ABSTRACT: It has always been a challenge to develop interventional therapies for *Mycobacterium tuberculosis*. Over the years, several attempts at developing such therapies have hit a dead-end owing to rapid mutation rates of the tubercular bacilli and their ability to lay dormant for years. Recently, cytochrome *bcc* complex (QcrB) has shown some promise as a novel target against the tubercular bacilli, with Q203 being the first molecule acting on this target. In this paper, we report the deployment of several ML-based approaches to design molecules against QcrB. Machine learning (ML) models were developed based on a data set of 350 molecules using three different sets of molecular features, i.e., MACCS keys, ECFP6 fingerprints, and Mordred descriptors. Each feature set was trained on eight ML classifier algorithms and optimized to classify molecules accurately. The support vector machine-based classifier using the ECFP6 feature set was found to be the best classifier in this study. Further, screening of the known imidazopyridine amide inhibitors demonstrated that the model correctly classified the most potent molecules as actives, hence validating the model for future applications.

INTRODUCTION

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtba) is a global health concern that is listed as one of the top 10 causes of death from a single infectious agent (ranking above HIV/AIDS since 2007). In 2019, an estimated 8.9–11.0 million people globally were affected by TB with a high death rate. Moreover, the ongoing SARS-COV-2 pandemic has confounded the progress made toward TB efforts. Traditional anti-TB drugs are known for their poor efficacy against nonreplicating bacteria as they target processes which are necessary for cell growth and replication with no effect on dormant tubercles. Consequently, treatments for both drug-susceptible TB (DS-TB) and drug-resistant TB (DR-TB) span several months. Furthermore, the current treatments for DR-TB are associated with low cure rates and high toxicity, thus necessitating the development of new and more efficacious drugs.

The novel target—cytochrome *bcc*-aa3 complex in the oxidative phosphorylation pathway has piqued the interest of researchers in this field. Bedaquiline (BDQ), the first FDA-approved antitubercular in 2012 for treating multidrug and extensively drug-resistant disease, targets the c-subunit of mycobacterial *F*<sub>1</sub>*F*<sub>0</sub> ATP synthase. Additionally, both imidazopyrimidine amide Q203, also known as telacebec, and lansoprazole sulide (LPZS) target the cytochrome *bcc* complex (QcrB), a subunit of the mycobacterial cyt-cc-cc-a3 oxidoreductase in the electron transport chain (ETC) Figure 1.

Mtba relies on the energetically efficient oxidative phosphorylation pathway to sustain its growth. This is evident from high-density mutagenesis and deletion studies that indicate
that \( \text{Mtb} \) is totally dependent on oxidative phosphorylation and cannot produce ATP by substrate level phosphorylation.\(^6\)

Mycobacteria possess two terminal oxidases which catalyze the two-electron reduction of oxygen atoms to water, namely, the proton-pump cyt-bcc-aa3 supercomplex and cytochrome \( bd \) oxidase (cyt-bd). Cytochrome \( bc \) is an intermediary in the terminal reduction of oxygen in the aerobic electron transport chains\(^3\) and transfers protons as part of the Q-cycle that differentiates it from cytochrome \( bd \). Cytochrome \( bd \) is a quinol oxidase that plays an important role in a number of physiological functions thereby allowing pathogenic and commensal bacteria to survive in anaerobic conditions. It has been shown that Q203 binds to the quinol oxidation site (\( Q_0 \)) in the cytochrome \( b \) subunit of complex III (QcrB) disrupting ATP formation to cause bacteriostasis. ETC is responsible for generating the proton motive force (PMF) by pumping protons across the membrane, and the energy from the PMF is used by ATP synthase to generate ATP molecules. As a continuous flow of PMF and ATP is essential even for the viability of nonreplicating Mtb, inhibition of this complex can eliminate nonreplicating subpopulations.\(^7\) The approval of BDQ for the treatment of DR-TB, and Q203 being in phase II clinical development, demonstrates that mycobacterial respiration could be valuable for therapeutic interventions.

One of the budding arms of computer-aided drug discovery is “machine learning” (ML). The approach of ML is quite different from the traditional physical models whose results solely rely on physical equations; on the other hand, machine learning uses algorithms that recognize patterns to establish relationships that help predict biological, chemical, and physical properties of novel compounds. ML techniques can be easily applied on enormous data sets that are quite unwieldy with physical models.\(^8\) To date, many ML techniques have been implemented to guide traditional experiments and have resulted in reducing both time and cost. In the past decade, machine learning tools have made a tremendous impact on quantitative structure activity relationship (QSAR) modeling. The ML tools have been refined and modified over time, and as a result, they are able to identify potential biologically active molecules from millions of possibilities, efficiently and easily.\(^8\)

The new approaches like multitarget QSAR (mt-QSAR)\(^9\),\(^9b\),\(^c\), multitasking quantitative-structure biological effect relationships (mt-QSBER)\(^10\),\(^10b\),\(^11\),\(^12\) for antiviral and antimicrobial activity, have been introduced that can simultaneously predict activities against multiple organisms. These mt-QSAR models work by constructing a drug–drug similarity complex network. The methods of mt-QSAR and mt-QSBER have been specifically applied to \textit{Mycobacterium tuberculosis},\(^13\) where the fragments contributing to the activity were recognized and new molecular entities identified.\(^10\),\(^14\)

The imidazopyridine amide Q203 was identified by Pethe et al. in 2013\(^{5a}\) from a set of 352 molecules that were tested against \textit{Mtb}.\(^15\) Several attempts to modify and discover a molecule more potent than Q203 have been unfruitful. The lack of an X-ray crystal structure of Mtb QcrB and the fact that homology models of the complex are not easy to build make structure-based approaches difficult. Thus, an ML-based QSAR (classification structure activity relationship) is a potential solution to this problem. Herein, we report the use of eight different ML algorithms to identify a pattern that will enable classification of molecules as active, moderately active, and inactive. To the best of our knowledge, this is an unprecedented attempt at using ML principles on QcrB inhibitors. We had previously reported the synthesis and SAR of benzyl piperezine ureas\(^15\) with \textit{in silico} studies suggesting QcrB as a plausible target. We have used the benzyl piperezine ureas (\( n = 55 \)) and molecules reported by Moraski et al.\(^17\),\(^17b\),\(^18\),\(^18b\) (\( n = 54 \)) as a validation set. The ML classifier has been made publicly accessible as a web-based app, Q-TB (https://github.com/CoutinhoLab/Q-TB.git), to enable researchers to evaluate their molecules as potential inhibitors of the QcrB complex of \textit{Mtb}.

### EXPERIMENTAL SECTION

#### Data Collection and Curation

A total of 352 compounds with their corresponding QIM (quantification of intracellular mycobacteria) values were taken from the patent WO 2011/113606 Al.\(^15\) The molecules were categorized as active (QIM <1 \( \mu M \)), moderately active (QIM 1–20 \( \mu M \)) and inactive (QIM >20 \( \mu M \)) as stated in the patent (Table 1). Compounds numbered 177 and 234 in the patent were discarded as they were found to be duplicates.

| class | active (class 1) | moderately active (class 2) | inactive (class 3) |
|-------|----------------|----------------------------|-------------------|
| range (QIM) | <1 \( \mu M \) | 1–20 \( \mu M \) | >20 \( \mu M \) |
| no. of compounds (% of total) | 214 (61%) | 58 (16.5%) | 78 (22.5%) |

The classification led to 214 molecules being defined as active, 58 as moderately active, and 78 as inactive compounds (Table 1). 20% of the 350 compounds (i.e., 70 molecules) were held back as an external validation set, while the rest were split into a training set (210 molecules) and a test set (70 molecules) at a ratio of 3:1. The 3 sets, i.e., the validation set, the training set, and the test set, had the same ratio of the three activity classes. An external validation was performed on the set of 55 molecules previously disclosed by our lab as mentioned earlier.
Table 2. Performance of the Individual Models on the Test Set*  

| models features_ML | 10-fold stratified cross validation | test set |    |    |    |    |    |
|-------------------|-----------------------------------|---------|----|----|----|----|----|
| maccs_lr          | 0.69                              | 0.70    | 0.41 | 0.91 | 0.75 | 0.82 | 0.69 | 0.30 |
| maccs_knn         | 0.65                              | 0.61    | 0.10 | 0.98 | 0.64 | 0.77 | 0.60 | 0.39 |
| maccs_svm         | 0.71                              | 0.71    | 0.44 | 0.93 | 0.77 | 0.84 | 0.67 | 0.28 |
| maccs_rf          | 0.68                              | 0.69    | 0.41 | 0.93 | 0.78 | 0.85 | 0.69 | 0.30 |
| maccs_xgb         | 0.68                              | 0.67    | 0.33 | 0.93 | 0.73 | 0.82 | 0.71 | 0.33 |
| maccs_gnb         | 0.64                              | 0.40    | 0.21 | 0.35 | 0.83 | 0.49 | 0.65 | 0.60 |
| maccs_dt          | 0.69                              | 0.70    | 0.38 | 0.88 | 0.75 | 0.81 | 0.65 | 0.30 |
| maccs_lda         | 0.66                              | 0.60    | 0.24 | 0.79 | 0.72 | 0.76 | 0.58 | 0.40 |
| ecp6_lr           | 0.73                              | 0.80    | 0.62 | 0.93 | 0.82 | 0.87 | 0.82 | 0.20 |
| ecp6_knn          | 0.69                              | 0.69    | 0.36 | 1.00 | 0.67 | 0.80 | 0.69 | 0.31 |
| ecp6_svm          | 0.71                              | 0.71    | 0.46 | 0.86 | 0.79 | 0.89 | 0.80 | 0.29 |
| ecp6_rf           | 0.75                              | 0.78    | 0.59 | 0.93 | 0.78 | 0.85 | 0.75 | 0.22 |
| ecp6_xgb          | 0.71                              | 0.73    | 0.47 | 0.91 | 0.75 | 0.82 | 0.74 | 0.27 |
| ecp6_gnb          | 0.59                              | 0.60    | 0.32 | 0.63 | 0.73 | 0.68 | 0.67 | 0.40 |
| ecp6_dt           | 0.70                              | 0.70    | 0.41 | 0.88 | 0.73 | 0.80 | 0.68 | 0.30 |
| ecp6_lda          | 0.67                              | 0.76    | 0.54 | 0.91 | 0.83 | 0.87 | 0.77 | 0.24 |
| des_lr            | 0.67                              | 0.67    | 0.35 | 0.88 | 0.73 | 0.80 | 0.78 | 0.33 |
| des_knn           | 0.67                              | 0.60    | 0.07 | 0.93 | 0.62 | 0.75 | 0.64 | 0.40 |
| des_svm           | 0.66                              | 0.7     | 0.41 | 0.91 | 0.75 | 0.82 | 0.78 | 0.30 |
| des_rf            | 0.72                              | 0.7     | 0.40 | 0.95 | 0.76 | 0.85 | 0.78 | 0.30 |
| des_xgb           | 0.7                                | 0.69    | 0.35 | 0.95 | 0.69 | 0.80 | 0.76 | 0.29 |
| des_gnb           | 0.59                              | 0.57    | 0.19 | 0.74 | 0.68 | 0.71 | 0.60 | 0.43 |
| des_dt            | 0.65                              | 0.57    | 0.25 | 0.67 | 0.74 | 0.71 | 0.62 | 0.43 |
| des_lda           | 0.57                              | 0.64    | 0.34 | 0.81 | 0.80 | 0.80 | 0.67 | 0.36 |

*The model is named by the features followed by the ML algorithm used. For example, maccs_rf is the rf classifier trained on maccs fingerprint descriptors.

**Feature Engineering, Calculation of Molecular Descriptors and Feature Selection.** A total of 1613 two-dimensional descriptors from 43 groups created on the molecular SMILES description of compounds in both the training and test sets were calculated with the python-based molecular descriptor calculator Mordred. The descriptors encoding physicochemical as well as topological properties were used to quantitatively represent each compound. Following descriptor calculation, descriptors with null values or with errors were discarded. As a result, a pruned list of 1331 descriptors was used for model building.

**Calculation of Molecular Features/Fingerprints.** In addition to the aforesaid descriptors (physicochemical and topological), two types of fingerprints, MACCS (molecular access system) keys (166 bits) and ECFP6 (extended-connectivity fingerprints) fingerprints (2047 bits), were calculated with RDKit (version 2020.09.1.0). Unlike descriptors, no preprocessing was applied to molecular fingerprints; all of them were used “as is” for model building.

**Model Building and Evaluation.** We explored eight ML algorithms, namely, Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGB), Gaussian Naïve Bayes (GBN), Decision Trees (DT), and Linear Discriminant Analysis (LDA). The classifiers were built using Python’s Scikit-learn toolkit. For all ML algorithms (except LDA and GNB), a grid-search approach with prediction accuracy as the objective was applied for optimization of hyperparameters, and the optimal parameters were determined during 10-fold cross-validation.

Logistic regression is an ML algorithm used to predict the probability of a target variable. KNN is based on the principle that all available data are stored, and a new point is classified based on the similarity to the stored data. SVM classifies the data points using a hyperplane and is effective in high-dimension spaces. An RF model builds multiple decision trees and merges them to obtain a more accurate and stable prediction. XGB is a decision tree-based ensemble ML algorithm that uses the gradient boosting framework. The Naïve Bayes method is a supervised machine learning algorithm based on the Bayes theory and operates on the assumption that each pair of features is independent. The DT classification works by splitting the data on different conditions and predicts the values of an unknown by learning the decision rules inferred by the data. LDA makes predictions by estimating the probability that a set of inputs belongs to a particular class.

**Validation Metrics.** To measure the performance of the model, several metrics, namely, accuracy (AC), sensitivity (SE) or recall, precision, specificity (SP), F1-score, receiver operating characteristic (ROC-AUC), Matthew’s correlation coefficient (MCC), and error rate (ERR), were calculated. For definitions of the metrics, the following terms (and their abbreviations) are used: number of true positives (TP), number of false positives (FP), number of true negatives (TN), and number of false negatives (FN).

Accuracy is a measure of model robustness and is given by

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

TN rate or specificity is the proportion of the actual negative cases correctly identified.
specification = \frac{TN}{(TN + FP)}

Recall or TP rate or sensitivity is the percentage of true class labels correctly identified by the model as true and is given by

recall = \frac{TP}{(TP + FN)}

Precision is the proportion of the positive cases that were correctly identified, given by

precision = \frac{TP}{(TN + FP)}

The F1 score is the harmonic mean of recall and precision, given by

F1 score = \frac{2 \times (precision \times recall)}{(precision + recall)}

MCC is a measure that takes into account the positives and negatives and is given by

MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}

Finally

ERR = \frac{FP + FN}{TP + TN + FN + FP}

Based on the TP and FP rate pairs, the corresponding AUC-ROC metric is calculated.

**Y-Randomization or Y-Scramble Tests.** Y-randomization tests are used to verify the "predictive power" of the model and are performed to evaluate any chance prediction. In each Y-randomization test, a new training set is generated by randomly shuffling the bioactivity labels (in the present case, active, moderately active, or inactive) while leaving the features data intact. A new model is built based on this shuffled training set with the same features, hyperparameters, and procedures of model training as the original ones. The resulting models are further evaluated on the test set with no random shuffling. If the new model performs worse than the one based on the original training set, it can safely be deduced that the performance of the classification model trained on the original training set is not accidental.

**Similarity Maps.** RDKit facilitates visualization of the atomic contributions to the predicted probability of the ML model. The "atomic weights" are generated by removing the bits belonging to the corresponding atom and comparing the similarity of the modification with the unmodified fingerprint. With the best performing model, the similarity maps were generated for compounds in the test set using the protocol provided by Riniker and Landrum.20

**RESULTS AND DISCUSSION**

**Classification Models.** Discerning Feature Set. Eight ML algorithms along with three kinds of features were used for model building. The optimal hyperparameters were set using GridSearch and were validated based on the AC, MCC, SE, PR, F1-score, AUC, and ERR metrics as listed in Table 2. The optimal parameters on which the models were built are listed in Table 3.

| model | parameters |
|-------|------------|
| LR    | C = 0.23, max_iter = 100, penalty = 12, solver = lbfgs |
| KNN   | n_neighbors = 5, leaf_size = 30, p = 2 |
| SVM   | C = 1.0, gamma = scale, kernel = poly |
| RF    | n_estimators = 100, criterion = gini, min_samples_split = 2 |
| XGB   | Booster = gtree, max_depth = 6, min_child_weight = 1 |
| GNB   | Priors = none, var_smoothing = 1e-09 |
| DT    | Criterion = gini, min_sample_split = 2, splitter = best |
| LDA   | Solver = svd, other parameters at default values |

**Table 3. Optimal Parameters for the Respective Models**

All models show equitable performance on the test set. The AUC values range from 0.58 to 0.82, and the MCC values span from 0.10 to 0.54. The mean value of SE is 0.86, and the error rate is 0.33; this indicates that all of the models can predict the “actives” more confidently than the moderately active and inactive molecules. The average AUC of models built with the MACCS features is 0.66, that with ECFP6 fingerprints 0.74, and that with Mordred descriptors 0.70. As the highest average AUC is returned for models with the ECFP6 features, this means that the ECFP6 features are able to recognize the underlying differences between structure and activity of the imidazopyridine amidines better than the other two feature sets.

**Leading ML Algorithm.** According to the metrics listed in Table 2, the SVM method is better than other ML methods in being able to classify the molecules with the least error rate. The SVM algorithm was applied to the data reserved for external validation. The results indicate that the SVM model built with the ECFP6 feature set performs the best.

**Y-Randomization Tests.** Y-scrambling or randomization was performed with the SVM algorithm. The classes were shuffled one hundred times while keeping the ECFP6 features intact. The performance of each "scrambled model" was examined on the test set, and the highest and lowest metrics are given in Table 4.

| models | AC  | MCC  | AUC  |
|--------|-----|------|------|
| Y1     | 0.63| 0.28 | 0.36 |
| Y2     | 0.43| −0.13| 0.65 |
| Y3     | 0.37| −0.21| 0.34 |

while lowest is 0.37; the MCC metric is highest at 0.28 and lowest at −0.21, and the highest and lowest values for AUC are 0.65 and 0.34, respectively. A table with the metrics of all 100 models generated is given in the Supporting Information (Section S4).

Overall, the AC, MCC, and AUC values for the randomization trials are lower than the corresponding values for the test set given in Table 4. This clearly indicates that the models after y-randomization are clearly unsatisfactory, and it can be safely concluded that the models built (Table 2) are not a result of chance correlation.

**Model Application.** Applicability Domain (AD). The purpose of an applicability domain is to determine the boundaries within which the model can make reliable predictions for compounds based on their similarity with the compounds on which the model was constructed. The

**Table 4. Three Y-Randomization Models Sampled from the Set of 100 to Show the Highest and Lowest Values**
compounds that satisfy the scope of the model are within the AD. In this study, the principal component analysis (PCA) bounding box was used to assess the AD of compounds contained in the training and testing sets. The ECFP6 fingerprints were used as input for the PCA; the resulting PCA bounding box scores are plotted in Figure 2. The data set was divided into internal and external sets, followed by the predictive model construction (for subsequent prediction on the external set), and it was also subjected to a 10-fold CV. As can be seen from Figure 2, the test and train compounds are within the AD of this model.

Chemical Space Analysis. The chemical space analysis is a key concept in drug discovery and helps to explore the characteristics differentiating active from moderately active and inactive molecules. The Lipinski’s rule of five (Ro5) enlists characteristics of drug-likeness for orally active drugs. The Ro5 filter was applied to the molecules in the training set. The molecular weight (MW), octanol–water partition coefficient (LogP), number of hydrogen bond acceptors (NumHAcceptors), and number of hydrogen bond donors (NumHDonors) were calculated using the RDKit library. According to the rules, all molecules in the training set fall within the Ro5 limits, i.e., MW < 500, LogP < 5, NumHAcceptors and NumHDonors < 10 (Figure 3). The box plots indicate that molecules classified as active have a higher average MW and LogP in contrast to the moderately active and inactive compounds. A scatter plot of MW as a function of LogP is shown in Figure 4, suggesting that the MW clusters in the range 400–500 Da, and the LogP ranges from 4.0 to 6.0. A large percentage of the active compounds have structures that are comparatively larger than

![PCA bounding box](https://example.com/image1.png)

**Figure 2.** PCA bounding box for assessing the applicability domain, internal training set (red), and external test set (purple).

![Chemical space analysis plots](https://example.com/image2.png)

**Figure 3.** Lipinski’s rule of five plots for QcrB inhibitors (training set). The bioactivity class 1 is the actives, class 2 the moderately actives, and class 3 the inactives. The plots are as follows: top left, bioactivity class vs molecular weight; top right, bioactivity class vs LogP; bottom left, bioactivity class vs number of hydrogen-atom acceptors; bottom right, bioactivity class vs number of hydrogen-atom donors.
the inactive compounds, as observed from the mean value of the box plots (Figure 3).

Model Applied to the Benzyl Piperazine Data Set. The data set disclosed by our group\(^1\) comprising 55 compounds has an activity span from 1 to \(>20\) μM. The molecules were classified according to the cutoffs used on the training and test sets. The ECFP6 feature was calculated for these 55 compounds, and the svm_ecfp6 model was applied. Twenty two compounds are predicted as moderately active and 33 inactive while none are predicted as active. Looking at the classification of the data set, 23 molecules have been predicted correctly. To put this prediction in context, we note that the benzyl piperazine molecules are in a chemical space distinct from the area occupied by the training set on which the model was established; this could be the source of the variance in the predictions.

Model Tested on the Imidazopyridine Amides Data Set. Moraski et al.\(^1\) have explored the SAR of imidazopyridine amides, which are postulated to act via inhibition of QcrB. A data set of 54 compounds reported by them was curated. The ECFP6 feature was calculated for these molecules, and the svm_ecfp6 model was constructed for these molecules. The model classifies as active all molecules which are reported as most potent in the respective publications. This indicates that the model is able to handle and predict compounds belonging to the imidazopyridine amides well. Similarity maps were generated, as shown in Figure 5 for all molecules predicted as active. The green contours highlight the core that is similar to Q203, which is the basis for the prediction. Beside this, the red contours highlight regions that have a positive influence and also add to the activity.

Model Tested on Other Chemical Data Sets. Chemical data sets apart from the imidazopyridine amides, which have been tested and identified as potential QcrB inhibitors as described in the literature, were curated and predicted using the leading model. Most compounds (Figure 6) were categorized satisfactorily. The Tanimoto similarity value for each molecule was calculated using RDKit with Q203 as the reference. The similarity values are given for each molecule in Figure 6. The most potent molecules—4, a trifluoromimidazo carboxamide;\(^\text{17a}\) 5, a pentafluorosulfanyl imidazo carboxamide;\(^\text{17a}\) and 6 and 8 from imidazopyridines\(^\text{22,23}\)—are reported to exhibit an inhibition profile like Q203; all of these molecules are correctly identified as belonging to class 1.
The molecules 7, an imidazothiazole carboxamide;9, an imidazopyridine;4, 10, with an aminoquinazoline core;12, 13, a pyrrolopyridinone;24 and 14, a phenoxyalkyl benzimidazole25 are less potent than Q203 and are correctly identified as belonging to class 2. Finally 11, a quinazoline molecule, poorly inhibits mycobacteria and has been identified as inactive or class 3.

The arylvinlypiperazine amide compound 14 is considered to be active; however, this was not correctly identified by the model. We suspect that this anomaly may be due to its binding mode which is reported to be different from Q203. Likewise, compound 15, a morpholino thiophene, is also predicted as inactive; this variance in prediction could be attributed either to the chemical space of its scaffold that is not a part of the training set or to the fact that its potency is lower than Q203.

The internal and external validation results suggest that the model shows reasonable accuracy in classifying bioactivity profiles of imidazopyridine amides and many other chemical classes, toward Mtb QcrB.

**Screening of the PubChem Database.** In a quest to find probable active molecules, a set of 211 molecules with the imidazopyridine core was retrieved from the PubChem database. The ECFP6 feature was calculated for all of the molecules. The molecules lie in the same chemical space as the training molecules. Further, the svm_ecfp6 model was run on this set of compounds. A total of 110 molecules are predicted as active, 35 as moderately active, and 66 as inactive molecules.

**Model Deployment as a Q-TB Web Application.** The ECFP6 model was deployed as an app named Q-TB (Figure 7) using streamlit, to enable researchers to test their compounds.
as probable QcrB inhibitors. The web application is enabled to accept a smile string in csv format as the molecular input. The name of the input file should be set to "Test.csv" with the column name having the structure in SMILES format as "Smiles"; any deviation from this will lead to an error. This is then submitted to the app. The ECFP6 descriptor is calculated using the RDKit package, and the svm_ecfp6 model is then applied to the input molecule; the app predicts the bioactivity class, which is labeled accordingly as class 1, class 2, or class 3. The app allows researchers with little to no background in machine learning to predict the activity of their compounds. The web app along with the manual can be accessed at https://github.com/CoutinhoLab/Q-TB.

**CONCLUSIONS**

ML is being widely used in all areas of science, including drug discovery to predict bioactivity and physicochemical properties. QcrB of Mtb is a novel target that is rigorously being explored for development of new anti-TB drugs. Our search for QcrB inhibitors of Mtb engaged the application of ML methods on a data set of 350 imidazopyridine amides curated from the literature. Three distinct classes of molecular descriptors were calculated, and eight different ML algorithms were applied to the data sets. Of the 24 models built, support vector machine was selected as the appropriate algorithm for classification of the data set based on various performance metrics. The model was further analyzed using a validation set. To complete validation of the ecfp6_svm model, Y-randomization was carried out. The model was applied to a known set of imidazopyridine amides, and it was found to correctly classify (according to the literature) all potent molecules as active. New QcrB inhibitors were identified using the model to predict the bioactivity on a data set downloaded from PubChem. Lastly, the classifier model was deployed as a web application for public usage.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01613.

Data set, imidazopyridine amide validation data set, benzylpiperazine validation data set, evaluation results of the model, and Y-scrambling results (PDF)

**AUTHOR INFORMATION**

Corresponding Author

Evans C. Coutinho — Department of Pharmaceutical Chemistry, Vasvik Research Centre, Bombay College of Pharmacy, Mumbai 400 098, India; Email: evans.coutinho@bcp.edu.in

Authors

Afreen A. Khan — Department of Pharmaceutical Chemistry, Vasvik Research Centre, Bombay College of Pharmacy, Mumbai 400 098, India; orcid.org/0000-0001-5895-3216

Sannidhi S. Poojary — Department of Pharmaceutical Chemistry, Vasvik Research Centre, Bombay College of Pharmacy, Mumbai 400 098, India

Ketki K. Bhave — Department of Pharmaceutical Chemistry, Vasvik Research Centre, Bombay College of Pharmacy, Mumbai 400 098, India

Santosh R. Nandan — Ambernath Organics Pvt. Ltd., Mumbai 400 093, India

Krishna R. Iyer — Department of Pharmaceutical Chemistry, Vasvik Research Centre, Bombay College of Pharmacy, Mumbai 400 098, India

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c01613

**Author Contributions**

S.S.P. and K.K.B. contributed equally. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. A.A.K.: Original idea, conducted research, wrote the code, analyzed the results, and drafted the paper. S.S.P. and K.K.B.: curated the data set and ran the code. S.R.N., K.R.I., and E.C.C.: analyzed the results, revised the paper and supervised the research.

**Funding**

This work was supported by the Amrut Mody Research Foundation (AMRF) at Bombay College of Pharmacy.

**Notes**

The authors declare no competing financial interest.

**Data and Software Availability**

The source code and manual for the application is hosted on GitHub: https://github.com/CoutinhoLab/Q-TB. The training, test, and validation data used to develop the model are included in the Supporting Information section and can also be found on https://github.com/CoutinhoLab/Q-TB/data.

**ACKNOWLEDGMENTS**

We thank Dr. Elvis Martis, Bombay College of Pharmacy, for his suggestions and for carefully reading the manuscript.

**REFERENCES**

(1) WHO. Global Tuberculosis Report, 2020.
(2) Togun, T.; Kampmann, B.; Stoker, N. G.; Lipman, M. Anticipating the impact of the COVID-19 pandemic on TB patients and TB control programmes. Ann. Clin. Microbiol. Antimicrob. 2020, 19 (1), 21.
(3) Poo, C. S.-Y.; Pethe, K.; Lupien, A. Oxidative Phosphorylation—an Update on a New, Essential Target Space for Drug Discovery in Mycobacterium tuberculosis. Appl. Sci. 2020, 10 (7), 2339.
(4) Arora, K.; Ochoa-Montano, B.; Tsang, P. S.; Blundell, T. L.; Dawes, S. S.; Mizrahi, V.; Bayliss, T.; Mackenzie, C. J.; Cleghorn, L. A.; Ray, P. C.; Wyatt, P. G.; Uh, E.; Lee, J.; Barry, C. E.; Boshoff, H. I. Respiratory flexibility in response to inhibition of cytochrome C oxidase in Mycobacterium tuberculosis. Antimicrob. Agents Chemother. 2014, 58 (11), 6962–5.
(5) (a) Pethe, K.; Bifani, P.; Jang, J.; Kang, S.; Park, S.; Ahn, S.; Jiricek, J.; Jung, J.; Jeon, H. K.; Cechetto, J.; Christophe, T.; Lee, H.; Kempf, M.; Jackson, M.; Lenaerts, A. J.; Pham, H.; Jones, V.; Seo, M. J.; Kim, Y. M.; Seo, M.; Seo, J. J.; Park, D.; Ko, Y.; Choi, I.; Kim, R.; Kim, S. Y.; Lim, S.; Yim, S. A.; Nam, J.; Kang, H.; Kwon, H.; Oh, C. T.; Cho, Y.; Jang, Y.; Kim, J.; Chua, A.; Tan, B. H.; Nanjundappa, M. B.; Rao, S. P.; Barnes, W. S.; Wintjens, R.; Walker, J. R.; Alonso, S.; Lee, S.; Kim, J.; Oh, S.; Oh, T.; Nehrbass, U.; Han, S. J.; No, Z.; Lee, J.; Brodin, P.; Cho, S. N.; Nam, K.; Kim, J. Discovery of Q203, a potent clinical candidate for the treatment of tuberculosis. Nat. Med. 2013, 19 (9), 1157–60. (b) Rybniker, J.; Vocat, A.; Sala, C.; Busso, P.; Pojer, F.; Benjak, A.; Cole, S. T. Lansomprazole Is An Antituberculous Prodrug Targeting Cytochrome bc 1. Nat. Commun. 2015, 6 (1), 1–8.
(6) Bald, D.; Villellas, C.; Lu, P.; Koul, A. Targeting energy metabolism in Mycobacterium tuberculosis, a new paradigm in antimycobacterial drug discovery. mBio 2017, 8 (2), No. e00272-17.
In Mycobacterium tuberculosis.

ACS Infect. Dis. 2019, 5 (2), 239–249.

Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today 2017, 22 (11), 1680–1685.

Speck-Planche, A.; Cordeiro, M. Multi-target Qsar approaches for modeling protein inhibitors. Simultaneous prediction of activities against biomacromolecules present in gram-negative bacteria. Current Topics In Medicinal Chemistry 2015, 15 (18), 1801–1813. (b) Prado-Prado, F. J.; de la Vega, O. M.; Uriarte, E.; Ubeira, F. M.; Chou, K.-C.; González-Díaz, H. Unified Qsar approach to antimicrobials. 4. Multi-target Qsar modeling and comparative multi-distance study of the giant components of antiviral drug complexes. Bioorg. Med. Chem. Lett. 2009, 19 (2), 569–575.

Kleandrea, V. V.; Speck-Planche, A. The Qsar paradigm in fragment-based drug discovery: from the virtual generation of target inhibitors to multi-scale modeling. Mini-Rev. Med. Chem. 2020, 20 (14), 1357–1374.

Planche, A. S. Advanced chemoinformatic models to speed up the discovery of antimicrobial agents: From protein inhibitors to antibacterial compounds and back. Universidade do Porto (Portugal), 2016.

Kleandrea, V. V.; Scotti, M. T.; Speck-Planche, A. Computational drug repurposing for antituberculosis therapy: Discovery of multi-strain inhibitors. Antibiotics 2021, 10 (8), 1005.

Speck-Planche, A.; V Kleandrea, V.; Luan, F.; Cordeiro, M. D. S. Advanced in silico approaches for drug discovery: Mining information from multiple biological and chemical data through mtq-Qsar and pt-qsprr Strategies. Curr. Med. Chem. 2017, 24 (16), 1667–1704.

Planche, A. S. Advanced chemoinformatic models to speed up the discovery of antimicrobial agents: From protein inhibitors to antibacterial compounds and back. Universidade do Porto (Portugal), 2016.

Kleandrea, V. V.; Scotti, M. T.; Speck-Planche, A. Computational drug repurposing for antituberculosis therapy: Discovery of multi-strain inhibitors. Antibiotics 2021, 10 (8), 1005.

Speck-Planche, A.; V Kleandrea, V.; Luan, F.; Cordeiro, M. D. S. Advanced in silico and virtual screening of multi-target inhibitors for proteins in Mycobacterium tuberculosis. Comb. Chem. High Throughput Screening 2012, 15 (8), 666–673.

No, Z.; Kim, J.; Brodin, P. B.; Seo, M. J.; Kim, Y. M.; Cechetto, J.; Jeon, H.; Genovesio, A.; Lee, S.; Kang, S.; Ewann, F. A.; Nam, J. Y.; Christophe, T.; Fenistestein, D.; Philipe, C.; Jamun, H.; Jiyeon, J. Anti- Infective Compounds. WO2011113606, 2011.

Satish, S.; Chitr, R.; Kori, A.; Sharma, B.; Putton, J.; Khan, A. A.; Desle, D.; Raikumar, K.; Korkegian, A.; Martins, E. A. F.; Iyer, K. R.; Coutinho, E. C.; Parish, T.; Nandan, S. Design, synthesis and SAR of antitubercular benzympiperazine ureas. Mol. Diversity 2022, 26 (1), 73–96.

Moraski, G. C.; Bristol, R.; Seeger, N.; Boshoff, H. L.; Tsang, P. S.; Miller, M. J. Preparation and Evaluation of Potent Pentfluorosulfanyl-Substituted Anti-Tuberculosis Compounds. ChemMedChem. 2017, 12 (14), 1108–1115. (b) Moraski, G. C.; Deboorseo, N.; Marshall, K. L.; Weaver, H. A.; Vandeputte, A.; Hastings, C.; Woolhiser, L.; Lenaerts, A. J.; Brodin, P.; Miller, M. J. Intracellular and in vivo evaluation of imidazo[2,1-b]thiazole-5-carboxamide anti-tuberculosis compounds. PLoS One 2020, 15 (1), No. e027724.

Moraski, G. C.; Markley, L. D.; Cramer, J.; Hipkind, P. A.; Boshoff, H.; Bailey, M. A.; Alling, T.; Ollinger, J.; Parish, T.; Miller, M. J. Advancement of imidazo [1,2-a] pyridines with improved pharmacokinetics and mAM activity vs. Mycobacterium tuberculosis. ACS Med. Chem. Lett. 2013, 4 (7), 675–679. (b) Moraski, G. C.; Markley, L. D.; Hipkind, P. A.; Boshoff, H.; Cho, S.; Franzblau, S. G.; Miller, M. J. Advent of Imidazol[1,2-a]pyridine-3-carboxamides with Potent Multi- and Extended Drug Resistant Antituberculosis Activity. ACS Med. Chem. Lett. 2011, 2 (6), 466–470.

Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. J. Cheminf. 2018, 10 (1), 4.

Rink, S.; Landrum, G. A. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. J. Cheminf. 2013, 5 (1), 43.

(a) Moraski, G. C.; Miller, P. A.; Bailey, M. A.; Ollinger, J.; Parish, T.; Boshoff, H. L.; Cho, S.; Anderson, J. R.; Mulugeta, S.; Franzblau, S. G. Putting tuberculosis (TB) to rest: transformation of the sleep aid, Ambien, and "anagrams" generated potent antituberculosis agents. ACS Infect. Dis. 2015, 1 (2), 85–90. (b) Moraski, G. C.; Oliver, A. G.; Markley, L. D.; Cho, S.; Franzblau, S. G.; Miller, M. J. Scaffold-switching: an exploration of 5,6-fused bicyclic heteroaromatics systems to afford antituberculosis activity akin to the imidazo[1,2-a]pyridine-3-carboxylate. Bioorg. Med. Chem. Lett. 2014, 24 (15), 3493–8.