ACS OMEGA

Open Access

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

Article

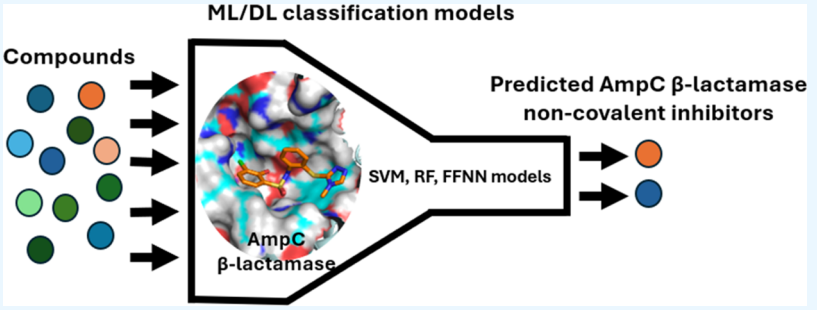
Machine Learning and Deep Learning Models for Predicting Noncovalent Inhibitors of AmpC β-Lactamase

Youcef Bagdad, Marion Sisquellas, Michel Arthur, and Maria A. Miteva*

Cite This: *ACS Omega* 2024, 9, 41334−41342

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Continuous use of antibiotics leads to the ability of bacteria to adapt by developing complex antibiotic resistance (AR) mechanisms. The synthesis of β-lactamases is a widely observed AR mechanism. The class C β-lactamase (AmpC) causes significant resistance toward β-lactam antibiotics, and new treatments are urgently needed. Noncovalent inhibitors have been developed against a broad spectrum of β-lactamases. In this study, we developed robust and accurate models for predicting noncovalent inhibitors of AmpC using large compound data sets and machine/deep learning modeling. We created support vector machine (SVM), random forest (RF), and feed-forward neural network (FFNN) classification models. The cross-validation (CV) accuracies varied between 80 and 82%, as combined models reached an accuracy of 83%. We analyzed the physicochemical characteristics of the noncovalent inhibitors and predicted the binding modes for some of them. Such models are helpful for identifying new noncovalent inhibitors in order to establish novel solutions against the growing resistance to standard β-lactam inhibitors. The best RF, SVM, and FFNN models for predicting noncovalent inhibitors of AmpC β-lactamase are available in the GitHub repository, https://github.com/UPCmctr/ML-DL-AmpC-B-lactamase.

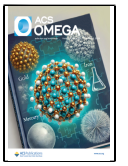
1. INTRODUCTION

Continuous use of antibiotics leads to the ability of bacteria to adapt by developing complex antibiotic resistance (AR) mechanisms. In particular, the production of β-lactamase enzymes is a widely observed AR mechanism. Among the numerous β-lactamases, Ampicillinase C (AmpC) (Figure 1), a class C β-lactamase,[1,2] shows significant resistance toward β-lactam antibiotics, including penicillin, cephalosporins, and carbapenems.[3] AmpC is expressed in gram-negative bacteria such as *Enterobacter cloacae* and *Pseudomonas aeruginosa*, among others.[2,4] Bacteria producing high levels of AmpC β-lactamase via alteration of regulatory genes are constantly increasing.[5]

Identifying new AmpC β-lactamase inhibitors is urgently needed to develop novel effective treatments against the resistant bacteria.[6] Conventional strategies utilize β-lactam inhibitors such as clavulanic acid, sulbactam, and tazobactam,[7] which are efficient for class A β-lactamases.[7,8] However, β-lactamases classes B, C (as AmpC β-lactamase),[8] and D[9,10] are slightly affected by these first-generation β-lactamase inhibitors. Currently, non-β-lactam inhibitors, such as diazabicyclooctane (DBO) derivatives (e.g., avibactam, relebactam) or boronic acid-based inhibitors (e.g., vaborbactam[11,12]), are employed to restore the activity of β-lactams against bacteria producing AmpC β-lactamase. In addition, noncovalent inhibitors are of particular interest as showing efficacy against a broad spectrum of β-lactamases,[13] including AmpC β-lactamases.[14,15]

Virtual screening has been recently used for discovering new noncovalent inhibitors of AmpC β-lactamase.[14,15] Lyu et al. performed large-scale virtual screening against AmpC, synthesized 44 diverse potential hits, and identified a phenolate compound as a novel strong inhibitor of AmpC with IC_{50} of 77 nM[15] (see Figure 1). Recently, machine learning (ML) and deep learning (DL) have also emerged as new strategies for the identification of β-lactamase inhibitors. The application of

Received: April 22, 2024
Revised: June 10, 2024
Accepted: June 13, 2024
Published: September 24, 2024



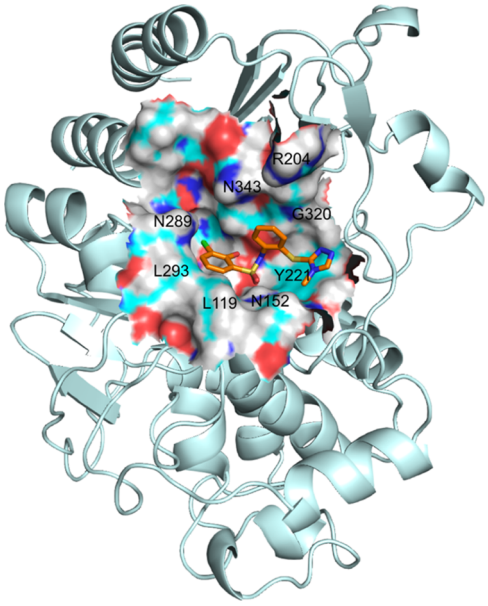


Figure 1. 3D structure of *E. coli's* AmpC β-lactamase cocrystallized with a noncovalent inhibitor ZINC549719643[15] (PDB ID 6DPT). The residues of the active site are shown as a surface in cyan atom type. The inhibitor is in sticks and colored in orange atom type.

ML/DL methodologies has been proven to be successful in drug discovery,[16] e.g., for identifying promising candidates for various therapeutic targets, e.g., enzymes and nuclear receptors, among others.[17−20] Stokes et al. utilized a deep neural network (DNN) to predict antibacterial candidates and discovered halicin showing a broad-spectrum bactericidal activity.[21] Li et al. built predictive models for identifying antibacterial compounds using support vector machines (SVM) and random forests (RF) that suggested numerous FDA-approved drugs as potential antibacterial agents.[22] Hsieh et al. employed QSAR and ML *k*-nearest-neighbors (*k*-NN) modeling for predicting new inhibitors of AmpC β-lactamase based on a small training set of 21 compounds.[23] Five weak noncovalent inhibitors of AmpC were found, the best one showing a K_i value of 135 μM. Recently, Anant et al. developed classification and regression models based on a large data set from ChEMBL using RF, logistic regression (LR), and support vector regression (SVR) to predict the potency of new compounds against AmpC.[24]

In this study, we developed robust and accurate models for predicting noncovalent inhibitors of AmpC using diverse compound data sets and ML (RF, SVM) and DL (up to three hidden layers were tested) modeling. We built RF,[25] SVM,[26] and feed-forward neural network (FFNN)[27,28] models. To enhance the predictive capabilities of our approach, we also combined models integrating the best SVM, RF, and FFNN models. We analyzed the physicochemical characteristics of the used noncovalent inhibitors and predicted the binding modes for some of them. Such models are helpful for identifying new noncovalent inhibitors in order to establish novel solutions against the growing resistance to standard β-lactam inhibitors.

2. MATERIALS AND METHODS

2.1. Data Set Preparation. The data sets were based on the PubChem bioassay ID 585 for the AmpC β-lactamase of *Escherichia coli*, where 70,699 molecules were used for high-throughput quantitative screening (qHTS). Compounds exhibiting an IC_{50} < 28 μM were taken as inhibitors, resulting in 130 inhibitors. The compounds showing <1% inhibition at a concentration of 28 μM were considered noninhibitors, resulting in 68,782 noninhibitors. Focusing here on drug-like compounds, we applied the same "soft" drug-like filtering and curation protocol for both inhibitors and noninhibitors (see the Supporting Information (SI) for details). That resulted in 97 inhibitors and ∼50,000 noninhibitors. For the non-inhibitors, we finally randomly chose 500 molecules to exceed the number of inhibitors by 5 fold. We performed principal component analysis (PCA) of the 500 randomly selected noninhibitors and 50,000 ones (see Figure S1 and details in SI), showing a good distribution of the 500 compounds in the chemical space of the 50,000 noninhibitors. The training set was built by randomly choosing 80% of the data set (68 inhibitors and 350 noninhibitors), while the remaining 20% (29 inhibitors and 150 noninhibitors) was used as the external test set. The lists of inhibitors included in the training and external test sets are provided in the SI.

2.2. Descriptor Calculation. We calculated 354 two-dimensional (2D) and three-dimensional (3D) physicochemical descriptors for the compounds of the training and external data sets using MOE software.[29] Descriptors with zero variance were removed. Descriptors exhibiting a Pearson correlation coefficient of 0.85 or higher were additionally analyzed. For each pair of correlated descriptors, we applied a Student's *t*-test to determine the more discriminative descriptor between inhibitors and noninhibitors. The descriptor showing a higher *p*-value was removed. Finally, 177 descriptors were retained.

2.3. Machine Learning Classification Modeling. We employed SVM and RF algorithms to develop classification models for the prediction of the AmpC inhibitors.

2.3.1. Importance of Descriptors. In the ML modeling, we included the best descriptors ranked based on their relative importance in predicting inhibitors of AmpC. The importance is measured by the Gini impurity index,[30] which is a measure of the probability of incorrectly classifying a randomly selected element in a data set. The Gini impurity index was calculated using an RF approach, constructing multiple decision trees. We built 1000 RF models on the training set using default hyperparameters (given in SI Table S1) and selected the subset of descriptors with the highest Gini importance.

2.3.2. RF Modeling. RF classification models were built using the Random Forest R library[31] of the R software. Multiple decision trees were built with bootstrap samples from the training data. To introduce diversity into the RF ensemble trees, each tree utilizes a random subset of descriptors to make decisions at its nodes. The classification was obtained by taking the results of all of the trees through a majority vote.

We first developed RF models using the default hyperparameters (see SI Table S1). Then, we optimized the RF models. This process was carried out through a search for the best combination of the hyperparameters: *ntree* (number of trees in the forest), *mtry* (number of selected descriptors at each node), and *sampsize*. The *ntree* parameter was explored within the range 128−1024 and *mtry* was tested within the range 5−50. The *sampsize* parameter was tuned for the number of inhibitors ranging from 60 to 80% of the total number of compounds and the number of noninhibitors from 70 to 120% of the inhibitor counts. This parameter is essential to handle imbalanced data to improve the model performance. The optimization of these hyperparameters was conducted by using

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

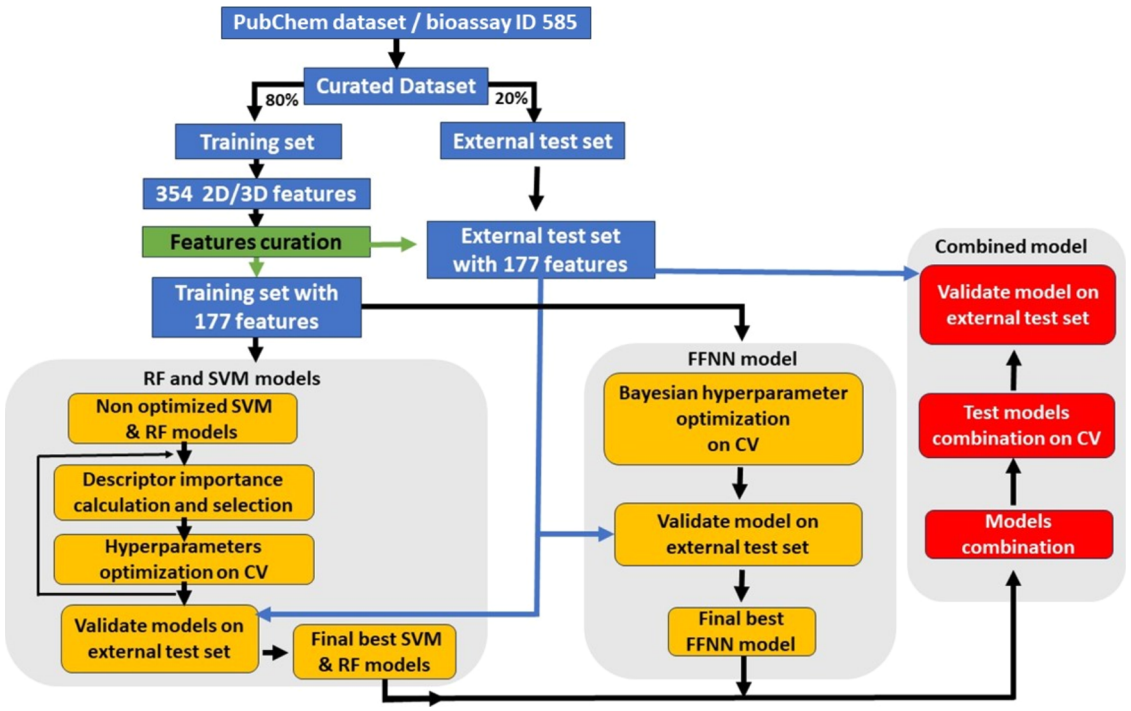


Figure 2. Workflow of the used machine learning and deep learning modeling. FFNN: feed-forward neural network, SVM: support vector machine, RF: random forest, and CV: cross validation.

a grid-based search. The 5-fold cross-validation (CV) procedure was repeated 10 times.

2.3.3. SVM Modeling. SVM methodology is based on minimization concepts in the statistical learning theory[32] and employs various kernel functions to project data into higher-dimensional spaces. This technique is especially effective in instances where data sets are not linearly separable. We employed SVM algorithms implemented in the R package with the Caret library[33] to build our classification models.

The initial SVM models were trained using default hyperparameters (see SI Table S1). The descriptors were centered around a zero mean and scaled to have a variance of one. We used the radial basis function kernel (SVM-Rad). Then, we optimized the hyperparameter to improve the SVM models. The cost parameter was optimized in the range of 2 to 2^{18}, where a higher cost value may lead to a more complex model and overfitting, while a lower value may lead to underfitting. The hyperparameter γ, influencing the decision boundary of the model, was optimized in the range of 2^{-14} to 2°. To address the imbalance of our data set, we also optimized the weight parameter. The optimization of the SVM hyperparameters was also undertaken using a grid-based search and employing a ten-time repeated 5-fold CV.

2.3.4. Additional Runs of Model Optimization. After obtaining the first optimized RF and SVM models, we carried out three successive runs of optimization. This involved recalculating the importance of the descriptors selected in the previous run and tuning the hyperparameters through CV.

2.4. Feed-Forward Neural Network Classification Modeling. Commonly referred to as multilayer perceptron (MLP), FFNNs constitute artificial neural networks of multiple interconnected layers of neurons, facilitating the propagation of information in a unidirectional manner from the input layer to the output layer. We constructed a classification predictive FFNN model using Python version 3.7.16[34] and employed TensorFlow V 2.11.0,[35,36] which incorporates the Keras API[37] for the model implementation.

Our FFNN model is based on a variety of hyperparameters requiring thorough optimization. These include the number of hidden layers, number of neurons per layer, batch size, optimizer type, class weights, and learning rate. To avoid overfitting in our models, we integrated the L2 regularizers in each hidden layer. These regularizers penalize higher weights and maintain the balance of the model. The coefficients associated with these L2 regularizers are also considered as hyperparameters entering the optimization (see SI Table S2). We defined the search space for each hyperparameter for efficient optimization. Each hidden layer within our architecture utilizes the rectified linear unit (ReLU) activation function, while the output layer employs the sigmoid function optimized for binary classification tasks (SI Table S2). To evaluate the performance of our FFNN model, we used a 5-fold cross-validation method, repeated 10 times, in which we calculated the metrics described in the SI.

Compared to RF and SVM modeling, FFNN integrates a greater number of hyperparameters to be optimized. This makes an exhaustive grid search of a map impractical. As a more efficient alternative, we employed Bayesian optimization (BO) using the GpyOpt Python library[38] based on probabilistic modeling for efficiently finding the best combination of hyperparameter values, ensuring the best-performing model. The optimization started by defining the ranges for each hyperparameter present in the initially created FFNN model and distributing random values for each hyperparameter within the defined ranges (SI Table S2). Our BO included 400 successive optimization cycles of Gaussian processes (GP)[39] and expected improvement (EI) acquisition function,[40,41] simplifying the objective function and thus ensuring accelerated finding of an optimized combination of hyperparameters at each optimization cycle.

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

The statistical metrics used to evaluate the predictive capabilities of all models are listed in SI.

2.5. Molecular Docking. To explore possible binding modes for some of the AmpC β-lactamase inhibitors present in our data set, we performed molecular docking on the chain B of the 3D structure of *E. coli* β-lactamase AmpC (PDB ID 4KZ5, resolution 1.35 Å, selected after analysis of the PDB structures 4KZ4, 1C3B, 1FCO, and 3BLS). We performed docking of two compounds using AutoDock Vina 1.1.2.[42] This software employs conformational gradient-based docking coupled with an empirical scoring function to predict the protein−ligand interaction energy in kcal/mol. The 3D conformations of the ligands were generated using CORINA software,[43] and their protonation states at pH 7 were assigned using Pipeline Pilot software v.20.1.[44] The protein and ligand structures were preprocessed with AutoDockTools,[45] the solvent molecules were removed, nonpolar hydrogens were merged, and Gasteiger charges were assigned. A grid box of 26 Å × 24 Å × 26 Å with a spacing of 1 Å and centered in the binding pocket was used for docking. The maximum number of binding modes was set to 20, and a global search exhaustiveness of 10 was utilized. The ligands were flexible, and the protein was kept rigid, except for residue Q120 of the binding pocket displaying different positions in the different crystal structures mentioned above. The docking results were analyzed and visualized using the software PyMol v 2.3.2.[46]

3. RESULTS AND DISCUSSION

In order to develop ML and DL models for predicting noncovalent inhibitors of AmpC β-lactamase, the following procedure was done: (i) we prepared data sets of noncovalent inhibitors and noninhibitors of AmpC β-lactamase; (ii) we calculated MOE physicochemical descriptors; (iii) we performed SVM, RF, and FFNN modeling; and (iv) we validated the models in the external test set (see Figure 2).

3.1. Chemical Space of Data Sets. Following the descriptor selection, 177 descriptors were retained for ML and DL modeling (see Section 2 for details). Principal component analysis (PCA) was performed on the 177 descriptors, and the projection of the training and external test sets is displayed in Figure 3A. The subspace depicted is defined by the first two principal components. Overall, the compounds of the training and test sets covered similar chemical spaces. Interestingly, the noninhibitors of the training and test sets cover the entire chemical space represented by the PCA, while the inhibitors are concentrated in a restricted area, which is a small part of the noninhibitor space (Figure 3B,C). The chemical space of the data sets is limited by the "soft" drug-like filter thresholds (see in Section 2.1), which define the applicability domain of our models.

3.2. Performance of ML Models for Predicting Inhibitors of AmpC. First, we developed preliminary RF and SVM models using the retained 177 MOE descriptors without optimizing the model parameters (the default parameters are listed in SI Table S1). These models displayed poor performances (shown in SI Tables S3 and S4). Very low sensitivity and an overfitting of specificity were obtained on the internal CV and external test sets. Such performance disparity arises from the imbalanced number of inhibitors and noninhibitors. Thus, we performed further optimization of the hyperparameters.

First optimization for both RF and SVM models: After ranking the 177 descriptors based on their importance (see

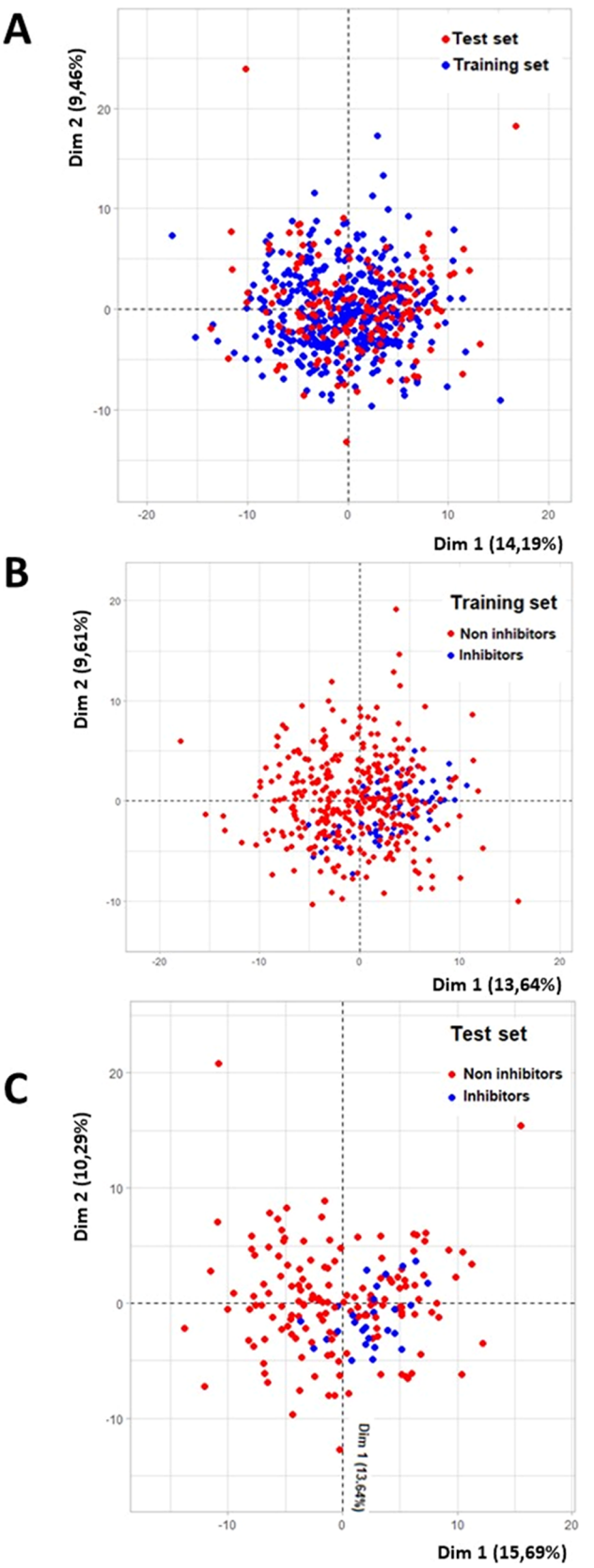


Figure 3. Chemical space of the training and external test sets is represented by the principal component analysis (PCA). (A) PCA of training versus test sets. (B) PCA of the inhibitors versus noninhibitors of the training set. (C) PCA of the inhibitors versus noninhibitors of the test set.

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

Table 1. Performance of the RF and SVM Models after the First Optimization

60 descriptors/1st optimization	method	balanced accuracy (%)	sensitivity (%)	specificity (%)	MCC (%)	AUC (%)
internal CV	RF	79	78	79	50	79
	SVM	78	79	77	45	78
external test set	RF	84	93	75	52	84
	SVM	82	86	77	49	82

Table 2. Performance of the RF and SVM Models after the Second Optimization

24 descriptors/2nd optimization	method	balanced accuracy (%)	sensitivity (%)	specificity (%)	MCC (%)	AUC (%)
internal CV	RF	79	78	80	48	79
	SVM	81	83	79	50	80
external test set	RF	85	93	77	57	87
	SVM	85	86	84	58	85

Table 3. Performance of the RF and SVM Models after the Third Optimization

12 descriptors/3rd optimization	method	balanced accuracy (%)	sensitivity (%)	specificity (%)	MCC (%)	AUC (%)
internal CV	RF	80	79	80	48	80
	SVM	81	82	80	50	81
external test set	RF	86	93	79	57	86
	SVM	86	90	83	59	86

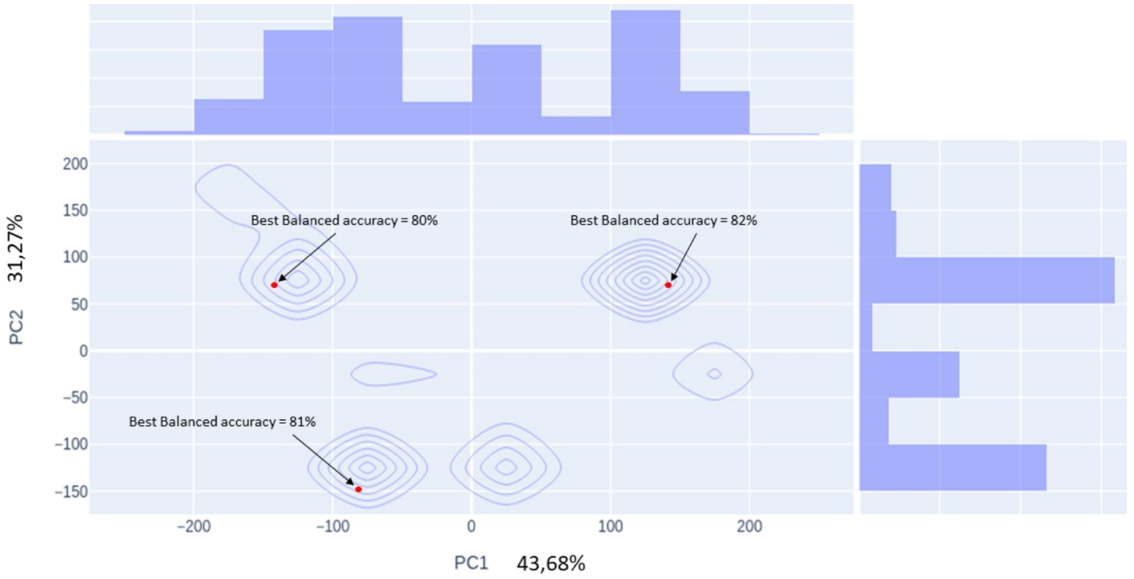


Figure 4. Density map and marginal histograms of the hyperparameter search in Bayesian optimization. The PC1 and PC2 represent the first and second principal components, respectively. The intensified research areas are marked by tight contours with three principal zones. The optimal hyperparameter combinations with the highest balanced accuracies in CV are indicated by red dots for each zone of density. Marginal histograms on top and right represent trial distribution across PC1 and PC2, respectively.

Section 2.3), we selected an importance threshold of 0.7, resulting in 60 descriptors (see SI Figure S2). The first optimization performed with the 60 descriptors allowed us to build better performing models. Notably, we achieved a significant improvement in balancing the sensitivity and specificity for both RF and SVM models (see Table 1).

Then, we proceeded with a second optimization where the importance of the 60 descriptors was recalculated using the previously optimized RF model. This process led to a reranking of the descriptors, choosing a new importance threshold of 0.5 (see SI Figure S3) and a reduced number of descriptors of 24. The newly optimized models, including 24 descriptors, improved the predictive performance (see Table 2) when compared to the first optimized models with 60 descriptors, in particular, the SVM models.

For the third optimization, the re-evaluation of the 24 descriptors' importance revealed that the importance of the descriptors ranked after the best 12 descriptors did not change significantly (see SI Figure S4). Consequently, the best 12 descriptors were retained for further modeling. The performance of the best predictive RF and SVM models is shown in Table 3.

The models built with 12 descriptors displayed a similar performance to those with 24 descriptors, slightly enhancing the SVM sensitivity on the external test set. In fact, the RF and SVM models optimized with 60, 24, and 12 descriptors show similar performance with a small improvement after each optimization. The models obtained after a fourth optimization with 6 descriptors (see SI Figure S5 and Tables S3 and S4) showed slightly worse performance. Taking into account that

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

the models based on 60 descriptors are more complex than the others, we considered the RF and SVM models optimized with 24 or 12 descriptors as the best ML predictive models developed.

3.3. Performance of FFNN Models for Predicting Inhibitors of AmpC. FFNN integrates a large number of hyperparameters that need to be optimized. We employed Bayesian optimization (BO), which is a powerful optimization technique based on probabilistic modeling for efficient exploration of the hyperparameter space in order to find the combination of hyperparameter values that would maximize the performance of the FFNN model.[47]

Figure 4 shows the outcomes of the BO applied to the hyperparameter search. In this density map, the first principal component (PC1) and the second one (PC2) account for approximately 75% of the total variance. The areas of intensified research marked on the map indicate a high concentration of hyperparameter trials. Notably, there are three main research zones, each corresponding to a set of conditions that were tested more frequently. The best balanced accuracies are denoted by red dots, highlighting the hyperparameter sets that exhibited the highest performance during the CV. These balanced accuracy values show slight variations, ranging from 80 to 82%, suggesting that the optimization identified several closely performing solutions. The area with the highest research density, encompassing the hyperparameter values that produced the top-performing model with a balanced accuracy of 82%, suggested that Bayesian optimization precisely targeted the most effective hyperparameters. This observation supports not only the relevance of this technique for refining model performance but also its efficiency in terms of computational time. These results confirm the successful applicability of BO as an appropriate method to efficiently tune hyperparameters for FFNN modeling. Among the developed FFNN models, the neural network with three hidden layers exhibited the best performances in CV (see Table 4 and SI Table S2 for the optimal hyperparameter values).

Table 4. Performance of the Best FFNN Model

177 descriptors/ FFNN	balanced accuracy (%)	sensitivity (%)	specificity (%)	MCC (%)	AUC (%)
internal CV	82	81	83	56	88
external test set	88	93	83	61	91

The developed FFNN and the optimized RF and SVM models (with 24 or 12 descriptors) show different performances both in CV and external testing. In CV, the FFNN outperformed the best models RF and SVM by achieving an AUC of 88% versus ~80% (see Tables 23−4). The same tendency is observed on the external test set. Interestingly, the three methods show equilibrated sensitivity and specificity. On the external test set, RF achieved a remarkable sensitivity of 93% but a lower specificity of 79%. SVM exhibited an excellent sensitivity of 90% and specificity of 83%. FFNN demonstrated excellent sensitivity of 93% and specificity of 83%, thus a better balance between the ability to detect true positives and true negatives. Validation of the best FFNN and the optimized RF, SVM models with 12 descriptors was also performed on an additional 1000 randomly chosen noninhibitors, confirming the excellent predictivity of these models (see Table S5 and Figure S6 in the SI).

In this study, we compared the performance of predictive models, including various numbers of descriptors. The SVM and RF models performed well with a set of 24 or 12 descriptors, whereas the FFNN model included a much broader spectrum of 177 descriptors. Although the optimized ML models (SVM and RF) and DL model (FFNN) demonstrated excellent predictive performance, each of them exhibits its own specificities. Figure 5 illustrates the chemical space of the entire external test set and the five inhibitors badly classified by the best RF, SVM, and FFNN models. The inhibitors CID 5309621 and 3601182, accurately classified by the SVM and RF models, were misclassified by the FFNN model. The FFNN model correctly identified the compound CID 5307628 as an inhibitor, incorrectly classified by the SVM and RF models. Interestingly, CID 5307628 is placed in the chemical space, mostly covered by the noninhibitors. Therefore, one may speculate that FFNN might better discriminate between positive and negative data when they share the same chemical space, being capable of detecting minor differences between the compounds.

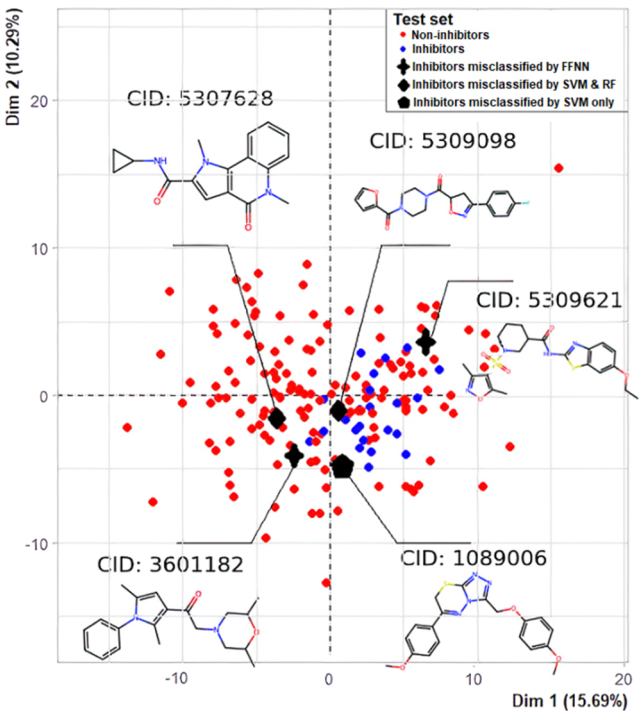


Figure 5. Chemical space of the external test set is represented by the principal component analysis (PCA). The inhibitors misclassified by the three models are displayed.

A combination of the three predictive methods (RF, SVM, and FFNN) could be a strategy to identify as many inhibitors as possible, but with the risk of decreasing the specificity. A combined model may consider a compound to be classified as an inhibitor if it is predicted as such by at least one of the three models. Such a model, combining the FFNN and the best SVM- and RF-optimized models with 12 descriptors, was evaluated (see SI Table S6). As expected, a significant improvement in the sensitivity was achieved (90% in CV), with a reduction of the specificity to 76% (in CV) due to the increase in the number of false positives. The application of the

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

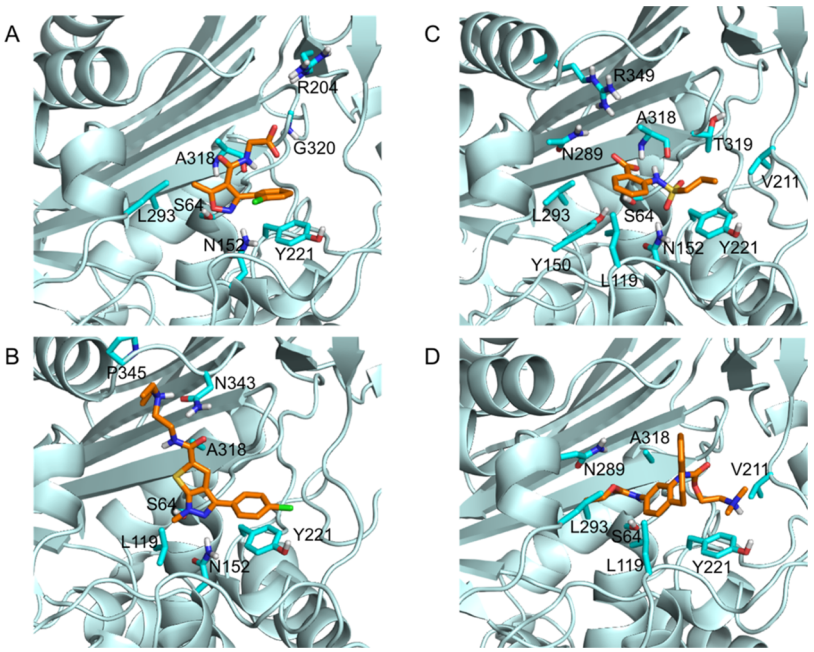


Figure 6. Binding poses of noncovalent inhibitors of AmpC β-lactamase. The active site residues interacting with the compounds are colored in cyan atom type. The inhibitors are colored in an orange atom type. (A) CID426788 cocrystallized with AmpC β-lactamase (PDB ID 4KZ5). (B) The top-ranked pose of CID5309397 docked into AmpC β-lactamase. (C) CID16770253 cocrystallized with AmpC β-lactamase (PDB ID 4KZ4). and (D) The top-ranked pose of CID6603584 docked into AmpC β-lactamase.

combined model on the external test set demonstrated a sensitivity of 100% and a specificity of 72%.

We can conclude that we obtained high-quality models predicting noncovalent inhibitors of AmpC with the RF, SVM, FFNN, and combined techniques. Our models largely outperform others recently reported by Anant et al.[24] In their work, the authors employed RF, LR, and SVR based on a large data set of ChEMBL containing >60 K compounds to create models for predicting the potency of compounds inhibiting AmpC. The best AUC achieved for their classification models was around 0.5. The reason for such unsatisfactory performance could be that their data set comprised <1% of inhibitors without undertaking any balancing procedures.

3.4. Interactions of AmpC β-Lactamase and Noncovalent Inhibitors. Experimental structures of AmpC with a bound inhibitor of our data set have not been reported by now. In order to predict binding modes for some of these inhibitors, we performed docking and compared the poses to the experimental binding modes of other noncovalent inhibitors of AmpC β-lactamase identified by NMR and SPR fragment-based screening.[48] Barelier et al. reported the crystal structure of the AmpC β-lactamase cocrystallized with two noncovalent fragments, CID426788 and CID16770253, showing strong inhibition (PDB IDs 4KZ5 and 4KZ4). By calculating the Tanimoto coefficient, we found that the inhibitor CID5309397 of our data set had a structural similarity with the cocrystallized inhibitor CID426788 of 44%, and the inhibitor CID6603584 of our data set had a structural similarity of 39% with the cocrystallized compound CID16770253. We focus here on the inhibitors similar to the cocrystallized ones to only analyze the docking poses because there is no clear correlation between the calculated docking scores and the similarity between the inhibitors and noninhibitors (see Figures S7 and S8 in SI).

Figure 6 shows the comparison of the experimental binding modes of the two cocrystallized inhibitors and the best ranked docking pose of the two inhibitors of our data set. The docking results suggest similar p-p staking between the halobenzene and Y221 for the two compounds CID426788 and CID5309397 and hydrogen bonding with S64 and N152. Similar compounds CID16770253 and CID6603584 are stabilized by hydrophobic contacts with Y221, L119, L293, and V211 and hydrogen bonding with S64 and N289. The similar interactions found by our docking results for the two compounds of our data set and those observed for the cocrystallized fragments reinforce the hypothesis for a similar inhibitory mechanism against the AmpC β-lactamase. The presence of several hydrophobic residues in the active site, L119, Y150, V211, Y221, and L293, and several hydrophilic residues, S64, N152, N289, and N343, shows the mixed nature of the pocket in terms of polarity. Notably, the two experimental binding modes show electrostatic interactions of the carboxylic groups of CID426788 and CID16770253 with the two important arginine residues R204 and R349, respectively, missing in the docked poses. This provides a possibility for further optimization of the noncovalent inhibitors.

We also performed an analysis of the inhibitors and noninhibitors of AmpC β-lactamase of our data set based on the best 24 descriptors derived from the second optimization. The differences in the most important 24 physicochemical descriptors between inhibitors and noninhibitors are illustrated by 24 boxplots in SI Figures S5 and S6. Notably, among the top 12 descriptors ranked after the first optimization (with 60 descriptors) and the second optimization (with 24 descriptors), 11 are the same, and among the top 12 descriptors ranked after the second optimization (24 descriptors) and the third optimization (12 descriptors), seven are the same,

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342

showing the stability of the optimized models after the second and third optimizations.

Four out of these seven descriptors are associated with the compound hydrophobicity: vsa_hyd, vsurf_d5, logP, PEOE_VSA-0, and one, vsurf_CW1 (SI Table S7). In the same line, GCUT_PEOE_0 and logS are lower for the inhibitors. Overall, the inhibitors have the tendency to be more hydrophobic than the noninhibitors. This is in agreement with the several hydrophobic residues of the binding pocket (Figures 1 and 6) (L119, Y150, V211, Y221, and L293) clustered in one area of the pocket.

Interestingly, two out of the seven most important descriptors are associated with the compound stability, E_ang and BCUT_PEOE_0, showing that the inhibitors are more stable than the noninhibitors (see SI Table S7 and Figures S9 and S10).

3.5. Conclusions. In this study, we developed high-quality predictive models for identifying noncovalent inhibitors of the AmpC β-lactamase enzyme using RF, SVM, and FFNN algorithms with cross-validation accuracies between 80 and 82%. FFNN slightly outperformed RF and SVM and was particularly effective in managing a large number of molecular descriptors without introducing noise or degrading performance. The created models can be used for screening large compound libraries for identifying new potential noncovalent inhibitors of AmpC β-lactamase for future development and optimization. The performed docking analyses permitted the prediction of the binding modes of two compounds, showing partial structural similarities with noncovalent inhibitors with known experimental binding modes, suggesting a similar inhibition mechanism and binding in the active site of AmpC. Our study can be helpful in establishing novel solutions against the growing resistance to standard β-lactam inhibitors.

■ ASSOCIATED CONTENT

Data Availability Statement

The best RF, SVM, and FFNN models for predicting noncovalent inhibitors of AmpC β-lactamase are available in the GitHub repository, https://github.com/UPCmctr/ML-DL-AmpC-B-lactamase.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c03834.

> Data sets preparation; PubChem CIDs of the training and test sets inhibitors; model performance metrics; hyperparameters of models; performance of models; descriptor importance; PCAs; descriptor, and docking score (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Maria A. Miteva − *Université Paris Cité, CNRS UMR 8038 CiTCoM, Inserm U1268 MCTR, 75006 Paris, France;* orcid.org/0000-0001-6895-1214; Email: maria.mitev@inserm.fr

Authors

Youcef Bagdad − *Université Paris Cité, CNRS UMR 8038 CiTCoM, Inserm U1268 MCTR, 75006 Paris, France*

Marion Sisquellas − *Université Paris Cité, CNRS UMR 8038 CiTCoM, Inserm U1268 MCTR, 75006 Paris, France*

Michel Arthur − *INSERM, Sorbonne Université, Université Paris Cité, Centre de Recherche des Cordeliers (CRC), 75006 Paris, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c03834

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the French ANR (project NASPEC), Université Paris Cité, and INSERM Institute for supporting this research. They also thank Prof. M. Ethève-Quelquejeu and Dr. L. Iannazzo (Université Paris Cité, CNRS) and Dr. K. Cariou (CNRS Chimie ParisTech, PSL University) for the helpful discussions.

■ REFERENCES

(1) Ambler, R. P. The Structure of Beta-Lactamases. *Philos. Trans. R. Soc. London, Ser. B* **1980**, *289* (1036), 321−331.

(2) Jacoby, G. A. AmpC β-Lactamases. *Clin. Microbiol. Rev.* **2009**, *22* (1), 161−182.

(3) Moya, B.; Dötsch, A.; Juan, C.; Blázquez, J.; Zamorano, L.; Haussler, S.; Oliver, A. β-Lactam Resistance Response Triggered by Inactivation of a Nonessential Penicillin-Binding Protein. *PLoS Pathog.* **2009**, *5* (3), No. e1000353.

(4) Llvermore, D. M.; Williams, R. J.; Llndridge, M. A.; Slack, R. C. B.; Williams, J. D. Pseudomonas Aeruginosa Isolates with Modified Beta-Lactamase Inducibility: Effects of Beta-Lactam Sensitivity. *Lancet* **1982**, *1*, 1466−1467.

(5) Tamma, P. D.; Doi, Y.; Bonomo, R. A.; Johnson, J. K.; Simner, P. J.; Antibacterial Resistance Leadership Group; et al. A Primer on AmpC β-Lactamases: Necessary Knowledge for an Increasingly Multidrug-Resistant World. *Clin. Infect. Dis.* **2019**, *69* (8), 1446−1455.

(6) Vena, A.; Castaldo, N.; Bassetti, M. The Role of New β-Lactamase Inhibitors in Gram-Negative Infections. *Curr. Opin. Infect. Dis.* **2019**, *32* (6), 638−646.

(7) Chen, J.; Shang, X.; Hu, F.; Lao, X.; Gao, X.; Zheng, H.; Yao, W. β-Lactamase Inhibitors: An Update. *Mini-Rev. Med. Chem.* **2013**, *13* (13), 1846−1861.

(8) Drawz, S. M.; Bonomo, R. A. Three Decades of Beta-Lactamase Inhibitors. *Clin. Microbiol. Rev.* **2010**, *23* (1), 160−201.

(9) Georgopapadakou, N. H. Beta-Lactamase Inhibitors: Evolving Compounds for Evolving Resistance Targets. *Expert Opin. Invest. Drugs* **2004**, *13* (10), 1307−1318.

(10) Watkins, R. R.; Papp-Wallace, K. M.; Drawz, S. M.; Bonomo, R. A. Novel β-Lactamase Inhibitors: A Therapeutic Hope against the Scourge of Multidrug Resistance. *Front. Microbiol.* **2013**, *4*, No. 392.

(11) Langley, G. W.; Cain, R.; Tyrrell, J. M.; Hinchliffe, P.; Calvopiña, K.; Tooke, C. L.; Widlake, E.; Dowson, C. G.; Spencer, J.; Walsh, T. R.; Schofield, C. J.; Brem, J. Profiling Interactions of Vaborbactam with Metallo-β-Lactamases. *Bioorg. Med. Chem. Lett.* **2019**, *29* (15), 1981−1984.

(12) Messner, K.; Vuong, B.; Tranmer, G. K. The Boron Advantage: The Evolution and Diversification of Boron's Applications in Medicinal Chemistry. *Pharmaceuticals* **2022**, *15* (3), No. 264.

(13) Romero, E.; Oueslati, S.; Benchekroun, M.; D'Hollander, A. C. A.; Ventre, S.; Vijayakumar, K.; Minard, C.; Exilie, C.; Tlili, L.; Retailleau, P.; Zavala, A.; Elisée, E.; Selwa, E.; Nguyen, L. A.; Pruvost, A.; Naas, T.; Iorga, B. I.; Dodd, R. H.; Cariou, K. Azetidinimines as a Novel Series of Non-Covalent Broad-Spectrum Inhibitors of β-Lactamases with Submicromolar Activities against Carbapenemases KPC-2 (Class A), NDM-1 (Class B) and OXA-48 (Class D). *Eur. J. Med. Chem.* **2021**, *219*, No. 113418.

(14) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-Based Discovery of a Novel, Noncovalent Inhibitor of AmpC Beta-Lactamase. *Structure* **2002**, *10* (7), 1013−1023.
(15) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224−229.
(16) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discovery* **2020**, *19* (5), 353−364.
(17) Xia, Z.; Wu, L.-Y.; Zhou, X.; Wong, S. T. Semi-Supervised Drug-Protein Interaction Prediction from Heterogeneous Biological Spaces. *BMC Syst. Biol.* **2010**, *4*, No. S6.
(18) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038−1040.
(19) Goldwaser, E.; Laurent, C.; Lagarde, N.; Fabrega, S.; Nay, L.; Villoutreix, B. O.; Jelsch, C.; Nicot, A. B.; Loriot, M.-A.; Miteva, M. A. Machine Learning-Driven Identification of Drugs Inhibiting Cytochrome P450 2C9. *PLoS Comput. Biol.* **2022**, *18* (1), No. e1009820.
(20) Dudas, B.; Bagdad, Y.; Picard, M.; Perahia, D.; Miteva, M. A. Machine Learning and Structure-Based Modeling for the Prediction of UDP-Glucuronosyltransferase Inhibition. *iScience* **2022**, *25* (11), No. 105290.
(21) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688−702.e13.
(22) Li, W.-X.; Tong, X.; Yang, P.-P.; Zheng, Y.; Liang, J.-H.; Li, G.-H.; Liu, D.; Guan, D.-G.; Dai, S.-X. Screening of Antibacterial Compounds with Novel Structure from the FDA Approved Drugs Using Machine Learning Methods. *Aging* **2022**, *14* (3), 1448−1472.
(23) Hsieh, J.-H.; Wang, X. S.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC Beta-Lactamase Binders vs. Decoys Using Classification kNN QSAR Modeling and Application of the QSAR Classifier to Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *22* (9), 593−609.
(24) Anant, P. S.; Gupta, P. Application of Machine Learning in Understanding Bioactivity of Beta-Lactamase AmpC. *J. Phys.: Conf. Ser.* **2022**, *2273* (1), No. 012005.
(25) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(26) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Appl.* **1998**, *13* (4), 18−28.
(27) Bebis, G.; Georgiopoulos, M. Feed-Forward Neural Networks. *IEEE Potentials* **1994**, *13* (4), 27−31.
(28) Leiva, M. U.; Arredondo, T. V.; Candel, D. C.; Dombrovskaia, L.; Agullo, L. C.; Seeger, M. P.; Vasquez, F. M. In *Feed-Forward Artificial Neural Network Based Inference System Applied in Bioinformatics Data-Mining*, International Joint Conference on Neural Networks; IEEE, 2009; pp 1744−1749.
(29) Molecular Operating Environment (MOE) | MOEsaic | PSILO, 2022 https://www.chemcomp.com/Products.htm. (accessed May 17, 2022).
(30) Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*; John Wiley & Sons, 2011.
(31) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.
(32) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273−297.
(33) Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Software* **2008**, *28*, 1−26.
(34) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
(35) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X. In *TensorFlow: A System for Large-Scale Machine Learning*, 12th USENIX Symposium on Operating Systems Design and Implementation; USENIX, 2016.
(36) TensorFlow Developers, 2022. DOI: 10.5281/zenodo.4758419.
(37) Gulli, A.; Pal, S. *Deep Learning with Keras: Implement Neural Networks with Keras on Theano and TensorFlow*; Packt Publishing, 2017.
(38) GPyOpt, 2023. https://github.com/SheffieldML/GPyOpt. (accessed June 07, 2023).
(39) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning; MIT Press: Cambridge, Mass, 2008.
(40) Wilson, J.; Hutter, F.; Deisenroth, M. In *Maximizing Acquisition Functions for Bayesian Optimization*, Advances in Neural Information Processing Aystems; Curran Associates, Inc., 2018.
(41) Christianson, R. B.; Gramacy, R. B. Robust Expected Improvement for Bayesian Optimization. *IISE Trans.* **2023**, 1−13, DOI: 10.1080/24725854.2023.2275166.
(42) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455−461.
(43) Release of 3D Structure Generator CORINA Classic Version 4.3.0 − MN-AM, 2023 https://mn-am.com/product-updates/release-of-3d-structure-generator-corina-classic-version-4-3-0/. (accessed August 09, 2023).
(44) Pipeline Pilot - BIOVIA - Dassault Systèmes, 2023. https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/. (accessed October 26, 2023).
(45) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785−2791.
(46) *The PyMOL Molecular Graphics System*, Version 2.0; Schrödinger, LLC.
(47) Wang, X.; Jin, Y.; Schmitt, S.; Olhofer, M. Recent Advances in Bayesian Optimization. *ACM Comput. Surv.* **2023**, *55* (13s), 1−36.
(48) Barelier, S.; Eidam, O.; Fish, I.; Hollander, J.; Figaroa, F.; Nachane, R.; Irwin, J. J.; Shoichet, B. K.; Siegal, G. Increasing Chemical Space Coverage by Combining Empirical and Computational Fragment Screens. *ACS Chem. Biol.* **2014**, *9* (7), 1528−1535.

https://doi.org/10.1021/acsomega.4c03834
ACS Omega 2024, 9, 41334−41342