Chemical Information

mycoCSM: using graph-based signatures to identify safe potent hits against Mycobacteria

Douglas Pires, and David B. Ascher

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.0c00362 • Publication Date (Web): 02 Jul 2020

mycoCSM: using graph-based signatures to identify safe potent hits against Mycobacteria

Douglas E. V. Pires[1,2,3,*] and David B. Ascher[1,2,4,*]

[1]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, 75 Commercial Rd, Melbourne VIC 3004.

[2]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, 30 Flemington Rd, Parkville VIC 3052.

[3]School of Computing and Information Systems, University of Melbourne, Parkville VIC 3052.

[4]Department of Biochemistry, University of Cambridge, 80 Tennis Ct Rd, Cambridge CB2 1GA.

*To whom correspondence should be addressed.

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

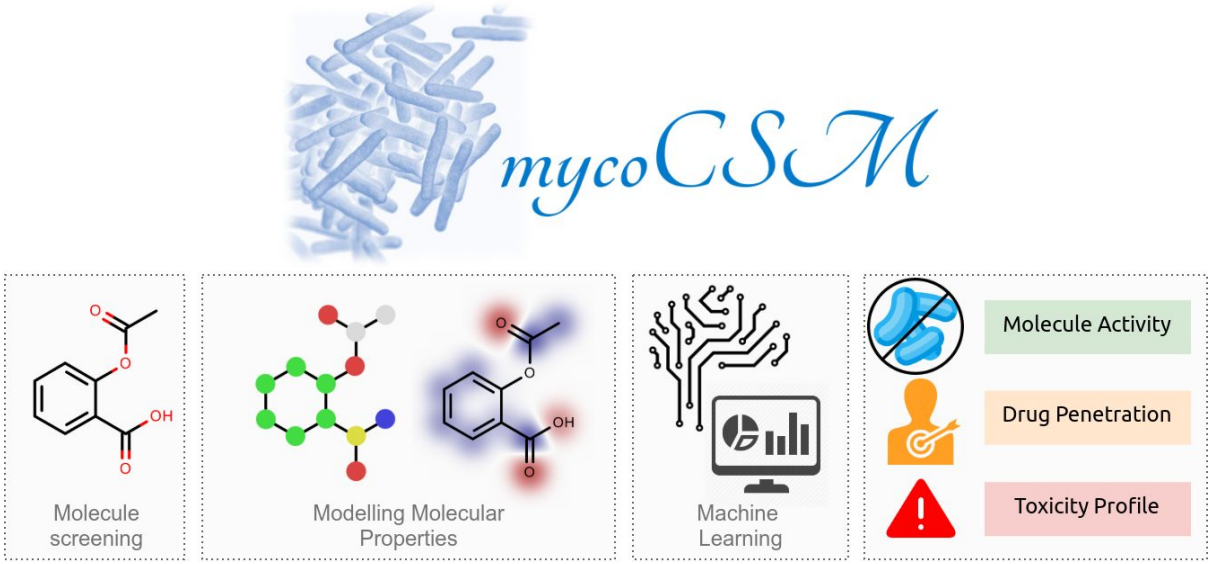
ABSTRACT

Development of new potent, safe drugs to treat Mycobacteria has proven to be challenging, with limited hit rates of initial screens restricting subsequent development efforts. Despite significant efforts and evolution of Quantitative Structure-Activity Relationship (QSAR) as well as machine learning-based models for computationally predicting molecule bioactivity, there is an unmet need for efficient and reliable methods for identifying biologically active compounds against mycobacterium that are also safe for humans. Here we have developed mycoCSM, a graph-based signature approach to rapidly identify compounds likely to be active against bacteria from the genus Mycobacterium, or against specific Mycobacteria species. mycoCSM was trained and validated on eight organism-specific and for the first time a general Mycobacteria data set, achieving correlation coefficients of up to 0.89 on cross-validation and 0.88 on independent blind tests, when predicting bioactivity in terms of Minimum Inhibitory Concentration (MIC). In addition, we also developed a predictor to identify those compounds likely to penetrate in necrotic tuberculosis foci, which achieved a correlation coefficient of 0.75. Together with a built-in estimator of the Maximum Tolerated Dose in humans, we believe this method will provide a valuable resource to enrich screening libraries with potent, safe molecules. To provide simple guidance in the selection of libraries with favourable anti-Mycobacteria properties, we have made mycoCSM freely available at: https://biosig.unimelb.edu.au/myco_csm.

KEYWORDS

Mycobacteria; Small molecule screening; Graph-based signatures; Machine-learning

Graphical Abstract



INTRODUCTION

Mycobacteria are a family of gram-positive bacilli, that are the causative agents of tuberculosis (*Mycobacterium tuberculosis*), leprosy (*Mycobacterium leprae*), and serious complications in cystic fibrosis (*Mycobacterium abscessus*). These infections are often hard to treat, in particular due to their unique cell wall, with full treatment regimens being time consuming, costly and associated with a range of side effects [1]. This has been further complicated by the spread of resistance against the major treatments, and that only a few antibiotics with new modes of action have been approved in the last 40 years [2-4]. There is, therefore, an increasing necessity for new and more efficient chemotherapies active against Mycobacteria.

Towards this, there have been coordinated efforts to perform and release the results from phenotypic screens and drug development efforts, leading to the accumulation of a large number of experimental data points of active and inactive compounds for different Mycobacteria species. However, most screening efforts are generally associated with a low hit-rate, and can only screen a fraction of the available chemical space. Further, it can be challenging to develop these molecules into potent, safe chemotherapies. The ability to rationally identify safe but potentially effective molecules computationally would significantly reduce development time and costs.

A few efforts to identify molecules likely to be effective against Mycobacterium tuberculosis have shown that such an approach could be effective, but have been limited by qualitative, QSAR and drug repurposing approaches [2, 5-9]. Particularly QSAR models have been focused on compound classes. Ragno and colleagues analysed the efficacy of antifungal pyrrole derivatives as antitubercular agents, deriving QSAR and molecular field analysis models[6], while Sivakumar and colleagues, have focused on developing QSAR models for chalcones and flavonoids[7]. More broadly applicable models are necessary. One example is the multitasking model based on quantitative-structure biological effect

relationships (mtk-QSBER) that enabled identification of antimycobacterial activity as well as their pharmacokinetics profile[9].

Previously we have shown that the application of graph-based signatures can be a very efficient way of representing molecular 3D space in order to accurately predict pharmacokinetic properties [10, 11] and the effects of mutations on protein structure and function [12-22]. Using this concept, here we developed a new machine learning method, mycoCSM, that for the first time can accurately predict molecules that are likely to be active against multiple Mycobacteria species, while remaining safe and well tolerated. Figure 1 depicts the general methodological workflow for mycoCSM.

RESULTS AND DISCUSSION

Correlating molecular properties with biological activity

A large and diverse data set of experimental Minimal Inhibitory Concentration (MIC) for molecules against 8 species of the genus Mycobacteria was collected from the literature, including Mycobacterium avium, Mycobacterium bovis, Mycobacterium fortuitum, Mycobacterium intracellulare, Mycobacterium kansasii, Mycobacterium phlei, Mycobacterium smegmatis, and Mycobacterium tuberculosis. This led to experimental MIC values for over 15,000 unique compounds across different organisms (Table 1). Figure S1 depicts the distribution of general physicochemical properties for molecules with anti-Mycobacteria activity, as well as the distribution of their biological activity measurements. Most of the molecules conformed to the Lipinski 'Rule of 5' [23], perhaps reflecting a bias in the original screening libraries.

To better understand what makes a good hit while searching for anti-Mycobacterial molecules, we initially evaluated whether any basic molecular properties (whose distribution was depicted in Figure S1) correlated strongly with biological activity. No strong correlation between molecular properties and biological activity was identified (Pearson's correlation of up to 0.27, data not shown), reflecting

a need for more sophisticated ways to model small-molecule geometry and chemistry. Interestingly, however, the top 10% of most active molecules (MIC < 1 μM) tended to have a slightly larger number of hydrogen bond acceptors, rings and a larger topological polar surface area (TPSA) (p-value < 0.001, using two-sample Kolmogorov-Smirnov test). This may represent the increasing molecular complexity needed in the evolution of hit to lead type molecules.

Predicting organism-specific activity

Predictive models were trained using supervised learning algorithms for each of the eight Mycobacterium species, using graph-based signatures and RDkit descriptors (Table S1) as evidence. The best performing models, after greedy forward feature selection, achieved Pearson's correlation coefficients during cross validation ranging from 0.80 (for *M. bovis*) to 0.89 (for *M. fortuitum*) (Table 2; Table S2). Figure 2 depicts the distribution of predicted vs. experimental MIC values per model for 10-fold cross validation, also highlighting the performances on 90% of the data (after 10% outlier removal). The models were further validated using independent blind tests (Figure S2). We observed, for every model, a consistent performance between cross validation and blind tests, indicating model generalization and reducing risk of overfitting. During blind tests, correlations ranged from 0.76 (*M. smegmatis*) and 0.88 (*M. avium*) (Table 2). Within our dataset we identified around 4,000 compounds with multiple separate experimental MIC measurements against *M. tuberculosis*. The Pearson's correlation between these separate experimental measurements was 0.78, suggesting that the predictive performance of our final models is comparable to the level of experimental variation observed, and the theoretical maximal achievable predictive performance.

We further investigated the performance of organism-specific methods on molecules that do not conform to Lipinski's rule of 5 (Ro5). Despite the bias in the training set towards Ro5 molecules, we saw no bias towards drug-like molecules, with similar performances between Ro5 (r = 0.80) and non-Ro5 molecules (r = 0.83) for the *M. tuberculosis* predictor.

To the best of our knowledge this is the first attempt at developing Mycobacterium species-specific bioactivity predictors, apart from qualitative predictions of *M. tuberculosis* activity [5, 24]. As mycoCSM quantitatively predicts bioactivity of compounds, allowing ranking and prioritization, performance comparison with other methods was done on a classification-by-regression manner. Prathipati et al. (2008) used an *M. tuberculosis* MIC < 5µM as a cutoff for labelling compounds as active, reporting an accuracy of up to 0.87 on their bayesian model. On the same data set, Yu and Wild (2012) reported a rule-based classification system, which achieved an F1-score of 0.74. By using the same cutoff, our model obtained an accuracy of 0.88 and F1-score of 0.72, comparable to previously reported performance.

Predicting drug penetration in *M. tuberculosis*

The effectiveness of drugs to treat M. tuberculosis has been linked to their ability to penetrate the cellular and necrotic regions of granulomas [25]. Poor drug penetration has been associated with poor diffusion through the caseous center, due to high protein binding in the caseum. Favourable caseum distribution is considered an important antitubercular drug property, therefore, in addition to predicting bioactivity, a model for predicting drug penetration in M. tuberculosis lesions was also developed. Using a data set of 279 compounds with experimentally characterised caseum distribution profiles, we investigated whether any molecular properties were associated with better drug penetration. We identified three main physiochemical properties that were correlated with favourable distribution. Compound molecular weight and surface were moderately predictive of caseum binding (r = -0.50), with larger molecules presenting lower fractions unbound and hence higher levels of caseum binding, while logP was also mildly predictive (r = -0.60), with more hydrophilic compounds displaying better distribution. We also identified a negative correlation between drug penetration and the negative logarithm of the MIC (r = -0.64, Figure S3), consistent with current

thoughts that more potent compounds (low MIC) are more likely to bind to caseum (low fraction unbound), enabling them to better penetrate and distribute into caseum.

This data set was then used to build a model capable of accurately predicting the caseum fraction unbound (%). mycoCSM achieved Pearson's correlation coefficient of up to 0.86 on 10-fold cross validation when predicting caseum fraction unbound, which was consistent with performance on other validation schemes (0.85 for 5-fold and 0.80 for 20-fold cross validation). The correlation increases to 0.95 when 10% of outliers are removed (Figure 3). The predictor was further evaluated on a blind test, achieving a correlation of r=0.90, consistent with cross-validation and comparable to previous efforts to predict caseum binding [26].

Building a general Mycobacteria predictor

Comparison of the molecules within each dataset revealed that there was a significant overlap of molecules with experimental MIC's in different species, with 64% of molecules tested in *M. avium*, *M. bovis*, *M. fortuitum*, *M. intracellulare*, *M. kansasii*, *M. phlei* and *M. smegmatis*, also tested in *M. tuberculosis* (Figure 4A). Interestingly, we observed a high correlation between MIC's for the same molecule between these different organisms (r=0.71, Figure 4B), which supported the feasibility of developing a general anti-Mycobacterium predictor. A genus level Mycobacterium training/test set was therefore also curated by combining all compounds with experimental MIC against any Mycobacterium, and averaging the MIC values for common molecules across species.

The M. tuberculosis model was used to predict the activities of all non-redundant compounds with experimentally measured MICs against the remaining 7 species. These predictions were correlated against the experimental measurement for that organism, revealing correlations ranging from 0.43 to 0.81. This provided further confidence in developing a general anti-Mycobacteria predictor. Building upon the 8 organism-specific predictors and data sets, we developed a general anti-Mycobacterium

predictor. mycoCSM achieved a correlation of 0.83 (RMSE of 0.52) on cross-validation (Figure 2), which was consistent with performance on an independent test set, 0.80 (RMSE of 0.55) (Figure S2). We further evaluated this final general predictor using MIC's of unique compounds against *Mycobacterium abscessus*, *Mycobacterium chelonae*, *Mycobacterium marinum*, and *Mycobacterium vaccae*, for which there was insufficient data to build species specific models. We observed correlations up to 0.89, demonstrating generalisation capabilities of our final model.

Myco-CSM Web server

Myco-CSM has been made available through an easy-to-use web interface at http://biosig.unimelb.edu.au/myco_csm, allowing users to submit molecule data sets for quick prioritization and screening (Figure 5). Users have the option to predict either organism-specific or general anti-Mycobacterial activity by submitting single molecules or batch-processing multiple molecules by providing molecules as SMILES strings. Users also have the option to calculate pharmacokinetic properties of selected molecules using pkCSM [10].

CONCLUSIONS

Here we present mycoCSM, a machine-learning based method for predicting safe, bioactive compounds for Mycobacteria. mycoCSM is capable, for the first time, of quantitatively predicting biologically active molecules for 8 Mycobacterium species as well as predicting molecules likely to be active across different species within the genus. mycoCSM also accompanies an estimator for Maximum Tolerated Dose in human, enabling the selection and enrichment of not only active but also safe compounds in screening libraries, and a model capable of predicting drug penetration in tubercular lesions. We have applied our method to the ChEMBL database to provide a rapid evaluation of commercially available compounds. Both the data sets used to train predictive models and ChEMBL screening results have been made available through a user-friendly web interface at: https://biosig.unimelb.edu.au/myco_csm. We believe mycoCSM would be an invaluable tool for

screening strategies in Mycobacteria and a platform from which similar initiatives for other relevant pathogens could be based upon.

METHODS

Data

Experimental Minimal Inhibitory Concentrations (MIC) values, given in Molar, for the Mycobacterium genus were collected from TIBLE [27] and ChEMBL [28] databases, comprising 19,684 experimental results against 8 distinct species. The penetration of antibiotics in necrotic tuberculosis lesions was also evaluated using a dataset of 279 compounds with experimentally measured avascular caseum binding and diffusion [26]. This data was used to build training and test datasets for training organism-specific predictive models as regression tasks as well as a general Mycobacterium predictor.

The logarithm of MIC100 values were averaged per molecule (based on ChEMBL identifiers) for each species, in order to generate organism-specific training/test sets containing at least 200 unique molecules. Each data set was divided into blind test (10% or at least 40 molecules) and training (the remaining 90% of the data). The resulting data sets and respective number of molecules are shown in Table 1.

Graph-based Signatures and Feature Engineering

Graphs are versatile mathematical abstractions to model entities and their relations, and have been proven intuitive and powerful tools for modelling small-molecule physicochemical properties. We have previously proposed the concept of graph-based signatures for modelling protein structures and the interactions with its partners as graphs and small-molecules [11-16, 18-22]. These have been successfully used as evidence to train and test a range of machine-learning based models, including the prediction of pharmacokinetic and toxicity profiles via the method pkCSM [10]. Here we adapted these signatures to model small-molecule activity against Mycobacteria (Supplementary Info). The main components

of the graph-based signatures are (i) distance-based patterns, represented as cumulative distribution functions of atom distances labelled based on their respective physicochemical properties (pharmacophores) and (ii) complementary physicochemical properties calculated using the RDKit cheminformatics library (Table S1) [29].

Identifying the best combination of attributes to train a predictive model is a challenging optimisation problem. To reduce noise and dimensionality, we employed feature selection via a Forward Greedy approach, by initially considering features individually and iteratively fixing the best performing ones. The main rationale behind applying this heuristic is its simplicity and relative efficiency (limited to generating a quadratic combination of features). It has also been shown that greedy feature selection improves generalisation performance, particularly for regression methods[30].

Model Selection and Validation

Several supervised machine learning methods for regression available on the scikit-learn Python library were assessed, including Random Forest, Extra Trees, Gaussian Process, Support Vector Machines, Gradient Boosting and XGBoost (Table S2). The best performing model was selected based on Pearson's correlation coefficient and Root Mean Squared Error (RMSE). Performance of predictive models was assessed under a 10-fold cross validation procedure with 10 bootstrap repetitions and using non-redundant blind tests. To validate the general predictor, organism-specific blind tests were compiled using MIC values available for other organisms (when from 50-200 unique molecules were available). Organisms with less than 50 molecules were combined within a single blind test (MIC values were averaged per molecule in both cases). Performance was also assessed on 90% of the data to investigate the effect of potential outliers. These were defined as the 10% worst predicted data points, (*i.e.*, the points further away from the regression line). For all data sets, the ensemble method Extra Trees was the best performing algorithm.

Web server

The web server front-end was developed using Bootstrap framework version 3.3.7 and the back-end was based on Python 2.7 via the Flask framework version 0.12.3 on a Linux server running Apache.

ACKNOWLEDGEMENTS

D.B.A and D.E.V.P were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundacao de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1]; the Jack Brockhoff Foundation [JBF 4186, 2016]; and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405]. Supported in part by the Victorian Government's OIS Pro-gram.

ASSOCIATED CONTENT

Supplementary Materials is available including the calculation of graph based signatures; distribution of compound properties (Figure S1); mycoCSM blind test results (Figure S2); the correlation between MIC activity and Caseum binding (Figure S3); a list of all the complementary features used in method development (Table S1); evaluation of the performance of different machine learning algorithms (Table S2).

REFERENCES

1. Hoagland, D. T.; Liu, J.; Lee, R. B.; Lee, R. E., New agents for the treatment of drug-resistant Mycobacterium tuberculosis. *Adv Drug Deliv Rev* **2016**, 102, 55-72.
2. Grzelak, E. M.; Choules, M. P.; Gao, W.; Cai, G.; Wan, B.; Wang, Y.; McAlpine, J. B.; Cheng, J.; Jin, Y.; Lee, H.; Suh, J. W.; Pauli, G. F.; Franzblau, S. G.; Jaki, B. U.; Cho, S., Strategies in anti-Mycobacterium tuberculosis drug discovery based on phenotypic screening. *J Antibiot (Tokyo)* **2019**, 72, 719-728.
3. Karmakar, M.; Rodrigues, C. H. M.; Holt, K. E.; Dunstan, S. J.; Denholm, J.; Ascher, D. B., Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One* **2019**, 14, e0217169.
4. Karmakar, M.; Trauer, J. M.; Ascher, D. B.; Denholm, J. T., Hyper transmission of Beijing lineage Mycobacterium tuberculosis: Systematic review and meta-analysis. *J Infect* **2019**, 79, 572-581.

5. Prathipati, P.; Ma, N. L.; Keller, T. H., Global Bayesian models for the prioritization of antitubercular agents. *J Chem Inf Model* **2008**, 48, 2362-70.
6. Ragno, R.; Marshall, G. R.; Di Santo, R.; Costi, R.; Massa, S.; Rompei, R.; Artico, M., Antimycobacterial pyrroles: synthesis, anti-Mycobacterium tuberculosis activity and QSAR studies. *Bioorg Med Chem* **2000**, 8, 1423-32.
7. Sivakumar, P. M.; Geetha Babu, S. K.; Mukesh, D., QSAR studies on chalcones and flavonoids as anti-tuberculosis agents using genetic function approximation (GFA) method. *Chem Pharm Bull (Tokyo)* **2007**, 55, 44-9.
8. Ekins, S.; Williams, A. J.; Krasowski, M. D.; Freundlich, J. S., In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov Today* **2011**, 16, 298-310.
9. Speck-Planche, A.; Cordeiro, M. N., Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. *Curr Top Med Chem* **2013**, 13, 1656-65.
10. Pires, D. E.; Blundell, T. L.; Ascher, D. B., pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J Med Chem* **2015**, 58, 4066-72.
11. Kaminskas, L. M.; Pires, D. E. V.; Ascher, D. B., dendPoint: a web resource for dendrimer pharmacokinetics investigation and prediction. *Sci Rep* **2019**, 9, 15465.
12. Pires, D. E.; Ascher, D. B.; Blundell, T. L., mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014**, 30, 335-42.
13. Pires, D. E.; Ascher, D. B.; Blundell, T. L., DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* **2014**, 42, W314-9.
14. Pires, D. E.; Ascher, D. B., CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res* **2016**, 44, W557-61.
15. Pires, D. E.; Ascher, D. B., mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res* **2016**, 44, W469-73.
16. Pires, D. E.; Blundell, T. L.; Ascher, D. B., mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci Rep* **2016**, 6, 29575.
17. Pires, D. E.; Chen, J.; Blundell, T. L.; Ascher, D. B., In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* **2016**, 6, 19848.
18. Pires, D. E. V.; Ascher, D. B., mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res* **2017**, 45, W241-W246.
19. Rodrigues, C. H.; Ascher, D. B.; Pires, D. E., Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res* **2018**, 46, W127-W132.
20. Rodrigues, C. H.; Pires, D. E.; Ascher, D. B., DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* **2018**, 46, W350-W355.
21. Myung, Y.; Rodrigues, C. H. M.; Ascher, D. B.; Pires, D. E. V., mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* **2020**, 36, 1453-1459.
22. Rodrigues, C. H. M.; Myung, Y.; Pires, D. E. V.; Ascher, D. B., mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res* **2019**, 47, W338-W344.
23. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1. *Advanced Drug Delivery Reviews* **2001**, 46, 3-26.
24. Yu, P.; Wild, D. J., Fast rule-based bioactivity prediction using associative classification mining. *J Cheminform* **2012**, 4, 29.
25. Hoff, D. R.; Ryan, G. J.; Driver, E. R.; Ssemakulu, C. C.; De Groote, M. A.; Basaraba, R. J.; Lenaerts, A. J., Location of intra- and extracellular M. tuberculosis populations in lungs of mice and guinea pigs during disease progression and after drug treatment. *PLoS One* **2011**, 6, e17550.

26. Sarathy, J. P.; Zuccotto, F.; Hsinpin, H.; Sandberg, L.; Via, L. E.; Marriner, G. A.; Masquelin, T.; Wyatt, P.; Ray, P.; Dartois, V., Prediction of Drug Penetration in Tuberculosis Lesions. *ACS Infect Dis* **2016**, 2, 552-63.
27. Malhotra, S.; Mugumbate, G.; Blundell, T. L.; Higueruelo, A. P., TIBLE: a web-based, freely accessible resource for small-molecule binding data for mycobacterial species. *Database (Oxford)* **2017**, 2017.
28. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, 40, D1100-7.
29. Landrum, G., RDKit: Open-source cheminformatics. **2006**.
30. Caruana, R.; Freitag, D., In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: New Brunswick, NJ, USA, 1994, pp 28–36.

TABLES

Table 1. Organism-specific and total unique compounds used to train and test myoCSM compiled based on ChEMBL and TIBLE databases.

Organism	**#Molecules (train)**	**#Molecules (blind test)**
M. avium	1,007	112
M. bovis	250	40
M. fortuitum	514	57
M. intracellulare	329	40
M. kansasii	900	100
M. phlei	190	40
M. smegmatis	1,903	212
M. tuberculosis	12,591	1400
Total unique compounds	**14,189**	**1577**

Table 2. Performance of the final mycoCSM models across training and non-redundant test sets.

	CV				Validation			
	Pearson	Kendall	Spearman	RMSE	Pearson	Kendall	Spearman	RMSE
M. avium	0.88	0.71	0.87	0.38	0.88	0.75	0.88	0.40
M. bovis	0.80	0.60	0.79	0.62	0.81	0.54	0.72	0.61
M. fortuitum	0.89	0.61	0.78	0.44	0.80	0.54	0.72	0.55
M. intracellulare	0.85	0.69	0.86	0.41	0.88	0.64	0.80	0.39
M. kansasii	0.87	0.70	0.87	0.42	0.83	0.66	0.84	0.45
M. phlei	0.85	0.68	0.86	0.44	0.79	0.64	0.81	0.60
M. smegmatis	0.84	0.64	0.81	0.52	0.76	0.53	0.70	0.56
M. tuberculosis	0.83	0.63	0.82	0.52	0.82	0.63	0.81	0.53
Mycobacterium	0.83	0.64	0.81	0.52	0.80	0.61	0.79	0.55

FIGURES

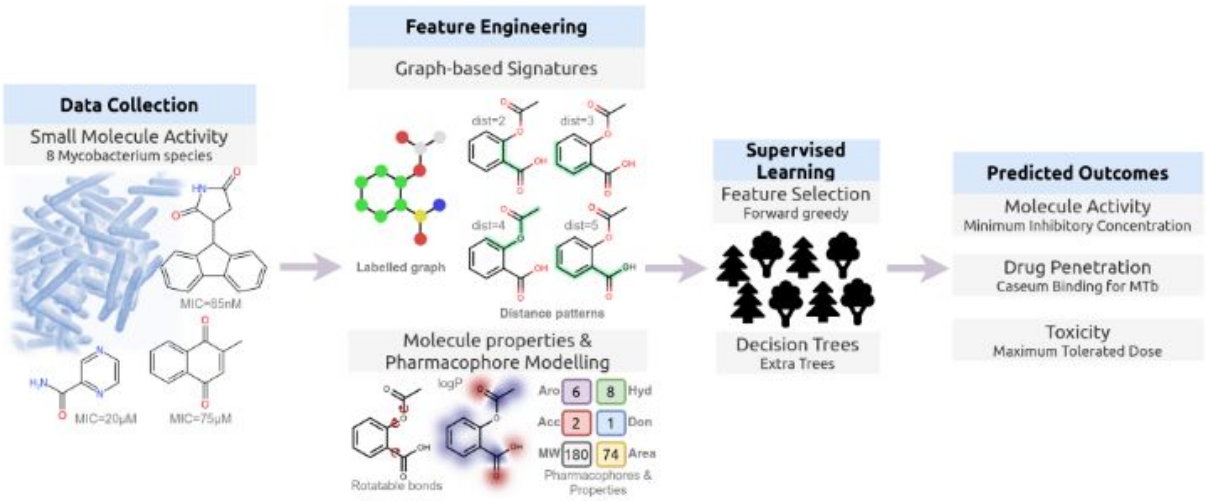


Figure 1. **mycoCSM workflow.** The developed method is composed of four main stages. During Data Collection, small molecule activity (in terms of Minimum Inhibitory Concentration) data was collected from the literature for eight different Mycobacteria species, in addition to drug penetration for M. tuberculosis. During Feature Engineering, two classes of features were derived: (i) graph-based signatures that aim to describe both small molecule geometry and physicochemical properties and (ii) general molecules properties and pharmacophores. These were then used as evidence to train and test predictive models via supervised learning. Models' performance was optimized using greedy feature selection. Finally, the best performing models have been made available through an easy-to-use web interface, also incorporating a toxicity filter for Maximum Tolerated Dose in Humans, allowing users to filter safer compounds.

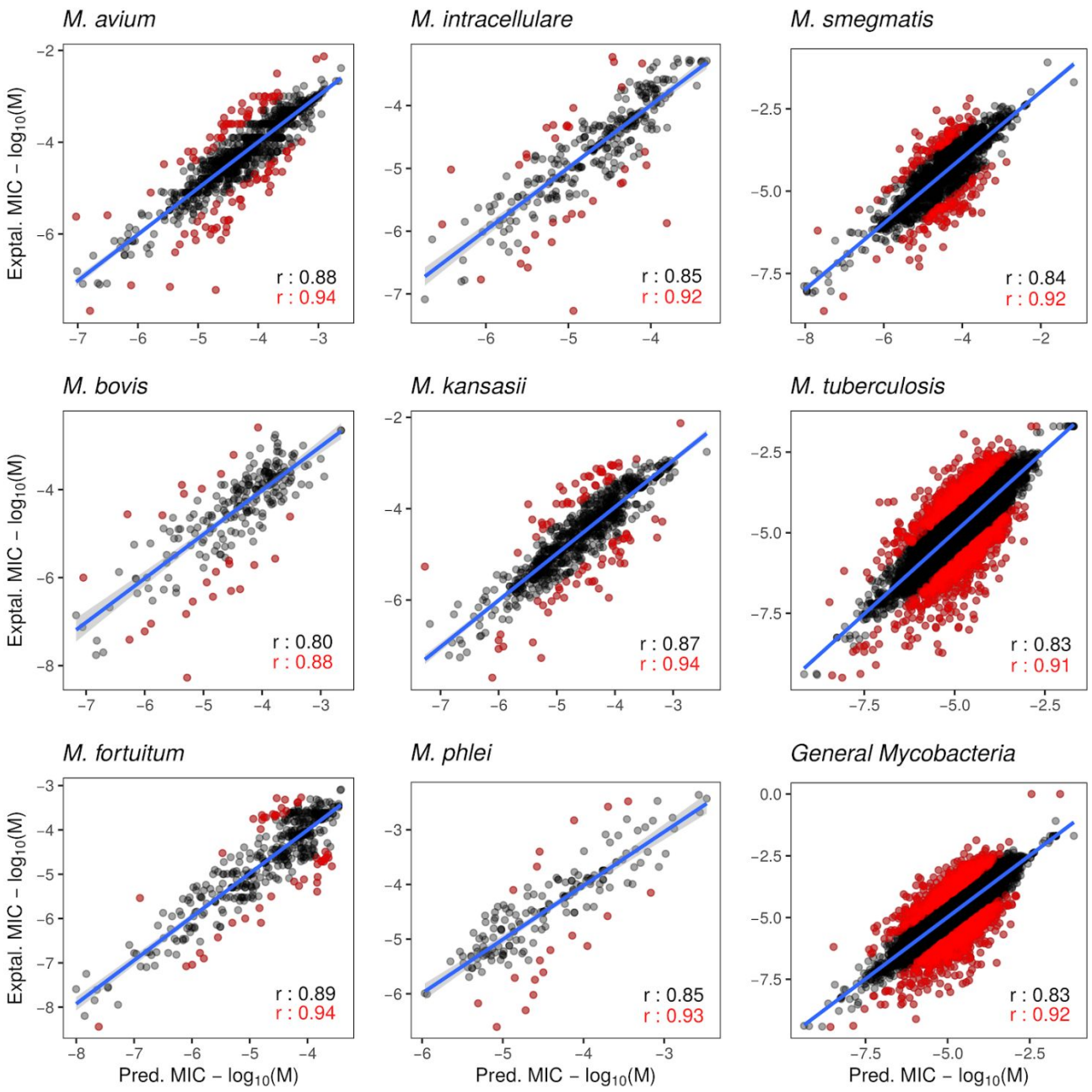


Figure 2. Performance of mycoCSM on cross validation. Scatter plots between experimental and predicted MIC values given in log10(Molar) for each of the eight organism-specific models as well as the general Mycobacteria model are shown. Pearson's correlation coefficient (r) are shown for each plot (in black for 100% of the data and in red for 90% of the data, after 10% outlier removal).

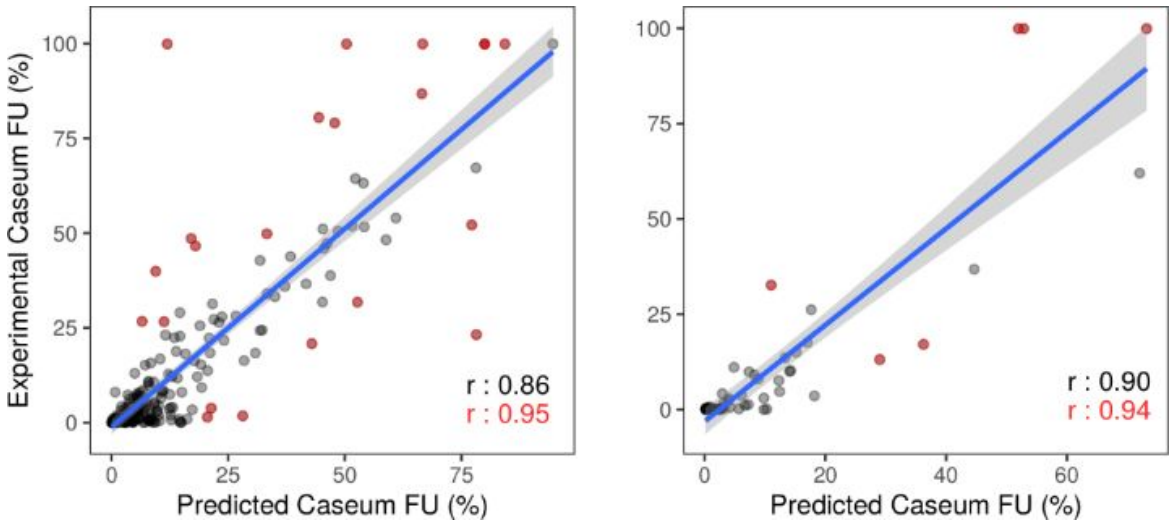


Figure. 3. Performance of mycoCSM on predicting compound penetration in tubercular lesions. The graphs present scatter plots of experimental and predicted caseum fraction unbound (as a percentage %) assessed under 10-fold cross-validation (left-hand side) and blind test (right-hand side). mycoCSM presented consistent performance on all experiments. Pearson's correlation coefficient (r) are shown for each plot (in black for 100% of the data and in red for 90% of the data, after 10% outlier removal).

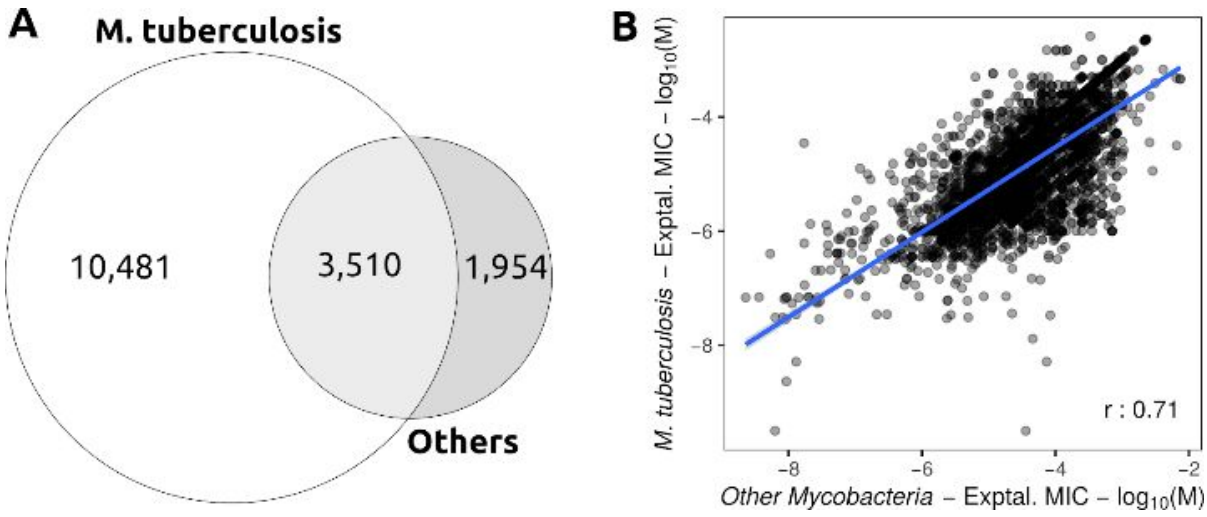


Figure 4. Comparison of compounds tested against different Mycobacteria species. (A) Venn diagram showing the overlap of molecules with experimental MICs in M. tuberculosis and the seven other species. (B) Correlation of experimental MICs between M. tuberculosis and the seven other species.

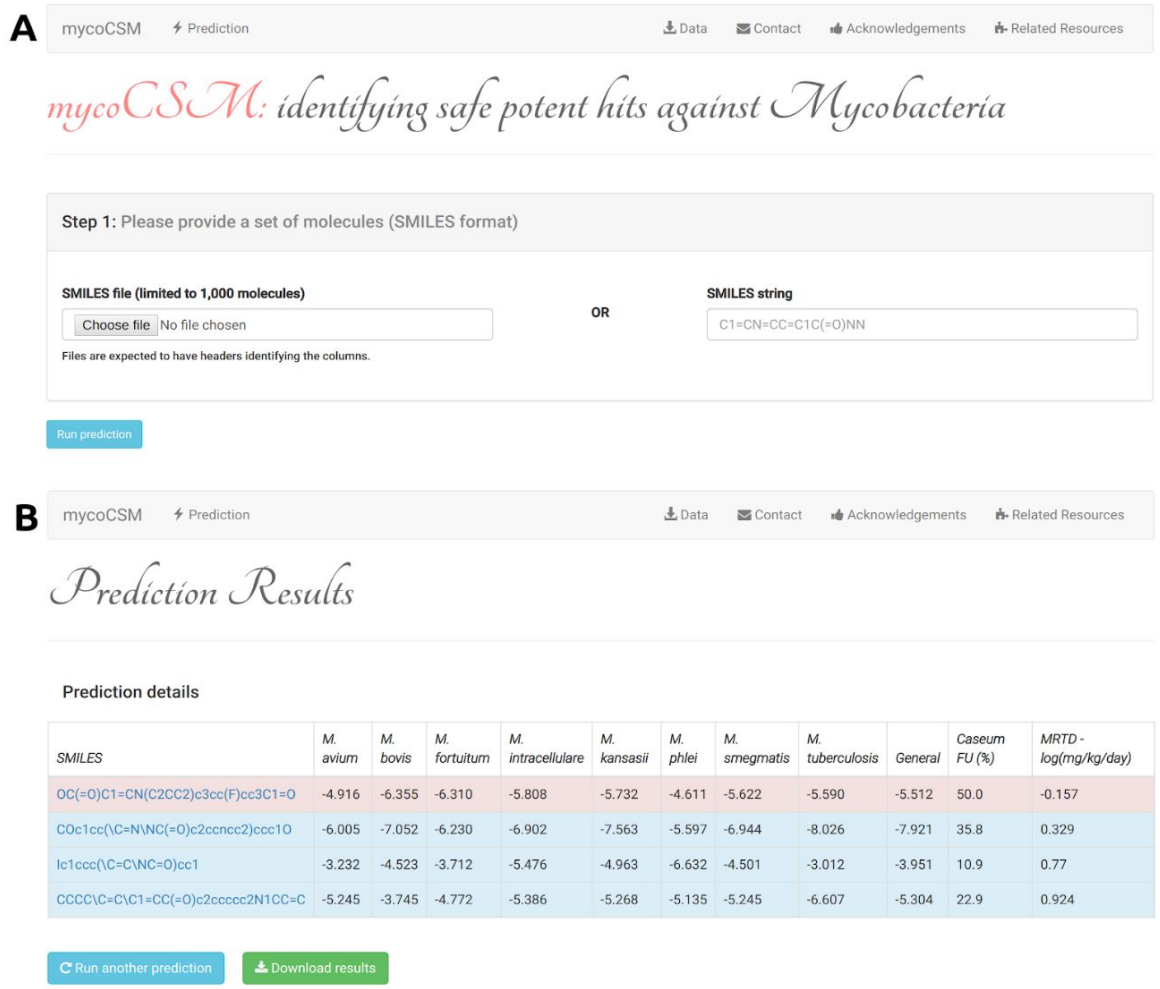


Figure 5. mycoCSM webserver interface. (A) shows the submission page for mycoCSM. Users have the option to either provide a compound represented as a SMILES string or a set of compounds as a SMILES file, for assessing multiple molecules. (B) shows the results page for multiple molecule submission. Results are presented in tabular format, including predic-tions for all 8 organism-specific models, the general Mycobacteria model and drug penetration. Maximum Recommended Tolerated Doses (MRTD) in human are also calculated using pkCSM and presented. Users have also the option to calculate other pharmacokinetic and toxicity properties of compounds of interest using the pkCSM platform.

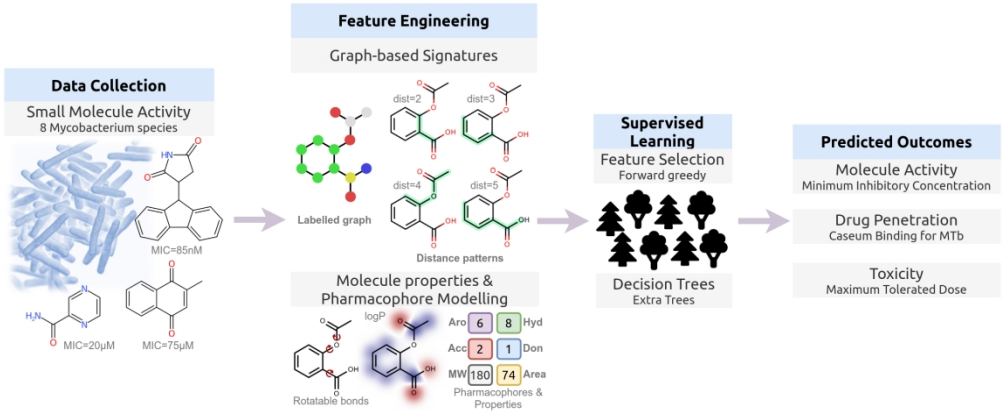


Fig. 1. mycoCSM workflow. The developed method is composed of four main stages. During Data Collection, small molecule activity (in terms of Minimum Inhibitory Concentration) data was collected from the literature for eight different Mycobacteria species, in addition to drug penetration for M. tuberculosis. During Feature Engineering, two classes of features were derived: (i) graph-based signatures that aim to describe both small molecule geometry and physicochemical properties and (ii) general molecules properties and pharmacophores. These were then used as evidence to train and test predictive models via supervised learning. Models' performance was optimized using greedy feature selection. Finally, the best performing models have been made available through an easy-to-use web interface, also incorporating a toxicity filter for Maximum Tolerated Dose in Humans, allowing users to filter safer compounds.

814x341mm (72 x 72 DPI)

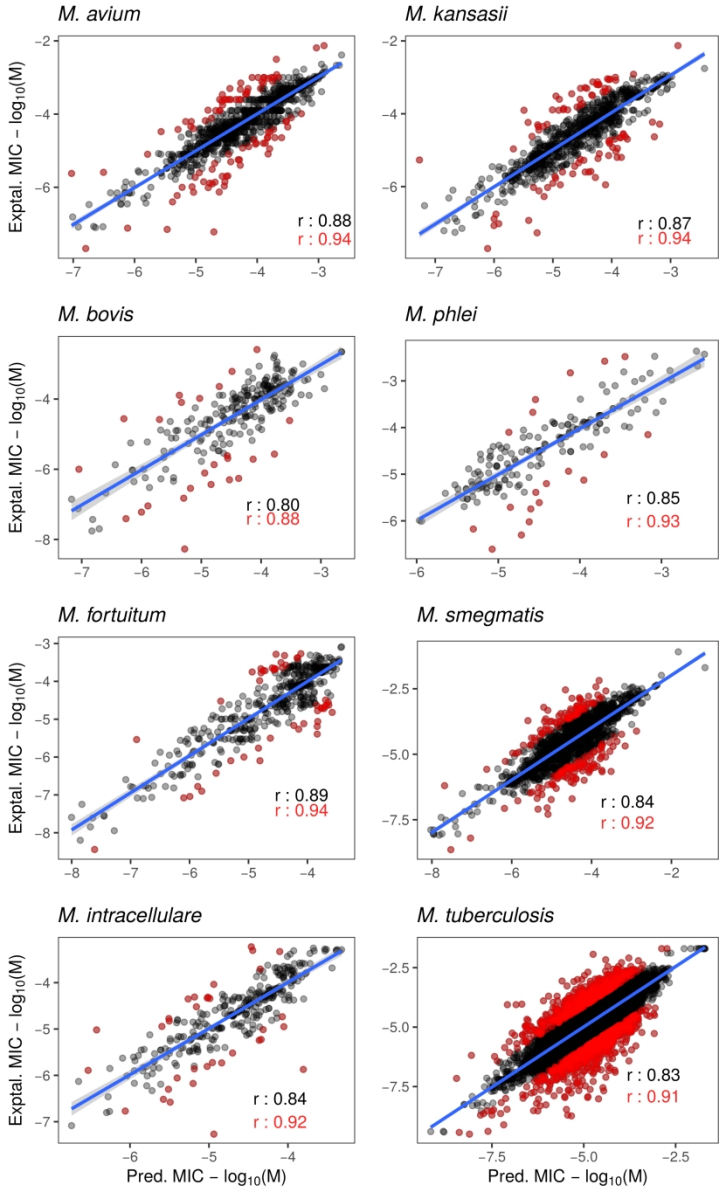


Fig. 2. Performance of mycoCSM on cross validation. Scatter plots between experimental and predicted MIC values given in log10(Molar) for each of the eight organism-specific models as well as the general Mycobacteria model are shown. Pearson's correlation coefficient (r) are shown for each plot (in black for 100% of the data and in red for 90% of the data, after 10% outlier removal).

162x269mm (400 x 400 DPI)

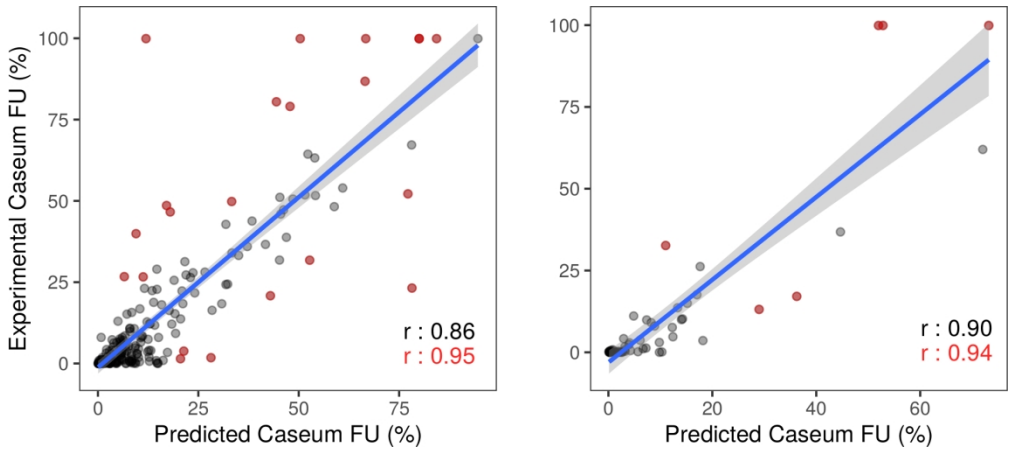


Fig. 3. Performance of mycoCSM on predicting compound penetration in tubercular lesions. The graphs present scatter plots of experimental and predicted caseum fraction unbound (as a percentage %) assessed under 10-fold cross-validation (left-hand side) and blind test (right-hand side). mycoCSM presented consistent performance on all experiments. Pearson's correlation coefficient (r) are shown for each plot (in black for 100% of the data and in red for 90% of the data, after 10% outlier removal).

168x74mm (300 x 300 DPI)

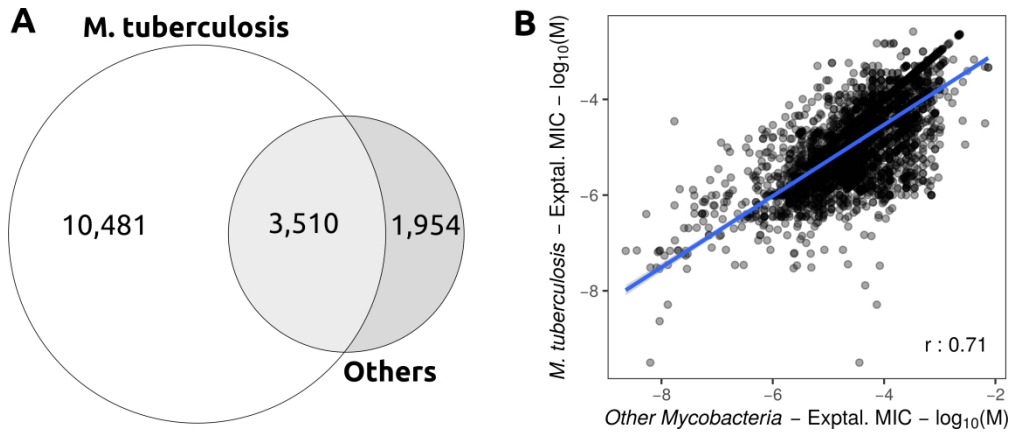


Fig. 4. Comparison of compounds tested against different Mycobacteria species. (A) Venn diagram showing the overlap of molecules with experimental MICs in M. tuberculosis and the seven other species. (B) Correlation of experimental MICs between M. tuberculosis and the seven other species.

381x160mm (300 x 300 DPI)

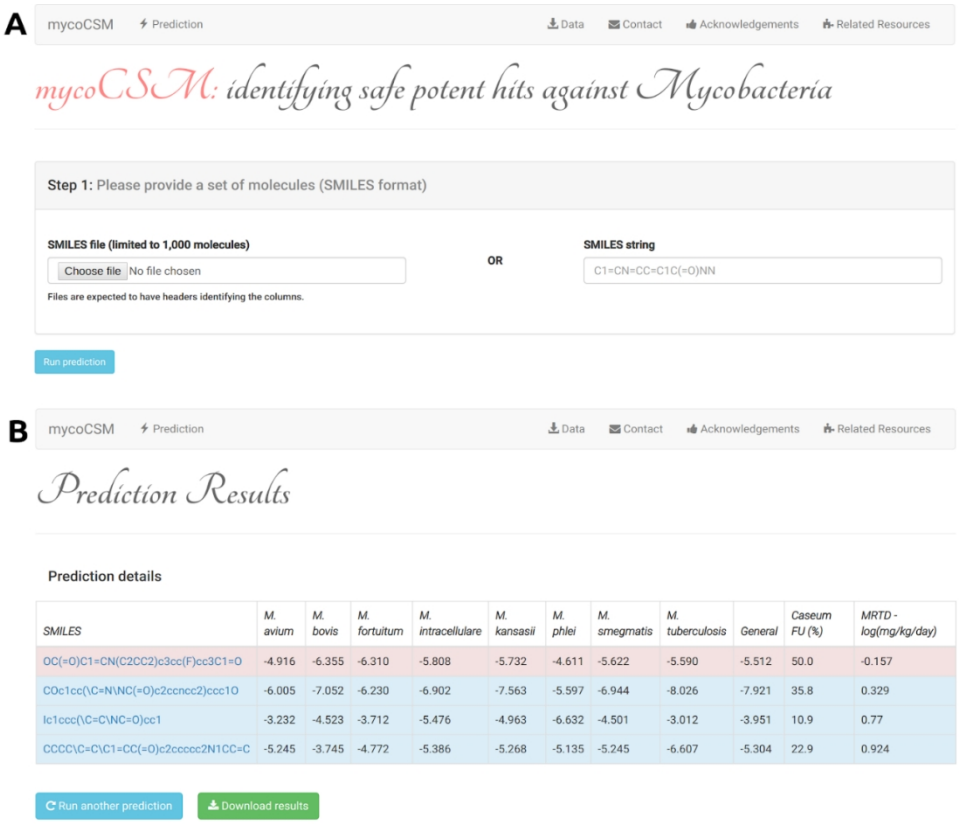


Fig. 5. mycoCSM webserver interface. (A) shows the submission page for mycoCSM. Users have the option to either provide a compound represented as a SMILES string or a set of compounds as a SMILES file, for assessing multiple molecules. (B) shows the results page for multiple molecule submission. Results are presented in tabular format, including predic-tions for all 8 organism-specific models, the general Mycobacteria model and drug penetration. Maximum Recommended Tolerated Doses (MRTD) in human are also calculated using pkCSM and presented. Users have also the option to calculate other pharmacokinetic and toxicity properties of compounds of interest using the pkCSM platform.

355x291mm (96 x 96 DPI)