DOI: 10.1002/cmdc.200600026

Ligand-Based Virtual Screening, Parallel Solution-Phase and Microwave-Assisted Synthesis as Tools to Identify and Synthesize New Inhibitors of *Mycobacterium tuberculosis*

Fabrizio Manetti,[a] Matteo Magnani,[a] Daniele Castagnolo,[a] Laura Passalacqua,[a] Maurizio Botta,*[a] Federico Corelli,[a] Manuela Saddi,[b] Delia Deidda,[b] and Alessandro De Logu[b]

In an attempt to identify new inhibitors of the growth of Mycobacterium tuberculosis *(MTB), the causative agent of tuberculosis, a procedure for the generation, design, and screening of a ligand-based virtual library was applied. This used both an in silico protocol centered on a recursive partitioning (RP) model described herein, and a pharmacophoric model for antitubercular agents previously generated by our research group. Two candidates emerged from databases of commercially available compounds, both characterized by a minimum inhibitory concentration (MIC) of 25 $\mu g\,mL^{-1}$. Based on these compounds, two series of derivatives were synthesized by both parallel solution-phase and microwave-assisted synthesis, leading to enhanced antimycobacterial activity. During both the design and synthesis, attention was focused on the efficient allocation of available resources with the aim of reducing the overall costs associated with calculation and synthesis.*

Introduction

Tuberculosis (TB), a chronic contagious disease caused by *Mycobacterium tuberculosis* (MTB), is one of the leading causes of death worldwide with approximately 3 million deaths and 8 million new cases each year.[1,2] According to the World Health Organization (WHO) one third of the world's population is infected with MTB, and nearly 30 million people will be infected in next 20 years. The dangerous spread of TB is mainly due to its association with HIV infection (about two-thirds of the patients infected with TB are also HIV seropositive),[3] and to the rapid development of multidrug-resistant (MDR) strains of MTB.[4,5] No TB-specific drug has been discovered in the last 40 years, and most of the current antitubercular compounds exhibit serious side effects and high-toxicity. Consequently there is an urgent need to discover new structural classes of antimycobacterial compounds in order to develop agents to replace or supplement the established drugs.[6] Many attempts have been made by our research group to identify new antitubercular hits with innovative chemical structures.[7–11]

Due to their ability to decrease the cost of drug discovery and development, computer-aided drug design approaches have emerged as useful tools to select, from a large library of possible compounds, a small subset for chemical synthesis and testing.[12] The computational synthesis of new molecules (which constitute the virtual library) is called "virtual library generation", while the selection of a subset of compounds from a larger library is termed "virtual library design".[13] Virtual library design and virtual screening could be also viewed as an attempt to merge the powerful tools of combinatorial chemistry and computer-aided drug design. Library design can also be applied as a filtering method (in silico or virtual screening, VS), by using computational models to filter databases or virtual libraries by the evaluation of specific biological properties of molecules.

In continuing our efforts to discover new chemical entities endowed with activity against MTB, we present the successful application of virtual library generation, design and screening tools for the discovery of new leads with antitubercular activity. Two compounds with potential activity against MTB were identified through a combination of computational approaches: a novel in silico screening protocol centered on a recursive partitioning (RP) model (RP-centered virtual screening protocol,[14] RPCVS) and a pharmacophoric model for antitubercular compounds previously generated by our research

[a] *Dr. F. Manetti, Dr. M. Magnani, Dr. D. Castagnolo, Dr. L. Passalacqua, Prof. M. Botta, Prof. F. Corelli*
Dipartimento Farmaco Chimico Tecnologico
Università degli Studi di Siena
Via Alcide de Gasperi 2, 53100 Siena (Italy)
Fax: (+39) 0577-234333
E-mail: botta@unisi.it

[b] *Dr. M. Saddi, Dr. D. Deidda, Prof. A. De Logu*
Sezione di Microbiologia Medica
Dipartimento di Scienze e Tecnologie Biomediche
Università di Cagliari
Viale Sant'Ignazio 38, 09123 Cagliari (Italy)

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

group.[7] For each lead compound a series of analogues was synthesized by parallel solution-phase and microwave-assisted synthesis, resulting in compounds with improved antimycobacterial activity.

The Computational Approach

The target of most antimycobacterial agents is still unknown at the molecular level,[15] and thus no three-dimensional structure of the putative target(s) of antitubercular agents is available. Consequently, to develop a VS procedure aimed at identifying new antitubercular compounds, we adopted a ligand-based (LB) VS approach[16] (LBVS, based on the use of known active compounds as templates), instead of the more popular structure-based VS (based mainly on docking algorithms).

Here we report the use of a LBVS approach both as a filtering method, based on Lipinski's "rule of five", and as an activity-based selection method (used as a RP-based technique for compound classification). Note that LBVS constitutes only a part of the overall computational protocol.

The first step of this protocol (Figure 1, step 1) consisted of building a learning set of 471 compounds (reported in the literature as inhibitors of MTB), and generating from this a population of classification structure–activity relationship (CSAR) models by means of RP analysis (step 2). From these the best model (in terms of predictive power toward a test set) was selected (step 3) and employed as a part of a more precise VS procedure based on the model itself (the RPCVS protocol)[14] incorporating two additional filters (Lipinski's rule, and a calculated penalty factor) as described below.

In a parallel exercise (Figure 2), a large fragment-based virtual library (more than 5 million compounds), the "large library" (LL) was built, based on a benzene core, (step 4). From this a smaller set (5000 molecules, the "small library", SL) was extracted by means of a library design procedure (step 5), aimed at decreasing the dimensions while retaining as much structural information of the original library as possible. The SL was subsequently filtered (step 6) by using the RPCVS protocol, resulting in a "targeted library" (TL) of 114 compounds, upon which a preliminary in silico ADME evaluation was also performed. Many commercially available databases were searched for suitable TL compounds (step 7), leading to the purchase of four members of the Asinex Gold Collection database (step 8).[17] One of these showed good anti-MTB activity (25 $\mu g\,mL^{-1}$). A subsequent in silico screening of the whole Asinex database, performed by using both our pharmacophoric model and the RPCVS protocol (Figure 3, step 9), led to the

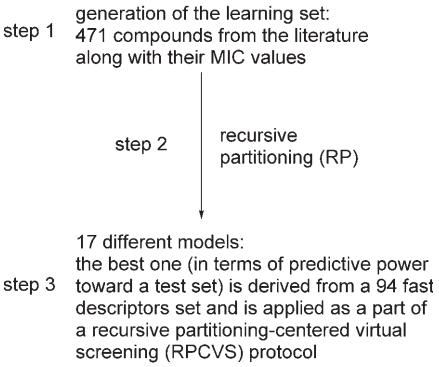


Figure 1. Computational steps leading from the learning set of antitubercular compounds to the recursive partitioning-centered virtual screening (RPCVS) protocol.

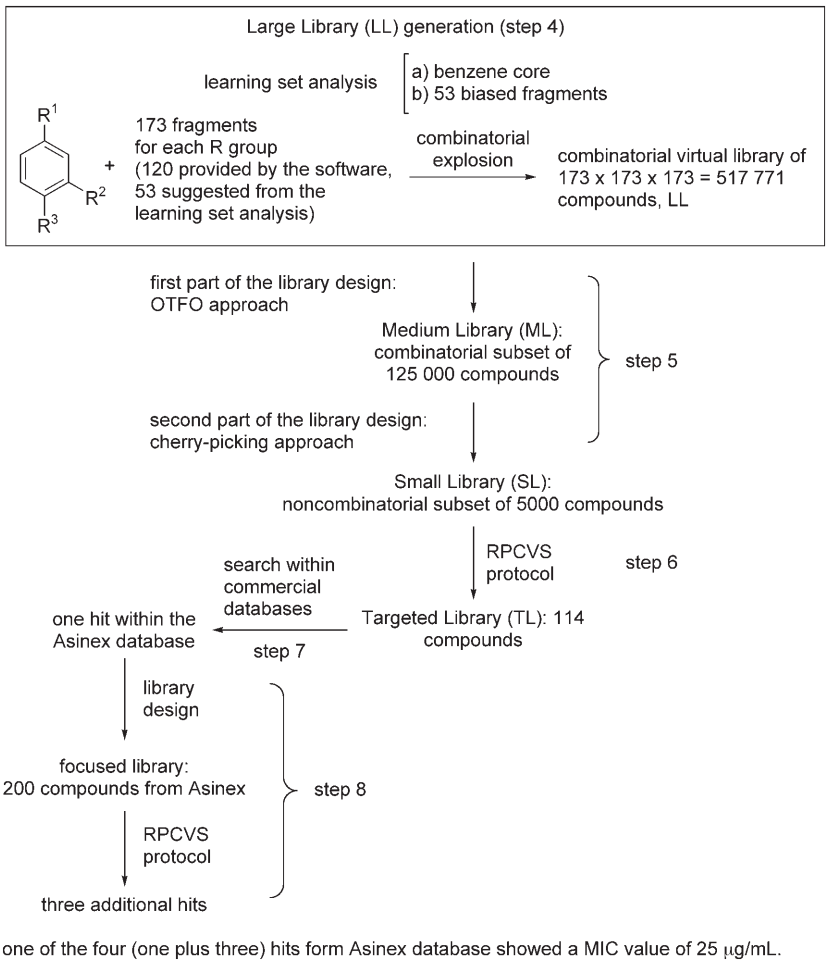


Figure 2. Computational steps leading from the large virtual library (LL) to the identification of four antimycobacterial hit compounds within the Asinex database.

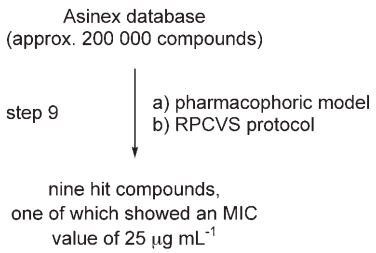


Figure 3. In silico screening of the whole Asinex database.

selection of a second set of nine compounds. From these a second lead compound was identified showing an anti-MTB MIC of 25 $\mu g\,mL^{-1}$. Importantly, at both the current and previous stages or our work, the search for virtual compounds in databases of real (commercially available) compounds was performed in order to obtain several hits for testing in a short time; this served to assess quickly the reliability of the VS procedure, and thus avoid time-consuming and costly synthesis.

Two series of analogues of the lead compounds were synthesized (step 10, Schemes 1–7), in order to obtain molecules with improved antitubercular activity (compared with their parent compounds), or data for a preliminary structure–activity relationship (SAR) study.

Step 1: generation of the learning set

A learning set of 471 compounds from structurally diverse chemical classes (imidazoles, pyridines, pyrroles, benzimidazoles, pyrazines, quinoxalines, quinolylhydrazones, benzanilides, etc.),[18] with their anti-MTB MIC values, was collated from literature published between 1997 and 2002. (When expressed as $\mu g\,mL^{-1}$, MIC values were converted into $\mu mol\,mL^{-1}$). Activity values spanned six orders of magnitude (5×10^{-4} to $\sim2\times10^{2}$ mM). However, such a database suffers from several limitations: since the mechanism of action at the molecular level is often unknown, the possibility cannot be excluded that compounds interact with different targets; and published activity values cannot be treated as homogeneous data, as these have been obtained from different MTB strains and with different protocols. Although biological data were deemed unsuited for a quantitative analysis, they were used in a semi-quantitative approach,[19] by using the RP classification technique.[20] Learning set compounds were categorized into three groups: "active", "medium-active" and "inactive". The boundaries between these class are shown in Table 1.

Table 1. Classification of compounds in the learning set.

Class	MIC Range [mM]	No. Compds	Percent
active	$MIC < 1\times10^{-2}$	88	18.68
medium-active	$1\times10^{-2} \le MIC < 4\times10^{-2}$	175	37.16
inactive	$4\times10^{-2} \le MIC$	208	44.16

Step 2: RP analysis of the learning set

Recursive partitioning is a statistical classification method well suited to the analysis of high-throughput screening data, as it can both account for multiple structurally diverse classes of compounds, and tolerate a certain amount of erroneous input data.[20] This approach divides data sets along decision tree structures through the use of two-state (absent/present, no/yes) molecular descriptors. During the partitioning of a set of active, medium-active and inactive compounds, RP attempts to identify descriptor pathways that significantly enrich some terminal nodes with correctly predicted compounds, irrespective of their activity class. The resulting RP models can then be used to search databases for additional active compounds. (Details on the RP method are provided in Supporting Information). Three distinct statistical parameters can be used to evaluate the quality of a RP model: a) Interclass prediction: only molecules in the corresponding class are considered. For each class this value represents the number of correctly predicted compounds as percentage of the total number of compounds known to belong to that class; b) Overall prediction: all molecules in the set are considered. For each class this value indicates the number of correctly predicted compounds as a percentage of compounds predicted to belong to that class; c) Enrichment: for each class, this value represents the percentage of compounds correctly predicted divided by the percentage of compounds belonging to that class.

A total of 392 descriptors, belonging to 12 different families, were calculated for all molecules. The dimensionality of the full set of descriptors was then reduced by principal component analysis (PCA), resulting in a significant decrease in the number of variables. 379 compounds (approximately 80% of the learning set) were then selected for the training set, while the remaining 92 compounds were used as a test set. Training set selection was performed with the aim of maximizing the structural diversity of the selected compounds, as encoded by the first three principal components (PCs), while maintaining for the training set a distribution profile similar to that of the learning set.

To identify the most appropriate set of descriptors, 12 RP models were generated on training set compounds, starting from each family of descriptors. Five additional RP models were obtained from other sets (derived from different combinations of the above 392 descriptors): the whole set, the set of 2D descriptors, the set of 3D descriptors, a default set provided by the program, and a set of 94 descriptors (1D or 2D descriptors, belonging to E-state keys, information-content, structural, thermodynamic, topological and ADME families) referred to by the program as "fast descriptors", being characterized by a very short computation time (a comprehensive list of all 94 descriptors is available in Supporting Information). The ability of such models to predict the activity class (active, medium-active, or inactive) of the 92 test set compounds was then evaluated on the basis of the statistical parameters described above: Class% ObsCorrect (interclass prediction), Overall% PredCorrect (overall prediction), and Enrichment. Being interested in an accurate prediction for the whole dataset, to

choose the RP model most suitable for our VS purposes we resorted to the analysis of the values that the overall prediction parameter (Overall % PredCorrect) assumed for all three classes.[21] Accordingly, we selected the RP model, derived from the set of 94 fast descriptors, which showed the highest average value of the overall prediction parameter (Table 2). This model consisted of 41 terminal nodes of which 13 were classified as active, 15 as medium-active, and 13 as inactive (statistical parameters relative to test set predictions are reported in Table 3). The corresponding decision tree is shown in Figure 4.

Table 2. Values of the Overall % ObsCorrect parameters for all the RP models generated starting from each of the 17 different available sets of descriptors.

Descriptor Set	"Active"	"Medium-Active"	"Inactive"	Average
Structural	44.44	57.14	71.74	57.77
Information	38.89	48.78	66.67	51.45
Spatial	40.00	41.18	68.42	49.87
E state keys	78.57	51.11	78.79	69.49
Topological	28.00	54.55	76.47	53.01
Thermodynamic	47.37	50.00	74.36	57.24
Electronic	30.77	52.17	72.73	51.89
QM	28.57	42.50	55.26	42.11
Catshape	30.00	48.48	61.54	46.67
Hypofit	47.62	65.38	60.00	57.67
3D fingerprint	54.55	66.67	77.50	66.24
3D feature count	66.67	58.82	67.35	64.28
Default descriptors	53.33	48.72	71.05	57.7
2D descriptors	55.56	57.58	78.05	63.73
3D descriptors	40.91	42.86	67.86	50.54
Fast descriptors	57.89	67.65	84.62	70.05
All	56.25	51.28	75.68	61.07

Table 3. Statistical parameters found for the fast descriptor family derived RP model applied to test set compounds.

Class	Class % ObsCorrect	Overall % PredCorrect	Enrichment
active	61.11	57.89	2.96
medium-active	71.88	67.65	1.94
inactive	78.57	84.62	1.85

Notably, the statistical parameters of Tables 2 and 3 referred to the test set, not only suggest the most suitable model derived from the fast descriptor family, but also demonstrate its good external predictive power. Moreover, the use of only fast descriptors (requiring minimum CPU time) makes the model particularly well suited to being applied to large sets of compounds.

Step 3: RP-centered VS (RPCVS) protocol

The RP model described above was used in a VS protocol in the search for antitubercular compounds. This consists of the following three filters: 1) Lipinski's rule of five: compounds with one or more violations to Lipinski's parameters are discarded. Owing to the high lipophilicity of the mycobacterial cell wall, the maximum value allowed for log *P* was raised from 5 to 6 so as not to penalize the more hydrophobic compounds, which (presumably) can more easily penetrate the cell wall. 2) RP model: each compound satisfying the first filter is classified as either "active", "medium-active" or "inactive" by the RP model. Only active compounds are retained. 3) Ranking criterion: compounds classified as active are ranked according to a "penalty factor" calculated on the basis of the similarity (encoded by molecular descriptors) to the most active compounds belonging to the training set (see Experimental Section). Molecules with the lowest penalty factor values are considered hit compounds.

The first three steps of the computational approach allowed us to set up the RPCVS protocol to screen databases of virtual and real compounds for new antitubercular agents. For this purpose, we generated a virtual library and submitted this to sequential filtering to decrease its complexity (number of possible hit compounds).

Step 4: virtual library generation: the large library (LL)

In general, VS has become attractive for the computational filtering of large databases of compounds, in order to evaluate their properties and thus identify preferred compounds and eliminate those having undesired features. However, VS is not limited to existing compounds: a virtual library can be generated, using a computational approach, with the aim of reducing the costs of the drug discovery process.[22]

We decided to generate a large virtual library by using an analogue building approach—the systematic insertion of a set of side chains at different positions of a core structure (Figure 2, step 4). This structure was selected on the basis of an analysis of the common chemical features of compounds belonging to the learning set, which showed that one benzene ring (or more) is common to 380 of the 471 compounds analyzed (over 80 %). This analysis also suggested the substitution pattern for the benzene core. Three substitution points (R groups) were defined at positions 1, 2, and 4 (R^1, R^2 and R^3 in Figure 2), with 173 substituents specified for each R group.[18] Of these, 120 fragments (with a wide structural variety) were provided by the program as an internal structural dictionary (a database of fragments), while the remaining 53 were derived from the analysis of the learning set compounds that allowed us to identify structural motifs and privileged substructures associated with antitubercular activity. The simultaneous use of these two distinct sets of substituents was aimed at both covering a wide chemical space and biasing the library toward compounds with antitubercular activity. The resulting library (Figure 5) consisted of over 5 million compounds, resulting from the systematic combination of all substituent groups at each position ($173 \times 173 \times 173 = 5\,177\,717$).

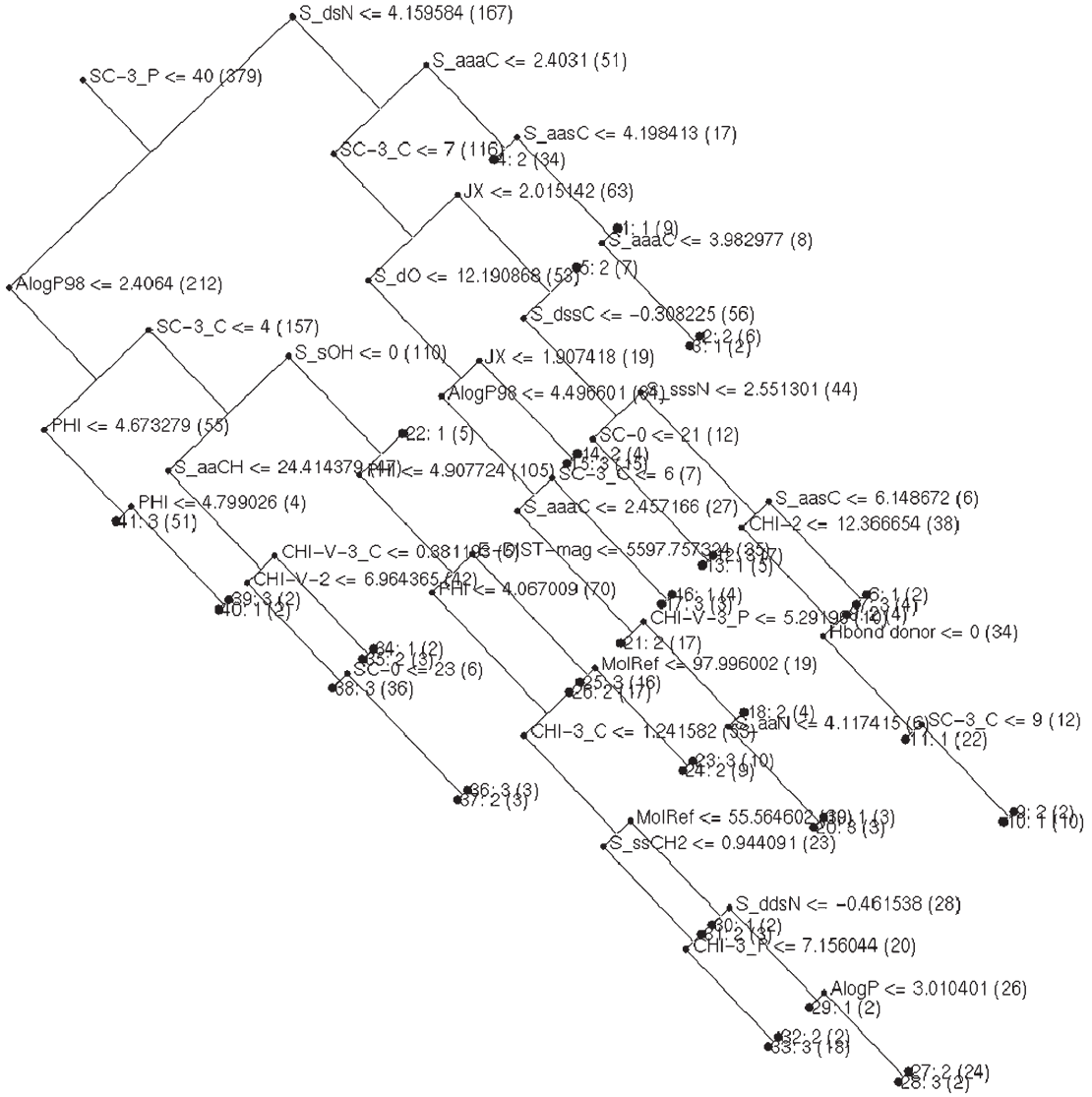
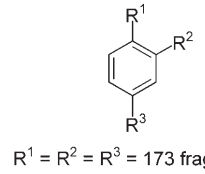


Figure 4. Decision tree derived from recursive partitioning applied to the fast descriptor family calculated on training set compounds. For each terminal node, the number of the node, the activity class (1: active; 2: medium-active; 3: inactive), and the number of compounds contained within the node are reported.



$R^1 = R^2 = R^3$ = 173 fragment set

Figure 5. Markush representation of the large library (LL).

Step 5: a two-part virtual library design: the small library (SL)

In view of the size of, and redundancy within the LL, a small subset (5000 compounds, SL) was extracted using a two part library design procedure. The SL was designed to cover uniformly the chemical space of the original library (to retain the maximum structural information) and to contain druglike molecules. Accordingly, molecules were selected by referring simultaneously to maximal structural diversity, and Lipinski's rule of five ($\log P = 6$), as design criteria.

In general, a library design protocol consists of the identification of an objective function (a weighted combination of terms, each reflecting a desired property of the reduced library), and an optimization step to identify reduced libraries that exhibit improved value of the objective function.[23] (Additional details of the library design are in Supporting Information.) A successful optimization procedure generates reduced libraries with maximally improved values for the objective function (that is, with optimized design properties). In both parts of our library design procedure the objective function consisted of two terms (one for each design criterion) accounting for the structural diversity of compounds (expressed by molecular descriptors), and for their drug-likeness (Lipinski's rule of five).

The conventional approach to library design includes the enumeration of structures and descriptor calculation for the entire library prior to the optimization procedure.[23,24] Enumeration of structures and properties for a library containing over 5 million compounds requires substantial computational and storage resources, increasing time and cost. Furthermore, in our plans, the library design procedure should not be limited, if possible, by the size of the virtual library, and thus should be applicable to even larger libraries. On the basis of such limitations, we adopted the "on-the-fly optimization" (OTFO) approach,[24] recently proposed as a suitable alternative to the conventional approach, especially when dealing with extremely large virtual libraries. Unlike the conventional approach, OTFO requires descriptor computation for only the reduced libraries selected during the optimization procedure. Consequently, calculation of descriptors and storage of data for all LL molecules are not necessary. Where, as in our case, descriptors used in the calculations are additive in nature (see below), these can be calculated without enumeration of structures, affording further time and storage savings.[25]

Unfortunately the OTFO approach is restricted to a combinatorial library design. Unlike the conventional approach, the resulting reduced library consists solely of combinations of the fragments selected for each R group; compounds cannot be selected individually ("cherry picked"). Consequently in the first part of the library design the OTFO approach was only applied to reduce the size of the LL, resulting in an intermediate library (the medium library (ML), Figure 2). The ML size was small enough to allow the application of the conventional approach in the second part of the library design. However, the ML should be as large as possible to maintain the maximum information content with respect to the parent library, thus reducing the limitation due to the combinatorial approach.

For the selection of ML from LL, structural diversity was assessed using BCI fingerprints (1052 bits)[26] as descriptors and "Fingerprints On Bits" as the diversity metric. Lipinski's parameters were used as upper bounds for molecular weight (M_r), number of hydrogen bond acceptors (HBA), number of hydrogen bond donors (HBD) and log *P* (Slog *P*)[27] descriptors. Owing to the high lipophilicity of the mycobacterium cell wall, the maximum value allowed for log *P* was increased from 5 to 6 to avoid the penalization of more lipophilic compounds, which presumably penetrate the cell wall more easily. Molecules whose descriptors exceeded one or more of the threshold values were likely to be discarded. As in this case, violation of the Lipinski's rule is not intended to exclude a compound, but contributes to a penalty function that reduces the possibility that it is retained in the reduced library.

A series of calculations suggested that a ML containing 50×50×50 compounds was the optimal solution, combining the most appropriate library dimension with an acceptable computational demand. 50 substituents were selected for each R group (combinatorial design) using a Monte Carlo (MC) algorithm as the optimization method.

Using this strategy the selection of SL from ML was achieved through a cherry picking method and was not limited by combinatorial constraints. The ML was entirely enumerated and 94 fast descriptors (the same as that used to build the RP model, above) were calculated for all 125000 compounds. PCA was then applied to reduce the number of variables and to remove redundant information encoded in the original descriptors. Structural diversity was evaluated in the space of the first eight PCs (accounting for over 70% variance in the original dataset), and the MaxMin function, belonging to distance-based methods of selection, was used as a diversity metric.[28] Lipinski's parameters were used as upper bounds for M_r, HBA, HBD, and log *P* (Alog *P*)[29] descriptors, penalizing molecules exceeding one or more of these. A 5000 compound noncombinatorial library was then selected (SL), using a genetic algorithm (GA) and a simulated annealing (SA) procedure for subset optimization.

In summary, the realization of the SL through two successive library design procedures allowed us to circumvent the limitations associated with both the OTFO and conventional approaches.

Figure 6 shows SL entries (black) uniformly spanning the descriptor space (first three PCs, accounting for about 50% of the original variance) occupied by the ML (grey), in agreement with the first selection objective represented by the maximal structural diversity encoded in the molecular descriptors.

With respect to the Lipinski's rule of five (representing the second design criterion), Figure 7 shows the distribution profiles of M_r, HBA, HBD and Alog *P* descriptors for the 5000 molecules of SL. All descriptors satisfy the above requirements, thus demonstrating the effectiveness of this selection procedure.

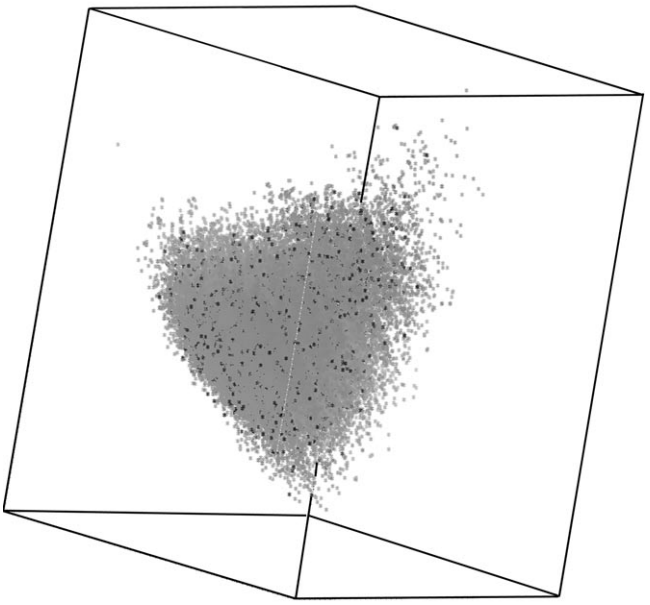


Figure 6. Projection of compounds belonging to both ML (grey) and SL (black) in the space of the first three PCs.

Step 6: virtual screening: the targeted library (TL)

The SL was submitted to the RPCVS protocol previously described. The RP model partitioned the 5000 entries of the SL into active (574), medium-active (1732) or inactive (2694) com-

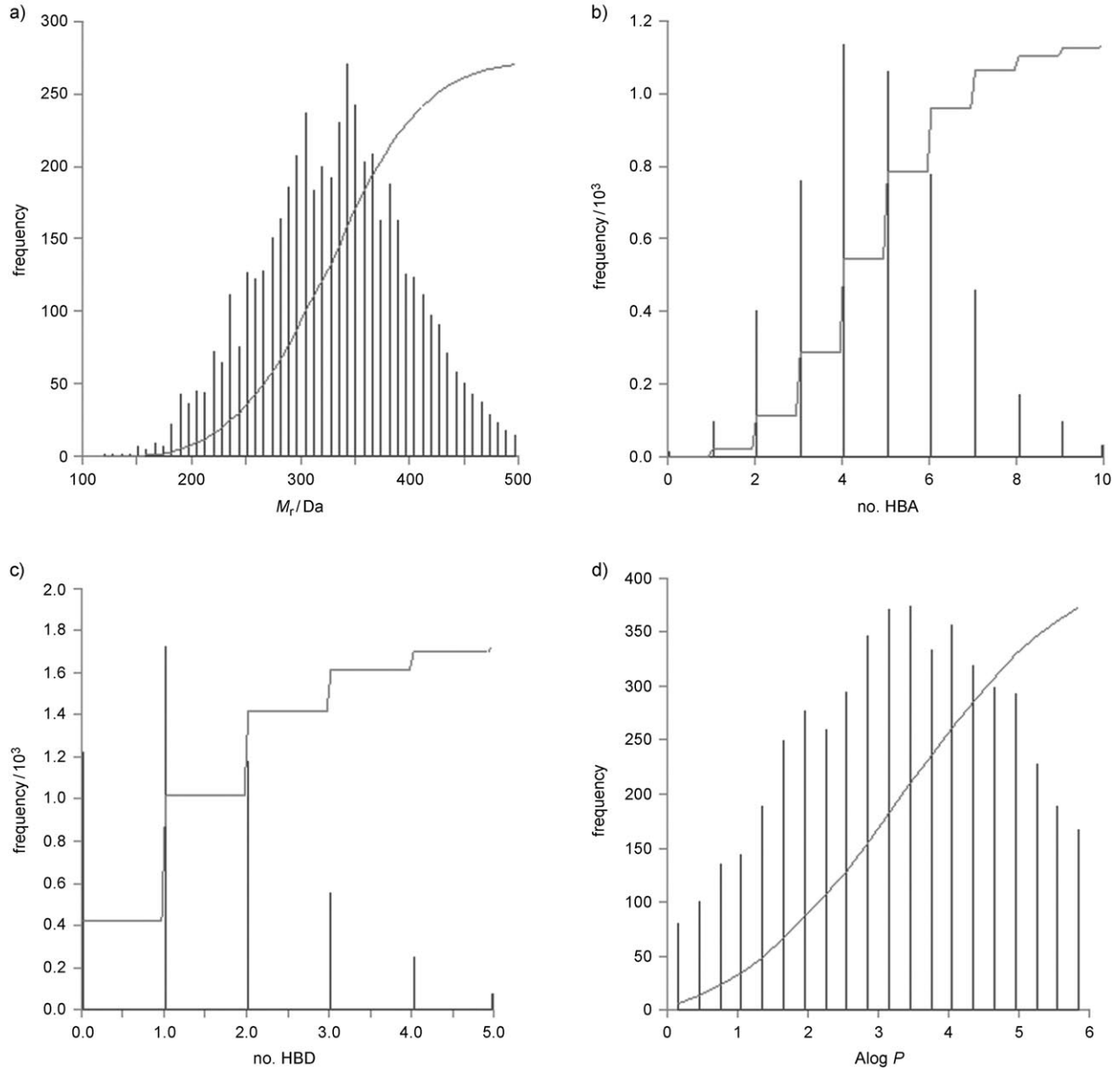


Figure 7. Distribution profiles of a) M_r, b) HBA, c) HBD, and d) Alog*P* descriptors for the 5000 molecules of SL.

pounds. Molecules predicted as active were then ranked according to the penalty factor, and those with a value higher than 1.0 were discarded.[30] Note that Lipinski's rule of five was not applied as a filtering function at this stage because the previous steps of the selection procedure (from LL to ML and from ML to SL) had already rejected compounds with unacceptable violations (see the distribution profiles of M_r, HBA, HBD and log*P* descriptors of SL compounds in Figure 7).

The resulting TL contained 114 compounds (about 2% of SL), classified as "active" and having a penalty factor less than 1.0. Figure 8 shows the projection of the TL (black) and SL (grey) compounds in the descriptor space.

A preliminary evaluation of the pharmacokinetic properties of the TL entries was performed by means of the Volsurf approach,[31] in order to assess their physicochemical properties. Permeability and solubility models implemented in Volsurf (Figure 9) suggested similar pharmacokinetic properties for all TL compounds (yellow), all of which are located in the high-permeability region (red: high, blue: low permeability) and medium- to low-solubility regions (red: high, white: medium, blue: low solubility).

Volsurf analysis confirmed the drug-likeness of these molecules and their lipophilic properties. This is in agreement with the criteria used to select the compounds and suggests that their physicochemical properties should facilitate transfer across the mycobacterium cell wall.

Step 7: database search of commercially available compounds

We checked commercial databases for the presence of TL virtual compounds, since these could be purchased and tested in

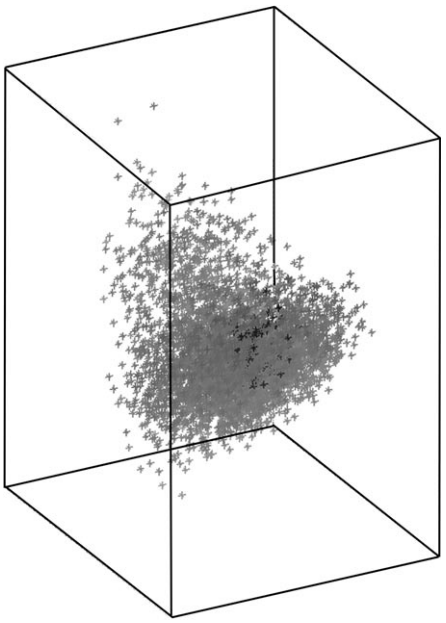


Figure 8. Projection of TL (black) and SL (grey) compounds in the descriptor space: molecules of the TL are located in a restricted region of the descriptor space relative to the whole SL.

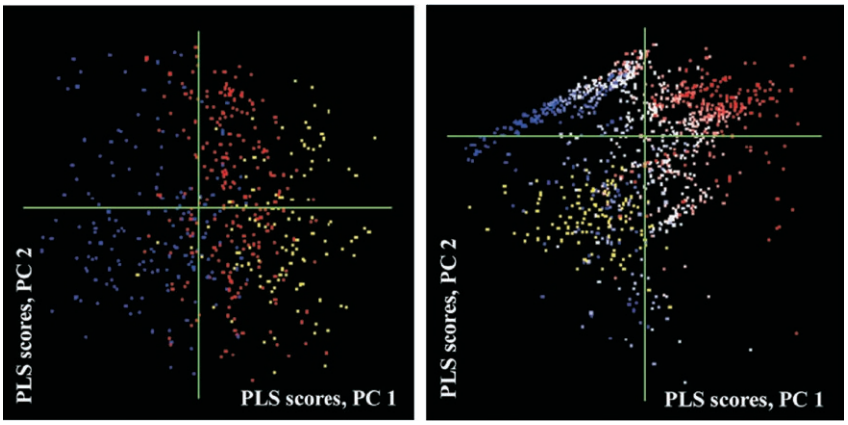
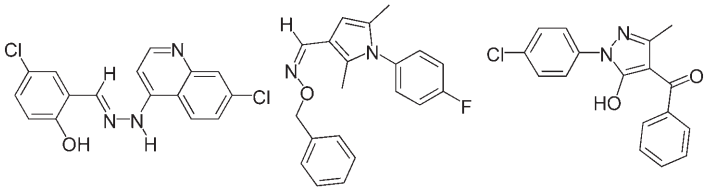


Figure 9. Evaluation of the pharmacokinetic properties of compounds belonging to TL (yellow). Left: TL compounds are located in the space of high-permeability compounds (red), whereas low-permeability compounds are in blue. Right: TL compounds are located in the space of low- to medium-solubility compounds (blue and white, respectively); high-solubility compounds are represented in red.

a very short time (essentially reduced to the time required for testing against MTB). This provided biological data useful for assessing the effectiveness and reliability of our VS protocol before undertaking the time-consuming and costly step of synthesis.

A number of various-sized databases of commercially available compounds were downloaded from the Internet,[18] providing approximately 2 million known compounds. A search for the 114 TL compounds was made, based on the projection of compounds belonging to both TL and the databases in a common descriptor space. Common compounds are identified by their co-location within the descriptor space. Unfortunately, the database search resulted in the identification of only one entry (compound **I**), found within the Asinex database.



I **II** **III**

Step 8: focused library from the Asinex database

In order to increase the number of compounds to be submitted to biological evaluation, we decided to select, from the Asinex database, a small-sized library based on the ten best ranking TL compounds (those with the lowest penalty factor). The Asinex database and the ten selected TL hits were projected onto a common descriptor space. (The position of each compound is determined by its calculated descriptor values.) On the basis of the similar property principle,[32] we assumed that Asinex molecules located very close to the preferred hits should be considered as candidates for testing. Accordingly, for each preferred hit, the 20 closest Asinex molecules were selected, resulting in a 200-member focused library that was then filtered through the RPCVS protocol:[14] 64 compounds were classified as active, 101 as medium-active, and 35 as inactive (penalty factors 0.237 to 1.417). From the 60 molecules both classified as active by RP and with a penalty factor of less than 1.0, three compounds with the best score and belonging to different structural classes were selected.

Concluding the selection process (Figure 2), we purchased four compounds from the Asinex database and evaluated these for their antitubercular activity. Compound **II** showed promising antimycobacterial activity (MIC = 25 $\mu g\,mL^{-1}$) and was therefore chosen as a lead compound for further development, while the remaining compounds, including **I**, were found to be inactive (MIC > 100 $\mu g\,mL^{-1}$).

Step 9: in silico screening of the whole Asinex Gold Collection database

Encouraged by the result of the biological assays, we decided to filter the whole Asinex database by combining our pharmacophoric model for antitubercular compounds (detailed in Supporting Information) and the RPCVS protocol[14] (Figure 3).

A preliminary treatment of the Asinex database with the pharmacophore model was required as it is a noncombinatorial database and too large to permit the application of a library design based on the protocol previously described. In fact, OTFO is ineffectual for noncombinatorial databases, while our computational resources impose a limit of 125 000 compounds (conventional approach, see above).

The Asinex database (>200 000 compounds) was filtered through the pharmacophoric model, rejecting all compounds unable to map simultaneously all the features of the model, The resulting library (43 845 entries, about 20% of the original database) was further filtered by the RPCVS protocol: only 447 compounds were retained showing no violation of Lipinski's rule of five ($\log P = 6$), classified as active by the RP model and with a penalty factor less than 1.0. Nine of these, showing the best fit against the pharmacophoric model and belonging to different chemical classes, were purchased. Biological assays revealed that compound **III** possessed a MIC value of 25 $\mu g\,mL^{-1}$, so this was selected as a second lead compound. Figure 10 shows the good superposition of compound **III** to the pharmacophoric model, used as the first filtering criterion.

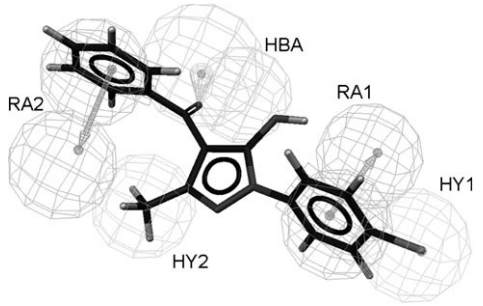


Figure 10. Compound **III** mapped to the pharmacophoric model. HY1, HI2: hydrophobic regions; RA1, RA2: aromatic rings; HBA: hydrogen bond acceptor.

By applying different in silico approaches we were able to identify two lead compounds active against MTB: compounds **II** and **III**. Compound **II** contains a pyrrole, and **III**, a pyrazole ring which is not unusual for antitubercular agents. However, despite the five-membered ring moiety, the structures of these candidates differ substantially from those of pyrroles and pyrazoles reported as antituberculosis compounds. Whereas pyrroles are well-established as antitubercular compounds,[11,15,33–35] pyrazoles have been poorly investigated thus far,[36] lending further value to our findings. Interestingly, both lead molecules contain a substructure motif falling within a more generic scaffold (Figure 11) which has been recently proposed as a minimum common bioactive substructure (MCBS) responsible for the activity of different chemical classes of antimycobacterial agents.[37]

Step 10: lead development: synthesis of new analogues of II and III

Since Lipinski's parameters were involved in the selection process, both **II** and **III** were expected to be suitable for drug development. In fact, as well as showing good anti-MTB activity, they should also have a good pharmacokinetic profile. Furthermore, their relatively simple chemical structures led us to conclude that the synthesis of their analogues would be feasible. Consequently a series of derivatives for both lead compounds was designed and synthesized with the purpose of obtaining compounds with improved antimycobacterial activity, and data for a preliminary SAR analysis.

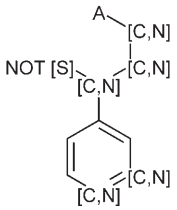
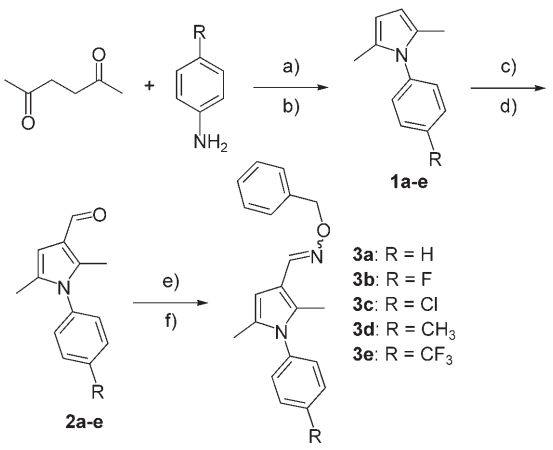


Figure 11. Minimum common bioactive substructure (MCBS) recently reported as responsible for antimycobacterial activity. Interestingly, both pyrrole and pyrazole derivatives identified by the computational protocol reported herein belong to this MCBS. A: any atom; [C,N]: carbon or nitrogen atom; NOT[S]: any atom but not sulfur.

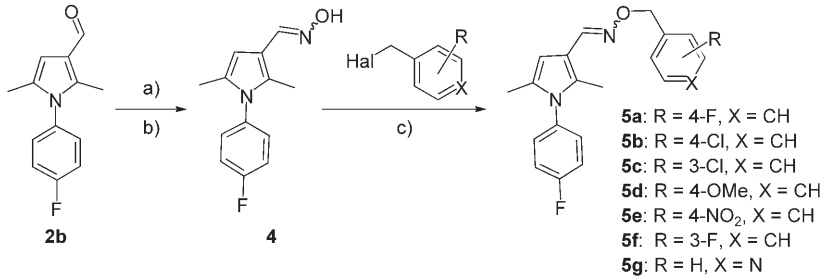
Compounds **3** and **5** were designed from **II** to assess the influence on anti-MTB activity of the electron-donating and electron-withdrawing capacity of small substituents. Compounds **9a** and **9b** were synthesized to investigate on the role of diverse alkyl groups at positions 2 and 5. To gain a preliminary insight on SAR of pyrazole derivatives, a series of congeneric compounds of **III** were obtained with different substituents at both the phenyl rings and at the hydroxy group at the position 5. The benzoyl moiety at the position 4 was also addressed: its phenyl ring was replaced by different (hetero)cyclic moieties (**11l**–**o**), and the carbonyl group was transformed into various substituted imino moieties (**12a**–**k**).

Chemistry. To synthesize efficiently a series of derivatives of our lead compounds, **II** and **III**, we set up a method for the solution-phase parallel synthesis of pyrrole and pyrazole analogues. 2,5-Dimethylpyrroles **3a**–**e** were synthesized using a Büchi Syncore parallel synthesizer in a three-step parallel procedure (Scheme 1). 2,5-Hexandione was placed in five reaction vessels and reacted with five different anilines in excess in order to drive the reactions to completion. Once the reactions had completed acidic solid-support scavenger (polymer-bound *p*-toluenesulfonic acid) was added to remove the excess anilines and pyrroles **1a**–**e** were obtained following a simple parallel filtration step (>95% purity). These were then subjected to a parallel Vilsmeier–Haack reaction with $POCl_3$ (in DMF, 100 °C, 2 h).[38] The reactions were performed in parallel, the solvents evaporated under vacuum, and the residues filtered through a silica gel pad to remove polymeric side products, resulting in **2a**–**e** (yield >90%). Compounds **2a**–**e** were submitted in parallel to treatment with benzylhydroxylamine in refluxing benzene and in the presence of 4 Å-molecular sieves to take the reactions to completion.[39] Acidic scavenger was then added to remove excess amine, leading to compounds **3a**–**e** after simple parallel filtration (>95% purity).

2b was treated with hydroxylamine to afford **4**, from which compounds **5a**–**g** (Scheme 2) were obtained in good yield by



Scheme 1. Parallel solution-phase synthesis of pyrroles **3a–e**: a) Syncore, $T = 130\,^{\circ}C$, 300 rpm, $t = 24$ h; b) toluene, *p*-toluenesulfonic acid, polymer-bound; c) $POCl_3$/DMF (1:1), DMF, 100 °C; d) filtration through silica gel pad; e) $PhCH_2ONH_2{\cdot}HCl$, Syncore, reflux, 300 rpm, $t = 3$ h; f) benzene, *p*-toluenesulfonic acid, polymer-bound.



Scheme 2. Parallel solution-phase synthesis of pyrroles **5a–g**: a) $HONH_2{\cdot}HCl$, Syncore, reflux, 300 rpm, $t = 3$ h; b) benzene, *p*-toluenesulfonic acid, polymer-bound; c) Syncore, $T = 25\,^{\circ}C$, 300 rpm, NaI, overnight.

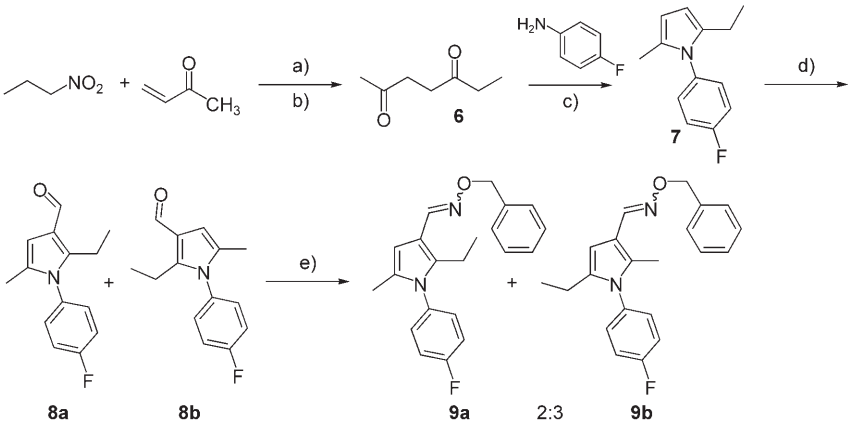
parallel reactions with seven different alkylating agents (chlorinated or brominated, RT). Compounds **3a–e**, **4**, and **5a–g** exist as *E*/*Z* isomers at the C=N bond. Compound **2b** was chosen as the intermediate for **5a–g** since, among pyrrole derivatives, the *p*-fluorophenyl group at N1 of the pyrrole ring has been shown to be a good substituent for antimycobacterial activity.[11]

To expand the structure–activity relationships for pyrrole derivatives, we also synthesized 2-ethyl-5-methylpyrroles (Scheme 3). In a separated reaction vessel 2,5-heptandione **6** was synthesized following a published procedure.[40] This was reacted with 4-fluoroaniline to afford 2-ethyl-5-methylpyrrole **7**, that was then converted in two steps (using the parallel procedure described above) into the two regioisomers **9a** and **9b** (ratio 2:3, as revealed by HPLC analysis). These compounds were separated by a semi-preparative HPLC column (simple chromatography was unsuccessful), resulting in sufficient amounts of each to confirm their structure by ^1H NMR spectroscopy. The 2:3 mixture was also submitted to preliminary antimycobacterial testing.

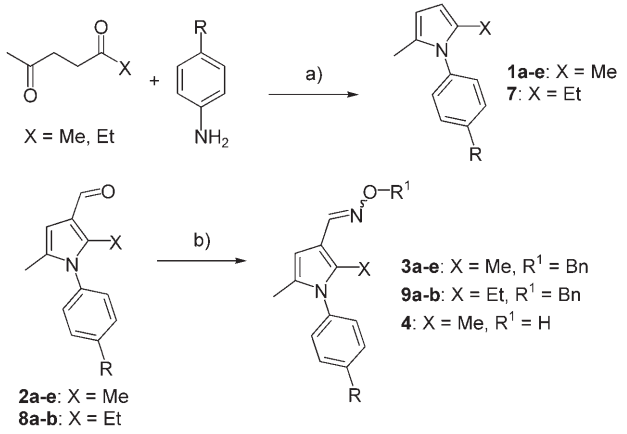
Based on our previous experience of microwave-assisted chemistry[41] and with the aim of decreasing reaction times, we also investigated the possibility of performing the previous reactions under microwave irradiation (Scheme 4). Compounds **1a–e**, **7**, **3a–e**, **9a–b**, and **4** were obtained in only 10 min instead of the 24 h (**1a–e** and **7**) and 3 h (**3a–e**, **9a–b**, and **4**) which are required by the normal procedure.

Analogues of the lead compound **III** were also synthesized by a parallel reaction approach. Pyrazolone **10a** was commercially available, while pyrazolones **10b–c** were prepared from ethylacetoacetate and arylhydrazines under microwave irradiation in few minutes (Scheme 5).[42] **III** was easily prepared from pyrazolone **10b** following a reported procedure,[43] and acylpyrazolones **11a** and **11e–o** were similarly synthesized from **10a–c** as follows. Pyrazolones were partitioned into 12 reaction vessels (**10a** into nine vessels, **10b** into one, and **10c** into two) on a Büchi Syncore parallel synthesizer and reacted in refluxing dioxane with different acyl chlorides in the presence of $Ca(OH)_2$. After completion of all reactions, dioxane was evaporated and compounds **11** were precipitated in 3 N HCl, filtered and recrystallized from EtOH (Scheme 6). Compounds **11a**, **11g**, **11h**, and **11j–o** were obtained from **10a**; **11e** was obtained from **10b**; and **11f** and **11i** were obtained from **10c**. Compounds **11b–d** were purchased.

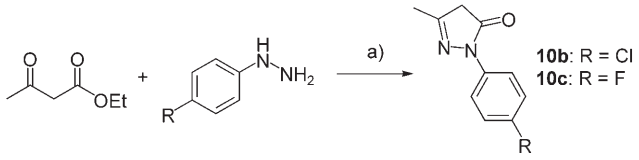
The possibility of carrying out these reactions under microwave irradiation was also investigated. Compounds **11a** and



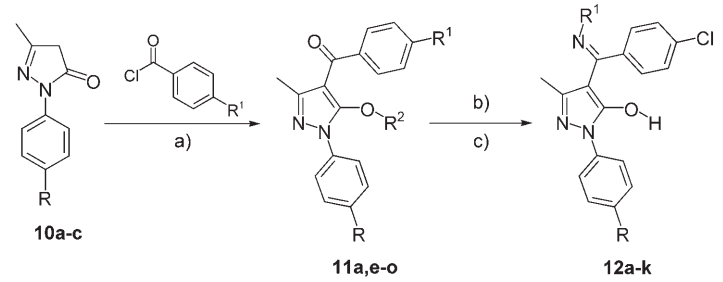
Scheme 3. Parallel solution-phase synthesis of pyrroles **9a** and **9b**: a) Al_2O_3, 0 °C; b) H_2O_2, K_2CO_3, 0 °C; c) Δ, Syncore, toluene, $T = 130\,^{\circ}C$, $t = 24$ h; d) $POCl_3$/DMF (1:1), DMF, 100 °C; e) $PhCH_2ONH_2{\cdot}HCl$, Syncore, benzene, reflux, 3 h.



Scheme 4. Microwave-assisted synthesis of pyrroles: a) Δ, MW, $T = 180\,^{\circ}C$, $t = 2\times5$ min; b) $HONH_2{\cdot}HCl$ or $PhCH_2ONH_2{\cdot}HCl$, MW, DMF, $T = 120\,^{\circ}C$, $t = 2\times5$ min.



Scheme 5. Microwave-assisted synthesis of pyrazolones **10b** and **10c**: a) MW, $T = 80\,^{\circ}C$, $t = 10$ min.

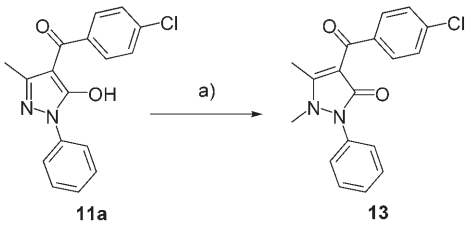


Scheme 6. Parallel solution-phase synthesis of pyrazoles **11a**,**e**–**o** and **12a**–**k**: a) Syncore, $Ca(OH)_2$, dioxane, reflux, 3 h or $Ca(OH)_2$, dioxane, MW, 5 min; b) $R^1{-}NH_2$, Syncore, Na_2SO_4, EtOH; c) *p*-toluenesulfonic acid, polymer-bound.

11e–**o** were obtained in good yield in only 5 min, instead of the 3 h required by the usual procedure (Scheme 6).

To enlarge the SAR for pyrazoles, we synthesized a series of derivatives, **12a**–**k**, where the carbonyl group of the C4 benzoyl moiety was replaced by functionally distinct imino groups. Pyrazoles **11a** and **11e** were partitioned into 11 reaction vessels (**11a** into seven vessels, **11e** into four) on a Büchi Syncore parallel synthesizer and reacted in refluxing ethanol with different primary amines in the presence of Na_2SO_4 (Scheme 6). Once the reactions were finished acidic solid-support scavenger (polymer-bound *p*-toluenesulfonic acid) was added to remove excess amine to obtain **12a**–**k** following a simple parallel filtration step (>90% purity).

The methyl pyrazolone **13** was synthesized from **11a** in refluxing toluene in the presence of Me_2SO_4 (Scheme 7) to investigate further the influence of the C5 substituent on anti-MTB activity.



Scheme 7. Synthesis of pyrazolone **13**: a) Me_2SO_4, toluene.

Biology

Compounds were assayed for their inhibitory activities toward *M. tuberculosis* H37Rv (ATCC 27294) and *M. avium* (ATCC 19421). The minimum inhibitory concentration ($\mu g\,mL^{-1}$) was determined for each compound (Table 4).

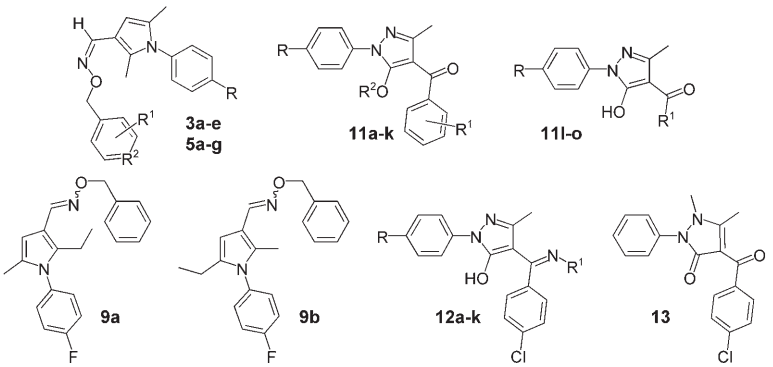
No improvement in activity was obtained for the pyrrole derivatives of **II**, so this group was not investigated further. However, MIC values for the pyrazoles decreased from 25 $\mu g\,mL^{-1}$ (lead compound **III**) to 12.5 (**11f**) and 6.25 $mg\,mL^{-1}$ (**11a** and **11e**). Owing to the small size of the pyrazole group, a structure–activity relationship analysis allowed only a few observations. Introduction of a *p*-chloro substituent of the benzoyl group at the position 4 led to enhanced activity relative to the parent compound (**11a**, **11e**, and **11f**), but a methyl or nitro substituent was detrimental (**11h** and **11b**). Modifications to the phenyl ring or the carbonyl group of the 4-benzoyl moiety led to a decrease in activity. Several of the new compounds were also tested for their inhibitory effect on *M. avium*, though none showed an activity better than 50 $\mu g\,mL^{-}$(**5d**).

Conclusions

A ligand-based virtual screening approach has allowed the identification of compounds exhibiting inhibitory activity toward MTB. This activity was optimized by synthesizing additional pyrazole derivatives by using parallel solution-phase and microwave-assisted synthesis approaches.

Both computational and chemical procedures were structured to optimize the efficiency of each step in terms of time and costs. For the calculations, both the use of fast descriptors (for generating a recursive partitioning model) and avoiding enumeration of the LL members contributed to a reduction of CPU demand. Searches of databases of commercially available compounds allowed the identification of putative hits. These were purchased and submitted to biological testing in order to validate computational models and avoid the time-consuming and costly step of synthesis. For the synthesis, pathways were designed for compounds identified as "actual hits" by application of efficient approaches (i.e., parallel solution-phase synthesis), using commercially available reagents. Reaction time has

Table 4. Antimycobacterial activity of compounds **3**, **5**, **9**, and **11**–**13**.



Compd	R	R^1	R^2	MIC [μg mL^{-1}] *M. tuberculosis*	MIC [μg mL^{-1}] *M. avium*
3a	H	H	CH	>100	ND
3b (**II**)	F	H	CH	25	ND
3c	Cl	H	CH	100	100
3d	CH_3	H	CH	>100	>100
3e	CF_3	H	CH	100	>100
5a	F	4-F	CH	100	>100
5b	F	4-Cl	CH	50	>100
5c	F	3-Cl	CH	100	>100
5d	F	4-OMe	CH	>100	50
5e	F	4-NO_2	CH	>100	>100
5f	F	3-F	CH	>100	>100
5g	F	H	N	25	ND
9a[a]				>100	ND
9b[a]				>100	ND
III	Cl	H	H	25	ND
11a	H	4-Cl	H	6.25	>100
11b[b]	H	4-NO_2	H	>100	>100
11c[b]	H	2-CH_3	H	50	>100
11d[b]	H	3-F	H	25	>100
11e	Cl	4-Cl	H	6.25	ND
11f	F	4-Cl	H	12.5	ND
11g	H	H	H	50	ND
11h	H	4-CH_3	H	50	ND
11i	F	H	benzoyl	25	ND
11j	H	4-CH_3	*p*-methylbenzoyl	>100	ND
11k	H	4-phenyl	H	50	ND
11l	H	3-piridynyl		>100	ND
11m	H	2-thienyl		50	ND
11n	H	2-naphthyl		25	ND
11o	H	2-styryl		50	ND
12a	H	benzyloxy		>100	ND
12b	H	benzyl		>100	ND
12c	H	4-fluorobenzyl		>100	ND
12d	H	methoxy		25	ND
12e	H	cyclopropyl		>100	ND
12f	H	propargyl		>100	ND
12g	H	propyl		>100	ND
12h	Cl	methoxy		25	ND
12i	Cl	propyl		>100	ND
12j	Cl	benzyloxy		>100	ND
12k	Cl	cyclopropyl		>100	ND
13				>100	ND
isoniazid				0.2	ND

[a] These compounds were tested as a 2:3 mixture (**9a**/**9b**). Considering their inactivity toward MTB, they were separated only to obtain each in sufficient quantity for the confirmation of structures. [b] Compounds purchased from Asinex.

been also optimized by the application of microwave-assisted chemistry.

The results reported herein should be considered preliminary. The intention has been to describe an efficient computational protocol for the virtual identification of compounds endowed with inhibitory activity toward the growth of MTB.

Both the identification of hit compounds by means of the computational model and the optimization of their activity

through chemical procedures have prompted us to plan the application of this virtual screening approach to all databases of commercially available compounds,[44] and the synthesis of additional pyrazole derivatives to enlarge the preliminary SAR analysis.

Glossary

classification structure–activity relationship	CSAR
genetic algorithm	GA
large library	LL
ligand-based	LB
ligand-based virtual screening	LBVS
medium library	ML
minimum common bioactive substructure	MCBS
minimum inhibitory concentration	MIC
Monte Carlo	MC
multidrug resistant	MDR
Mycobacterium tuberculosis	MTB
number of hydrogen bond acceptors	HBA
number of hydrogen bond donors	HBD
on the fly optimization	OTFO
principal component analysis	PCA
principal components	PC
recursive partitioning	RP
recursive partitioning centered virtual screening	RPCVS
simulated annealing	SA
small library	SL
structure–activity relationship	SAR
targeted library	TL
tuberculosis	TB
virtual screening	VS

Experimental Section

Computational details: RP analysis (descriptor computation for learning set compounds, training set selection and RP models generation), library generation, and library design were performed by Cerius^2 software (version 4.8.1).[45] Predictions of permeability and solubility of TL molecules were made with Volsurf (version 3.0.11).[31] The search of the Asinex database (through the pharmacophoric model) was performed by using Catalyst software (version 4.8).[45] All calculations were carried out on a SGI Origin 300 server, R14000, 4×500 MHz.

Building the RP model: RP models were generated for the training set using the following parameters:

- Type of weighting: by observation, i.e., each compound is considered of equal importance to the model, rather than each class;
- Splitting method: Gini Impurity, i.e., the scoring function that determines how groups are partitioned into statistically distinct nodes (and consequently where to split in a growing tree) tries to minimize the impurity of nodes resulting from the split;
- Pruning factor: no pruning, i.e., no pruning procedure is applied to reduce the tree depth;
- Minimum number of samples per node: 2, i.e., each node must contain at least two compounds;
- Maximum number of knots per split: no Knot limit, i.e., the maximum number of ways a descriptor range can be split before statistical significance is evaluated is not defined;
- Maximum tree depth: 10, i.e., the maximum number of node splits that can yield a terminal node is 10;
- Number of cross-validation groups: 10, i.e., 10 random groups are used for cross-validation test.

For each model, the following statistical parameters referred to the test set were calculated: interclass prediction (Class % ObsCorrect), overall prediction (Overall % PredCorrect) and enrichment.

Computation of the penalty factor: The penalty factor was calculated from the 19 most active compounds of the training set (MIC $< 1.5\times10^{-3}$ mM). For these, 94 descriptors (the same used to derive the RP model) were calculated and their mean values were used to define a penalty function.[23] This function was employed to assign a penalty to each molecule: the greater the difference between a descriptor value of a compound and the average value of the reference compounds, the higher the penalty value associated to the compound.

Library generation: The "large library" (LL) was generated with the Analog Builder tool. In a benzene core, three substituents (R^1, R^2, and R^3) were inserted at positions 1, 2, and 4, respectively. For each R group, a set of 173 (120 predefined + 53 biased) substituents was defined, resulting in a combinatorial library of $173\times173\times173 = 5\,177\,717$ compounds. To reduce CPU demand the LL was not enumerated but encoded in a small file (in *rg* format), used as the input to the following computational step (the OTFO procedure).

Library design: *1) Selection of the medium library (ML).* Combinatorial selection of the ML was performed through the OTFO approach, implemented in the LibX module. The Monte Carlo algorithm (the optimization method) was set to run for 10 000 steps with at least 1000 idle steps (no improvement in the optimal result found) before stopping, with a temperature factor of 300 K. *2) Selection of the small library (SL).* Design of the SL was based on the "cherry picking" method, within the conventional approach. A genetic algorithm (GA) followed by a simulated annealing (SA) procedure were applied as optimization methods. The initial population for the GA consisted of 20 random individuals (each corresponding to a subset of 5000 molecules) and the evolution was allowed to proceed for a maximum of 3000 generations, with 100 idle generations before stopping. Parent selection was performed by selecting the best fitting from a random selection of three individuals. For the evolution, the probability of crossover was 20 %, and the probability of mutation was 80 %. The three best fitting individuals resulting from GA were refined by a SA procedure (Table 5). The individual with the best final fitness was then chosen as SL.

Table 5. Simulated annealing protocol used to generate the small library (SL).

Cycle	Steps	*T* [°C]	Idle Steps	Mutations
1	5000	100	500	10
2	5000	10	500	10
3	5000	0	500	10
4	5000	0	500	5
5	5000	0	500	2
6	5000	0	500	1
7	5000	0	500	1

Projection on pharmacokinetic models: TL Compounds were converted from 2D to 3D and then subjected to energy minimization, in order to obtain reliable three-dimensional structures (no additional analysis was performed, since Volsurf descriptors are "hardly influenced by conformational sampling and averaging").[46] Volsurf descriptors were calculated starting from GRID molecular interaction fields derived for water, dry and carbonyl oxygen probes. These were used to project molecules onto two quantitative models provided by the program: the permeability (referred to Caco2 cells) and the solubility models.

Search within databases of commercially available compounds: Databases (SDF format) were downloaded from the Internet in and imported into Cerius^2. The database search method involved the computation of a set of 94 fast descriptors (the same as used to build the RP model) for both the TL and checked databases. PCA was then applied and molecules belonging to the two different sets were projected onto a common descriptor space (first three PC). The use of a large number of descriptors assured that each molecule would be characterized unequivocally, so that two molecules located in the same point of the descriptor space (i.e., with equal values for all 94 descriptors) are expected to be the same molecule. On the other hand, the use of fast descriptors allowed the searching of large databases with minimal computational time. After projection on the space of the first three PCs, the two libraries were compared using the LibCompare module, allowing the identification of molecules with null distance between them. The method described above was applied to each database downloaded from the web and led to the identification of one molecule of the TL within the Asinex database.

Database search by means of the pharmacophoric model for antitubercular compounds: The pharmacophoric model[7] was used as a three-dimensional query to filter the Asinex database, (>200 000 compounds). The database was downloaded as an SDF file and converted into a Catalyst database (i.e., a conformational model consisting of a maximum of 250 conformers was generated for each compound, so as to reproduce the flexibility of the molecule during the database search). The database search was then performed using the 'fast flexible search' method implemented in Catalyst, which retrieves compounds able to map the three-dimensional query represented by the pharmacophoric model, and finds the best fit among the conformers. Only molecules mapping all the pharmacophoric features were considered as hits resulting in the identification of about 44 000 compounds.

Chemistry: Reagents were obtained from commercial suppliers and used without further purification. Dioxane was dried over sodium/benzophenone prior to use. Anhydrous reactions were performed under a positive pressure of dry N_2. Merck silica gel 60 was used for flash chromatography (23–400 mesh). ^{1}H NMR spectra were measured at 200 MHz on a Bruker AC200F spectrometer. Chemical shifts are reported relative to $CDCl_3$ at $\delta = 7.24$ ppm and tetramethylsilane at $\delta = 0.00$ ppm. Büchi Syncore polyvap was used for parallel synthesis, filtration and evaporation.

HPLC and MS analysis: The purity of compounds was assessed by reversed-phase liquid chromatography and a mass spectrometer (Agilent series 1100 LC/MSD) with a UV detector at $\lambda = 254$ nm and an electrospray ionization (ESI) source. All the solvents were HPLC grade (Fluka). MS data were obtained using an Agilent 1100 LC/MSD VL system (G1946C, 0.4 mL min^{-1} flow rate, methanol/water binary solvent (95:5)). UV detection was monitored at $\lambda = 254$ nm. MS data were acquired in positive mode scanning over the mass range of 50–1500. The following ion source parameters were used: drying gas flow: 9 mL min^{-1}; nebulizing pressure: 40 psig; drying gas temperature: 350 °C.

Microwave irradiation experiments: Microwave irradiation was conducted using a CEM Discover Synthesis Unit (CEM Corp., Matthews, NC). The machine consists of a continuous focused microwave power delivery system with operator-selectable power output (0 to 300 W). The temperature of the contents of the vessels was monitored using a calibrated infrared temperature sensor mounted under the reaction vessel. All the experiments were performed using the stirring option whereby the contents of the vessel are stirred by means of rotating magnetic plate located below the floor of the microwave cavity and a Teflon-coated magnetic stir bar in the vessel.

*Parallel synthesis of 2,5-dimethyl-1-aryl-1H-pyrroles **1a–e** and 2-ethyl-5-methyl-1-(4-fluorophenyl)-1H-pyrrole **7**; general procedure:* 2,5-Hexandione (1 mL, 8.5 mmol), divided into five vessels, and 2,5-heptandione (900 mg, 7 mmol) were placed in the Büchi Syncore and dissolved in toluene (5 mL). The appropriate anilines (1.2 equiv mol^{-1}) were added, and the reaction mixtures were heated (130 °C, 300 rpm, 24 h). To the cooled solutions, *p*-toluenesulfonic acid polymer bound scavenger was added (0.4 equiv mol^{-1}) and the mixtures stirred (300 rpm, 2 h, RT). The reaction mixtures were filtered in parallel with a specific filtration unit (Büchi filtration unit for R-24, cat. no. 015695285) and the scavengers washed twice with CH_2Cl_2 (10 mL). The solvents were completely evaporated in the same apparatus. Compounds **1a–e** and **7** (obtained in quantitative yields) were identified by LC/MS analysis and proved to be pure enough (>95%) for the following step.

*Microwave-accelerated synthesis of 2,5-dimethyl-1-aryl-1H-pyrroles **1a–e** and 2-ethyl-5-methyl-1-(4-fluorophenyl)-1H-pyrrole **7**; general procedure:* 2,5-Hexandione (1 mmol) or 2,5-heptandione (1 mmol) and the appropriate anilines (1.2 equiv mol^{-1}) were mixed in oven-dried pressure vials with magnetic stir bars. The vessels were placed in the microwave oven and heated twice (180 °C, 5 min) under microwave irradiation. To the cooled solutions, *p*-toluenesulfonic acid polymer bound scavenger was added (0.4 equiv/mol) and the mixtures were stirred for 2 h at room temperature. The reaction mixtures were filtered and the scavengers were washed twice with CH_2Cl_2 (10 mL). The solvent was removed under reduced pressure affording **1a–e** and **7**.

*Parallel synthesis of 2,5-dimethyl-1-aryl-3-formylpyrroles **2a–e** and 2-ethyl-5-methyl-1-(4-fluorophenyl)-3-formylpyrroles **8a–b**; general procedure:* Phosphorous oxychloride (6 mmol) was dropped into ice-cooled *N*,*N*-dimethylformamide (12 mL) under stirring and an N_2 atmosphere. The mixture was kept at room temperature (15 min), then partitioned equally into six vessels in the Büchi Syncore. Solutions of **1a–e** and **7** in *N*,*N*-dimethylformamide (5 mL) were then added to the vessels under N_2 atmosphere and heated (100 °C, 300 rpm,3 h). After cooling, 30% NaOH was added to each solution until it became alkaline, and the resulting mixture stirred (15 min, 300 rpm). CH_2Cl_2 (15 mL) was then added to each and the mixtures stirred (10 min). The water layers were removed with a specific filtration unit (Büchi filtration unit for R-24, cat. no. 015695285) and the remaining organic layers were evaporated to dryness in the apparatus. The solid residues were filtered through a pad of silica gel (petroleum ether/ethyl acetate (3:1) as eluent) to remove polymeric side products, affording pure compounds **2a–e** and **8a–b** (70–85% yield).[47]

*Parallel synthesis of 2,5-dimethyl-1-aryl-3-(benzylhydroxyamino)-methylene-pyrroles **3a–e**, 2-ethyl-5-methyl-1-(4-fluorophenyl)-3-(benzylhydroxyamino)methylene-pyrrole **9a**, 2-methyl-5-ethyl-1-(4-fluoro-*

phenyl)-3-(benzylhydroxyamino)methylene-pyrrole ***9b****, and 2,5-dimethyl-1-(4-fluorophenyl)-3-(hydroxyamino)methylene-pyrrole* ***4****; general procedure:* Compounds **2a**–**e** and **8a**–**b** (0.5 mmol) were placed in the Büchi Syncore and dissolved in benzene. Activated 4 Å-molecular sieves were added. Benzylhydroxylamine (1.5 equiv mol^{-1}) or hydroxylamine hydrochloride (1.5 equiv mol^{-1}) was then added and the reactions heated at reflux (300 rpm, 3 h). *p*-toluenesulfonic acid polymer bound scavenger was added (1 equiv mol^{-1}) to the cooled solutions and the mixtures stirred (300 rpm, 2 h, RT). The reaction mixtures were filtered in parallel with a specific filtration unit and the scavengers were washed twice with CH_2Cl_2 (10 mL). The solvents were evaporated completely in the same apparatus. Compounds **3a**–**e**, **4** (all in quantitative yield) and **9a**–**b** were identified by LC/MS analysis and proved to be >95% pure. The final products **3a**–**e** were further purified by flash chromatography (petroleum ether/ethyl acetate 4:1) to afford a >98% purity. **4** was used in the next step without further purification, while **9a** and **9b** were separated by using a semipreparative Agilent ZORBAX ODS 9.4×250 mm 5-μm HPLC column (acetonitrile/water 60:40) to confirm their structure.

Microwave-accelerated synthesis of 2,5-dimethyl-1-aryl-3-(benzylhydroxyamino)methylene-pyrroles ***3a****–****e****, 2-ethyl-5-methyl-1-(4-fluorophenyl)-3-(benzylhydroxyamino)methylene-pyrrole* ***9a****, 2-methyl-5-ethyl-1-(4-fluorophenyl)-3-(benzylhydroxyamino)methylene-pyrrole* ***9b****, and 2,5-dimethyl-1-(4-fluorophenyl)-3-(hydroxyamino)methylene-pyrrole* ***4***: Compounds **2a**–**e** and **8a**–**b** (0.2 mmol) were dissolved in benzene in pressure vials equipped with magnetic stir bar. Benzylhydroxylamine (1.5 equiv mol^{-1}) or hydroxylamine chloridrate (1.5 equiv mol^{-1}) was added. Activated 4 Å-molecular sieves were added to the mixtures and the vessels placed in the microwave oven and heated twice (120 °C, 5 min) under microwave irradiation. *p*-toluenesulfonic acid polymer bound scavenger was added (1 equiv mol^{-1}) to the cooled solutions, and the mixtures stirred (2 h, RT). The reaction mixtures were filtered and the scavengers washed twice with CH_2Cl_2 (10 mL). The solvent was removed under reduced pressure affording **3a**–**e**, **4**, and **9a**–**b**.

Parallel synthesis of 2,5-dimethyl-1-aryl-3-(arylmethylhydroxyamino)-methylene-pyrroles ***5a****–****g****; general procedure:* Compound **4** (100 mg, 0.43 mmol), was divided into seven vessels, placed in the Büchi Syncore and dissolved in dry THF (5 mL). NaH (2 equiv mol^{-1}) was added and the reaction mixtures stirred (300 rpm, 1 h). The appropriate benzyl chloride or benzyl bromide (1 equiv mol^{-1}) and NaI (cat.) were then added, and the resulting mixtures were stirred overnight (300 rpm, RT). Water (5 mL) and AcOEt (5 mL) were added and the resulting mixtures were stirred (1 h). The organic layers were then filtered out in parallel and evaporated to dryness to afford crude compounds **5a**–**g**, which were purified by flash chromatography (petroleum ether/ethyl acetate (4:1)) to afford the final products (60–75% yield).

Microwave synthesis of 1-chlorophenyl-3-methyl-5-pyrazolone ***10b*** *and 1-fluorophenyl-3-methyl-5-pyrazolone* ***10c***: Ethylacetoacetate (3 mmol), the appropriate phenylhydrazine (3 mmol) and EtOH (5 mL) were mixed in oven-dried pressure vials equipped with magnetic stir bars. The vessels were placed in the microwave oven and heated (80 °C, 10 min) under microwave irradiation. The solutions were cooled and ethanol evaporated under vacuum. Diethyl ether was added and the crystals filtered. The products were purified by flash chromatography (petroleum ether/ethyl acetate (2:1)) to afford the products **10b**–**c**.

Parallel synthesis of pyrazoles ***11a****–****o****; general procedure:* Pyrazolones **10a** partitioned into 12 vessels (150 mg, 0.86 mmol), **10b** (160 mg, 0.76 mmol), and **10c** divided into two different vessels (113 mg, 0.58 mmol), were placed in the Büchi Syncore and dissolved in dioxane (10 mL). $Ca(OH)_2$ (2 equiv mol^{-1}) and the appropriate acyl chlorides (2 equiv mol^{-1} for **11i** and **11j**, 1 equiv mol^{-1} otherwise) were added. The reaction mixtures were refluxed (300 rpm,3 h). The cooled solutions were completely evaporated in under vacuum. 3 N HCl was added to precipitate crude compounds **11a**–**o** which were then filtered and recrystallized from EtOH (50–70% yield after crystallization).

Microwave-accelerated synthesis of pyrazoles ***11a****–****o****; general procedure:* Pyrazoles (1 mmol in 5 mL dioxane), $Ca(OH)_2$ and the appropriate acyl chlorides (1 equiv mol^{-1} or 2 equiv mol^{-1}), were mixed in oven-dried pressure vials equipped with magnetic stir bars. The vessels were placed in the microwave oven and heated (100 °C, 5 min) under microwave irradiation. After cooling, 3 N HCl was added, the solid phase removed by filtration, washed with water and recrystallized from EtOH to afford **11a**–**o**.

Parallel synthesis of pyrazoles ***12a****–****k****; general procedure:* Compounds **11a**, partitioned into seven vessels (200 mg, 0.64 mmol), and **11e**, divided into four different vessels (200 mg, 0.57 mmol), were placed in the Büchi Syncore and dissolved in ethanol (10 mL). Na_2SO_4 (1.5 equiv mol^{-1}) and the appropriate amine (1.5 equiv mol^{-1}) were then added. The reaction mixtures were refluxed (300 rpm, 4 h). *p*-toluenesulfonic acid polymer bound scavenger was added (1 equiv mol^{-1}) to the cooled solutions and the mixtures stirred (300 rpm, 2 h, RT). The reaction mixtures were filtered with a specific filtration unit and the scavengers were washed twice with CH_2Cl_2 (10 mL). The solvents were evaporated completely in the apparatus. Compounds **12a**–**k** (all in >70% yield) were identified by LC/MS analysis and proved to be >90% pure. The final products **12a**–**k** were further purified by flash chromatography (hexane/ethyl acetate (4:1)) yielding a purity >98%.

Synthesis of 4-(4-chlorobenzoyl)-1,2-dihydro-1-methyl-2-phenyl-3H-pyrazol-3-one ***13****:* A mixture of **11a** (200 mg, 0.64 mmol) and Me_2SO_4 (0.18 mL, 1.92 mmol) was heated in refluxing toluene (2 h). The mixture was then poured into water, stirred (15 min) made alkaline with 2 N NaOH and extracted exhaustively with CH_2Cl_2. The organic phases were washed twice with brine, dried (Na_2SO_4) and evaporated. The residue was recrystallized from toluene to afford pure **13**.

Biological assays: *Compounds:* Initial stock solutions of each compound and isoniazid (INH; Sigma Chemicals, St. Louis, MO, USA), employed as reference drug, were made in DMSO at 10 or 20 mg mL^{-1} and stored at −20 °C. Further dilutions were made in Middlebrook 7H9 broth (Difco Laboratories, Detroit, MI, USA). To avoid interference by the solvent, the highest DMSO concentration was 0.5%.[48]

Mycobacterial strains: M. tuberculosis H37Rv (ATCC 27294) and *M. avium* (ATCC 19421) were used in this study. These were maintained on Löwenstein–Jensen (bioMérieux, Marcy l'Étoile, France) agar slants until needed.

Antimicrobial susceptibility testing: MIC values were determined by a standard twofold agar dilution method.[49] Middlebrook 7H11 agar (1 mL, Difco Laboratories) supplemented with 10% oleic acid–albumin–dextrose–catalase (OADC) enrichment (Difco Laboratories) containing the testing compound, or INH, in 24-multiwell plates at concentrations between 0.09 and 100 μg mL^{-1}, was inoculated with 10 μL of a suspension containing *M. tuberculosis* H37Rv 1.5×10^5 cfu mL^{-1} obtained as described below. Final inoculum was 1.5×10^3 per well. Plates were incubated for 28 days, and MIC values

were read as minimal concentrations of compound completely inhibiting visible growth of mycobacteria.

Inoculum preparation: Suspension of *M. tuberculosis* to be used for antimicrobial susceptibility testing was prepared by inoculating the organism grown on Löwenstein-Jensen slants in tubes containing 7H9 broth supplemented with 10% albumin-dextrose-catalase (ADC) enrichment (Difco Laboratories) and Tween-80 (0.05% *v/v*). The suspension was incubated aerobically (14 days). The cells were then washed, suspended in 7H9 broth, shaken, and sonicated in an ultrasonicator until visible clumps were disrupted (typically 15–30 s). The suspension was then diluted in 7H9 broth to a turbidity of 1 McFarland and finally diluted in the same medium to 1.5×10^5 cfu mL^{-1}.

Acknowledgements

Financial support from the Italian Ministero dell'Istruzione, dell' Università e della Ricerca (PRIN 2005037820) is gratefully acknowledged. We thank Asinex for partial financial support for this work. We are also indebted to Professor Thierry Langer (University of Innsbruck) for hosting one of us (M.M.) in his laboratory. Professor G. Cruciani (University of Perugia) is also acknowledged for kindly providing us with the program Volsurf.

Keywords: drug design · ligand-based virtual screening · medicinal chemistry · molecular descriptors · molecular diversity

[1] D. Maher, M. Raviglione, *Clin. Chest Med.* **2005**, *26*, 167 – 182.
[2] World Health Organization: *Global Tuberculosis Control: Surveillance, Planning, Financing WHO Report*, **2004**, Geneva, Switzerland.
[3] World Health Organization: The World Health Organization Global Tuberculosis Program, **2003**. http://www.who.int/gtb/.
[4] M. A. Espinal, *Tuberculosis* **2003**, *83*, 44 – 51.
[5] a) T. R. Frieden, S. S. Munsiff, *Clin. Chest Med.* **2005**, *26*, 197 – 205; b) W. J. Burman, *Clin. Chest Med.* **2005**, *26*, 283 – 294.
[6] R. J. O'Brien, P. P. Nunn, *Am. J. Respir. Crit. Care Med.* **2001**, *1635*, 1055 – 1058.
[7] F. Manetti, F. Corelli, M. Biava, R. Fioravanti, G. C. Porretta, M. Botta, *Farmaco* **2000**, *55*, 484 – 491.
[8] M. Biava, R. Fioravanti, G. C. Porretta, D. Deidda, G. Lampis, R. Pompei, A. Tafi, F. Manetti, *Med. Chem. Res.* **2002**, *12*, 50 – 66.
[9] M. Biava, G. C. Porretta, D. Deidda, R. Pompei, F. Manetti, A. Tafi, *Bioorg. Med. Chem.* **2003**, *11*, 515 – 520.
[10] M. Biava, G. C. Porretta, D. Deidda, R. Pompei, A. Tafi, F. Manetti, *Bioorg. Med. Chem.* **2004**, *12*, 1453 – 1458.
[11] M. Biava, G. C. Porretta, G. Poce, D. Deidda, R. Pompei, A. Tafi, F. Manetti, *Bioorg. Med. Chem.* **2005**, *13*, 1221 – 1230.
[12] S. Meiering, O. Inhoff, J. Mies, A. Vincek, G. Garcia, B. Kramer, M. Dormeyer, R. L. Krauth-Siegel, *J. Med. Chem.* **2005**, *48*, 4793 – 4802.
[13] J. H. van Drie, M. S. Lajiness, *Drug Discovery Today* **1998**, *3*, 274 – 283.
[14] Recursive partitioning-centered virtual screening (RPCVS) protocol is a term used in the text and in Figures 1, 2, and 3 to outline a three-step virtual screening procedure consisting of: 1) the application of Lipinski's rule of five (due to the high lipophilicity of the mycobacterial cell wall, log *P* was raised from 5 to 6), 2) the classification of compounds according to the recursive partitioning (RP) model, and 3) the computation of the penalty factor.
[15] R. J. O'Brien, M. Spigelman, *Clin. Chest Med.* **2005**, *26*, 327 – 340.
[16] F. L. Stahura, J. Bajorath, *Curr. Pharm. Des.* **2005**, *11*, 1189 – 1202.
[17] Further information is available at: http://www.asinex.com/prod/gold.html.
[18] Separate lists containing: 1) references to papers in which compounds have been collected to build the learning set, 2) fragments used in the generation of LL, and 3) databases of commercially available compounds will be provided upon request.
[19] A similar approach was recently reported (Ref. [37]), based on the classification of compounds as active, moderately active, and inactive.
[20] Aimed at elaborating a ligand-based virtual screening procedure, the objective of the first part of our work was the collection of a huge set of compounds along with their MIC values suitable for the generation of a statistical model, without reference to their (known or unknown) mechanisms of action. Consistent with this collection criterion, we analyzed data by means of a recursive partitioning technique, which is able, in principle, not only to account for multiple structurally diverse classes of compounds and to tolerate a certain amount of erroneous input data, but also to deal with different mechanisms of action, as reported in: a) D. Horvath, *SAR QSAR Environ. Res.* **2001**, *12*, 181 – 212; b) S. S. Young, C. E. Keefer, Computation and analysis of large chemistry data sets. Chemical Data Analysis in the Large: The Challenge of the Automation Age. International Workshop, May 22nd–26th, **2000**, Bozen, Italy. http://www.beilstein-institut.de/bozen2000/proceedings/young/young.htm; c) A. Rusinko III; M. W. Farmen, C. G. Lambert, P. L. Brown, S. S. Young, *J. Chem. Inf. Model.* **1999**, *39*, 1017 – 1026.
[21] Calculation of the overall prediction parameter provides information on the accuracy of the model in predicting properties of the whole set of compounds.
[22] T. Langer, R. D. Hoffmann, *Curr. Pharm. Des.* **2001**, *7*, 509 – 527.
[23] R. D. Brown, M. Hassan, M. Waldman, *J. Mol. Graphics Modell.* **2000**, *18*, 427 – 437.
[24] E. A. Jamois, C. T. Lin, M. Waldman, *J. Mol. Graphics Modell.* **2003**, *22*, 141 – 149.
[25] J. M. Barnard, G. M. Downs, A. von Scholley-Pfab, R. D. Brown, *J. Mol. Graphics Modell.* **2000**, *18*, 452 – 463.
[26] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141 – 142.
[27] S. A. Wildman, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868 – 873.
[28] M. Hassan, J. P. Bielawski, J. C. Hempel, M. Waldman, *Mol. Diversity* **1996**, *2*, 64 – 74.
[29] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A* **1998**, *102*, 3762 – 3772.
[30] Penalty factor values in the SL range from 0.373 to 5.530. The threshold value of the penalty factors was arbitrarily defined, so as to result in a reasonably restricted number of selected compounds. 668 compounds of the SL had a penalty factor value less than 1 (about 13%).
[31] Details of the software can be found at the Molecular Discovery web page: http://www.moldiscovery.com/soft volsurf.php.
[32] M. A. Johnson, G. M. Maggiora, *Concepts and applications of molecular similarity*, Wiley, New York, **1990**.
[33] a) R. Ragno, G. R. Marshall, R. Di Santo, R. Costi, S. Massa, R. Pompei, M. Artico, *Bioorg. Med. Chem.* **2000**, *8*, 1423 – 1432, and references cited therein; b) G. Sbardella, A. Mai, M. Artico, R. Loddo, M. G. Setzu, P. La Colla, *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1537 – 1541; c) A. Nayyar, R. Jain, *Curr. Med. Chem.* **2005**, *12*, 1873 – 1886.
[34] As an example, BM212 was one of the first pyrrole derivatives, described in: D. Deidda, G. Lampis, R. Fioravanti, M. Biava, G. C. Porretta, S. Zanetti, R. Pompei, *Antimicrob. Agents Chemother.* **1998**, *42*, 3035 – 3037.
[35] A novel pyrrole compound, LL4858, is currently in development by Lupin Limited (US 2005 026128A1, WO 2004 026828).
[36] a) G. Menozzi, L. Merello, P. Fossa, S. Schenone, A. Ranise, L. Mosti, F. Bondavalli, R. Loddo, C. Murgioni, V. Mascia, P. La Colla, E. Tamburini, *Bioorg. Med. Chem.* **2004**, *12*, 5465 – 5483; b) M. G. Mamolo, D. Zampieri, V. Falagiani, L. Vio, E. Banfi, *Farmaco* **2001**, *56*, 593 – 599.
[37] I. Ghosh, O. Prakash, *J. Chem. Inf. Model.* **2006**, *46*, 17 – 23.
[38] G. C. Porretta, M. Biava, R. Fioravanti, A. Villa, N. Simonetti, *Farmaco* **1991**, *46*, 987 – 995.
[39] R. Fioravanti, M. Biava, G. C. Porretta, G. Lampis, C. Maullu, R. Pompei, *Med. Chem. Res.* **1999**, *9*, 249 – 266.
[40] R. Ballini, M. Petrini, E. Marcantoni, G. Rosini, *Synthesis* **1988**, *3*, 231 – 233.
[41] a) M. Radi, E. Petricci, G. Maga, F. Corelli, M. Botta, *J. Comb. Chem.* **2005**, *7*, 117 – 122; b) E. Petricci, C. Mugnaini, M. Radi, F. Corelli, M. Botta, *J.*

Org. Chem. **2004**, *69*, 7880 – 7887; c) D. Castagnolo, M. L. Renzulli, E. Galletti, F. Corelli, M. Botta, *Tetrahedron: Asymmetry* **2005**, *16*, 2893 – 2896.
[42] D. Villemin, B. Labiad, *Synth. Commun.* **1990**, *20*, 3213 – 3218.
[43] W. Holzer, R. M. Claramunt, M. Pérez-Torralba, D. Guggi, T. Brehmer, *J. Org. Chem.* **2003**, *68*, 7943 – 7950.
[44] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.
[45] Cerius^2 and Catalyst are distributed by Accelrys, Inc., San Diego, CA.
[46] G. Cruciani, P. Crivori, P.-A. Carrupt, B. Testa, *J. Mol. Struct.* **2000**, *503*, 17 – 30.
[47] Compounds **8a** and **8b** proved inseparable by flash chromatography; they were used in the next step as a mixture.
[48] C. Jagannath, V. M. Reddy, P. R. Gangadharam, *J. Antimicrob. Chemother.* **1995**, *35*, 381 – 390.
[49] A. De Logu, V. Onnis, B. Saddi, C. Congiu, M. L. Schivo, M. T. Cocco, *J. Antimicrob. Chemother.* **2002**, *49*, 275 – 282.

Received: February 2, 2006
Revised: May 22, 2006
Published online on August 7, 2006

