Discovering New Agents Active against Methicillin-Resistant *Staphylococcus aureus* with Ligand-Based Approaches

Ling Wang,† Xiu Le,† Long Li,† Yingchen Ju,† Zhongxiang Lin,‡ Qiong Gu,*† and Jun Xu*†

†Research Center for Drug Discovery and Institute of Human Virology, School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China
‡College of Chemical Engineering, Nanjing Forestry University, Nanjing 210037, China

Supporting Information

ABSTRACT: To discover new agents active against methicillin-resistant *Staphylococcus aureus* (MRSA), *in silico* models derived from 5451 cell-based anti-MRSA assay data were developed using four machine learning methods, including naive Bayesian, support vector machine (SVM), recursive partitioning (RP), and k-nearest neighbors (kNN). A total of 876 models have been constructed based on physicochemical descriptors and fingerprints. The overall predictive accuracies of the best models exceeded 80% for both training and test sets. The best model was employed for the virtual screening of anti-MRSA compounds, which were then validated by a cell-based assay using the broth microdilution method with three types of highly resistant MRSA strains (ST239, ST5, and 252). A total of 12 new anti-MRSA agents were confirmed, which had MIC values ranging from 4 to 64 mg/L. This work proves the capacity of combined multiple ligand-based approaches for the discovery of new agents active against MRSA with cell-based assays. We think this work may inspire other lead identification processes when cell-based assay data are available.

INTRODUCTION

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a major cause of patient morbidity and mortality and its associated health care costs.¹⁻³ The emergence of new pathogenic strains has led to the recognition of community-associated MRSA (CA-MRSA), as well as hospital-associated MRSA (HA-MRSA).⁴⁻⁵ In the last century, the high prevalence of MRSA across the world and the paucity of effective drugs prompted the increased use of vancomycin, despite its poor bioavailability and associated toxicity.⁶ This resulted in the emergence of vancomycin-intermediate *S. aureus* and vancomycin-resistant *S. aureus,*⁷ Anti-MRSA drug discovery is impaired by many issues, such as efficacy, toxicity, adverse drug reactions, and multidrug resistance as well as a lack of detailed information about the modes of actions for the chemotherapeutic agents.⁸⁻⁹ Over the last two decades, genomics-based antibacterial target discovery programs have made significant progress.¹⁰⁻¹¹ In 1995, the sequencing of the first complete bacterial genome heralded a new era of antibacterial drug discovery. This provided the tools to search for new antibacterial drug targets from entire genomes. Target-based approaches (protein screening) became major tools for discovering anti-MRSA agents; however, novel anti-MRSA drugs were not available in the market for clinical use.⁸⁻⁹ The discovery of type II fatty acid synthesis and peptide deformylase inhibitors are successful examples of this target-based approach. Three *S. aureus* FabI inhibitors (AFN-1252,¹ Fab-001¹² and CG400549¹³) and *S. aureus* peptide deformylase inhibitor (GSK1322322¹⁴) are currently in clinical trial for use in MRSA infections. A recent study has suggested that type II fatty acid synthesis is not a suitable antibiotic target for the MRSA infection because *S. aureus* uses fatty acids directly from the host serum rather than from *de novo* synthesis.¹⁵ Debate about this topic is ongoing.¹⁶⁻¹⁷ The identification of novel targets requires the characterization of MRSA-specific biochemical pathways. But the rational design of new anti-MRSA agents via a target-based approach is complex, and many metabolic processes are unknown. Target-based approaches are seen as “not delivering the pipeline” in a timely manner, and intense efforts should be continued. Therefore, other approaches should be attempted because of the impending dire situation without effective antibiotics.⁹⁻¹⁸⁻²¹

Compared to target-based approaches, the traditional whole cell-based screening approach (phenotypic screening) is an old but indispensable method to discover new anti-MRSA agents. A cell-based approach validates if the target-active agent interaction...
In vitro by data. Four data mining methods (naive Bayes screening data, we collected 5451 cell-based anti-MRSA assay we were inspired to predict potential anti-MRSA agents based new anti-MRSA agents. The virtual screening hits were veri
validation. In addition, based on cross validation, a test set validation, and an external test set machine, recursive partitioning, and k-nearest neighbors) were
results, the descriptors with more than 95% zero values or zero
process resulted in 182 (from MOE) and 252 (from DS)
MOE represents 21 descriptors from MOE calculations, and DS represents 29 descriptors from Discovery Studio calculations.

<table>
<thead>
<tr>
<th>MATERIALS AND METHODS</th>
</tr>
</thead>
</table>

**Cell-Based Anti-MRSA Assay Data.** The cell-based anti-MRSA screening data set was extracted from the ChEMBL database (version 17) and refined with the following criteria: (1) Only cell based assay data were selected. (2) Only MIC assay values based on MRSA strains were kept, and other assay data were excluded, e.g., MIC, MIC, MIC, IC, and IC. (3) Duplicate data and compounds without detailed assay values were removed. This process generated a data set of 5451 compounds and their cell-based anti-MRSA activities. The MIC values in this data set ranged from 0.000002 to 334955.781 μM (11 orders of magnitude). There were 2066 active compounds in the data set below the MIC threshold of 5 μM (a cutoff for hit-to-lead activity studies). The detailed results of choosing a MIC threshold are available in Figure S1 of the Supporting Information.

The structures of the compounds were downloaded and checked against the original published papers. Each molecular structure record experienced preprocessing, washing counterions, adding hydrogen atoms, and optimization by molecular mechanics with the MMFF94 force field by means of the MOE program (version 2010.10, Chemical Computing Group, Inc., Canada). The structural data were saved in a MACCS SDF file and a SMILES file. Finally, the entire data set was randomly split into a training set (4088) and test set (1363). The data are available in Table S1 of the Supporting Information.

**Molecular Descriptors Calculations.** The topological descriptors were calculated using MOE and DS 3.5 (Discovery Studio, version 3.5, Accelrys, Inc., San Diego, CA, U.S.A.). This process resulted in 182 (from MOE) and 252 (from DS) molecular descriptors for each cell-based anti-MRSA agent.

**Molecular Descriptors Selection.** To generate meaningful results, the descriptors with more than 95% zero values or zero variances were removed. To reduce noise and avoid bias, Pearson correlation analyses were carried out to eliminate the descriptors that were weakly correlated (Pearson correlation coefficient <0.1) with anti-MRSA activity or the descriptors that were highly correlated (Pearson correlation coefficient >0.9) with other descriptor(s).

A stepwise variable selection method via linear regression analysis was performed for the remaining descriptors. The linear regression analysis of the anti-MRSA activity and the first molecular descriptor was performed to generate an initial equation. Then, additional molecular descriptors were added to the regression equation one by one. A significance test was conducted for every new regression equation and each descriptor in the equation. If the new regression equation was not "statistically significant" following the addition of a new descriptor, the new descriptor would be removed. The linear regression process was recursively executed until all descriptors that had been examined. All linear regression analyses are performed in SPSS 17.0. Consequently, 21 descriptors (from MOE) and 29 descriptors (from DS 3.5) were chosen and are listed in Table 1.

**Calculation of Molecular Fingerprints.** Molecular fingerprints are an abstract representation of certain structural features of a molecule, which are stored in a bit map that can be used for QSAR modeling. With DS 3.5, we calculated 28 fingerprints, including the SciTeP fingeripnts (ECFP, FCFP, and LCFP with diameters of 4, 6, 8, 10, and 12) and the Daylight fingerprints (EFPF, FFPP, and LPPF with diameters of 4, 6, 8, 10, and 12, if applicable).

**Modeling Methods.** Naive Bayesian (NB), support vector machine (SVM), recursive partitioning (RP), and k-nearest neighbors (kNN) methods were employed to build ligand-based models for the virtual screening campaigns. The NB and RP models were constructed with DS 3.5. The SVM model was built with LIBSVM 3.17 package. The k-N model was built with Orange 2.0 (http://www.ailab.si/orange/).

**Naive Bayesian (NB).** Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability and a "likelihood function" derived from a probability model for the data to be observed. Bayesian inference computes the posterior probability directly based on the following core function,

$$ P(c|x) = \frac{P(c)P(x|c)}{\sum_i P(i)P(x|i)} $$

where $P(c)$ is the initial degree of belief in $c$, $P(x)$ is the initial degree of belief in $x$, $P(c|x)$ is the degree of belief having accounted for $x$, $P(x|c)$ is the degree of belief having accounted for $c$, and $c$ and $x$ represent independent molecules ($i, j = 1, 2, ..., N$). Detailed descriptions of the naive Bayesian method can be found in previously published literature.
Support Vector Machine (SVM). The SVM was first developed by Vapnik \(^{28}\) for pattern recognition to minimize structural risk under the frame of the VC theory. Each molecule is represented by an eigenvector \(t\), and the selected patterns, \(t_1, t_2, \ldots, t_n\) make up the components of \(t\). For SVM training, the category label \(y\) was added. The \(i\)th molecule in the data set is defined as \(M_i = (t_i, y_i)\), where \(y_i = 1\) for the active category and \(y_i = 0\) for the inactive category. SVM gives a classifier

\[
 f(t) = \text{sgn}\left\{ \frac{1}{2} \sum_{i \neq j} a_i a_j K(t_i, t_j) + b \right\}
\]

where \(a_i\) is the coefficient to be learned, and \(K\) is a kernel function. The coefficient \(a_i\) and \(b\) are determined by maximizing the Lagrangian expression

\[
 \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j K(t_i, t_j)
\]

under the following conditions

\[
 0 \leq a_i \leq C \quad \text{and} \quad \sum_{i=1}^{n} y_i a_i = 0
\]

In the present study, a Gaussian radial basis function (RBF) kernel was used to build models. The two parameters of the SVM \((C, \gamma)\) for each model were selected using the autosearching program “grid” through a 5-fold cross validation in LibSVM.

Recursive Partitioning (RP). RP is a statistical method for multivariable analysis. It creates a decision tree that strives to correctly classify members of the population based on a dichotomous dependent variable (e.g., active or inactive class).

Figure 1. Training and testing sets of anti-MRSA have covered broad chemical diversity. (a) Chemical space of anti-MRSA training compounds (4088) and testing compounds (1363). (b) Chemical space comparison of compounds in anti-MRSA data set, DrugBank, and Specs database. MRSACOM: compounds with activity against MRSA. MW: molecular weight.
and a set of independent variables (e.g., molecular properties and fingerprints). A 5-fold cross-validation scheme was used to determine the degree of pruning required for the best predictive performance. Detailed descriptions of the RP method can be found in the literature.29,30

k-Nearest Neighbor (k-NN). The k-nearest neighbor algorithm (k-NN) is a method to classify objects based on the closest examples in the feature space. In k-NN, the Euclidean distance between an unclassified vector \( x \) and each individual vector \( x_i \) in the training set is calculated using the following formula

\[
D = \sqrt{x - x_i^2}
\]

A total of \( k \) number of vectors nearest to the vector \( x \) are used to determine the class of that unclassified vector. The class of the majority of the \( k \)-nearest neighbors is decided as the predicted class of the unclassified vector \( x \). In the present study, the nearness is measured by Euclidean distance metrics, and the parameter of \( k = 5 \) (default parameter) was used.

Validating Performances of Models. A 5-fold cross-validation scheme was employed to validate the accuracy and robustness of the models. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE, prediction accuracy for active compound against MRSA), specificity (SP, prediction accuracy for inactive compound), overall predictive accuracy (Q), and Matthews correlation coefficient (C) were calculated. The area values under the receiver operating characteristic (ROC) curves were also calculated.31 The performances of multiple machine learning approaches can be found in other studies.32

Results and Discussion

Chemical Space and Structural Diversity Analysis. The chemical space of the training and testing data sets influences the predictive ability of the in silico models. One way to view the diversity is to depict compounds in a two-dimensional space using molecular weight (MW) and ALogP as shown in Figures 1a and b. Figure 1a indicates that the training set and the test compounds are distributed over a wide range of MW (61–4500 Da) and ALogP (−30–20) values. By comparing the chemical diversity of the 5451 anti-MRSA agents against the chemical diversity of DrugBank38 and the Specs database39 (Figure 1b), we determined that the chemical diversity of DrugBank and the Specs database is included in the diversity space of our anti-MRSA data set. The SCA plot40 further confirms that the anti-MRSA compounds are structurally more diverse than the compounds in the Specs database and similar to the compounds in DrugBank (Table S2, Supporting Information).

Descriptors Highly Correlated with Anti-MRSA Activity. A number of molecular properties, such as lipophilicity, hydrogen bonding ability, molecular flexibility, and molecular volume, have been useful for QSAR, QSRR, and ADME predictions.30,41 To identify the descriptors that are significantly associated with anti-MRSA activity, we conducted Student’s t-tests (p-value, Table 2). Table 2 indicates that the means of molecular mass (MM), hydrogen bond acceptor (HBA), sums of N+O, and Molecular_PolarSASA (MPSASA) values between the active and inactive compounds are significantly different (p-values of 4.725 \( \times 10^{-7} \), 8.647 \( \times 10^{-3} \), 1.382 \( \times 10^{-27} \), and 8.171 \( \times 10^{-24} \), respectively). Furthermore, the anti-MRSA activity index (LogMIC + 6) and the eight descriptors are highly correlated. The correlation coefficients between the anti-MRSA activity index and MM, HBA, N + O, and MPSASA are 0.203, 0.242, 0.197, and 0.200, respectively, for the 5451 anti-MRSA compounds and 0.167, 0.201, 0.161, and 0.184, respectively, for the 2066 highly active (MIC < 5 \( \mu \text{M} \))
compounds. Previous studies reported that the lipophilicity (logP) was important for antibacterial activity.\(^{46,47}\) However, this is not observed in our case (the correlation coefficient between anti-MRSA activity index and AlogP is 0.086, and the \(p\)-value is \(2.075 \times 10^{-7}\)). Perhaps, it was not enough to previously focus on one scaffold (2-(4-substituted phenyl)-3(2H)-isothiazolones and 1-benzylbenzimidazole derivatives) to draw such conclusions. Furthermore, the agent active against \textit{Salmonella typhimurium} or \textit{Escherichia coli} and an anti-MRSA agent may demonstrate different correlations for logP and antibacterial activity. LogS, HBD, and MFPSA measurements cannot significantly differentiate the active anti-MRSA agents from the inactive ones because of low correlation coefficient values and high \(p\)-values (Table 2).

**Performance of Descriptor-Based Models.** Naive Bayesian (NB), support vector machine (SVM), recursive partitioning (RP), and k-nearest neighbors (kNN) models were built based upon the descriptors (physicochemical properties of the compounds) selected by the feature reduction methods (Table 1). The 5-fold cross-validation process was used to evaluate the robustness of the model. The models were validated with a testing data set comprising 1363 compounds (503 active and 860 inactive).

The performance validation results of the descriptor-based models are listed in Table 3. According to the Matthews correlation coefficient (MCC) value from the training set, the SVM_DS, SVM_MOE, kNN_DS, and kNN_MOE models have high overall prediction accuracies (0.919, 0.929, 0.991, and 0.992, respectively). The performance validation results of the SVM_DS, SVM_MOE, kNN_DS, and kNN_MOE models with the training and testing data are consistent. The best model is SVM_MOE with a good Matthews correlation coefficient (\(C = 0.62\)) and suitable overall prediction accuracy (\(Q = 0.83\)), sensitivity (74.8%), and specificity (87.0%) based upon the test data set (1363 compounds).

**Performance of Fingerprint-Based Models.** When building a RP model, the depth of the decision tree controls the complexity of a model. Increasing the depth may improve accuracy but may also result in overfitting.\(^{30}\) The optimized depth can be identified by recursively validating the models with different combinations of the training and testing data. In the present study, the tree depths between 3 and 30 were tried, which

| descriptors | TP  | FN  | TN  | FP  | SP  | C   | Q   | TP  | FN  | TN  | FP  | SP  | C   | Q   |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| NB_DS       | 1038| 525 | 1655| 870 | 0.664| 0.655| 0.311|0.659|293  | 210 | 613 | 247 | 0.583| 0.713| 0.291|0.665|
| NB_MOE      | 968 | 595 | 1952| 573 | 0.619| 0.773| 0.393|0.714|309  | 194 | 628 | 232 | 0.614| 0.730| 0.340|0.687|
| RP_DS       | 1350| 213 | 2084| 441 | 0.864| 0.825| 0.675|0.840|376  | 127 | 640 | 220 | 0.748| 0.744| 0.478|0.745|
| RP_MOE      | 1354| 209 | 2069| 456 | 0.866| 0.819| 0.671|0.837|387  | 116 | 647 | 213 | 0.769| 0.752| 0.507|0.759|
| SVM_DS      | 1362| 201 | 2394| 131 | 0.871| 0.948| 0.827|0.919|365  | 138 | 756 | 104 | 0.726| 0.879| 0.614|0.822|
| SVM_MOE     | 1395| 168 | 2403| 122 | 0.893| 0.952| 0.849|0.929|376  | 127 | 748 | 112 | 0.748| 0.870| 0.621|0.825|
| kNN_DS      | 1544| 19  | 2508| 17  | 0.988| 0.993| 0.981|0.991|387  | 116 | 706 | 154 | 0.769| 0.821| 0.582|0.802|
| kNN_MOE     | 1544| 19  | 2511| 14  | 0.988| 0.994| 0.983|0.992|393  | 110 | 722 | 138 | 0.781| 0.840| 0.614|0.818|

\(^{a}\text{RP, recursive partitioning; NB, na} \grave{\text{i}}\text{ve Bayesian; SVM, support vector machine, and kNN, k-nearest neighbors. MOE represents 21 descriptors from MOE calculations, and DS represents 29 descriptors from Discovery Studio calculations. TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; SE, sensitivity; SP, specificity; Q, overall predictive accuracy, and C, Matthews correlation coefficient. For RP and NB methods, the most important descriptors (self-select by RP and NB) are the same with the optimized 21 MOE descriptors and 29 DS descriptors, respectively.}
resulted 756 RP models (Figure 2). The LPFP_12 fingerprint is not adopted in RP methods because it is time consuming to establish a RP model based on this fingerprint. For the NB model building, we tried 28 NB models based on 28 fingerprints (Figure 3). The 5-fold cross-validation process was applied to measure the robustness of these fingerprint-based models.

As shown in Figure 2, based upon training and testing data, the Matthews correlation coefficient (C) value varies with the decision tree depth. The optimized tree depth varies along with the fingerprints as well. The top five RP models derived from the 27 fingerprints of 1363 highly active anti-MRSA agents are listed in Table 4. Table 4 indicates that the most favored fingerprint set for modeling is EPFP_10. For RP modeling, 23 is the optimized tree depth. The top five RP models derived from the 27 fingerprints of 1363 highly active anti-MRSA agents are listed in Table 4. Table 4 indicates that the most favored fingerprint set for modeling is EPFP_10. For RP modeling, 23 is the optimized tree depth. The average numbers of compounds per leaf (1−23) are 0, 0, 0, 23, 206, 176, 114, 192, 572, 202, 82, 588, 84, 293, 60, 42, 48, 10, 57, 13, 10, 1269, and 47, respectively. The corresponding RP model has a sensitivity of 0.803, specificity of 0.803, and overall prediction accuracy of 80.3%. The model evaluation results from both the training set and testing set are consistent. The AUC values for the training and testing sets are 0.913 and 0.866, respectively (Table 4).

Similar to the RP analysis, the performance of the NB classifiers are different based on different fingerprints and the diameter of the fingerprints (Figure 3). As shown in Figure 3, the Matthews correlation coefficient (MCC) value varies with fingerprint diameter. This trend is also observed in RP modeling (Figure 2). As shown in Table 4, the best NB classifier is derived from fingerprint LCFP_12 with a sensitivity of 0.845, specificity of 0.920, and overall prediction accuracy of 89.2%. The best NB model validated with the testing set has a sensitivity of 0.742, specificity of 0.895, and overall prediction accuracy of 83.9%. For the NB models, the AUC values for the training and testing sets are 0.869 and 0.874, respectively. Compared with the RP models, the NB models have better prediction ability for both the training set and test set (Table 4).

As shown in Figure 3a, the MCC value increases as the fingerprint diameter increases for the training set, while the MCC values do not always increase for the test set (Figure 3b). For
example, the MCC value of ECFP increases significantly from diameter 4 to 12 for the training set, while diameter 6 (ECFP_6) is the best choice. Our findings are consistent with Li’s results.32 Therefore, the best length of the fingerprint for classification models should be determined by the MCC values from the test set.

Performance of Models Based on Combinations of Descriptors and Fingerprints. Molecular descriptors (physicochemical) can depict the properties of an entire molecule, but they cannot characterize the important substructures or the molecular fragments that play a key role in anti-MRSA activity. A fingerprint can make up for this shortcoming. Therefore, combinations of molecular descriptors and fingerprints were used simultaneously to establish in silico models. The NB methodology was employed because it was superior to the RP method, according to the performance results of descriptor- and fingerprint-based models (Tables 3 and 4). A total of 56 models were constructed based upon combinations of the descriptors and the fingerprints. The performance validation results are summarized in Table 5. According the MCC values, all combinational NB models exhibit much better performances than those of sole descriptor-based NB models (Table 3 and Figure 4). Compared with sole fingerprint-based NB models, some combinational NB models, e.g., MOE+ECFP_8, MOE+ECFP_10, and MOE+ECFP_12, show better performance results. However, the performances of some combinational NB models were not improved (e.g., MOE+ECFP_4, MOE+ECFP_6, and MOE+LCFP_12). The major reason may be caused by the complementarity of the molecular descriptor and the fingerprint. If the complementarity of the molecular descriptor and the fingerprint is bigger than the contribution of the sole fingerprint, the combined model will show better performance and vice versa. The same trends were also observed for the NB models based on combinations of DS descriptors and fingerprints (Figure 4).

Table 5 lists the performance validation results with the testing data set (1363 compounds) for the top five NB models using combinations of descriptors and fingerprints. The best NB model was derived from 21 MOE descriptors combined with the LCFP_12 fingerprint and has a sensitivity of 0.846, specificity of 0.920, and overall prediction accuracy of 89.1% (validated with the training set). For the testing data, the best NB model has a sensitivity of 0.773, specificity of 0.864, and overall prediction accuracy of 83.1%. The AUC values are 0.872 and 0.878 for the training and test sets, respectively.

Validating Models with External Testing Data. To access their reliability and usefulness, the models were further validated by an external test data set containing 63 active and 154 inactive compounds collected from recent publications (Table S3, Supporting Information), which were not included in the previously described training and test sets. The external data have novel scaffolds (such as new amphiphilic anthone-based compounds targeting the bacterial membrane, tetrahydropryan-based compounds targeting bacterial topoisomerase, 4-nitripropyrrole-based 1,3,4-oxadiazole derivatives, and alpha-triazolyl chalcone derivatives targeting calf thymus DNA) that are not found in the training and test sets. Nine models have been validated with the external data, and the results are listed in Table 6. Most of the models achieve approximately 75% overall prediction accuracy. The top three models are kNN_MOE, NB_LCFP_12, and NB_LCFP_10, which achieve sensitivities of 0.825, 0.714, and 0.714, specificities of 0.766, 0.799, and 0.805, and overall prediction accuracies of

| Table 5. Performances of Validation Results for Top Five NB Models Using Combinations of Descriptors and Fingerprints |
|----------------------------------|--|---|---|---|---|---|---|---|
| models                          | TP  | FN  | TN  | FP  | SE  | SP  | Q  | AUC |
| training set                    |     |     |     |     |     |     |     |     |
| MOE+LCFP_12                    | 1322| 241 | 2322| 203 | 0.846| 0.920| 0.769| 0.872| 0.891|
| MOE+ECFP_12                    | 1340| 223 | 2296| 223 | 0.857| 0.911| 0.769| 0.870| 0.891|
| MOE+LCFP_10                    | 1307| 248 | 2258| 248 | 0.844| 0.899| 0.772| 0.869| 0.881|
| MOE+ECFP_10                    | 1394| 233 | 2261| 233 | 0.841| 0.896| 0.772| 0.869| 0.877|
| DS+LCFP_10                     | 1351| 212 | 2239| 299 | 0.864| 0.882| 0.751| 0.873| 0.826|
| DS+ECFP_10                     | 1387| 196 | 2231| 207 | 0.857| 0.881| 0.751| 0.873| 0.826|
| DS+ECFP_12                     | 1387| 196 | 2231| 207 | 0.857| 0.881| 0.751| 0.873| 0.826|
| MOE represents 21 descriptors from MOE calculations, and DS represents 29 descriptors from Discovery Studio calculations. TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; SE, sensitivity; SP, specificity; Q, overall prediction accuracy; C, Matthews correlation coefficient; and AUC, area under the receiver operating characteristic curve. |
78.3%, 77.4%, and 77.9%, respectively. Therefore, the models are consistent, reliable, and useful. Moreover, some tested external compounds are new scaffolds that were collected based on different targets, indicating that scaffold hopping can be carried out via virtual screening based on our models. Moreover, any compound with anti-MRSA activity via a target-based approach can be predicted through our in silico cell-based models. In other words, in silico cell-based anti-MRSA models (possibly referred to as target-network models, which are data based on data from multiple known targets and unknown mechanistic data) can cover the target-based approach (e.g., 3D-QSAR, pharmaco- phore, and docking models).

### Validating Models with New Types (MIC$_{50}$ and MIC$_{90}$) of External Data

The previous models were built based upon structure MIC data. Now, we validated them with new types (MIC$_{50}$ and MIC$_{90}$) of external data to confirm that the models are indeed predictive. The MIC$_{50}$ and MIC$_{90}$ assay data were extracted from the ChEMBL database$^{23}$ and reconfigured in the same format as the MIC data used in the previous training and test data sets. Consequently, we collected 284 compounds with MIC$_{50}$ data and 348 compounds with MIC$_{90}$ data (Tables S4 and S5, Supporting Information). Due to different assay type values (MIC$_{50}$ and MIC$_{90}$), the active cutoff values of MIC$_{50}$ and MIC$_{90}$ were set to 1, 5, 10, 15, 20, 25, and 30 μM, respectively. For example, when a cutoff value of MIC$_{50}$ is 1 μM, compounds were considered to be “active” in our study as their reported MIC$_{50}$ assay values were below 1 μM and vice versa. The SVM_MOE, kNN_MOE, NB_LCFP_12, and NB_MOE+LCFP_12 models were employed for predicting the MIC$_{50}$ and MIC$_{90}$ assay data at different active cutoff values. The results are shown in Figure 5.

### Table 6. Performance of Nine in Silico Cell-Based Models on External Test Set

<table>
<thead>
<tr>
<th>models</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>SE</th>
<th>SP</th>
<th>C</th>
<th>AUC</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN_MOE</td>
<td>52</td>
<td>11</td>
<td>118</td>
<td>36</td>
<td>0.825</td>
<td>0.766</td>
<td>0.547</td>
<td>0.852</td>
<td>0.783</td>
</tr>
<tr>
<td>SVM_MOE</td>
<td>52</td>
<td>11</td>
<td>86</td>
<td>68</td>
<td>0.825</td>
<td>0.558</td>
<td>0.350</td>
<td>0.743</td>
<td>0.636</td>
</tr>
<tr>
<td>SVM_DS</td>
<td>48</td>
<td>16</td>
<td>98</td>
<td>56</td>
<td>0.750</td>
<td>0.636</td>
<td>0.352</td>
<td>0.754</td>
<td>0.670</td>
</tr>
<tr>
<td>NB_LCFP_12</td>
<td>45</td>
<td>18</td>
<td>123</td>
<td>31</td>
<td>0.714</td>
<td>0.799</td>
<td>0.488</td>
<td>0.845</td>
<td>0.774</td>
</tr>
<tr>
<td>NB_LCFP_10</td>
<td>45</td>
<td>18</td>
<td>124</td>
<td>30</td>
<td>0.714</td>
<td>0.805</td>
<td>0.496</td>
<td>0.848</td>
<td>0.779</td>
</tr>
<tr>
<td>NB_ECFP_10</td>
<td>46</td>
<td>17</td>
<td>110</td>
<td>44</td>
<td>0.730</td>
<td>0.714</td>
<td>0.409</td>
<td>0.851</td>
<td>0.719</td>
</tr>
<tr>
<td>NB_MOE+LCFP_12</td>
<td>46</td>
<td>17</td>
<td>121</td>
<td>33</td>
<td>0.730</td>
<td>0.786</td>
<td>0.487</td>
<td>0.842</td>
<td>0.770</td>
</tr>
<tr>
<td>NB_MOE+ECFP_12</td>
<td>45</td>
<td>18</td>
<td>122</td>
<td>32</td>
<td>0.714</td>
<td>0.792</td>
<td>0.481</td>
<td>0.839</td>
<td>0.770</td>
</tr>
<tr>
<td>NB_MOE+LCFP_10</td>
<td>45</td>
<td>18</td>
<td>120</td>
<td>34</td>
<td>0.714</td>
<td>0.779</td>
<td>0.466</td>
<td>0.838</td>
<td>0.760</td>
</tr>
</tbody>
</table>

RP, recursive partitioning; NB, naïve Bayesian; SVM, support vector machine; and kNN, k-nearest neighbors. MOE represents 21 descriptors from MOE calculations, and DS represents 29 descriptors from Discovery Studio calculations. TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; SE, sensitivity; SP, specificity; Q, overall predictive accuracy; C, Matthews correlation coefficient; and AUC, area under the receiver operating characteristic curve.
According to the C value from the different cutoff values (Figure 5a and c), the best cutoff values are 10 and 15 μM for MIC$_{50}$ and MIC$_{90}$, respectively. In the best cutoff value distribution, most models can achieve an overall prediction accuracy of approximately 70% (Figure 5b and d). Our results also suggest that the best cutoff values of 10 and 15 μM for MIC$_{50}$ and MIC$_{90}$ are consistent with the cell assay type rule (for the same active compound, MIC$_{50}$ < MIC$_{90}$). All these results illustrate that the cell-based models developed in the present study can predict other cell-based assay results (MIC$_{50}$ and MIC$_{90}$) and exhibit a general ability of prediction.

**Favorable and Unfavorable Fragments for Anti-MRSA Activity.** A shown in Table 4, a NB model (NB_LCFP_12) exhibits the best predictive performance. Consequently, it possesses the best features correlated with anti-MRSA activity. These features (fingerprints) were translated into topological fragments using the DS 3.5 program and are depicted in Figure S2 of the Supporting Information, where the favorable and unfavorable fragments for anti-MRSA activity are depicted and ranked with Bayesian scores.

By analyzing the fragments with positive contributions to anti-MRSA activity (Figure S2a, Supporting Information), it is quite interesting that approximately half of the fragments have nitrogen atoms encoded in saturated rings, and nearly half of the fragments (G11-G20) are core structures. These fragments may be “support scaffolds” that assist in maintaining the active conformation and forming favorable hydrophobic interactions with the anti-MRSA targets. Analysis of the unfavorable fragments for anti-MRSA activity (Figure S2b, Supporting Information) revealed that most fragments contain imides connected with hydrophobic saturated rings or aliphatic chains. Compounds that contain these fragments may not show anti-MRSA activity.

**Applications of in Silico Cell-Based Models and Case Study in Virtual Screening.** On the basis of important information from the in silico cell-based anti-MRSA models, there are at least four applications in drug discovery research. In the simplest sense, the favorable fragments presented in Figure S2a of the Supporting Information can be used as queries for screening compound libraries. Furthermore, the results of the models could be useful for the design and optimization of compounds with anti-MRSA activity by replacing unfavorable fragments with favorable fragments, removing inactive fragments altogether, or adding active fragments to other fragments with promising anti-MRSA activity. In addition, in silico cell-based anti-MRSA models are well suited as tools for virtual screening. Last but not least, cell-based anti-MRSA models can be employed for the design of focused libraries enriched in anti-MRSA compounds starting from any drug-like compound collection, and focused anti-MRSA libraries can be used for any target-based high-throughput assay screening or virtual screening project to avoid a full-scale screening. Hits from focused libraries may show both enzyme inhibition activity and cell line activity against MRSA.

In the present study, a case study in virtual screening was carried out to search new anti-MRSA agents. The NB_LCFP_12 model is elected as the screening engine. The virtual compound library is the Guangdong Small Molecule Tangible Library (GSMTL), which has approximately 7500 compounds. The virtual screening protocol is depicted in Figure S3 of the Supporting Information. The NB_LCFP_12 model selected 887 hits from the GSMTL. Among the hits, the compounds with a molecular weight less than 200 were discarded, which resulted in 440 virtual hits. Furthermore, we removed the compounds with the old scaffolds (149) that were included in the training and testing sets and the compounds with simple scaffolds (72), such as sole benzene ring. Consequently, the virtual hits were further refined, and this resulted in 219 hits. The 219 hits were ranked by EstPGood score (estimate prediction good score from NB method, Table S6, Supporting Information). Finally, 56 compounds were selected based upon their ranking and their availability for in vitro cell-based microbiological assays.

**In Vitro Cell-Based Microbiological Studies.** Vancomycin and ampicillin sodium were used as positive controls. A total of
56 compounds were assayed. Of these, 12 compounds were considered as active agents against MRSA strains (Table 7 and Figure 6), and their MIC values ranged from 4 to 256 mg/L. Compounds (1, 5, 7, 11, and 12) exhibited good activity against three MRSA strains with MIC values below 32 mg/L. These activities were superior or comparable to ampicillin sodium (a common antibacterial drug). These experiments proved that the cell-based model has the capability to identify an anti-MRSA lead.

New anti-MRSA agents based on two major approaches, the discovery of new compounds with a new mechanism of action or the discovery of novel scaffolds for known targets, 9, 18 can

<table>
<thead>
<tr>
<th>Table 7. In Vitro Cell-Based Anti-MRSA Assay Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MIC (mg/L)</strong></td>
</tr>
<tr>
<td><strong>compound no.</strong></td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>vancomycin b</td>
</tr>
<tr>
<td>ampicillin b</td>
</tr>
</tbody>
</table>

*S. aureus: ATCC29213. b Positive control drugs.

Figure 6. Bioassay confirmed anti-MRSA agents.
overcome multiresistant strains (e.g., MRSA). In Table 7, ampicillin sodium exhibits excellent activity against ATCC29213 (MIC = 1 mg/L) but moderate activity against the MRSA strains (ST239, ST5, and 252, MIC = 32 mg/L) due to mutations or modifications of the penicillin-binding proteins (PBPs), the target for β-lactams. Among the active compounds discovered in the present study, most compounds show comparable activity against the Staphylococcus aureus ATCC29213 standard (Table 7), suggesting that these active candidates kill bacteria via a new mechanism of action or a novel scaffolds for known targets. To the best of our knowledge, these 12 compounds have not been previously reported as anti-MRSA agents. On the basis of the analyses above, these lead compounds are worthy for further study.

### CONCLUSIONS

In silico models for the prediction of anti-MRSA agents were developed based on the data from 5451 cell-based anti-MRSA assays with optimized 2D physicochemical descriptors and fingerprints. The models were successfully cross-validated with internal and external data sets. The applications of in silico cell-based anti-MRSA models were proposed. The best model was elected for an anti-MRSA virtual screening campaign, which selected 56 hits from the GSMTL database. The hits were biologically screened with in vitro cell-based microbiological assays, which revealed 12 new anti-MRSA agents. This work demonstrated that in silico cell-based models can efficiently identify novel anti-MRSA agents. The cell-based biological assay data are useful for building predictive virtual screening models. Therefore, this approach may be applied for other lead identification processes.

### ASSOCIATED CONTENT

#### Supporting Information

Distribution of MCC values and overall accuracy values based on different active cutoff values using ECFP_6 and LCFP_4 fingerprints (Figure S1), important favorable and unfavorable fragments for anti-MRSA activity obtained from Bayesian classifiers (Figure S2), schematic representation of anti-MRSA compounds discovery strategy (Figure S3), detailed information on training set and test set and their predicted results based on NB_LCPF_12 model (5451, Table S1), structural diversity comparison of the compounds from COMDECOM, DrugBank, and WDI databases (Tables S2), detailed information on 217 external tested compounds, 284 MIC<sub>50</sub> and 348 MIC<sub>90</sub> assay compounds (Tables S3, S4, S5), and EstPGood score of 219 compounds from GSMTL database and 56 hits selected for in vitro microbiological assay (Table S6). This material is available free of charge via the Internet at http://pubs.acs.org.

### AUTHOR INFORMATION

#### Corresponding Authors

*Phone: +86-20-39943077. Fax: +86-20-39943077. E-mail: guqiong@mail.sysu.edu.cn (Q.G.).*  
*Phone: +86-20-39943023. Fax: +86-20-39943023. E-mail: junxu@biochemomes.com (J.X.).

#### Author Contributions

L. Wang and X. Le contributed equally to this work.

#### Notes

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

This work was supported by the introduction of the innovative R&D team program of Guangdong Province (2009010058), National High Technology Research and Development Program of China (863 Program, 2012AA020307), and Special Funding Program for the National Supercomputer Center in Guangzhou (2012Y2-00045, 2013Y2-00045, 20120000037).

### REFERENCES


