-spectrum antimicrobial peptide brevinin-1gha from the frog skin secretion of *hylarana guentheri* and its analogues. *Toxins* **2018**, *10*, 413.
(47) Dong, Z.; Luo, W.; Zhong, H.; Wang, M.; Song, Y.; Deng, S.; Zhang, Y. Molecular cloning and characterization of antimicrobial peptides from skin of *Hylarana guentheri*. *Acta Biochim. Biophys. Sin.* **2017**, *49*, 450−457.
(48) Jin-Jiang, H.; Jin-Chun, L.; Min, L.; Qing-Shan, H.; Guo-Dong, L. The design and construction of K11: a novel α-helical antimicrobial peptide. *Int. J. Microbiol.* **2012**, *2012*, 764834.
(49) Mangoni, M. Temporins, anti-infective peptides with expanding properties. *Cell. Mol. Life Sci.* **2006**, *63*, 1060−1069.
(50) Xie, Z.; Wei, H.; Meng, J.; Cheng, T.; Song, Y.; Wang, M.; Zhang, Y. The analogs of temporin-GHa exhibit a broader spectrum of antimicrobial activity and a stronger antibiofilm potential against Staphylococcus aureus. *Molecules* **2019**, *24*, 4173.
(51) Mishra, B.; Wang, G. The importance of amino acid composition in natural AMPs: an evolutional, structural, and functional perspective. *Front. Immunol.* **2012**, *3*, 221.
(52) Zouhir, A.; Jridi, T.; Nefzi, A.; Ben Hamida, J.; Sebei, K. Inhibition of methicillin-resistant Staphylococcus aureus (MRSA) by antimicrobial peptides (AMPs) and plant essential oils. *Pharm. Biol.* **2016**, *54*, 3136−3150.
(53) Shang, L.; Li, J.; Song, C.; Nina, Z.; Li, Q.; Chou, S.; Wang, Z.; Shan, A. Hybrid antimicrobial peptide targeting Staphylococcus aureus and displaying anti-infective activity in a murine model. *Front. Microbiol.* **2020**, *11*, 1767.
(54) Zarena, D.; Mishra, B.; Lushnikova, T.; Wang, F.; Wang, G. The π configuration of the WWW motif of a short Trp-rich peptide is critical for targeting bacterial membranes, disrupting preformed biofilms, and killing methicillin-resistant Staphylococcus aureus. *Biochemistry* **2017**, *56*, 4039−4043.
(55) Sharma, A.; Kapoor, P.; Gautam, A.; Chaudhary, K.; Kumar, R.; Chauhan, J. S.; Tyagi, A.; Raghava, G. P. Computational approach for designing tumor homing peptides. *Sci. Rep.* **2013**, *3*, 1607.
(56) Nagpal, G.; Usmani, S. S.; Dhanda, S. K.; Kaur, H.; Singh, S.; Sharma, M.; Raghava, G. P. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci. Rep.* **2017**, *7*, 42851.
(57) Kumar, R.; Chaudhary, K.; Chauhan, J. S.; Nagpal, G.; Kumar, R.; Sharma, M.; Raghava, G. P. An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* **2015**, *5*, 12512.
(58) Chaudhary, K.; Kumar, R.; Singh, S.; Tuknait, A.; Gautam, A.; Mathur, D.; Anand, P.; Varshney, G. C.; Raghava, G. P. A web server and mobile app for computing hemolytic potency of peptides. *Sci. Rep.* **2016**, *6*, 22843.
(59) Klein, P.; Kanehisa, M.; DeLisi, C. Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1984**, *787*, 221−226.
(60) Wang, Z.; Wang, G. APD: the antimicrobial peptide database. *Nucleic Acids Res.* **2004**, *32*, D590−D592.
(61) Zimmerman, J.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170−201.
(62) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **1996**, *3*, 842−848.

https://doi.org/10.1021/acsomega.2c04305
ACS Omega 2022, 7, 32653−32664