JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

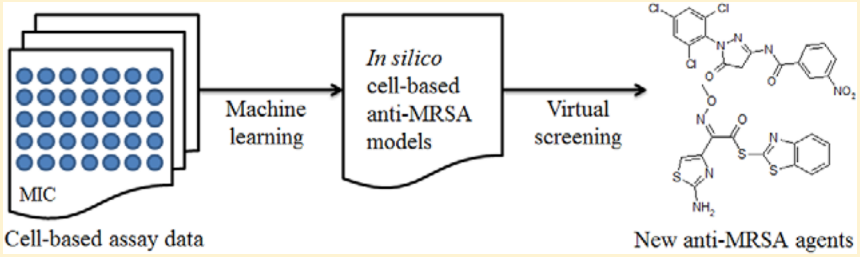
Discovering New Agents Active against Methicillin-Resistant *Staphylococcus aureus* with Ligand-Based Approaches

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 naïve 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-001[12] and CG400549[13]) and *S. aureus* peptide deformylase inhibitor (GSK1322322[14]) 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

Received: April 26, 2014

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Molecular Descriptors Used in This Work

class	number	descriptor
MOE	21	a_ICM, BCUT_SMR_3, PEOE_VSA-2, opr_leadlike, b_rotR, a_acid, VDistEq, SMR_VSA1, a_nO, SlogP_VSA3, logS, PEOE_VSA+1,VAdjMa, PEOE_VSA_FPOS, SlogP_VSA4, GCUT_PEOE_0, opr_violation,GCUT_PEOE_1, BCUT_SLOGP_1, PEOE_VSA+2, GCUT_SMR_1
DS	29	ALogP, Apol, Molecular_Mass, Molecular_Solubility, HBD_Count, NPlusO_Count, Num_AtomClasses, Num_Chains, Num_ExplicitBonds, Num_H_Acceptors, Num_Hydrogens, Num_NegativeAtoms, Num_PositiveAtoms, Num_Rings5, Num_StereoAtoms, Num_StereoBonds, SIC, Num_TerminalRotomers, Molecular_FractionalPolarSASA, Molecular_PolarSASA, BIC, CHI_V_1, CIC, IAC_Mean, IAC_Total, IC, JY, PHI, SC_3_C

MOE represents 21 descriptors from MOE calculations, and DS represents 29 descriptors from Discovery Studio calculations.

has anti-MRSA functionality. A validated compound from a cell-based assay has to bind to its target(s) and play roles in subsequent events, such as accumulation at its site of action.[22] Cell-based approaches cannot confirm modes of action at the target molecular level; however, it provides primary data for ligand structure−activity relationship (SAR) studies. Therefore, we were inspired to predict potential anti-MRSA agents based upon mining whole cell-based screening data.

To develop *in silico* models from the cell-based anti-MRSA screening data, we collected 5451 cell-based anti-MRSA assay data. Four data mining methods (naïve Bayesian, support vector machine, recursive partitioning, and k-nearest neighbors) were employed to construct models for anti-MRSA agent prediction. The performances of the models were successful evaluated by cross validation, a test set validation, and an external test set validation. In addition, based on *in silico* cell-based anti-MRSA models, a virtual screening campaign was carried out to search for new anti-MRSA agents. The virtual screening hits were verified by *in vitro* cell-based anti-MRSA assays.

MATERIALS AND METHODS

Cell-Based Anti-MRSA Assay Data. The cell-based anti-MRSA data set was extracted from the ChEMBL database (version 17)[23] 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_{50}, MIC_{90}, IC_{50}, and K_i. (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[24] 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[25] 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"[24,25] following the addition of a new descriptor, the new descriptor would be removed. The linear regression process was recursively executed until all descriptors had been examined. All linear regression analyses are performed in SPSS 17.0.[24] 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 SciTegic fingerprints (ECFP, FCFP, and LCFP with diameters of 4, 6, 8, 10, and 12) and the Daylight fingerprints (EPFP, FPFP, and LPFP with diameters of 4, 6, 8, 10, and 12, if applicable).

Modeling Methods. Naïve 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.[26] The *k*-NN model was built with Orange 2.0 (http://www.ailab.si/orange/).

Naïve 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,

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{\sum_j Pc_jP(x|j)}$$

where $P(c_i)$ is the initial degree of belief in c_i, $P(x)$ is the initial degree of belief in x, $P(c_i|x)$ is the degree of belief having accounted for x, $P(x|c_i)$ is the degree of belief having accounted for c_i, and c_i and x represent independent molecules ($i, j = 1, 2, ..., N$). Detailed descriptions of the naïve Bayesian method can be found in previously published literature.[27]

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

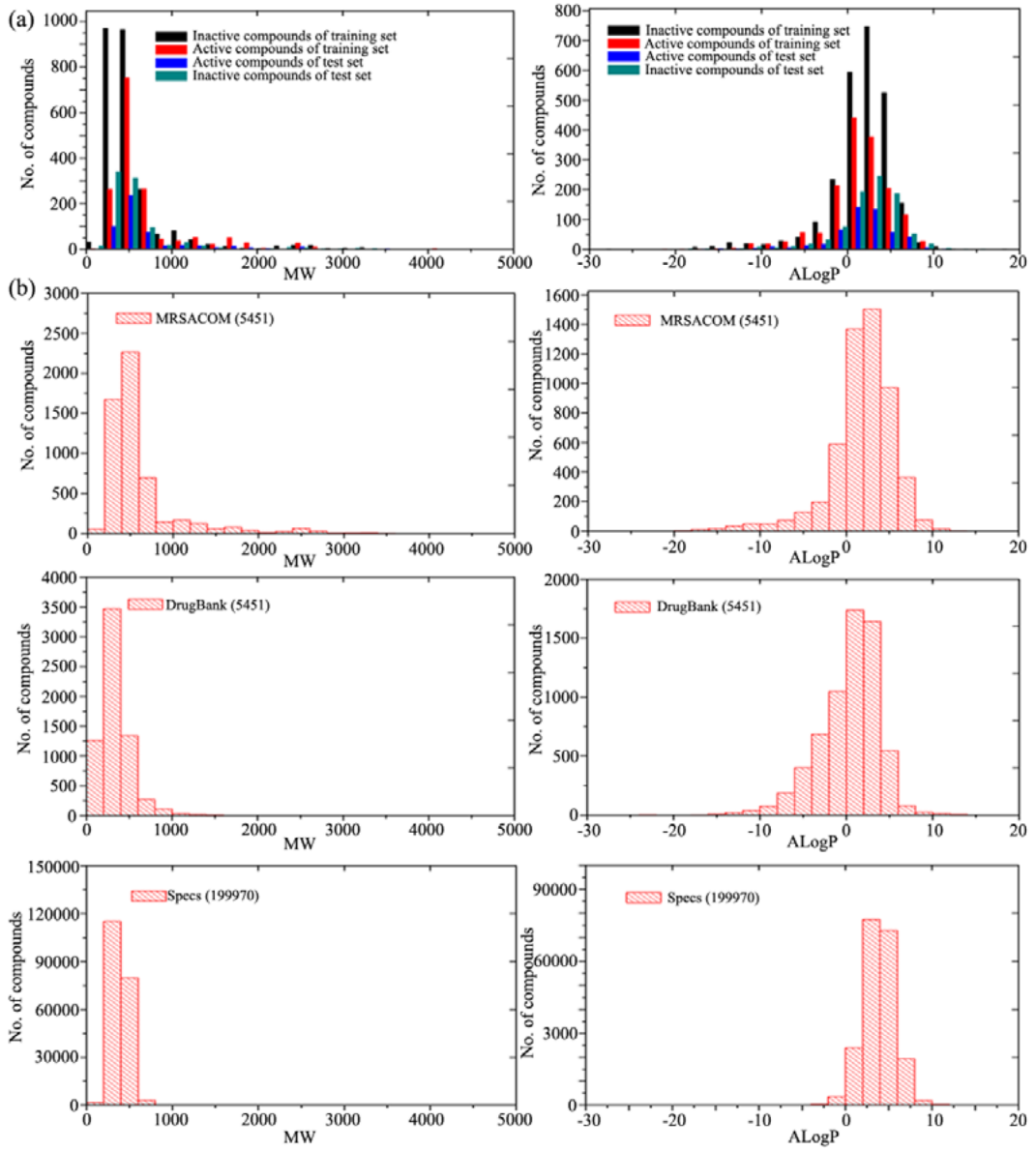


Figure 1. Training and testing sets of anti-MRSA have covered broad chemical diversity. (a) Chemical space of anti-MRSA training compounds (4088) and testing compounds (1363). (b) Chemical space comparison of compounds in anti-MRSA data set, DrugBank, and Specs database. MRSACOM: compounds with activity against MRSA. MW: molecular weight.

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, ..., t_n make up the components of **t**. For SVM training, the category label y was added. The ith 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

$$f(\mathbf{t}) = \mathrm{sgn}\left\{\frac{1}{2}\sum_{i=n}^{n} a_i K(t_i, t) + b\right\}$$

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

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

under the following conditions

$$0 \le a_i \le C \quad \text{and}$$

$$\sum_{y_i=1}^{n} y_i a_i = 0$$

In the present study, a Gaussian radial basis function (RBF) kernel was used to build models. The two parameters of the SVM (C,γ) 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)

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

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.[32−35] Additionally, the actual predication power of each model was evaluated with three independent external testing sets.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Q} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

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

***In Vitro* Cell-Based Microbiological Studies.** The three MRSA strains (ST239, ST5, and 252) and the standard *Staphylococcus aureus* (ATCC29213) acquired from the Chinese Center for Disease Control and Prevention (CDC, China) were used in the cell-based screening campaign for agents against MRSA. The positive control agents were vancomycin and ampicillin sodium. The MIC values were determined using a broth microdilution method (Mueller−Hinton broth) based on the National Committee for Clinical Laboratory Standards (NCCLS).[36,37] The starting concentrations of the tested compounds (dissolved in DMSO) were 256 mg/L and then their working concentrations ranged from 256 to 0.25 mg/L. The solution containing 10 μL of compound was added to 90 μL of bacterial culture (1×10^6 CFU/mL) in the first well of a flat-bottomed 96-well tissue culture plate (JET Biofil, JET Biochemicals Intl., Inc., Canada). The solution was then double diluted. The bacterial culture solution containing the appropriate compound (50 μL) was discarded at the last well to ensure that there was a 100 μL volume of bacterial culture in every well. A set of tubes containing only inoculated broth was kept as a control. The plate was incubated at 37 °C overnight in an electro-heating standing temperature cultivator before the measurement of the absorbance value. The optical density values at 600 nm were measured using a multifunction microplate reader (PowerWave XS2, BioTek Instruments, Inc., U.S.A.). Each experiment in the anti-MRSA assay was replicated twice to define the MIC values.

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 DrugBank[38] and the Specs database[39] (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 plot[40] 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, QSPR, and ADME predictions.[30,41−45] 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 4.725×10^{-9}, 8.647×10^{-38}, 1.382×10^{-22}, and 8.171×10^{-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 μM)

Table 2. Correlation Coefficients and *p*-Values of Anti-MRSA Activity and Descriptors Derived from Anti-MRSA Compound Data Set

descriptor	R_1[a]	R_2[b]	*p*-value[c]
AlogP	0.086	0.121	2.075×10^{-7}
MM	0.203	0.167	4.725×10^{-10}
LogS	0.151	0.122	5.290×10^{-9}
HBD	0.132	0.157	9.932×10^{-8}
HBA	0.242	0.201	8.647×10^{-38}
N + O	0.197	0.161	1.382×10^{-22}
MFPSA	0.106	0.110	4.649×10^{-10}
MPSASA	0.200	0.184	8.171×10^{-24}

[a] R_1 represents the linear correlation between descriptor value and anti-MRSA activity index (LogMIC + 6; 5451 compounds). [b] R_2 represents the linear correlation coefficient for a descriptor and the anti-MRSA activity index (2066 compounds with MIC values <5 μM). [c] *p*-value represents the distributions for active compounds (2066) and inactive compounds (3385).

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Performance Validation Results of Descriptor-Based Models[a]

	training set								test set							
descriptors	TP	FN	TN	FP	SE	SP	*C*	Q	TP	FN	TN	FP	SE	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]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, 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.

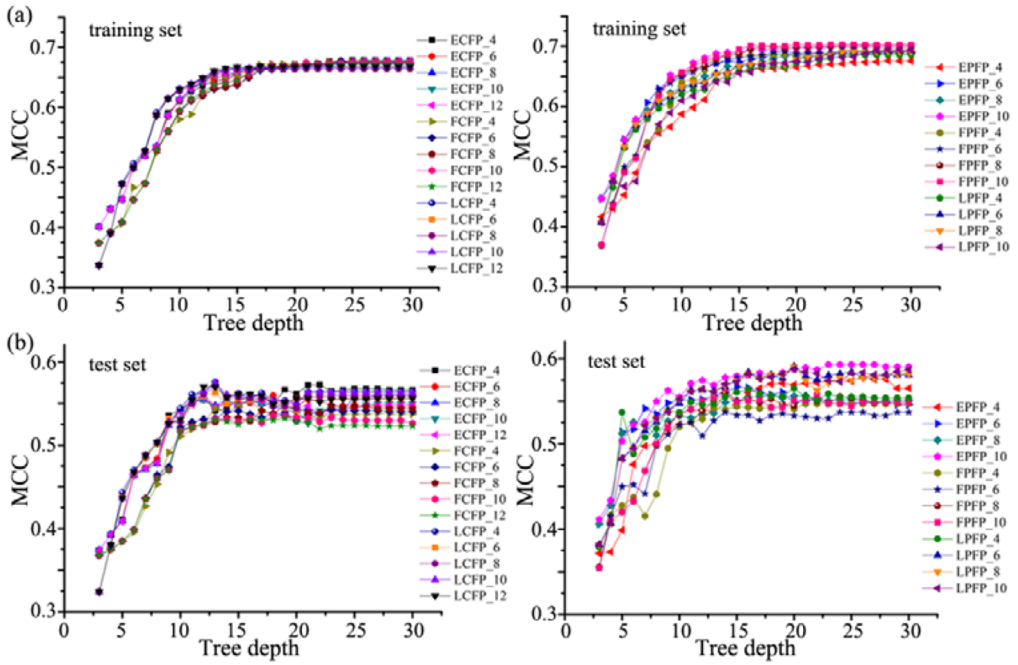


Figure 2. Matthews correlation coefficient (MCC) versus the tree depth for (a) training set and (b) test set. The 756 RP models were constructed.

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 2.075×10^{-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 *Salmonella typhimurium* or *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. Naïve 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

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

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.

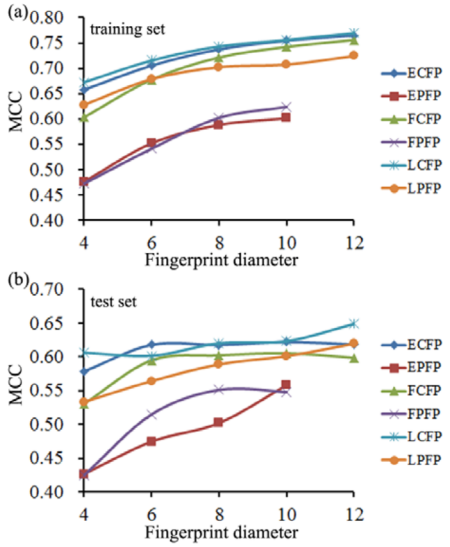


Figure 3. Matthews correlation coefficient (MCC) versus the diameter of the fingerprint set: (a) training set and (b) test set. A total of 28 NB models were constructed based on 28 different fingerprint sets.

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 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% based upon the training set. 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

Table 4. Performances of Top Five RP and NB Models Based on 28 Fingerprints

	training set								test set							
models	TP	FN	TN	FP	SE	SP	C	AUC	TP	FN	TN	FP	SE	SP	C	AUC	Q
RP_EPFP_10	1357	206	2130	395	0.868	0.844	0.699	0.913	404	99	691	169	0.803	0.803	0.593	0.866	0.803
RP_LPFP_6	1286	277	2175	350	0.823	0.861	0.679	0.899	382	121	719	141	0.759	0.836	0.591	0.835	0.808
RP_LPFP_8	1279	284	2181	344	0.818	0.864	0.677	0.896	381	122	720	140	0.757	0.837	0.591	0.836	0.808
RP_LPFP_10	1283	280	2180	345	0.821	0.863	0.679	0.894	381	122	718	142	0.757	0.835	0.588	0.834	0.806
RP_EPFP_4	1328	235	2110	415	0.850	0.836	0.673	0.913	398	105	688	172	0.791	0.800	0.578	0.850	0.797
NB_LCFP_12	1321	242	2324	201	0.845	0.920	0.770	0.869	373	130	770	90	0.742	0.895	0.649	0.874	0.839
NB_LCFP_10	1284	279	2336	189	0.821	0.925	0.756	0.868	372	131	754	106	0.740	0.877	0.623	0.873	0.826
NB_ECFP_10	1295	268	2322	203	0.829	0.920	0.755	0.867	383	120	740	120	0.761	0.860	0.622	0.870	0.824
NB_LCFP_8	1275	288	2321	204	0.816	0.919	0.743	0.867	365	138	760	100	0.726	0.884	0.620	0.871	0.825
NB_LPFP_12	1230	333	2331	194	0.787	0.923	0.724	0.845	361	142	764	96	0.718	0.888	0.619	0.849	0.825

NB, naïve Bayesian; and RP, recursive partitioning. The best tree depth is 23 for RP models (EPFP_10), 20 for RP models (LPFP_6), 20 for RP models (LPFP_8), 20 for RP models (LPFP_10), and 26 for RP models (EPFP_4). 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.

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

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,[48] tetrahydropyran-based compounds targeting bacterial topoisomerase,[49] 4-nitropyrrole-based 1,3,4-oxadiazole derivatives,[50] and alpha-triazolyl chalcone derivatives targeting calf thymus DNA[51]) 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

	training set								test set						
models	TP	FN	TN	FP	SE	SP	C	AUC	Q	TP	FN	TN	FP	SE	SP	C	AUC	Q
MOE+LCFP_12	1322	241	2322	203	0.846	0.920	0.769	0.872	0.891	389	114	743	117	0.773	0.864	0.637	0.878	0.831
MOE+ECFP_12	1340	223	2296	223	0.857	0.911	0.769	0.870	0.891	384	119	747	113	0.763	0.869	0.634	0.875	0.830
MOE+LCFP_10	1307	256	2308	217	0.836	0.914	0.754	0.871	0.884	396	107	732	128	0.787	0.851	0.633	0.877	0.828
DS+LCFP_12	1367	196	2233	292	0.875	0.884	0.751	0.869	0.881	392	111	737	123	0.779	0.857	0.633	0.881	0.828
DS+LCFP_10	1351	212	2226	299	0.864	0.882	0.739	0.869	0.875	392	111	734	126	0.779	0.853	0.629	0.880	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 predictive accuracy; C, Matthews correlation coefficient; and AUC, area under the receiver operating characteristic curve.

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

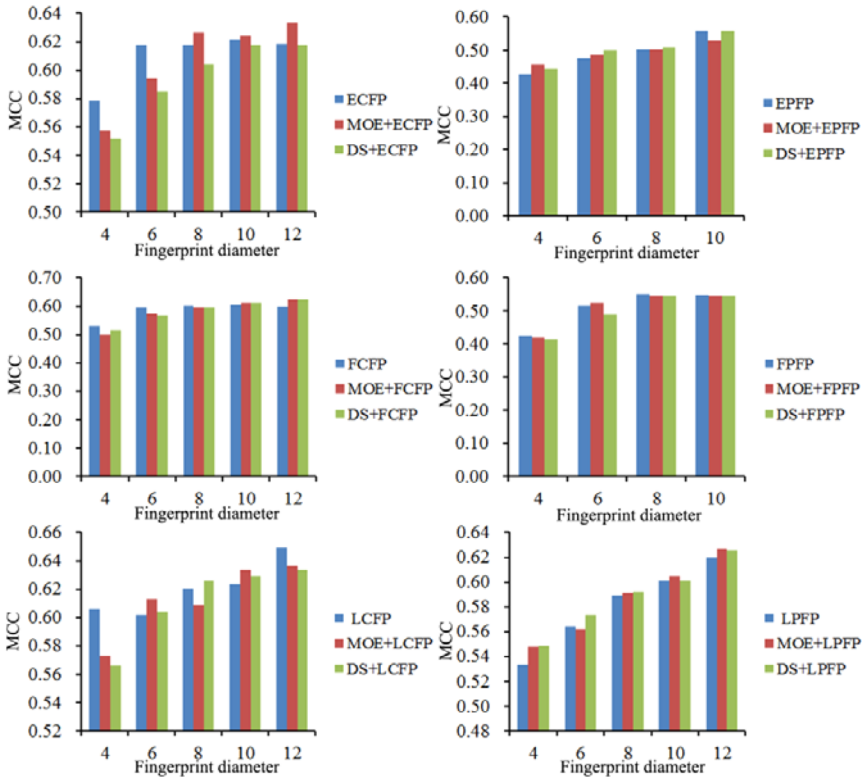


Figure 4. Performances of NB models based upon the combinations of the descriptors and fingerprints validated with the testing data (1363 compounds). MCC: Matthews correlation coefficient. MOE represents 21 descriptors from MOE calculations. DS represents 29 descriptors from Discovery Studio calculations.

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

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

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.

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, pharmacophore, 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.

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

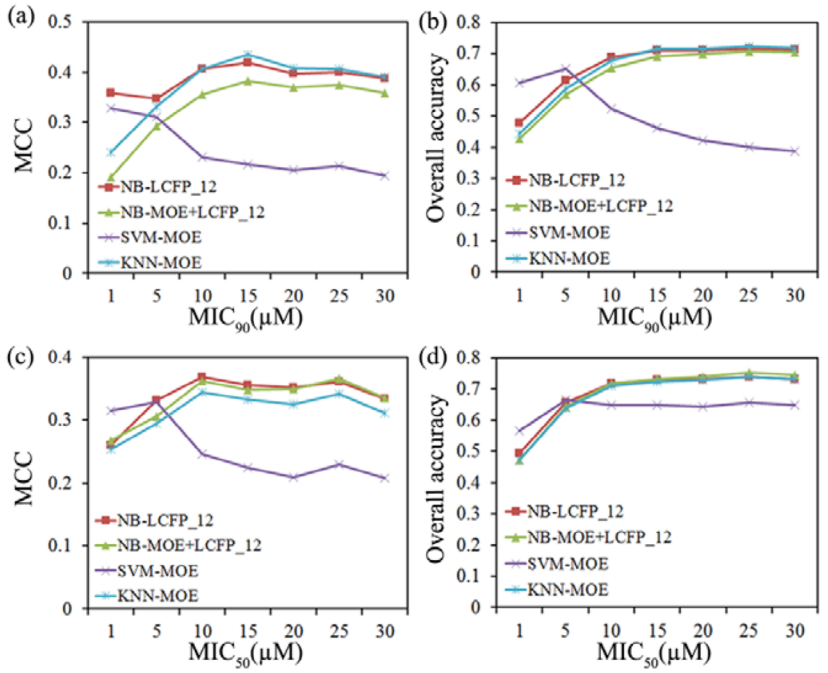


Figure 5. Validating the models with new types (MIC_{50} and MIC_{90}) of external data.

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),[52] 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

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 7. *In Vitro* Cell-Based Anti-MRSA Assay Results

	MIC(mg/L)				
compound no.	MRSA ST239	MRSA ST5	MRSA 252	*S. aureus*[a]	MIC range
1	32	32	32	32	32−32
2	64	256	256	256	64−256
3	64	64	64	64	64−64
4	64	32	16	32	16−32
5	8	16	8	16	8−16
6	>256	32	128	>256	32→256
7	8	4	4	4	4−8
8	128	128	64	8	8−128
9	64	64	16	16	16−64
10	32	64	32	64	32−64
11	8	8	16	16	8−16
12	32	32	32	32	32−32
vancomycin[b]	0.5	0.25	0.5	0.25	0.25−0.5
ampicillin[b]	32	32	32	1	1−32

[a]*S. aureus:* ATCC29213. [b]Positive control drugs.

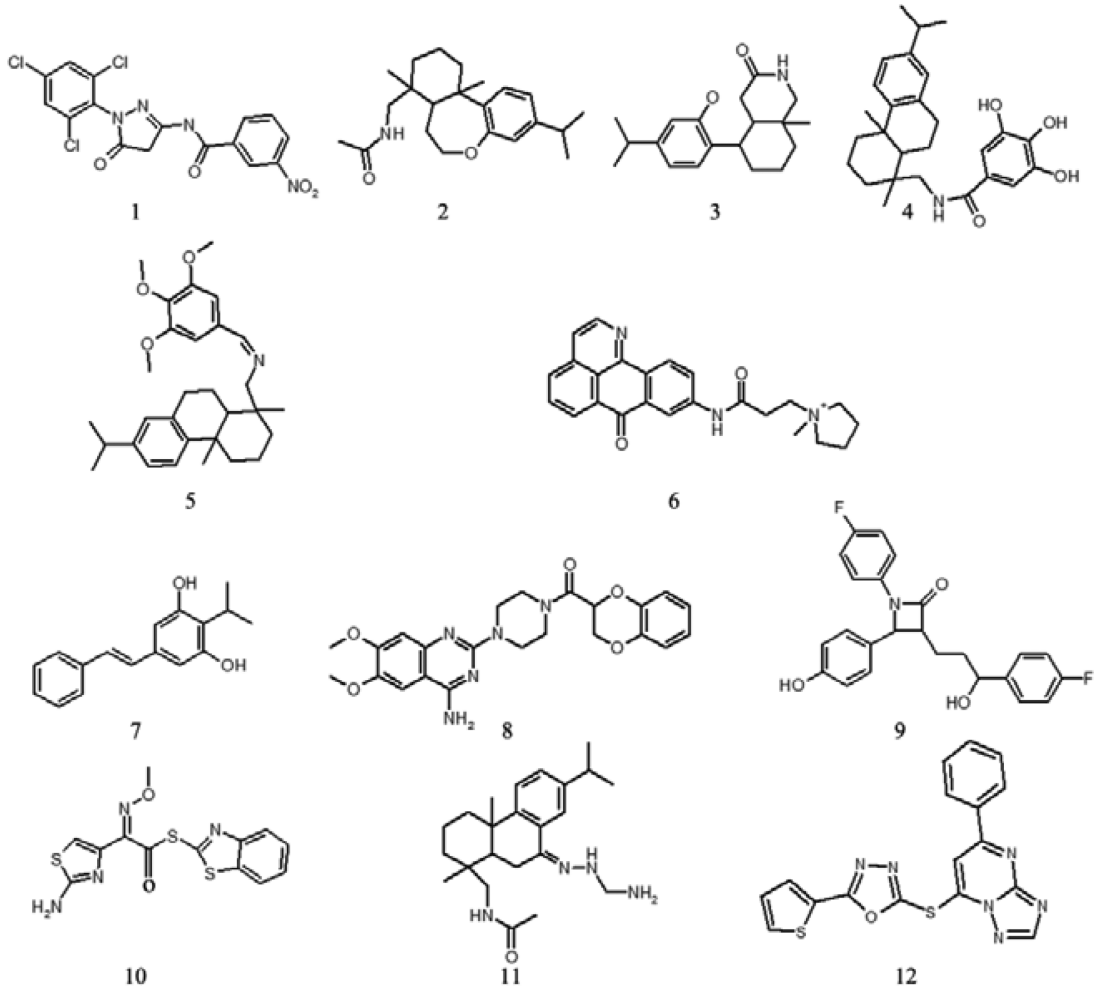


Figure 6. Bioassay confirmed anti-MRSA agents.

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

dx.doi.org/10.1021/ci500253q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

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.[9] 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_{50} and 348 MIC_{90} 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 (2012Y2-00048/2013Y2-00045, 201200000037).

REFERENCES

(1) Kaplan, N.; Albert, M.; Awrey, D.; Bardouniotis, E.; Berman, J.; Clarke, T.; Dorsey, M.; Hafkin, B.; Ramnauth, J.; Romanov, V.; Schmid, M. B.; Thalakada, R.; Yethon, J.; Pauls, H. W. Mode of action, in vitro activity, and in vivo efficacy of AFN-1252, a selective antistaphylococcal FabI inhibitor. *Antimicrob. Agents Chemother.* **2012**, *11*, 5865−5874.
(2) Nair, R.; Ammann, E.; Rysavy, M.; Schweizer, M. L. Mortality among patients with methicillin-resistant Staphylococcus aureus USA300 versus non-USA300 invasive infections: A meta-analysis. *Infect. Control. Hosp. Epidemiol.* **2014**, *1*, 31−41.
(3) Nguyen, D. B.; Lessa, F. C.; Belflower, R.; Mu, Y.; Wise, M.; Nadle, J.; Bamberg, W. M.; Petit, S.; Ray, S. M.; Harrison, L. H.; Lynfield, R.; Dumyati, G.; Thompson, J.; Schaffner, W.; Patel, P. R. Program, E. I. Invasive methicillin-resistant *Staphylococcus aureus* infections among patients on chronic dialysis in the United States, 2005−2011. *Clin. Infect. Dis.* **2013**, *10*, 1393−1400.
(4) Qiao, Y. H.; Dong, F.; Song, W. Q.; Wang, L. J.; Yang, Y. H.; Shen, X. Z. Hospital- and community-associated methicillin-resistant *Staphylococcus aureus*: A 6-year surveillance study of invasive infections in Chinese children. *Acta. Paediatrica.* **2013**, *11*, 1081−1086.
(5) Chambers, H. F. Community-associated MRSA-Resistance and virulence converge. *N. Engl. J. Med.* **2005**, *14*, 1485−1487.
(6) Levine, D. P. Vancomycin: A history. *Clin. Infect. Dis.* **2006**, S5−12.
(7) Hiramatsu, K.; Aritaka, N.; Hanaki, H.; Kawasaki, S.; Hosoda, Y.; Hori, S.; Fukuchi, Y.; Kobayashi, I. Dissemination in Japanese hospitals of strains of *Staphylococcus aureus* heterogeneously resistant to vancomycin. *Lancet* **1997**, *9092*, 1670−1673.
(8) Kumar, K.; Chopra, S. New drugs for methicillin-resistant *Staphylococcus aureus*: An update. *J. Antimicrob. Chemother.* **2013**, *7*, 1465−1470.
(9) Silver, L. L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* **2011**, *1*, 71−109.
(10) Forsyth, R. A.; Haselbeck, R. J.; Ohlsen, K. L.; Yamamoto, R. T.; Xu, H.; Trawick, J. D.; Wall, D.; Wang, L.; Brown-Driver, V.; Froelich, J. M.; C, K. G.; King, P.; McCarthy, M.; Malone, C.; Misiner, B.; Robbins, D.; Tan, Z.; Zhu, Z.; Zy, Y.; Carr, G.; Mosca, D. A.; Zamudio, C.; Foulkes, J. G.; Zyskind, J. W. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **2002**, *6*, 1387−1400.
(11) Miesel, L.; Greene, J.; Black, T. A. Genetic strategies for antibacterial drug discovery. *Nat. Rev. Genet.* **2003**, *6*, 442−456.
(12) Escaich, S.; Prouvensier, L.; Saccomani, M.; Durant, L.; Oxoby, M.; Gerusz, V.; Moreau, F.; Vongsouthi, V.; Maher, K.; Morrissey, I.; Soulama-Mouze, C. The MUT056399 inhibitor of FabI is a new antistaphylococcal compound. *Antimicrob. Agents Chemother.* **2011**, *10*, 4692−4697.
(13) Park, H. S.; Yoon, Y. M.; Jung, S. J.; Kim, C. M.; Kim, J. M.; Kwak, J. H. Antistaphylococcal activities of CG400549, a new bacterial enoyl-acyl carrier protein reductase (FabI) inhibitor. *J. Antimicrob. Chemother.* **2007**, *3*, 568−574.
(14) Ross, J. E.; Scangarella-Oman, N. E.; Miller, L. A.; Sader, H. S.; Jones, R. N. Determination of disk diffusion and MIC quality control ranges for GSK1322322, a novel peptide deformylase inhibitor. *J. Clin. Microbiol.* **2011**, *11*, 3928−3930.
(15) Brinster, S.; L, G.; Staels, B.; Trieu-Cuot, P.; Gruss, A.; Poyart, C. Type II fatty acid synthesis is not a suitable antibiotic target for Gram-positive pathogens. *Nature* **2009**, 83−86.
(16) Balemans, W.; Lounis, N.; Gilissen, R.; Guillemont, J.; Simmen, K.; Andries, K.; Koul, A. Essentiality of FASII pathway for *Staphylococcus aureus*. *Nature* **2010**, *7279*, E3 , discussion E4..

(17) Hafkin, A. K. B. Is there a future for FabI inhibitors as antibacterial agents? *Clin. Invest.* **2013**, *8*, 707−709.
(18) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discovery* **2007**, *1*, 29−40.
(19) Gwynn, M. N.; Portnoy, A.; Rittenhouse, S. F.; Payne, D. J. Challenges of antibacterial discovery revisited. *Ann. N.Y. Acad. Sci.* **2010**, 5−19.
(20) Gilbert, I. H. Drug discovery for neglected diseases: Molecular target-based and phenotypic approaches. *J. Med. Chem.* **2013**, *20*, 7719−7726.
(21) Sams-Dodd, F. Target-based drug discovery: Is something wrong? *Drug Discovery Today* **2005**, *2*, 139−147.
(22) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian models for the prioritization of antitubercular agents. *J. Chem. Inf. Model.* **2008**, *12*, 2362−2370.
(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic. Acids. Res.* **2012**, *40*, D1100−D1107.
(24) Fang, J. S.; Yang, R. Y.; Gao, L.; Zhou, D.; Yang, S. Q.; Liu, A. L.; Du, G. H. Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. *J. Chem. Inf. Model.* **2013**, *11*, 3009−3020.
(25) Wang, L.; Wang, M. L.; Yan, A. X.; Dai, B. Using self-organizing map (SOM) and support vector machine (SVM) for classification of selectivity of ACAT inhibitors. *Mol. Divers.* **2013**, *1*, 85−96.
(26) Chang, C.-C.; Lin C.-J. LIBSVM: A Library for Support Vector Machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed September 8, 2013).
(27) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *1*, 166−178.
(28) Vapnik, V. N. An overview of statistical learning theory. *IEEE. Trans. Neural. Netw.* **1999**, *5*, 988−999.
(29) De'ath, G.; Fabricius, K. E. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* **2000**, *11*, 3178−3192.
(30) Chen, L.; Li, Y. Y.; Zhao, Q.; Peng, H.; Hou, T. J. ADME Evaluation in Drug Discovery. 10. Predictions of p-Glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharmacol.* **2011**, *3*, 889−900.
(31) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *5*, 412−424.
(32) Li, D.; Chen, L.; Li, Y.; Tian, S.; Sun, H.; Hou, T. ADMET Evaluation in drug discovery. 13. Development of in silico prediction models for p-glycoprotein substrates. *Mol. Pharmaceutics* **2014**, *11*, 716−726.
(33) Ling, W.; Lei, C.; Zhihong, L.; Minghao, Z.; Qiong, G.; Jun, X. Predicting mTOR inhibitors with a classifier using recursive partitioning and naïve Bayesian approaches. *PLoS One* **2014**, *9*, e95221.
(34) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *3*, 655−669.
(35) Klepsch, F.; Vasanthanathan, P.; Ecker, G. F. Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors. *J. Chem. Inf. Model.* **2014**, *1*, 218−229.
(36) Performance Standards for Antimicrobial Disk Susceptibility Tests. Approved Standard M2-A6. National Committee for Clinical Laboratory Standards: Wayne, PA, 1997.
(37) European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Dieases (ESCMID). EUCAST definitive document E.DEF 3.1: Determination of minimum inhibitory concentrations (MICs) of antibacterial agents by agar dilution. *Clin. Microbiol. Infect.* **2000**, *9*, 509−515.
(38) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids. Res.* **2011**, *39*, D1035−D1041.
(39) Specs: Chemistry Solutions for Drug Discovery. http://www.specs.net/ (accessed March 1, 2010).
(40) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *24*, 5311−5320.
(41) Tian, S.; Li, Y. Y.; Wang, J. M.; Zhang, J.; Hou, T. J. ADME Evaluation in Drug Discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Mol. Pharmacol.* **2011**, *3*, 841−851.
(42) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian models for the prioritization of antitubercular agents. *J. Chem. Inf. Model.* **2008**, *12*, 2362−2370.
(43) McIntyre, T. A.; Han, C.; Davis, C. B. Prediction of animal clearance using naive Bayesian classification and extended connectivity fingerprints. *Xenobiotica* **2009**, *7*, 487−494.
(44) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *7*, 682−686.
(45) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *3*, 501−511.
(46) Rezaee, S.; Khalaj, A.; Adibpour, N.; Saffary, M. Correlation between lipophilicity and antimicrobial activity of some 2-(4-substituted phenyl)-3(2H)-isothiazolones. *Daru, J. Fac. Pharm., Tehran Univ. Med. Sci.* **2009**, *4*, 256−263.
(47) Podunavac-Kuzmanovic, S. O.; Cvetkovic, D. D.; Barna, D. J. The effect of lipophilicity on the antibacterial activity of some 1-benzylbenzimidazole derivatives. *J. Serb. Chem. Soc.* **2008**, *10*, 967−978.
(48) Zou, H.; Koh, J. J.; Li, J.; Qiu, S.; Aung, T. T.; Lin, H.; Lakshminarayanan, R.; Dai, X.; Tang, C.; Lim, F. H.; Zhou, L.; Tan, A. L.; Verma, C.; Tan, D. T.; Chan, H. S.; Saraswathi, P.; Cao, D.; Liu, S.; Beuerman, R. W. Design and synthesis of amphiphilic xanthone-based, membrane-targeting antimicrobials with improved membrane selectivity. *J. Med. Chem.* **2013**, *6*, 2359−2373.
(49) Surivet, J. P.; Zumbrunn, C.; Rueedi, G.; Hubschwerlen, C.; Bur, D.; Bruyere, T.; Locher, H.; Ritz, D.; Keck, W.; Seiler, P.; Kohl, C.; Gauvin, J. C.; Mirre, A.; Kaegi, V.; Dos, S. M.; Gaertner, M.; Delers, J.; Enderlin-Paput, M.; Boehme, M. Design, synthesis, and characterization of novel tetrahydropyran-based bacterial topoisomerase inhibitors with potent anti-gram-positive activity. *J. Med. Chem.* **2013**, *18*, 7396−7415.
(50) Rane, R. A.; Bangalore, P.; Borhade, S. D.; Khandare, P. K. Synthesis and evaluation of novel 4-nitropyrrole-based 1,3,4-oxadiazole derivatives as antimicrobial and anti-tubercular agents. *Eur. J. Med. Chem.* **2013**, *70*, 49−58.
(51) Yin, B. T.; Yan, C. Y.; Peng, X. M.; Zhang, S. L.; Rasheed, S.; Geng, R. X.; Zhou, C. H. Synthesis and biological evaluation of alpha-triazolyl chalcones as a new type of potential antimicrobial agents and their interaction with calf thymus DNA and human serum albumin. *Eur. J. Med. Chem.* **2014**, 148−159.
(52) Wang, L.; Gu, Q.; Zheng, X.; Ye, J.; Liu, Z.; Li, J.; Hu, X.; Hagler, A.; Xu, J. Discovery of new selective human aldose reductase inhibitors through virtual screening multiple binding pocket conformations. *J. Chem. Inf. Model.* **2013**, *53*, 2409−2422.