Global Bayesian Models for the Prioritization of Antitubercular Agents

Philip Prathipati, Ngai Ling Ma,* and Thomas H. Keller

Novartis Institute for Tropical Diseases, 10 Biopolis Road, #05-01 Chromos 138670, Singapore

Received April 25, 2008

To aid the creation of novel antituberculosis (antiTB) compounds, Bayesian models were derived and validated on a data set of 3779 compounds which have been measured for minimum inhibitory concentration (MIC) in the *Mycobacterium tuberculosis* H37Rv strain. The model development and validation involved exploring six different training sets and 15 fingerprint types which resulted in a total of 90 models, with active compounds defined as those with MIC $< 5\ \mu$M. The best model was derived using Extended Class Fingerprints of maximum diameter 12 (ECFP_12) and a few global descriptors on a training set derived using Functional Class Fingerprints of maximum diameter 4 (FCFP_4). This model demonstrated very good discriminant ability in general, with excellent discriminant statistics for the training set (total accuracy: 0.968; positive recall: 0.967) and a good predictive ability for the test set (total accuracy: 0.869; positive recall: 0.789). The good predictive ability was maintained when the model was applied to a well-separated test set of 2880 compounds derived from a commercial database (total accuracy: 0.73; positive recall: 0.72). The model revealed several conserved substructures present in the active and inactive compounds which are believed to have incremental and detrimental effects on the MIC, respectively. Strategies for enhancing the repertoire of antiTB compounds with the model, including virtual screening of large databases and combinatorial library design, are proposed.

INTRODUCTION

Tuberculosis, a common and deadly infectious disease caused by *Mycobacterium tuberculosis* (M Tb), is primarily an illness of the respiratory system which is spread by coughing and sneezing. Among one-third of the world's population infected with M Tb, there are about 8 million new cases and 3.1 million deaths reported annually.[1−7] According to recent estimates, around 50 million people are infected with drug-resistant forms of tuberculosis (TB).[6,7] Despite several decades of research, the outlook for new TB drugs is bleak, as antituberculosis (antiTB) drug discovery suffers from many issues. Apart from efficacy, toxicity, adverse drug reactions, and multidrug resistance, there is a profound ignorance in the mode of action of these agents.[1−7] Thus, novel tools are needed to facilitate the optimization of existing antiTB compounds and the identification of new compound classes.

The activity of a drug in a cell-based assay depends not only on the affinity to its target but also on the sequence of events that effect the accumulation at its site of action. Thus, the SAR (structure−activity relationship) of the compounds obtained in a cell-based assay, with diverse modes of action and pharmacophoric patterns, often require computational methods that can deal with such complexities.[8] In view of the gaps in our knowledge of the M Tb target space in general and the mode of action of compounds in particular, the application of a direct target-based drug design or a systems biology approach to explain the variation in a cell-based assay is presently very difficult.[9−11] Hence, indirect approaches like quantitative structure−activity relationships (QSAR), which can explain the variation in the activities of a diverse class of compounds, are highly desirable.[8]

The importance of modeling in antiTB drug discovery efforts has been revealed by numerous QSAR, pharmacophore, and/or docking studies. Most of the QSAR studies are local models which focus on specific classes of antiTB compounds,[12] with only three studies reporting global QSAR models to explain the variation of minimum inhibitory concentration (MIC) for a range of compound classes.[13−15] *García-García* et al. reported two discriminant analysis models (derived from 71 and 60 compounds, respectively) and a quantitative MIC model based on 45 compounds.[13] With these models, a database of 5000 commercial compounds was screened, in which five virtual hits were confirmed to have MIC below 50 μM.[13] The hologram QSAR study by *Prakash* and *Ghosh* employed a training set of 120 compounds (including 18 standard antiTB compounds) and reported five models derived on compounds from five different clusters.[14] The classification QSAR model reported by *Manetti* et al. is the most exhaustive reported so far, with the model derived from 471 compounds belonging to more than 10 classes. With this model, two compounds with MIC of 25 μg L^{-1} have been identified.[15]

Global QSAR models derived using many classes of compounds, though difficult to derive, interpret, and validate, overcome many idiosyncrasy issues which local (mode of action based) models fail to address.[16,17] Among the existing QSAR approaches that are suited to the development of global models, Bayesian models are considered to be one of the best, as they are robust, faster, more reproducible, and amenable to visual interpretation than other techniques.[18−21] With a substantially larger data set of antiTB compounds (3779) than all the three previous global QSAR studies,[13−15]

* Corresponding author phone: (65) 6722-2932; fax: (65)6722-2910; e-mail: ida.ma@novartis.com.

10.1021/ci800143n CCC: $40.75
Published on Web 11/17/2008

efforts were undertaken in the present study to capture the structure-MIC data with Bayesian technique.

METHODS

Data Set. A data set of ~4K antiTB compounds with National Institute of Allergy and Infectious Diseases (NIAID) ID and various M Tb H37Rv activities were collected from the NIAID Web site.[22] The structures and other annotations including synonyms (which include class information supplied by the depositors) were added via files downloaded from Pubchem.[23] Within this data set, a subset of 3779 compounds having MIC values (ranging from 0.00000316 to 4094 μM i.e. eight-order of magnitude, Supporting Information Table S1), was used in our current study. Among these, 1886 compounds were considered to be "active" in our study as their reported MIC were below 5 μM. Such a cutoff value appeared to be a reasonable starting point for hit-to-lead activity, and, in view of the noise level in the data set, the choice of 5 μM would seem justified.

Issues in the Development of Global QSAR Models. The predictive ability of any mathematical model is generally dependent on three aspects: training set design,[24] choice of descriptors/fingerprints,[25−28] and choice of modeling/statistical method.[29,30] While training set design and choice of descriptors is highly subjective, the choice of modeling methods can be decided upon by the quality (noise) and quantity of the activity data. The present data set with 3779 structure-MIC data points was curated by the NIAID from different sources like journals, patents, meeting reports, etc. and may be noisy in many ways (e.g., assay variability, interlab assay variability, and omissions/commissions during the curation of the database). In addition, it contains a wide variety of chemical classes apparently acting via diverse modes of action. Thus, in view of its large size, structural diversity, and "noise" in the activity data, methods like Bayesian modeling are the most appropriate for this data set.[18−20]

Bayesian Models in Pipeline Pilot. Bayesian analysis and model building were implemented using the Scitegic Pipeline Pilot Laplacian-corrected Bayesian classifier algorithm.[18−21] This implementation of Bayesian statistics uses information from both the active ("good") and inactive ("bad") compounds in the training set and removes features from the model which are deemed not to be important.

Fingerprints for Training Set Design. Diverse training sets of 1260 compounds (approximately 1/3rd of the data set) were selected using hierarchical cluster analysis with two classes of fingerprints: (1) structural fingerprints: Scitegic Functional Class Fingerprints of maximum diameter 4 (FCFP_4)[21] and Sybyl's molecular hologram fingerprint of length 401 (H401)[31] and (2) text fingerprints: functional group fingerprints (FG),[32] domain fingerprints (DM),[33,34] target fingerprints (TG),[35] and a combination of functional group and target fingerprints (FG-TG). For the hologram fingerprints, the hierarchical cluster analysis module implemented in Sybyl software[36] was used, while for the other five fingerprints, the relocation method based on the maximal dissimilarity partitioning implemented as the cluster data component in Pipeline Pilot[21] was used.

These different fingerprints were chosen as representatives for the various fingerprint classes,[38] with the aim to assess which type of fingerprint is more appropriate for describing the pharmacophoric similarity of antiTB compounds with cellular activities. The discussion of FCFP has been deferred to the section "Fingerprints Used for the Development of the Bayesian Models". Molecular hologram fingerprints are extensions of 2D fingerprints, which contain all possible molecular fragments within a molecule and the number of times each unique fragment occurs.[31] The functional group FG fingerprints, also known as chemical ontology annotations, are based on chemical functional groups assignment with the program *checkmol*.[37] The domain DM and target TG fingerprints were recently developed in our group in attempting to map the chemical space to the TB biology space.[35] In brief, DM fingerprints are the five most probable (based on normalized probability scores) predicted Interpro domain[39] a compound is associated with. These probabilities are based on predictions from Bayesian models trained to capture the ligand-target associations from the WOMBAT database.[33,34] The predicted domain associations were then extrapolated to M Tb targets to form the TG fingerprint, using the Interpro domain-M Tb protein association information.[35]

Fingerprints Used for the Development of the Bayesian Models. In all, 15 fingerprints of various maximum diameters viz. 4, 6, 8, 10, and 12 and belonging to three fingerprint types available in Pipeline Pilot i.e. Extended-Connectivity Fingerprints (ECFP), Functional Class Fingerprints (FCFP), and Sybyl atom types Extended-Connectivity Fingerprints (SCFP) were used for Bayesian model development.[21] All these fingerprints belong to the circular substructure class of fingerprints. Circular substructure is a fragment descriptor where each atom is represented by a string of extended connectivity values that are calculated up to a specified diameter of bonds, using a modified Morgan algorithm, and these fingerprints differ from each other with respect to atom abstraction or representation.[38,40,41]

RESULTS AND DISCUSSIONS

Measures of Discriminative Ability. A known problem in the validation of QSARs is that none of the statistical parameters measured on the training set on their own correlate with the predictive ability against external test sets, which is one of the ultimate measures of the usefulness of a model for virtual screening.[42−44] To overcome this, several parameters (Table 1) were estimated for the training and the test sets, and the averages of these parameters were also determined. Based on the sum of average total accuracy (of the training and test set) and average positive recall (of the training and test set), the best model (highest sum) was chosen.

Training Set Design for Data Sets with Cellular Activity. Training set design is an experiment to identify the most diverse compounds from a data set for QSAR model development.[24,43] Typically, the sources of SAR data used in a QSAR study are either from isolated proteins or from cellular systems. SAR data from isolated proteins contain ligands that are associated with a target and often share a common pharmacophore. Hence, for this type of study, it is sufficient to assess the ligand diversity in terms of its structural fingerprints or descriptors. However, for ligands identified in a cell-based assay (as in the present study), diversity should ideally be measured using their biologically

Table 1. Comparison of the Statistics of the Best Model with All 90 Derived Models

	total accuracy[a]	positive recall[b]	negative recall[c]	precision[d]	EF[e]
Training Set					
all[f]	0.812−0.968	0.738−0.967	0.783−0.985	0.726−0.978	1.668−2.504
best[g]	0.968	0.967	0.969	0.951	2.491
best (n = 1890)[h]	0.924	0.887	0.960	0.954	1.965
best (n = 2520)[i]	0.838	0.806	0.875	0.878	1.666
Test Set					
all[f]	0.414−0.790	0.387−0.827	0.313−0.932	0.652−0.826	0.879−1.829
best[g]	0.784	0.797	0.769	0.794	1.507
best (n = 1890)[h]	0.758	0.712	0.798	0.758	0.471
best (n = 2520)[i]	0.756	0.690	0.816	0.771	1.629
Cross-Validation					
all[f]	0.500−0.810	0.561−0.784	0.292−0.920	0.500−0.910	1.000−1.984
best[g]	0.723	0.692	0.742	0.624	1.634

[a] *Total accuracy* = $(tp + tn)/(tp + fp + tn + fn)$. [b] *Positive recall* = $(tp)/(tp + fn)$. [c] *Negative recall* = $(tn)/(tp + fn)$. [d] *Precision* = $(tp)/(tp + fp)$. [e] *EF* = $(precision)/((tp + fn)/(tp + fp + tn + fn))$; with *tp*, *tn*, *fp*, and *fn* being the number of true positives, true negatives, false positives, and false negatives, respectively. [f] The range of the parameter for the 90 models. [g] The best model (obtained from diversity analysis of FCFP_4 fingerprints and derived using the ECFP_12 fingerprints, ALogP, molecular weight, number of hydrogen bond donor and acceptor, number of rotatable bonds and molecular fractional polar surface area), with training set size of 1260. [h] The best model, with training set size, n = 1890. [i] The best model, with training set size, n = 2520.

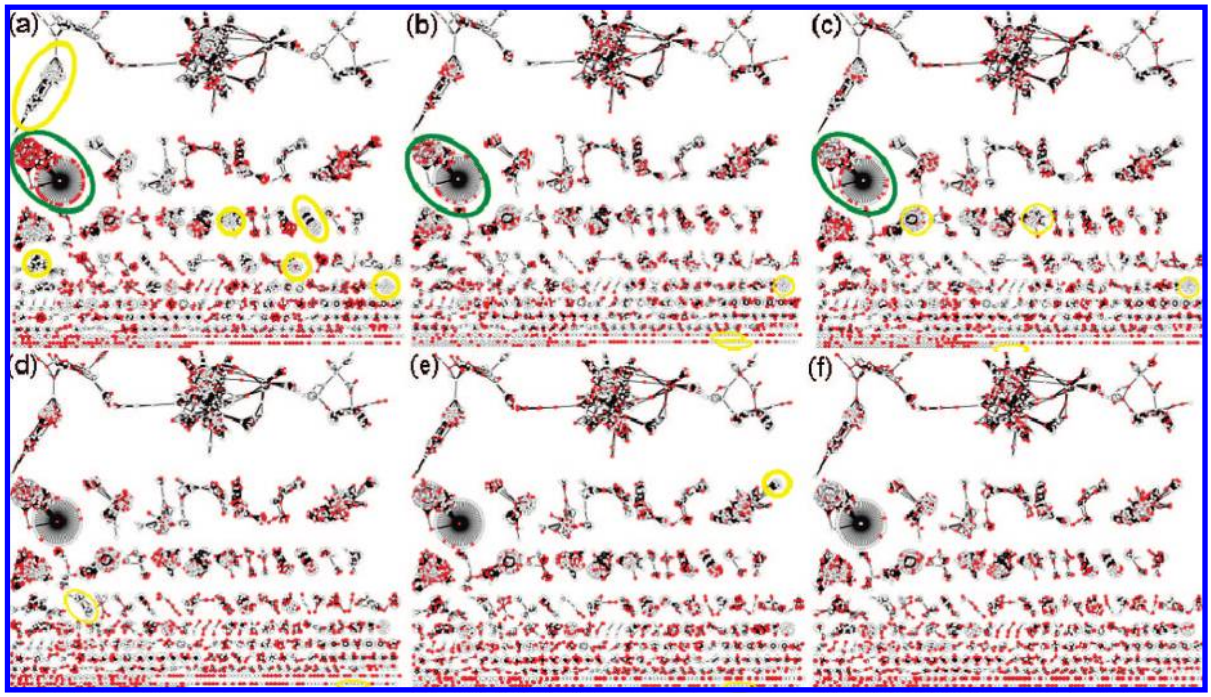

Figure 1. Pair-wise similarity plots as chemical space maps derived using ECFP_6 fingerprints. The six different plots reveal diverse compounds selected using the following fingerprints (a) hologram fingerprint of length 401 (H401), (b) Domain (DM), (c) Target (TG), (d) Functional group (FG), (e) functional group-target (FG-TG), and (f) FCFP_4. Nodes represent various antiTB agents. Edges between two nodes represent a ECFP_6 Tanimoto similarity >70% between molecules, with clusters being nodes connected by edges. The active compounds are highlighted in red, with the over-represented and under-represented regions in the map highlighted in green and yellow, respectively.

relevant 3D pharmacophoric fingerprints, which capture the diversity in both chemical and target space. Though 3D pharmacophore measurements are most appropriate and acceptable for this endeavor, they are often difficult to determine because of factors like computational scalability of handling multiple 3D conformers, ambiguities in the determination of the bioactive conformations, etc. Furthermore, none of the studies[24] point unambiguously to an approach for the training set selection from inhibitors identified in a cell-based assay. Thus, in the present study, six different fingerprints (FCFP_4, H401, FG, DM, TG, and FG-TG discussed in the section "Fingerprints for Training Set Design") were used to select different training sets.

Using Tanimoto similarity >70% as criteria, the fraction of compounds in the test set that is similar to the training set is low for H401 (∼0.4), medium for DM, TG, and FG-TG (∼0.6), and high for FG and FCFP_4 (∼0.8). Similar information is visually captured in the pairwise similarity map (Figure 1). The distribution of (and relationships between) scaffolds in screening libraries has recently been analyzed by Shelat and Guy[45] using an organic spring-embedded graph algorithm.[46] Instead of only the scaffold,

we analyze the similarity of the entire molecule (as encoded by ECFP_6 fingerprint) with the same methodology to generate the pairwise similarity map. These pairwise similarity maps are plots that depict the inter- and intracluster (with similarity >70%) distribution of molecules. In the plot, each compound is represented by a node. Similar compounds, connected by edges, are clustered to form "islands", with the smaller clusters located at the lower part of the plot. The chosen diverse compounds based on the different fingerprints are highlighted red in Figure 1a-f. Ideally, each cluster should have at least one chosen compound as representative but should not have too many. Hence, clusters with no representative compounds chosen are under-presented (highlighted by yellow circles in Figure 1), and those with too many compounds chosen are over-represented (highlighted by green circles in Figure 1). The pairwise similarity map clearly suggested that the diverse compounds selected using molecular hologram fingerprints (H401) are the least representative, with smaller clusters significantly under-represented (Figure 1a). While the representativeness of diverse compounds selected using the domain, target, functional group (Figure 1b−d) reveal a gradual improvement, the diverse compounds selected with FG-TG fingerprints (a combination of chemical ontology annotations and target predictions) (Figure 1e) and FCFP_4 fingerprints (Figure 1f) are the most representative of the entire data set.

Bayesian Model Development and Validation. The purpose of the present study was to derive a predictive Bayesian model capable of discriminating active (MIC < 5 μM) from inactive antiTB agents described in the training and test set. Using the six different training sets described above and 15 fingerprint types viz. ECFP_4, ECFP_6, ECFP_8, ECFP_10, ECFP_12, FCFP_4, FCFP_6, FCFP_8, FCFP_10, FCFP_12, SCFP_4, SCFP_6, SCFP_8, SCFP_10, and SCFP_12, a total of 90 Bayesian models were derived. The statistical parameters measuring the discriminative ability of these 90 models are presented in Figure 2.

For various training sets, the performance of different fingerprint classes are quite similar. However, for the test set, the positive recall of the 15 models developed using the H401 derived training set are consistently low (Figure 2b). This is probably related to the poor representativeness of the H401 training set and thus reinforces the importance of training set design discussed above and suggesting that chemical space maps like those presented in Figure 1 can be a powerful visual tool. Comparing the DM with the TG fingerprint, the former fingerprint appears to be better (Figure 2a). The TG fingerprints, as described above, were extrapolated from the DM fingerprints using the Interpro domain-M Tb protein association information.[35] As domains are believed to be the basic unit of protein evolution,[47] the domain fingerprints probably provide a more direct link between the chemical and biological space and hence a better fingerprint. Models developed from a training set of diverse compounds with FCFP_4 have superior total accuracy (Figure 2a) and high average positive recall (Figure 2b), closely followed by models developed on training sets derived using FG-TG fingerprints. Interestingly, models derived from ECFP fingerprints with diameter larger than 4 (i.e., 6, 8, 10, and 12) performed very poorly with the FG derived training set, which was rectified in the FG-TG derived set. However, the reason behind this improvement is unclear.

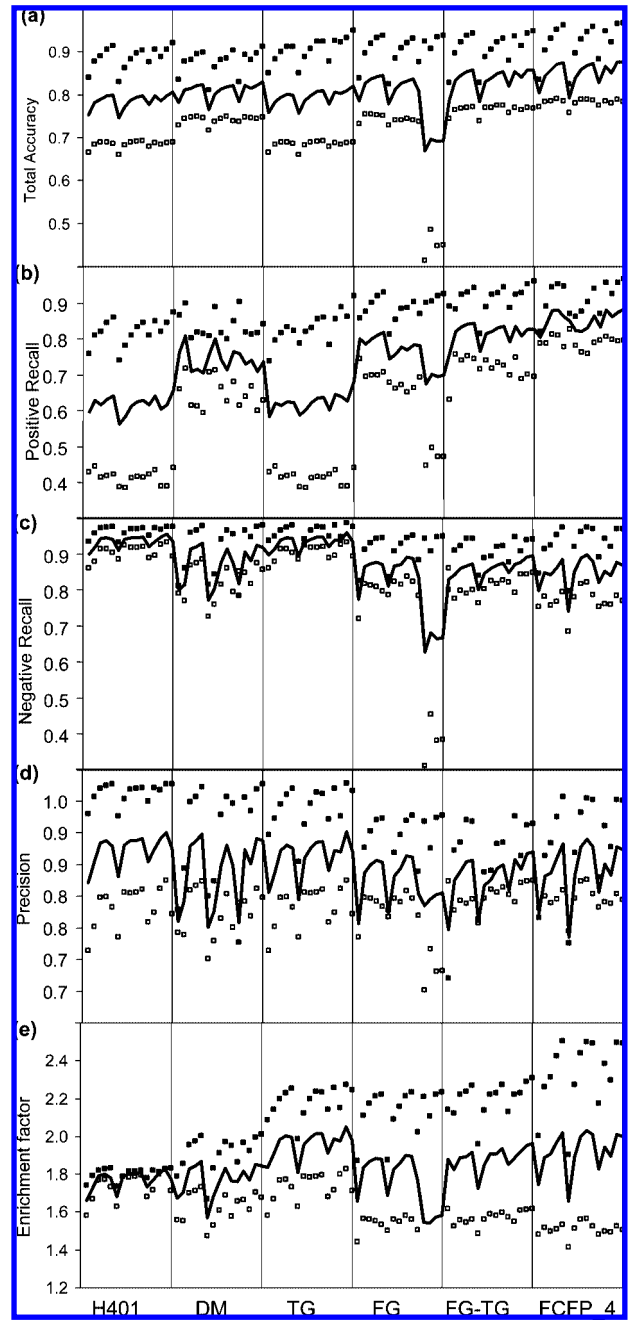

Figure 2. Comparison of (a) total accuracy, (b) positive recall, (c) negative recall, (d) precision, and (e) enrichment factor for the 90 models: six training set derived from the six fingerprints (H401, DM, TG, FG, FG-TG, FCFP_4), and each training set subjected to Bayesian modeling of the 15 fingerprints, in the order of SCFP (length 4 to 12), FCFP (length 4 to 12), and ECFP (length 4 to 12). The discriminative statistical parameters for the training set and test set is presented as a solid square (■) and a open square (□), respectively, with the solid line, representing the average of the training and test sets.

The best Bayesian model, developed on a training set obtained from diversity analysis of FCFP_4 fingerprints and derived using the ECFP_12 fingerprints plus Lipinski-type descriptors (ALogP, molecular weight, number of hydrogen bond donor and acceptor, number of rotatable bonds) and molecular fractional polar surface area as implemented in Pipeline Pilot,[21] was chosen based on optimum prediction accuracy for actives (total positive recall) and total accuracy values. Table 1 summarizes the key statistical discriminative

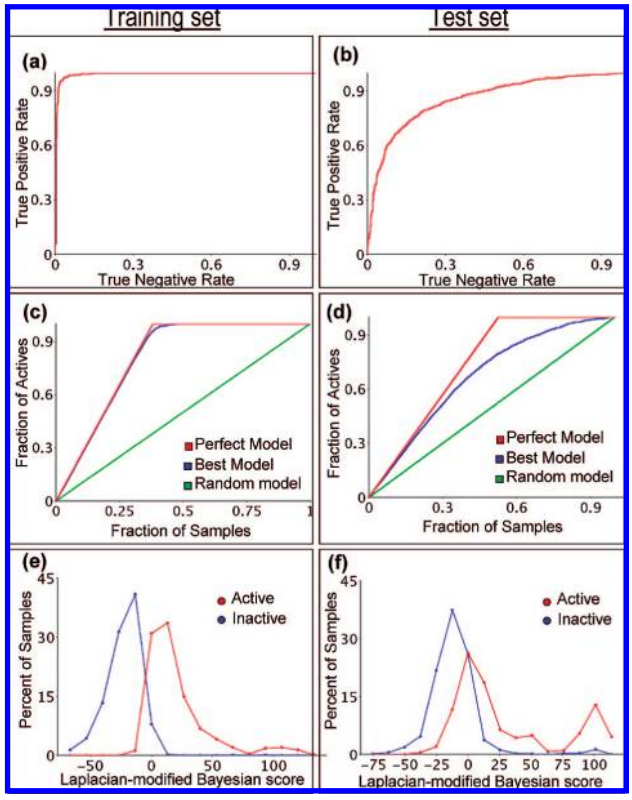

Figure 3. The performance of the best model depicted graphically via (a) Receiver Operating Characteristic (ROC) plot for training set, (b) ROC plot for test set, (c) enrichment plot for training set, (d) enrichment plot for test set, (e) frequency plot of active/inactive for training set, and (f) frequency plot of active/inactive for test set.

parameters of this model, in comparison with the performance of all 90 derived models. The chosen model has the highest average prediction accuracy for actives (average positive recall: 0.882) and the second highest total prediction accuracy (average total accuracy: 0.876). This model also had acceptable average prediction accuracy values for inactives (average negative recall: 0.869), average enrichment factor (average EF: 1.999), and average Receiver Operating Characteristic (ROC) score (average ROC score: 0.926). The ROC plot (Figure 3a-b) revealed that the model presents excellent prediction accuracy for the training set (Figure 3a, ROC score: 0.993) and good prediction accuracy for the test set (Figure 3b, ROC score: 0.858). The enrichment plots (Figure 3c-d) reveal very high enrichment rates for the compounds of the training set and reasonably good enrichment values for the compounds of the test set. With the best model, 50% of the training set actives would be found by screening 19.2% of the training set, compared to 19.1% for the perfect model and 50% for a random model, with very similar enrichment observed for the test set actives. Finally, the distribution curve for the training set suggests that this model can segregate the actives from the inactives: while the Laplacian-modified Bayesian score for the actives is mostly between (10 to 80), the inactives are found predominantly between (−50 to 10) (Figure 3e), even though the discriminatory power is only fair for the test set (Figure 3f). In summary, all statistical parameters suggested that the chosen Bayesian antiTB compounds model is of high quality.

To assess the stability of the chosen model, two other training sets were derived identically, but with a different number of molecules ($n = 1890$: approximately half the data set and $n = 2520$: approximately two-thirds of the data set). The comparisons of the discriminative ability of the three models derived from these training sets reveal a superior optimal prediction accuracy of the chosen model (Table 1). Interestingly, the discriminative ability for both the training and test sets appeared to deteriorate with an increasing number of molecules in the training set. This may be attributed to the bias in the weighing scheme for over-represented substructures in the data set.

In addition, this model successfully classified several well-known antiTB agents in the tuberculosis pipeline and compounds under development (Supporting Information Table S2). With 13 out of a total of 44 compounds belonging to the training set, the model recalled 86% of the actives (five false negatives, Figure 4b: TMC207, thiacetazone, viomycin, FAS20013, and clotrimazole) and 86% of inactives (one false positive, Figure 4a: mefloquine). The inaccurate classification by Bayesian models are primarily attributed to outliers i.e. compounds whose end point measurements might be faulty; or compounds where theory of similar-structure-similar-activity is violated; or compounds that do not fall into any natural cluster (singlets). In order to gain insight into the misclassification, pairwise similarity maps of these six misclassified compounds with the 3779 antiTB compounds of the data set were constructed. Among the six misclassified compounds, TMC207 and mefloquine were absent from the pairwise similarity maps as they had no structural similarity (>70%) with any of the antiTB compounds of the data set. Similar to the pairwise similarity map displayed in Figure 1, the nodes (circles) in Figure 5 represents individual compounds, with edges connecting similar compounds. However, in Figure 5, the red circles highlight the active compounds (MIC < 5 μM), with the misclassified compounds depicted as yellow hexagons. Figure 5a reveals that the misclassified compounds are found in the minor clusters, and the incorrect classification of thiacetazone, viomycin, FAS20013, and clotrimazole can be attributed to the fact that they belong to clusters whose members are completely or mostly inactive (Figure 5b-e). For example, thiacetazone (MIC = 0.0423 μM: Figure 5b, ID: 126) was in the same cluster with one active compound (MIC = 0.393 μM, Figure 5b, ID: 1208) and four inactives (MIC > 17 μM, Figure 5b: ID 1500-1503).

We have also validated our model on an external test set derived from a commercial database. The original database containing 9416 compounds with antibacterial end points collected from journals and patents, with ~3600 compounds with MICs from ~10 wild type *Mycobacterium tuberculosis* strains. Compounds common to the NIAID data sets were removed. For the remaining compounds where multiple MIC values were reported, an average value was obtained, and those with standard deviation were more than three removed. Furthermore, we applied this data set to the model published in ref 13 (see below for further discussion) and found that the descriptors of ~40 compounds cannot be obtained. Hence, the external test set consists of 2880 entries with averaged MICs (Supporting Information Table S3) from the original commercial database. As less than 1% of the external test set is 70% similar (ECFP_4 fingerprint) to the compounds in the training set, suggesting that the two data sets are well-separated. Despite the low similarity between the training and the external test set, the

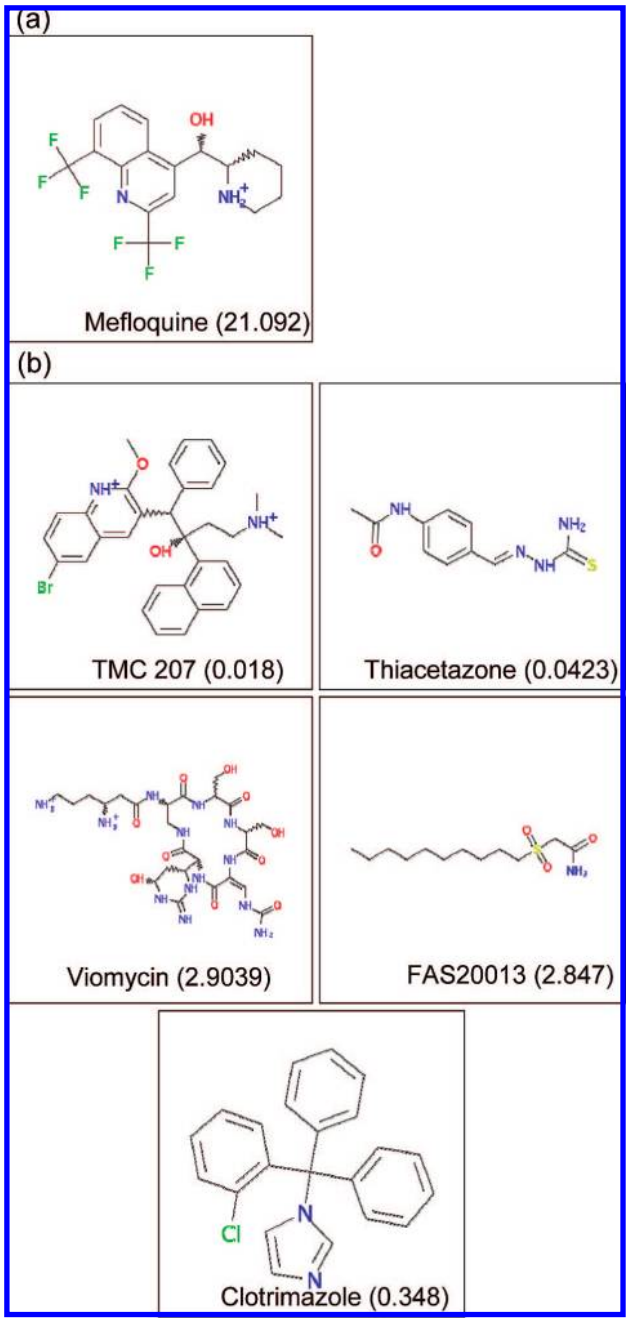

Figure 4. (a) False positive and (b) false negative predicted compounds among the 44 standard antiTB agents, predicted by the best model. The experimental MIC (μM) is included in brackets.

proposed model successfully discriminates the actives and inactives (total accuracy: 0.73; positive recall: 0.72) purchased from GVKBio.[48]

In order to gauge whether using a larger training set uncovers new information, we apply the published model by *García- García et al.*[13] to the 2880 data set. In prioritizing antiTB compounds, ref 13 suggested the use of four equations. As the criteria for actives/inactives (eqs 1 and 2) differs from our model here, we decided to make use of eqs 3 and 4, which relates the structures to MIC. To obtain the MIC via eq 3, a proprietary program is needed of which we cannot access. In order to compare, the predicted MIC from eq 4 (in μg L^{-1}) were converted to μM, and compounds less than 5 μM are considered active. Using this approach, the positive and negative recalls using eq 4[13] were found to

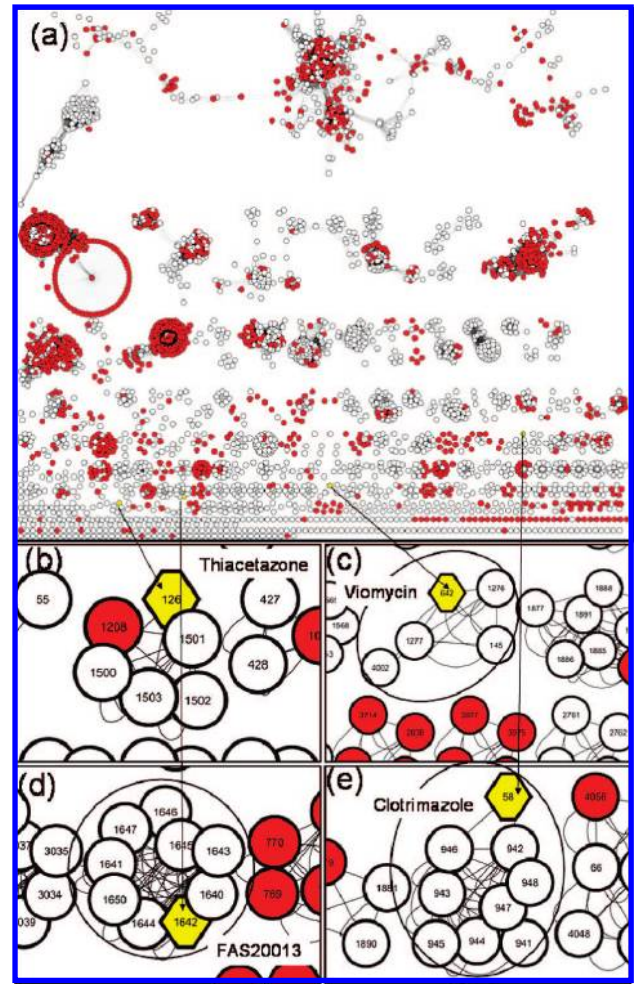

Figure 5. (a) Chemical space map of antiTB compounds related by pairwise similarity: nodes represent various antiTB agents; edges between two nodes represent a ECFP_6 Tanimoto similarity >70% between molecules, with clusters being nodes connected by edges. Inactives are depicted as open circles and actives (MIC < 5μM) as red circles. The clusters containing misclassified compounds (yellow hexagons) are depicted in parts (b)-(e) for thiacetazone, viomycin, FAS20013, and clotrimazole, respectively. The numbers in the figure correspond to the "TB data set ID" in Supporting Information Table S1.

be 0.44 and 0.43, respectively, thus, suggesting that our model has indeed uncovered new information.

Model Interpretation. Though Bayesian models are predominantly applied as virtual screening tools, they may also be useful for elucidating the essential structural and physicochemical requirements for activity. While the majority of 3D QSAR and pharmacophore elucidation approaches use three-dimensional steric and electrostatic potentials or chemical feature based pharmacophoric fingerprints, the model reported in this paper uses 2D fingerprints (atom types and connectivity), in addition to several global physiochemical descriptors in the Bayesian model development. In view of these differences, the current Bayesian model reveals essential features in terms of abstract substructural features that discriminate active and inactive molecules,[18−21] which is not necessarily the conserved array of 3D pharmacophoric features which are usually identified in a 3D QSAR/ pharmacophoric study. This list of substructures preferentially present in actives and inactives (Supporting Information-Figure S1) can serve as a look-up guide to chemists/ computational chemists during a hit-to-lead or lead optimi-

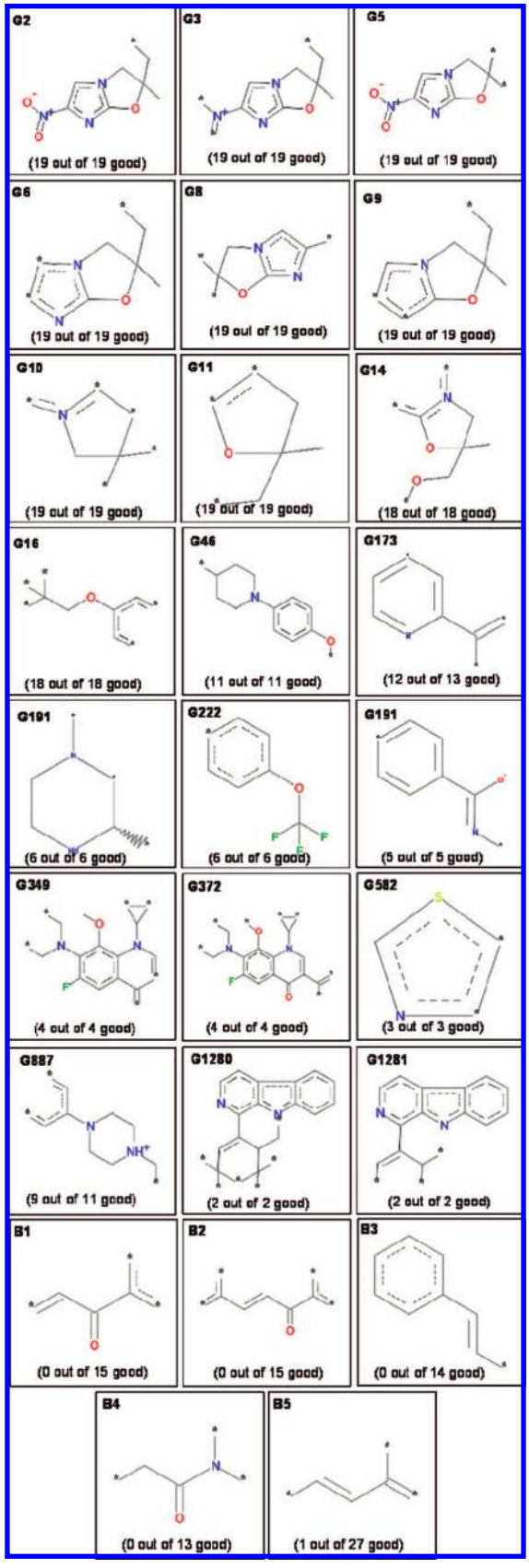

Figure 6. (a) Selected substructures with incremental effect, prefixed with "G", on MIC (b) selected substructures with detrimental effect, prefixed with "B", on MIC, predicted by the best model. The frequency of their occurrences in active (good) molecules is given in bracket, with * represents any atom. The complete list of the top 500 substructures identified to have incremental and detrimental effects on MIC is deposited as Supporting InformationFigure S1.

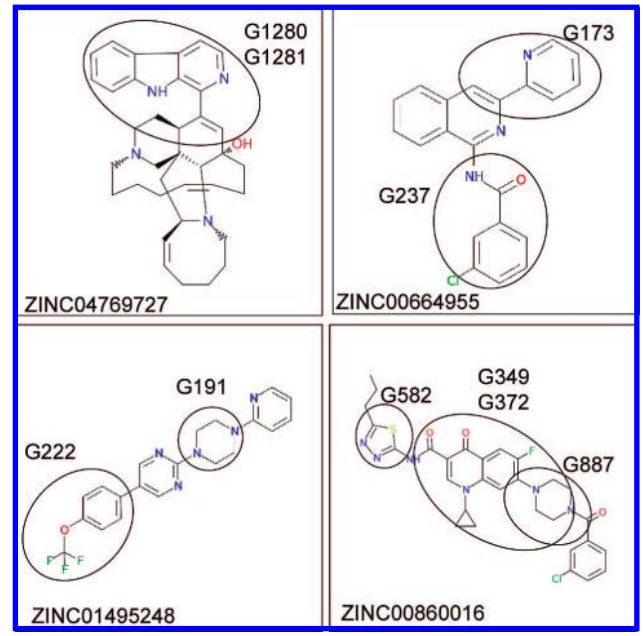

Figure 7. Some compounds prioritized from the ZINC database, with the substructures having an incremental effect on MIC (predicted by the best model) highlighted by a circle. The list of 50 diverse compounds predicted to be active with the best model is deposited in Supporting InformationFigure S2.

zation campaign. In general, substructures that have an incremental effect on MIC (Figure 6a) include the following: substituted imidazooxazoles (G2, G3, G5, G6, G8, and G9), pyrrols (G10), furans (G11), oxazoles (G14), and phenoxy (G16), while substructures such as hydroxyl ketones (B1), chalcones (B2), phenyl ethelenes (B3), tertiary amides (B4), dienes (B5), etc. appeared to have a detrimental effect on the MIC (Figure 6b).

Application of the Model. In the simplest sense, the active substructures presented in Supporting InformationFigure S1 can be used as queries for compound libraries. Furthermore, the Bayesian model itself is well-suited for virtual screening. As an illustration, the selected model was used to screen the ZINC database[49] with the list of prioritized compounds deposited in Supporting InformationFigure S2. Some interesting compounds are highlighted in Figure 7: ZINC04769727 is a known as an anticancer agent from the NCI anticancer set,[50] while the other three (ZINC00664955, ZINC01495248, and ZINC00860016) are "hybrids" of active substructures. The design of hybrid drugs, i.e. compounds that incorporate two drug pharmacophores into a single molecule, has been substantially reviewed,[51−53] with DU1302 (that incorporates substructures from antimalarial trioxanes and quinolines) recently reported as a new generation antimalarial.[54]

In addition, the results of the model would be useful for the design and optimization of compounds with M Tb cellular activity, either by replacing inactive substructures with actives, removing inactive substructures altogether, or adding active substructures to small fragments with promising cellular activity. Such strategies are especially useful in lead optimization campaigns on compounds identified in cellular screens, where structure activity relationships are often difficult to interpret and maintaining cellular activity can be a challenge.[55] For example, the anti TB compound OPC-67683[56−58] currently in a phase II trial has a MIC 10-fold lower than that of CGI17341 (Figure 8). While the two compounds

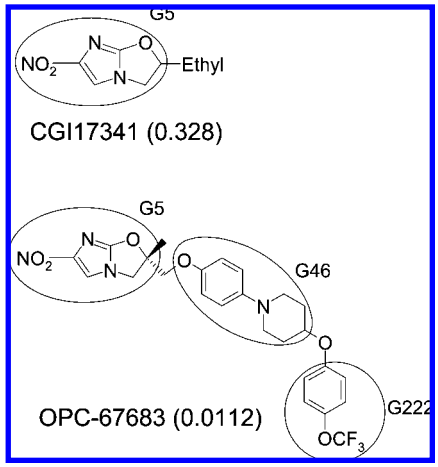

Figure 8. CGI17341 and OPC-67683, with the substructures having an incremental effect on MIC (predicted by the best model) highlighted by a circle. The experimental MIC (μM) is included in brackets.

shared the same nitroimidazole scaffold (G5), OPC-67683 has incorporated two good features identified here (G46 and G222) which is not present in CGI17341. So even though the 4-piperdine (G46) was originally introduced to improve oral bioavailability, it probably also helped in the reduction of the MIC of CGI17341.

CONCLUSIONS

Bayesian classification models which discriminate active (MIC < 5 μM) from inactive molecules were derived and validated with a data set of 3779 compounds which have been measured for MIC in the *Mycobacterium tuberculosis* H37Rv strain. Since training set design and choice of the fingerprints critically effect the predictive ability of the QSAR models, the model development and validation involved exploring six different training sets and 15 fingerprint types which resulted in a total of 90 models. The best model, prioritized using average total accuracy and positive recall, was derived using Extended Class Fingerprints of maximum diameter 12 (ECFP_12) and a few global descriptors on a training set derived using Functional Class Fingerprints of maximum diameter 4 (FCFP_4). This model demonstrated very good discriminant ability in general, with excellent discriminant statistics for the training set (total accuracy: 0.968; positive recall: 0.967) and a good predictive ability for the test set (total accuracy: 0.869; positive recall: 0.789). Among a set of 44 well-known antiTB agents, only six were misclassified, and the reasons for failure were identified. In addition, the model demonstrated good predictive ability (total accuracy: 0.73; positive recall: 0.72) against a well-separated test set ($n = 2880$, with less than 1% of the compounds have ECFP_4 similarity >70% with respect to the compounds of the training set). With such encouraging results, the model was used to screen the ZINC databases, and several compounds which might be of interest to the TB drug discovery community were highlighted.

ACKNOWLEDGMENT

The authors thank Dr. Shahul Nilar and the Accelrys Pipeline Pilot support team for their comments and support. The generous help from Dr. Jesus Vicente de Julián-Ortiz in discussions and providing the program to obtain their published model is gratefully acknowledged.

Supporting Information Available: Data set of 3779 compounds from National Institute of Allergy and Infectious Diseases (NIAID) with their MIC values (Table S1), a data set of 44 well-known antiTB agents in the tuberculosis pipeline and compounds under development with their MIC values (Table S2), a data set of 2880 compounds from GVK with their average MIC against TB wide type strains (Table S3), the list of substructures preferentially present in actives and inactives identified by the best model (Figure S1), and the list of 50 diverse compounds prioritized by the best model from the ZINC database (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lenaerts, A. J.; Degroote, M. A.; Orme, I. M. Preclinical testing of new drugs for tuberculosis: current challenges. *Trends Microbiol.* **2008**, *16*, 48–54.
(2) Tomioka, H. Current status of some antituberculosis drugs and the development of new antituberculous agents with special reference to their in vitro and in vivo antimicrobial activities. *Curr. Pharm. Des.* **2006**, *12*, 4047–4070.
(3) Williams, K. J.; Duncan, K. Current strategies for identifying and validating targets for new treatment-shortening drugs for TB. *Curr. Mol. Med.* **2007**, *7*, 297–307.
(4) Check, E. After decades of drought, new drug possibilities flood TB pipeline. *Nat. Med.* **2007**, *13*, 266.
(5) Salomon, J. A.; Lloyd-Smith, J. O.; Getz, W. M.; Resch, S.; Sanchez, M. S.; Porco, T. C.; Borgdorff, M. W. Prospects for advancing tuberculosis control efforts through novel therapies. *PLoS Med.* **2006**, *3*, e273.
(6) Freire, M. C. Opportunities for overcoming tuberculosis: new treatment regimens. *World Hosp. Health Serv.* **2006**, *42*, 34–37.
(7) Spigelman, M.; Gillespie, S. Tuberculosis drug development pipeline: progress and hope. *Lancet* **2006**, *367*, 945–947.
(8) Selassie, C. D. The History of Quantitative Structure Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 6th ed.; Abraham, D. J., Ed.; John Wiley and Sons Publishers: New York, 2003; Vol. 1, pp 1−48.
(9) Ginsberg, A. M.; Spigelman, M. Challenges in tuberculosis drug research and development. *Nat. Med.* **2007**, *13*, 290–294.
(10) Takata, T.; Winzeler, E. A. Genomics, systems biology and drug development for infectious diseases. *Mol. BioSyst.* **2007**, *3*, 841–848.
(11) Fischer, H. P. Towards quantitative biology, integration of biological information to elucidate disease pathways and to guide drug discovery. *Biotechnol. Annu. Rev* **2005**, *11*, 1–68.
(12) (a) Desai, B.; Sureja, D.; Naliapara, Y.; Shah, A.; Saxena, A. K. Synthesis and QSAR studies of 4-substituted phenyl-2,6-dimethyl-3,5-bis-N- (substituted phenyl)carbamoyl-1,4-dihydropyridines as potential antitubercular agents. *Bioorg. Med. Chem.* **2001**, *9*, 1993–1998. (b) Saquib, M.; Gupta, M. K.; Sagar, R.; Prabhakar, Y. S.; Shaw, A. K.; Kumar, R.; Maulik, P. R.; Gaikwad, A. N.; Sinha, S.; Srivastava, A. K.; Chaturvedi, V.; Srivastava, R.; Srivastava, B. S. C-3 Alkyl/Arylalkyl-2,3-dideoxy Hex-2-enopyranosides as Antitubercular Agents: Synthesis, Biological Evaluation, and QSAR Study. *J. Med. Chem.* **2007**, *50*, 2942–2950. (c) Bagchi, M. C.; Maiti, B. C.; Mills, D.; Basak, S. C. Usefulness of graphical invariants in quantitative structure-activity correlations of tuberculostatic drugs of the isonicotinic acid hydrazide type. *J. Mol. Model.* **2004**, *10*, 102–111. (d) Ventura, C.; Martins, F. Application of quantitative structure-activity relationships to the modeling of antitubercular compounds. 1. The hydrazide family. *J. Med. Chem.* **2008**, *51*, 612–624. (e) Pasqualoto, K. F.; Ferreira, E. I.; Santos-Filho, O. A.; Hopfinger, A. J. Rational design of new antituberculosis agents: receptor-independent four-dimensional quantitative structure-activity relationship analysis of a set of isoniazid derivatives. *J. Med. Chem.* **2004**, *47*, 3755–3764. (f) Kiritsy, J. A.; Yung, D. K.; Mahony, D. E. Synthesis and quantitative structure-activity relationships of some antibacterial 3-formylrifamycin SV N-(4-substituted phenyl)piperazinoacethydrazones. *J. Med. Chem.* **1978**, *21*, 1301–1307. (g) Gossman, W.; Oldfield, E. Quantitative structure--activity relations for gammadelta T cell activation by phosphoantigens. *J. Med. Chem.* **2002**, *45*, 4868–4874. (h) Rugutt, J. K.; Rugutt, K. J. Relationships between molecular properties and antimycobacterial activities of steroids. *Nat. Prod. Lett.* **2002**, *16*, 107–113. (i) Aparna,

V.; Jeevan, J.; Ravi, M.; Desiraju, G. R.; Gopalakrishnan, B. 3D-QSAR studies on antitubercular thymidine monophosphate kinase inhibitors based on different alignment methods. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1014–1020. (j) Schaper, K. J.; Pickert, M.; Frahm, A. W. Substituted xanthones as antimycobacterial agents. Part 3: QSAR investigations. *Arch. Pharm. (Weinheim)* **1999**, *332*, 91–102.
(13) García-García, A.; Galvez, J.; Julian-Ortiz, J. V.; García-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. Search of chemical scaffolds for novel antituberculosis agents. *J. Biomol. Screen.* **2005**, *10*, 206–214.
(14) Prakash, O.; Ghosh, I. Developing an antituberculosis compounds database and data mining in the search of a motif responsible for the activity of a diverse class of antituberculosis agents. *J. Chem. Inf. Model.* **2006**, *46*, 17–23.
(15) Manetti, F.; Magnani, M.; Castagnolo, D.; Passalacqua, L.; Botta, M.; Corelli, F.; Saddi, M.; Deidda, D.; De Logu, A. Ligand-based virtual screening, parallel solution-phase and microwave-assisted synthesis as tools to identify and synthesize new inhibitors of Mycobacterium tuberculosis. *ChemMedChem* **2006**, *1*, 973–989.
(16) Yuan, H.; Wang, Y.; Cheng, Y. Local and global quantitative structure-activity relationship modeling and prediction for the baseline toxicity. *J. Chem. Inf. Model.* **2007**, *47*, 159–169.
(17) Yuan, H.; Wang, Y. Y.; Cheng, Y. Y. Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. *J. Mol. Graphics Modell.* **2007**, *26*, 327–335.
(18) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463–4470.
(19) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356–4359.
(20) Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 139–144.
(21) *Pipeline Pilot version 6.1.5*; SciTegic: San Diego, CA, 2007.
(22) National Institutes of Health, National Institute of Allergy and Infectious Diseases (NIAID), Division of AIDS, HIV/OI/TB Therapeutics Database. http://chemdb.niaid.nih.gov/struct_search/oi/OI_search.asp# (accessed April 29, 2007).
(23) United States National Library of Medicine, National Institutes of Health, National Center for Biotechnology Information. Pubchem FTP SDF download site. ftp://ftp.ncbi.nih.gov/pubchem/Compound/CURRENT-Full/SDF/ (accessed January 15, 2008).
(24) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357–369.
(25) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.
(26) Schellhammer, I.; Rarey, M. TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223–238.
(27) Oloff, S.; Zhang, S.; Sukumar, N.; Breneman, C.; Tropsha, A. Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J. Chem. Inf. Model.* **2006**, *46*, 844–851.
(28) Vogt, M.; Bajorath, J. Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chem. Biol. Drug Des.* **2008**, *71*, 8–14.
(29) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49*, 7169–7181.
(30) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019–2024.
(31) Sheridan, R. P. The centroid approximation for mixtures: calculating similarity and deriving structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456–1469.
(32) Feldman, H. J.; Dumontier, M.; Ling, S.; Haider, N.; Hogue, C. W. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett.* **2005**, *579*, 4685–4691.
(33) Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M.; Mikhailov, D.; Clemons, P. A.; Davies, J. W. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screening* **2007**, *10*, 719–731.
(34) Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G. W.; Tao, C. Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **2008**, *4*, 59–68.
(35) Prathipati, P.; Bender, A.; Ma, N. L.; Manjunatha, U. H.; Nilar, S.; Keller, T. H. Unpublished results.
(36) *SYBYL, version 7.0*; Tripos Inc.: St. Louis, MO, 2007.
(37) Haider, N. *Checkmol version 0.4*;University of Vienna: 2003−2007. http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html (accessed January 20, 2008).
(38) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
(39) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Buillard, V.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Daugherty, L.; Dibley, M.; Finn, R.; Fleischmann, W.; Gough, J.; Haft, D.; Hulo, N.; Hunter, S.; Kahn, D.; Kanapin, A.; Kejariwal, A.; Labarga, A.; Langendijk-Genevaux, P. S.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Orengo, C.; Petryszak, R.; Selengut, J. D.; Sigrist, C. J.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H.; Yeats, C. New developments in the InterPro database. *Nucleic Acids Res.* **2007**, *35*, D224−D228.
(40) *SciTegic Pipeline Pilot Chemistry Collection: Basic Chemistry User Guide*; Accelrys Software Inc.: San Diego, CA, March 2008.
(41) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682–686.
(42) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
(43) Saxena, A. K.; Prathipati, P. Comparison of MLR, PLS and GA-MLR in QSAR analysis. *SAR QSAR. Environ. Res* **2003**, *14*, 433–445.
(44) Prathipati, P.; Saxena, A. K. Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 39–51.
(45) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442–446.
(46) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.
(47) Breinbauer, R.; Vetter, I. R.; Waldmann, H. From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem., Int. Ed.* **2002**, *41*, 2879–2890.
(48) GVK Bio. http://www.gvkbio.com (accessed Oct 2008).
(49) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(50) Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubenstein, L.; Plowman, J.; Boyd, M. R. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.
(51) Christiaans, J. A. M.; Timmerman, H. Cardiovascular hybrid drugs: combination of more than one pharmacological property in one single molecule. *J. Pharm. Sci.* **1996**, *4*, 1–22.
(52) Meunier, B. Hybrid molecules with a dual mode of action: dream or reality. *Acc. Chem. Res.* **2008**, *41*, 69–77.
(53) Viegas-Junior, C.; Danuello, A.; da, S. B. V.; Barreiro, E. J.; Fraga, C. A. Molecular hybridization: a useful tool in the design of new drug prototypes. *Curr. Med. Chem* **2007**, *14*, 1829–1852.
(54) Dechy-Cabaret, O.; Benoit-Vical, F.; Robert, A.; Meunier, B. Preparation and antimalarial activities of "trioxaquines", new modular molecules with a trioxane skeleton linked to a 4-aminoquinoline. *ChemBioChem* **2000**, *1*, 281–283.
(55) Silver, L. L. A retrospective on the failures and successes of antibacterial drug discovery. *IDrugs* **2005**, *8*, 651–655.
(56) Sacchettini, J. C.; Rubin, E. J.; Freundlich, J. S. Drugs versus bugs: in pursuit of the persistent predator Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* **2008**, *6*, 41–52.
(57) Sasaki, H.; Haraguchi, Y.; Itotani, M.; Kuroda, H.; Hashizume, H.; Tomishige, T.; Kawasaki, M.; Matsumoto, M.; Komatsu, M.; Tsubouchi, H. Synthesis and antituberculosis activity of a novel series of optically active 6-nitro-2,3-dihydroimidazo[2,1-b]oxazoles. *J. Med. Chem.* **2006**, *49*, 7854–7860.
(58) Matsumoto, M.; Hashizume, H.; Tomishige, T.; Kawasaki, M.; Tsubouchi, H.; Sasaki, H.; Shimokawa, Y.; Komatsu, M. OPC-67683, a nitro-dihydro-imidazooxazole derivative with promising action against tuberculosis in vitro and in mice. *PLoS Med.* **2006**, *3*, e466.

CI800143N