Briefings in Bioinformatics, 2025, **26(1)**, bbae696
https://doi.org/10.1093/bib/bbae696
Problem Solving Protocol

Machine learning–enabled virtual screening indicates the anti-tuberculosis activity of aldoxorubicin and quarfloxin with verification by molecular docking, molecular dynamics simulations, and biological evaluations

Si Zheng[1,2,‡], Yaowen Gu[3,‡], Yuzhen Gu[4,‡], Yelin Zhao[5], Liang Li[5], Min Wang[2], Rui Jiang[6], Xia Yu[4,*], Ting Chen[1,*], Jiao Li[2,*]

[1]Institute for Artificial Intelligence & Department of Computer Science and Technology, Tsinghua University, Haidian District, Beijing 100084, China
[2]Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Chaoyang District, Beijing 100020, China
[3]Department of Chemistry, New York University, New York, NY 10027, United States
[4]National Clinical Laboratory on Tuberculosis, Beijing Key Laboratory on Drug-Resistant Tuberculosis, Beijing Chest Hospital, Capital Medical University, Tongzhou District, Beijing 101149, China
[5]Institute of Medicinal Biotechnology, Chinese Academy of Medical Sciences and Peking Union Medical College, Xicheng District, Beijing 100050, China
[6]Department of Automation, Tsinghua University, Haidian District, Beijing 100084, China

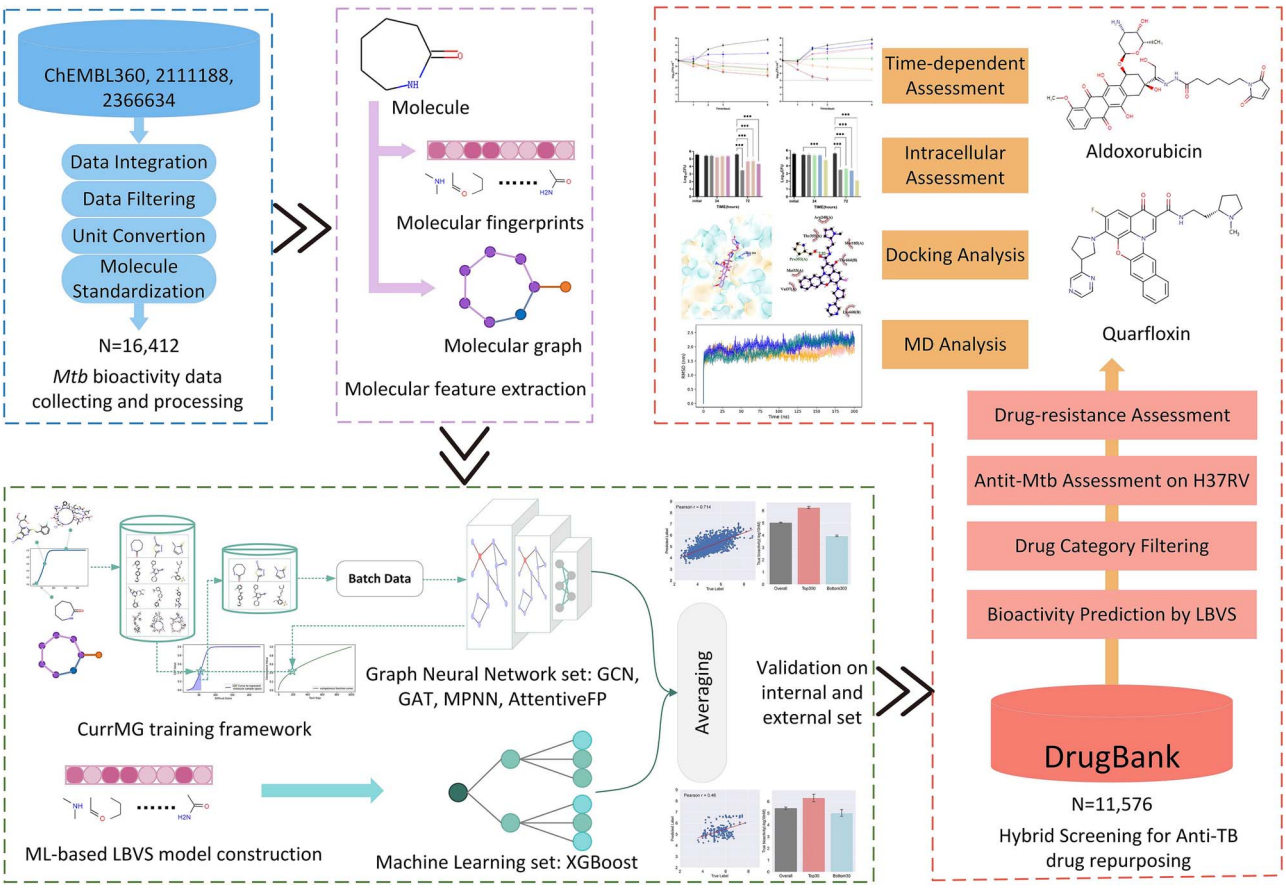
[*]Corresponding authors. Jiao Li, Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Chaoyang District, Beijing 100020, China. E-mail: jiao_h_li@163.com; Ting Chen, Institute for Artificial Intelligence & Department of Computer Science and Technology, Tsinghua University, Haidian District, Beijing 100084, China. E-mail: tingchen@mail.tsinghua.edu.cn; Xia Yu, National Clinical Laboratory on Tuberculosis, Beijing Key Laboratory on Drug-Resistant Tuberculosis, Beijing Chest Hospital, Capital Medical University, Tongzhou District, Beijing 101149, China. E-mail: yuxia@mail.ccmu.edu.cn
[‡]Si Zheng, Yaowen Gu, and Yuzhen Gu contributed equally to this work.

Abstract

Drug resistance in *Mycobacterium tuberculosis* (Mtb) is a significant challenge in the control and treatment of tuberculosis, making efforts to combat the spread of this global health burden more difficult. To accelerate anti-tuberculosis drug discovery, repurposing clinically approved or investigational drugs for the treatment of tuberculosis by computational methods has become an attractive strategy. In this study, we developed a virtual screening workflow that combines multiple machine learning and deep learning models, and 11 576 compounds extracted from the DrugBank database were screened against Mtb. Our screening method produced satisfactory predictions on three data-splitting settings, with the top predicted bioactive compounds all known antibacterial or anti-TB drugs. To further identify and evaluate drugs with repurposing potential in TB therapy, 15 screened potential compounds were selected for subsequent computational and experimental evaluations, out of which aldoxorubicin and quarfloxin showed potent inhibition of Mtb strain H37Rv, with minimal inhibitory concentrations of 4.16 and 20.67 μM/mL, respectively. More inspiringly, these two compounds also showed antibacterial activity against multidrug-resistant TB isolates and exhibited strong antimicrobial activity against Mtb. Furthermore, molecular docking, molecular dynamics simulation, and the surface plasmon resonance experiments validated the direct binding of the two compounds to Mtb DNA gyrase. In summary, our effective comprehensive virtual screening workflow successfully repurposed two novel drugs (aldoxorubicin and quarfloxin) as promising anti-Mtb candidates. The verification results provide useful information for the further development and clinical verification of anti-TB drugs.

Received: July 16, 2024. **Revised:** October 16, 2024. **Accepted:** December 17, 2024

Graphical Abstract



Keywords: machine learning; ligand-based virtual screening; drug repurposing; antitubercular; *Mycobacterium tuberculosis*

Introduction

Tuberculosis (TB) is a severe infectious disease with global health concern caused by the bacterium *Mycobacterium tuberculosis* (Mtb). TB is the leading cause of infectious disease worldwide, with an estimated 1.3 million deaths per year [1]. Although the extensive use of anti-Mtb drugs such as rifampicin (RIF), isoniazid (INH), and pyrazinamide has reduced the development of TB, the emergence of multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) in recent years has rendered existing antibiotics ineffective and brought new threats to tuberculosis control [2]. Drug resistance is responsible for a quarter of Mtb deaths and has become a significant challenge to the effective treatment and control of TB. Recently, the newly approved antitubercular agents, such as bedaquinoline, pretomanid, and delamanid, or their combined use with linezolid have shown the potential to enhance the treatment of drug-resistant tuberculosis. Still, <20% of persons are likely to benefit from such treatment due to the high medical cost, low accessibility, and insufficient validation of efficacy [3, 4]. In addition, clinical resistance to these new drugs has already been observed [5]. Thus, there is still an urgent need to develop new drugs or repurpose currently approved drugs faster and more efficiently to combat TB [6].

Existing or under development anti-Mtb drugs target various structures and pathways of the bacterium to combat tuberculosis, especially the enzymes involved in cell wall synthesis, metabolic pathways, DNA replication, etc. For example, the biosynthesis of cell wall components such as Rv3806c, DprE1, and the Emb proteins have become attractive targets for potent anti-Mtb drug research and development [7–9]. Moreover, DNA gyrase is also a validated target for antitubercular drug discovery [10]. The current research landscape in anti-drug compound development is diverse and rapidly evolving, and computational methods have revolutionized the process by enabling faster and more efficient identification of drug candidates [11, 12]. Among them, ligand-based virtual screening (LBVS) is a widespread computational screening method that compares a library of compounds with a known active ligand and predicts the bioactivities of new molecules based on known bioactivities of other molecules [13, 14]. LBVS offers accessibility and convenience to rapidly screen potentially bioactive molecules from large compound libraries in the early stage of drug discovery with low costs [15]. Several previous studies have demonstrated the value of LBVS methods in the discovery of anti-Mtb drugs. For instance, Naz *et al.* discovered an α-tryptophan synthase inhibitor with anti-Mtb bioactivity by pharmacophore model–based LBVS, molecular docking, and molecular dynamics simulation methods [16]. Zhu *et al.* proposed an LBVS workflow with three rounds consisting of 3D shape and pharmacophore matching and a topological shape and pharmacophore fingerprint algorithm to

discover anti-Mtb hits targeting the microbial enzyme 1-deoxy-D-xylulose-5-phosphate synthase [17]. Hassam *et al.* trained multiple machine learning (ML)–based LBVS models for anti-Mtb drug discovery targeting pantothenate synthetase [18]. Lane *et al.* organized an Mtb bioactivity prediction dataset and validated the superior performances of ML-based LBVS models on binary anti-Mtb bioactivity prediction tasks [19]. Ngidi *et al.* used virtual screening methods to identify Accolate as the best potential drug against the fatty acid degradation protein D32 (FadD32) in Mtb [20]. Overall, the implementation of ML in LBVS methods is regarded as an advanced approach, showing improved and competitive prediction performance in both bioactivity regression and classification prediction tasks [19, 21]. However, most of the current research lacks sufficient subsequent experimental evidence to prove the validity of the screened candidates. Moreover, an integrated multistep screening workflow combining advanced ML-based LBVS, classical structure-based virtual screening (e.g. molecular docking and molecular dynamics simulation), and wet laboratory experimental assessments would contribute significantly to improving anti-Mtb drug discovery.

In this study, we integrated multiple machine learning and deep learning models to establish LBVS to repurpose existing drugs for Mtb inhibition. Specifically, public anti-Mtb bioassay data were collected from ChEMBL [22] to construct anti-Mtb bioactivity prediction models, and a set of approved or investigational drugs obtained from DrugBank [23] were screened against Mtb. Then, the top-ranked screening hits resulted from a multistep filtering process were selected. Extensive *in vitro* bioassay experiments revealed that aldoxorubicin and quarfloxin clearly inhibited the growth of both the *Mycobacterium tuberculosis* H37Rv strain [with minimal inhibitory concentration (MIC) values of 4.16 and 20.67 μM/mL, respectively], and other drug-resistant strains (with MIC values of 4.16–41.34 μM/mL). Finally, molecular docking, molecular dynamics simulation analysis and surface plasmon resonance (SPR) were initiated to predict and validate the binding modes of the repurposed drugs.

Materials and methods

The overall workflow of this study is shown in Fig. 1. Specifically, anti-Mtb bioactivity data were collected from public databases, preprocessed, and integrated. Then, a set of ML models including XGBoost and graph neural networks (GNNs) were used to predict the MICs of known bioactive/inactive ligands using multiple chemical descriptors (molecular fingerprints and molecular graphs), followed by both internal and external validation. Overall, 11 576 compounds obtained from DrugBank were screened using these LBVS methods and sequentially filtered by drug category and bioassessments. Finally, aldoxorubicin and quarfloxin were selected and validated for their anti-Mtb bioactivity and drug-target binding relationships.

Data preparation

In this study, we first collected publicly available anti-Mtb bioassay data from the ChEMBL database, including ChEMBL360 (n = 168 749), ChEMBL2111188 (n = 23 554), and ChEMBL2366634 (n = 2022). Then, we followed our previous preprocessing strategies [24] to assemble these raw anti-Mtb bioassay data into an integrated dataset consisting of compounds with Mtb bioactivity annotations for subsequent virtual screening model training. (1) The "Standard Relation" is "=" (n = 176 256); (2) the "Standard Type" is "MIC" (n = 30 982); (3) the "Standard Units" are {"μM", "μmol/L", "μM/L", "μMl^{-1}", "10^{-6} mol/L"} and {"μg/ml"} (n = 30 308); (4) the bioactivities of those bioassay data with "Standard Units" of "μg/ml" were converted to "μM" by dividing by their molecular weights and then multiplying by 1000; (5) all bioactivities were converted to "$-\log_{10}$M"; (6) the molecular simplified molecular input line entry specification structures were normalized with MolVS [25]; (7) the duplicate molecules were removed (n = 16 412).

We also introduced the anti-Mtb bioassay dataset (n = 258) from Lane *et al.* [26] as the external validation set to evaluate the screening performance of the developed computational models for virtual screening compared to the random strategy. In addition, to construct a drug repurposing library, we collected 11 576 drugs as well as their chemical structures from DrugBank. The molecules in these datasets were preprocessed in the same manner as those in the ChEMBL dataset.

Molecular feature extraction

Specifically, to construct the XGBoost model for LBVS, we extracted two of the most widely used fingerprint-based descriptors, Extended Connectivity FingerPrint (ECFP4) [27] and the Molecular ACCess System (MACCS) fingerprint [28]. ECFP4 is a Morgan fingerprint with a radius of 2, and molecules are decomposed into substructures with the radius distance for each atom. As the radius expands, ECFP includes all identifiers found in both the previous and current iterations. All identifiers corresponding to different unique substructures were encoded by 2048 bits via one-hot encoding to represent the molecular features. MACCS predefines 166 unique substructures, the features of which were also encoded using a one-hot encoding scheme.

To design models based on GNNs for virtual screening, we used the DGL-lifesci package [29] to calculate the molecular atom and bond features and then construct molecular graphs. The atom features include the atomic elements, atomic degrees, number of implicit hydrogens, formal charges, number of radical electrons, atom hybridization, aromatics, and total hydrogens. All atom features were encoded with a one-hot encoding schema. The bond features entail bond type, conjugation, rings, and stereo configuration, in which the bond types and the stereo configuration were also encoded with a one-hot encoding schema.

Construction of ML models for LBVS

In detail, we first split the integrated anti-Mtb bioassay data from ChEMBL into a training dataset and an internal validation dataset with a splitting ratio of 8:2 based on three splitting strategies (random splitting, scaffold splitting, activity cliff splitting). Then, all ChEMBL data were integrated as the training set for model development, and the anti-Mtb bioassay dataset from Lane *et al.* [26] was used for external validation.

We used the Scikit-learn package [30] to construct an XGBoost model (XGB) for LBVS, which predicts the antitubercular bioactivities for the given query molecules based on their ECFP4 or MACCS features. We also proposed GNNs, including the graph convolutional network (GCN) [31], graph attention network (GAT) [32], message passing neural network (MPNN) [33], and AttentiveFP [34], for LBVS with the molecular graphs and their atom and bond features as inputs and antitubercular bioactivities as outputs. Moreover, for GNN model training, we used a curriculum learning strategy (CurrMG) [35] based on our previous studies to enhance model convergence and filter the noise. The DGL-lifesci package was used for GNN model construction. The optimization and training details of the above ML models are shown in Additional file 1.

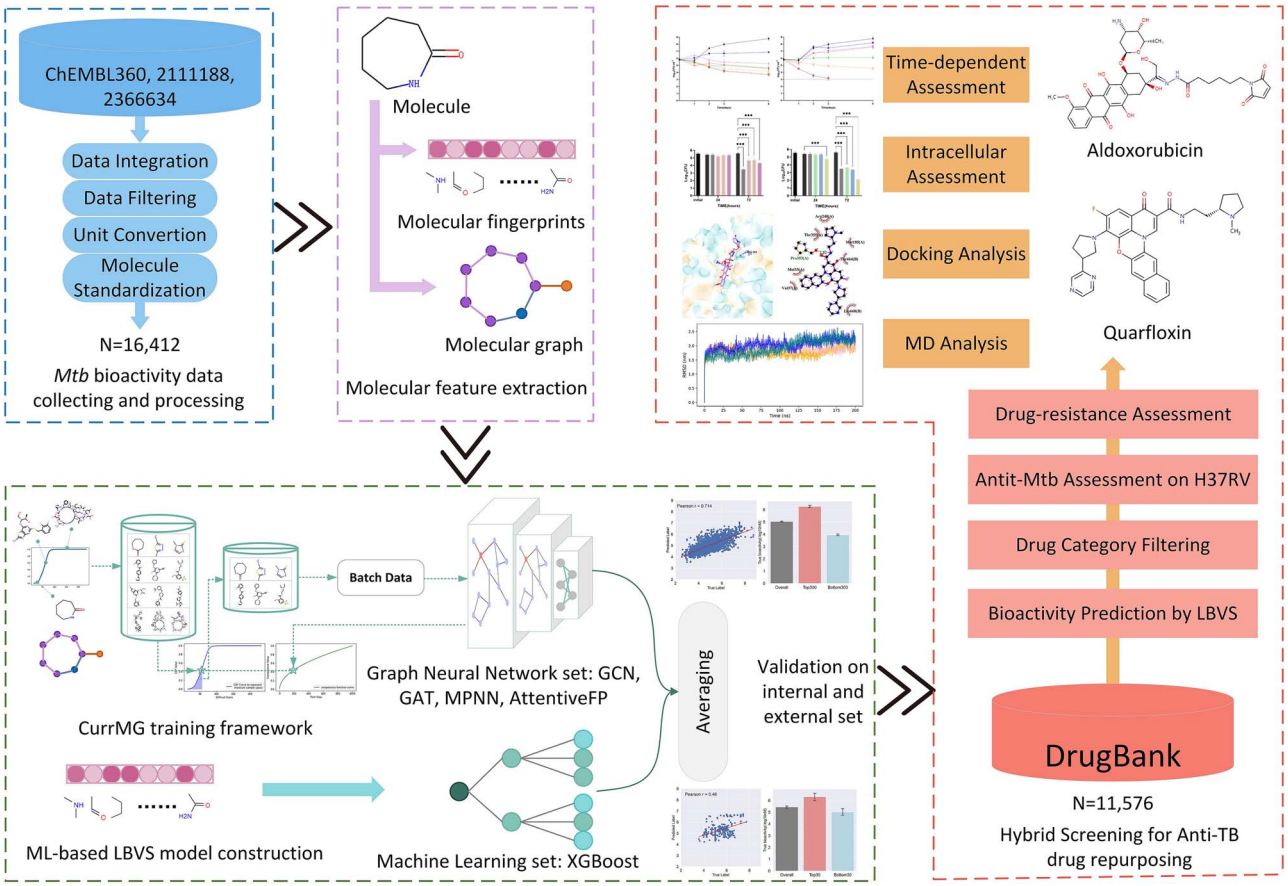


Figure 1. The overall workflow of this study.

Finally, we aggregated the prediction results of the LBVS models with a soft voting averaging strategy. Here, a leave-one-model-out strategy was employed to exclude the poorest performing model, and the remaining LBVS models were then combined to form an ensemble for final predictions and output generation.

Model evaluation

Both internal and external validations were used to evaluate the performance of the constructed LBVS models. As antitubercular bioactivity prediction is a regression task, we used the coefficient of determination (R^2), mean square error (MSE), and mean absolute error (MAE) [36] to assess the prediction performance of our LBVS models, which can be formulated as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y_i})^2} \quad (1)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (3)$$

where N denotes the sample size and y_i, $\hat{y}_i$, and $\overline{y_i}$ are the true label, predicted label, and overall average label, respectively. Meanwhile, the MSE_{cliff} was adopted as the activity cliff splitting metric to determine the performance of LBVS model in predicting bioactivities for sensitive molecules, which is formulated as

$$MSE = \frac{1}{N_c}\sum_{i=1}^{N_c}(y_i - \hat{y}_i)^2 \quad (4)$$

where N_c represents the number of activity cliff molecules in the test set. Moreover, taking 6-$\log_{10}$M as the threshold to differentiate bioactive/inactive molecules [13, 37], we also adopted the enrichment factor (EF@N) and hit rate (HR@N) to evaluate the efficiency of our LBVS models in discovering bioactive molecules among *top N* predictions. Taking $N = 300$ as an example, EF@300 and HR@300 can be formulated as

$$EF@300 = \frac{TP_{300}/300}{TP/N} \quad (5)$$

$$HR@300 = 100\% \times \frac{TP_{300}}{300} \quad (6)$$

where TP_{300} and TP denote the number of true positive bioactive molecules among the top 300 predictions and all predictions, respectively.

Reference strain and clinical isolates

The *M. tuberculosis* reference strain H37Rv ATCC27294 was obtained from the American Type Culture Collection. Clinical MDR-TB isolates stored in the Biobank in Beijing Chest Hospital (Beijing, China) were tested to investigate their susceptibility *in vitro*. The isolates were first used on Löwenstein–Jensen (LJ) medium and tested with MPT64 antigen to confirm the presence of the *M. tuberculosis* complex (Hangzhou Genesis Biodetection and Biocontrol Co., Ltd., China).

Determination of the minimum inhibitory concentration of selected compounds

Quarfloxin, pibrentasvir, sonidegib, batefenterol, glecaprevir, sitravatinib, oglemilast, losmapimod, nirogacestat, velsecorat,

sapitinib, sunitinib, and lenacapavir were purchased from TargetMol, USA. Aldoxorubicin and elexacaftor were purchased from MedChemExpress, USA. In addition, aldoxorubicin acted as a prodrug of doxorubicin, and thus, the MIC of doxorubicin was also determined. All inhibitors were dissolved in dimethyl sulfoxide (DMSO) to generate stock solutions with a concentration of 5 or 10 mg/mL. The broth microdilution method was performed according to the guidelines of the Clinical and Laboratory Standards Institute. Middlebrook 7H9 broth (Becton, Dickinson) containing 10% oleic acid–albumin–dextrose–catalase (OADC) was used to determine the MICs of *M. tuberculosis* and the MDR-TB isolates. The inoculum was prepared with fresh cultures grown on LJ medium. The tested drug concentrations ranged from 0.049 to 100 μg/mL. Briefly, *M. tuberculosis* and the MDR-TB isolates were scraped from LJ medium, homogenized, and adjusted to 1 McFarland standard. Then, the suspensions were diluted and inoculated into a 96-well microtiter plate to achieve a final bacterial load of 10^5 colony-forming units (CFUs) per well. The plates were then incubated at 37 °C for 7 days, after which 30 μL of resazurin (0.02%, wt/vol) was added to each well, and the plates were reincubated for an additional 24 h at 37 °C [38, 39]. A change from blue to pink or purple indicated bacterial growth. The MIC was defined as the lowest concentration of antibiotic that prevented a color change. In order to facilitate comparison of different drugs' efficiency, we also changed the nanograms per milliliter to micromolar per milliliter for the determined MIC value.

Time–kill curves

Individual Eppendorf tubes with 100 μL of Middlebrook 7H9 plus OADC growth supplement (BD Bioscience) and 0.05% Tween 80 containing six 2-fold increasing concentrations of each antibiotic (from 1×, 2×, 4×, 8×, to 16× MIC) were cultured with the inoculum (density ~10^5–10^6 CFU/mL) at 37 °C. An additional tube without a drug was included as a growth control, and evaluation of INH was used as positive control. The Eppendorf tubes were inoculated with 100 μL of the previously prepared inoculum at a starting concentration of 5×10^5 CFU/mL and incubated for 7 days at 37 °C. At predetermined time points (0, 1, 2, 3, and 6 days), 30 μL was taken from each tube, and dilutions (ranging from 10^{-1} to 10^{-6}) were prepared. Then, 10 μL of each dilution was plated directly on Middlebrook 7H10 supplemented with OADC agar for CFU counting. Agar plates were incubated at 37 °C, and the number of CFUs was determined 3 weeks later.

Intracellular killing and concentration–kill assays

THP-1 cells were seeded at 5×10^5 cells/well in a 24-well plate and induced to differentiate into macrophages with phorbol myristate acetate (100 ng/mL) for 48 h. The cells were infected at a multiplicity of infection of 5:1 with Mtb H37Rv (ATCC27294). After 4 h of infection at 37 °C under 5% CO_2, the cells were gently washed three times with prewarmed 1× PBS to remove the extracellular bacteria. For the intracellular killing assay, RPMI complete medium containing quinfloxin or aldoxorubicin at concentrations of 1× MIC, 2× MIC, and 4× MIC was used. A culture medium containing DMSO was used as a negative control, and RPMI 1640 medium with INH [3 μg/mL (21.88 μM/mL)] was used as a positive control. At 24 and 72 h postinfection, the macrophages were extensively washed with PBS and lysed with 0.1% Triton X-100. The number of CFUs was determined by plating serial dilutions of the lysates on 7H10 agar plates. The bacterial survival rate was calculated using the following formula: viability = (CFUs of bacteria treated with quinfloxin, aldoxorubicin, or INH/CFUs of the bacteria treated with DMSO) × 100%.

Molecular docking

To further investigate the potential protein–ligand interactions of the predicted bioactive candidates in Mtb, Autodock Vina [40] was used to optimize the molecular binding conformers and predict the binding modes. The docking coordinates (67.98, 38.72, 34.61) were adopted from a reference crystal structure (PDB ID: 7UGW). The docking grid box, the exhaustiveness, and the number of sampled poses were set as 20, 64, and 20, respectively. After the docking procedure finished, vina scores (kcal/mol) of the top 10 poses were reported to evaluate the binding free energies. To explore the binding modes of protein–ligand pairs, we adopted LigPlot+ [41] to visualize the interactions (hydrogen bonds and hydrophobic contacts) between proteins and ligands on 2D and 3D structures.

Molecular dynamics simulation

Molecular dynamics simulation was performed by Amber [42]. The initial configurations for protein–ligand complex were generated by CHARMM-GUI [43] with a general Amber force field [44]. A unit cell with water as solvent was defined (box size 15 Å), and the complex was neutralized by ions (Na^+ and Cl^-). Then, an energy minimization was executed to relax the structure and guarantee appropriate geometry in the system. The NVT and NPT ensemble equilibrations were successively conducted with the 300-K fixed temperature and 200-ps running. After that, a 200-ns molecular dynamic simulation was produced to simulate the trajectories of complex and protein–ligand binding interactions.

Surface plasmon resonance analysis

The surface plasmon resonance assay was used to examine the interaction between the DNA gyrase and two selected compounds. As shown in previous studies, DNA gyrase is composed of two subunits, GyrA and GyrB [45]. Here, the recombinant GyrA and GyrB subunits of *M. tuberculosis* DNA gyrase was purchased from TargetMol, USA. All SPR experiments were performed on a BIAcore T200 biosensor system (GE Healthcare Life Sciences, Piscataway, NJ, USA) at 25 °C using a CM5 chip (Lot: 10343798). The bindings of gyreA and gyreB at different concentrations of aldoxorubicin and quarfloxin were performed in 1× PBS-P (including 5% DMSO) (GE Healthcare Life Sciences) with a contact time of 100 s. The dissociation time was 100 s after each binding reaction. Afterward, aldoxorubicin was regenerated with pH 2.5 glycine for 60 s and quarfloxin in water for 10 s. All data were analyzed by the kinetic model in the Biacore T200 Evaluation Software 2.0 (GE Healthcare, US), and K_d was applied to evaluate the binding affinity.

Results

Molecular property distribution analysis

The integrated bioassay data from ChEMBL contained 16 412 records of compounds that inhibited Mtb along with their MIC values. According to the Quantitative Estimate of Drug-likeness (QED) rules, the distributions of eight physiochemical and structural properties were calculated, and their correlations with antitubercular activity were determined (Fig. 2). As shown in Fig. 2, there were certain proportions of molecules from the ChEMBL dataset that did not meet the requirements of drug-likeness mentioned in the QED rules, especially for molecular properties such as AlogP (26.88%), HBA (52.02%), and ALERTS (43.89%). It is also clear from Fig. 2 that the molecular properties such as

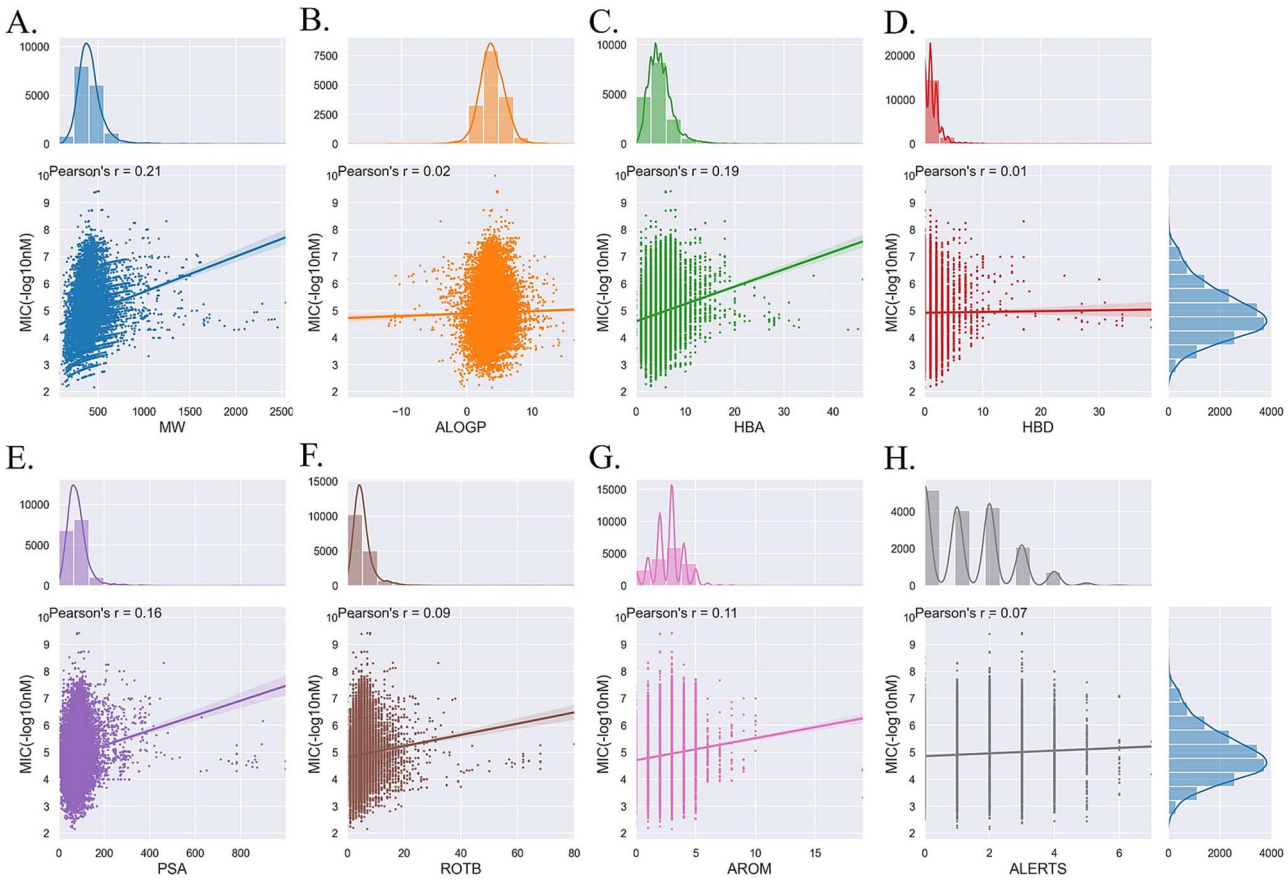


Figure 2. QED property distributions of the compounds from the ChEMBL dataset and their correlations with MIC. (A) Molecular weight (MW), (B) AlogP, (C) number of hydrogen bond acceptors (HBA), (D) number of hydrogen bond donors (HBD), (E) polar surface area (PSA), (F) number of rotatable bonds (ROTB), (G) number of aromatic rings (AROM), and (H) number of alert structures (ALERTS).

MW (Pearson's $r = 0.21$), HBA (0.19), and PSA (0.16) have a weak correlation with antitubercular bioactivity. These findings suggested that taking simple molecular properties as descriptors cannot accurately identify molecules with antitubercular activity. We also evaluated the consistency of the molecular property distributions between the integrated ChEMBL training dataset and DrugBank repurposing library (Additional file 2: Fig. S1). Similar distributions of molecular properties, such as HBA, PSA, and ROTB, were observed, while the molecular property distributions for the DrugBank dataset showed a left bias for MW, AlogP, AROM, and ALERTS. The overall distributions between the ordinary molecular descriptor distributions of the collected ChEMBL dataset and the DrugBank repurposing library were consistent.

We further explored key molecular substructures and scaffolds among the anti-Mtb bioactive molecules by fingerprint-based molecular clustering and grouping. Here, pairwise bulk Tanimoto similarities for each two bioactive molecules ($-\log_{10}M > 6$) in the ChEMBL dataset were calculated using ECFP4 molecular fingerprints, and Butina clustering (cutoff = 0.4) was used for molecular clustering. The common molecular scaffolds of the top five clusters (number of molecules >30) are shown in Fig. 3A. Furthermore, the t-SNE method was used to visualize the chemical space distances and groups of bioactive or bioinactive molecules (Fig. 3B). From Fig. 3B, we can see that the bioinactive molecules evenly fill all chemical space, and three observable groups of bioactive molecules presenting substructures similar to those of the common molecular scaffolds shown in Fig. 3A were observed. Furthermore, we also compared the chemical spatial distributions of molecules from DrugBank with those from ChEMBL. As shown in Fig. 3C, the molecules from the DrugBank dataset covered most of the chemical space that the molecules from the ChEMBL dataset filled, including the spaces of bioactive groups.

Construction and validation of ML models for LBVS

ML models and a soft averaging ensemble approach (ENSEM) were used for LBVS, and their anti-Mtb bioactivity prediction performance was evaluated on the internal validation dataset (ChEMBL) with three data-splitting strategies and the external validation dataset (Lane *et al.*) with all ChEMBL data as training set. The results in Table 1 show that the developed models achieved acceptable performance on the internal validation dataset. Comparatively, deep learning methods such as GCN_TS, GAT_MCE, and MPNN_AB outperform AttentiveFP_MCE and traditional machine learning methods like XGBoost in terms of the R^2, MAE, and MSE metrics, with GCN_TS showing the best predictive performance. During external validation, all models exhibited a significant drop in performance, among which AttentiveFP_MCE had the worst performance on the internal validation dataset but the best performance on the external validation dataset, and machine learning methods exhibit better predictive performance than GCN_TS, GAT_MCE, and MPNN_AB [Table 1, Table 2, Additional file 3 (Table S3, Table S4)]. One possible reason for this result might be the low structural similarities between the bioactive molecules in the ChEMBL dataset and those in the Lane *et al.* dataset. While deep learning can capture complex patterns and perform well

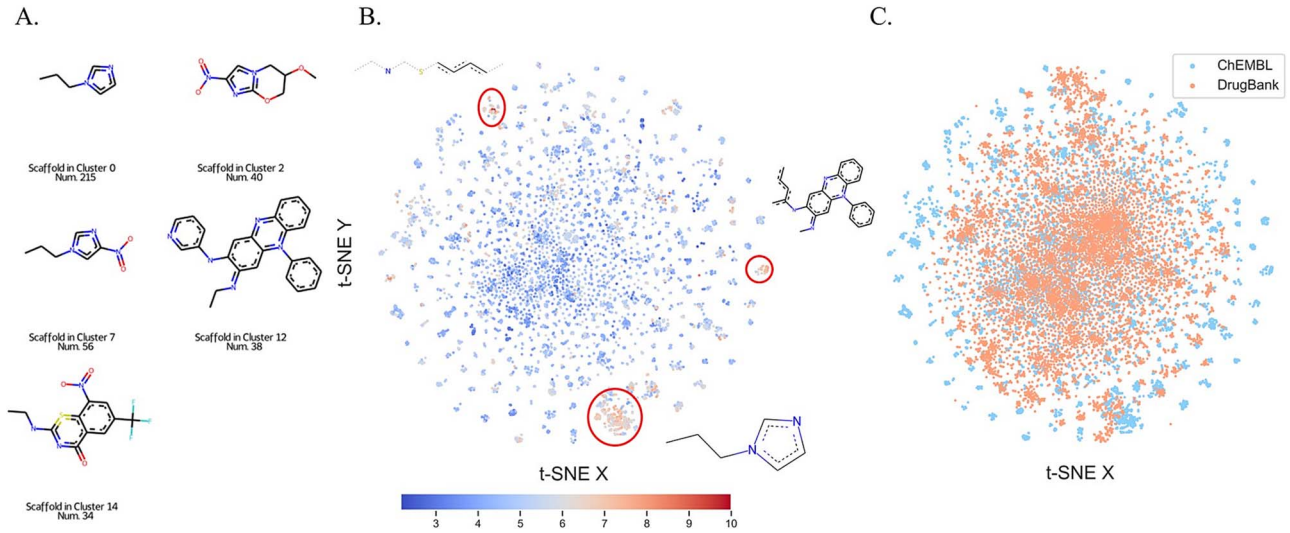


Figure 3. Chemical structure space analysis. (A) Representative molecular scaffolds in the top 5 bioactive molecular clusters. (B) T-SNE plot data and common chemical substructures in three representative bioactive molecular clusters in the ChEMBL dataset (red dots: bioactive molecules; blue dots: bioinactive molecules). (C) T-SNE plot data from the ChEMBL and DrugBank datasets.

Table 1. Evaluation of the performance of six machine learning models and the ensemble approach for LBVS on the ChEMBL internal validation dataset with random splitting

Model	R^2	MAE	MSE	EF@300	Hit@300
XGB_ECFP	0.336	0.636	0.641	3.431	61.667
XGB_MACCS	0.301	0.669	0.655	3.116	56.000
GCN_TS	0.520	0.460	0.521	3.672	66.000
GAT_MCE	0.490	0.489	0.543	3.765	67.667
MPNN_AB	0.467	0.511	0.548	3.413	61.333
AttentiveFP_MCE	0.238	0.730	0.677	2.838	51.000
ENSEM	0.610	0.475	0.376	3.913	70.333

on training data, its performance may degrade significantly when faced with new data that differs from the training set. In contrast, machine learning models tend to generalize better under such conditions, as they are less reliant on large datasets and have fewer parameters.

We implemented a leave-one-model-out strategy to select ensemble of LBVS models, as shown in Additional file 3 (Table S1, Table S2), excluding GAT_MCE that slightly enhanced ENSEM's performance across all validation sets. Thus, the remaining LBVS models (XGB_ECFP4, XGB_MACCS, GCN_TS, MPNN_AB, and AttentiveFP_MCE) were then combined to form an ensemble for final predictions. Generally, the ensemble approach achieved the highest performance in terms of R^2, MAE, and MSE (Table 1, Table 2). We further showed the scatter plots of the predicted anti-Mtb MIC values of the ensemble model and the true anti-Mtb MIC values of the compounds in the internal/external validation datasets (Fig. 4A, Fig. 4B). The scatter plots for other ML models for LBVS are shown in Additional file 2 (Fig. S2, Fig. S3). Moreover, the performance results for two more challenging data-splitting approaches (scaffold split, activity cliff split) revealed the solidarity and robustness of these LBVS models on unknown and structural sensitive samples. For instance, the results presented in Additional file 3 (Table S3) revealed that all developed models experienced significant performance degradation under the scaffold splitting setting. However, the ensemble approach consistently demonstrated the highest performance in terms of R^2, MAE, and MSE. Additional file 3 (Table S4) showed the performance outcomes for activity cliff splitting, highlighting that while all LBVS models exhibited moderate performance, the ensemble model outperformed others by achieving the lowest prediction errors between ordinary molecules and activity cliff molecules (by comparing MSE and MSE_{cliff}). This suggests that the ensemble approach is particularly effective at distinguishing sensitive samples characterized by minor structural differences yet significant bioactivity discrepancies. Consequently, the ensemble approach was used for the following anti-Mtb drug virtual screening.

To further measure the efficiency of these constructed ML models in practical virtual screening scenarios, we calculated four metrics (EF@300 and Hit@300 for the internal validation dataset and EF@30 and Hit@30 for the external validation dataset) for each model, which are listed in Table 1, Table 2, and Additional file 3 (Table S3, Table S4). The results indicated that the constructed models could enrich bioactive molecules among the top predictions with desirable enrichment efficiencies on both the internal validation dataset and the external validation dataset, and the hit rates were also acceptable. We also plotted the average MIC values of the overall, top 300 predicted, bottom 300 predicted, top 30 predicted, and bottom 30 predicted molecules to determine the bioactivity distribution differences among the screened, excluded, and overall subsets (as shown in Fig. 4D, Fig. 4E). The results showed that the anti-Mtb bioactivities of the top predicted molecules are significantly higher than those of the overall molecules and bottom predicted molecules, indicating notable ranking performance for the proposed virtual screening models.

Table 2. Evaluation of the performance of six machine learning models and the ensemble approach for LBVS on the Lane *et al.* external validation dataset

Model	R^2	MAE	MSE	EF@30	Hit@30
XGB_ECFP	0.141	0.571	0.607	2.736	70.000
XGB_MACCS	0.063	0.623	0.624	2.345	60.000
GCN_TS	0.059	0.626	0.623	2.867	73.333
GAT_MCE	−0.531	1.018	0.782	2.606	66.667
MPNN_AB	−0.073	0.713	0.669	2.606	66.667
AttentiveFP_MCE	0.173	0.550	0.600	2.606	66.667
ENSEM	0.165	0.584	0.555	2.736	70.000

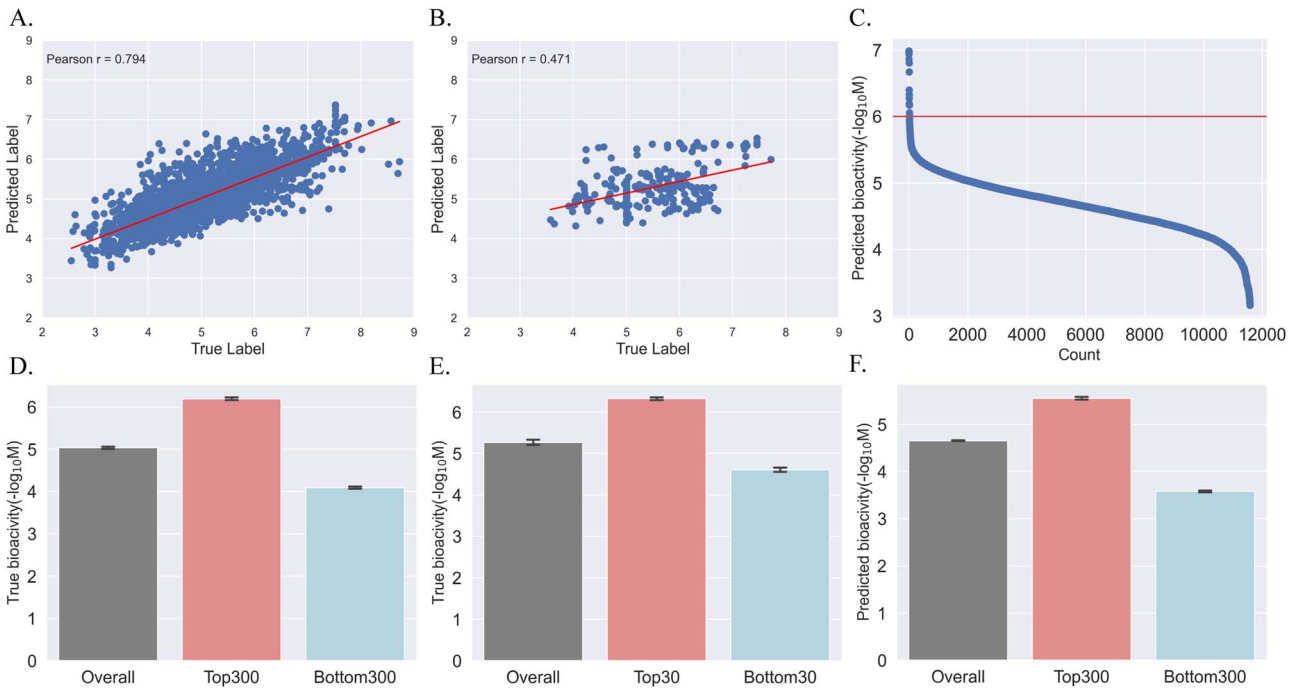


Figure 4. (A) Scatter plot showing the prediction results of the ENSEM on the ChEMBL internal validation dataset with random splitting (*x*-axis: actual MIC values; *y*-axis: EnSEM predicted MIC values). (B) Scatter plot showing the prediction results of the ensemble model on the Lane *et al.* external validation dataset (*x*-axis: actual MIC values; *y*-axis: EnSEM predicted MIC values). (C) Rank-ordered predicted bioactivities for DrugBank molecules. (D) Average bioactivities of the overall, top 300 ranked, and bottom 300 ranked molecules in the ChEMBL internal validation dataset. (E) Average bioactivities of the overall, top 30 ranked, and bottom 30 ranked molecules in the Lane *et al.* external validation dataset. (F) Average predicted bioactivities of the overall, top 300 ranked, and bottom 300 ranked molecules in the DrugBank repurposing dataset.

Screening of anti-Mtb compounds in the DrugBank database

We further retrained the ML models for LBVS on the whole ChEMBL dataset and then used the ensemble approach to combine the predictions of the retrained models to screen anti-Mtb molecules in the DrugBank database. The predicted rank-ordered bioactivities are shown in Fig. 4C. We first adopted −6 $\log_{10}$M as the threshold to identify bioactive molecules, and Table 3 shows the 15 top-ranked bioactive candidates for anti-Mtb drugs. As expected, the majority of them were anti-Mtb or antibacterial drugs. For instance, rifampicin (RIF), delamanid, pretomanid, rifamycins (rifabutin and rifapentine), and macozinone are all major hits active against Mtb, which also proved the effectiveness of our constructed LBVS methods for anti-Mtb drug repurposing [46, 47] (Table 3).

To get the tractable number of hits remaining for our filtering steps and analysis, we then expanded the screening range with relatively looser bioactivity restrictions (top 200 screened hits) to identify more novel potential anti-Mtb drugs (Additional file 3: Table S5). While a small portion of the ranked bioactive candidates already exist in the ChEMBL dataset (Additional file 2: Fig. S4), our LBVS method successfully identified known anti-*Mtb* drugs that were not previously seen by model, such as rifapentine, rifalazil, and isoniazid (Additional file 3: Table S5). This demonstrates the feasibility of our approach in identifying novel anti-Mtb drug candidates. The overall distribution of the predicted bioactivities and the differences between the top predictions and overall/bottom predictions are shown in Fig. 4C and Fig. 4F. To further reduce the size of the screened candidate pool, we conducted drug category filtering to exclude antibiotics or those drugs with the indications of antibacterial infection and anti-tuberculosis infection, as we intended to find novel potential anti-Mtb indications for existing drugs. Finally, considering the availability of the compounds, 15 candidates were selected for further *in vitro* anti-Mtb assessment (Table 4).

MIC values of the selected candidates against Mtb reference strain H37Rv and MDR-TB isolates

To evaluate the accuracy of prediction results, we tested the MICs of the 15 selected candidates against the Mtb reference

Table 3. The 15 top-ranked bioactive anti-Mtb candidates screened from the DrugBank database

DrugBank	Name	CAS	Predicted bioactivity ($-\log_{10}$M)	Drug categories in DrugBank
DB14821	Macozinone	1377239-83-2	6.986	Antibacterial; anti-infective; treatment of tuberculosis
DB11753	Rifamycin	6998-60-3	6.967	Antibacterial; anti-infective; antimycobacterials
DB00615	Rifabutin	72559-06-9	6.937	Antibacterial; anti-infective; antimycobacterials; antibiotics, antitubercular
DB11637	Delamanid	681492-22-8	6.861	Antimycobacterials
DB01201	Rifapentine	61379-65-5	6.804	Antibacterial; anti-infective; antibiotics, antitubercular; antimycobacterials
DB16312	TNP-2092	922717-97-3	6.673	Lactams, macrocyclic; quinolines
DB01045	Rifampicin	13292-46-1	6.668	Antibacterial; anti-infective; antibiotics, antitubercular; antimycobacterials
DB05154	Pretomanid	187235-37-6	6.400	Antimycobacterials
DB04934	Rifalazil	129791-92-0	6.332	Antibacterial; anti-infective; antibiotics, antitubercular
DB01220	Rifaximin	80621-81-4	6.274	Antibacterial; anti-infective
DB08903	Bedaquiline	843663-66-1	6.193	Antibacterial; anti-infective; antimycobacterials
DB15213	25-Desacetylrifapentine	79039-56-8	6.175	Antibacterial; lactams, macrocyclic; rifamycins
DB00845	Clofazimine	2030-63-9	6.172	Antibacterial; anti-infective; antimycobacterials
DB04220	CGP 4832	113303-81-4	6.057	Lactams, macrocyclic
DB00218	Moxifloxacin	151096-09-2	6.010	Antibacterial; anti-infective

Table 4. The 15 selected anti-Mtb candidates and their MICs against Mtb reference strain H37Rv

DrugBank	Name	CAS	Predicted bioactivity ($-\log_{10}$M)	Drug categories in DrugBank	MIC (μg/mL)	MIC (μM/mL)
DB06638	Quarfloxin	865311-47-3	5.660	Oxazines; quinolines	12 500	20.67
DB13878	Pibrentasvir	1353900-92-1	5.582	Anti-infectives for systemic use; antivirals for systemic use	>100	>89.83
DB09143	Sonidegib	956697-53-3	5.554	Antineoplastic; benzene derivatives	>100	>205.97
DB15444	Elexacaftor	2216712-66-0	5.534	Cystic fibrosis transmembrane conductance regulator correctors	25 000	41.83
DB12526	Batefenterol	743461-65-6	5.530	Anticholinergic; muscarinic antagonists; quinolines	25 000	33.77
DB13879	Glecaprevir	1365970-03-1	5.516	Anti-infectives for systemic use; antivirals for systemic use	100 000	119.21
DB15036	Sitravatinib	1123837-84-2	5.514	Aniline compounds	100 000	158.81
DB12375	Oglemilast	778576-62-8	5.506	–	100 000	193.69
DB12270	Losmapimod	585543-15-3	5.505	Cycloparaffins; p38 mitogen-activated protein kinases	>100	>260.78
DB06013	Aldoxorubicin	1361644-26-9	5.504	Anthracyclines; antineoplastic	3.125	4.16
DB12005	Nirogacestat	1290543-63-3	5.471	Gamma secretase inhibitors and modulators	100 000	204.23
DB16347	Velsecorat	1196509-60-0	5.438	Dioxins and dioxin-like compounds; pyrazoles	>100	>164.85
DB12183	Sapitinib	848942-61-0	5.416	Heterocyclic compounds, fused-ring	>100	>211.00
DB01268	Sunitinib	557795-19-4	5.382	Antineoplastic; antineoplastic agents	50 000	125.48
DB15673	Lenacapavir	2189684-44-2	5.331	Anti-HIV; anti-infectives for systemic use; antivirals for systemic use	100 000	103.28

strain H37Rv. As shown in Table 4, among these, aldoxorubicin showed the highest antibacterial activity against H37Rv with MIC = 4.16 μM/mL. Considering aldoxorubicin is a prodrug of doxorubicin (CAS: 23214-92-8) bound to a peptide that binds albumin when entering the bloodstream, we also tested the MIC value of doxorubicin (MIC = 10.78 μM/mL). Besides, quarfloxin, elexacaftor, and batefenterol also showed relatively moderate antibacterial activity with MICs $\leq$50 μM/mL (or MICs $\leq$ 25 μg/mL) [48], while the remaining 11 inhibitors showed no activity against growth of H37Rv (Table 4). Next, the top 4 inhibitors (aldoxorubicin, quarfloxin, elexacaftor, and batefenterol) were chosen to test their antibacterial activity against MDR isolates (the drug resistance susceptibility of the MDR-TB strains are shown in Additional file 3: Table S6). Ultimately, aldoxorubicin exhibited potent activity against MDR isolates with MICs ranging from 4.16 to 16.65 μM/mL, while quarfloxin showed moderate antibacterial activity against MDR isolates with MICs ranging from 10.34 to 41.34 μM/mL (Table 5).

Table 5. MICs of aldoxorubicin, quarfloxin, batefenterol, and elexacaftor against MDR-TB isolates

Name	MIC (μM/mL) in clinical multidrug-resistant Mtb				
	MDR-TB 34 789	MDR-TB 34 832	MDR-TB 34 796	MDR-TB 34 816	MDR-TB 34 786
Quarfloxin	10.34	41.34	20.67	41.34	20.67
Elexacaftor	83.66	83.66	83.66	83.66	83.66
Batefenterol	33.77	67.55	67.55	67.55	135.09
Aldoxorubicin	4.16	8.33	8.33	8.33	16.65

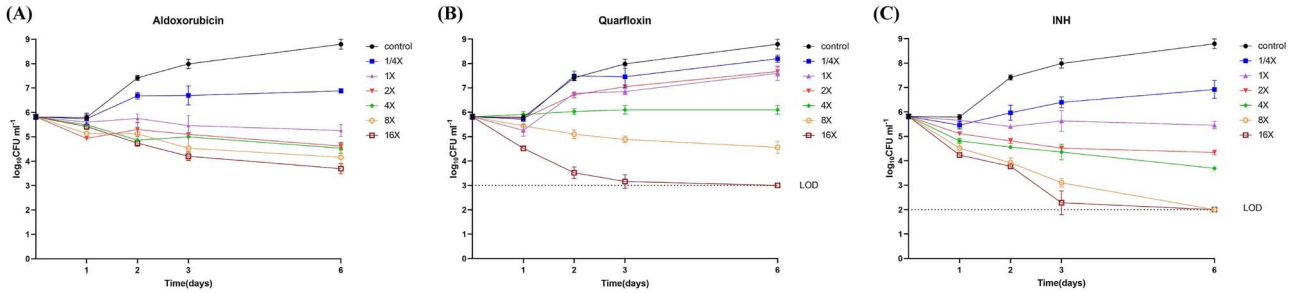


Figure 5. Time–kill curves of aldoxorubicin (A), quarfloxin (B), and INH (C) against Mtb H37Rv. Antibiotic concentrations are presented as different symbols. The INH was used as a positive control.

Time–kill curve assessments

We further evaluated the antimicrobial activity of aldoxorubicin and quarfloxin by time–kill curve analysis. In general, the patterns of H37Rv growth and killing by aldoxorubicin and quarfloxin were moderate with apparent concentration-dependent features. Killing effects were observed at all the tested concentrations in the first day for both aldoxorubicin and quarfloxin. Compared with the growth control, the time–kill curves of aldoxorubicin showed that the bacteria counts could decrease by ~1 $\log_{10}$ CFU/mL at a concentration of ≥2× MIC on day 1 (Fig. 5A). From approximately day 2 to day 6, exposure to ≥1× MIC of aldoxorubicin effectively decreased the bacterial density. On day 1, compared with the growth control, quarfloxin acquired about 0.5 $\log_{10}$ CFU/mL at 1× MIC (Fig. 5B). Quarfloxin exhibited strong antibacterial activity, especially at concentrations ≥16× MIC, and the CFUs of Mtb were all under the limit of detection on day 3. In contrast, INH demonstrated sustained antimycobacterial activity, and the bacterial counts can decrease by 1 $\log_{10}$ CFU mL^{-1}at a drug concentration of 4× MICs on day 1 (Fig. 5C), which was comparable with aldoxorubicin at the concentration of 2× MICs on day 1.

Intracellular bacterial effect of aldoxorubicin and quarfloxin

As shown in Fig. 6, after 24 h of incubation with aldoxorubicin, there was no significant change in bacterial viability. At 72 h, the bacterial survival rates decreased compared with the negative control. At a concentration of 1× MIC, aldoxorubicin acquired a mean 0.88 $\log_{10}$ CFU mL^{-1} decrease (Fig. 6A). Comparatively, quarfloxin exhibited effective inhibition at a concentration of 4× MIC at 24 h posttreatment (Fig. 6B). After 72 h posttreatment, quarfloxin showed a comparable antimycobacterial effect to INH at a concentration of 1× MIC. Briefly, quarfloxin at the concentration of 1× MIC made a 1.92 $\log_{10}$ CFU mL^{-1} decrease of *Mtb*, whereas INH had a 2.07 $\log_{10}$ CFU mL^{-1} decrease. At a concentration of 4× MIC, the CFUs were below the detection limit (Fig. 6B).

Protein–ligand binding analysis

We performed molecular docking to study the potential anti-Mtb mechanisms of aldoxorubicin and quarfloxin and their protein–ligand binding modes. Based on the previously reported targets of aldoxorubicin and quarfloxin in other species [49, 50], DNA gyrase in *M. tuberculosis* (PDB ID: 7UGW) [10] was selected as the potential target, and the protein structure was downloaded from the PDB database. Molecular docking was performed with AutoDock Vina to discover the binding modes of the two candidates. The Vina score docking results for aldoxorubicin and quarfloxin are listed in Table 6 and Table 7, respectively, with a more negative value indicating a higher affinity. The results demonstrated that aldoxorubicin and quarfloxin exhibited relatively higher molecular docking scores with DNA gyrase, ranging from −6.499 to −7.494 kcal/mol, compared to the reference compound evybactin, a known DNA gyrase inhibitor that binds to a site overlapping with synthetic thiophene poisons [10]. Evybactin showed a lower docking score of −5.834 kcal/mol (Additional file 3: Table S7). Moreover, protein–ligand interaction analysis was conducted with LigPlot+. The 3D structure of DNA gyrase, along with the binding poses and interaction profiles of the two candidate compounds and the reference compound with DNA gyrase, are shown in Fig. 7. The binding poses revealed that aldoxorubicin and quarfloxin docked into the same hydrophobic binding pocket as evybactin (Fig. 7A and 7B). Aldoxorubicin formed hydrogen bonds with residues PRO353(A) and ARG354(A), while quarfloxin formed a single hydrogen bond with PRO353(A). In contrast, the crystal structure of Mtb DNA gyrase bound to evybactin revealed hydrogen bonds with ASP350(A), ARG354(A), and TYR364(A). Notably, the shared binding pocket and residues (PRO353 and ARG354) involved in interactions with aldoxorubicin, quarfloxin, and evybactin are also targeted by azaindole or chlorophenyl groups in thiophenes—a synthetic class of gyrase poisons that act via an allosteric mechanism [10, 51]. Therefore, aldoxorubicin and quarfloxin may bind to the same allosteric site targeted by thiophenes in Mtb DNA gyrase, suggesting a possible structural basis for their anti-Mtb bioactivities.

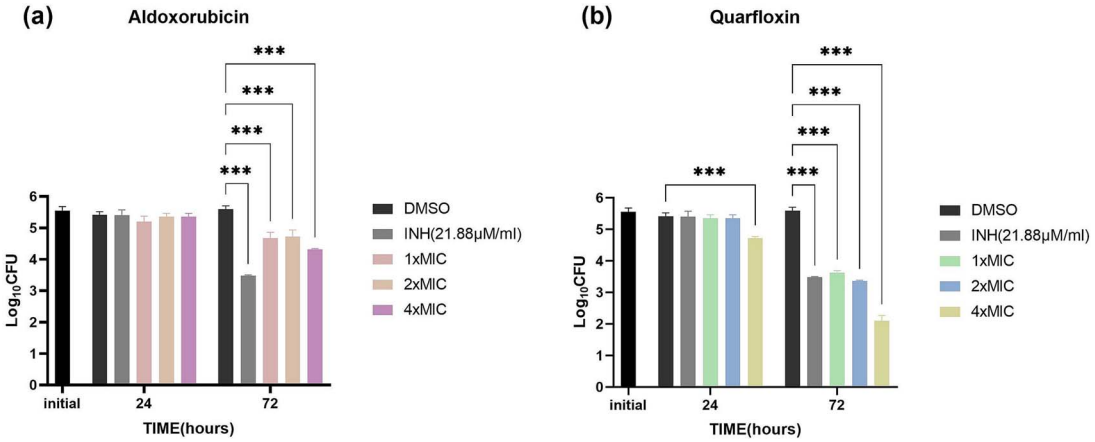


Figure 6. The intracellular survival rate of Mtb H37Rv after aldoxorubicin and quarfloxin exposure. (A) Aldoxorubicin group: infected macrophages treated with aldoxorubicin at 1× MIC, 2× MIC, and 4× MIC. (B) Quarfloxin group: infected macrophages treated with quarfloxin at 1× MIC, 2× MIC, and 4× MIC. The INH group served as the positive control. ***$P < .001$.

Table 6. Molecular docking results of aldoxorubicin (DrugBank ID: DB06013) to Mtb DNA gyrase (PDB ID: 7UGW)

Docking pose	Vina score (kcal/mol)	RMSD l.b.	RMSD u.b.
1	−6.499	0.000	0.000
2	−6.291	2.748	10.370
3	−6.255	1.298	1.613
4	−6.169	1.475	2.589
5	−6.163	2.217	2.880
6	−6.118	2.993	11.030
7	−6.054	2.527	3.571
8	−5.896	1.968	2.292
9	−5.860	2.578	3.797
10	−5.838	2.653	11.700

Table 7. Molecular docking results of quarfloxin (DrugBank ID: DB06638) to Mtb DNA gyrase (PDB ID: 7UGW)

Docking pose	Vina score (kcal/mol)	RMSD l.b.	RMSD u.b.
1	−7.494	0.000	0.000
2	−7.462	2.805	5.662
3	−7.458	2.783	5.973
4	−7.432	1.930	1.965
5	−7.387	2.067	5.069
6	−7.277	3.578	7.178
7	−7.052	2.137	5.381
8	−7.017	3.320	10.340
9	−6.860	2.371	2.830
10	−6.838	3.254	10.620

Analysis of the molecular dynamics simulation

Molecular dynamics (MD) simulations were performed to further assess the stability of the docked protein–ligand complex and their interactions throughout the entire simulation trajectories. The system properties during equilibration (Additional file 2: Fig. S5) indicate that all simulated structures were optimized and equilibrated. Following the MD production phase, root mean square deviation (RMSD) analysis was performed to assess the overall trajectory fluctuations and conformational stability of the Mtb DNA gyrase monomer and the three protein–ligand complexes: Mtb DNA gyrase–aldoxorubicin, –quarfloxin, and –evybactin. The time evolution of RMSD values (Fig. 8A) showed that the motions of all complexes converged after 100 ns. Notably, aldoxorubicin and quarfloxin complexes exhibited lower RMSD values compared to the reference ligand (evybactin) and the DNA gyrase monomer, indicating greater stability and fewer structural changes throughout the simulation period.

Then, we calculated root mean square fluctuation (RMSF) to evaluate residue flexibility across the four MD simulations (Fig. 8B). The Mtb DNA gyrase monomer exhibited higher fluctuations, while the Mtb DNA gyrase–aldoxorubicin and Mtb DNA gyrase–evybactin complexes significantly stabilized the protein structure, particularly around key residues ASP350, PRO353, ARG354, and TYR364, which form hydrogen bonds with evybactin. In contrast, the Mtb DNA gyrase–quarfloxin complex showed similar fluctuations to the monomer, indicating that quarfloxin provided only moderate stabilization of protein's residue flexibility.

Hydrogen bonds play a crucial role in mediating the binding interactions between ligands and proteins within protein–ligand complexes. The time evolution analysis of hydrogen bonds formed between the candidate ligands and Mtb DNA gyrase revealed that aldoxorubicin consistently formed between 0 and 3 hydrogen bonds throughout the MD production, while evybactin generally formed 0 to 2 hydrogen bonds, and quarfloxin typically maintained a single hydrogen bond. This comparison underscores potential differences in binding affinity, with aldoxorubicin demonstrating the strongest affinity for Mtb DNA gyrase, followed by evybactin and quarfloxin.

The overall MD simulation results confirm the thermodynamic stability and intramolecular interactions of the DNA gyrase–aldoxorubicin and DNA gyrase–quarfloxin complexes [52]. These findings are consistent with the molecular docking and experimental results, where aldoxorubicin demonstrated a higher Vina score and greater efficacy compared to quarfloxin.

Assessment of binding capacity by surface plasmon resonance

Subsequently, through surface plasmon resonance, the two selected compounds were evaluated for their binding ability to DNA gyrase. As depicted in Fig. 9A and Fig. 9B, it displayed the concentration-dependent binding of aldoxorubicin with GyrA

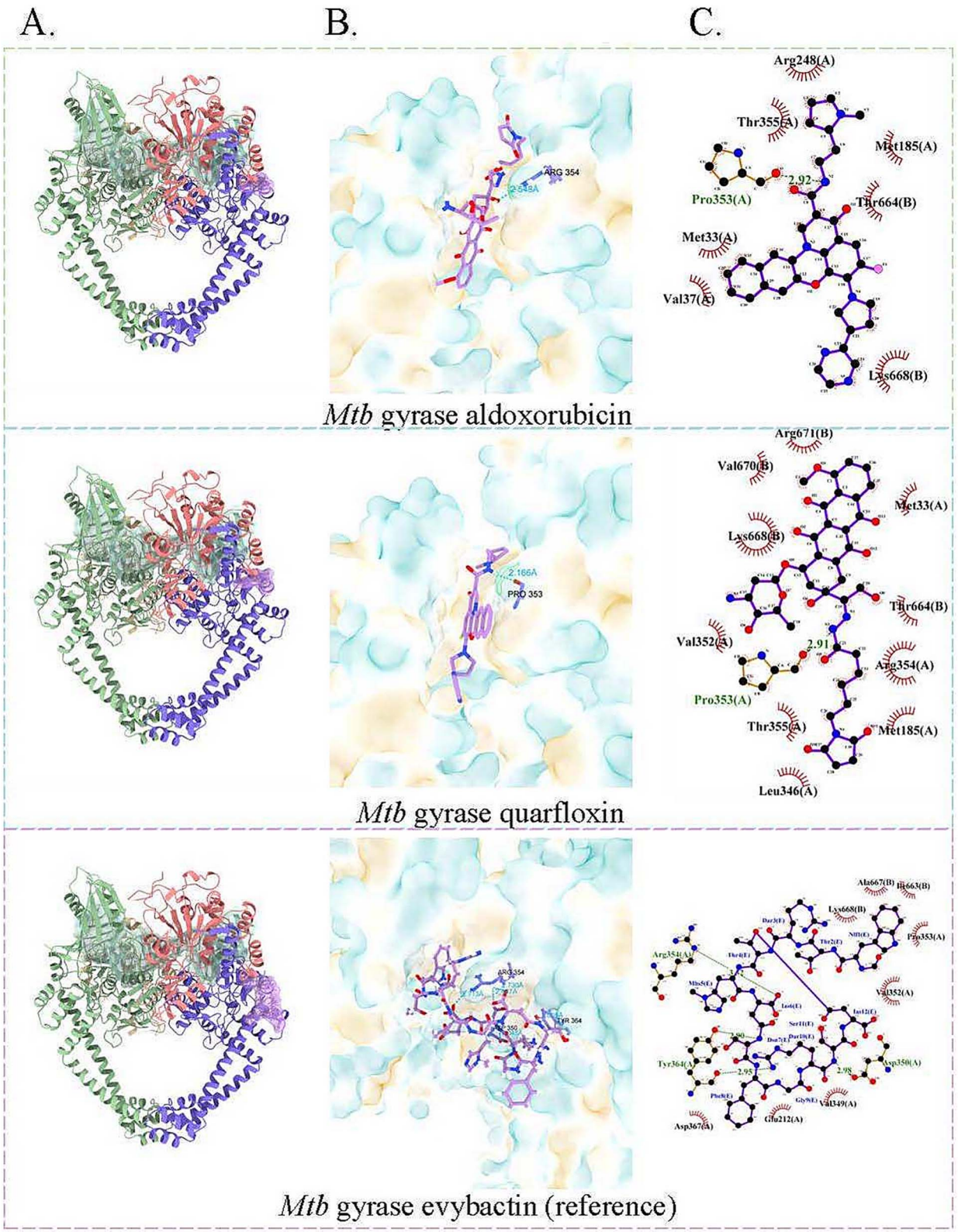


Figure 7. The protein–ligand binding modes of two identified ligands (aldoxorubicin and quarfloxin) and reference ligand (evybactin). (A) The global protein–ligand complexes of Mtb DNA gyrase (PDB ID: 7UGW) and ligands. (B) The local protein–ligand complexes and binding modes of Mtb DNA gyrase and three ligands, with protein hydrophobic surface and binding residues shown. (C) The ligand-centered binding mode profiles for three ligands.

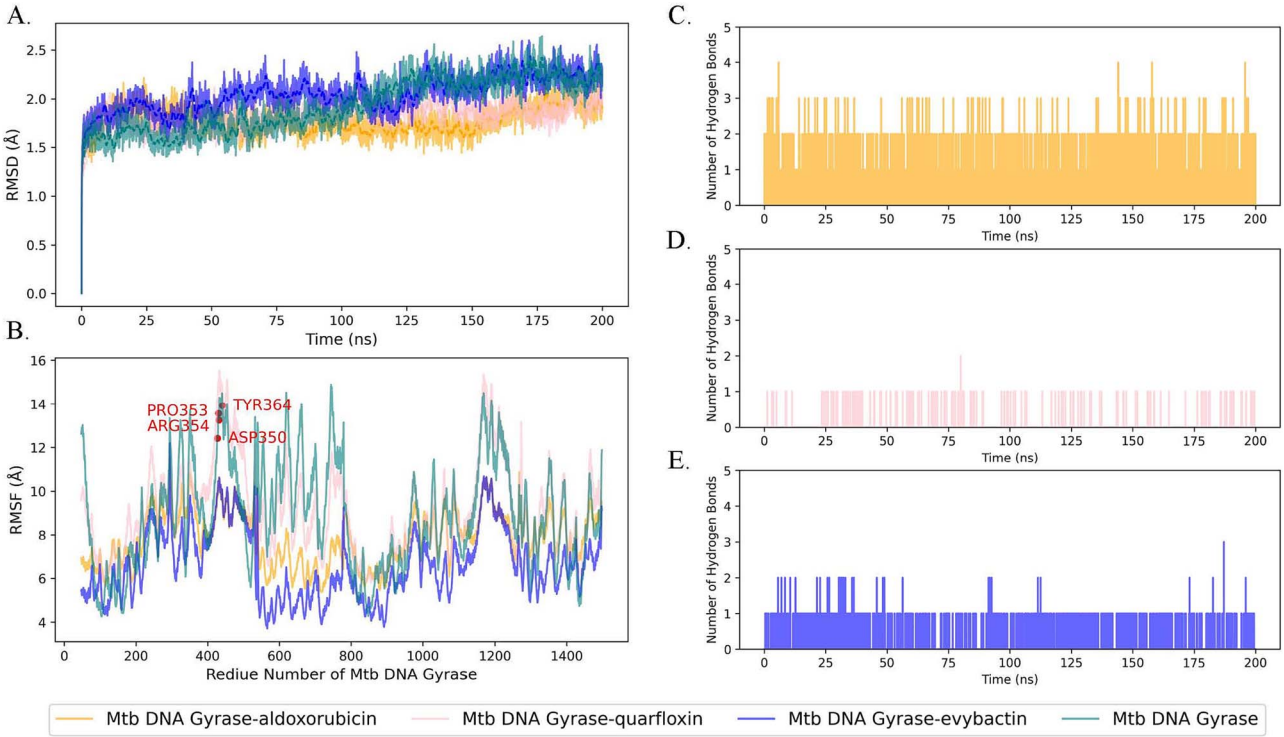


Figure 8. The MD simulation analysis results for Mtb DNA gyrase monomer and three protein–ligand complexes (Mtb DNA gyrase–aldoxorubicin, –quarfloxin, and –evybactin). (A) Time evolution of RMSD during MD production; (B) RMSF for each residues in Mtb DNA gyrase during MD production. The 1–46 numbered residues representing DNA molecule were removed. (C) Time evolution of hydrogen bonds formed between aldoxorubicin and protein. (D) Time evolution of hydrogen bonds formed between quarfloxin and protein. (E) Time evolution of hydrogen bonds formed between evybactin and protein.

and GryB with K_d value of 0.3997 and 0.3242 μM, respectively, while Fig. 9C and Fig. 9D displayed the concentration-dependent binding of quarfloxin with GyrA and GryB with K_d value of 2.891 and 30.1 μM, respectively. These results verified the direct strong binding between aldoxorubicin and the DNA gyrase, as well as between quarfloxin and the DNA gyrase.

Conclusion and discussion

In conclusion, our comprehensive virtual screening study brings insights into ML-led LBVS workflow design and provides experimental evidence of the anti-Mtb bioactivities of two predicted novel compounds, aldoxorubicin and quarfloxin. The overall results could benefit recent studies on computational flow construction and optimization of LBVS for anti-Mtb drug discovery. The initial antimycobacterial bioactivity dataset was organized from the ChEMBL database. Six ML models for anti-Mtb bioactivity prediction were trained and validated, followed with an ensemble approach to integrate the established machine learning techniques. For anti-Mtb drug repurposing, molecules in the DrugBank database were screened and the majority of known anti-Mtb drugs could be predicted with high precision. Then, further drug category filtering and experimental validation resulted two candidates (aldoxorubicin and quarfloxin) with anti-Mtb potentials, and their drug-target binding modes were further verified by molecular dynamics simulation and SPR experiments.

Among the predicted two candidates, aldoxorubicin exhibited high antimicrobial activity against Mtb both *in vitro* and in cell lines. As a prodrug of doxorubicin, aldoxorubicin has been successfully tested against glioblastoma and soft-tissue sarcoma [53, 54]. Moreover, in our study, doxorubicin, the active form of aldoxorubicin, also exhibited an MIC (10.78 μM/mL) comparable to that of aldoxorubicin (4.16 μM/mL). Doxorubicin, which has been in clinical use for >2 decades, has a very wide antitumor application spectrum; it has been used against cancers such as breast and ovarian carcinoma, sarcoma, and many other solid tumors [55]. Previous studies have shown that the Cmax of doxorubicin ranged from 14.66 to 303.46 μM/mL (corresponding to dosages of 20 to 340 mg/m^2), which corresponds to 1.4- to 28-fold the MIC of Mtb H37Rv [56]. Generally, Cmax/MIC is considered an index for the efficacy of drugs *in vivo*. Thus, the Cmax plus the low MIC values obtained in this study also implies the potential efficacy of doxorubicin in the treatment of tuberculosis. Meanwhile, compared to the known Mtb DNA gyrase binder evybactin, our newly discovered aldoxorubicin demonstrated a competitive binding affinity ($K_d = 0.3997$ μM vs. $IC_{50} = 1$ μM) for a similar protein pocket, previously identified as an allosteric site on the gyrase protein [10]. However, despite its potential role as an allosteric inhibitor, aldoxorubicin showed relatively lower anti-Mtb bioactivity compared to evybactin [MIC value: 3.125 μg/mL (4.16 μM/mL) vs. 0.25 μg/mL] [10]. Given these preliminary findings, it is important to experimentally compare aldoxorubicin with evybactin to further explore aldoxorubicin's therapeutic potential and underlying mechanisms of action. Quarfloxin is an investigational drug for different malignant tumors; it was originally derived from the fluoroquinolone class of compounds to target G-quadruplex within ribosomal DNA [57, 58]. Recent studies showed that some stable G-quadruplexes in the genome of *M. tuberculosis* were in the promoter region of genes belonging to definite functional categories [59]. Previous studies showed that the stable G-quadruplex structure could be transformed into its duplex conformation by reverse gyrase; the potential interactions

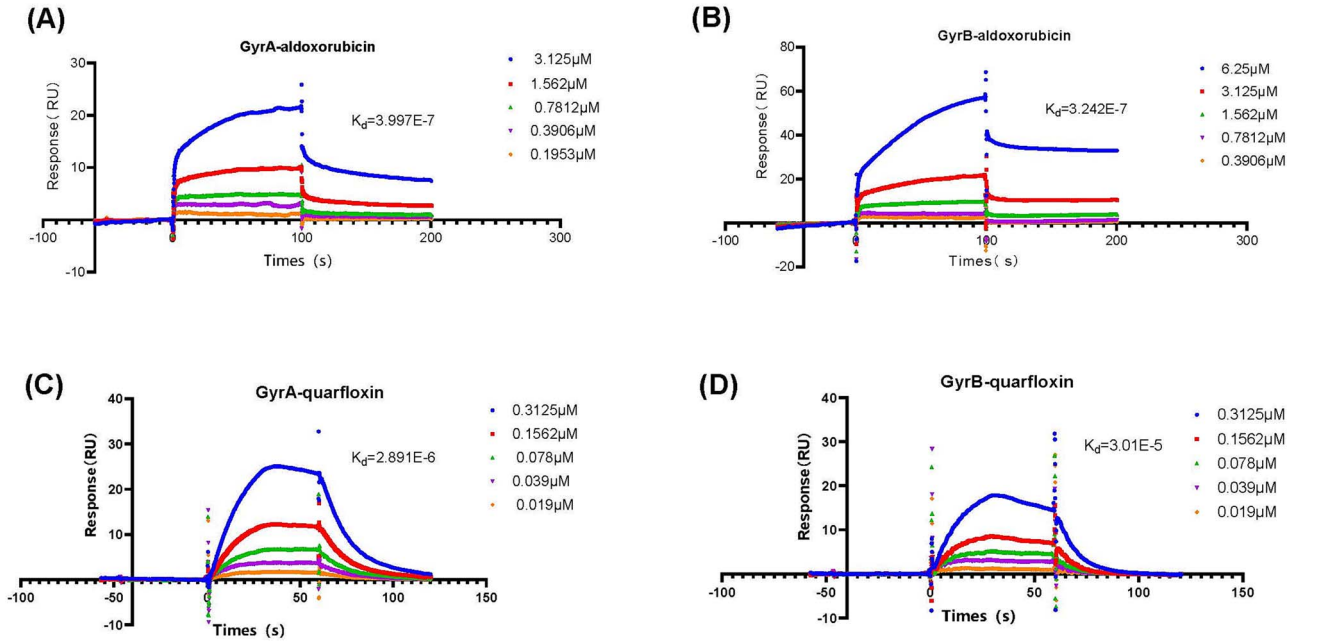


Figure 9. SPR analysis of aldoxorubicin and quarfloxin with DNA gyrase. (A) Evaluation of interaction between aldoxorubicin and GyrA. (B) Evaluation of interaction between aldoxorubicin and GyrB. (C) Evaluation of interaction between quarfloxin and GyrA. (D) Evaluation of interaction between quarfloxin and GyrB.

of quarfloxin with DNA gyrase found in this study may help to further explain the anti-Mtb mechanism of this compound.

Besides, there are certain limitations in this study. First, although the performances in terms of the enrichment factor and hit rate are moderate, the R^2, MAE, and MSE values of the LBVS methods on the external validation dataset are weak. These results implied that our trained LBVS models may have a relatively limited adaptive domain and chemical structural preference, this be alleviated by introducing the data from more bioactive anti-Mtb compounds. Second, the repurposing dataset can be further expanded to discover more candidates. Considering sample accessibility and cost, this study focused on only anti-Mtb drug repurposing with approved or investigational drugs. Considering that most of the top predicted compounds in DrugBank were known anti-Mtb drugs, therefore it would be promising to screen more investigational drugs to expand the range of compounds for repurposing. To address this issue, we also used our LBVS methods to predict the anti-Mtb bioactivities of molecules in the Drug Repurposing Hub [60] to promote the identification of more potential anti-Mtb drugs and enhance the application and precision of machine learning–led anti-Mtb drug virtual screening. Third, further studies are needed to explore and validate potential additional Mtb targets for the identified compounds. This will help deepen our understanding of the pharmacological mechanisms of action for both compounds.

Key Points

- We have created and validated an effective machine learning–enabled virtual screening workflow to repurpose existing drugs with anti-Mtb efficiency.
- We virtual screened the DrugBank dataset against Mtb and successfully identified two novel drugs (aldoxorubicin and quarfloxin) with significant antibacterial activity against *M. tuberculosis* H37Rv strain and other multidrug-resistant TB isolates.
- In our study, molecular docking analysis indicated the direct binding of the two promising compounds with Mtb DNA gyrase, which were then validated by molecular dynamics analysis and the surface plasmon resonance experiments.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the IMICAMS supporting Fund [2024YT03], the CAMS Innovation Fund for Medical Sciences [2021-I2M-1-056], the National Key R&D Program of China [2021YFF1201303, 2022YFC2703105], and Funds from Guoqiang Institute of Tsinghua University.

Data availability

The training dataset (ChEMBL), screening and predicting datasets (DrugBank, Drug Repurposing Hub) with predicted Mtb bioactivities, and related data processing and model training codes are available at https://github.com/gu-yaowen/Anti-TB.

References

1. Furin J, Cox H, Pai M. Tuberculosis. *Lancet* 2019;**393**:1642–56. https://doi.org/10.1016/S0140-6736(19)30308-3.
2. Abubakar I, Zignol M, Falzon D. *et al.* Drug-resistant tuberculosis: time for visionary political leadership. *Lancet Infect Dis* 2013;**13**:529–39. https://doi.org/10.1016/S1473-3099(13)70030-6.

3. Conradie F, Bagdasaryan TR, Borisov S. *et al.* Bedaquiline-pretomanid-linezolid regimens for drug-resistant tuberculosis. *N Engl J Med* 2022;**387**:810–23. https://doi.org/10.1056/NEJMoa2119430.
4. Li Y, Sun F, Zhang W. Bedaquiline and delamanid in the treatment of multidrug-resistant tuberculosis: promising but challenging. *Drug Dev Res* 2019;**80**:98–105. https://doi.org/10.1002/ddr.21498.
5. He W, Liu C, Liu D. *et al.* Prevalence of mycobacterium tuberculosis resistant to bedaquiline and delamanid in China. *J Glob Antimicrob Resist* 2021;**26**:241–8. https://doi.org/10.1016/j.jgar.2021.06.007.
6. Waller NJE, Cheung CY, Cook GM. *et al.* The evolution of antibiotic resistance is associated with collateral drug phenotypes in mycobacterium tuberculosis. *Nat Commun* 2023;**14**:1517. https://doi.org/10.1038/s41467-023-37184-7.
7. Gao S, Wu F, Gurcha SS. *et al.* Structural analysis of phosphoribosyltransferase-mediated cell wall precursor synthesis in mycobacterium tuberculosis. *Nat Microbiol* 2024;**9**:976–87. https://doi.org/10.1038/s41564-024-01643-8.
8. Chikhale RV, Barmade MA, Murumkar PR. *et al.* Overview of the development of DprE1 inhibitors for combating the menace of tuberculosis. *J Med Chem* 2018;**61**:8563–93. https://doi.org/10.1021/acs.jmedchem.8b00281.
9. Zhang L, Zhao Y, Gao Y. *et al.* Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol. *Science* 2020;**368**:1211–9. https://doi.org/10.1126/science.aba9102.
10. Imai Y, Hauk G, Quigley J. *et al.* Evybactin is a DNA gyrase inhibitor that selectively kills mycobacterium tuberculosis. *Nat Chem Biol* 2022;**18**:1236–44. https://doi.org/10.1038/s41589-022-01102-7.
11. Stokes JM, Yang K, Swanson K. *et al.* A deep learning approach to antibiotic discovery. *Cell* 2020;**180**:688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.
12. Kong W, Midena G, Chen Y. *et al.* Systematic review of computational methods for drug combination prediction. *Comput Struct Biotechnol J* 2022;**20**:2807–14. https://doi.org/10.1016/j.csbj.2022.05.055.
13. Gu Y, Li J, Kang H. *et al.* Employing molecular conformations for ligand-based virtual screening with equivariant graph neural network and deep multiple instance learning. *Molecules* 2023;**28**:5982. https://doi.org/10.3390/molecules28165982.
14. Kong W, Wang W, An J. Prediction of 5-hydroxytryptamine transporter inhibitors based on machine learning. *Comput Biol Chem* 2020;**87**:107303. https://doi.org/10.1016/j.compbiolchem.2020.107303.
15. Jung S, Vatheuer H, Czodrowski P. VSFlow: an open-source ligand-based virtual screening tool. *J Chem* 2023;**15**:40. https://doi.org/10.1186/s13321-023-00703-1.
16. Naz S, Farooq U, Khan S. *et al.* Pharmacophore model-based virtual screening, docking, biological evaluation and molecular dynamics simulations for inhibitors discovery against alpha-tryptophan synthase from mycobacterium tuberculosis. *J Biomol Struct Dyn* 2021;**39**:610–20. https://doi.org/10.1080/07391102.2020.1715259.
17. Zhu D, Johannsen S, Masini T. *et al.* Discovery of novel drug-like antitubercular hits targeting the MEP pathway enzyme DXPS by strategic application of ligand-based virtual screening. *Chem Sci* 2022;**13**:10686–98. https://doi.org/10.1039/D2SC02371G.
18. Hassam M, Shamsi JA, Khan A. *et al.* Prediction of inhibitory activities of small molecules against pantothenate synthetase from mycobacterium tuberculosis using machine learning models. *Comput Biol Med* 2022;**145**:105453. https://doi.org/10.1016/j.compbiomed.2022.105453.
19. Lane TR, Urbina F, Rank L. *et al.* Machine learning models for mycobacterium tuberculosis in vitro activity: prediction and target visualization. *Mol Pharm* 2022;**19**:674–89. https://doi.org/10.1021/acs.molpharmaceut.1c00791.
20. Ngidi NTP, Machaba KE, Mhlongo NN. In silico drug repurposing approach: investigation of mycobacterium tuberculosis FadD32 targeted by FDA-approved drugs. *Molecules* 2022;**27**:668. https://doi.org/10.3390/molecules27030668.
21. Ye Q, Chai X, Jiang D. *et al.* Identification of active molecules against mycobacterium tuberculosis through machine learning. *Brief Bioinform* 2021;**22**:bbab068. https://doi.org/10.1093/bib/bbab068.
22. Gaulton A, Bellis LJ, Bento AP. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7. https://doi.org/10.1093/nar/gkr777.
23. Wishart DS, Feunang YD, Guo AC. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82. https://doi.org/10.1093/nar/gkx1037.
24. Yaowen G, Si Z, Fengchun Y. *et al.* GNN-MTB: an anti-mycobacterium drug virtual screening model based on graph neural network. *Data Anal Knowl Discov* 2022;**6**:93–102. https://doi.org/10.11925/infotech.2096-3467.2022.0196.
25. Swain M. MolVS: molecule validation and standardization. Apr 12, 2018; Version 0.1.1. Available from: https://github.com/mcs07/MolVS.
26. Lane T, Russo DP, Zorn KM. *et al.* Comparing and validating machine learning models for mycobacterium tuberculosis drug discovery. *Mol Pharm* 2018;**15**:4346–60. https://doi.org/10.1021/acs.molpharmaceut.8b00083.
27. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54. https://doi.org/10.1021/ci100050t.
28. Polton DJ. Installation and operational experiences with MACCS (Molecular Access System). *Online Review* 1982;**6**:235–42. https://doi.org/10.1108/eb024099.
29. Li M, Zhou J, Hu J. *et al.* DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega* 2021;**6**:27233–8. https://doi.org/10.1021/acsomega.1c04017.
30. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
31. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907.
32. Veličković P, Cucurull G, Casanova A. *et al.* Graph attention networks. 2017arXiv:1710.10903. https://doi.org/10.48550/arXiv.1710.10903.
33. Gilmer J, Schoenholz SS, Riley PF. *et al.* Neural message passing for quantum chemistry. 2017arXiv:1704.01212. https://doi.org/10.48550/arXiv.1704.01212.
34. Xiong Z, Wang D, Liu X. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**:8749–60. https://doi.org/10.1021/acs.jmedchem.9b00959.
35. Gu Y, Zheng S, Xu Z. *et al.* An efficient curriculum learning-based strategy for molecular graph learning. *Brief Bioinform* 2022;**23**:bbac099. https://doi.org/10.1093/bib/bbac099.
36. Li X, Krumholz HM, Yip W. *et al.* Quality of primary health care in China: challenges and recommendations. *Lancet* 2020;**395**:1802–12. https://doi.org/10.1016/S0140-6736(20)30122-7.
37. Isigkeit L, Chaikuad A, Merk D. A consensus compound/bioactivity dataset for data-driven drug design and chemogenomics. *Molecules* 2022;**27**:2513. https://doi.org/10.3390/molecules27082513.
38. Yu X, Zhu R, Geng Z. *et al.* Nosiheptide harbors potent in vitro and intracellular inhibitory activities against

mycobacterium tuberculosis. *Microbiol Spectr* 2022;**10**:e0144422. https://doi.org/10.1128/spectrum.01444-22.
39. Coeck N, de Jong BC, Diels M. *et al.* Correlation of different phenotypic drug susceptibility testing methods for four fluoroquinolones in mycobacterium tuberculosis. *J Antimicrob Chemother* 2016;**71**:1233–40. https://doi.org/10.1093/jac/dkv499.
40. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61. https://doi.org/10.1002/jcc.21334.
41. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* 2011;**51**:2778–86. https://doi.org/10.1021/ci200227u.
42. Case DA, Aktulga HM, Belfon K. *et al.* AmberTools. *J Chem Inf Model* 2023;**63**:6183–91. https://doi.org/10.1021/acs.jcim.3c01153.
43. Brooks BR, Brooks CL, III, Mackerell AD, Jr. *et al.* CHARMM: the biomolecular simulation program. *J Comput Chem* 2009;**30**:1545–614. https://doi.org/10.1002/jcc.21287.
44. Wang J, Wolf RM, Caldwell JW. *et al.* Development and testing of a general amber force field. *J Comput Chem* 2004;**25**:1157–74. https://doi.org/10.1002/jcc.20035.
45. Kolaric A, Germe T, Hrast M. *et al.* Potent DNA gyrase inhibitors bind asymmetrically to their target using symmetrical bifurcated halogen bonds. *Nat Commun* 2021;**12**:150. https://doi.org/10.1038/s41467-020-20405-8.
46. Kim M, Johnson CE, Schmalstig AA. *et al.* A long-acting formulation of rifabutin is effective for prevention and treatment of mycobacterium tuberculosis. *Nat Commun* 2022;**13**:4455. https://doi.org/10.1038/s41467-022-32043-3.
47. Mok J, Lee M, Kim DK. *et al.* 9 months of delamanid, linezolid, levofloxacin, and pyrazinamide versus conventional therapy for treatment of fluoroquinolone-sensitive multidrug-resistant tuberculosis (MDR-END): a multicentre, randomised, open-label phase 2/3 non-inferiority trial in South Korea. *Lancet* 2022;**400**:1522–30. https://doi.org/10.1016/S0140-6736(22)01883-9.
48. Thongdee P, Hanwarinroj C, Pakamwong B. *et al.* Virtual screening identifies novel and potent inhibitors of mycobacterium tuberculosis PknB with antibacterial activity. *J Chem Inf Model* 2022;**62**:6508–18. https://doi.org/10.1021/acs.jcim.2c00531.
49. Buzun K, Bielawska A, Bielawski K. *et al.* DNA topoisomerases as molecular targets for anticancer drugs. *J Enzyme Inhib Med Chem* 2020;**35**:1781–99. https://doi.org/10.1080/14756366.2020.1821676.
50. Yang DZ, Okamoto K. Structural insights into G-quadruplexes: towards new anticancer drugs. *Future Med Chem* 2010;**2**:619–46. https://doi.org/10.4155/fmc.09.172.
51. Chan PF, Germe T, Bax BD. *et al.* Thiophene antibacterials that allosterically stabilize DNA-cleavage complexes with DNA gyrase. *Proc Natl Acad Sci U S A* 2017;**114**:E4492–500. https://doi.org/10.1073/pnas.1700721114.
52. Bepari AK, Reza HM. Identification of a novel inhibitor of SARS-CoV-2 3CL-PRO through virtual screening and molecular dynamics simulation. *Peerj* 2021;**9**:e11261. https://doi.org/10.7717/peerj.11261.
53. Seetharam M, Kolla KR, Ganjoo KN. Aldoxorubicin therapy for the treatment of patients with advanced soft tissue sarcoma. *Future Oncol* 2018;**14**:2323–33. https://doi.org/10.2217/fon-2018-0047.
54. Da Ros M, Iorio AL, De Gregorio V. *et al.* Aldoxorubicin and temozolomide combination in a xenograft mice model of human glioblastoma. *Oncotarget* 2018;**9**:34935–44. https://doi.org/10.18632/oncotarget.26183.
55. Zhao H, Yu J, Zhang R. *et al.* Doxorubicin prodrug-based nanomedicines for the treatment of cancer. *Eur J Med Chem* 2023;**258**:115612. https://doi.org/10.1016/j.ejmech.2023.115612.
56. Unger C, Häring B, Medinger M. *et al.* Phase I and pharmacokinetic study of the (6-maleimidocaproyl)hydrazone derivative of doxorubicin. *Clin Cancer Res* 2007;**13**:4858–66. https://doi.org/10.1158/1078-0432.CCR-06-2776.
57. Bates PJ, Laber DA, Miller DM. *et al.* Discovery and development of the G-rich oligonucleotide AS1411 as a novel treatment for cancer. *Exp Mol Pathol* 2009;**86**:151–64. https://doi.org/10.1016/j.yexmp.2009.01.004.
58. Carvalho J, Mergny JL, Salgado GF. *et al.* G-quadruplex, friend or foe: the role of the G-quartet in anticancer strategies. *Trends Mol Med* 2020;**26**:848–61. https://doi.org/10.1016/j.molmed.2020.05.002.
59. Perrone R, Lavezzo E, Riello E. *et al.* Mapping and characterization of G-quadruplexes in mycobacterium tuberculosis gene promoter regions. *Sci Rep* 2017;**7**:5743. https://doi.org/10.1038/s41598-017-05867-z.
60. Corsello SM, Bittker JA, Liu Z. *et al.* The drug repurposing hub: a next-generation drug library and information resource. *Nat Med* 2017;**23**:405. https://doi.org/10.1038/nm.4306.

Briefings in Bioinformatics, 2025, **26(1)**, bbae696
https://doi.org/10.1093/bib/bbae696
Problem Solving Protocol