JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Looking Back to the Future: Predicting *in Vivo* Efficacy of Small Molecules versus *Mycobacterium tuberculosis*

Sean Ekins,*[,†,‡] Richard Pottorf,[§] Robert C. Reynolds,[∥] Antony J. Williams,[⊥] Alex M. Clark,[#] and Joel S. Freundlich*[,§,▽]

[†]Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, California 94010, United States

[‡]Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, North Carolina 27526, United States

[§]Department of Pharmacology & Physiology, Rutgers University − New Jersey Medical School, 185 South Orange Avenue, Newark, New Jersey 07103, United States

[∥]Department of Chemistry, University of Alabama at Birmingham, 1530 Third Avenue South, Birmingham, Alabama 35294-1240, United States

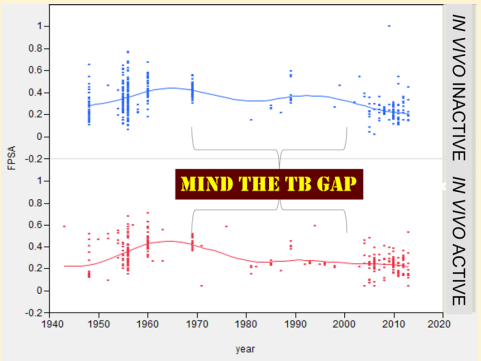
[⊥]Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, North Carolina 27587, United States

[#]Molecular Materials Informatics, 1900 St. Jacques #302, Montreal, Quebec, Canada H3J 2S1

[▽]Department of Medicine, Center for Emerging and Reemerging Pathogens, Rutgers University − New Jersey Medical School, 185 South Orange Avenue, Newark, New Jersey 07103, United States

Supporting Information

ABSTRACT: Selecting and translating *in vitro* leads for a disease into molecules with *in vivo* activity in an animal model of the disease is a challenge that takes considerable time and money. As an example, recent years have seen whole-cell phenotypic screens of millions of compounds yielding over 1500 inhibitors of *Mycobacterium tuberculosis* (*Mtb*). These must be prioritized for testing in the mouse *in vivo* assay for *Mtb* infection, a validated model utilized to select compounds for further testing. We demonstrate learning from *in vivo* active and inactive compounds using machine learning classification models (Bayesian, support vector machines, and recursive partitioning) consisting of 773 compounds. The Bayesian model predicted 8 out of 11 additional *in vivo* actives not included in the model as an external test set. Curation of 70 years of *Mtb* data can therefore provide statistically robust computational models to focus resources on *in vivo* active small molecule antituberculars. This highlights a cost-effective predictor for *in vivo* testing elsewhere in other diseases.



INTRODUCTION

Drug discovery involves a considerable effort in the selection and translation of *in vitro* leads into molecules with *in vivo* efficacy in an animal model of the disease. Our collective memories are often short when it comes to decades of research in a single therapeutic area, let alone research of over more than a half-century, and yet the past may hold many insights to aid us in drug discovery efforts in both the present and future. As an example, from the 1940s to the 1960s,[1] significant efforts led to first- and second-line drugs for *Mycobacterium tuberculosis* (*Mtb*), which is the causative agent of tuberculosis (TB). This disease has infected approximately 2 billion people and kills 1.3 million people annually.[2] We critically need next-generation active small molecules as tools to query essential infection biology to drive novel therapies. Chemical probes can enable interrogation of *Mtb* pathways essential to *in vivo* infection. Next generation drugs must lack cross resistance to current therapeutics, shorten treatment, and address drug−drug interactions with co-administered treatments.[3−6] Many molecules have been assessed as to their ability to modulate *Mtb* infection in mice. These data reside in numerous journals and reports that are not readily accessible despite today's electronic media and databases. Unfortunately, much of the early (pre-1970s) structure−activity relationship (SAR) data from both *in vitro* and *in vivo* models appears to have been neglected. Typically, these data are only unearthed on a compound-by-compound basis when we rediscover[7,8] an agent that was already known from decades ago.[9,10]

The field's interest in this historical data may have been unfortunately diminished due to the recent upsurge in whole-cell phenotypic high-throughput screens (HTS) for novel antituberculars that have seen several million compounds tested.[11−17] Commercial vendor and in-house libraries[4,13−15] have been assayed, leading to the clinical candidate SQ-109[18] and a diarylquinoline hit that was optimized to the drug bedaqui-

Received: February 8, 2014
Published: March 25, 2014

line.[19] However, these successful outcomes from screening represent anomalies as the road from hit to drug invokes words such as "valley of death".[20] *Mtb* HTS hit rates are usually below 1%,[11,21] and most hits require a significant amount of chemical evolution in an effort to identify a probe let alone a drug discovery lead. To date we estimate about 1500 *in vitro Mtb* hits of interest have been derived from one laboratory alone,[13−15,22] while GSK has recently published another 177 promising *in vitro* actives.[12] Many laboratories have also described *Mtb* hits; so in total, there are likely close to 2000 compounds from recent efforts that may require triage before advancement of the most promising through the discovery pipeline.

To significantly impact the TB field, these hits, or their evolved analogs, ideally must demonstrate *in vivo* activity in an animal model. Rarely is the specific *Mtb* target for these compounds known before *in vivo* testing. The mouse model of TB infection is considered important for comparative assessment of different treatments and optimization of TB drug dosing schedules.[23] There has also been considerable development of the acute[24] and chronic mouse models[25] of *Mtb* infection. Given current resource limitations that are magnified for a neglected disease such as TB, we are faced with a dilemma: How do we efficiently select among the thousands of hits to decide which to carry forward for *in vivo* efficacy assessment?

This report details a novel approach to address this critical issue through the curation of 773 molecules that have been tested in the last 70 years in the mouse TB infection model. This *in vivo* data has never before been curated. We present detailed analyses of the physiochemical and structural properties of both active and inactive molecules as well their chemical property space coverage in an effort to guide the future design of novel antitubercular chemical probes and drugs. Furthermore, we leverage our experience with machine learning models for *in vitro* activity[8,22,26−29] to construct and statistically validate these computational models to predict *in vivo* efficacy in the *Mtb*-infected mouse model. The computational models are further validated through the correct prediction of 8 of 11 known *in vivo* actives absent from the training set. The models are also applied to score a set of 177 *in vitro* drug leads recently reported by GlaxoSmithKline (GSK)[12] to aid in their prioritization for *in vivo* assessment.

EXPERIMENTAL SECTION

Molecule: Curation, Drawing, Quality Assessment, and Storage. Various search terms were used in PubMed to retrieve papers with compounds tested in murine acute and chronic *Mtb* infection models. For example, "tuberculosis and in vivo and mouse", "Tuberculosis and efficacy and mouse", and "comparison and antituberculosis and mouse". The same search terms were also used in SciFinder (CAS, Columbus, OH) and Web of Knowledge (Thomson Reuters). Individual journals were also searched online (e.g., Tuberculosis, Journal of Medicinal Chemistry, and PLOS journals). The *In vivo* data was manually curated, and structures were sketched using the Mobile Molecular DataSheet (MMDS) iOS app,[30] ChemDraw (Perkin-Elmer, Waltham MA), or downloaded from ChemSpider (www.chemspider.com) and combined with pertinent data fields. The data has been made publically available in the CDD TB database (Collaborative Drug Discovery, Inc., Burlingame, CA).[31] Molecules were classed as active/inactive, and this was generally based on the data in the publications. For example, a reduction of log CFU in lung greater or equal to 1 log was considered active. The initial assembled data set was shown to contain duplicates by utilizing the ACD/ChemFolder version 12 software program (Advanced Chemistry Development, http://www.acdlabs.com/products/km/ackm/chemfolder/). Utilizing the ability to check for duplicates and incorrect structures (valence errors, pentavalent carbons, missing stereochemistry), identified structure issues were manually curated. A total of 18 compounds were either removed from the originally assembled data set or edited to deal with the identified errors.

Molecular Property Distribution. AlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular fractional polar surface area were calculated from input SD files using Discovery Studio 3.5 (San Diego, CA).

Principal Components Analysis with *in Vitro* Hits and TB Mobile Data. We compared the 773 compounds with the previously described 745 compounds with known *Mtb* targets collated from the literature[27] and available in TB Mobile (version 1)[32] that were utilized to generate a principal components analysis (PCA) plot with the interpretable descriptors selected previously (AlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular fractional polar surface area) for machine learning. This PCA model represents essentially the published target-chemistry property space for *Mtb*. We also compared 1429 *Mtb* hits (active and nontoxic only, from the NIH screens where IC_{90} < 10 μg/mL or 10 μM and a selectivity index (SI) greater than 10 where the SI is calculated from SI = CC_{50}/IC_{90}) to show how they covered the target-chemistry property space. These analyses can be compared with those previously published which focused on *in vitro Mtb* data.[29]

Building and Validating Machine Learning Models with Mouse *Mtb in Vivo* Data. We have previously described the generation and validation of the Laplacian-corrected Bayesian classifier models developed from *Mtb* growth inhibition screens of small molecule libraries[8,28] using Discovery Studio 3.5.[33−37] This approach was utilized with the literature data curated in the course of this study. The following molecular descriptors were used and were calculated from input SD files: molecular function class fingerprints of maximum diameter 6 (FCFP_6),[38] AlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular fractional polar surface area. Models were validated using leave-one-out cross validation in which each sample was left out one at a time, a model was built using the remaining samples, and that model was utilized to predict the left-out sample. Each model was internally validated, ROC plots were generated, and the XV ROC AUC calculated. The Bayesian model was additionally evaluated by leaving out 50% of the data and rebuilding the model 100 times using a custom protocol for validation to generate the ROC AUC, concordance, specificity, and selectivity as described previously.[8,28] The internal ROC value represents the training set value while the external ROC represents the test set molecules left out. We also compared the resulting Bayesian model with SVM and RP Forest and single tree models built with the same molecular descriptors in Discovery Studio. For SVM models, we calculated interpretable descriptors in Discovery Studio and then used Pipeline Pilot to generate the FCFP_6 descriptors followed by integration with R.[39] RP Forest and RP Single Tree models used the standard protocol in Discovery Studio. In the case of RP Forest models, 10

Table 1. Mean (Standard Deviation) of Molecular Descriptors for *in Vivo N* = 773 *in Vivo Mtb* Data set, Comparing Actives and Inactives[a]

	MW	AlogP	HBD	HBA	Num Rings	Num Arom Rings	FPSA	RBN
active (N = 362)	417.25 ± 454.39	3.11 ± 2.71[b]	1.49 ± 2.17	6.68 ± 8.33	2.96 ± 2.09[b]	1.72 ± 1.46	0.29 ± 0.13[b]	7.84 ± 22.48
inactive (N = 411)	386.95 ± 440.40	3.89 ± 4.88	1.39 ± 1.86	5.75 ± 6.20	2.51 ± 2.70	1.90 ± 2.37	0.31 ± 0.14	8.09 ± 16.05

[a]MWT = molecular weight, HBD = hydrogen bond donor, HBA = hydrogen bond acceptor, Num Rings = number of Rings, Num Arom Rings = number of aromatic rings, FPSA = fractional polar surface area, and RBN = rotatable bond number. Fractional polar surface area (FPSA) = total partially positively charged molecular surface area divided by the total molecular surface area. [b]$p < 0.05$.

trees were created with bagging. Bagging is short for "Bootstrap AGgregation". For each tree, a bootstrap sample of the original data is taken, and this sample is used to grow the tree. A bootstrap sample is a data set of the same size as the original one but in which the same data record can be included multiple times. RP Single Trees had a minimum of 10 samples per node and a maximum tree depth of 20. In all cases, five-fold cross validation (leave out 20% of the database five times) was used to calculate the ROC for the models generated.

Model Predictions for Additional Compounds Identified after Model Building. Eleven compounds active in the mouse *in vivo* model were identified from a 1950s compilation.[40] Of these, only seven were not in the 773 compounds training set. A further four active compounds[41,42] resulted in 11 compounds that were predicted with the computational models developed (Table 4). For each molecule, the closest distance to the training set was also calculated using the Bayesian model in the calculated properties protocol method in Discovery Studio (a value of zero represents a molecule in the training set, while larger values are further from the training set).

Predictions for GSK Compounds. 177 *Mtb* leads were recently disclosed by GSK[12] and represent a promising set of small molecules for further exploration as potential antitubercular drug candidates. The GSK set was scored with all of the *in vivo* models generated in this study. The mean closest distance to the training set was also calculated for the 177 compounds to provide an idea of similarity to the training set. These data were calculated from the outputs of each of the Bayesian models. For each test set molecule a score for closest distance to training set was calculated using Discovery Studio (described earlier). We averaged this number across the 177 molecules, where the "closeness" of a compound to the training set scales inversely with the value. The maximum Tanimoto similarity for each molecule versus the training set was also calculated using MDL fingerprints. Consensus predicted active compounds were identified across all four machine learning models. These compounds were then evaluated using TB Mobile and clustering in Discovery Studio with the 745 compounds in this data set to infer potential targets.[32] ADME/Tox properties for these compounds were generated using Discovery Studio ADMET predictors and custom Bayesian models for PXR, hERG,[43] etc.

Scaffold Analysis Using SAR Table. Scaffold analysis was performed using the SAR Table app[44] (for iOS-based devices such as iPhones and iPads), which provides a user interface for drawing scaffolds and substituents and specifying activity data. It also provides access to analysis functionality such as scaffold-substructure matching, structure–activity model generation, data visualization, and manuscript figure creation.

Statistical Analysis. Means for descriptor values for active and inactive compounds were compared by two tailed *t*-tests with JMP v. 8.0.1 (SAS Institute, Cary, NC).

RESULTS

Data Set Curation, Molecular Property Analysis, and Data Visualization. A total of 773 molecules were collated from the literature for the first time to our knowledge using various search terms in PubMed, SciFinder, and Web of Knowledge (see Experimental Section), for which there was data in acute or chronic mouse models of *Mtb* infection (Table S1, Supporting Information). Following convention, an "active" compound minimally exhibited a 1 $\log_{10}$ reduction in *Mtb* colony-forming units (CFUs) in the lungs as compared to no-drug control. Occasionally, other types of analyses required our best scientific judgment on the active/inactive boundary; for example, the work of Denny and co-workers[45−50] used an activity ratio versus PA-824. We considered a value >0.4 as active given the demonstrated 2.5−3 $\log_{10}$ reduction in CFUs in the lungs by PA-824 (27−32). Older publications relied on extension of survival compared to negative and positive controls.[9] Fortuitously for machine learning model construction, the data set was divided almost equally between actives (N = 362) and inactives (N = 411). Through analysis of simple molecular descriptors (see Experimental Section) we sought to gain insight as to why active and inactive compounds behaved differently (Table S2 and Figure S1, Supporting Information). Among the 773 compounds, statistically significant differences were found between the number of rings, their AlogP, and fractional polar surface area (Table 1). It should be noted that the standard deviations are large for the *in vivo* data and may stem from the heterogeneity of structures (e.g., between very small molecules, large macrolide antibiotics, and calixarenes) and likely published experimental methods (e.g., different mice strains, dosing concentrations, dosing period, etc.). We would not advise using individual properties alone (like calculated logP) to differentiate *in vivo* active compounds as there were temporospatial effects for the descriptors, suggesting the addition of further compounds over time increases or decreases differences observed (Figure S1, Supporting Information).

PCA can be used to understand multi-dimensional data represented by the multiple molecular descriptors representing the molecular properties of the training set and shows overlap in this property space between active and inactive compounds (Figure 1A). Our analyses can also be compared to those previously performed with *Mtb in vitro* data.[29] This overlap would also suggest some complexity in using these individual molecular descriptors alone to distinguish *in vivo* active compounds. Approved drugs for TB are distributed in this same chemical property space with much larger intravenous TB drugs (representing generally larger molecules with different molecular properties (less hydrophobic)[31]) separated out of the main cluster (Figure 1B).

PCA using molecules known to inhibit specific *Mtb* targets[32] suggests that the *in vivo* compounds possess good coverage of known target chemistry property space and extend well outside of it (Figure 2A). However, it should be noted these targets do

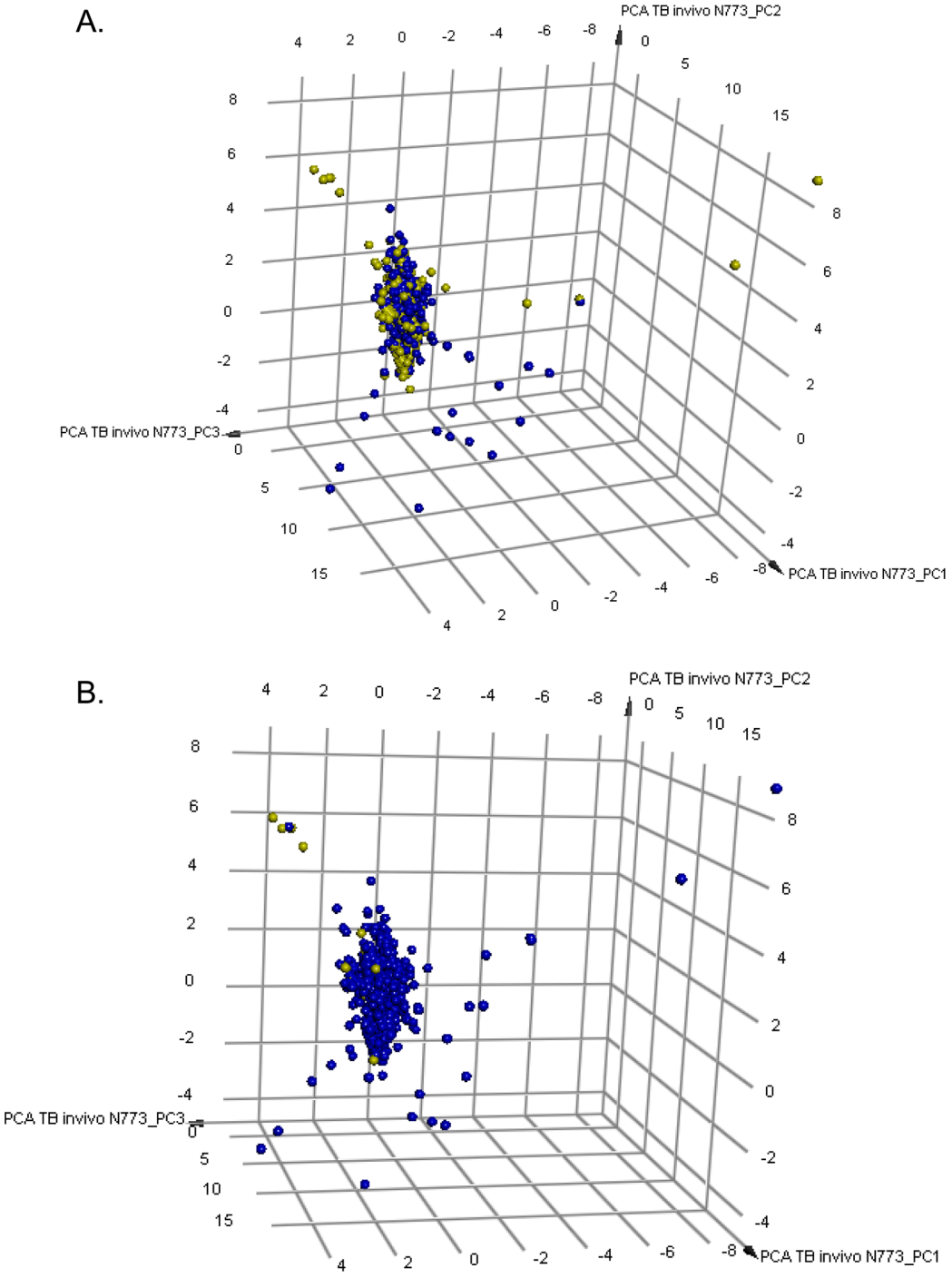


Figure 1. Coverage of *Mtb in vivo* molecule property space: (A) $N = 773$ compounds showing how some actives (yellow) are outside the major cluster and represent more diverse molecules. 3PCs describe 87% of variance. (B) Highlighting known first and second line TB drugs and others used against the disease (bedaquiline, moxifloxacin, ofloxacin, sparfloxacin, imipenem, gatifloxacin, rifampin, pyrazinamide, rifalazil, rifapentine, rifabutin, levofloxacin, clarithromycin, amikacin, kanamycin, streptomycin, capreomycin IA, ethambutol, ethionamide, isoniazid, and meropenem). Most *Mtb* drugs (yellow) are hidden in the large blue cluster; top left-hand cluster is amikacin, capreomycin IA, kanamycin, and streptomycin.

not broadly cover the chemistry property space of the *Mtb* metabolome (using molecules extracted from BioCyc[51,52]) (Figure 2B), and the active *in vivo* compounds also only cover a fraction of the *Mtb* metabolome PCA plot (Figure 2C).

When the same *in vivo* compounds are compared to known *Mtb in vitro* actives,[13−15,22] they are also well distributed over the PCA plot (Figure 3A). The majority of the *in vivo* actives overlap with the *in vitro* actives[13−15,22] suggesting a good coverage of the chemistry property space (and likely requirement for similar molecular features or descriptors), with a small number of the *in vivo* actives exploring property space distinct from that of the *Mtb in vitro* actives from recently described screening efforts. This graphical tool can demonstrate differential coverage of chemical property space and may be utilized to explore how candidate molecules compare to molecules previously assayed *in vivo*.

A significant component of the data set (51 of 773 molecules; Table S1, Supporting Information) contains the triazine central scaffold. These compounds were initially brought to our

dx.doi.org/10.1021/ci500077v | *J. Chem. Inf. Model.* 2014, 54, 1070−1082

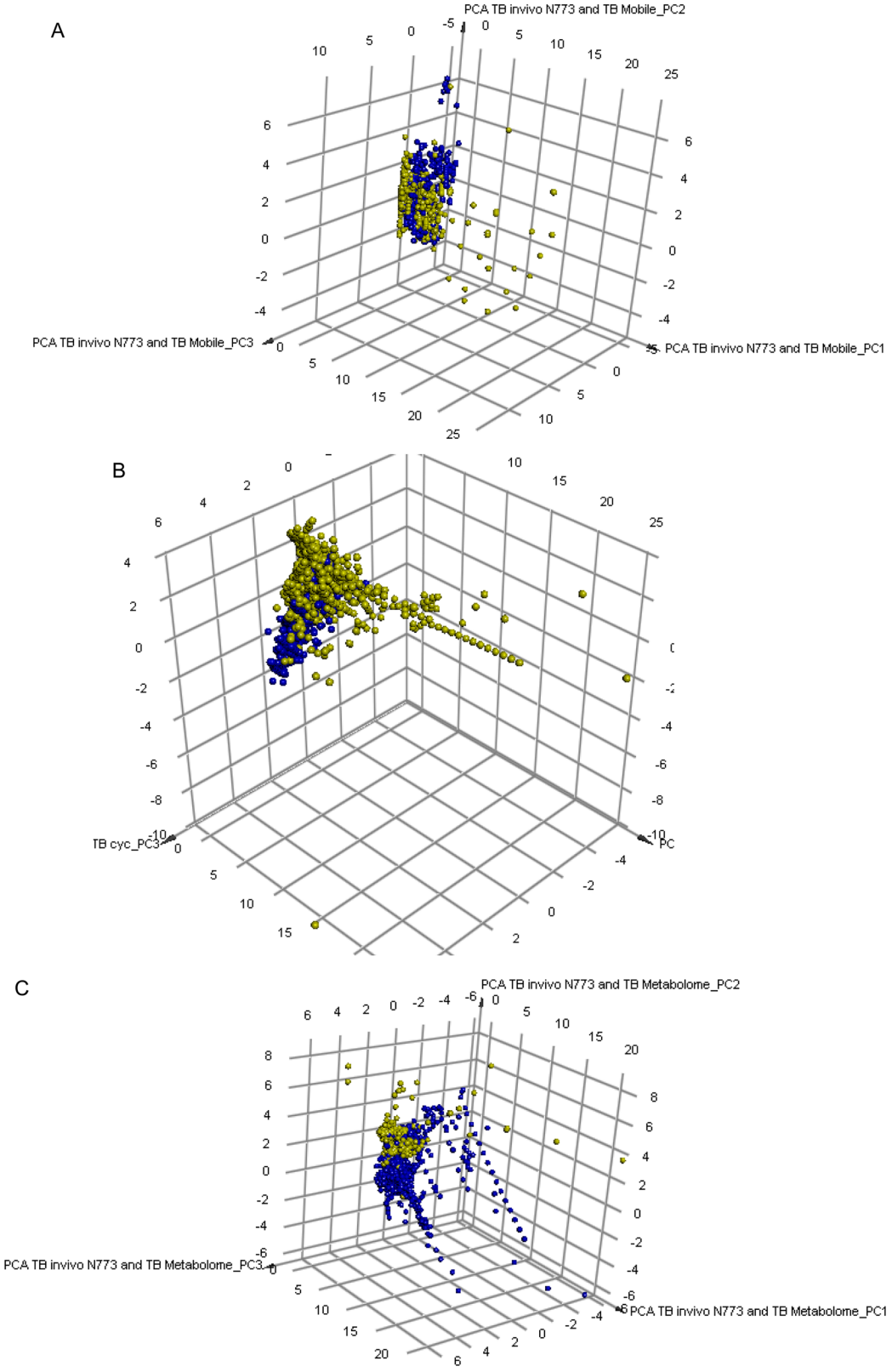


Figure 2. Coverage of *Mtb* target molecule property space: (A) 745 TB Mobile molecules (blue) with annotated targets and 773-member TB *in vivo* training set (yellow) PCA; 3PCs explain 88% of variance. (B) Comparison of TB target molecule property space using data from TB Mobile (blue) and 1770 *Mtb* metabolites (yellow) using data from BioCyc.[51] 3PCs explain 89% of the variance. (C) Comparison of 1770 *Mtb* metabolites (blue) and 773-member TB *in vivo* data set (yellow); 3PCs explain 87% of the variance.

attention when we identified TCMDC-125802 ((*E*)-6-(2-((5-nitrofuran-2-yl)methylene)hydrazinyl)-N^2,N^4-diphenyl-1,3,5-triazine-2,4-diamine, Figure 4A; MIC of 62.5 ng/mL against *Mtb*)[8] through our machine learning models for *in vitro* antitubercular activity. A literature search highlighted *one report* of the antitubercular activity of TCMDC-125802 and related

dx.doi.org/10.1021/ci500077v | *J. Chem. Inf. Model.* 2014, 54, 1070−1082

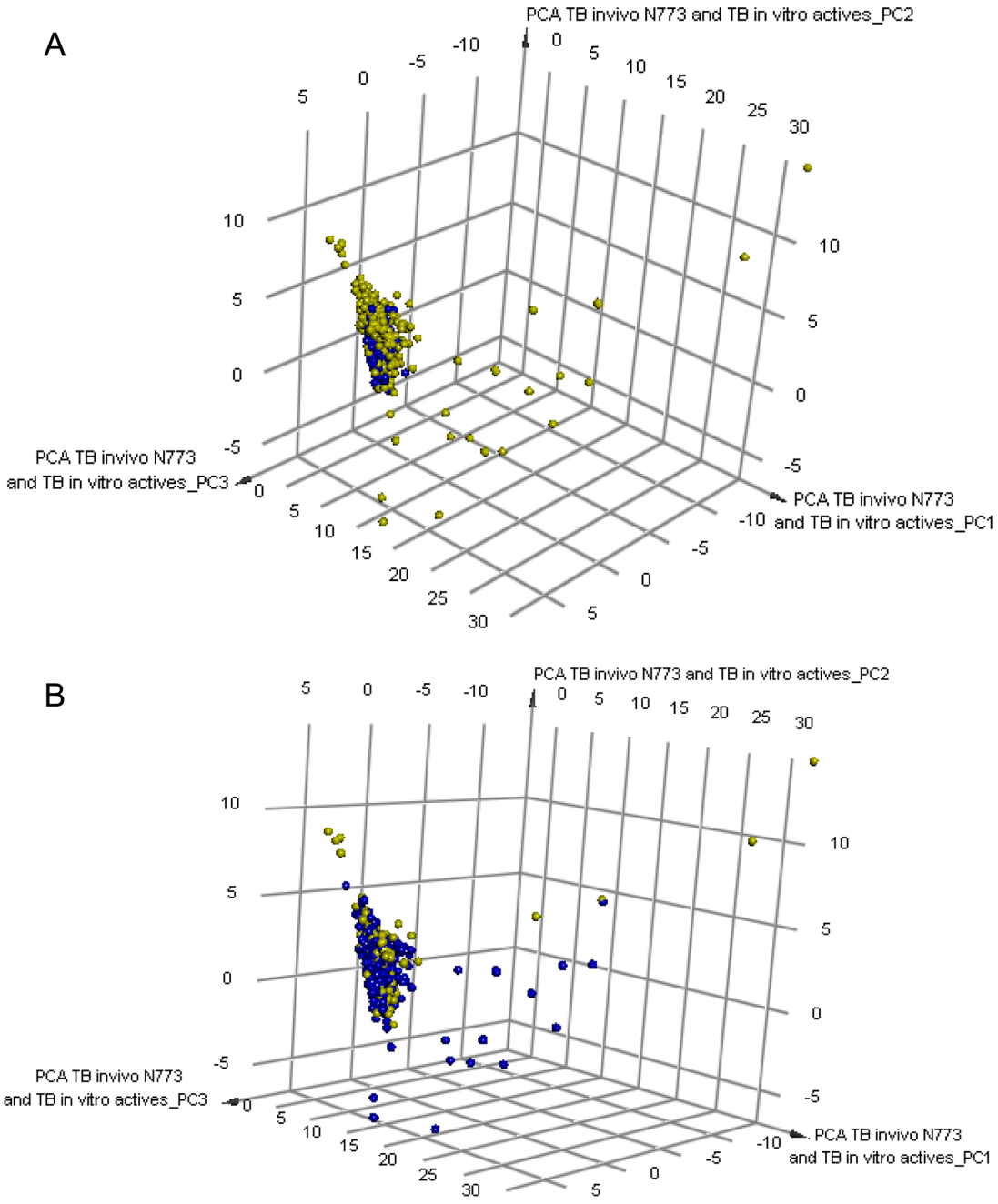


Figure 3. Coverage of *Mtb in vitro* growth inhibitor chemistry property space. (A) 1429 TB *in vitro* actives (blue) and 773 molecule TB *in vivo* data set (yellow) PCA; 3PCs explain 83.7% of variance. Aminoglycosides are shown toward the top of the plot. (B) Highlighting the TB *in vivo* active compounds only (yellow).

triazines from 1969.[53] Figure 4 shows a matrix correlation plot of triazine substituents R^1 versus R^2 for this published data set based on *Mtb in vitro* and *in vivo* data (Figure 4A).[9] The data visualization quickly hones in on the features important at R^1 and R^2 (*i*-propylamino, and *n*-propylamino groups) in cells that are green (Figure 4B). This type of approach could be useful for interrogating other structure−activity relationships within the complete *in vivo* data set.

Machine Learning Models for *Mtb in vivo* Data. Machine learning models are educated by example and provide an excellent opportunity to discern actives from inactives by both their physiochemical and structural properties.[54] They have an additional benefit of enabling prediction of additional sets of small molecules with a significant degree of accuracy.[8,22,28,29] Previously, we have reported the validation of support vector machine (SVM), recursive partitioning (RP) Single Tree, and RP Forest models to compare with Bayesian models of *in vitro* antitubercular efficacy with acceptable (selectivity index >10) Vero cell cytotoxicity.[29] These types of models (Bayesian, SVM, and RP) are commonly used for drug discovery applications in virtual screening and balance fitting the training set data with external predictive capability outside of the training set's chemical property space. Such approaches have been described by us in some detail previously,[29,55] and the machine learning with the *Mtb in vivo* data paralleled our practices with previous *Mtb in vitro* data sets.[8,22,28,29] We utilized FCFP_6 fingerprints[55,56] and the following set of readily interpretable molecular descriptors: ALogP,[57] molecular weight, number of H-bond

dx.doi.org/10.1021/ci500077v | J. Chem. Inf. Model. 2014, 54, 1070−1082

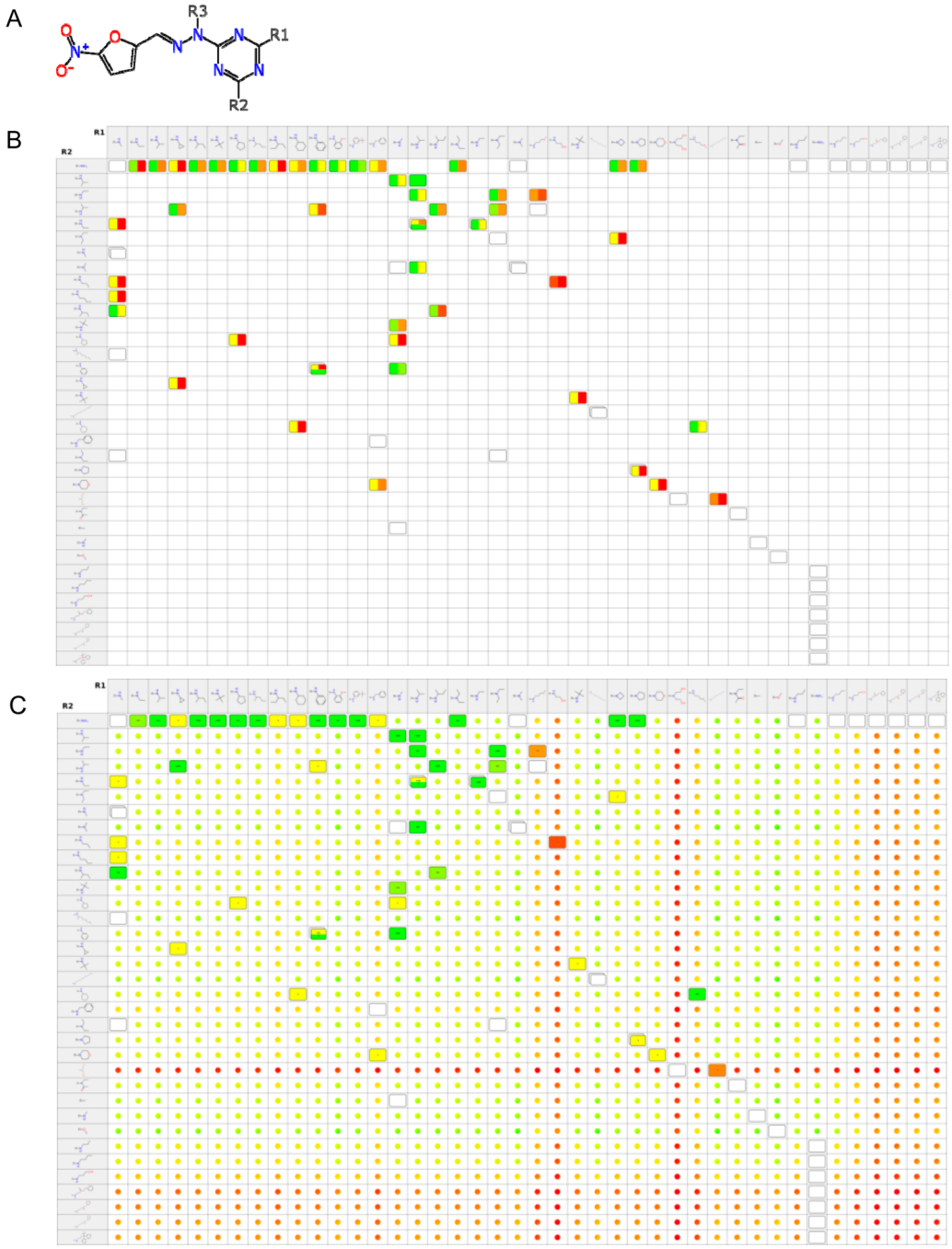


Figure 4. (A) Triazine Markush structure for analogs of TCMDC-125802 (R^1 = R^2 = NHPh; R^3 = H). (B) Matrix correlation plot showing cells with *Mtb in vitro* (left) and *in vivo* (right) data. (C) Solid cells are used to show assayed compounds, and colored dots for activity estimates for hypothetical compounds using internally generated predictions. Green is a favorable. Red is unfavorable. Yellow is intermediate.

Table 2. Mean (± sd) Leave-One-Out and Leave-Out 50% × 100 Cross Validation of Bayesian Models[a]

leave-one-out ROC	leave-out 50% × 100 external ROC score	leave-out 50% × 100 internal ROC score	leave-out 50% × 100 concordance	leave-out 50% × 100 specificity	leave-out 50% × 100 sensitivity
0.77	0.72 ± 0.02	0.74 ± 0.02	66.91 ± 2.24	74.23 ± 8.96	58.46 ± 9.19

[a]ROC = receiver operator characteristic. Best split −2.195.

donors, number of H-bond acceptors, number of rings, number of aromatic rings, number of rotatable bonds, and molecular fractional polar surface area (FPSA). The Bayesian model statistics for 773 molecules, generated by leaving out 50% of the data and rebuilding the model 100 times using a custom protocol for validation to produce the cross-validated receiver operator curve area under the curve (XV ROC AUC), concordance, specificity, and sensitivity as described previously,[8,28] are shown in Table 2. The leave-out 50% × 100 external ROC score was 0.72, while the concordance (66.91), specificity (74.23), and sensitivity (58.46) suggested a bias toward predicting inactive compounds. The 773 molecule Bayesian model provides almost

dx.doi.org/10.1021/ci500077v | J. Chem. Inf. Model. 2014, 54, 1070−1082

identical ROC AUC values with leave-one-out (0.77), leave-out 50% × 100 (0.72), and the five-fold cross validation (0.73, Tables 2 and 3). With five-fold cross validation (leave out 20% × 5), the

Table 3. Individual Machine Learning Model Cross Validation Receiver Operator Curve Statistics for 773 Molecules Tested in the Mouse *in Vivo* Model for *Mtb*[a]

RP Forest (out of bag ROC)	RP Single Tree (with five-fold cross validation ROC)	SVM (with five-fold cross validation ROC)	Bayesian (with five-fold cross validation ROC)
0.75	0.71	0.77	0.73

[a]Bayesian five-fold cross validation has sensitivity = 66.3%, specificity = 90.3%, and concordance = 79.0%.

concordance (79.0%), specificity (90.3%), and sensitivity (66.3%) also suggested a bias toward predicting inactive compounds (Table 3), although all the values are higher compared to leave out 50% × 100 fold. This result highlights the importance of testing different hold out groups and illustrates the model stability based on these values. Using the FCFP-6 descriptors, the top 20 substructure descriptors consistent with both activity and relative lack of cytotoxicity all are derived from the riminophenazine core (Figure S2, Supporting Information), while features of inactives are derived from the thioester, 2,6-disubstituted phenol, and guanidine chemotypes (Figure S3, Supporting Information). This result is not surprising given that clofazimine and all 10 of its analogs were active. If we removed these 11 compounds, the effect on the five-fold cross validation statistics was minimal (Figures S4 and S5,Supporting Information) but the clofazamine-related active features were replaced with others, such as *O*,*S*-disubstituted carbonothioate, 2,3-dihydro-7*H*-[1,4]oxazino[2,3,4-*ij*]quinolin-7-one, and features derived from the fluoroquinolones (Figures S4 and S5, Supporting Information). All machine learning methods showed comparable ROC AUC values (0.71−0.77) using five-fold internal validation for the mouse *in vivo Mtb* data set (Table 3). The SVM model has the best statistics based on the five-fold cross validation with a ROC value of 0.77. In summary, these results suggest that statistically valid computational models can be derived that could be used for predicting new molecules.

Model Predictions for Additional Compounds Identified after Model Building. A compilation of 3500 compounds tested against *Mtb* in the 1950s[40] was identified after the initial data compilation for the training set. Within this data set, 11 compounds were tested and were all active in the mouse *in vivo* model.[40] Of these, only seven were not in the 773 compounds we compiled. Three recent manuscripts described four additional active compounds.[41,42,58] Together these 11 compounds not in our training set (Figure S6, Supporting Information) were scored with all the computational models developed. The Bayesian model correctly predicted 8 out of 11 actives and outperformed the other models (Table 4).

Model Predictions for GSK Compounds To Prioritize for *in Vivo* Testing. The 177 *Mtb* leads (actives *in vitro*) recently disclosed by GSK[12] were scored with the machine learning models. When PCA is performed on these compounds and the *in vivo* data training set, the GSK compounds appear to be relatively localized in just a part of the *in vivo* data set chemistry property space (Figure S7A, Supporting Information). The predicted human intestinal absorption using AlogP and PSA suggests that the *in vivo* data set is quite divergent (Figure S8A, Supporting Information), while the GSK compounds are tightly clustered in a more drug-like area of property space for these two descriptors (Figure S8B, Supporting Information). The Bayesian, SVM, RP Forest, and RP Single Tree models classed 85, 133, 41, and 85 compounds as active *in vivo*, respectively (Table S3, Supporting Information). There were statistically significant differences between predicted active and inactive compounds when one looked at the FPSA and hydrogen bond acceptor counts, as these had lower mean values in predicted active compounds (Table S4, Supporting Information). Twenty-four predicted actives were common across all four models (Table S5, Supporting Information). These 24 compounds were analyzed separately along with the *in vivo* TB data set and appear to reside well within the chemistry property space of the *in vivo* TB data set molecules (Figure S7B, Supporting Information). The putative targets for these compounds were also assessed using TB Mobile[32] and clustered with the TB mobile data set. This result highlighted several compounds as likely mycolic acid transporter (MmpL3) and ubiquinol cytochrome C reductase (QcrB) inhibitors as well as one as a potential dihydrofolate reductase (DHFR) inhibitor (Table S5, Supporting Information). The mean closest distance to the training set was 0.49 (range 0.09−0.74, where larger numbers are more dissimilar, and a value of zero represents the molecule is in the training set). We are aware of a single report of *in vivo* data for these compounds, which is an assay of GSK 1589673A.[59]

DISCUSSION

A recent review has described a timeline for the introduction of the antitubercular drugs.[60] Several of these date back to the 1940s and 1950s and span up to the late 1960s. Only recently (2012) has a new drug, namely, bedaquiline, been approved. Although there are several drugs in clinical trials, the pipeline is relatively thin for a disease where drug resistance has a significant impact and extensively drug resistant *Mtb* is present in nearly 60 countries.[61] The last 10−15 years have witnessed an upswing in high-throughput screening in an attempt to identify molecules that modulate perceived essential targets[62,63] or from phenotypic screening in whole cells.[12−17] The result has been the discovery of about 2000 *in vitro Mtb* hits and perhaps 100s of promising leads. The next hurdle is likely the selection of appropriate compounds to test in the mouse *in vivo* model of infection. The history of this model itself dates back to the 1940s, and even though it has limitations in extrapolating to humans,[64] it is the only animal model that has been validated with human subjects in guiding TB drug development.[25] The mouse represents an expensive medium-throughput model and a bottleneck in screening when used for rank-ordering compounds. On the basis of a recent literature analysis of publications over a 12-year period, there was a five-fold increase in use of the TB mouse model from 1997 to 2009.[65] Our own analysis looks at a much longer time period, collecting data from over 70 years and illustrates that between 1970 to 2000 there was a gap in the publication of mouse *in vivo* data (Figure S1, Supporting Information), with just 55 compounds retrieved in this period out of the total 773. We see no change in this previously reported increased utilization of the mouse model based on the large number of papers describing *in vivo* data for approximately 200 compounds from the last four years (Table S1, Supporting Information), and it is therefore imperative that we question the current workflow and ask how greater cost- and time-efficiencies can be achieved. These data may also suggest some degree of publication bias toward actives.

Table 4. Test Set of *in Vivo* Active Compounds Not in the TB *in Vivo* Models[a]

name (number or abbreviation relates to original nomenclature)	Forest	Single Tree	Bayesian score	Bayesian class	closest distance	SVM	ref
1070 − anisaldehyde, thiosemicarbazone	0	0	−3.02	0	0.35	0	40
1493 − 1-(*p*-methoxybenzyl)-3-thiosemicarbazide	0	0	−1.19	1	0.46	0	40
2403 − *p*-nitrobenzaldehyde, thiosemicarbazone	0	0	−3.70	0	0.44	0	40
2406 − D-threo-*α*,*α*-dichloro-N-[*β*-hydroxy-*α*-(hydroxymethyl)-*p*-nitrophenethyl] acetamide(chloromyecetin)	0	0	1.54	1	0.53	1	40
2875 − nicotinamide	0	1	−1.06	1	0.40	1	40
viomycin	0	1	10.95	1	0.27	1	40
neomycin	1	0	10.11	1	0.01	1	40
PCIH − 2-pyridylcarboxaldehyde *p*-nitrobenzoyl hydrazine	1	1	−1.53	1	0.39	1	41
Cpd 3 − N-(2-fluoroethyl)-1-((6-methoxy-5-methylpyrimidin-4-yl)-methyl)-1H-pyrrolo[3,2-b]pyridine-3-carboxamide	1	1	−0.44	1	0.56	1	42
Cpd 4 −- N-(cyclopropylmethyl)-1-((6-methoxy-5-methylpyrimidin-4-yl)methyl)-1H-pyrrolo[3,2-b]pyridine-3-carboxamide	0	0	−2.06	1	0.53	0	42
indoleamide 3	0	0	−6.20	0	0.31	1	58

[a]Prediction scores 1 = active, 0 = inactive.

While we need to better understand how *in vitro* efficacy, absorption, metabolism, distribution and excretion (ADME), and *in vivo* pharmacokinetic profiles are linked to *in vivo* antitubercular activity, we also need to computationally learn from the collective experiences of the considerable number of generally small molecules (773 in this study alone) that have been tested in the mouse model. These data serve as the training set for an *educated* computational model to prioritize hits for *in vivo* testing, minimizing the assessment of compounds likely to fail. Several computational models are available to *predict in vitro* ADME properties in addition to other physicochemical properties.[66−75] While we and others have utilized machine learning models to predict *Mtb* activity *in vitro*,[8,22,26−28,31,33,54,76−78] the studies to date have not analyzed *Mtb in vivo* data in mice. However, mathematical modeling of pathogen and host interactions pertinent to latent infection have been reported.[79] In total, our utilization of machine learning methods with *Mtb* has covered external model validation, hit discovery, lead optimization, data set fusion, and now *in vivo* data analysis.[8,13,22,28,33,76] One reason for this previous lack of modeling *Mtb in vivo* is perhaps because the achievement of *in vivo* efficacy is multi-factorial due to complex interactions between *Mtb*, host, and drug. Additionally, the data to conduct this *in vivo* modeling exercise had not been previously curated until now.

The curation of this *in vivo* data, as with all literature data capture, must be conducted with attention paid to structure deposition errors. For example, after testing in mice, it was suggested that 27753-RP was a highly promising antituberculosis drug.[80] Subsequently, in the absence of supportive data, it was stated that griselimycin(e) while effective *in vitro* against *Mtb* was inactive *in vivo*. We were unable to find any further information on these compounds in the literature beyond their structures. In the process, we noted that the griselimycin structure does not define all of the stereocenters. Finding 27753-RP proved another challenge, as it was characterized by the wrong molecular formula $C_{67}H_{11}N_{10}O_{11}$ (the very low number of hydrogens was noted as compared with griselimycin $C_{57}H_{96}N_{10}O_{12}$). Searching ChemSpider for $C_{67}H_{116}N_{10}O_{11}$ retrieved the compound as the synonyms list contained "27753R.P.". The additional periods in the compound name rendered it invisible to previous searches for the literature name of "22753-RP." Again, stereocenters were not defined for the structure. This compound may deserve further characterization because it was active against a rifampicin-resistant strain of *Mtb*, and it raises the concern that other important molecules active against *Mtb* may be hidden in publications and databases while being obscured by synonyms or frank errors.

Our retrospective analysis of the data for small molecules (monotherapy only) in *Mtb*-infected mice has been critical to the realization of what physicochemical properties and chemical features best describe the actives (Figure S2, Supporting Information) as well as inactives (Figure S3, Supporting Information). This analysis is important because many promising *in vitro* active compounds do not show *in vivo* activity.[7,8] The ring count was significantly higher in *in vivo* active compounds, and the AlogP and fractional polar surface area were lower than in inactives. Intriguingly, our analysis of *Mtb in vitro* HTS results has shown actives to have a higher calculated logP than inactives.[31,76] We cannot also discount possible differences between methods for predicting logP. The reliance on a single molecular descriptor may be suboptimal. The curated *in vivo* data was also analyzed using multiple interpretable descriptors and machine learning models in this study. All four of the methods we have described resulted in similar receiver operator curve (ROC) values (0.71−0.77), indicative of potentially useful models. These models can score small molecules (absent in the training set) to prioritize antitubercular hits and leads for crucial *in vivo* studies. We evaluated the models with 11 known *in vivo* actives and found the Bayesian model outperformed the SVM and tree methods (Table 4). The Bayesian model correctly identified 8 out of 11 small molecules, absent from the training set, as *in vivo* actives. Compounds like the recently published indoleamide 3[58] (Table 4, Figure S6, Supporting Information) may have been poorly predicted due to the absence of features such as the cyclooctyl ring in the training set molecules, and the most similar compound (missing this feature) was classed as inactive. This points to the need for a structurally diverse training set and understanding the threshold similarity distance for making a prediction.[29] Scoring of a set of 177 *in vitro* active compounds from GSK[12] with the different machine learning models showed some variation in the number predicted as likely active *in vivo*. Twenty-four compounds were predicted as consensus actives (Table S5, Supporting Information). This result represents a potential data set for testing the *in vivo* model predictions. Their further prioritization for *in vivo* study could also rely on predicted ADME/Tox profiles and *Mtb* targets (Table S6, Supporting Information). Additionally, with the attainment of more

extensive SAR for these chemical series, one could apply the scaffold analysis approach illustrated herein for the triazines.

We have described a data curation and machine learning process that could be further expanded to capture the remaining *public Mtb in vivo* data that we have been unable to find to date. As older articles prior to 1970 may not have an abstract in PubMed, we also used SciFinder and Web of Knowledge to expand our possible range of journals covered to extract as many molecules as possible. Ultimately, training and test set molecules were extracted from 119 references (Table S1, Supporting Information). Another consideration when curating data from old papers is that there are few clearly drawn structures, and it can be a very complex and tedious process even for experienced medicinal chemists to definitively identify structures. For this reason, some structures and data were excluded because of this uncertainty or structural ambiguity.

While prospectively testing the models would be preferred, the high cost of *in vivo* experiments (currently ~$5000 per compound) renders this practically difficult, and much of this testing is coordinated by the NIH at academic laboratories. Traditionally, the use of computational machine learning models in other areas for which testing is relatively expensive (e.g., ADME/Tox models like hERG requiring patch clamping data[81−83]) has involved prospective testing on a long time scale (10 years) as more data is generated by different groups and as more higher throughput techniques are developed. This also enables iterative model building and updating to expand the chemistry property space covered, while also expanding the scope of algorithms tested and compared with each other.

In conclusion, we suggest that the machine learning methods described here utilized for modeling *Mtb in vivo* data could also be applied to other diseases for which similar information from a pertinent animal model is available. Currently, to our knowledge, there are no databases that cover or curate *in vivo* animal model data, and other sources like PubChem focus almost exclusively on *in vitro* HTS data. Several limitations of the current study include poor extrapolation from mouse (or other animal models) to human due to likely ADME differences.[25,64] In addition, we have not yet pursued modeling combination therapies and potential synergies and antagonisms of experimental antituberculars with known drug treatments. We have also not removed prodrugs from our data set, and we have combined data from several dosing routes, both of which may add to the noise in the models. It is important to note that the four computational models do not *per se* replace *in vivo* studies but instead they may enable prioritization of molecules that are likely to perform well *in vivo*. This approach has obvious benefits for decreasing costs and reducing animal testing[65] as well as continuing to enrich and accelerate the tuberculosis drug discovery process by learning from prior data. It may also be a useful approach for design of antituberculars or at the very least prioritizing compounds for host-derived therapeutics for TB.[84] This work also calls for more sharing of data so that we can avoid repeating the discoveries and failures of the past in the future.

■ ASSOCIATED CONTENT

Supporting Information

Seven figures, six tables, and supplemental references. All computational models are available from the authors upon request. All molecules used in the models are available in CDD, ChemSpider, and FigShare. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: ekinssean@yahoo.com (S.E.). Phone: (973) 972-7165. Fax: (973) 972-7950.

*E-mail: freundjs@njms.rutgers.edu (J.S.F.). Phone: (973) 972-7165. Fax: (973) 972-7950.

Author Contributions

S.E. and J.S.F. contributed equally.

Notes

The authors declare the following competing financial interest(s): S.E. is a consultant for Collaborative Drug Discovery, Inc.

■ ACKNOWLEDGMENTS

S.E. acknowledges colleagues at CDD, in particular Anna Coulon Spektor for vault assistance, and Dr. Malabika Sarker at SRI for supplying the list of TB metabolites from BioCyc. Dr. Barbara Laughon (NIAID) is kindly acknowledged for pointing us to the compilation of data by Youmans et al. Accelrys is kindly acknowledged for providing Discovery Studio and Dr. Katalin Nadassy for her support. The models were created in Discovery Studio. S.E. acknowledges that the earlier Bayesian models, which informed this work, were developed with support from Award Number R43 LM011152-01 "Biocomputation across distributed private datasets to enhance drug discovery" from the National Library of Medicine. TB Mobile and the associated dataset used in this study were developed with funding from Award Number 2R42AI088893-02 "Identification of novel therapeutics for tuberculosis combining cheminformatics, diverse databases and logic based pathway analysis" from the National Institutes of Allergy and Infectious Diseases. J.S.F. acknowledges funding from NIH/NIAID (2R42AI088893-02) and Rutgers University−NJMS.

■ ABBREVIATIONS USED

ADME, absorption, metabolism, distribution, and excretion; CFUs, colony-forming units; DHFR, dihydrofolate reductase; FCFP_6, molecular function class fingerprints of maximum diameter 6; GSK, GlaxoSmithKline; HTS, high-throughput screening; hERG, human ether-a-go-go related-gene; MMDS, Mobile Molecular DataSheet; *Mtb*, *Mycobacterium tuberculosis*; NIH, National Institutes of Health; PCA, principal components analysis; PXR, pregnane X-receptor; SAR, structure−activity relationship; RP, recursive partitioning; SI, selectivity index; SVM, support vector machine; TB, tuberculosis; XV ROC AUC, cross-validated receiver operator curve area under the curve

■ REFERENCES

(1) Nuermberger, E. L.; Spigelman, M. K.; Yew, W. W. Current development and future prospects in chemotherapy of tuberculosis. *Respirology* **2010**, *15* (5), 764−778.

(2) Global Tuberculosis Report 2013. World Health Organization. http://www.who.int/tb/publications/global_report/en/.

(3) Zhang, Y. The magic bullets and tuberculosis drug targets. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 529−564.

(4) Ballel, L.; Field, R. A.; Duncan, K.; Young, R. J. New small-molecule synthetic antimycobacterials. *Antimicrob. Agents Chemother.* **2005**, *49*, 2153−2163.

(5) Ma, Z.; Lienhardt, C.; McIlleron, H.; Nunn, A. J.; Wang, X. Global tuberculosis drug development pipeline: The need and the reality. *Lancet* **2010**, *375* (9731), 2100−2109.

(6) Wong, E. B.; Cohen, K. A.; Bishai, W. R. Rising to the challenge: New therapies for tuberculosis. *Trends Microbiol.* **2013**, *21* (9), 493−501.
(7) Gold, B.; Pingle, M.; Brickner, S. J.; Shah, N.; Roberts, J.; Rundell, M.; Bracken, W. C.; Warrier, T.; Somersan, S.; Venugopal, A.; Darby, C.; Jiang, X.; Warren, J. D.; Fernandez, J.; Ouerfelli, O.; Nuermberger, E. L.; Cunningham-Bussel, A.; Rath, P.; Chidawanyika, T.; Deng, H.; Realubit, R.; Glickman, J. F.; Nathan, C. F. Nonsteroidal anti-inflammatory drug sensitizes *Mycobacterium tuberculosis* to endogenous and exogenous antimicrobials. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (40), 16004−16011.
(8) Ekins, S.; Reynolds, R.; Kim, H.; Koo, M.-S.; Ekonomidis, M.; Talaue, M.; Paget, S. D.; Woolhiser, L. K.; Lenaerts, A. J.; Bunin, B. A.; Connell, N.; Freundlich, J. S. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. *Chem. Biol.* **2013**, *20*, 370−378.
(9) Bruhin, H.; Buhlmann, X.; Hook, W. H.; Hoyle, W.; Orford, B.; Vischer, W. Antituberculosis activity of some nitrofuran derivatives. *J. Pharm. Pharmacol.* **1969**, *21* (7), 423−33.
(10) Hoffmann, K.; Onoz, E. Inhibitory effect of oxyphenbutazone against *Mycobacterium tuberculosis* in vitro. *Arzneimittelforschung* **1969**, *19* (2), 241−242.
(11) Rao, S. P.; Lakshminarayana, S. B.; Kondreddi, R. R.; Herve, M.; Camacho, L. R.; Bifani, P.; Kalapala, S. K.; Jiricek, J.; Ma, N. L.; Tan, B. H.; Ng, S. H.; Nanjundappa, M.; Ravindran, S.; Seah, P. G.; Thayalan, P.; Lim, S. H.; Lee, B. H.; Goh, A.; Barnes, W. S.; Chen, Z.; Gagaring, K.; Chatterjee, A. K.; Pethe, K.; Kuhen, K.; Walker, J.; Feng, G.; Babu, S.; Zhang, L.; Blasco, F.; Beer, D.; Weaver, M.; Dartois, V.; Glynne, R.; Dick, T.; Smith, P. W.; Diagana, T. T.; Manjunatha, U. H. Indolcarboxamide is a preclinical candidate for treating multidrug-resistant tuberculosis. *Sci. Transl. Med.* **2013**, *5* (214), 214ra168.
(12) Ballell, L.; Bates, R. H.; Young, R. J.; Alvarez-Gomez, D.; Alvarez-Ruiz, E.; Barroso, V.; Blanco, D.; Crespo, B.; Escribano, J.; Gonzalez, R.; Lozano, S.; Huss, S.; Santos-Villarejo, A.; Martin-Plaza, J. J.; Mendoza, A.; Rebollo-Lopez, M. J.; Remuinan-Blanco, M.; Lavandera, J. L.; Perez-Herran, E.; Gamo-Benito, F. J.; Garcia-Bustos, J. F.; Barros, D.; Castro, J. P.; Cammack, N. Fueling open-source drug discovery: 177 small-molecule leads against tuberculosis. *ChemMedChem* **2013**, *8*, 313−321.
(13) Ananthan, S.; Faaleolea, E. R.; Goldman, R. C.; Hobrath, J. V.; Kwong, C. D.; Laughon, B. E.; Maddry, J. A.; Mehta, A.; Rasmussen, L.; Reynolds, R. C.; Secrist, J. A., 3rd; Shindo, N.; Showe, D. N.; Sosa, M. I.; Suling, W. J.; White, E. L. High-throughput screening for inhibitors of *Mycobacterium tuberculosis* H37Rv. *Tuberculosis (Edinb)* **2009**, *89*, 334−353.
(14) Maddry, J. A.; Ananthan, S.; Goldman, R. C.; Hobrath, J. V.; Kwong, C. D.; Maddox, C.; Rasmussen, L.; Reynolds, R. C.; Secrist, J. A., 3rd; Sosa, M. I.; White, E. L.; Zhang, W. Antituberculosis activity of the molecular libraries screening center network library. *Tuberculosis (Edinb)* **2009**, *89*, 354−363.
(15) Reynolds, R. C.; Ananthan, S.; Faaleolea, E.; Hobrath, J. V.; Kwong, C. D.; Maddox, C.; Rasmussen, L.; Sosa, M. I.; Thammasuvimol, E.; White, E. L.; Zhang, W.; Secrist, J. A., 3rd High throughput screening of a library based on kinase inhibitor scaffolds against *Mycobacterium tuberculosis* H37Rv. *Tuberculosis (Edinb)* **2012**, *92*, 72−83.
(16) Grant, S. S.; Kawate, T.; Nag, P. P.; Silvis, M. R.; Gordon, K.; Stanley, S. A.; Kazyanskaya, E.; Nietupski, R.; Golas, A.; Fitzgerald, M.; Cho, S.; Franzblau, S. G.; Hung, D. T. Identification of novel inhibitors of nonreplicating *Mycobacterium tuberculosis* using a carbon starvation model. *ACS Chem. Biol.* **2013**, *8* (10), 2224−2234.
(17) Stanley, S. A.; Grant, S. S.; Kawate, T.; Iwase, N.; Shimizu, M.; Wivagg, C.; Silvis, M.; Kazyanskaya, E.; Aquadro, J.; Golas, A.; Fitzgerald, M.; Dai, H.; Zhang, L.; Hung, D. T. Identification of novel inhibitors of *M. tuberculosis* growth using whole cell based high-throughput screening. *ACS Chem. Biol.* **2012**, *7*, 1377−1384.
(18) Lee, R. E.; Protopopova, M.; Crooks, E.; Slayden, R. A.; Terrot, M.; Barry, C. E., 3rd Combinatorial lead optimization of [1,2]-diamines based on ethambutol as potential antituberculosis preclinical candidates. *J. Comb. Chem.* **2003**, *5* (2), 172−187.
(19) Andries, K.; Verhasselt, P.; Guillemont, J.; Gohlmann, H. W.; Neefs, J. M.; Winkler, H.; Van Gestel, J.; Timmerman, P.; Zhu, M.; Lee, E.; Williams, P.; de Chaffoy, D.; Huitric, E.; Hoffner, S.; Cambau, E.; Truffot-Pernot, C.; Lounis, N.; Jarlier, V. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* **2005**, *307* (5707), 223−227.
(20) Collins, F. S. Reengineering translational science: The time is right. *Sci. Transl. Med.* **2011**, *3* (90), 90cm17.
(21) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10* (3), 188−195.
(22) Ekins, S.; Freundlich, J. S.; Hobrath, J. V.; White, E. L.; Reynolds, R. C. Combining computational methods for hit to lead optimization in *Mycobacterium tuberculosis* drug discovery. *Pharm. Res.* **2013**, *31*, 414−435.
(23) Chang, K. C.; Leung, C. C.; Grosset, J.; Yew, W. W. Treatment of tuberculosis and optimal dosing schedules. *Thorax* **2011**, *66* (11), 997−1007.
(24) Lenaerts, A. J.; Gruppo, V.; Brooks, J. V.; Orme, I. M. Rapid in vivo screening of experimental drugs for tuberculosis using gamma interferon gene-disrupted mice. *Antimicrob. Agents Chemother.* **2003**, *47* (2), 783−5.
(25) Franzblau, S. G.; DeGroote, M. A.; Cho, S. H.; Andries, K.; Nuermberger, E.; Orme, I. M.; Mdluli, K.; Angulo-Barturen, I.; Dick, T.; Dartois, V.; Lenaerts, A. J. Comprehensive analysis of methods used for the evaluation of compounds against *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* **2012**, *92* (6), 453−488.
(26) Ekins, S.; Freundlich, J. S. Validating new tuberculosis computational models with public whole cell screening aerobic activity datasets. *Pharm. Res.* **2011**, *28*, 1859−1869.
(27) Sarker, M.; Talcott, C.; Madrid, P.; Chopra, S.; Bunin, B. A.; Lamichhane, G.; Freundlich, J. S.; Ekins, S. Combining cheminformatics methods and pathway analysis to identify molecules with whole-cell activity against Mycobacterium tuberculosis. *Pharm. Res.* **2012**, *29*, 2115−2127.
(28) Ekins, S.; Reynolds, R. C.; Franzblau, S. G.; Wan, B.; Freundlich, J. S.; Bunin, B. A. Enhancing hit identification in *Mycobacterium tuberculosis* drug discovery using validated dual-event Bayesian models. *PLoS One* **2013**, *8*, e63240.
(29) Ekins, S.; Freundlich, J. S.; Reynolds, R. C. Fusing dual-event datasets for *Mycobacterium tuberculosis* machine learning models and their evaluation. *J. Chem. Inf. Model.* **2013**, *53*, 3054−3063.
(30) Clark, A. M. Mobile Molecular DataSheet (MMDS). http://molmatinf.com/products.html#mmds.
(31) Ekins, S.; Bradford, J.; Dole, K.; Spektor, A.; Gregory, K.; Blondeau, D.; Hohman, M.; Bunin, B. A collaborative database and computational models for tuberculosis drug discovery. *Mol BioSyst.* **2010**, *6*, 840−851.
(32) Ekins, S.; Clark, A. M.; Sarker, M. TB Mobile: A mobile app for anti-tuberculosis molecules with known targets. *J. Cheminform.* **2013**, *5*, 13.
(33) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian models for the prioritization of antitubercular agents. *J. Chem. Inf. Model.* **2008**, *48* (12), 2362−2370.
(34) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2* (6), 861−873.
(35) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J. Chem. Inf. Model.* **2006**, *46* (5), 1945−1956.
(36) Hassan, M.; Brown, R. D.; Varma-O'brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *10* (3), 283−299.

(37) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10* (7), 682−686.
(38) Jones, D. R.; Ekins, S.; Li, L.; Hall, S. D. Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). *Drug Metab. Dispos.* **2007**, *35* (9), 1466−1475.
(39) The R Project for Statistical Computing. http://www.r-project.org/.
(40) Youmans, G. P.; Doub, l.; Youmans, A. S. *The Bacteriostatic Activity of 3500 Organic Compounds for Mycobacterium tuberculosis var. Hominis*; National Research Council: Washington, DC, 1953.
(41) Ellis, S.; Kalinowski, D. S.; Leotta, L.; Huang, M. L.; Jelfs, P.; Sintchenko, V.; Richardson, D. R.; Triccas, J. A. Potent anti-mycobacterial activity of the pyridoxal isonicotinoyl hydrazone analogue, 2-pyridylcarboxaldehyde isonicotinoyl hydrazone: A lipophilic transport vehicle for isonicotinic acid hydrazide. *Mol. Pharmacol.* **2014**, *85*, 269−278.
(42) Shirude, P. S.; Shandil, R.; Sadler, C.; Naik, M.; Hosagrahara, V.; Hameed, S.; Shinde, V.; Bathula, C.; Humnabadkar, V.; Kumar, N.; Reddy, J.; Panduga, V.; Sharma, S.; Ambady, A.; Hegde, N.; Whiteaker, J.; McLaughlin, R. E.; Gardner, H.; Madhavapeddi, P.; Ramachandran, V.; Kaur, P.; Narayan, A.; Guptha, S.; Awasthy, D.; Narayan, C.; Mahadevaswamy, J.; Vishwas, K.; Ahuja, V.; Srivastava, A.; Prabhakar, K.; Bharath, S.; Kale, R.; Ramaiah, M.; Choudhury, N. R.; Sambandamurthy, V. K.; Solapure, S.; Iyer, P. S.; Narayanan, S.; Chatterji, M. Azaindoles: Noncovalent DprE1 inhibitors from scaffold morphing efforts, kill *Mycobacterium tuberculosis* and are efficacious in vivo. *J. Med. Chem.* **2013**, *56*, 9701−9708.
(43) Ekins, S. Progress in computational toxicology. *J. Pharmacol. Toxicol. Methods* **2014**, *69*, 115−140.
(44) Clark, A. M. SAR Table. http://molmatinf.com/products.html#sartable.
(45) Blaser, A.; Palmer, B. D.; Sutherland, H. S.; Kmentova, I.; Franzblau, S. G.; Wan, B.; Wang, Y.; Ma, Z.; Thompson, A. M.; Denny, W. A. Structure-activity relationships for amide-, carbamate-, and urea-linked analogues of the tuberculosis drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]oxy}-6,7-dihydro-5H-imidazo[2,1-b][1, 3]-oxazine (PA-824). *J. Med. Chem.* **2012**, *55* (1), 312−326.
(46) Thompson, A. M.; Sutherland, H. S.; Palmer, B. D.; Kmentova, I.; Blaser, A.; Franzblau, S. G.; Wan, B.; Wang, Y.; Ma, Z.; Denny, W. A. Synthesis and structure-activity relationships of varied ether linker analogues of the antitubercular drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]oxy}-6,7-dihydro-5h-imidazo[2,1-b][1, 3]-oxazine (PA-824). *J. Med. Chem.* **2012**, *54* (19), 6563−6585.
(47) Palmer, B. D.; Thompson, A. M.; Sutherland, H. S.; Blaser, A.; Kmentova, I.; Franzblau, S. G.; Wan, B.; Wang, Y.; Ma, Z.; Denny, W. A. Synthesis and structure-activity studies of biphenyl analogues of the tuberculosis drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]oxy}-6,7-dihydro-5H-imidazo[2,1-b][1, 3]oxazine (PA-824). *J. Med. Chem.* **2010**, *53* (1), 282−294.
(48) Thompson, A. M.; Blaser, A.; Anderson, R. F.; Shinde, S. S.; Franzblau, S. G.; Ma, Z.; Denny, W. A.; Palmer, B. D. Synthesis, reduction potentials, and antitubercular activity of ring A/B analogues of the bioreductive drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]-oxy}-6,7-dihydro-5H-imidazo[2,1-b][1, 3]oxazine (PA-824). *J. Med. Chem.* **2009**, *52* (3), 637−645.
(49) Thompson, A. M.; Sutherland, H. S.; Palmer, B. D.; Kmentova, I.; Blaser, A.; Franzblau, S. G.; Wan, B.; Wang, Y.; Ma, Z.; Denny, W. A. Synthesis and structure-activity relationships of varied ether linker analogues of the antitubercular drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]oxy}-6,7-dihydro-5h-imidazo[2,1-b][1, 3]-oxazine (PA-824). *J. Med. Chem.* **2011**, *54* (19), 6563−6585.
(50) Kmentova, I.; Sutherland, H. S.; Palmer, B. D.; Blaser, A.; Franzblau, S. G.; Wan, B.; Wang, Y.; Ma, Z.; Denny, W. A.; Thompson, A. M. Synthesis and structure-activity relationships of aza- and diazabiphenyl analogues of the antitubercular drug (6S)-2-nitro-6-{[4-(trifluoromethoxy)benzyl]oxy}-6,7-dihydro-5H-imidazo[2,1-b][1, 3]-oxazine (PA-824). *J. Med. Chem.* **2010**, *53*, 8421−8439.
(51) Caspi, R.; Foerster, H.; Fulcher, C. A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S. Y.; Shearer, A. G.; Tissier, C.; Walk, T. C.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2008**, *36* (Database issue), D623−D631.
(52) Galagan, J. E.; Sisk, P.; Stolte, C.; Weiner, B.; Koehrsen, M.; Wymore, F.; Reddy, T. B.; Zucker, J. D.; Engels, R.; Gellesch, M.; Hubble, J.; Jin, H.; Larson, L.; Mao, M.; Nitzberg, M.; White, J.; Zachariah, Z. K.; Sherlock, G.; Ball, C. A.; Schoolnik, G. K. TB database 2010: overview and update. *Tuberculosis (Edinb)* **2010**, *90* (4), 225−235.
(53) Bruhin, H.; Buhlmann, X.; Hook, W. H.; Hoyle, W.; Orford, B.; Vischer, W. Antituberculosis activity of some nitrofuran derivatives. *J. Pharm. Pharmacol.* **1969**, *21* (7), 423−433.
(54) Ekins, S.; Freundlich, J. S. Computational models for tuberculosis drug discovery. *Methods Mol. Biol.* **2013**, *993*, 245−262.
(55) Ekins, S.; Casey, A. C.; Roberts, D.; Parish, T.; Bunin, B. A. Bayesian models for screening and TB Mobile for target inference with Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **2014**, *94* (2), 162−169.
(56) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24* (7), 805−815.
(57) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem.* **1998**, *102*, 3762−3772.
(58) Lun, S.; Guo, H.; Onajole, O. K.; Pieroni, M.; Gunosewoyo, H.; Chen, G.; Tipparaju, S. K.; Ammerman, N. C.; Kozikowski, A. P.; Bishai, W. R. Indoleamides are active against drug-resistant *Mycobacterium tuberculosis*. *Nat Commun* **2013**, *4*, 2907.
(59) Remuinan, M. J.; Perez-Herran, E.; Rullas, J.; Alemparte, C.; Martinez-Hoyos, M.; Dow, D. J.; Afari, J.; Mehta, N.; Esquivias, J.; Jimenez, E.; Ortega-Muro, F.; Fraile-Gabaldon, M. T.; Spivey, V. L.; Loman, N. J.; Pallen, M. J.; Constantinidou, C.; Minick, D. J.; Cacho, M.; Rebollo-Lopez, M. J.; Gonzalez, C.; Sousa, V.; Angulo-Barturen, I.; Mendoza-Losana, A.; Barros, D.; Besra, G. S.; Ballell, L.; Cammack, N. Tetrahydropyrazolo[1,5-a]pyrimidine-3-carboxamide and N-benzyl-6′,7′-dihydrospiro[piperidine-4,4′-thieno[3,2-c]pyran] analogues with bactericidal efficacy against *Mycobacterium tuberculosis* targeting MmpL3. *PLoS One* **2013**, *8* (4), e60933.
(60) Villemagne, B.; Crauste, C.; Flipo, M.; Baulard, A. R.; Deprez, B.; Willand, N. Tuberculosis: The drug development pipeline at a glance. *Eur. J. Med. Chem.* **2012**, *51*, 1−16.
(61) Dheda, K.; Migliori, G. B. The global rise of extensively drug-resistant tuberculosis: is the time to bring back sanatoria now overdue? *Lancet* **2012**, *379* (9817), 773−775.
(62) Sassetti, C. M.; Rubin, E. J. Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (22), 12989−94.
(63) Sassetti, C. M.; Boyd, D. H.; Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **2003**, *48* (1), 77−84.
(64) Dartois, V.; Barry, C. E., 3rd A medicinal chemists' guide to the unique difficulties of lead optimization for tuberculosis. *Bioorg. Med. Chem. Lett.* **2013**, *23* (17), 4741−4750.
(65) Franco, N. H.; Correia-Neves, M.; Olsson, I. A. Animal welfare in studies on murine tuberculosis: assessing progress over a 12-year period and the need for further improvement. *PLoS One* **2012**, *7* (10), e47723.
(66) Guner, O. F.; Bowen, J. P. Pharmacophore modeling for ADME. *Curr. Top. Med. Chem.* **2013**, *13* (11), 1327−1342.
(67) Gombar, V. K.; Hall, S. D. Quantitative structure-activity relationship models of clinical pharmacokinetics: Clearance and volume of distribution. *J. Chem. Inf. Model.* **2013**, *53* (4), 948−957.
(68) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* **2012**, *52* (11), 3099−3105.

(69) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25* (6), 533−54.
(70) Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discovery* **2011**, *10* (3), 197−208.
(71) Gupta, R. R.; Gifford, E. M.; Liston, T.; Waller, C. L.; Bunin, B.; Ekins, S. Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties. *Drug Metab. Dispos.* **2010**, *38*, 2083−2090.
(72) Ekins, S.; Williams, A. J. Precompetitive preclinical ADME/Tox data: Set it free on the Web to facilitate computational model building to assist drug development. *Lab Chip* **2010**, *10*, 13−22.
(73) Ekins, S.; Honeycutt, J. D.; Metz, J. T. Evolving molecules using multi-objective optimization: Applying to ADME. *Drug Discovery Today* **2010**, *15*, 451−460.
(74) Lagorce, D.; Sperandio, O.; Galons, H.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinform.* **2008**, *9*, 396.
(75) Ekins, S.; Swaan, P. W. Computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. *Rev. Comput. Chem.* **2004**, *20*, 333−415.
(76) Ekins, S.; Kaneko, T.; Lipinksi, C. A.; Bradford, J.; Dole, K.; Spektor, A.; Gregory, K.; Blondeau, D.; Ernst, S.; Yang, J.; Goncharoff, N.; Hohman, M.; Bunin, B. Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*. *Mol. BioSyst.* **2010**, *6*, 2316−2324.
(77) Periwal, V.; Rajappan, J. K.; Jaleel, A. U.; Scaria, V. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC Res. Notes* **2011**, *4*, 504.
(78) Periwal, V.; Kishtapuram, S.; Consortium, O. S.; Scaria, V. Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. *BMC Pharmacol.* **2012**, *12* (1), 1.
(79) Lin, P. L. Modeling pathogen and host: In vitro, in vivo and in silico models of latent Mycobacterium tuberculosis infection. *Drug Discovery Today: Dis. Models* **2005**, *2*, 149−154.
(80) Toyohara, M. Aspects of the antituberculous activity of 27753-RP, a new semisynthetic derivative of griselimycine. *Ann. Inst. Pasteur/Microbiol.* **1987**, *138* (6), 737−744.
(81) Chekmarev, D. S.; Kholodovych, V.; Balakin, K. V.; Ivanenkov, Y.; Ekins, S.; Welsh, W. J. Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem. Res. Toxicol.* **2008**, *21* (6), 1304−1314.
(82) Ekins, S.; Balakin, K. V.; Savchuk, N.; Ivanenkov, Y. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J. Med. Chem.* **2006**, *49* (17), 5059−5071.
(83) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **2002**, *301* (2), 427−434.
(84) Hawn, T. R.; Matheson, A. I.; Maley, S. N.; Vandal, O. Host-directed therapeutics for tuberculosis: Can we harness the host? *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 608−627.

■ NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on April 3, 2014, with minor text errors. The corrected version was published ASAP on April 4, 2014.