JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Machine Learning Model Analysis and Data Visualization with Small Molecules Tested in a Mouse Model of *Mycobacterium tuberculosis* Infection (2014−2015)

Sean Ekins,[*,†,‡,∇] Alexander L. Perryman,[§,∇] Alex M. Clark,[∥] Robert C. Reynolds,[⊥] 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 and Neuroscience, Rutgers University−New Jersey Medical School, Newark, New Jersey 07103, United States

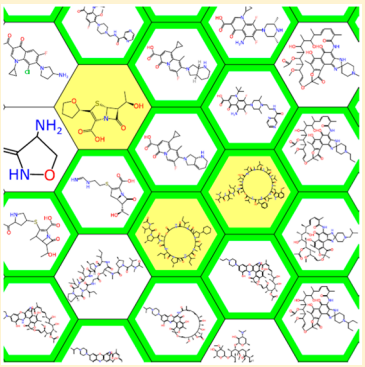
[∥]Molecular Materials Informatics, Inc., 1900 St. Jacques #302, Montreal, Quebec H3J 2S1, Canada

[⊥]Division of Hematology and Oncology, Department of Medicine, and Department of Chemistry, College of Arts and Sciences, University of Alabama at Birmingham, 1530 Third Avenue South, Birmingham, Alabama 35294-1240, United States

[#]Division of Infectious Diseases, Department of Medicine, and the Ruy V. Lourenço Center for the Study of Emerging and Re-emerging Pathogens, Rutgers University−New Jersey Medical School, Newark, New Jersey 07103, United States

Supporting Information

ABSTRACT: The renewed urgency to develop new treatments for *Mycobacterium tuberculosis* (*Mtb*) infection has resulted in large-scale phenotypic screening and thousands of new active compounds *in vitro*. The next challenge is to identify candidates to pursue in a mouse *in vivo* efficacy model as a step to predicting clinical efficacy. We previously analyzed over 70 years of this mouse *in vivo* efficacy data, which we used to generate and validate machine learning models. Curation of 60 additional small molecules with *in vivo* data published in 2014 and 2015 was undertaken to further test these models. This represents a much larger test set than for the previous models. Several computational approaches have now been applied to analyze these molecules and compare their molecular properties beyond those attempted previously. Our previous machine learning models have been updated, and a novel aspect has been added in the form of mouse liver microsomal half-life (MLM $t_{1/2}$) and *in vitro*-based *Mtb* models incorporating cytotoxicity data that were used to predict *in vivo* activity for comparison. Our best *Mtb in vivo* models possess fivefold ROC values > 0.7, sensitivity > 80%, and concordance > 60%, while the best specificity value is >40%. Use of an MLM $t_{1/2}$ Bayesian model affords comparable results for scoring the 60 compounds tested. Combining MLM stability and *in vitro Mtb* models in a novel consensus workflow in the best cases has a positive predicted value (hit rate) > 77%. Our results indicate that Bayesian models constructed with literature *in vivo Mtb* data generated by different laboratories in various mouse models can have predictive value and may be used alongside MLM $t_{1/2}$ and *in vitro*-based *Mtb* models to assist in selecting antitubercular compounds with desirable *in vivo* efficacy. We demonstrate for the first time that consensus models of any kind can be used to predict *in vivo* activity for *Mtb*. In addition, we describe a new clustering method for data visualization and apply this to the *in vivo* training and test data, ultimately making the method accessible in a mobile app.



INTRODUCTION

Tuberculosis (TB) is a major infectious disease that unfortunately knows no geographic boundaries and accounts for approximately 9 million new cases and 1.5 million deaths each year.[1] TB and its etiological agent, *Mycobacterium tuberculosis* (*Mtb*), continue to be the focus of intense international efforts to develop new tools for the control and ultimate elimination[2] of this devastating disease that is increasingly associated with resistance to first- and second-line drugs.[3] The discovery of new TB drug candidates with novel mechanisms of action is of fundamental importance in this regard. The majority of funding for TB research still comes from the NIH NIAID and the Bill and Melinda Gates Foundation. In the past, the European Commission has also funded TB research in the FP7 Program (although nowhere near the levels of the aforementioned organizations). However, no funding for TB small-molecule drug discovery is foreseen in the EC's Horizon 2020 Program over the next few years. These cuts in funding highlight the need to increase the efficiency of tuberculosis small-molecule drug

Received: January 4, 2016
Published: June 22, 2016

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

discovery. Analysis of the recent pipeline at TB Alliance[4] and elsewhere[5] reveals that while there are ~27 projects in preclinical stages and 13 in clinical trials in phases 1−3, only one project is in phase 4 (Figure 1). This indicates a suboptimal pipeline. Of the

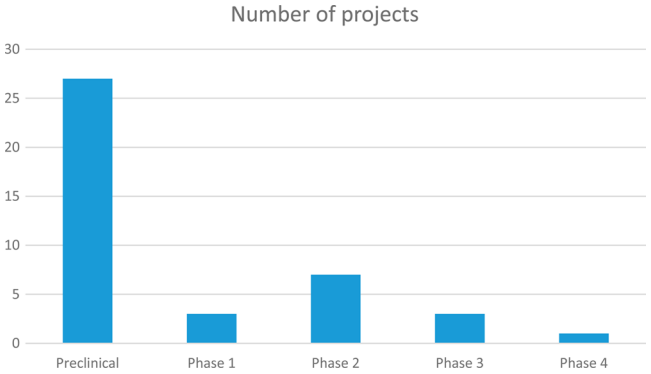


Figure 1. Global TB pipeline using data from TB Alliance and the Working Group on New TB Drugs Drug Pipeline.

latter clinical-stage compounds, several do not seem to have progressed since earlier analyses. It is incredibly concerning that we do not have more new molecules in clinical stages, especially with the prescribing limitations surrounding bedaquiline and delamanid because of their cardiovascular side effects from hERG inhibition.[6,7]

A major hurdle to progressing molecules into the clinic is identifying compounds that have activity in the mouse models of *Mtb* infection.[8] Mice have been used since the 1940s to test drug efficacy.[9] Of course, the mouse cannot completely model the complex pathology observed in humans, and its drug metabolism and pharmacokinetics also may differ. A recent questionnaire polled different TB laboratories and found that most use BALB/C, C57BL/6, or Swiss mice.[9] It was concluded that the mouse model may be most useful for rank ordering of compounds to select a drug regimen. However, some laboratories also supplement the *in vivo* mouse data with *in vivo* rabbit or marmoset studies of *Mtb* infection. We previously published a comprehensive assessment of over 70 years of literature resulting in modeling of 773 compounds reported to modulate *Mtb* infection in mice.[8] Our detailed analyses of the physiochemical and structural properties of both active and inactive molecules as well as their chemical property space coverage revealed new insights. Furthermore, we used machine learning models to correctly predict *in vivo* efficacy in the *Mtb*-infected mouse model for eight of 11 compounds. We identified gaps corresponding to the discovery and approval of new compounds,[10] as highlighted by the 40 years between the approvals of rifampicin and bedaquiline, suggesting that we can learn from earlier drug discovery. Furthermore, there were clear peaks in *in vivo* testing in the 1940s and 1950s, and it now appears that recent testing in mouse *in vivo* seems to have peaked,[10] also leading to concerns about the future health of the TB drug pipeline.

Since that report, we have collated an additional 60 molecules from more recent data published in 2014 and 2015 and further evaluated and validated the earlier models with this much larger test set that was not previously available. In addition, we have used recently published models based on *in vitro* data to create consensus models to predict *in vivo* activity. Further we describe a new clustering method for data visualization and apply it to all of the *in vivo* data gathered to date. Our goal is to continue to develop and validate various computational approaches for predicting and visualizing *in vivo* activity data in the *Mtb* mouse model, enabling the prediction of new compounds that are the most promising for advancement.

EXPERIMENTAL SECTION

Data Collection. The original data used in the initial published model were curated and quality-assessed as described previously.[8] Literature searching in the 2014−2015 time frame was performed using PubMed, and data curation was also described previously.[8] We present only new molecules that were not included in the earlier *in vivo* paper as assessed using similarity of training and test compounds based on the Tanimoto similarity distance metric[11−13] in Discovery Studio (a value of 0 represents the molecule being in the model). Distance is a generalization to continuous properties of the Tanimoto distance for binary fingerprints: $D = 1 - \sum x_{iA}x_{iB}/[\sum(x_{iA})^2 + \sum(x_{iB})^2 - \sum x_{iA}x_{iB}]$. Possible values range from 0 to 1.3333. As also described earlier, molecules were classified as active in the mouse model if they demonstrated at least 1 $\log_{10}$ reduction in colony-forming units (CFU) (or in some cases a statistically significant reduction in CFU).[8]

Test Set 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 structural data (SD) files using Discovery Studio 4.1 (San Diego, CA).[8]

Principal Component Analysis with *in Vivo* Test Set Compounds and TB Mobile Data. In order to assess the applicability domain of the 60 new *in vivo* molecules and the 784 compounds in the *in vivo Mtb* training set, we used the union of these sets to generate a principal component analysis (PCA) plot based on 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. We also compared the 60 new compounds tested in the *in vivo* mouse *Mtb* model to the previously described 805 compounds with known *Mtb* targets collated from the literature[14] and available in TB Mobile (version 2).[15] This PCA model essentially represents the published target-chemistry property space for *Mtb*.[15]

Building and Validating Machine Learning Models with Mouse *Mtb in Vivo* Data. We have previously described the generation and validation of the Laplacian-corrected naïve Bayesian classifier models developed from the mouse *Mtb* infection *in vivo* model using Discovery Studio 3.5.[16−20] We have now updated the Bayesian, tree, and support vector machine (SVM) models using Discovery Studio 4.1. In addition to the eight molecular descriptors listed in the previous section, the molecular function class fingerprints of maximum diameter 6 (FCFP_6) was added as the ninth descriptor.[21] Computational 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, the receiver operator characteristic (ROC) plots were generated, and the areas under the cross-validated ROC curves (XV ROC AUC) were calculated. Fivefold cross-validation (leave out 20% of the data set and repeat five times) was also performed, as was leave out 50% × 100-fold cross-validation. We compared the resulting Bayesian model with SVM, recursive partitioning forest (RP Forest), and RP Single Tree models built with the same set

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

Table 1. Means and Standard Deviations of Molecular Descriptors for the New *in Vivo Mtb* Dataset ($N = 60$), Comparing Actives and Inactives[a]

	MW	AlogP	HBD	HBA	Num Rings	Num Arom Rings	FPSA	RBN
active ($N = 41$)	493.88 ± 219.81	3.65 ± 3.00	2.07 ± 2.70[b]	7.22 ± 2.95	3.63 ± 0.83	2.15 ± 0.96	0.26 ± 0.08	7.10 ± 3.18
inactive ($N = 19$)	427.83 ± 72.78	3.63 ± 1.91	1.05 ± 0.97	6.21 ± 2.17	3.68 ± 1.06	2.53 ± 0.90	0.24 ± 0.10	6.47 ± 1.87

[a]MW = molecular weight; HBD = number of hydrogen-bond donors; HBA = number of hydrogen-bond acceptors; Num Rings = number of rings; Num Arom Rings = number of aromatic rings; FPSA = fractional polar surface area (sum of areas of the polar regions of the molecular surface divided by the total molecular surface area); RBN = number of rotatable bonds. [b]$p < 0.05$.

of molecular descriptors in Discovery Studio. For SVM models[22,23] we calculated interpretable descriptors in Discovery Studio and then used Pipeline Pilot to generate the FCFP_6 descriptors, followed by integration with R.[24] RP Forest[25−28] and RP Single Tree models used the standard protocol in Discovery Studio. In the case of RP Forest models, 10 trees were created with bootstrap aggregation ("bagging"). 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 total size as the original one, but a subset of the data records can be included multiple times (i.e., each tree is built with a slightly different subset of the original set, and each tree's set can contain duplicates). RP Single Tree models had a minimum of 10 samples per node and a maximum tree depth of 20. In all cases, fivefold cross-validation was used to calculate the ROC for the models generated. CDD Models (Collaborative Drug Discovery, Inc., Burlingame, CA) was also utilized to build a Bayesian model using just the open source FCFP_6 descriptors and threefold cross-validation as described previously.[29] This provides an approach for generating models that can be shared between researchers and used in mobile apps, thereby making the models more accessible.[30−33]

Mouse *Mtb* Infection Model Predictions for Compounds Identified after Model Building. From the data curation in this study, 60 compounds were identified from the literature (2014−2015) that were tested in *Mtb* infected mice (Supplemental Data 1). These were predicted with the mouse *Mtb* infection computational machine learning models previously reported as well as the updated models with 784 compounds. For each molecule, the closest distance to the training set for each model was also calculated using the "calculate molecular properties" protocol in Discovery Studio, in which a value of zero represents a molecule in the training set while larger values indicate that a molecule is more different than the training set.

***In Vivo* Activity Predictions with Previous *in Vitro*-Trained Bayesian Models.** Previously generated Bayesian models for mouse liver microsomal half-life (MLM $t_{1/2}$)[34] and dual-event models that combine *in vitro Mtb* activity and Vero cell cytotoxicity[35,36] (e.g., the Tuberculosis Antimicrobial Acquisition and Coordinating Facility (TAACF-CB2) and Molecular Libraries Small Molecule Repository (MLSMR) data sets) were used either alone or in a novel consensus workflow to predict the *in vivo Mtb* activity for the 60 compounds identified from the literature (2014−2015) that were tested in mice. The sort by two attributes features in Discovery Studio and Excel were used to organize the data, followed by tabulating the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), to enable calculation of the external statistics and enrichment factors.

Clustering of Mouse *Mtb in Vivo* Data. Honeycomb clustering (Molecular Materials Informatics, Inc., Montreal, Canada) is a greedy layout method for arranging structures on a plane in a meaningful way. A single reference compound is selected by the user as the focal point, and this is placed on a hexagonal grid pattern. For each compound, the ECFP_6 fingerprints are determined, and for all similarity comparisons, the Tanimoto coefficient is used as the metric. The six compounds most similar to the reference compound are arranged in the six available positions immediately adjacent to the focus to form the initial flower-petal starting point. The compound that is most similar to the reference compound is placed immediately above ("north"), and all possible permutations of the remaining five neighbors are considered and evaluated by summing the pairwise similarities between radially adjacent neighbors. The permutation with the highest score is used, and thus, the placement positions for the first seven compounds are fixed. The remaining compounds are ordered by decreasing similarity to the reference compound, and each is evaluated in turn. At each step the next compound is placed irreversibly: all of the unoccupied hexagons that are adjacent to at least one already-placed compound are considered and evaluated according to a score. The hexagon with the highest score is taken to be the position for this compound. The score is calculated by determining the average similarity of the compound to each of its putative neighbors. An additional "density fudge factor" of 0.01 per neighbor is added to balance out what would otherwise be a tendency to minimize the neighbor count, i.e., to prevent overfavoring long, spindly branches. For positions where there is just a single neighbor, there are three positions that may be occupied by neighbors of the neighbors, and each of these that is occupied is compared with the current compound: for the position directly opposite, the score is increased by its similarity to the current compound multiplied by 0.001, whereas for the other two positions the multiplier is 0.002. This additional term encourages the arrangement of compounds to "bend" in the direction that encourages higher similarity. This approach was used with the complete training and test set for compounds tested in the mouse *Mtb* infection model.

Statistical Analysis. Means for descriptor values for active and inactive compounds were compared by two-tailed *t* test with JMP version 8.0.1 (SAS Institute, Cary, NC). We also evaluated several additional alternative statistics for the test set, including Youden's J statistic,[37] Matthews' correlation coefficient (MCC),[38] the F_1 score,[39] and κ,[40,41] which are given by the following expressions:

$$J = \text{sensitivity} + \text{specificity} - 1$$

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

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

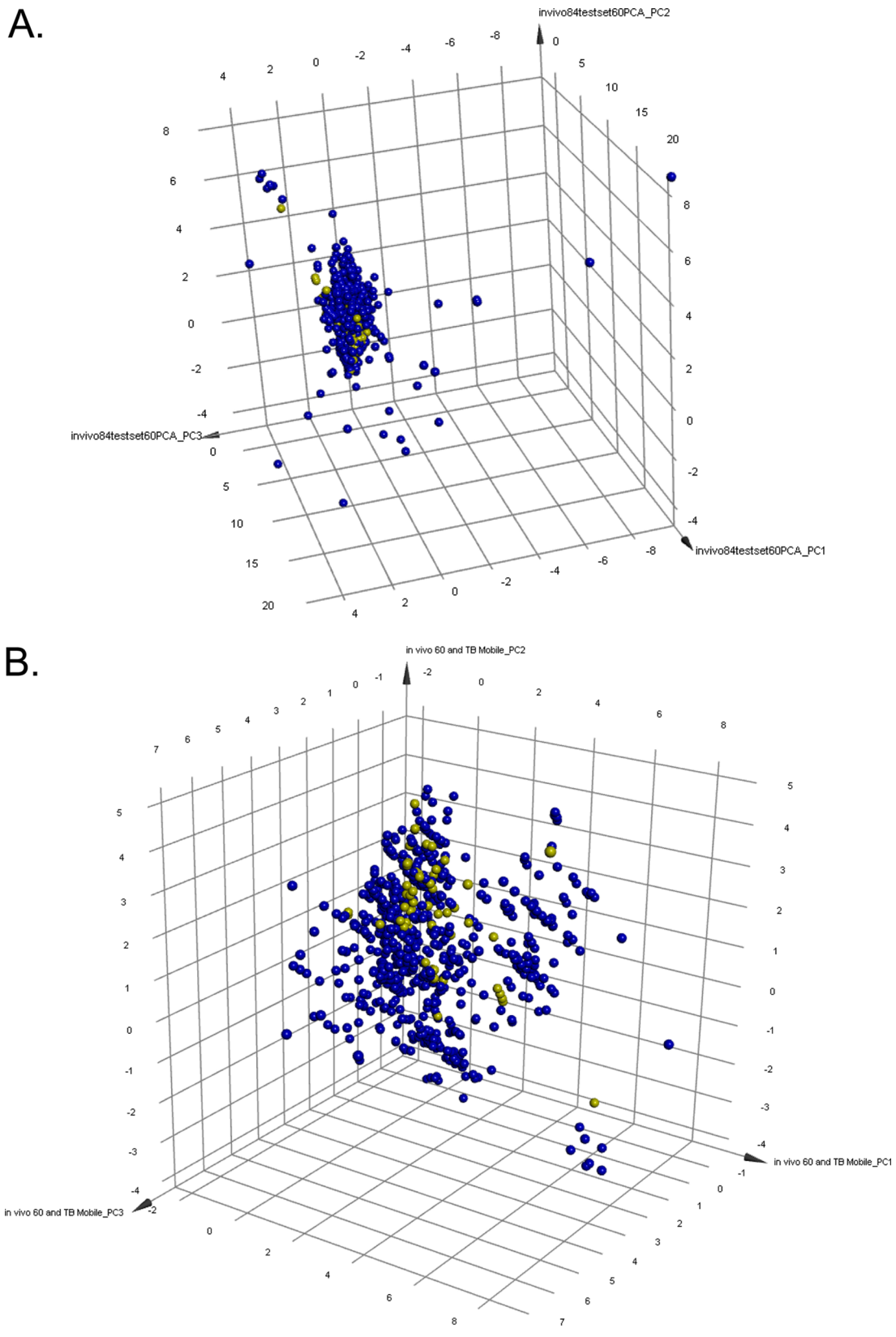


Figure 2. (A) Principal component analysis (PCA) of the updated training set for the *in vivo* model (blue) and compounds tested *in vivo* in 2014 and 2015 (yellow). Three principal components explain 86.9% of the variance. (B) PCA of TB Mobile 2 compounds ($N = 805$, blue) and compounds tested *in vivo* in 2014 and 2015 (yellow). Three principal components explain 87.5% of the variance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement, obtained using the observed data to calculate the probabilities of each

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

observer randomly saying each category. If the raters are in complete agreement, then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by p_e), $\kappa \leq 0$.

RESULTS

Molecular Property Distribution. The 60 compounds identified from the literature (2014−2015) that were tested in *Mtb*-infected mice were collated in this study (Table S1 and Supplemental Data 1) and analyzed with respect to eight simple and interpretable molecular descriptors used previously (Table 1). The only difference between active and inactive compounds was that the number of hydrogen-bond donors was statistically significantly higher for the *in vivo* active compounds.

Principal Component Analysis with *in Vivo* Data, *in Vitro* Hits, and TB Mobile Data. The PCA analysis of the complete set of 844 molecules with *in vivo* data indicated that the majority of the 60 new molecules are within the area of the training set (Figure 2A), suggesting similar chemical property coverage. PCA with the same descriptors and compounds for which we have previously collected information on targets in *Mtb* suggested that these 60 molecules tested *in vivo* in mice are in the same regions of physiochemical property space as the prior compounds (Figure 2B).

Building and Validating Machine Learning Models with Mouse *Mtb* Data. The original $N = 773$ data set used to build a Discovery Studio Bayesian model had a leave-one-out ROC of 0.77 and fivefold cross-validation value of 0.73 (Supplemental Data 2). The updated $N = 784$ Discovery Studio Bayesian model (combining the initial training set and test set[8]) for *in vivo Mtb* activity had fivefold cross-validation ROC values > 0.70 (Table 2), which were comparable to those published previously.[8] The fivefold cross-validation model ROC values (Supplemental Data 3) were comparable with the leave out 50% × 100-fold ROC values, although the concordance, specificity, and sensitivity were lower in the latter case (Table S2).

Table 2. Fivefold Cross-Validation ROC AUC Values for the Updated ($N = 784$) *in Vivo* Machine Learning Models

Bayesian[a]	SVM	Single Tree	Forest
0.733	0.77	0.72	0.74

[a]Bayesian leave-one-out cross-validation = 0.772.

Model Predictions for Additional Compounds Identified after Model Building. The 60 molecules collated in this study (Table S1 and Supplemental Data 1) were used as an external test set for the original $N = 773$ Bayesian model (Supplemental Data 1), which produced a poor ROC score (0.554). The updated $N = 784$ Bayesian model performed similarly (Supplemental Data 3) and displayed a poor ROC (0.558). The set of 60 molecules included 20 PA-824 analogues (mean closest distance = 0.26 with a standard deviation of 0.07; Table S3); all were predicted by these Bayesian models as actives (including seven *in vivo* inactives), which may reflect some bias based on similarity to active PA-824 analogues. New models including SVM (Supplemental Data 4), Single Tree and Forest models had similar ROC values (Table S4). When the sensitivity, specificity, and concordance data for the 60 molecules are compared across the different models, there are subtle differences. For example, the RP Forest model has highest sensitivity (85.4%) and concordance (66.7%). The CDD Bayesian, relying solely on the open source FCFP_6 descriptors, was similar (82.9% and 61.7% for sensitivity and concordance, respectively). The SVM model has the highest specificity (42.1%) (Table 3).

Table 3. External Statistics for the *in Vivo* TB Machine Learning Models Tested on the New *in Vivo* Mouse TB Data

machine learning model	sensitivity (%)	specificity (%)	concordance (%)
TB *in vivo* $N = 773$ Bayesian	70.7	36.8	60.0
TB *in vivo* $N = 784$ Bayesian	78.0	10.5	56.7
RP Forest TB *in vivo* $N = 784$	85.4	26.3	66.7
Best RP Tree TB *in vivo* $N = 784$	78.0	21.1	60.0
TB *in vivo* $N = 784$ SVM	68.3	42.1	60.0
TB *in vivo* $N = 784$ CDD Bayesian[a]	82.9	15.8	61.7

[a]External statistics were calculated from the results of the Bayesian modeling tool on Collaborative Drug Discovery using a cutoff score of >0.65, which produced an internal sensitivity of 0.7 and an internal specificity of 0.67.

***In Vivo* Activity Predictions with Prior *in Vitro* Bayesian Models.** Previously generated Bayesian models for MLM $t_{1/2}$, dual-event models for *Mtb in vitro* activity and Vero cell cytotoxicity, and a consensus approach were used to predict mouse *in vivo* activity (Table 4). The dual-event Bayesian models displayed poor sensitivity (<35%) and concordance (≤50%), but two had sufficient specificity (78.9%). The full $t_{1/2}$ MLM stability model with just FCFP_6 descriptors and the pruned $t_{1/2}$ MLM model with all nine descriptors produced sensitivity (78%), specificity (26.3%), and concordance (61.7%) data for the 60 molecules that were comparable to those for the $N = 784$ *in vivo* Bayesian model (Table 4). When the MLM stability Bayesian model and a dual-event *in vitro* Bayesian model agreed that a compound was either "good" or "bad," the concordance (overall accuracy) values for the *in vivo* predictions were significantly improved relative to the original dual-event model (from 43.3% to 58.8%, from 48.3% to 62.5%, and from 50.0% to 60.6%; Table 4). The confusion matrices for all of the models are also shown for clarity (Table 5).

Enrichment Factors. All of the computational models were used to calculate enrichment factors (EFs) for the test set (Table 6). The best enrichment in the hit rate (positive predicted value, or PPV) for the top-scoring 10% of compounds was 1.22, which was obtained for the *in vivo* Bayesian models, the Combined TB *in vitro* dual-event Bayesian, and the TAACF-CB2 *in vitro* dual-event Bayesian. The highest overall PPV (78.6%) was seen for the consensus model that involved using both the combined TB dual-event Bayesian and the full $t_{1/2}$ MLM stability Bayesian (just FCFP_6). The second-best PPV (77.8%) was also produced by a consensus approach (the TAACF-CB2 dual-event *in vitro* Bayesian plus the full $t_{1/2}$ MLM stability Bayesian). These two consensus approaches had better overall hit rates (and thus better overall enrichment factors) than the machine learning models that were trained with *in vivo* mouse *Mtb* data alone. Our analysis of additional statistics illustrates that some models perform better on the basis of some statistics versus others (Table 7). For example, the consensus models perform best on the basis of κ and MCC, whereas the combined dual-event Bayesian does best on the basis of Youden's J statistic and the RP Forest model performs best on the basis of the F_1 score (Table 7).

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

Table 4. External Statistics for the Dual-Event Bayesian (*in Vitro Mtb* Efficacy and Non-cytotoxicity in Vero Cells) and Mouse Liver Microsomal Stability Bayesian Models Tested on the New *in vivo* Mouse TB Data

machine learning model	sensitivity (%)	specificity (%)	concordance (%)
full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	78.0	26.3	61.7
pruned $t_{1/2}$ MLM stability Bayesian (all nine descriptors)	78.0	26.3	61.7
TAACF-CB2 dual-event Bayesian	26.8	78.9	43.3
combined TB dual-event Bayesian	34.1	78.9	48.3
MLSMR dual-event Bayesian	34.1	57.9	50.0
consensus:[a] TAACF-CB2 dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	7 true positives when predictions agree	3 true negatives when predictions agree	58.8
consensus:[a] combined TB dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	11 true positives when predictions agree	4 true negatives when predictions agree	62.5
consensus:[a] MLSMR dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	17 true positives when predictions agree	3 true negatives when predictions agree	60.6
modified consensus:[b] TAACF-CB2 dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	7 true positives when predictions agree (17.1%)	17 true negatives (89.5%)	40.0
modified consensus:[b] combined TB dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	11 true positives when predictions agree (26.8%)	16 true negatives (84.2%)	45.0
modified consensus:[b] MLSMR dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	17 true positives when predictions agree (41.5%)	13 true negatives (68.4%)	50.0

[a]For the initial consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive or false positive. Similarly, both models had to classify a compound as bad/inactive for it to be considered as a true negative or false negative. Since the combination of the models agreed on the classification only for a subset of the test set, the overall sensitivity and overall specificity are not applicable. However, the overall concordance is still relevant and was calculated as (number of true positives + number of true negatives)/(number of compounds on which both models agreed on the good or bad classification). [b]For the modified consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive. However, if either model classified a compound as bad/inactive, it was defined as a true negative or false negative (depending on its experimental value). Thus, the modified consensus approaches made predictions for all of the test compounds.

Clustering of Mouse *Mtb* in Vivo Data. The new honeycomb clustering approach (Figure 3A; an enlarged version is shown in Figure S1) provides a map of the *in vivo* active and inactive compounds according to structural similarity. The 60 additional test set molecules cluster near similar molecules, such as the macrolides (Figure 3B). This approach can also be used to infer activity on the basis of similarity:[11−13] if a compound is surrounded by other active molecules, this might suggest that it too is active. For example, the macrolides cyclogriselimycin and ecumicin in the test set are surrounded by other active compounds (Figure 3B).

DISCUSSION

An abundance of data from large *in vitro* phenotypic screens against *Mtb* exists in the public domain[42−46] that can readily be used to assist future drug discovery. Since machine learning methods can learn from past data, we have extensively applied Bayesian and other machine learning algorithms to model *Mtb* inhibition and Vero cell cytotoxicity. We have pioneered the use of dual-event data sets, which use both dose−response data for whole-cell antitubercular activity[47,48] and Vero cell cytotoxicity.[35,36,42,49−51] We have also used the same approach to model mouse *Mtb in vivo* data[8] and most recently MLM stability.[34] In addition, we have recently developed a freely available mobile app called TB Mobile[15] that displays over 800 *Mtb*-active molecule structures and their targets, with links to associated data. This tool was recently enhanced by adding target-specific Bayesian models to rank probable targets.[15] Such Bayesian modeling approaches with FCFP_6 fingerprints have also been integrated into the CDD Vault software,[29] and the algorithms were made open source and applied to large numbers of data sets.[31] Our combined efforts in this area indicate that such machine learning models are a valuable resource and can be used prospectively to suggest molecules to test not only against *Mtb* but also for other diseases.[52,53]

Since our earlier study compiling compounds tested in the mouse *Mtb* model,[8] we have continued to collect and curate data with the aim of testing the machine learning models developed and improving upon them. We have now added to our database 60 unique molecules published in 2014−2015 (Table S1 and Supplemental Data 1), of which 41 were classified as actives. The large percentage of active compounds (~60%) in this new set may represent a publication bias, as noted in our prior analysis for data up to 2014.[8] Simple molecular property analysis suggested that only the number of hydrogen-bond donors was significantly higher for the 41 *in vivo* active compounds (Table 1). These 60 molecules broadly cover a similar physiochemical property space as the training set of 784 molecules (Figure 2A) and the over 800 molecules collated in TB Mobile[15] (Figure 2B). These results suggest that we are likely in the applicability domain of the data set. The closest similarity values calculated with the N = 784 Bayesian model (mean closest distance = 0.40 with a standard deviation of 0.15, where a value of zero connotes identity and larger values indicate greater difference; Figure S2) are quite low, suggesting that most are within the applicability domain of the model.

In this study, we have updated the machine learning models and also evaluated whether models for other properties, such as the combined *in vitro* bioactivity and Vero cell cytotoxicity, MLM $t_{1/2}$, or a consensus of these models, could also predict compounds likely to have *in vivo* activity. When comparing different machine learning approaches such as Bayesian, SVM, and recursive partitioning, we observed little difference based on internal fivefold ROC (Table 2), although predictions for the external test set produced some slight differences in sensitivity, specificity, and concordance (Table 3). These show that the sensitivity values are generally much higher than the specificity

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

Table 5. Confusion Matrices Produced When the Machine Learning Models Were Tested on the New *in Vivo* Mouse TB Data

Legend	
true positives	false positives
false negatives	true negatives
TB *in Vivo N* = 773 Bayesian	
29	12
12	7
TB *in Vivo N* = 784 Bayesian	
32	17
9	2
RP Forest TB *in Vivo N* = 784	
35	14
6	5
Best RP Tree TB *in Vivo N* = 784	
32	15
9	4
TB *in Vivo N* = 784 SVM	
28	11
13	8
TB *in Vivo N* = 784 CDD Bayesian[a]	
34	16
7	3
Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)	
32	14
9	5
Pruned $t_{1/2}$ MLM Stability Bayesian (All Nine Descriptors)	
32	14
9	5
TAACF-CB2 Dual-Event Bayesian	
11	4
30	15
Combined TB Dual-Event Bayesian	
14	4
27	15
MLSMR Dual-Event Bayesian	
19	8
22	11
Consensus:[b] TAACF-CB2 Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[d]	
7	2
5	3
Consensus:[b] Combined TB Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[e]	
11	3
6	4
Consensus:[b] MLSMR Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[f]	
17	6
7	3
Modified Consensus:[c] TAACF-CB2 Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[g]	
7	2
34	17
Modified Consensus:[c] Combined TB Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[g]	
11	3
30	16
Modified Consensus:[c] MLSMR Dual-Event Bayesian + Full $t_{1/2}$ MLM Stability Bayesian (Just FCFP_6)[g]	
17	6
24	13

[a]External statistics were calculated using the Bayesian modeling tool on Collaborative Drug Discovery with a cutoff score of >0.65, which produced an internal sensitivity of 0.7 and an internal specificity of 0.67. [b]For the consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive or false positive (depending on the experimental value of the compound). Similarly, both models had to classify a compound as bad/inactive for it to be considered as a true negative or false negative. [c]For the modified consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive. However, if either model classified a compound as bad/inactive, it was defined as a true negative or false negative (depending on its experimental value). [d]Coverage = 17/60 = 28%. [e]Coverage = 24/60 = 40%. [f]Coverage = 33/60 = 55%. [g]Coverage = 60/60 = 100%.

values, which is a reversal of what we observed for the fivefold ROC for the Bayesian model training sets (Supplemental Data 1 and 2). This likely suggests that we would have some difficulty classifying inactives with these models while being able to select actives. The updated training set is fairly well balanced, as it was before.[8] Obviously the use of a computational approach to select compounds presents advantages in likely reducing follow-up costs and lowering numbers of mice used. The $t_{1/2}$ MLM stability models follow the same trend with the sensitivity being much higher than the specificity (Table 4), while the *in vitro Mtb* bioactivity TAACF-CB2 dual-event, combined TB dual-event, and MLSMR dual-event models have higher specificity than sensitivity for the test set molecules (Table 4).

Surprisingly, we found that Bayesian models built with *in vitro* MLM $t_{1/2}$ data sets could produce external statistics similar to those for the *in vivo Mtb* models (Table 4), as dual-event models that included *in vitro Mtb* bioactivity and cytotoxicity had positive predicted values of >70% and the best overall hit rates were obtained with a consensus of an *Mtb in vitro* model and the MLM $t_{1/2}$ model (Table 6). *In vivo* models and *in vitro* models in the best cases individually had enrichment factors of 1.22 for the top-scoring 10% of compounds, and the combined TB dual-event *in vitro* Bayesian achieved the best enrichment factor of 1.32 for the top-scoring 10 compounds. Because of the large percentage of active compounds in this new *in vivo* test set (60%), the maximum enrichment factor that a "perfect" model could produce was 1.46. On the basis of these results (Table 6), it seems that the better PPV hit rates and enrichment factors for two of the consensus approaches are due to (a) superior specificity (filtering out compounds likely to be MLM-unstable, *Mtb*-inactive, and/or cytotoxic gave ~2 times the best specificity of an *in vivo*-trained model) while (b) not having high false positive rates for the consensus approaches. The use of additional external statistics suggested that no single model performed best across all (Table 7). Therefore, these may not be as useful as considering the enrichment factors for test set evaluation (Table 6).

Perhaps part of the accuracy of the MLM Bayesian for predicting *in vivo* activity is based on the fact that at least 190 of the 894 compounds in the full $t_{1/2}$ MLM Bayesian training set were from *Mtb* and malaria projects (i.e., 20 stable compounds out of 42 were from *Mtb* projects and 49 stable compounds out of 148 were from malaria projects). Antimalarial compounds sometimes display activity against *Mtb*,[36,54] and researchers are

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

Table 6. External Enrichment Factors in Hit Rates for the Machine Learning Models Tested on the New *in Vivo* Mouse TB Data

machine learning model	overall hit rate (PPV)[a,c]	enrichment factors[b,c]		
		for top 10%	for top 10 compounds	for top 20%
TB *in vivo* N = 773 Bayesian	29/41 (70.7%)	1.22	1.17	0.98
TB *in vivo* N = 784 Bayesian	32/49 (65.3%)	1.22	1.17	0.98
RP Forest TB *in vivo* N = 784	35/49 (71.4%)	0.98	1.02	0.98
TB *in vivo* N = 784 SVM	28/39 (71.8%)	N/A	N/A	N/A
TB *in vivo* N = 784 CDD Bayesian[d]	34/50 (68.0%)	1.22	1.02	1.10
full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	32/46 (69.6%)	0.98	0.88	0.98
pruned $t_{1/2}$ MLM stability Bayesian (all nine descriptors)	32/46 (69.6%)	0.73	0.88	0.98
TAACF-CB2 dual-event Bayesian	11/15 (73.3%)	1.22	1.17	1.10
combined TB dual-event Bayesian	14/18 (77.8%)	1.22	1.32	1.22
MLSMR dual-event Bayesian	19/27 (70.4%)	0.73	0.88	0.98
consensus: TAACF-CB2 dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	7/9 (77.8%); overall EF = 1.14	N/A	N/A	N/A
consensus: combined TB dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	11/14 (78.6%); overall EF = 1.15	N/A	N/A	N/A
consensus: MLSMR dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	17/23 (73.9%); overall EF = 1.08	N/A	N/A	N/A

[a]The hit rate (positive predicted value = PPV) was calculated as (number of true positives)/(number of true positives + number of false positives). [b]The enrichment factor was calculated as (hit rate in %)/(% of *in vivo* active compounds in the external test set). Since 41 of the 60 compounds in this external test set (68.3%) were active, the maximum enrichment factor that a perfect model could achieve would be 100%/68.3% = 1.46. [c]Since each original consensus model and the corresponding "modified consensus" model have the same number of true positives and false positives, their hit rates and enrichment factors are equivalent. [d]External statistics were calculated from the results of the Bayesian modeling tool on Collaborative Drug Discovery using a cutoff score of >0.65, which produced an internal sensitivity of 0.7 and an internal specificity of 0.67.

Table 7. Additional External Statistics for the Machine Learning Models Tested on the New *in Vivo* Mouse TB Data[a]

machine learning model	κ	MCC	J	F_1
TB *in vivo* N = 773 Bayesian	0.08	0.08	0.08	0.71
TB *in vivo* N = 784 Bayesian	−0.13	−0.14	−0.11	0.71
RP Forest TB *in vivo* N = 784	0.13	0.14	**0.12**	**0.78**
TB *in vivo* N = 784 SVM	0.10	0.10	0.10	0.70
TB *in vivo* N = 784 CDD Bayesian	−0.01	−0.02	−0.01	**0.75**
full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	0.05	0.05	0.04	0.74
pruned $t_{1/2}$ MLM stability Bayesian (all nine descriptors)	0.05	0.05	0.04	0.74
TAACF-CB2 dual-event Bayesian	0.04	0.06	0.06	0.39
combined TB dual-event Bayesian	0.10	0.13	**0.13**	0.47
MLSMR dual-event Bayesian	0.04	0.04	0.04	0.56
consensus:[b] TAACF-CB2 dual-event Bayesian + full$t_{1/2}$ MLM stability Bayesian (just FCFP_6)	0.13	**0.17**	N/A	0.67
consensus:[b] combined TB dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	**0.18**	**0.20**	N/A	0.71
consensus:[b] MLSMR dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	**0.19**	0.04	N/A	0.72
modified consensus:[c] TAACF-CB2 dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	0.05	0.09	0.07	0.28
modified consensus:[c] combined TB dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	0.08	0.12	0.11	0.40
modified consensus:[c] MLSMR dual-event Bayesian + full $t_{1/2}$ MLM stability Bayesian (just FCFP_6)	0.08	0.09	0.10	0.53

[a]The top two scores for each particular type of external statistic are shown in bold. [b]For the initial consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive or false positive. Similarly, both models had to classify a compound as bad/inactive for it to be considered as a true negative or false negative. Consequently, these workflows made active/inactive classifications on only a subset of the test set. [c]For the modified consensus approaches, both types of Bayesian models had to classify a compound as good/active for it to be considered as a true positive. However, if either model classified a compound as bad/inactive, it was defined as a true negative or false negative (depending on its experimental value). Thus, the modified consensus approaches made predictions for all of the test compounds.

unlikely to devote the time and money to MLM stability assays unless a compound displays promising therapeutic activity. Thus, perhaps some of the chemical features and properties that are deemed favorable by the MLM stability Bayesian implicitly incorporate both stability and efficacy. Greatly simplified, it is reasonable to correlate *in vivo* efficacy in the mouse with compounds that *at minimum* are metabolically stable with regard to phase I metabolism in the mouse and inhibit the growth of *Mtb* under *in vitro* conditions mimicking aspects of their actual pathologic environs. Support for this notion is provided by the fact that when evaluating an external test set of known antitubercular molecules, the MLM stability Bayesian recognized most TB drugs as metabolically stable.[34]

It is interesting that although the TAACF-CB2 and combined TB dual-event Bayesian models had poor overall external sensitivity and concordance values, they displayed enrichment factors for the top-scoring 10% of compounds that were equivalent or superior to the *in vivo* models. When these two *in vitro Mtb* dual-event models were combined with the $t_{1/2}$ MLM stability model to make a novel consensus workflow, the best overall hit rates (and thus the best overall enrichment factors) were achieved (Table 6). This highlights the importance of

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

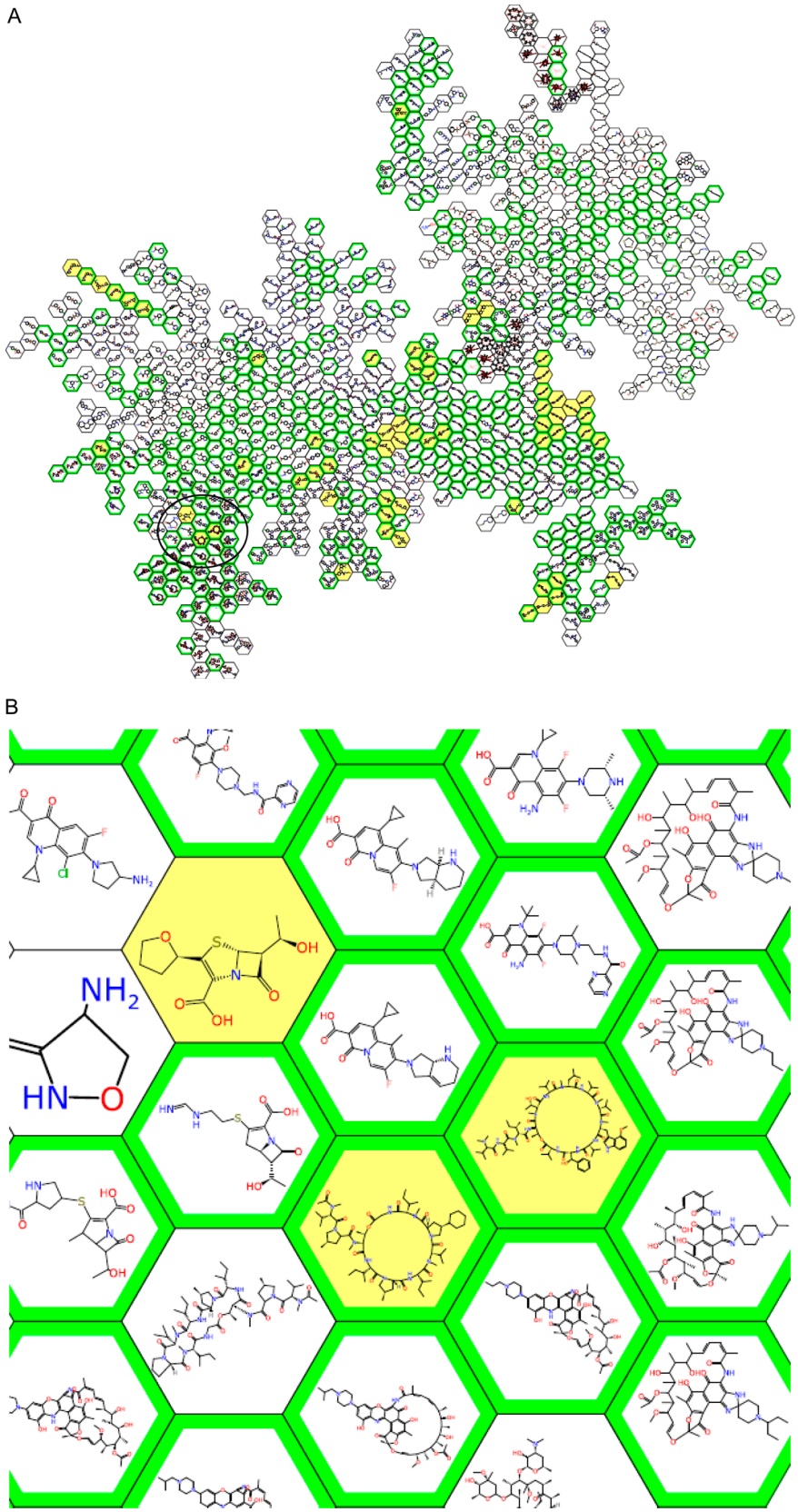


Figure 3. Honeycomb clustering of TB *in vivo* data from 2014 and 2015. Yellow hexagons highlight the compounds from 2014 and 2015, and green outlines signify *in vivo* active compounds. (A) Complete map of compounds in the training and testing sets. (B) Enlarged view of the section marked with the black circle in (A), highlighting cyclogriselimycin and ecumicin.

specificity and of studying more than just the overall external statistics, especially since most laboratories will only be able to perform *in vivo Mtb* studies in mice for a very small number of candidate compounds. In summary, this suggests that together

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

the experimental *in vitro Mtb* and $t_{1/2}$ MLM stability models are a good predictor of *in vivo Mtb* efficacy in the mouse model and that a consensus machine learning model may be a useful alternative, at least on the basis of this particular test set as described.

While humans can visualize quite complex data, there have been many approaches to data visualization to produce maps that can explain the terrain of biological data, including Kohonen networks, Sammon maps, and many other approaches.[55] In this study we have described a new data visualization approach called honeycomb visualization that was used to cluster the *in vivo* training set and the new external test set (Figures 3 and S1). This shows, for example, how macrocyclic compounds cluster together (Figure 3B) and how the 60 new compounds are dispersed around the map with local clusters of similar analogues. Such an approach, while also relying on the same descriptors, illustrates an alternative way to assess predictions or structure−activity relationships and has been implemented recently in a new iOS free mobile app called PolyPharma[56] to demonstrate how machine learning and this visualization can be used outside of the desktop. In addition, the use of tools such as CDD Models can enable the sharing of Bayesian models such that they can be run in freely available mobile apps, making the models more accessible.[30]

In summary, this work adds further weight to the use of machine learning approaches to predict *in vivo Mtb* activity in the mouse efficacy model. We have not seen this kind of approach taken with other disease models and data sets, so this still ranks as a difficult but interesting problem to address. Our previous reported *Mtb in vivo* model[8] was tested with only a small test set of 11 molecules, while we now report one that has 60 molecules. For the first time we have shown that MLM stability predictions using a Bayesian model could be a useful adjunct to applying a specific *Mtb in vivo* mouse machine learning model to predict efficacy. This is potentially of importance because it is likely that MLM stability models are based on more structurally diverse molecules than just antitubercular efficacy models. Having more modeling options for predicting *in vivo* activity is of value because of the relatively small data set of *in vivo Mtb* values publicly available at this time. The *in vitro* and cytotoxicity models' training sets are larger, and these with the MLM $t_{1/2}$ models can be additionally used for predicting *in vivo* activity based on our test set of 60 recently published molecules. There is no shortage of *in vitro* screening hits (there are likely many thousands across the various public−private partnership, NIH-funded, and commercial screens), and we would propose that computational models such as those described in this study be made available, shared,[29] and utilized alongside other selection criteria (medicinal chemistry heuristics or gut feeling) prior to selecting compounds for testing in the animal model in order to expedite TB research and save both time and money. Along these lines, the 177 GSK open source *Mtb* leads were scored with the top two consensus dual-event and MLM Bayesian workflows, and the intersection of their top predictions was analyzed in order to suggest compounds that should be prioritized for *in vivo* assays (Tables S5−S7). In due course we will use our models to make prospective predictions prior to *in vivo* testing in the mouse *Mtb* model in our own laboratories, and these results will be reported in the future.

In conclusion, we have built on our previous *Mtb in vivo* and *in vitro* modeling studies[8,35,36,47,48,50,51,57,58] to suggest a combination of published data and machine learning models that can be used to harness limited research resources and increasingly contracting funding for TB research. With continual pressure to identify novel *in vivo* active antituberculars to supplement the currently depleted clinical pipeline, these machine learning models could be considered.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00004.

Supplemental Data 2−4, Tables S1−S7, and Figures S1 and S2 (PDF)
Enlarged Figure S1 (PDF)
Supplemental Data 1 (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ekinssean@yahoo.com. Phone 215-687-1320.

Author Contributions

▽S.E. and A.L.P contributed equally.

Notes

The authors declare the following competing financial interest(s): S.E., A.M.C., and R.C.R. are consultants for Collaborative Drug Discovery, Inc.

All of the computational models are available from the authors upon request. The structures of the molecules used in this study are available in CDD (as described previously) or in previous papers.

■ ACKNOWLEDGMENTS

S.E. acknowledges colleagues at CDD. Biovia is kindly acknowledged for providing Discovery Studio and Pipeline Pilot to S.E. and J.S.F. This work was supported by Award 9R44TR000942-02 "Biocomputation across distributed private datasets to enhance drug discovery" from the NIH National Center for Advancing Translational Sciences. S.E. acknowledges that TB Mobile and the associated dataset used in this study were developed partially with funding from Award 2R42AI088893-02 "Identification of novel therapeutics for tuberculosis combining cheminformatics, diverse databases and logic based pathway analysis" from the National Institute of Allergy and Infectious Diseases. The work was partially supported by a grant from the European Community's Seventh Framework Programme (Grant 260872, MM4TB Consortium) to S.E. We acknowledge Dr. Iain Old and Dr. Giovanna Riccardi for stimulating some of the ideas discussed on funding. J.S.F. acknowledges funding from NIH/NIAID (2R42AI088893-02) and Rutgers University−NJMS. J.S.F., S.E., and A.L.P. were supported by funding from NIH/NIAID (1U19AI109713) for the "Center to develop therapeutic countermeasures to high-threat bacterial agents," from the NIH Centers of Excellence for Translational Research (CETR).

■ ABBREVIATIONS USED:

CFU, colony-forming units; FCFP_6, molecular function class fingerprints of maximum diameter 6; hERG, human ether-a-go-go related channel; *Mtb*, *Mycobacterium tuberculosis*; NIAID, National Institute of Allergy and Infectious Diseases; NIH, National Institutes of Health; PCA, principal component analysis; PPV, positive predicted value; RP, recursive partitioning; SVM, support vector machine; TB, tuberculosis; XV ROC AUC, area under the cross-validated receiver operator characteristic curve

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

REFERENCES

(1) World Health Organization. Global Tuberculosis Report 2014. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1 (accessed Jan 4, 2016).

(2) Lonnroth, K.; Migliori, G. B.; Abubakar, I.; D'Ambrosio, L.; de Vries, G.; Diel, R.; Douglas, P.; Falzon, D.; Gaudreau, M. A.; Goletti, D.; et al. Towards Tuberculosis Elimination: An Action Framework for Low-Incidence Countries. *Eur. Respir. J.* **2015**, *45*, 928−952.

(3) Jakab, Z.; Acosta, C. D.; Kluge, H. H.; Dara, M. Consolidated Action Plan to Prevent and Combat Multidrug- and Extensively Drug-Resistant Tuberculosis in the Who European Region 2011−2015: Cost-Effectiveness Analysis. *Tuberculosis (Oxford, U. K.)* **2015**, *95* (Suppl. 1), S212−S216.

(4) Tb Alliance. Our Pipeline. http://www.tballiance.org/pipeline/pipeline.php (accessed Jan 4, 2016).

(5) Working Group on New TB Drugs. Drug Pipeline. http://www.newtbdrugs.org/pipeline.php (accessed Jan 4, 2016).

(6) Esposito, S.; Bianchini, S.; Blasi, F. Bedaquiline and Delamanid in Tuberculosis. *Expert Opin. Pharmacother.* **2015**, *16*, 2319−2330.

(7) Chahine, E. B.; Karaoui, L. R.; Mansour, H. Bedaquiline: A Novel Diarylquinoline for Multidrug-Resistant Tuberculosis. *Ann. Pharmacother.* **2014**, *48*, 107−115.

(8) Ekins, S.; Pottorf, R.; Reynolds, R. C.; Williams, A. J.; Clark, A. M.; Freundlich, J. S. Looking Back to the Future: Predicting in Vivo Efficacy of Small Molecules Versus Mycobacterium Tuberculosis. *J. Chem. Inf. Model.* **2014**, *54*, 1070−1082.

(9) 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.; et al. Comprehensive Analysis of Methods Used for the Evaluation of Compounds against Mycobacterium Tuberculosis. *Tuberculosis (Oxford, U. K.)* **2012**, *92*, 453−488.

(10) Ekins, S.; Nuermberger, E. L.; Freundlich, J. S. Minding the Gaps in Tuberculosis Research. *Drug Discovery Today* **2014**, *19*, 1279−1282.

(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf Comp Sci.* **1998**, *38*, 983−996.

(12) Willett, P. Similarity-Based Approaches to Virtual Screening. *Biochem. Soc. Trans.* **2003**, *31*, 603−606.

(13) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(14) 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.

(15) Clark, A. M.; Sarker, M.; Ekins, S. New Target Predictions and Visualization Tools Incorporating Open Source Molecular Fingerprints for Tb Mobile 2.0. *J. Cheminf.* **2014**, *6*, 38.

(16) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian Models for the Prioritization of Antitubercular Agents. *J. Chem. Inf. Model.* **2008**, *48*, 2362−2370.

(17) 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*, 861−873.

(18) 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*, 1945−1956.

(19) Hassan, M.; Brown, R. D.; Varma-O'brien, S.; Rogers, D. Cheminformatics Analysis and Learning in a Data Pipelining Environment. *Mol. Diversity* **2006**, *10*, 283−299.

(20) 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*, 682−686.

(21) 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*, 1466−1475.

(22) Chang, C. C.; Lin, C. J. *LIBSVM: A Library for Support Vector Machines*, 2001.

(23) Christianini, N.; Shawe-Taylor, J. *Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, U.K., 2000.

(24) http://www.r-project.org/ (accessed Jan 4, 2016).

(25) Hawkins, D. M.; Young, S. S.; Rusinko, A. I. Analysis of Large Structure Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.

(26) Therneau, T. M.; Atkinson, E. J. *An Introduction to Recursive Partitioning Using the Rpart Routines*; Department of Health Sciences Research, Mayo Clinic: Rochester, MN, 1997.

(27) Chen, X.; Rusinko, A., III; Young, S. S. Recursive Partitioning Analysis of a Large Structure-Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(28) Rusinko, A., 3rd; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(29) Clark, A. M.; Dole, K.; Coulon-Spektor, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S. Open Source Bayesian Models: 1. Application to Adme/Tox and Drug Discovery Datasets. *J. Chem. Inf. Model.* **2015**, *55*, 1231−1245.

(30) Ekins, S.; Clark, A. M.; Wright, S. H. Making Transporter Models for Drug-Drug Interaction Prediction Mobile. *Drug Metab. Dispos.* **2015**, *43*, 1642−1645.

(31) Clark, A. M.; Ekins, S. Open Source Bayesian Models: 2. Mining a "Big Dataset" to Create and Validate Models with Chembl. *J. Chem. Inf. Model.* **2015**, *55*, 1246−1260.

(32) Clark, A. M.; Dole, K.; Coulon-Spektor, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S. Open Source Bayesian Models. 1. Application to Adme/Tox and Drug Discovery Datasets. *J. Chem. Inf. Model.* **2015**, *55*, 1231−1245.

(33) Clark, A. M.; Dole, K.; Ekins, S. Open Source Bayesian Models. 3. Composite Models for Prediction of Binned Responses. *J. Chem. Inf. Model.* **2016**, *56*, 275−285.

(34) Perryman, A. L.; Stratton, T. P.; Ekins, S.; Freundlich, J. S. Predicting Mouse Liver Microsomal Stability with "Pruned' Machine Learning Models and Public Data. *Pharm. Res.* **2016**, *33*, 433−449.

(35) 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.

(36) 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.; et al. Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery. *Chem. Biol.* **2013**, *20*, 370−378.

(37) Youden, W. J. Index for Rating Diagnostic Tests. *Cancer* **1950**, *3*, 32−35.

(38) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

(39) Van Rijsbergen, C. J. *Information Retrieval*, 2nd ed.; Butterworth: London, 1979.

(40) Carletta, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Comput. Linguist.* **1996**, *22*, 249−254.

(41) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37−46.

(42) Ekins, S.; Freundlich, J. S.; Hobrath, J. V.; Lucile White, E.; Reynolds, R. C. Combining Computational Methods for Hit to Lead Optimization in Mycobacterium Tuberculosis Drug Discovery. *Pharm. Res.* **2014**, *31*, 414−435.

(43) 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.; et al. High Throughput Screening of a Library Based on

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343

Kinase Inhibitor Scaffolds against Mycobacterium Tuberculosis H37rv. *Tuberculosis (Oxford, U. K.)* **2012**, *92*, 72−83.
(44) 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.; et al. Antituberculosis Activity of the Molecular Libraries Screening Center Network Library. *Tuberculosis (Oxford, U. K.)* **2009**, *89*, 354−363.
(45) 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.; et al. High-Throughput Screening for Inhibitors of Mycobacterium Tuberculosis H37rv. *Tuberculosis (Oxford, U. K.)* **2009**, *89*, 334−353.
(46) Ballell, L.; Bates, R. H.; Young, R. J.; Alvarez-Gomez, D.; Alvarez-Ruiz, E.; Barroso, V.; Blanco, D.; Crespo, B.; Escribano, J.; Gonzalez, R.; et al. Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. *ChemMedChem* **2013**, *8*, 313−321.
(47) Ekins, S.; Kaneko, T.; Lipinski, C. A.; Bradford, J.; Dole, K.; Spektor, A.; Gregory, K.; Blondeau, D.; Ernst, S.; Yang, J.; et al. Analysis and Hit Filtering of a Very Large Library of Compounds Screened against Mycobacterium Tuberculosis. *Mol. BioSyst.* **2010**, *6*, 2316−2324.
(48) 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.
(49) 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 (Oxford, U. K.)* **2014**, *94*, 162−169.
(50) 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.
(51) Ekins, S.; Freundlich, J. S.; Reynolds, R. C. Are Bigger Data Sets Better for Machine Learning? Fusing Single-Point and Dual-Event Dose Response Data for Mycobacterium Tuberculosis. *J. Chem. Inf. Model.* **2014**, *54*, 2157−2165.
(52) Ekins, S.; Lage de Siqueira-Neto, J.; McCall, L.-I.; Sarker, M.; Yadav, M.; Ponder, E. L.; Kallel, E. A.; Kellar, D.; Chen, S.; Arkin, M.; et al. Machine Learning Models and Pathway Genome Data Base for Trypanosoma Cruzi Drug Discovery. *PLoS Neglected Trop. Dis.* **2015**, *9*, e0003878.
(53) Ekins, S.; Freundlich, J.; Clark, A.; Anantpadma, M.; Davey, R.; Madrid, P. Machine Learning Models Identify Molecules Active against Ebola Virus in Vitro. *F1000Research* **2015**, *4*, 1091.
(54) Vilcheze, C.; Baughn, A. D.; Tufariello, J.; Leung, L. W.; Kuo, M.; Basler, C. F.; Alland, D.; Sacchettini, J. C.; Freundlich, J. S.; Jacobs, W. R., Jr. Novel Inhibitors of Inha Efficiently Kill Mycobacterium Tuberculosis under Aerobic and Anaerobic Conditions. *Antimicrob. Agents Chemother.* **2011**, *55*, 3889−3898.
(55) Ivanenkov, Y. A.; Savchuk, N. P.; Ekins, S.; Balakin, K. V. Computational Mapping Tools for Drug Discovery. *Drug Discovery Today* **2009**, *14*, 767−775.
(56) Clark, A. M. PolyPharma. https://itunes.apple.com/us/app/polypharma/id1025327772?mt=8 (accessed Jan 4, 2016).
(57) Perryman, A. L.; Stratton, T. P.; Ekins, S.; Freundlich, J. S. Predicting Mouse Liver Microsomal Stability with "Pruned" Machine Learning Models and Public Data. *Pharm. Res.* **2016**, *33*, 433−449.
(58) Ekins, S.; Freundlich, J. S. Validating New Tuberculosis Computational Models with Public Whole Cell Screening Aerobic Activity Datasets. *Pharm. Res.* **2011**, *28*, 1859−1869.

DOI: 10.1021/acs.jcim.6b00004
J. Chem. Inf. Model. 2016, 56, 1332−1343