Screening of Antibacterial Compounds With Novel Structure From The FDA Approved Drugs Using Machine Learning Methods

Wen-Xing Li
Southern Medical University https://orcid.org/0000-0001-9984-8439
Xin Tong
Kunming University of Science and Technology
Peng-Peng Yang
Kunming University of Science and Technology
Yang Zheng
Kunming University of Science and Technology
Ji-Hao Liang
Kunming University of Science and Technology
Gong-Hua Li
Kunming Institute of Zoology Chinese Academy of Sciences https://orcid.org/0000-0002-9311-6613
Dao-Gang Guan
Southern Medical University https://orcid.org/0000-0002-1479-9808
Shao-Xing Dai (daisx@lpbr.cn)
Kunming University of Science and Technology https://orcid.org/0000-0002-6919-2034

Research article

Keywords: antibacterial compound, drug repositioning, machine learning, structural similarity, virtual screening

Posted Date: October 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-951331/v1

Version of Record: A version of this preprint was published at Aging on February 12th, 2022. See the published version at https://doi.org/10.18632/aging.203887.

Abstract

Background: Due to the lack of new antibiotics in recent years, bacterial resistance has increasingly become a serious problem globally. The aim of this study is to construct an antibacterial compound predictor using machine learning methods to screen potential antibacterial drugs.

Methods: Active and inactive antibacterial compounds were acquired from the ChEMBL and PubChem database, which were used to construct benchmark datasets. The antibacterial compound predictor is constructed using the support vector machine (SVM), random forest (RF), and multi-layer perception (MLP) methods. We predicted the antibacterial activity of the Food and Drug Administration (FDA) approved drugs in the DrugBank database and screened novel antibacterial drugs through structural similarity analysis.

Results: In the initial screen process, the results suggested that the benchmark dataset based on FP2 molecular fingerprints, along with the SVM, RF, and MLP methods showed excellent prediction performance (mean AUC > 0.9 for all models). Using the combination of these three models, a total of 957 potential antibacterial drugs were predicted. Most of the predicted drugs showed low structural similarities compared with the FDA approved antibacterial drugs. We finally screened 9 predicted antibacterial drugs with novel structures including 2 anti-tumor drugs (cyclophosphamide and ifosfamide), 2 ophthalmic drugs (apraclonidine and echothiophate) and 5 anesthetics (desflurane, enflurane, isoflurane, methoxyflurane, and sevoflurane).

Conclusions: This study provides a new insight for predicting antibacterial compounds with novel structures by using FDA approved drugs. The predicted compounds with novel structures are worthy of further experimental verification in the future.

1. Background

Antibiotics are a fast and effective way to deal with bacterial infections. However, with the widespread use of antibiotics, bacteria are also constantly evolving and a large number of pathogens have emerged that can resist these drugs [1]. As the research and development of novel antibiotics by pharmaceutical companies has drastically decreased in recent years, bacterial resistance has increasingly become a serious problem [2]. Therefore, the development of novel and highly efficient antibiotics is an urgent issue. High-throughput screening has been the dominant approach of antimicrobial drug development in the industry in the past few decades [3, 4]. However, due to the long development time, huge cost, and low efficiency of this method [5], computer-aided drug design techniques have become a promising method in the discovery of novel antibacterial drugs [6].

In recent years, machine learning methods have shown tremendous potential in the process of drug discovery and development [7]. Multiple machine learning-based methods effectively improved the accuracy of drug-target interaction prediction [8]. Especially in the early phases of drug discovery, the use of machine learning methods significantly reduces time and effort in drug discovery and development [9].

In other areas of drug discovery, deep learning is a promising method for the prediction of molecular properties and the de novo generation of suggestions for new molecules [10]. These technologies may have fundamentally changed the process of identifying new molecules and/or repurposing old drugs [11].

Multiple machine learning methods are widely used in ligand-based and receptor-based antibacterial drug discovery [6, 12–17]. By using chemoinformatics methods to extract the molecular characteristics of short peptides, studies have shown that support vector machine (SVM) can accurately predict the antibacterial activity of short peptides [12, 13] and the genetic characteristics of antibiotic resistance in specific pathogens [18]. The combination using random forest (RF) and genome-based analysis approaches promoted phenotypic antibacterial drug discovery [14] and revealing potential antibiotic resistance genes [15]. In recent years, emerging deep neural network methods have facilitated the discovery of antibacterial molecules with unique structures from massive data [16]. Furthermore, due to the limitations of a single method, the combination using multiple machine learning methods showed excellent performance in antibacterial compounds discovery [16, 17] and predicting the bacterial genetic mutations on drug resistance [19].

Although the popularization of machine learning methods has greatly shortened the discovery of antibacterial lead compounds, there are still require long-term studies from the identified lead compounds to clinical applications, especially experiment on drug safety [20]. Therefore, a new use for old drugs may be a way to resolve current antibiotic resistance [21]. The current Food and Drug Administration (FDA) approved antibiotics can be divided into multiple categories according to the core scaffolds, and a variety of semi-synthetic antibiotics are based on these scaffolds [22]. Due to the increased bacterial resistance to specific scaffold structures, it is a promising way to develop antibiotics with novel structures. In this study, we combined using multiple machine learning methods to build the antibacterial compound predictor, and identified structure novel small-molecule antibacterial compounds from the FDA approved drugs.

2. Methods

2.1 Antibacterial compounds collection

Compounds that performed antimicrobial activity test were collected from ChEMBLdb (version 25, https://www.ebi.ac.uk/chembl/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/) databases. A total of 83768 compounds were obtained, 8001 of these compounds has a clear IC50 value, and others only have an inactive label. The IC50 cutoff value of antibacterial activity was defined by curve fitting the IC50 values of all compounds. Compounds with IC50 less than 10 μmol/L were generally considered as active antibacterial compounds [23–27], the curve fitting results also suggest that this cutoff is reasonable (Supplementary Figure 1). Based on the curve fitting results, compounds with IC50 higher than 10 μmol/L were considered as inactive antibacterial compounds. Pybel, a python wrapper of OpenBabel [28, 29] was used to access the SMILES string of compounds and calculate molecular fingerprint which represents the presence or absence of particular substructures in the molecule. Multiple types of molecular fingerprints

of all compounds were calculated. Benchmark datasets were build based on the following steps: (1) remove duplicate compounds; (2) remove compounds with a molecular weight greater than 1000; (3) remove compounds with molecular fingerprint similarities higher than 0.9 between the active and inactive antibacterial compounds. Finally, we got a positive dataset including 2708 active antibacterial compounds and a negative dataset including 78620 inactive antibacterial compounds. All active antibacterial compounds have IC50 values whereas only 1893 inactive antibacterial compounds have IC50 values.

2.2 Construction of the benchmark dataset

There is a large difference in the number of compounds between the positive and negative datasets. The positive dataset contains 2708 active antibacterial compounds. To balance the number of compounds between the positive and negative datasets, the filtered negative dataset contains 1893 inactive antibacterial compounds with IC50 values, the remaining quantity difference was random selected from the inactive antibacterial compounds only with an inactive label. Considering the uncertainty of random selection, we repeated 10 times for negative dataset extract. Therefore, the filtered datasets including one positive dataset and 10 negative datasets, each negative data set is combined with the positive data set for subsequent analysis. Next, the molecular fingerprint is calculated for the positive dataset and all repeated negative datasets. The following types of molecular fingerprints were calculated including FP2, FP3, FP4, DLFP, MACCS, ECFP2, ECFP4, ECFP6, FCFP2, FCFP4, and FCFP6. Several start-of-the-art chemoinformatics approaches were also calculated such as mol2vec [30], SMILES2Vec [31], and FP2VEC [32]. The features of each compound were presented by the binary bits of the different types of molecular fingerprints or vectors and these features were used for machine learning modeling (Supplementary Table 1). All these benchmark datasets were used for preliminary screening of applicable machine learning models.

2.3 Initial screening of machine learning methods

In order to choose the appropriate machine learning methods to construct the anti-bacterial compound prediction model, we evaluate the predictive performance of different machine learning methods including k-nearest neighbor (KNN), logistic regression (LR), linear support vector classifier (LSVC), random forest (RF), gradient boosting regression tree (GBRT), support vector machine (SVM), and multi-layer perception (MLP). In the initial screening process, each machine learning method used benchmark datasets constructed from different molecular fingerprints for training and prediction with default parameters. The benchmark dataset was split to the training set (accounting for 80%) and the validation set (accounting for 20%), and then performed a 5-fold cross-validation test. The results suggested that the benchmark dataset based on FP2 molecular fingerprints, along with the SVM, RF, and MLP methods showed excellent prediction accuracy among all machine learning methods and molecular fingerprints combinations, whereas the accuracy fluctuates greatly among different machine learning methods in the benchmark dataset based on vector features (Supplementary Figure 2). Therefore, the benchmark dataset based on FP2 molecular fingerprints, and the RF, SVM and MLP methods were selected in the subsequent analysis.

2.4 Parameter selection of the SVM, RF, and MLP models

The SVM, RF, and MLP models for antibacterial compounds prediction were built using the svm, ensemble, and neural_network module in the scikit-learn Python library (version: 0.20.0, https://scikit-learn.org/stable/). A parameter grid search strategy was used to choose the optimal parameter "gamma" for the kernel function and regularization parameter "C" for the SVM model, the optimal number of trees (parameter "n_estimators") for the RF model, and the optimal hidden layer sizes and alpha for the MLP model. The other parameters of the above three models use default values. The benchmark dataset was randomly split to the training and validation set (accounting for 80%) and the test set (accounting for 20%) using the train_test_split function in the scikit-learn. The 5-fold cross-validation method was used to evaluate the generalization performance of the model with specified parameters in the training and validation set. The cross-validation accuracy was calculated for model evaluation. After cross-validation, a temporary model was built using the training and validation set and calculated the area under the curve (AUC) for the receiver operating characteristic (ROC) curve in the test set. Considering that there may be similar compounds in the split datasets, dataset split and cross-validation were repeated 10 times, which may reduce the impact of similar compounds on the prediction performance of these models. For each given parameter, the mean cross-validation accuracy and mean AUC was calculated. The optimal model was selected by comparing the maximum mean cross-validation accuracy under different parameters. If there were multiple models with the same mean accuracy, the model with the maximum AUC was considered to be the optimal model.

2.5 Performance evaluation

The optimal SVM, RF, and MLP models were used for performance evaluation. The confusion matrix was calculated using the results of the optimal cross-validation test. The true positive (TP) indicates the number of correctly predicted active antibacterial compounds, the true negative (TN) indicates the number of correctly predicted inactive antibacterial compounds, the false positive (FP) indicates the number of inactive antibacterial compounds predicted as active antibacterial compounds, and the false negative (FN) indicates the number of active antibacterial compounds predicted as inactive antibacterial compounds. We calculated the following quality indices: accuracy = (TP + TN) / (TP + TN + FP + FN), precision = TP / (TP + FP), sensitivity = TP / (TP + FN), specificity = TN / (TN + FP), and F1 score = 2 * TP / (2 * TP + FP + FN). Mean squared error (MSE) was calculated for all three models. Because the filtered datasets including one positive dataset and 10 negative datasets, the average of 10 calculations of these quality indices and AUC were used to evaluate the SVM, RF, and MLP model performance. A model with high scores (≥ 0.8) of accuracy, precision, F1 score, and AUC was considered to be an effective model.

2.6 Antibacterial small-molecule drugs prediction

The final SVM, RF, and MLP models were built using the benchmark dataset with the optimal parameters. All these three models were used to predict antibacterial activity for approved small-molecule drugs. We compared the prediction performance of a single model and a combination of different models. The candidate antibacterial drugs were defined as the drugs that showed antibacterial activity in all the SVM,

RF, and MLP models. Drug information was acquired from the DrugBank database (https://www.drugbank.ca/) [33]. We first filtered the drugs with approved status but not withdrawn yet, then removed the drugs with molecular weight ≥ 1000. Finally, 2315 approved small-molecule drugs were screened to perform antibacterial activity prediction. The predicted active antibacterial drugs excluding FDA approved antibacterial drugs were defined as novel antibacterial drugs.

2.7 Structural similarity analysis

FP2 molecular fingerprint similarity was calculated among all novel antibacterial drugs and FDA approved antibacterial drugs. The overlap between fingerprints is quantified as a measure of molecular similarity using the Tanimoto coefficient (Tc). The predicted drugs with average and maximum molecular fingerprint similarity less than 0.1 and 0.2 were considered to be structurally novel. Furthermore, previous literature reported several core scaffolds shared by most antibacterial compounds [22]. The flexible maximum common substructure algorithms in the fmcsR package [34] in R was used to identify whether the core scaffolds exist in the predicted antibacterial drugs.

3. Results

3.1 Development process of antibacterial compound predictor

The development process of the antibacterial compound predictor is shown in Figure 1. The first step of the antibacterial compound predictor is to prepare the benchmark dataset using screened active and inactive antibacterial compounds from the ChEMBL and the PubChem database. Then, the SVM, RF, and MLP methods were used to build models using the benchmark dataset. Using the parameter grid search and 5-fold cross-validation strategy, the optimal parameters of these three models were determined (Table 1, Supplementary Figure 3-5). After training, parameter optimization, and model evaluation, the optimal SVM, RF, and MLP models were established. The final antibacterial compound predictor includes the combination of the optimal three models. The integrated model was used to predict the antibacterial activity of FDA approved small-molecule drugs from the DrugBank database.

Table 1
Optimal parameters and prediction performance of different machine learning methods.

	Support vector machine	Random forest	Multi-layer perception
Optimal parameters	gamma: 0.01 C:10	n_estimators: 750	hidden_layer_sizes: 512 alpha: 0.0001
Accuracy	0.852 ± 0.002	0.849 ± 0.004	0.847 ± 0.004
Precision	0.854 ± 0.004	0.868 ± 0.004	0.850 ± 0.007
Sensitivity	0.850 ± 0.004	0.822 ± 0.007	0.845 ± 0.003
Specificity	0.854 ± 0.005	0.875 ± 0.004	0.850 ± 0.009
F1 score	0.852 ± 0.002	0.844 ± 0.005	0.847 ± 0.003
AUC	0.926 ± 0.002	0.932 ± 0.002	0.920 ± 0.002
MSE	0.148 ± 0.002	0.151 ± 0.004	0.153 ± 0.004
Abbr: AUC, area under the curve; MSE, mean squared error.			
Parameters of predictive performance were displayed as mean ± standard deviation.			

3.2 High performance of the SVM, RF, and MLP models

The overall performance of the SVM, RF and MLP models was quantified by multiple classification evaluation indicators including accuracy, precision, sensitivity, specificity, F1 score, AUC and MSE (Table 1). The mean values of accuracy, precision, sensitivity, specificity, and F1 score of the three models at around 0.85. The mean values of AUC of these models were higher than 0.92 (ROC curves of these models shown in Supplementary Figure 6). The mean squared error of the three models is around 0.15. These indicate that all three models showed high effectiveness in antibacterial compounds prediction. Furthermore, the data showed that the standard deviations of these indicators are very small, suggesting that different negative datasets do not affect the overall performance of these models.

3.3 Prediction of candidate antibacterial small-molecule drugs

All approved small-molecule drugs in the DrugBank database were used to screen for potential antibacterial compounds through the antibacterial compound predictor. The results showed that there are large differences in the number of drugs in isolated prediction intervals among the SVM, RF, and MLP models. There are more compounds in the probability intervals at both ends and fewer compounds in the middle intervals in the MLP model, whereas the distribution of predicted probabilities showed the opposite trend in the RF and SVM models (Figure 2A). There were 1482, 1539, and 1398 predicted active antibacterial drugs in the single SVM, RF, and MLP models. A total of 1090 drugs showed antibacterial activity shared by all three models (Figure 2B). The single model and the combination of the two models

predicted relatively more active antibacterial compounds, and there is more overlap with FDA approved drugs (Supplementary Figure 7). Among the prediction results by the combination of the three models, 133 antibacterial drugs were FDA approved (Figure 2C). Our results suggested that both the single models and the combination of multiple models all showed excellent prediction performance (Supplementary Table 2). Furthermore, for the remaining 957 drugs, many of them belong to benzene and substituted derivatives (184 drugs) and steroids and steroid derivatives (116 drugs), few drugs belong to other categories (Figure 2D).

3.4 Structural similarity of the predicted antibacterial drugs

Molecular fingerprint similarity was calculated between the predicted and FDA approved antibacterial drugs. The predicted antibacterial drugs that are not approved for marketing were defined as novel predicted antibacterial drugs. There were low overall similarities between approved antibacterial drugs and novel predicted antibacterial drugs (Supplementary Figure 8). There were 873 novel-predicted antibacterial drugs that showed average similarities ≤ 0.2 to all approved antibacterial drugs (Figure 3A). According to previous reports [22], we identified 8 representative core scaffolds from the FDA approved antibacterial drugs (Supplementary Table 3). 906 predicted compounds do not contain any core scaffold (Figure 3B). Only 51 (5.3 %) of the predicted compounds showed a high overlap coefficient with core scaffolds (Supplementary Table 4). These indicate that most of the predicted antibacterial drugs are structurally novel.

3.5 Novel predicted antibacterial drugs

There were 9 novel-predicted drugs with an average similarity less than 0.1 and a maximum similarity less than 0.2 to all approved antibacterial drugs, and these drugs all showed high predicted probability in SVM, RF, and MLP models (Table 2). Details of these 9 drugs listed in Supplementary Table 5. Among these drugs, cyclophosphamide (DB00531) and ifosfamide (DB01181) are anticancer drugs that were used to treat a variety of hematological tumors and solid tumors. Apraclonidine (DB00964) is used to relieve postsurgical ocular hypertension. Echothiophate is used for the treatment of subacute or chronic angle-closure glaucoma. The other 5 drugs are mainly used in general anesthesia, such as enflurane (DB00228), isoflurane (DB00753), methoxyflurane (DB01028), desflurane (DB01189), and sevoflurane (DB01236).

Table 2
The prediction results of 9 antibacterial drugs with low structural similarities.

DrugBank ID	Name	Predicted probability			Structural similarity
		SVM	RF	MLP	(mean (min-max))[1]
DB00228	Enflurane	0.741	0.544	0.916	0.055 (0.000-0.119)
DB00531	Cyclophosphamide	0.571	0.518	0.902	0.086 (0.010-0.150)
DB00753	Isoflurane	0.698	0.536	0.980	0.055 (0.000-0.120)
DB00964	Apraclonidine	0.514	0.514	0.501	0.093 (0.013-0.198)
DB01028	Methoxyflurane	0.770	0.504	0.913	0.048 (0.000-0.143)
DB01057	Echothiophate	0.703	0.518	0.864	0.072 (0.017-0.143)
DB01181	Ifosfamide	0.589	0.515	0.888	0.095 (0.010-0.172)
DB01189	Desflurane	0.732	0.546	0.975	0.055 (0.000-0.150)
DB01236	Sevoflurane	0.538	0.517	0.934	0.060 (0.000-0.162)

Abbr. SVM, support vector machine; RF, random forest; MLP, multi-layer perception.

[1]The structural similarities were calculated between the novel predicted antibacterial drugs and FDA-approved antibacterial drugs.

4. Discussion

Exploring the antibacterial activity of the approved drugs may be an effective way of screening new antibiotics. It is an effective approach by using machine learning methods to predict active antibacterial compounds [16, 18, 35]. The accuracy of the prediction model is affected by many factors, such as the quality of the benchmark datasets [36], the representative molecular characteristics of the compounds [37], the applicable machine learning models [6], and the optimized model parameters [38]. This study collected a large amount of experimental data on the antibacterial activity of compounds from the ChEMBL and PubChem databases. By comparing the prediction accuracy of multiple machine learning models on benchmark datasets constructed based on different molecular fingerprints, our results showed that the average prediction accuracy of SVM, RF, and MLP models are higher than other machine learning methods, and the FP2 molecular fingerprints is more representative than other fingerprints. Therefore, it is reasonable to construct the antibacterial compound predictor through building the benchmark datasets by calculating the FP2 molecular fingerprint of the compounds, and combining the RF, SVM, and MLP models. However, the model constructed in this study did not achieve the desired prediction performance (only 133 of the 206 FDA approved antibacterial drugs have been successfully predicted). This is probably because the benchmark datasets collected data from multiple sources and requires a more

effective data integration strategy. Furthermore, it is worth noting that parameter optimization can only slightly improve (approximately 1%) the prediction accuracy of the different machine learning models.

Through structural similarity analysis of the predicted active antibacterial drugs, we screened 9 drugs with novel structures. Apraclonidine is mainly used for the prevention and treatment of post-surgical intraocular pressure (IOP) elevation, and it is also indicated for the short-term adjunctive treatment of glaucoma [39]. Echothiophate is used in the treatment of subacute or chronic angle-closure glaucoma and in some cases it is also used as accommodative esotropia [40]. Cyclophosphamide and ifosfamide are widely used broad-spectrum anticancer drugs [41, 42]. Studies showed that cyclophosphamide can inhibit bacterial translocation of the gastrointestinal tract [43] and reduce the abundance of lactobacilli and enterococci [44] in mice. Desflurane, enflurane, isoflurane, methoxyflurane, and sevoflurane are widely used volatile anesthetics [45, 46], most of these anesthetics have demonstrated antibacterial properties *in vitro* [47–50]. An early in *vitro* experiment showed that methoxyflurane and isoflurane exhibited excellent antibacterial activity, while enflurane had less effect on a few pathogens [48]. The resistance experiment to a variety of bacteria showed that isoflurane has higher antibacterial activity than sevoflurane [49]. Based on these reports, the antibacterial compound screening method used in this study is credible.

There are still many difficulties in the discovery of antibacterial compounds in silico. Firstly, the prediction accuracy is affected by the size and quality of the benchmark dataset. The definition of the active or inactive antibacterial compounds in this study is based on the in vitro experimental data. However, most of the screened active antibacterial compounds have not yet entered clinical trials, the human safety and clinical effectiveness of these compounds are still unclear [51]. Then, the compounds in this study were characterized by molecular fingerprints, whereas this method cannot reflect the complete structural features of given compounds and is not suitable for macromolecular compounds [37]. Next, machine learning models need further optimization. The prediction accuracy of the SVM, RF, and MLP models in this study is around 0.85, optimizing these models may be able to obtain higher prediction accuracy. Lastly, considering that compounds may produce different types of molecules during the metabolic process, computational simulation of the drug metabolic process [52] in the human will make the predictions more convincing.

5. Conclusions

In summary, this study provides a new insight for predicting antibacterial compounds with novel structures by using approved drugs. The existing approach could be extended by different augmentation methods (such as compound augmentation by graph or molecular description) with different machine learning state-of-the-art methods such as deep-learning methods. There are still many challenges and opportunities in using machine learning to predict antibacterial compounds. With the development of big data technology, the continuous optimization of machine learning models and algorithms, and the discovery of more antibacterial active compounds and drugs, it is foreseeable that the prediction of antibacterial compounds in the future will achieve higher accuracy and credibility.

Abbreviations

SVM: support vector machine

RF: random forest

MLP: multi-layer perception

FDA: Food and Drug Administration

TP: true positive

TN: true negative

FP: false positive

FN: false negative

ROC: receiver operating characteristic

AUC: area under the curve

MSE: mean squared error

Declarations

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Start-up Fund of Kunming University of Science and Technology (grant No. KKZ3201927005 for SXD), and the Yunnan Fundamental Research Projects (grant No. 2019FB050 and 202105AD160008 for SXD), the Start-up Fund of Southern Medical University (grant No. G820282016 for DGG), and the National Natural Science Foundation of China (grant No. 31501080 and 32070676 for DGG, No. 32160153 for SXD). The funding bodies had no role in the design of the study, collection, analysis and interpretation of data, or in writing the manuscript.

Authors' contributions

WXL and SXD designed the study. WXL, GHL, and SXD collected the data. WXL, XT, PPY, YZ, and DGG designed the method and analyzed the data. WXL, JHL, and SXD wrote the manuscript. All authors read and approved the submitted version.

Acknowledgements

Not applicable.

References

1. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Pulcini C, Kahlmeter G, Kluytmans J, Carmeli Y et al (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis 18(3):318–327
2. Jackson N, Czaplewski L, Piddock LJV (2018) Discovery and development of new antibacterial drugs: learning from experience? J Antimicrob Chemother 73(6):1452–1459
3. Pereira DA, Williams JA (2007) Origin and evolution of high throughput screening. Br J Pharmacol 152(1):53–61
4. Mishra KP, Ganju L, Sairam M, Banerjee PK, Sawhney RC (2008) A review of high throughput technology for the screening of natural products. Biomed Pharmacother 62(2):94–98
5. Tommasi R, Brown DG, Walkup GK, Manchester JI, Miller AA (2015) ESKAPEing the labyrinth of antibacterial discovery. Nat Rev Drug Discovery 14(8):529–542
6. Durrant JD, Amaro RE (2015) Machine-learning techniques applied to antibacterial drug discovery. Chem Biol Drug Des 85(1):14–21
7. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery 18(6):463–477
8. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y (2016) Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform 17(4):696–712
9. D'Souza S, Prema KV, Balaji S (2020) Machine learning models for drug-target interactions: current knowledge and future directions. Drug Discov Today 25(4):748–756
10. Walters WP, Barzilay R: **Applications of Deep Learning in Molecule Generation and Molecular Property Prediction**. *Acc Chem Res* 2020
11. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, Hickey AJ, Clark AM (2019) Exploiting machine learning for end-to-end drug discovery and development. Nat Mater 18(5):435–441
12. Lata S, Sharma BK, Raghava GP (2007) Analysis and prediction of antibacterial peptides. BMC Bioinform 8:263
13. Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett 20(2):180–186

14. Zoffmann S, Vercruysse M, Benmansour F, Maunz A, Wolf L, Blum Marti R, Heckel T, Ding H, Truong HH, Prummer M et al (2019) Machine learning-powered antibiotics phenotypic drug discovery. Scientific reports 9(1):5013
15. Rahman SF, Olm MR, Morowitz MJ, Banfield JF: **Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome**. *mSystems* 2018, **3**(1)
16. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackerman Z et al (2020) A Deep Learning Approach to Antibiotic Discovery. Cell 180(4):688–702 e613
17. Yang XG, Chen D, Wang M, Xue Y, Chen YZ (2009) Prediction of antibacterial compounds by machine learning approaches. J Comput Chem 30(8):1202–1211
18. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, Heckmann D, Anand A, Yang L, Nizet V et al (2018) Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nature communications 9(1):4306
19. Niehaus KE, Walker TM, Crook DW, Peto TEA, Clifton DA: **Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis**. In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Valencia*. IEEE (2014) 618-621
20. Hefti FF (2008) Requirements for a lead compound to become a clinical candidate. BMC Neurosci 9(Suppl 3):S7
21. Savoia D (2016) New Antimicrobial Approaches: Reuse of Old Drugs. Curr Drug Targets 17(6):731–738
22. Fair RJ, Tor Y (2014) Antibiotics and bacterial resistance in the 21st century. Perspect Medicin Chem 6:25–64
23. Choi KE, Balupuri A, Kang NS: **The Study on the hERG Blocker Prediction Using Chemical Fingerprint Analysis**. *Molecules* 2020, 25(11)
24. Lind AP, Anderson PC (2019) Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS One 14(7):e0219774
25. Deng YH, Wang NN, Zou ZX, Zhang L, Xu KP, Chen AF, Cao DS, Tan GS (2017) Multi-Target Screening and Experimental Validation of Natural Products from Selaginella Plants against Alzheimer's Disease. Front Pharmacol 8:539
26. Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Dogan T (2020) DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 11(9):2531–2557
27. Li GH, Huang JF (2012) CDRUG: a web server for predicting anticancer activity of chemical compounds. Bioinformatics 28(24):3334–3335
28. O'Boyle NM, Morley C, Hutchison GR (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem Cent J 2:5

29. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: An open chemical toolbox. J Cheminform 3:33
30. Jaeger S, Fulle S, Turk S (2018) Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model 58(1):27–35
31. Ozturk H, Ozkirimli E, Ozgur A (2018) A novel methodology on distributed representations of proteins using their interacting ligands. Bioinformatics 34(13):i295–i303
32. Jeon W, Kim D (2019) FP2VEC: a new molecular featurizer for learning molecular properties. Bioinformatics 35(23):4979–4985
33. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research 34(Database issue):D668–D672
34. Wang Y, Backman TW, Horan K, Girke T (2013) fmcsR: mismatch tolerant maximum common substructure searching in R. Bioinformatics 29(21):2792–2794
35. Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrubbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO et al (2019) A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell 177(6):1649–1661 e1649
36. Beltran JA, Aguilera-Mendoza L, Brizuela CA (2018) Optimal selection of molecular descriptors for antimicrobial peptides classification: an evolutionary feature weighting approach. BMC Genom 19(Suppl 7):672
37. Muegge I, Mukherjee P (2016) An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 11(2):137–148
38. Diaz I (2020) Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. Biostatistics 21(2):353–358
39. Harasymowycz P, Royer C, Cui AX, Barbeau M, Jobin-Gervais K, Mathurin K, Lachaine J, Beauchemin C: **Short-term efficacy of latanoprostene bunod for the treatment of open-angle glaucoma and ocular hypertension: a systematic literature review and a network meta-analysis**. *Br J Ophthalmol* 2021
40. Kini MM, Dahl AA, Roberts CR, Lehwalder LW, Grant WM (1973) Echothiophate, pilocarpine, and open-angle glaucoma. Archives of ophthalmology 89(3):190–192
41. Emadi A, Jones RJ, Brodsky RA (2009) Cyclophosphamide and cancer: golden anniversary. Nat Rev Clin Oncol 6(11):638–647
42. Matz EL, Hsieh MH (2017) Review of Advances in Uroprotective Agents for Cyclophosphamide- and Ifosfamide-induced Hemorrhagic Cystitis. Urology 100:16–19
43. Suzuki T, Itoh K, Hagiwara T, Nakayama H, Honjyo K, Hirota Y, Kaneko T, Suzuki H (1996) Inhibition of bacterial translocation from the gastrointestinal tract of mice injected with cyclophosphamide. Curr Microbiol 33(2):78–83
44. Alexander JL, Wilson ID, Teare J, Marchesi JR, Nicholson JK, Kinross JM (2017) Gut microbiota modulation of chemotherapy efficacy and toxicity. Nature reviews Gastroenterology hepatology

14(6):356–365

45. Kharasch Evan D, Thummel Kenneth E (1993) Identification of cytochrome P450 2E1 as the predominant enzyme catalyzing human liver microsomal defluorination of sevoflurane, isoflurane, and methoxyflurane. Anesthesiology 79(4):795–807
46. Terrell Ross C, Warner David S (2008) The invention and development of enflurane, isoflurane, sevoflurane, and desflurane. Anesthesiology 108(3):531–533
47. Horton JN, Sussman M, Mushin WW (1970) The antibacterial action of anaesthetic vapours. Br J Anaesth 42(6):483–487
48. Giorgi A, Parodi F, Piacenza G, Mantellini E, Salio M, Cremonte LG, Grosso E (1986) [Antibacterial and antifungal activity of isoflurane and common anesthetic gases]. Minerva Med 77(42-43):2007–2010
49. Martinez-Serrano M, Geronimo-Pardo M, Martinez-Monsalve A, Crespo-Sanchez MD (2017) Antibacterial effect of sevoflurane and isoflurane. Rev Esp Quimioter 30(2):84–89
50. Imbernon-Moya A, Ortiz-de Frutos FJ, Sanjuan-Alvarez M, Portero-Sanchez I, Merinero-Palomares R, Alcazar V (2017) Topical sevoflurane for chronic venous ulcers infected by multi-drug-resistant organisms. Int Wound J 14(6):1388–1390
51. Pogodin PV, Lagunin AA, Rudik AV, Druzhilovskiy DS, Filimonov DA, Poroikov VV (2019) AntiBac-Pred: A Web Application for Predicting Antibacterial Activity of Chemical Compounds. J Chem Inf Model 59(11):4513–4518
52. Roy H, Nandi S (2019) In-Silico Modeling in Drug Metabolism and Interaction: Current Strategies of Lead Discovery. Curr Pharm Des 25(31):3292–3305

Figures

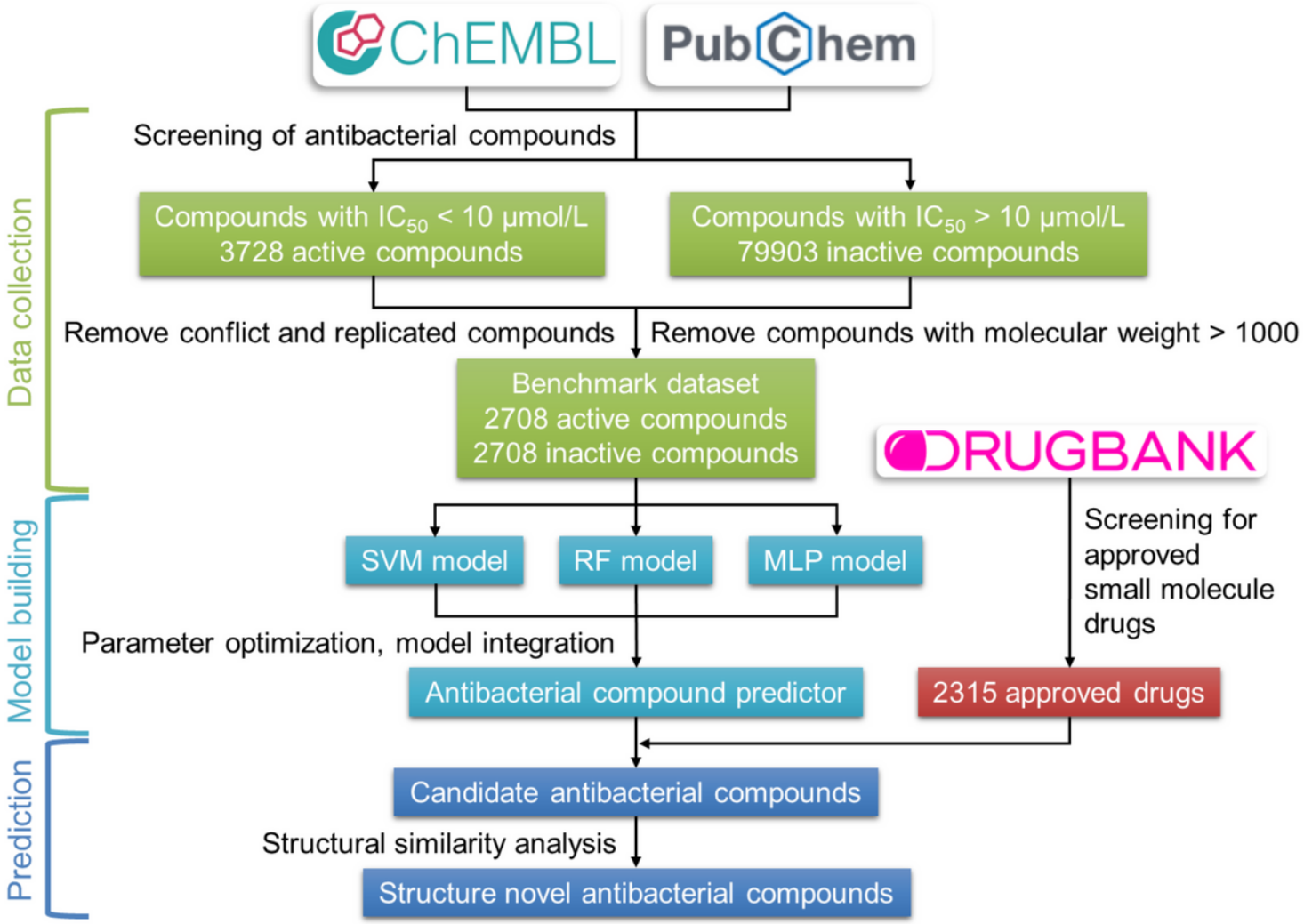

Figure 1

Flow chart of the construction of the antibacterial compound prediction model. The benchmark dataset was built using the active and inactive antibacterial compounds downloaded from the ChEMBL and the PubChem database. The combination of SVM, RF and MLP methods were used to construct the antibacterial compounds predictor, which is used to predict the antibacterial activity of approved small-molecule drugs from the DrugBank database.

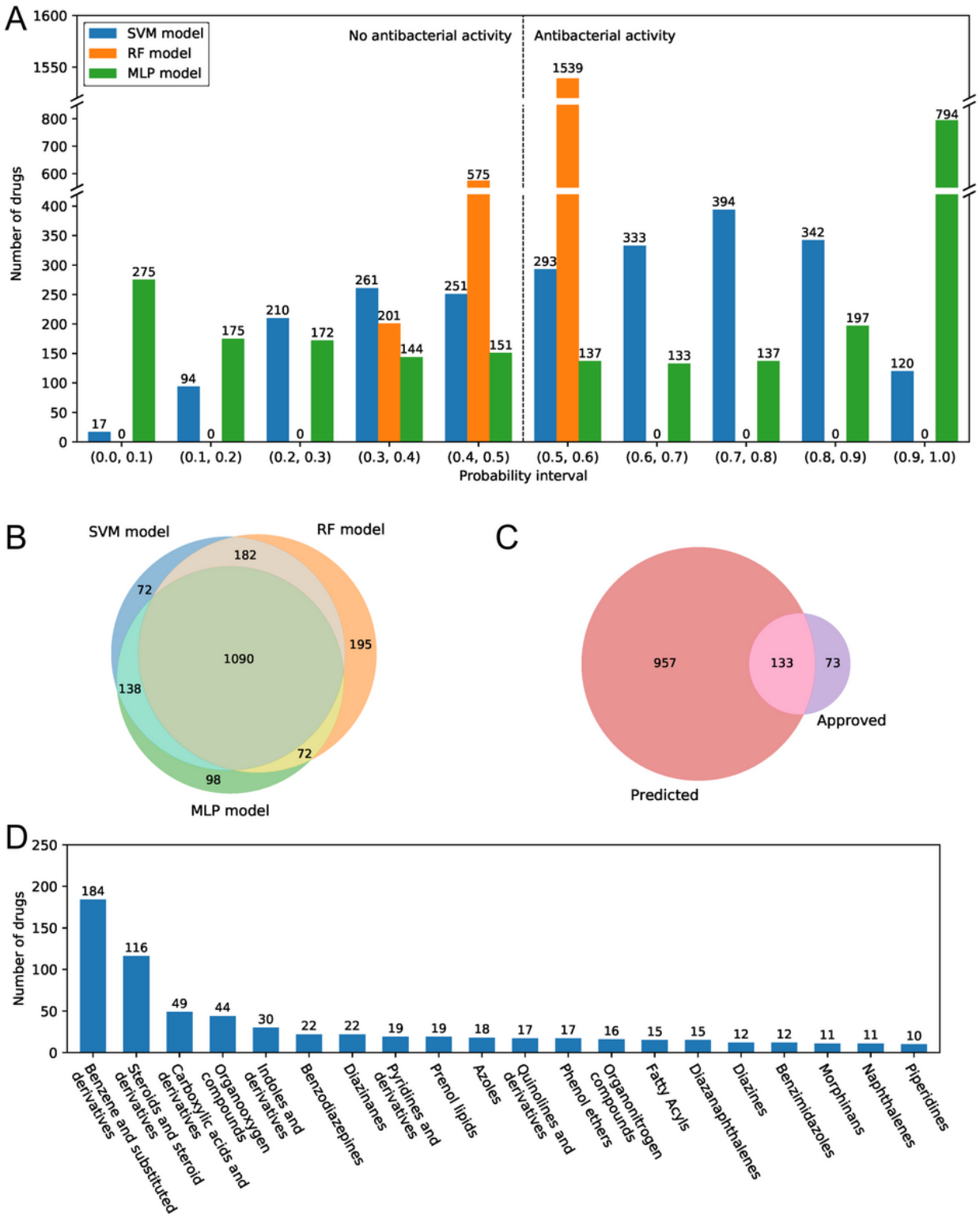

Figure 2

Antibacterial prediction results of approved small-molecule drugs. (A) The probability of predicted antibacterial activity for all small-molecule drugs in the SVM, RF, and MLP models. A drug with a probability value greater than 0.5 is considered an active antibacterial compound. (B) Venn diagram of the predicted antibacterial drugs in three machine learning models. (C) Venn diagram of the predicted

antibacterial drugs and FDA approved antibacterial drugs. (D) The top 20 categories of the 957 predicted novel antibacterial drugs.

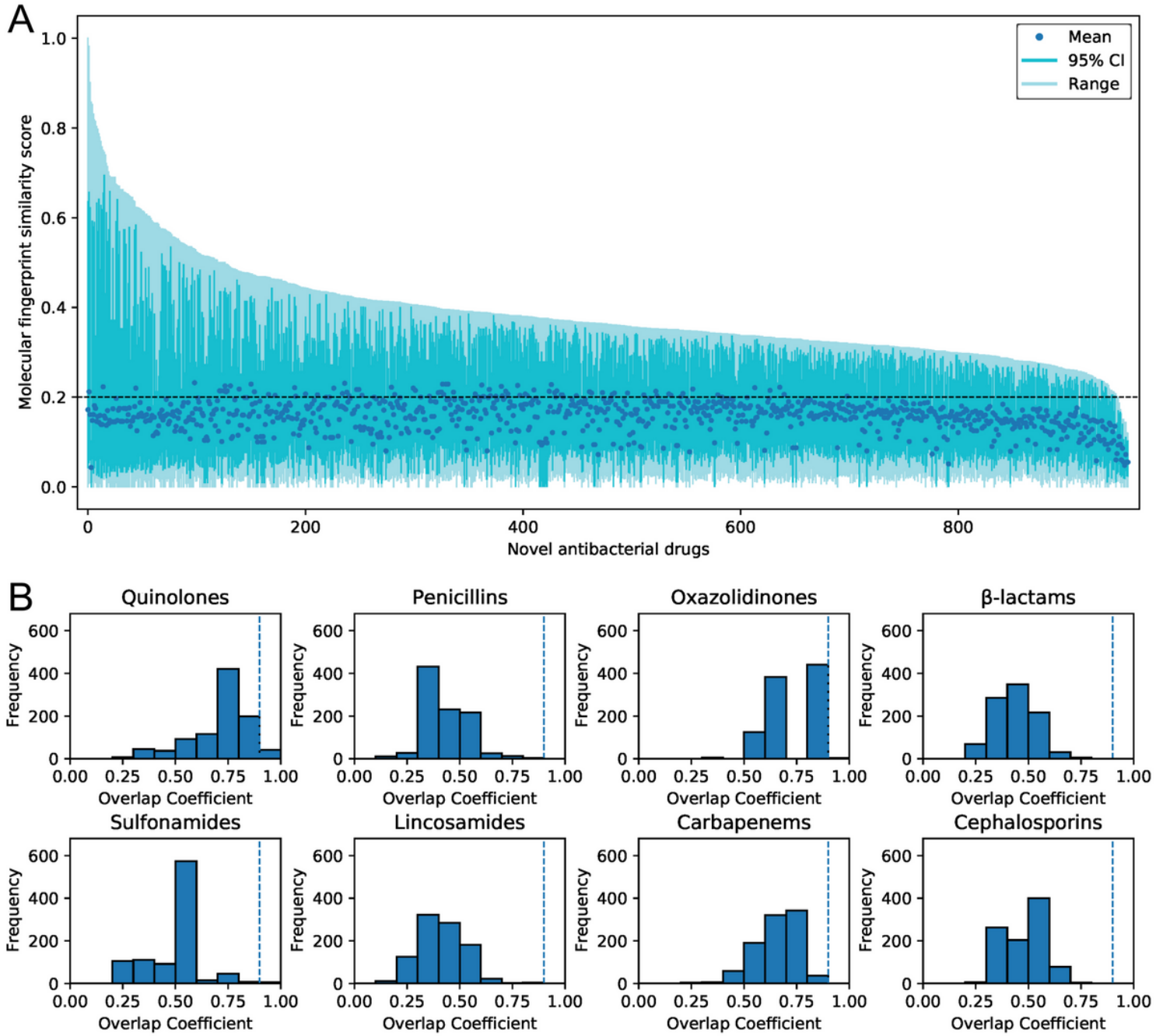

Figure 3

The similarity of the predicted antibacterial drugs and FDA approved antibacterial drugs. (A) The molecular fingerprint similarity of 957 predicted novel antibacterial drugs and 206 FDA approved antibacterial drugs. The average similarities between most of the predicted drugs and approved drugs were less than 0.2. (B) Substructure similarity between novel predicted antibacterial drugs and core scaffolds of approved antibacterial drugs. Compounds with an overlap coefficient higher than 0.9 are considered to have high substructure similarity.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx