Published in final edited form as:

Pharm Res. 2016 February ; 33(2): 433–449. doi:10.1007/s11095-015-1800-5.

Predicting Mouse Liver Microsomal Stability with “Pruned” Machine Learning Models and Public Data

Alexander L. Perryman^a, Thomas P. Stratton^b, Sean Ekins^{c,d}, and Joel S. Freundlich^{a,b,*}

^aDivision of Infectious Disease, 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

^bDepartment of Pharmacology & Physiology, Rutgers University-New Jersey Medical School, Newark, New Jersey 07103, United States

^cCollaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, North Carolina 27526, United States

^dCollaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, California 94010, United States

Abstract

Purpose—Mouse efficacy studies are a critical hurdle to advance translational research of potential therapeutic compounds for many diseases. Although mouse liver microsomal (MLM) stability studies are not a perfect surrogate for *in vivo* studies of metabolic clearance, they are the initial model system used to assess metabolic stability. Consequently, we explored the development of machine learning models that can enhance the probability of identifying compounds possessing MLM stability.

Methods—Published assays on MLM half-life values were identified in PubChem, reformatted, and curated to create a training set with 894 unique small molecules. These data were used to construct machine learning models assessed with internal cross-validation, external tests with a published set of antitubercular compounds, and independent validation with an additional diverse set of 571 compounds (PubChem data on percent metabolism).

Corresponding author: Joel S. Freundlich, Ph.D., Departments of Pharmacology & Physiology and Medicine, The Ruy V. Lourenço Center for Emerging and Re-emerging Pathogens, Rutgers University-New Jersey Medical School, Medical Sciences Building, I-503, 185 South Orange Ave., Newark, NJ 07103, Phone: 973-972-7165; Fax: 973-972-1141; freundjs@rutgers.edu.

Supplementary Material Available: Supplemental information consists of 14 figures (*i.e.*, a workflow describing the percent compound left set, many internal and external ROC curves, and PCA plots), 8 tables and commentary (describing internal and external results of SVM and RP-Random Forest models, as well as additional half-life Bayesians that were created using different types of 2D topological fingerprints and numbers of bins), and the sdf files of the curated, full and pruned versions of the MLM half-life and percent compound left datasets. These curated sdf files contain the PubChem CID numbers, structural information, MLM stability data, qualifiers/notes (such as < or > and comments on the details of certain assay methods or the series of compounds of which that molecule is a member), our binary classification (1 = stable and 0 = unstable), and the AID reference numbers that cite the source of the assay results on PubChem for each and every compound utilized. This material is available free of charge via the Internet at <http://link.springer.com/journal/11095/>.

Conflicts of Interest

S.E. is a consultant for Collaborative Drug Discovery Inc.

Results—“Pruning” out the moderately unstable/moderately stable compounds from the training set produced models with superior predictive power. Bayesian models displayed the best predictive power for identifying compounds with a half-life > 1 hour.

Conclusions—Our results suggest the pruning strategy may be of general benefit to improve test set enrichment and provide machine learning models with enhanced predictive value for the MLM stability of small organic molecules. This study represents the most exhaustive study to date of using machine learning approaches with MLM data from public sources.

Keywords

Bayesian model; machine learning; metabolic stability; mouse liver microsomal stability; translational research

Introduction

Efficacy studies in mice are a critical hurdle to advance translational research of therapeutic compounds for many diseases, including tuberculosis,[1–4] malaria,[5,6] cancer,[7–11] Alzheimer’s disease,[12] and many other therapeutic areas[13–15] before clinical trials in humans are considered. A compound of interest for mouse *in vivo* studies should ideally be soluble, have good absorption (if orally dosed), and must also exhibit metabolic stability. A sufficient oral dosage must survive first-pass clearance through the liver before reaching the biological target(s). The majority of drugs as well as other xenobiotics undergo phase I metabolism, mediated by the cytochrome P450 (P450) family of heme-thiolate enzymes predominantly in the liver as well as extrahepatically,[16] which causes most drug-drug interactions.[17,18] These enzymes can either inactivate or activate molecules, which may be followed by secondary or tertiary metabolism.[19] While mouse liver microsomal (MLM) stability studies are not a perfect surrogate for mouse *in vivo* studies of metabolic clearance, they are a widely used initial cell-based model system that can correlate well with human liver microsomal (HLM) stability[20] and *in vivo* activity in mice.[21]

Predicting metabolism using computational methodologies is also valuable for prioritizing compounds. Several different methods have been reported for predicting P450-mediated metabolism.[17,18,22,23] Initially, quantitative structure-metabolism relationships[24–27] used small sets of similar molecules and a few molecular descriptors. These evolved into quantitative structure-activity relationships (QSAR)[28–33] for the human P450s, which showed that lipophilicity (logP or molecular refractivity) was important. Steric, electronic, and molecular shape properties were also shown to be significant, which led to more sophisticated models for human enzymes.[34–39]

QSAR and machine learning methods have also been used to model metabolic stability, beyond focusing on individual enzymes. A recursive partitioning model containing 875 proprietary molecules with HLM stability was used to predict the clearance of 41 drugs.[40] A k-nearest neighbour model was used with metabolic stability data from human S9 homogenate for 631 proprietary compounds to classify metabolism of 100 molecules.[41] A set of 130 calcitriol analogs was used to develop QSAR models that prioritized 244 virtual analogs—correctly predicting metabolic stability for 17 of the 20 selected analogs.[42]

Random forest models were generated for 27,697 proprietary molecules with rat liver microsomal clearance data and were found to perform better than Bayesians on an external dataset of 10,011 compounds at Pfizer.[43] Bayesian models were used to predict rat, human and mouse microsomal stability data for proprietary datasets consisting of 15,355, 4,184 and 1,617 molecules, respectively, and had accuracies 75%.[44] As a proof of concept using open source descriptors and modeling tools,[45] models were compared using Pfizer's HLM dataset consisting of approximately 230,000 proprietary compounds. The models were tested with an additional set of 2,310 compounds and were nearly comparable (*e.g.*, positive predicted value of 0.80 and 0.82). A recent study described Collaborative Drug Discovery, Inc. (CDD) Bayesian models for mouse (364 molecules) and human intrinsic clearance (743 and 1100 molecules) using data from ChEMBL, with internal 3-fold cross-validation ROC values of 0.80–0.92.[46]

In the current study, we explored the development of machine learning models, built using public data from PubChem BioAssay,[47,48] that enhance the probability of identifying compounds with MLM stability. We compared the accuracy of Naïve Bayesian classifiers, Support Vector Machines, and Recursive-Partitioning Random Forest models. Using the same descriptors for these approaches, Bayesian models displayed the best accuracy in external test and validation studies, with good internal and external statistics, ROC curves, and enrichment factors. We also explored the effect of “pruning” moderately unstable compounds from the training set, to study whether removing potentially ambiguous information from the model improves predictive power.

Materials and Methods

Molecular Datasets: Retrieval, Re-formatting, Curation, and Pruning

Previously published machine learning models to predict stability in liver microsomes (from mice, rats, and/or humans) were based on very large, in-house, proprietary datasets generated at different laboratories within Pfizer or Wyeth.[43–45,49] Since the details of the structures and metabolic stabilities of the compounds in these datasets have not been publicly disclosed, we searched for available, free data that could be used to train and test our models. The PubChem BioAssay database[48] (<http://www.ncbi.nlm.gov/pcassay>) was queried for all results that contained the phrase “mouse liver microsomes,” which retrieved a total of 950 different sets of assay results. The descriptions and tables of results on PubChem for all 950 sets were manually investigated, and the pertinent results were downloaded as structural data files (sdf format) and comma-separated value (csv) files. The sdf files contained the 2D structural information and PubChem CID (Compound Identification) numbers but no assay results. The csv files contained the assay results and the PubChem CID numbers for the compounds but no structural information. These different types of data were combined manually, by entering the assay results into the corresponding rows of the spreadsheet that Discovery Studio 4.0, (BIOVIA, San Diego, CA),[50] created from the sdf files. The PubChem BioAssay results selected were generally extracted from the literature and deposited into PubChem by ChEMBL.[51]

In one of the first sets of half-life values that were entered into the spreadsheet, the stability values in the csv file from PubChem appeared unlikely: the compounds were listed as

having half-life values on the order of hundreds of hours. The data on PubChem (AID 312986) incorrectly listed the units as hours, when the original publication of the assay results listed the half-life units as minutes. AID 661287 had the same issue with incorrect unit labels. Consequently, for all sets of assay results that were selected for our datasets, the original manuscript from which the data were extracted was manually inspected (*i.e.*, the tables, text, methods, and Supplementary Material sections of the original papers were examined, to make sure that the units were correct and that each value listed in the csv file actually appeared in the results of the paper). This inspection process also revealed that one set of assay results from PubChem (AID 552687) incorrectly listed the values from the column of another type of data (Pgp efflux ratios) that was next to the half-life values as the half-life values. AID 720987 incorrectly listed rat liver microsomal stability data as MLM data. This highlights the importance of a rigorous data inspection and curation process. Although inspecting all of the original sources of these data slowed down the data entry and curation process, it was critical to building datasets with which we had high confidence.

Of the 950 sets of assay results, we chose to focus on two types of data, which were used to create two independent datasets: (A) half-life values in mouse liver microsomes (reported in either minutes or hours) and (B) percent compound remaining (or percent compound metabolized) in mouse liver microsomes, after a specified amount of time (*i.e.*, the methods for the assay had to state the duration of the stability test). Since phase I metabolism was our primary concern, we only utilized the results of assays that were either performed in the presence of NADPH (or an NADPH-regenerating system) or that contained both NADPH and UDP-glucuronosyl transferase (such as AID 643411). (A) 103 different sets of assay results on the half-life of a compound in mouse liver microsomes were accessed, of which 99 sets had data that were utilized. Three sets of assay results were discarded, because they just listed the half-life values as >40 min. (which was too ambiguous for our purposes). The fourth set was discarded, because the assay was performed without NADPH. The 99 sets of assay results on half-life values that were selected encompassed 894 unique small molecules (Fig. 1). All half-life values were used as minutes. (B) 78 different sets of assay results on the percent compound left (or the percent compound metabolized) in mouse liver microsomes were obtained from PubChem BioAssay.[47,48] Each percent compound metabolized value was subtracted from 100 to obtain the percent compound remaining value, which was then entered into the spreadsheet. A total of 571 unique small molecules were used to create this second dataset, which contained these percent compound left values (Fig. S1). For both the half-life dataset and the percent compound left dataset, the “qualifier” data (*i.e.*, < or > symbols that pertained to the stability results) were recorded in the spreadsheet as a separate column, to ensure proper sorting of the data.

Duplicate compounds were removed from within each dataset. In almost all cases, when a compound was present in the results of more than one assay, the stability values were very similar (which meant that it did not matter which instance of the compound was kept and which instance was deleted). In these cases, the additional entries were removed, and a single instance of that compound was kept in the dataset. However, in a few cases the conditions of the assay changed (*e.g.*, the concentration of the compound was increased from 1 μ M in one assay (AID 623834) to 10 μ M in another assay (AID 623836)), and the

assay results also changed—an unstable compound in one assay appeared to be stable in a different assay. In these ambiguous cases, all instances of those compounds (*e.g.*, PubChem CID numbers 46,219,863; 46,219,973; and 46,220,295) were removed from the dataset.

Training, Test, and Validation Sets

Since our goal was to develop a model that could help identify compounds suitable for *in vivo* studies of potential antitubercular or antimalarial therapeutics in mice, we set a high bar when classifying compounds as either metabolically “stable” or “unstable”. For the mouse liver microsome half-life dataset, compounds with a $t_{1/2} < 59.5$ min were classified as “unstable,” and compounds with a $t_{1/2} \geq 59.5$ min were defined as “stable” (Fig. 1). The “full” half-life dataset on mouse liver microsomal stability thus contains 262 “stable” compounds (29.3% of the set) and 632 “unstable” compounds (70.7% of the set). The “full” percent-compound left stability dataset (Fig. S1) has 109 “stable” compounds (19.1% of this independent set), which had $\geq 50\%$ remaining after 60 min in mouse liver microsomes, and it has 462 “unstable” compounds (80.9%), which had $< 50\%$ left after either 60, 40, 30, 20, 15, or 10 min in mouse liver microsomes (*i.e.*, assays that were performed for at least 60 min were used to identify stable and unstable compounds, but assays that had a shorter duration were only used to provide unstable compounds, since the amount a compound is metabolized does not always scale linearly with time).

Bayesian models are binary classifiers—they are trained on datasets that are divided into two categories, such as “active” and “inactive,” “toxic” and “non-toxic,” or “stable” and “unstable”.[44,49,52–70] However, experimental data on compounds tend to be more continuous (*i.e.*, they represent an analog spectrum, not a digital “yes” or “no” categorization). We hypothesized that moderately stable/moderately unstable compounds might contain ambiguous information (or disinformation) that could confuse the machine learning process and decrease the predictive power of the model developed. Consequently, we chose to compare the accuracy of the model produced with the “full” half-life dataset to a model produced by a dataset in which these moderately unstable/moderately stable compounds were removed (or “pruned”). This produced the “pruned” half-life dataset of 759 compounds, which contains the same 262 stable compounds as the “full” half-life set with a $t_{1/2} \geq 59.5$ min (now 34.5% of the pruned dataset), but it only has the 497 unstable compounds (65.5%) that had a $t_{1/2} < 30$ min (*i.e.*, compounds were removed if they displayed a $t_{1/2}$ between 30 and 59.5 minutes). Following the same logic, the “pruned” percent compound left dataset has the same 109 stable compounds as the “full” percent-compound left set (now 20% of the pruned set), and it has 437 unstable compounds (80%), which had $< 25\%$ left after 60 min in mouse liver microsomes (*i.e.*, compounds were removed if they had 25 to 49% left after 60 min or if they had 43 to 48% after 40 min).

For the Bayesian models tested with the Collaborative Drug Discovery database (CDD, Burlingame, CA. <http://collaborativedrug.com>),[71] one compound was not uploaded properly into the CDD Vault for either half-life dataset, which was likely due to an issue that results when multiple tautomers of one compound are mapped to the same unique CDD ID #. Thus, for only the “CDD Bayesian” model results presented, those half-life training sets

contained 261 “stable” compounds and the aforementioned numbers of “unstable” compounds.

The full and pruned half-life datasets were each used as independent training sets to create different types of machine learning models in Pipeline Pilot 9.1 (BIOVIA, San Diego, CA). [50] These models were also used to calculate Bayesian scores and the “closest distance to the training set” (a measurement of structural similarity that compares each member in a test set to every member in the training set) values on compounds in Discovery Studio 4.0 (using the “calculate molecular property” tool).[50] The machine learning models created with the full half-life and pruned half-life sets were initially tested on the “Dartois 2015 set” of known antitubercular compounds. The independent “Dartois set” contains 30 antitubercular drugs with known clearance values from mouse liver microsomes.[72] The sdf file for this test set was manually created (using structures from <http://www.ChemSpider.com>) and curated; it has 27 “stable” compounds and 3 “unstable” compounds by our criteria. Each type of machine learning model created with the full half-life and pruned half-life sets was also assessed with a 2nd independent test set (*i.e.*, a validation set), the full percent-compound left set (Fig. 1 and Fig. S1). An alternative strategy that involved using the full percent-compound left and the pruned percent compound left sets as the initial training sets, followed by using the Dartois 2015 set as the test set and the full half-life set as the validation set was also investigated.

Principal Component Analyses to Characterize Chemical Properties

The chemical property space encompassed by each dataset was compared to the other datasets by concatenating their relevant sdf files together and performing a Principal Component Analysis (PCA) in Pipeline Pilot 9.1 and Discovery Studio 4.0.[50] These PCA studies were performed using established protocols.[73,74] Briefly, eight different interpretable molecular descriptors were calculated and utilized to characterize the chemical properties of each compound: ALogP, molecular weight, number of rings, number of aromatic rings, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, and molecular fractional polar surface area. Pipeline Pilot then combined these eight descriptors in different ways, with different weights on each term, to reduce the dimensionality of the data into three Principal Components that describe most of the variance in the compounds.

Bayesian Models to Predict Mouse Liver Microsomal Stability

Bayesian models build a binary classifier by combining different sets of descriptors, with different weights on each descriptor, in different ways, to build a model that best reproduces the known trends in the training set. The sdf files of the full half-life and pruned half-life training sets were utilized to build two different Naïve Bayesian models (hereafter called “Bayesian models”) in Pipeline Pilot 9.1.[50] Our Bayesian models used 9 different interpretable descriptors: ALogP, molecular weight, number of rings, number of aromatic rings, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, molecular fractional polar surface area, plus the “molecular function class fingerprints of maximum diameter 6” (FCFP_6).[75] Thus, 8 of the 9 descriptors describe the physiochemical properties of the entire compound, while the FCFP_6 fingerprints

describe the 2D topology of different subregions of each compound (from the functional groups adjacent to each atom up to and including 6 rings, or dimensions, of topological connections). Different features and subregions of the “good” compounds are mixed and matched to build the model and to score new compounds. However, due to the results of our Principal Component Analyses, we also investigated additional Bayesian models (including the “CDD Bayesians”) that were constructed using just FCFP_6 (i.e., the other 8 descriptors were excluded). We constructed these “Laplacian-corrected Naïve Bayesian classifier” models using our previously published protocols.[4,53,55,56,60,73,76,77] Models were internally validated using 5-fold cross-validation in Pipeline Pilot, which involves leaving out a random selection of 20% of the compounds, building a model with the remaining 80% of the dataset, and evaluating the model with the set of 20% that was initially left out. This process was repeated five times, and the following “internal” statistics for the best model were calculated by Pipeline Pilot[50]: ROC rating (Receiver Operator Characteristic curve’s quality), ROC score (the area under the curve of the ROC plot), sensitivity, specificity, and concordance. When a Bayesian model was then assessed by using an independent “test set” to investigate its predictive power, Pipeline Pilot calculated the same type of statistics as “external” metrics. After a Bayesian model was built in Pipeline Pilot, that model was then utilized with the “calculate molecular properties” tool within Discovery Studio 4.0[50] to calculate the Bayesian scores for each compound in the test and validation sets, to enable a manual assessment of the “stability hit rates” (i.e., the “positive predicted values,” which are the number of true positives divided by the sum of the number of true positives and the number of false positives) and enrichment factors for the compounds that displayed the top Bayesian scores.

To test the feasibility of making our Bayesian models publicly accessible, and to test the accuracy of the slightly different algorithms utilized, additional versions of our “full” half-life and “pruned” half-life Bayesian models were constructed on the Collaborative Drug Discovery database (CDD, Burlingame, CA. <http://collaborativedrug.com>),[71] using a beta test of the first version of the “build QSAR model” CDD Models software.[46,78] These “CDD Bayesian Models” used only one descriptor, an open source version of the FCFP_6 topological fingerprint.[79] In addition, instead of the 5-fold cross-validation protocol employed in Pipeline Pilot, the CDD Bayesian Models are constructed and internally validated using 3-fold cross-validation (in which 33.3% of the test set is left out, a model is built, the model is tested on the remaining 66.6% of the compounds, and the process is repeated three times). The software computes the overall, internal ROC score, but instead of providing the overall, internal sensitivity and specificity, it provides different sensitivity and specificity values that would be produced when using a particular Bayesian score as the cut-off.

Support Vector Machine (SVM) and Recursive Partitioning (RP-Forest) Models to Predict Mouse Liver Microsomal Stability

We also generated SVM and RP Random Forest models with the same 9 molecular descriptors in Discovery Studio. SVM models find a plane that best separates a description of the two classes of compounds in the training set. For complicated datasets, algorithms called kernel functions are used to transform the descriptors of a dataset into a higher

dimensional space, to better separate the “good” and “bad” compounds. Support vectors are constructed that use a minimal number of the “good” and “bad” compounds to define a boundary to the hyper plane (the multi-dimensional plane that separates the transformed descriptions of the compounds in higher dimensional space), in a way that attempts to minimize error while maximizing generalizability. As described previously,[1] for SVM models we calculated interpretable descriptors in Discovery Studio, then used Pipeline Pilot to generate the FCFP_6 descriptors, followed by integration with R.[80] The parameters used were as follows: SVM-Type = C-classification; SVM-Kernel = radial; Cost = 2; Gamma = 0.007352941.

Recursive Partitioning Random Forest models are an extension of the decision tree approach that minimizes errors from over-fitting by generating multiple trees and randomizing the training sets and sets of descriptors that each tree utilizes to reproduce the trends in the training set. “Bagging” (bootstrap aggregation) with replacement is used to generate a new, modified version of the training set for each tree: each tree tends to get a random subset of the original training set, with ~ 2/3 unique compounds and ~ 1/3 duplicates. When each tree is “induced” (created), it receives a different subset of the descriptors and will place different weights on the descriptors, which can also be used in different orders in the different hierarchies of the levels of nodes within each tree. Each tree then develops a different model to reproduce its training set. When the collection of these trees, the Random Forest, is then used to score compounds, the ultimate classification of each compound depends on a majority vote from all of the trees. RP Forest models were calculated using the standard protocol in Discovery Studio (Tree options: Minimum samples per node = 10; Maximum tree depth = 20; Split method= Gini; Weighting Method = By Class; Maximum Knots per property = 20; Maximum look ahead depth = 0; Maximum generic depth = 0). In the case of RP Forest models, ten trees were created with “bagging.” Since the same 9 descriptors that were used with the Bayesian and SVM models were used in the RP Forest approach, 10 trees seemed appropriate, to avoid over-fitting. In all cases, 5-fold cross-validation (see the description in the Bayesian methods) was used to calculate the ROC curve and statistics for the models generated.

Results

Internal, Five-fold Cross-validation Statistics

For the Bayesian (Table I) as well as Support Vector Machine (SVM; Table S-I) and RP Random Forest (Table S-I) models investigated, “pruning” out the moderately unstable/moderately stable compounds from the training set produced a model that displayed superior predictive power in internal five-fold cross-validation, as compared to the model created from the respective original full training set. The Bayesian model created from the pruned half-life training set displayed a better ROC score, or area under the ROC curve, than the full half-life Bayesian (0.870 versus 0.835), a similar sensitivity (93.1 vs. 92.7), a much higher specificity (80.9 vs. 72.2), and a much better concordance (85.1 vs. 78.2). Similarly, the pruned half-life Bayesian also displayed a slightly better ROC curve than the full half-life Bayesian in the results of five-fold cross-validation (Fig. S2). The same trends were

observed for the ROC scores for the pruned versus full SVM models and for the pruned vs. full RP Random Forest models during five-fold cross-validation (see Table S-I).

Statistics from External Test and Validation Studies of the Bayesian Models

Similar to these trends observed with the internal test, the pruned half-life Bayesian also displayed a better ROC curve than the full half-life Bayesian in the external test with the independent Dartois 2015 set[72] of 30 known antitubercular compounds (Fig. S3). The pruned half-life Bayesian also displayed a better ROC score than the full half-life Bayesian when they were tested on this independent, external set (0.778 vs. 0.704; Table II). The other statistics were equivalent for the pruned and full half-life Bayesian models in this test, with an external sensitivity of 81.5, specificity of 33.3 (due to correctly predicting only 1 of the 3 unstable compounds as unstable), concordance of 76.7, and positive predicted value (or “stability hit rate”) of 22/24 (92%). Thus, for this external test with antitubercular compounds, “pruning” the moderately unstable/moderately stable compounds from the training set improved or maintained the predictive power, depending on the particular metrics used to judge performance.

A Principal Component Analysis (PCA) was performed to compare the chemical property space sampled by the full and pruned half-life training sets (Fig. S4 and Fig. S5), to study whether any trends exist amongst the moderately unstable compounds that were pruned. The pruned compounds, however, were not obvious outliers; in fact, no significant clustering occurred between the majority of the stable compounds, the very unstable compounds, and the compounds that were removed. Similarly, no significant clustering occurred to differentiate most of the very stable, stable, moderately unstable, and very unstable compounds, either (Fig. S5).

In addition to the overall statistics that demonstrated that the pruned Bayesian model was more accurate than the full Bayesian model, which also had good predictive power when evaluating the Dartois 2015 set of antituberculars, the positive predicted values (or “stability hit rates”) for the top-scoring compounds displayed a similar trend. For the 15 antituberculars in this set that received the top Bayesian scores, the full half-life Bayesian correctly ranked all 15 as stable (i.e., this model did not make a mistake until it reached the 16th highest-scoring compound). 18 out of the 20 top-scoring compounds (90%) were true positive predictions by the full half-life Bayesian. This predictive power was even higher for the pruned half-life Bayesian: of the 18 compounds that received its highest scores, all 18 were correctly identified as stable, and 19 out of the 20 top-scoring compounds (95%) were true positives. Only 3 of the 30 compounds in the Dartois 2015 set are unstable; thus, 90% are stable (which is likely atypical for most sets). Consequently, the best enrichment factor that could possibly be achieved for this test set by a perfect model is $100\%/90\% = 1.11$. Although our half-life Bayesian models did not perform flawlessly, they did achieve this 1.11 enrichment factor for the half of this set that received the top Bayesian scores.

When compared to the Dartois 2015 set of antituberculars, the pruned and full half-life training sets used to create these Bayesian models are very different, with respect to both chemical property space (see the PCA comparisons in Fig. S6 and Fig. S7) and 2D structural similarity. Using the “closest distance” between each compound in the test set compared to

every member of the training set, the average closest distance between the full half-life training set and the Dartois 2015 set of antituberculars is 1.15, with a minimum closest distance of 0.52 and a maximum of 2.89. A “closest distance” of 0 corresponds to two equivalent compounds. The pruned half-life training set and the Dartois test set displayed an average closest distance of 1.16, with a minimum closest distance of 0.52 and a maximum closest distance of 2.90. Using the more conventional metric, according to the maximum Tanimoto similarity displayed between each member of this test set compared to every member of the training set (as calculated with the CDD Models software),^[71] the full half-life training set had an average similarity of 0.178 to the Dartois 2015 set, with a minimum of 0.098 and a maximum of 0.250. A Tanimoto similarity of 1.0 indicates that two compounds are equivalent, with smaller values corresponding to increasing dissimilarity. The pruned half-life training set had average, maximum, and minimum Tanimoto similarity values for the Dartois 2015 set that were identical to the values displayed by the full half-life training set. Thus, the pruning process did not make the pruned half-life set more similar to the Dartois 2015 test set, but it did make the pruned half-life Bayesian more accurate against it.

Since the Dartois 2015 set had such a high proportion of stable compounds (90%), we chose to validate these models with an additional independent, external set that was unbalanced in the other direction—the full percent compound left set (Fig. 1 and Fig. S1). This is a large, diverse, and very challenging validation set, in which only 19.1% of the compounds were stable in the presence of MLM. Unlike the aforementioned trends from the internal, five-fold cross-validation studies and the external test with the Dartois 2015 set, the full half-life Bayesian displayed slightly better accuracy in the external validation studies with the full percent compound left set according to some of the statistics, and it displayed similar predictive power to the pruned half-life Bayesian according to other statistics (Table S-III). In this validation study, the full half-life Bayesian had a similar external ROC score to the pruned half-life Bayesian (0.785 vs. 0.777; not a significant difference in our experience). The overall shapes of the ROC curves were very similar and were both fairly accurate (Fig. S9). The full half-life Bayesian had a similar external concordance (56.2% vs. 55.2%) and stability hit rate (28.2% vs. 26.3%), but it had a better external specificity (49.8% vs. 44.8%). However, the full half-life Bayesian and the pruned half-life Bayesian displayed the exact same sensitivity (83.5%).

For the 30 or 50 compounds that received the top Bayesian scores (i.e., the top 5 or 9%), the full and pruned half-life Bayesians both displayed similarly good enrichment factors, with slightly better enrichment displayed by the full half-life model. Of the compounds in the full percent compound left validation set that received the top scores from the full half-life Bayesian, 21 out of 30 (70.0%) were actually stable, as were 33 out of 50 (66.0%). Thus, for the top-scoring 30 compounds, the full half-life Bayesian displayed an enrichment factor of $70.0\%/19.1\% = 3.66$, while for the top 50 the enrichment factor was 3.46. Similarly, of the compounds in this validation set that received the top Bayesian scores from the pruned half-life Bayesian, 20 out of 30 (66.7%) and 31 out of 50 (62.0%) were actually stable, giving it enrichment factors of 3.49 and 3.25, respectively.

The entire distribution of pruned half-life and full half-life Bayesian scores versus the proportion of experimentally stable or unstable compounds that received a particular score is displayed in Fig. 2. According to this histogram comparison (Fig. 2.A), instead of using the top N compounds, if a pruned half-life Bayesian score cut-off of 0 is used to filter this validation set, then 78.0% of the experimentally stable compounds would remain, while 67.3% of the unstable compounds would be removed. For perspective, the compound with the 50th highest ranking had a pruned half-life Bayesian score of 4.79. The compound with the 50th highest ranking had a full half-life Bayesian score of 3.88. The corresponding histogram comparison with the full half-life Bayesian scores (Fig. 2.B) displayed a very skewed bimodal distribution of scores for the compounds that are known to be experimentally stable. If a full half-life Bayesian score cut-off of 0 is used to filter this validation set, only 37.6% of the stable compounds would be retained, while 57.4% of the unstable compounds would be discarded. Thus, this comparison of histograms firmly establishes the superiority of the pruned half-life Bayesian model over the full half-life Bayesian model (both constructed using 9 descriptors) for this validation study. These good external ROC scores, sensitivity values, and enrichment factors were displayed with this challenging validation set, despite the fact that these half-life training sets and the full percent compound left validation set generally have very different structures (see Supplementary Material Commentary on Structural Comparison of the Half-life and Percent Compound Left Sets and the Effects of Removing Duplicate Compounds on Enrichment Factors).

Using the Percent Compound Left Validation Set as a Training Set to Construct a Bayesian

To be thorough, the full and pruned percent compound left datasets were also used to create Bayesian models, which were then tested with the independent “Dartois 2015” set of antituberculars and with the full half-life set (Fig. S1). Although the full and pruned percent compound left Bayesians both displayed excellent ROC curves during internal five-fold cross-validation (Fig. S10), they both performed quite poorly on the Dartois test set of 30 antitubercular compounds (Fig. S11). Both percent compound left Bayesians also failed the validation study in which they scored the full half-life dataset, due to displaying ROC curves that were similar to random chance (see Fig. S12). Consequently, we focused all of our remaining efforts on using the half-life datasets to build machine learning models.

Comparing the Performance of Pruned vs. Full SVM and Random Forest Models

The full and pruned half-life datasets were also used to build two versions of a Support Vector Machine model and two different RP-Random Forest models. The results of internal five-fold cross-validation studies indicated that pruning the training set can also improve the performance of other types of machine learning models (Table SI). In the external test with the Dartois 2015 set of antituberculars, pruning the training set produced the most significant improvement in accuracy overall for the SVM models (Table S-II). The pruning strategy produced less clear trends for the RP Random Forests results with the external Dartois 2015 test set. Overall, the full Forest performed better than the pruned Forest in external tests (Tables S-II and S-III), but both Random Forest models were not accurate at correctly identifying compounds that are stable in the presence of MLM.

Importance of Different Descriptors When Constructing Bayesian Models

The different PCA comparisons (Fig. S4, S5, and S8) did not indicate that any significant clustering occurred that could help distinguish most stable from unstable compounds. Since these PCA are performed using 8 of the 9 descriptors utilized in the machine learning models (*i.e.*, everything except the FCFP_6 descriptor, which describes the 2D topology), it seemed to suggest that these 8 descriptors might not be contributing much to the predictive power that the half-life Bayesians displayed in the internal and external studies.

Consequently, we explored the construction of full and pruned half-life Bayesians that utilized only one descriptor, FCFP_6, and compared them to the aforementioned Bayesians that were built with all 9 descriptors. These FCFP_6-only Bayesian models were also constructed in Pipeline Pilot 9.1. Similar to our previous results with the 9-descriptor-based Bayesians, these FCFP_6-only, pruned half-life Bayesians displayed better internal statistics in cross-validation studies than the corresponding full Bayesians (Table I). The FCFP_6-only, pruned half-life Bayesian displayed better internal ROC scores (0.869 vs. 0.835), better specificity (76.3% vs. 71.4%), better concordance (82.3% vs. 77.7%), and similar sensitivity (93.9% vs. 93.1%) to the corresponding full half-life Bayesian. In the external tests with the Dartois 2015 set of antituberculars (Table S-IV), the FCFP_6-only, pruned half-life Bayesian displayed a slightly better external ROC score than the corresponding full Bayesian (0.815 vs. 0.790), but all of the other external statistics were equivalent between the full and pruned models (regardless of whether they were built with 9 descriptors or 1).

Similarly, when the enrichment factors for the molecules in the percent compound left validation set that received the top 50 Bayesian scores were compared, the Bayesian models constructed with only FCFP_6 actually displayed better predictive power than the corresponding Bayesians constructed using all nine descriptors (Table II). The enrichment factors for the top 50-scoring compounds were quite good for the 9-descriptor-based half-life Bayesians built in Pipeline Pilot 9.1 (3.46 for the full model and 3.25 for the pruned model). However, the enrichment factors for the top 50-scoring compounds were even better for the FCFP_6-only Bayesian models (4.29 for the full half-life model and 3.98 for the pruned model). Thus, when selecting <10% of the top-scoring compounds (*i.e.*, for subsequent ordering or synthesis and experimental evaluation), the full half-life Bayesian constructed with only FCFP_6 as the sole descriptor seems superior.

The utility of using all 9 descriptors, instead of only using FCFP_6, was not apparent until we compared the histograms of the Bayesian scores for experimentally stable vs. unstable compounds from these four types of models (*i.e.*, pruned and full half-life Bayesians constructed using either nine descriptors or only FCFP_6). In this histogram analysis, the full half-life Bayesian constructed with all 9 descriptors was clearly flawed (Fig. 2), since it displayed a skewed bimodal distribution that only retained 37.6% of stable compounds and removed 57.4% of unstable compounds when using a cut-off score 0 to filter the percent compound left validation set. Conversely, the full half-life Bayesian constructed with only FCFP_6 displayed the expected distribution pattern and retained 76.1% of the stable compounds, while removing 65.4% of the unstable compounds using a cut-off score 0 as the filter. Removing (or “pruning”) those 8 descriptors, which did not help distinguish stable from unstable compounds in the PCA plots, improved the predictive power of the FCFP_6-

only, full half-life Bayesian model in this validation study (as compared to the full half-life Bayesian that used all 9 descriptors). However, the pruned half-life Bayesian constructed using all 9 descriptors performed the best overall according to this metric, harvesting 78.0% of the stable compounds and removing 67.3% of the unstable compounds when using a cut-off score 0 as the filter, while the pruned FCFP_6-only half-life Bayesian model retained 73.4% of the stable compounds and removed 64.7% of the unstable compounds. Thus, when it came to evaluating a large, diverse, challenging validation set with only 19.1% stable compounds, the pruned 9 descriptor-based Bayesian displayed the best predictive power at harvesting a large fraction of the stable compounds, while simultaneously removing over 2/3 of the unstable compounds. Although the 8 descriptors used to perform PCA do not obviously distinguish stable from unstable compounds, when the FCFP_6 topological descriptor is combined with those 8 descriptors, the accuracy of the pruned half-life Bayesian model for predicting MLM stability is enhanced as a tool for the initial filtering of a large library (Fig. 2).

Structural Features Associated with MLM Stability or Instability

One distinct advantage of the Bayesian modeling approach, as compared to SVM and RP-Random Forest models, is that Bayesian models constructed in Pipeline Pilot produce a characterization of the “good” and “bad” molecular features (*i.e.*, the 2D fingerprint-based fragments that are associated with the “good”/stable compounds and with the “bad”/unstable compounds). Pertaining to the pruned half-life Bayesian model, the features most frequently associated with MLM stability are shown in Fig. 3, while the types of substructures prevalent in unstable molecules are displayed in Fig. 4. Substructural features associated with stability included pyridone and saturated analogs, quinoline/quinolone and saturated analogs, N-substituted pyrrole and saturated analogs, β -hydroxy ether, fluoroanilines (2-substituted, 3-substituted, and 2,3-disubstituted), and hydroxyamine/hydroxamic acid. Substructural features associated with a lack of MLM stability included five-membered heterocycles with a pendant carboxamide, azepinoindoles with various substituents, and *meta*-substituted aryl ethers/phenols. Importantly, these features either consistent or inconsistent with MLM stability are different from the features that were published in previous studies of liver microsomal stability.[44,49] Thus, new structural insights were provided regarding possible structure-stability relationships, which should be investigated experimentally in the future.

Internal and External Evaluation of Open Source “CDD Bayesian” Models

Similar to our previous results, the open source FCFP_6-only, pruned half-life CDD Bayesian displayed better internal statistics in cross-validation studies than the corresponding full half-life Bayesian (Table III). The pruned half-life CDD Bayesian displayed a better internal ROC score (0.86 vs. 0.83), a better specificity (66% vs. 58% or 76% vs. 72%, using the same sensitivity-based cut-off), and a slightly better ROC curve (Fig. S14) than the corresponding full half-life CDD Bayesian.

In the external tests with the Dartois 2015 set of antituberculars (Table III), the pruned half-life Bayesians constructed in CDD produced the exact same statistics as the corresponding full Bayesian. They were both also quite accurate in this external test set, with concordance

values of 67% using the less stringent score cut-off and 90% using the more stringent score cut-off, with stability hit rates (positive predicted values) of 17/17 (100%) and 26/28 (93%), respectively.

The CDD Bayesians performed significantly worse with the percent compound left validation set. The stability hit rates (positive predicted values) were 32% or 33% using the less stringent score cut-off for the pruned or full half-life CDD Bayesians, respectively, and both produced a 21% hit rate using the more stringent score cut-off. The enrichment factors for the top 50-scoring compounds in the validation set were unsatisfactory for the Bayesians constructed in CDD (1.15 for both the full and pruned half-life models; see also Supplementary Material Commentary: Structural Comparison of the Half-life and Percent Compound Left Sets and the Effects of Removing Duplicate Compounds on Enrichment Factors).

Discussion

Efficacy studies in infected mice are a critical hurdle to advance the translational research of potential therapeutic compounds. We have witnessed this first hand during our work on different projects developing small molecules for the treatment of tuberculosis[56] or malaria.[77] While the mouse is not an absolutely predictive model for translating these[1–4] or other diseases[7,8,11,12,14] to human studies, it is relatively cheap and widely accessible. One of the key criteria for obtaining sufficient *in vivo* efficacy is displaying high metabolic stability (or low to moderate clearance of the compound).[44,81] One early way to assess this metabolic stability or intrinsic clearance is to incubate the compound with MLM and evaluate the amount of parent compound remaining. While this is not a perfect predictor for the mouse *in vivo*, it is faster and cheaper than conducting an *in vivo* study. Even more time- and cost-efficient still is the use of computational models that can help prioritize compounds to synthesize or purchase, before pursuing experimental tests.

Pharmaceutical companies have generated increasingly larger, internally consistent datasets that are highly predictive for ADME/Tox properties.[44,82–85] Unfortunately for those in academia or small companies, these models are not accessible, and very rarely are the underlying data made public or the models shared due to concerns over intellectual property. [46] Fortunately, several sources of public data relevant to metabolic stability exist in ChEMBL[51,86,87] or PubChem,[47,48] containing data from numerous publications, including studies from the National Institutes of Health.[88–91] In the current study we focused on the ‘MLM half-life’ and ‘percent compound left’ data in PubChem BioAssay. [48]

From the academic perspective the ability to construct and share models, such as those for predicting metabolic stability, is highly desirable so that resources in multiple labs can be focused on testing the compounds that have a higher probability of meeting the required metrics or decision gates. Consequently, our curated version of the MLM datasets (*i.e.*, our full and pruned versions of both the half-life and percent compound left data) are being made available as sdf files in the Supplementary Material and can also be shared as CDD Models.

In the current study we mainly focused on using the Naïve Bayesian classifier algorithm for model building. This strategy is based on our and others' extensive use of this method to build predictive models for ADME/Tox properties [46,59,64,92,93] as well as models for activity against *Mycobacterium tuberculosis* (*Mtb*) [1,55,69,73,94]. For means of comparison, we also tested our datasets with Support Vector Machine and RP-Random Forest models.

The models in this study were evaluated with two different external, independent sets to test and then validate their accuracy. In our external test and validation studies, our half-life Bayesian models displayed better predictive power than the SVM and Random Forest models that were built with the same descriptors, especially regarding sensitivity (*i.e.*, correctly identifying MLM stable compounds). Our half-life Bayesian models were perhaps too permissive overall (*i.e.*, they classify too many compounds in the entire set as stable), but they display sufficient stringency when examining the compounds that receive the top Bayesian scores (*i.e.*, they had good enrichment factors and histogram-based filtering). In contrast, the half-life SVM and Random Forest models tended to classify a vast majority of compounds as unstable (*i.e.*, they were too restrictive), which makes them less useful for our intended purposes.

Our goal is to increase the probability of finding stable compounds, as opposed to focusing on eliminating most compounds that might be unstable. Consequently, when using an MLM stability Bayesian as the first filtering step (before applying other types of metrics or methods as filters) for a large library of compounds, we advise using the pruned half-life Bayesian with 9 descriptors, based on the histogram analyses. However, if MLM stability is the sole metric for selecting a small percentage of compounds, then the full half-life Bayesian constructed with only FCFP_6 in Pipeline Pilot is advised, based on the enrichment factors. In our opinion, this represents the optimal strategy for utilization of our MLM datasets and machine learning models, although some expected stable features (*e.g.*, aryl fluorides) and unstable features (*e.g.*, aliphatic ester) may be noted and leveraged during a compound optimization campaign.

More extensive Random Forest models could be built in the future that use sets of several hundred different (generally non-interpretable) descriptors and hundreds of different trees; thus, the accuracy of Random Forest models built from our curated MLM training sets could likely be improved. More accurate SVM models that use different kernel functions could also be pursued in the future. But in the present study we wanted to compare different types of machine learning models that were constructed using the same interpretable descriptors, to avoid adding extra variables.

When we utilized the initial beta version of the CDD Models software to construct half-life Bayesian models, the internal statistics were accurate, and they performed very well with the Dartois 2015 test set of antituberculars. However, their performance declined with the percent compound left validation set, especially when considering the enrichment factors. This drop in performance might be due to differences in the nature of the different libraries, since only 19% of the compounds in the validation set were stable. Regardless, it highlights the utility of studying a wide range of metrics (such as histograms, positive predicted values,

and enrichment factors) when evaluating the accuracy of Bayesian models, instead of relying on the ROC score, sensitivity, specificity, and concordance values. It also indicates that further improvements in the CDD Bayesian modeling protocol might be warranted, such as adding the capability to use the 8 additional descriptors or perhaps more as needed (instead of only one—the open source FCFP_6 descriptor for 2D topology).

Since Bayesian, SVM, and Random Forest models are all binary classifiers, we hypothesized that pruning out potentially ambiguous information from the training set might enable a more informed education process for these machine learning models. Admittedly, this is a somewhat counterintuitive notion, since one might wonder how training a machine learning model with less information could actually make it more accurate. To the best of our knowledge, we are the first to explore how pruning the moderate compounds from a training set affects the accuracy of Bayesian models. Although it might be considered similar to margin detection used previously with machine learning models,[95] margin detection was used to remove compounds that were near the classification boundary from the test sets—not from a training set used to construct the model. Our pruned half-life Bayesian, SVM, and Random Forest models all displayed better predictive power in internal cross-validation studies than the full half-life models. Similarly, in the external test with the independent set of 30 antitubercular compounds, the sensitivity, concordance, and positive predicted values (i.e., stability hit rates) were all improved or maintained for the three types of pruned half-life models, as compared to their respective full half-life models. When the histograms of the Bayesian scores for the experimentally stable versus unstable compounds were compared, the pruned half-life Bayesian model constructed with nine descriptors was clearly superior to the corresponding full half-life Bayesian. Overall, these results support the notion that our novel pruning strategy may be of general benefit, warranting further testing by the machine learning model community at large.

The few previous studies that used machine learning to model liver microsomal stability were performed at Pfizer[43,49] or at Wyeth (now Pfizer),[44] using proprietary data from each company. A third study was performed by Pfizer and collaborators at CDD.[45] The datasets from these previous studies were proprietary and were never disclosed to the research community. The only other recent work in this area used a smaller set of mouse intrinsic clearance data from ChEMBL, likely a subset of the current PubChem dataset.[46] This and the current work utilized CDD Models to build models from microsomal data that could be shared with the community, in contrast to the earlier models published by pharmaceutical companies. The Supplementary Material may be consulted for how to obtain a copy of the curated datasets developed and utilized in the present study.

Conclusions

Our validation results suggest our Bayesian MLM models may have value for prospective predictions. The training sets used to build our pruned and full half-life models were very different than the compounds in the Dartois set of antituberculars (Figure S6) and almost all of the compounds in the full percent compound left validation set, in terms of chemical property space (for the antituberculars) and 2D Tanimoto similarity or structural distance (for both external sets). Yet our Bayesian models displayed good overall accuracy and

stability hit rates with these independent, external assessments. In particular, the “pruning,” or removal, of potentially ambiguous data in the training set led to a more predictive model, highlighting a strategy for further evaluation. This study represents the most exhaustive study to date using machine learning approaches with MLM data from public sources. We are currently evaluating the utility of these models with a variety of small molecules under investigation in our laboratories as anti-infective agents. Through making the curated MLM datasets available, we hope to encourage additional evaluation of our models in the scientific community.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J.S.F., S.E., and A.L.P. were supported by Award Number 1U19AI109713 NIH/NIAID for the “Center to develop therapeutic countermeasures to high-threat bacterial agents,” from the National Institutes of Health: Centers of Excellence for Translational Research (CETR). S.E. and J.S.F. were also supported in part by Award Number 9R44TR000942-02 “Biocomputation across distributed private datasets to enhance drug discovery” from the National Center for Advancing Translational Sciences. We thank Dr. John Piwinski for suggesting that an MLM $t_{1/2}$ of 60 min was ideal, but a $t_{1/2}$ of 30 minutes was not significantly unfavorable. S.E. kindly acknowledges Alex Clark, Molecular Materials Informatics, Inc. and Krishna Dole and colleagues at Collaborative Drug Discovery, Inc., for their development of CDD Models. We thank Thomas Mayo at BIOVIA (formerly known as Accelrys, Inc.) for providing S.E. and J.S.F. with Discovery Studio and Pipeline Pilot. We also thank Jodi Shaulsky at BIOVIA and Katalin Nadassy for assistance with setting up and maintaining the license server and Pipeline Pilot server.

Abbreviations Used

ADME/Tox	absorption, metabolism, distribution, excretion and toxicity
CDD	Collaborative Drug Discovery
FCFP₆	molecular function class fingerprints of maximum diameter 6
HLM	human liver microsomal stability
HTS	High Throughput Screens
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
PPV	positive predictive value
QSAR	Quantitative Structure-Activity Relationships
ROC	receiver-operator characteristic
SAR	Structure Activity Relationship
SVM	Support Vector Machine

References

1. Ekins S, Pottorf R, Reynolds RC, Williams AJ, Clark AM, Freundlich JS. Looking back to the future: Predicting in vivo efficacy of small molecules versus *Mycobacterium tuberculosis*. *J Chem Inf Model*. 2014; 54:1070–1082. [PubMed: 24665947]

2. Franzblau SG, DeGroote MA, Cho SH, Andries K, Nuermberger E, Orme IM, Mdluli K, Angulo-Barturen I, Dick T, Dartois V, Lenaerts AJ. Comprehensive analysis of methods used for the evaluation of compounds against *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)*. 2012; 92:453–488. [PubMed: 22940006]
3. Dartois V, Barry CE 3rd. A medicinal chemists' guide to the unique difficulties of lead optimization for tuberculosis. *Bioorg Med Chem Lett*. 2013; 23:4741–4750. [PubMed: 23910985]
4. Ekins S, Nuermberger EL, Freundlich JS. Minding the gaps in tuberculosis research. *Drug Discovery Today*. 2014
5. Lotharius J, Gamon-Benito FJ, Angulo-Barturen I, Clark J, Connelly M, Ferrer-Bazaga S, Parkinson T, Viswanath P, Bhandokar B, Rautela N, Bharath S, Duffy S, Avery VM, Mohrle JJ, Guy RK, Wells T. Repositioning: The fast track to new anti-malarial medicines? *Malar J*. 2014; 13:143. [PubMed: 24731288]
6. Kaushansky A, Mikolajczak SA, Vignali M, Kappe SH. Of men in mice: The success and promise of humanized mouse models for human malaria parasite infections. *Cell Microbiol*. 2014; 16:602–611. [PubMed: 24506682]
7. Goyama S, Wunderlich M, Mulloy JC. Xenograft models for normal and malignant stem cells. *Blood*. 2015
8. Hayes SA, Hudson AL, Clarke SJ, Molloy MP, Howell VM. From mice to men: Gemms as trial patients for new nscL therapies. *Semin Cell Dev Biol*. 2014; 27:118–127. [PubMed: 24718320]
9. Morton JP, Sansom OJ. Myc-y mice: From tumour initiation to therapeutic targeting of endogenous MYC. *Mol Oncol*. 2013; 7:248–258. [PubMed: 23523308]
10. Koren S, Bentires-Alj M. Mouse models of pik3ca mutations: One mutation initiates heterogeneous mammary tumors. *FEBS J*. 2013; 280:2758–2765. [PubMed: 23384338]
11. Kirma NB, Tekmal RR. Transgenic mouse models of hormonal mammary carcinogenesis: Advantages and limitations. *J Steroid Biochem Mol Biol*. 2012; 131:76–82. [PubMed: 22119744]
12. Millington C, Sonogo S, Karunaweera N, Rangel A, Aldrich-Wright JR, Campbell IL, Gyengesi E, Munch G. Chronic neuroinflammation in Alzheimer's disease: New perspectives on animal models and promising candidate drugs. *Biomed Res Int*. 2014; 2014:309129. [PubMed: 25025046]
13. Ford Siltz LA, Viktorova EG, Zhang B, Kouliavskaya D, Dragunsky E, Chumakov K, Isaacs L, Belov GA. New small-molecule inhibitors effectively blocking picornavirus replication. *J Virol*. 2014; 88:11091–11107. [PubMed: 25008939]
14. Charbogne P, Kieffer BL, Befort K. 15 years of genetic approaches in vivo for addiction research: Opioid receptor and peptide gene knockout in mouse models of drug abuse. *Neuropharmacology*. 2014; 76(Pt B):204–217. [PubMed: 24035914]
15. Cachat A, Villaudy J, Rigal D, Gazzolo L, Duc Dodon M. Mice are not Men and yet... How humanized mice inform us about human infectious diseases. *Med Sci (Paris)*. 2012; 28:63–68. [PubMed: 22289832]
16. Paine MF, Khalighi M, Fisher JM, Shen DD, Kunze KL, Marsh CL, Perkins JD, Thummel KE. Characterization of interintestinal and intrainestinal variations in human CYP3A-dependent metabolism. *J Pharmacol Exp Ther*. 1997; 283:1552–1562. [PubMed: 9400033]
17. Afzelius L, Arnby CH, Broo A, Carlsson L, Isaksson C, Jurva U, Kjellander B, Kolmodin K, Nilsson K, Raubacher F, Weidolf L. State-of-the-art tools for computational site of metabolism predictions: Comparative analysis, mechanistical insights, and future applications. *Drug Metab Rev*. 2007; 39:61–86. [PubMed: 17364881]
18. Jolivet LJ, Ekins S. Methods for predicting human drug metabolism. *Adv Clin Chem*. 2007; 43:131–176. [PubMed: 17249383]
19. Williams JA, Hyland R, Jones BC, Smith DA, Hurst S, Goosen TC, Peterkin V, Koup JR, Ball SE. Drug-drug interactions for udp-glucuronosyltransferase substrates: A pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab Dispos*. 2004; 32:1201–1208. [PubMed: 15304429]
20. Quintieri L, Fantin M, Palatini P, De Martin S, Rosato A, Caruso M, Geroni C, Floreani M. In vitro hepatic conversion of the anticancer agent nemorubicin to its active metabolite PNU-159682 in mice, rats and dogs: A comparison with human liver microsomes. *Biochem Pharmacol*. 2008; 76:784–795. [PubMed: 18671948]

21. Palmer BD, Thompson AM, Sutherland HS, Blaser A, Kmentova I, Franzblau SG, Wan B, Wang Y, Ma Z, Denny WA. 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; 53:282–294. [PubMed: 19928920]
22. Crivori P, Poggesi I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur J Med Chem*. 2006; 41:795–808. [PubMed: 16644065]
23. Stjernschantz E, Vermeulen NP, Oostenbrink C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Expert Opin Drug Metab Toxicol*. 2008; 4:513–527. [PubMed: 18484912]
24. Hansch C. Quantitative relationships between lipophilic character and drug metabolism. *Drug Metabolism Reviews*. 1972; 1:1–13.
25. Hansch C. The qsar paradigm in the design of less toxic molecules. *Drug Metab Rev*. 1984; 15:1279–1294. [PubMed: 6398775]
26. Hansch C, Lien EJ, Helmer F. Structure--activity correlations in the metabolism of drugs. *Arch Biochem Biophys*. 1968; 128:319–330. [PubMed: 5698027]
27. Hansch C, Zhang L. Quantitative structure-activity relationships of cytochrome P-450. *Drug Metab Rev*. 1993; 25:1–48. [PubMed: 8449144]
28. Lewis DF. Quantitative structure-activity relationships in substrates, inducers, and inhibitors of cytochrome P4501 (CYP1). *Drug Metab Rev*. 1997; 29:589–650. [PubMed: 9262943]
29. Lewis DF. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: Towards the prediction of human P450 substrate specificity and metabolism. *Biochem Pharmacol*. 2000; 60:293–306. [PubMed: 10856424]
30. Lewis DF. Structural characteristics of human P450s involved in drug metabolism: QSARs and lipophilicity profiles. *Toxicology*. 2000; 144:197–203. [PubMed: 10781888]
31. Lewis DF, Eddershaw PJ, Dickins M, Tarbit MH, Goldfarb PS. Structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate. *Chem Biol Interact*. 1998; 115:175–199. [PubMed: 9851289]
32. Lewis DF, Eddershaw PJ, Dickins M, Tarbit MH, Goldfarb PS. Erratum to structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate. *Chemico Biol Interact*. 1999; 117:187.
33. Lewis DF, Jacobs MN, Dickins M. Compound lipophilicity for substrate binding to human P450s in drug metabolism. *Drug Discov Today*. 2004; 9:530–537. [PubMed: 15183161]
34. Fuhr U, Strobl G, Manaut F, Anders EM, Sorgel F, Lopez-de-Brinas E, Chu DT, Pernet AG, Mahr G, Sanz F, et al. Quinolone antibacterial agents: Relationship between structure and in vitro inhibition of the human cytochrome P450 isoform CYP1A2. *Mol Pharmacol*. 1993; 43:191–199. [PubMed: 8429824]
35. Jones JP, Korzekwa KR. Predicting the rates and regioselectivity of reactions mediated by the p450 superfamily. *Methods Enzymol*. 1996; 272:326–335. [PubMed: 8791792]
36. Jones JP, Korzekwa KR. Predicting intrinsic clearance for drugs and drug candidates metabolized by aldehyde oxidase. *Mol Pharm*. 2013; 10:1262–1268. [PubMed: 23363487]
37. Locuson CW, Wahlstrom JL. Three-dimensional quantitative structure-activity relationship analysis of cytochromes P450: Effect of incorporating higher-affinity ligands and potential new applications. *Drug Metab Dispos*. 2005; 33:873–878. [PubMed: 15843486]
38. Sorich MJ, McKinnon RA, Miners JO, Winkler DA, Smith PA. Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J Med Chem*. 2004; 47:5311–5317. [PubMed: 15456275]
39. Dajani R, Cleasby A, Neu M, Wonacott AJ, Jhoti H, Hood AM, Modi S, Hersey A, Taskinen J, Cooke RM, Manchee GR, Coughtrie MW. X-ray crystal structure of human dopamine sulfotransferase, SULT1A3. Molecular modeling and quantitative structure-activity relationship analysis demonstrate a molecular basis for sulfotransferase substrate specificity. *J Biol Chem*. 1999; 274:37862–37868. [PubMed: 10608851]
40. Ekins S. In silico approaches to predicting drug metabolism, toxicology and beyond. *Biochem Soc Trans*. 2003; 31:611–614. [PubMed: 12773166]

41. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem.* 2003; 46:3013–3020. [PubMed: 12825940]
42. Jensen BF, Sorensen MD, Kissmeyer AM, Bjorkling F, Sonne K, Engelsen SB, Norgaard L. Prediction of in vitro metabolic stability of calcitriol analogs by QSAR. *J Comput Aided Mol Des.* 2003; 17:849–859. [PubMed: 15124933]
43. Chang C, Duignan DB, Johnson KD, Lee PH, Cowan GS, Gifford EM, Stankovic CJ, Lepsey CS, Stoner CL. The development and validation of a computational model to predict rat liver microsomal clearance. *J Pharm Sci.* 2009; 98:2857–2867. [PubMed: 19116953]
44. Hu Y, Unwalla R, Denny RA, Bikker J, Di L, Humblet C. Development of qsar models for microsomal stability: Identification of good and bad structural features for rat, human and mouse microsomal stability. *J Comput Aided Mol Des.* 2010; 24:23–35. [PubMed: 19937264]
45. Gupta RR, Gifford EM, Liston T, Waller CL, Hohman M, Bunin BA, Ekins S. Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties. *Drug Metabolism and Disposition.* 2010; 38:2083–2090. [PubMed: 20693417]
46. Clark AM, Dole K, Coulon-Spektor A, McNutt A, Grass G, Freundlich JS, Reynolds RC, Ekins S. Open source bayesian models. 1. Application to ADME/Tox and drug discovery datasets. *J Chem Inf Model.* 2015
47. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. Pubchem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 2009; 37:W623–633. [PubMed: 19498078]
48. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. Pubchem's bioassay database. *Nucleic Acids Res.* 2012; 40:D400–412. [PubMed: 22140110]
49. Lee PH, Cucurull-Sanchez L, Lu J, Du YJ. Development of in silico models for human liver microsomal stability. *J Comput Aided Mol Des.* 2007; 21:665–673. [PubMed: 17599241]
50. BIOVIA. Discovery Studio modeling environment. San Diego, CA: BIOVIA; 2013.
51. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Kruger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP. The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* 2014; 42:D1083–1090. [PubMed: 24214965]
52. Ekins S.; Freundlich, J. Computational models for tuberculosis drug discovery. In: Kortagere, S., editor. *In silico models for drug discovery.* Humana Press; 2013. p. 245-262.
53. Ekins S, Freundlich J. Validating new tuberculosis computational models with public whole cell screening aerobic activity datasets. *Pharm Res.* 2011; 28:1859–1869. [PubMed: 21547522]
54. Ekins S, Freundlich J, Hobrath J, Lucile White E, Reynolds R. Combining computational methods for hit to lead optimization in Mycobacterium tuberculosis drug discovery. *Pharm Res.* 2013; 1–22. [PubMed: 23179779]
55. Ekins S, Reynolds RC, Franzblau SG, Wan B, Freundlich JS, Bunin BA. Enhancing hit identification in Mycobacterium tuberculosis drug discovery using validated dual-event bayesian models. *PLoS ONE.* 2013; 8:e63240. [PubMed: 23667592]
56. Ekins S, Reynolds RC, Kim H, Koo M-S, Ekonomidis M, Talaue M, Paget SD, Woolhiser LK, Lenaerts AJ, Bunin BA, Connell N, Freundlich JS. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. *Chemistry & biology.* 2013; 20:370–378. [PubMed: 23521795]
57. Langdon SR, Mulgrew J, Paolini GV, van Hoorn WP. Predicting cytotoxicity from heterogeneous data sources with Bayesian learning. *Journal of Cheminformatics.* 2010; 2.
58. Prathipati P, Ma NL, Keller TH. Global Bayesian models for the prioritization of antitubercular agents. *Journal of Chemical Information and Modeling.* 2008; 48:2362–2370. [PubMed: 19053518]
59. Ekins S, Williams AJ, Xu JJ. A predictive ligand-based bayesian model for human drug-induced liver injury. *Drug Metabolism and Disposition.* 2010; 38:2302–2308. [PubMed: 20843939]

60. Ekins S, Freundlich JS, Hobarth JV, Lucile White E, Reynolds RC. Combining computational methods for hit to lead optimization in Mycobacterium tuberculosis drug discovery. *Pharm Res*. 2014; 31:414–435. [PubMed: 24132686]
61. Langdon SR, Mulgrew J, Paolini GV, van Hoorn WP. Predicting cytotoxicity from heterogeneous data sources with Bayesian learning. *J Cheminform*. 2010; 2:11. [PubMed: 21143909]
62. Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Hohman M, Bunin BA. A collaborative database and computational models for tuberculosis drug discovery. *Mol Biosyst*. 2010; 6:840–851. [PubMed: 20567770]
63. Ekins S, Kaneko T, Lipinski CA, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Ernst S, Yang J, Goncharoff N, Hohman MM, Bunin BA. Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis. *Mol Biosyst*. 2010; 6:2316–2324. [PubMed: 20835433]
64. Ekins S. Progress in computational toxicology. *J Pharmacol Toxicol Methods*. 2013; 69:115–140. [PubMed: 24361690]
65. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *Chem Med Chem*. 2007; 2:861–873. [PubMed: 17477341]
66. Klon AE, Lowrie JF, Diller DJ. Improved naïve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model*. 2006; 46:1945–1956. [PubMed: 16995725]
67. Hassan M, Brown RD, Varma-O'Brien S, Rogers D. Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers*. 2006; 10:283–299. [PubMed: 17031533]
68. Rogers D, Brown RD, Hahn M. Using extended-connectivity fingerprints with laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen*. 2005; 10:682–686. [PubMed: 16170046]
69. Ekins S, Freundlich JS, Reynolds RC. 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. [PubMed: 24968215]
70. Xia X, Maliski EG, Gallant P, Rogers D. Classification of kinase inhibitors using a Bayesian model. *J Med Chem*. 2004; 47:4463–4470. [PubMed: 15317458]
71. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov Today*. 2009; 14:261–270. [PubMed: 19231313]
72. Lakshminarayana SB, Huat TB, Ho PC, Manjunatha UH, Dartois V, Dick T, Rao SP. Comprehensive physicochemical, pharmacokinetic and activity profiling of anti-tb agents. *J Antimicrob Chemother*. 2015; 70:857–867. [PubMed: 25587994]
73. Ekins S, Freundlich JS, Reynolds RC. Fusing dual-event data sets for mycobacterium tuberculosis machine learning models and their evaluation. *Journal of Chemical Information and Modeling*. 2013; 53:3054–3063. [PubMed: 24144044]
74. Perryman AL, Yu W, Wang X, Ekins S, Forli S, Li SG, Freundlich JS, Tonge PJ, Olson AJ. A virtual screen discovers novel, fragment-sized inhibitors of Mycobacterium tuberculosis InhA. *J Chem Inf Model*. 2015
75. Jones DR, Ekins S, Li L, Hall SD. Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). *Drug Metab Dispos*. 2007; 35:1466–1475. [PubMed: 17537872]
76. Ekins S, Williams AJ, Krasowski MD, Freundlich JS. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today*. 2011; 16:298–310. [PubMed: 21376136]
77. Anderson JW, Sarantakis D, Terpins J, Santha Kumar TR, Tsai H-C, Kuo M, Ager AL, Jacobs WR Jr, Schiehsler GA, Ekins S, Sacchettini JC, Jacobus DP, Fidock DA, Freundlich JS. Novel diaryl ureas with efficacy in a mouse model of malaria. *Bioorganic & Medicinal Chemistry Letters*. 2013; 23:1022–1025. [PubMed: 23313245]
78. Clark AM, Ekins S. Open source Bayesian models. 2. Mining a “big dataset” to create and validate models with ChEMBL. *J Chem Inf Model*. 2015

79. Clark AM, Sarker M, Ekins S. New target prediction and visualization tools incorporating open source molecular fingerprints for tb mobile 2. 0. *Journal of Cheminformatics*. 2014; 6:38–54.
80. Anon R. R: The R project for statistical computing. 2014. <http://www.r-project.org>
81. Di L, Kerns EH, Hong Y, Kleintop TA, McConnell OJ, Huryn DM. Optimization of a higher throughput microsomal stability screening assay for profiling drug discovery candidates. *J Biomol Screen*. 2003; 8:453–462. [PubMed: 14567798]
82. Lombardo F, Obach RS, Dicapua FM, Bakken GA, Lu J, Potter DM, Gao F, Miller MD, Zhang Y. A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human. *J Med Chem*. 2006; 49:2262–2267. [PubMed: 16570922]
83. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J Med Chem*. 2004; 47:1242–1250. [PubMed: 14971904]
84. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data. *J Med Chem*. 2002; 45:2867–2876. [PubMed: 12061889]
85. Lombardo F, Shalaeva MY, Tupper KA, Gao F. ElogD(oct): A tool for lipophilicity determination in drug discovery. 2. Basic and neutral compounds. *J Med Chem*. 2001; 44:2490–2497. [PubMed: 11448232]
86. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012; 40:D1100–1107. [PubMed: 21948594]
87. Papadatos G, Overington JP. The ChEMBL database: A taster for medicinal chemists. *Future Med Chem*. 2014; 6:361–364. [PubMed: 24635517]
88. Sun H, Veith H, Xia M, Austin CP, Tice RR, Huang R. Prediction of cytochrome P450 profiles of environmental chemicals with QSAR models built from drug-like molecules. *Mol Inform*. 2012; 31:783–792. [PubMed: 23459712]
89. Sun H, Veith H, Xia M, Austin CP, Huang R. Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. *J Chem Inf Model*. 2011; 51:2474–2481. [PubMed: 21905670]
90. Veith H, Southall N, Huang R, James T, Fayne D, Artemenko N, Shen M, Inglese J, Austin CP, Lloyd DG, Auld DS. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat Biotechnol*. 2009; 27:1050–1055. [PubMed: 19855396]
91. MacArthur R, Leister W, Veith H, Shinn P, Southall N, Austin CP, Inglese J, Auld DS. Monitoring compound integrity with cytochrome P450 assays and qHTS. *J Biomol Screen*. 2009; 14:538–546. [PubMed: 19483146]
92. Litterman NK, Lipinski CA, Bunin BA, Ekins S. Computational prediction and validation of an expert's evaluation of chemical probes. *J Chem Inf Model*. 2014; 54:2996–3004. [PubMed: 25244007]
93. Dong Z, Ekins S, Polli JE. Structure-activity relationship for FDA approved drugs as inhibitors of the human sodium taurocholate cotransporting polypeptide (NTCP). *Mol Pharm*. 2013; 10:1008–1019. [PubMed: 23339484]
94. Ekins S, Freundlich JS, Hobrath JV, Lucile White E, Reynolds RC. Combining computational methods for hit to lead optimization in Mycobacterium tuberculosis drug discovery. *Pharm Res*. 2014; 31:414–435. [PubMed: 24132686]
95. Ekins, S.; Embrechts, MJ.; Breneman, CM.; Jim, K.; Wery, J-P. Novel applications of kernel-partial least squares to modeling a comprehensive array of properties for drug discovery. In: Ekins, S., editor. *Computational toxicology: Risk assessment for pharmaceutical and environmental chemicals*. Hoboken, NJ: John Wiley and Sons; 2007. p. 403-432.

Figure 1. Workflow used to build, test, and validate the Bayesian models for predicting mouse liver microsomal stability

The PubChem BioAssay database was queried for data on the stability that different small molecule compounds displayed in the presence of mouse liver microsomes (MLM). The summaries and tables of results on PubChem were investigated, to determine which assay results and structural data files (sdf) to obtain. The cited publication for each set of MLM half-life results was examined, to ensure that the primary literature presented the same set of half-life values with the same units as the comma separated value (csv) file that was obtained from PubChem. This verification stage revealed several mistakes in the csv files, which were manually corrected before the MLM half-life data were entered into a spreadsheet that Discovery Studio 4.0 (BIOVIA) created from the sdf files retrieved from PubChem. Only 99 of the 103 sets of assay results were utilized, since three assays just reported $t_{1/2}$ values of >40 min (which was too ambiguous), and one assay was performed without NADPH (which means that Phase I metabolism was not measured). The full half-life set (A1) of MLM data contained 262 compounds (29.3%) with a $t_{1/2} \geq 60$ min that we defined as being "stable" in MLM. It contained 632 (70.7%) with a $t_{1/2} < 60$ min, which we defined as "unstable." The pruned half-life set (A2) included the same 262 "stable" compounds (now 34.5%), but it only contained the 497 (65.5%) "unstable" compounds that had a $t_{1/2} < 30$ min. (*i.e.*, compounds with $30 \leq t_{1/2} < 60$ min were removed). Two Bayesian classifier models were built in Pipeline Pilot 9.1 (BIOVIA)—one using the full half-life set and the other using the pruned half-life set. Both Bayesian models were then tested with the Dartois 2015 set of 30 antitubercular compounds, followed by a validation study with the independent set of MLM stability data that were reported on PubChem as the % compound remaining after incubation in the presence of MLM (see Figure S1).

A

B

C

D

Figure 2. Distribution of Bayesian scores produced when the half-life MLM stability models were used to score the full % compound left validation set

(A) The score produced by the pruned half-life Bayesian model constructed using 9 descriptors is plotted versus the percentage of experimentally stable (green diamonds) or unstable compounds (red circles) which received that score. A majority of the compounds known to be stable received positive (favorable) pruned $t_{1/2}$ Bayesian scores. Similarly, most of the compounds known to be unstable in MLM received negative (unfavorable) Bayesian scores when evaluated by the pruned half-life MLM model. (B) The corresponding histograms produced with the 9-descriptor-based full half-life Bayesian model, displaying experimentally stable (green squares) vs. unstable compounds (red circles). (C) The histograms of the Bayesian scores for the experimentally stable (green diamonds) vs. unstable compounds (red circles), as calculated by the pruned half-life Bayesian that was constructed using only FCFP_6. (D) The corresponding histograms for the FCFP_6-only,

full half-life Bayesian, with stable compounds (green squares) versus unstable compounds (red circles).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Structural features associated with MLM stability, as identified by the pruned half-life Bayesian model created in Pipeline Pilot

The 2D structural features that were most frequently associated with compounds that displayed MLM stability are displayed. Of the three types of machine learning models investigated, only Bayesian models output these structural guidelines. These features of stability were identified by the pruned half-life Bayesian model that was created in Pipeline Pilot using all 9 descriptors (including FCFP_6). The * symbols are wild cards (which can represent any atom type). Aromatic rings are drawn with the pi electrons in a delocalized, dashed fashion (instead of rendering them as alternating single and double bonds).

Figure 4. Structural characteristics associated with instability in MLM, as identified by the pruned half-life Bayesian model created in Pipeline Pilot

The 2D structural features that were most frequently associated with compounds that displayed MLM instability are displayed. These instability liabilities were identified by the pruned half-life Bayesian model that was created in Pipeline Pilot using all 9 descriptors (including FCFP_6). The * symbols are wild cards (which can represent any atom type). Aromatic rings are drawn with the pi electrons in a delocalized, dashed fashion (instead of rendering them as alternating single and double bonds).

Internal statistics from five-fold cross-validation studies performed when creating different machine learning models to predict MLM stability in Pipeline Pilot, using either 9 descriptors or just 1 (FCFP_6).

Table 1

Half-Life Bayesian	ROC score	ROC rating ^c	Sensitivity %	Specificity %	Concordance %
Full t _{1/2} with 9 ^a	0.835	good	92.7	72.2	78.2
Pruned t _{1/2} with 9	0.870	good	93.1	80.9	85.1
Full t _{1/2} with 1 ^b	0.835	good	93.1	71.4	77.7
Pruned t _{1/2} with 1	0.869	good	93.9	76.3	82.3

Notes:

^a“With 9” indicates that all 9 descriptors were utilized when creating that Bayesian model, while

^b“with 1” means that only the FCFP_6 fingerprints that describe 2D topology were used.

^cThe “ROC rating” is a qualitative grading system that Pipeline Pilot outputs, which ranges from: fail < poor < fair < good < excellent.

Table II

Comparing the half-life Bayesian models, in terms of the enrichment factors for the compounds in the full % compound left validation set that received the top Bayesian scores.

Half-Life Bayesian	# of True Positives in Top-Scoring Compounds ^a	% of True Positives in Top-Scoring 30 or 50 Compounds	Enrichment Factor vs. Random ^b
Full t _{1/2} with 9	21/30 33/50 (or 32/50)	70.0% 66.0% (or 64.0%)	3.66 3.46 (or 3.35)
Pruned t _{1/2} with 9	20/30 31/50 (still 31/50)	66.7% 62.0% (still 62.0%)	3.49 3.25 (still 3.25)
Full t _{1/2} with 1	24/30 41/50 (still 41/50)	80.0% 82.0% (still 82.0%)	4.19 4.29 (still 4.29)
Pruned t _{1/2} with 1	24/30 38/50 (still 38/50)	80.0% 76.0% (still 76.0%)	4.19 3.98 (still 3.98)

Notes:

^a Eight or nine compounds were present in the pruned and full half-life sets, respectively, and the full percent compound left set. Only one of these duplicate compounds was present in the 50 compounds from the percent compound left set that received the top half-life Bayesian scores. If that duplicate compound is removed, then the compound with the 51st top score becomes the last member of the top 50. The values in parentheses reflect how removing the duplicate affects the stability hit rates and enrichment factors for these different Bayesian models.

^b The Enrichment Factors were calculated by dividing the % of true positives in the top-scoring compounds by 19.1%. Since 19.1% of the compounds in the full % compound left set were stable, this value reflects the random chance of selecting a stable compound from that set. An Enrichment Factor of 1 represents no improvement over random chance.

Table III

Comparing the internal and external test and validation statistics for the full and pruned half-life Bayesian models constructed with CDD.

CDD Bayesian	ROC score	Sensitivity % ^a	Specificity %	Concordance %	Stability Hit Rate ^b
Full $t_{1/2}$ internal 3-fold cross-validation	0.83	80 90	72 58		
Pruned $t_{1/2}$ internal 3-fold cross-validation	0.86	80 90	76 66		
Full $t_{1/2}$ vs. Dartois 2015 test set		63 96	67 33	67 90	17/17 26/28
Pruned $t_{1/2}$ vs. Dartois 2015 test set		63 96	67 33	67 90	17/17 26/28
Full $t_{1/2}$ vs. % compound left validation set		50 89	76 19	71 33	54/163 (33%) 97/470 (21%)
Pruned $t_{1/2}$ vs. % compound left validation set		49 86	76 21	71 34	53/165 (32%) 94/458 (21%)

Notes: the internal concordance values are not an output of the CDD Bayesian models, but it was possible to manually calculate the external concordance values for the test and validation sets.

^aTwo sets of values are listed for the CDD Bayesians results with the Dartois 2015 set, because two different cut-off values were used to score these known antituberculars (*i.e.*, the Bayesian score cut-off values that produced internal sensitivity values of 0.8 and 0.9 were used, respectively).

^bThe “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives.