

Rational Design of New Antituberculosis Agents: Receptor-Independent Four-Dimensional Quantitative Structure–Activity Relationship Analysis of a Set of Isoniazid Derivatives

Kerly F. M. Pasqualoto,^{*,§} Elizabeth I. Ferreira,[§] Osvaldo A. Santos-Filho,[‡] and Anton J. Hopfinger[‡]

LAPEN, Departamento de Farmácia, Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, Av. Prof. Lineu Prestes 580, Bloco 13 sup., São Paulo, SP, 05508-900, Brasil, and Laboratory of Molecular Modeling and Design (M/C-781), College of Pharmacy, The University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231

Received January 31, 2004

A 4D-QSAR analysis was carried out for a set of 37 hydrazides whose minimum inhibitory concentrations against *M. tuberculosis* var. *bovis* were evaluated. These ligands are thought to act like isoniazid in mycolic acid biosynthesis. Results indicate that nonpolar groups in the acyl moiety of ligands markedly decrease biological activity. Molecular modifications of the ligand NAD moiety, including nonpolar groups and hydrogen bond donor and acceptor groups, seemingly improve ligand interactions with amino acid residues of the InhA active site.

Introduction

Tuberculosis (TB) is a chronic infectious disease caused by a mycobacteria of the “tuberculosis complex”, including *Mycobacterium bovis*, *Mycobacterium africanum*, and mainly *Mycobacterium tuberculosis*.^{1,2} According to World Health Organization, the global and regional incidence per year is nearly 2 million TB cases in sub-Saharan Africa, nearly 3 million TB cases in Southeast Asia, over a quarter of a million TB cases in Eastern Europe, and nearly 161 800 new TB cases annually in Brazil.^{3,4} From 2002 to 2020, nearly one billion more people will be newly infected, about 150 million people will get sick, and approximately 36 million will die from TB if control is not strengthened.⁴

The pandemic of AIDS has had a major impact on the worldwide TB problem. One-third of the increase in the incidence of TB in the past 5 years can be attributed to coinfection with HIV. Another factor contributing to the rise in TB and responsible for the increased death rate is the emergence of new strains of *M. tuberculosis* resistant to some or all current anti-TB drugs, so-called multidrug-resistant TB (MDR-TB).⁴

Considering drug resistance, the serious side effects of some current anti-TB drugs, and the lack of efficacy of current treatment in immunodepressed patients, it is still necessary to search for new antimycobacterial agents.² The identification of novel targets needs the identification of biochemical pathways specific to mycobacteria and related organisms. Many unique metabolic processes occur during the biosynthesis of mycobacterial cell wall components.⁵ One of these attractive targets for the rational design of new antituberculosis agents is the mycolic acids, the major components of the cell wall of *M. tuberculosis*.⁶

Mycolic acids are high molecular weight α -alkyl, β -hydroxy fatty acids covalently linked to arabino-

galactan.^{6–8} Differences in mycolic acid structure may affect the fluidity and permeability of an asymmetric lipid bilayer that would explain the different sensitivity levels of various mycobacterial species to lipophilic inhibitors.⁹

Enzymes that form the biosynthetic apparatus for fatty acid production, the fatty acid synthase (FAS), are considered ideal targets for designing new antibacterial agents. The difference between the molecular organization of FAS found in most bacteria and mammals^{8,10,11} is the reason for this assumption. Enoyl-acyl carrier protein reductase is a key regulatory step in fatty acid elongation and catalyzes the NADH-dependent stereospecific reduction of α,β -unsaturated fatty acids bound to the acyl carrier protein.^{12–14}

The crystal structure of the *M. tuberculosis* enoyl-acyl carrier protein reductase, named InhA, in complex with cofactor NADH and the inhibitor isoniazid (INH) was isolated by Rozwarski and co-workers (1998) (PDB entry code 1zid). They showed that the drug mechanism of action in *M. tuberculosis* involves a covalent attachment of the activated form of the drug (isonicotinic acyl anion or radical) to the carbon at position 4 of the nicotinamide ring of NADH bound within the active site of InhA, resulting in the formation of an acylpyridine/NAD adduct.¹⁵ The crystal structure of the complex between isonicotinic acyl/NAD and InhA provides a basis for the design of agents that inhibit InhA without needing a KatG drug activation.^{6,15}

With the purpose of contributing to the rational design of tuberculostatic leads, a mechanism-based study through a computer-assisted molecular design (CAMD) methodology was performed with analogues of INH, a drug that acts on mycolic acid biosynthesis. In this paper we report the receptor-independent four-dimensional quantitative structure–activity relationship (RI 4D-QSAR) analysis of a set of 37 hydrazides, which were evaluated with the same biological assay and probably would act like the lead INH. Although the geometry of the biomacromolecule target is available,

* To whom correspondence should be addressed. Phone: 55-11-3091 3793. Fax: 55-11-3815 4418. E-mail: kerly@netpoint.com.br.

[§] Universidade de São Paulo.

[‡] The University of Illinois at Chicago.

Table 1. Structures and Biological Activities of the 37 Hydrazides Series^a

R—CONHNH₂											
Compound	R	pMIC	Compound	R	pMIC	Compound	R	pMIC	Compound	R	pMIC
INH1		4.20	INHd2		3.82	INHd31		3.40	INHd43		3.40
INHd20		3.22	INHd14		3.22	INHd46		2.82	INHd37		2.70
INHd15		2.52	INHd23		2.52	INHd29		2.00	INHd16		1.92
INHd18		1.92	INHd44		1.82	INHd25		1.92	INHd30		1.92
Idv130		1.70	INHd27		1.52	INHd22		1.52	INHd34		1.40
INHd42		1.10	INHd47		1.00	Idv107		0.65	INHd45		0.70
Idv126		0.60	INHd19		0.52	Idv125		0.52	INHd41		0.40
INHd49		0.22	INHd48		0.22	Idv90		4.22	Idv128		2.82
Idv124		2.70	Idv131		2.00	Idv132		1.82	Idv136		0.52
INHd51		0.22									

^a Activity was measured as the minimum inhibitory concentration (MIC) against strains of *M. tuberculosis* var. *bovis* at 310 K and given as pMIC (see refs 19–22). The test set comprises the compounds Idv90, Idv128, Idv124, Idv131, Idv132, Idv136, and INHd51 (underlined letters). The others 30 compounds constitute the training set (bold letters). INHd = aromatic, heteroaromatic, and ring-substituted hydrazides, isoniazid derivatives; Idv = heterocyclic acid hydrazides and derivatives.

the RI formalism was applied in this study because of uncertainty in the binding mode of the ligands. 4D-QSAR analysis¹⁶ has been used to develop 3D pharmacophore models because of its capability of exploring large degrees of both conformational and alignment freedoms in the search for the active conformation and binding mode, respectively, of each compound investigated.

The hypothesized active conformations resulting from 4D-QSAR analysis can be used as structure design templates, which include their deployment as the molecular geometries of each ligand in a structure-based ligand–receptor binding research. This is the theme of

our future study where the training set (Table 1) and the three-dimensional structure of enoyl-*acp* reductase from *M. tuberculosis*, InhA¹⁵ (PDB entry code 1zid), will be used to generate a receptor-dependent three-dimensional quantitative structure–activity relationship (RD 3D-QSAR) model applying the free energy force field (FEFF) 3D-QSAR ligand–receptor binding formalism, as proposed by Tokarski and Hopfinger.^{17,18}

Methods

Biological Data. A series of 37 hydrazides, including the drug isoniazid (isonicotinic acid hydrazide, INH1),

Table 2. Ten Operational Steps in Performing an RI 4D-QSAR Analysis

step	description of the step operation
1	Generate the reference grid cell lattice and initial 3D models for all compounds in the training set.
2	Select the trial set of interaction pharmacophore elements, IPEs.
3	Perform a conformational ensemble sampling of each compound to generate its conformational ensemble profile, CEP.
4	Select a trial alignment.
5	Place each conformation of each compound in the reference grid cell lattice according to the alignment, and record the grid cell occupancy descriptor, GCOD, for each IPE and choose an occupancy measure for the CEP.
6	Perform a partial least-squares (PLS) data reduction of the entire set of GCODs against the biological activity measures.
7	Use the most highly weighted PLS GCODs and any other user-selected descriptors for the initial basis set in a genetic algorithm (GA) 4D-QSAR model optimization.
8	Return to step 4 and repeat steps 4–7 unless all trial alignments have been included in the analysis.
9	Select the optimum 4D-QSAR model with respect to alignment and any of the methodology parameters.
10	Select the low-energy conformer state, from the CEP set, for each compound that predicts the maximum activity using the optimum 4D-QSAR model as the “active” conformation.

were selected from refs 19–22. Biological activities were evaluated as the minimum inhibitory concentration, MIC ($\mu\text{g/mL}$), against strains of *M. tuberculosis* var. *bovis* at 310 K.^{19–22} The minimum inhibitory concentrations of these compounds were converted to molar units and then expressed in negative logarithmic units, pMIC ($-\log \text{MIC}$). The pMIC values are given in Table 1 and comprise the set of dependent variables in the 4D-QSAR analysis. The range in activity for the analogues in Table 1 is about 5 (0.22–4.70) pMIC units. The training set contains 30 hydrazides (Table 1) and comprises 10 active compounds (INH1, INHd2, INHd31, INHd43, INHd20, INHd14, INHd46, INHd37, INHd15, INHd23), 10 compounds with medium activity (INHd29, INHd16, INHd18, INHd44, INHd25, INHd30, Idv130, INHd22, INHd27, INHd34), and 10 inactive compounds (INHd42, INHd47, Idv107, INHd45, Idv126, INHd19, Idv125, INHd41, INHd49, INHd48). Additionally, seven compounds were selected as an external validation set. That is, these seven analogues represent a test set of compounds not included in developing the 3D-QSAR models. The test set comprises three active compounds (Idv90, Idv128, Idv124), two compounds with medium activity (Idv131, Idv132), and two inactive compounds (Idv136, INHd51). These additional seven compounds, and their respective pMIC values, are also listed in Table 1.

4D-QSAR Analysis. The current methodology formulation of 4D-QSAR analysis consists of 10 operational steps, which are given in Table 2.¹⁶ The implementation of this formalism for the analogues listed in Table 1 is described below.

It was presumed that all compounds investigated in this study would act like the lead drug isoniazid.¹⁵ After the hydrazide group is lost, the activated form (acylpyridine anion or radical) would be covalently attached to the C4 of the nicotinamide ring of the cofactor NAD, resulting in the formation of an acylpyridine/NAD adduct, which is a strongly bound inhibitor.

Step 1. The three-dimensional structures of each of the 37 analogues (Table 1) in their neutral forms were constructed using the HyperChem 6.0 software.²³ The crystallized structure of the isonicotinic acyl/NAD adduct in the active site of the enoyl-*acp* reductase from *M. tuberculosis*, InhA (PDB entry code 1zid, 2.7 Å resolution), was used as a geometry reference in the building up of all ligands. Each structure was energy-

minimized using the HyperChem 6.0 MM+ force field without any restriction. The Molsim 3.0 program²⁴ was also used for the optimization of each structure investigated. Partial atomic charges were computed using the AM1²⁵ semiempirical method, also implemented in the HyperChem program.

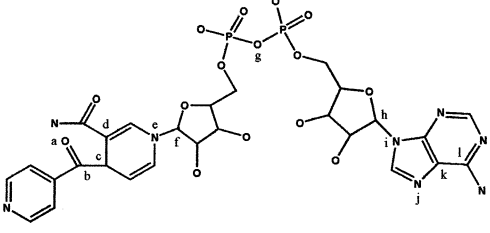
The structures modeled as described above were used as the initial structures in each molecular dynamics simulation, MDS,²⁶ used to construct the conformational ensemble profile, CEP, of each ligand.

Step 2. Seven types of atomic groups were used to define the interaction pharmacophore elements, IPEs, in this analysis. The atoms of each ligand of the training set are partitioned into seven classes: polar atoms of positive charge (p^+), polar atoms of negative charge (p^-), nonpolar atoms (np), hydrogen bond acceptor (hba), hydrogen bond donor (hbd), atoms in aromatic system (ar), and no differentiation of all-atom occupancy (a or any).

Step 3. The minimized structures of each of the compounds bound to the cofactor (NAD) (ligands) were used as the initial structures in each MDS employed to construct the conformational ensemble profile (CEP) of each ligand. The Molsim 3.0 program²⁴ was used to perform the MDS and to generate the trajectories for, in turn, deriving the CEP. The MDS protocol employed 100 000 steps for each ligand, the step size was 0.001 ps (1 fs), and the simulation temperature was 310 K, the same used in the biological assay.^{19–21,27} An output trajectory file was saved every 20 simulation steps to generate a CEP consisting of 5000 conformations. 4D-QSAR analysis does not use a single conformation in constructing a 4D-QSAR model, but rather the intrinsic conformational flexibility of each compound is taken into account through its CEP.

Step 4. The alignments are selected to explore each major “part” of a molecule, as well as the possible combinations of the major parts of a molecule. The current 4D-QSAR algorithm considers only the unrestricted three-ordered atom match alignment rule.

In this study, seven different three-ordered-atom alignments were selected to cover the entire bonding topology of the common chemical structure of the training set. The atom numbers, and the corresponding letter sequences, for each alignment are listed in Table 3 using compound INH1 bound to the cofactor NAD (ligand INH1) as example.

Table 3. Set of Trial Alignments Used in Constructing the 4D-QSAR Models^a


alignment	1st atom (atom number)	2nd atom (atom number)	3rd atom (atom number)
1	a (46)	b (45)	c (42)
2	b (45)	d (38)	e (36)
3	a (46)	c (42)	f (35)
4	b (45)	f (35)	g (23)
5	a (46)	g (23)	h (12)
6	i (13)	j (15)	k (16)
7	b (45)	g (23)	l (17)

^a The compound INH1 bound to NAD (ligand INH1) is used as an example.

Step 5. Each conformation from the CEP, consisting of the 5000 conformations generated by the MDS sampling for each ligand, was placed in a reference grid cell space according to the trial alignment under consideration. In this study, the selected size of the cubic grid cell was 1 Å and the size of the overall grid cell lattice was chosen to enclose each ligand of the training set. The atom occupancy of each grid cell is a descriptor in 4D-QSAR analysis, and as described in step 2, these grid cell occupancy descriptors (GCODs) were computed for each of seven IPE atom types.

No reference standard ligand¹⁶ was used to compute the GCODs. Thus, the normalized absolute occupancy of each grid cell by each IPE atom type over the CEP for a given alignment forms a unique set of QSAR descriptors (GCODs). The GCODs were computed and used as the basis set of trial 4D-QSAR descriptors in the 4D-QSAR analysis.

Step 6. A 4D-QSAR analysis generates an enormous number of trial QSAR descriptors, GCODs, because of the large number of grid cells and the seven IPEs. Partial least squares (PLS) regression analysis²⁸ is used to perform a data reduction fit between the observed dependent variable (the experimental biological activities, pMIC; Table 1) and the corresponding GCOD values for each of the trial (seven) alignments. In this study, a variance-filtering constraint was applied to the entire set of GCODs prior to the PLS analysis. GCODs having a variance (self-variance) over the set of analogues less than 2.0 (prechosen fraction) were eliminated. The automated reduction data by the PLS analysis provides the selection of the descriptors having the highest individual weightings to the observed biological activity measures.

Step 7. The 200 most highly weighted PLS GCODs (generated in step 6) were used to form the trial basis set for the genetic algorithm (GA) analysis.²⁹ In this study, the genetic function approximation (GFA)³⁰ was employed in 4D-QSAR model building and optimization. The GFA optimizations were initiated using 200 randomly generated 4D-QSAR models. Mutation probability over the crossover optimization cycle was set at

10%. The smoothing factor alters the balance between the number of independent variables, GCODs, in the models and the reduction in least-squares error and controls the overfitting. Smoothing factor values of 1–2.5 were tested in order to determine the optimal number of descriptors in the 4D-QSAR models^{31–33} based on Friedman's lack-of-fit, LOF,³⁴ which is a penalized least-squares measure. There is no way to know in advance how many GCODs should be part of a QSAR model; it depends on the ligand–receptor system. Generally, larger and/or more flexible ligands have a greater number of GCODs. However, the “rules” of statistics impose a limit of four to five observations (compounds) per descriptor.

The diagnostic measures used to analyze the resultant 4D-QSAR models generated by the GFA include descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or dependent variables (biological activity measures), number of significant models, and indices of model significance including the correlation coefficient, r^2 , leave-one-out cross-validation correlation coefficient, q^2 , and LOF.^{30,33} In this study, the ligands of the training set whose differences in experimental (Exp_BA) and predicted (Pred_BA) activities exceeded 2.0 standard deviations, SD, from the mean of a model were considered as outliers.

Other descriptors not derived from the 4D-QSAR analysis, like $\log P$ (the water/octanol partition coefficient), molar refractivity (MR) etc., can be added to the trial basis set at the start of this step.³⁵ Considering that lipophilicity is a major determinant of pharmacokinetic and pharmacodynamic properties of drug molecules,³⁶ the $\log P$ values were calculated (ClogP) for all ligands of the training set and were included in the GA analysis. The Ghose, Pritchett, and Crippen method (1998)³⁷ was used to obtain the ClogP values (HyperChem 6.0).

Step 8. Steps 4–7 were repeated until all (seven) trial alignments were included in the 4D-QSAR analysis.

Step 9. The inspection and evaluation of the entire population of 4D-QSAR models are made in this step. The purpose is to identify the “best” 4D-QSAR models with respect to alignment. In this study, the top 10 models for each of the seven alignments were selected by the 4D-QSAR program,³⁵ and their statistical measures were evaluated.

Each alignment considered will lead to a particular best 4D-QSAR model for that specific alignment. The alignment corresponding to the 4D-QSAR model with the overall highest r^2 and q^2 measures for all alignments tested is selected as the best alignment. For the best alignment, a cross-correlation matrix of the residuals in error (observed less predicted activities) between pairs of the top 10 4D-QSAR models, based on their q^2 , is built.^{16,35} This is done to determine if the top 10 4D-QSAR models are providing common, or distinct, structure–activity information. In other words, it is possible to identify the set of unique best 4D-QSAR models. Pairs of models with highly correlated residuals of fit ($R \approx 1$) are judged to be nearly the same model, while pairs of models with poorly correlated residuals ($R < 0.5$) are distinct from one another. Also, the linear cross-correlation matrix of the GCODs for the best 4D-QSAR

model for the best alignment is built to determine if these significant GCODs are correlated to one another.

External Validation. The seven compounds of the test set were not included in the building of the 4D-QSAR models, but they were used to validate the best QSAR model constructed from the training set and to evaluate its prediction capacity.³⁵ The predicted activity value (pMIC) of each ligand in the test set was calculated using the equation of the best model or alignment by substitution of the GCODs values found for the Cartesian coordinates indices of the reference grid cell space, $GCi(x,y,z)$, which represents each GCOD of each ligand in the test set and its respective position in the grid space.

Step 10. The final step of the 4D-QSAR formalism is to hypothesize the “active” conformation of each ligand in the training set. This is achieved by first identifying all conformer states sampled for each ligand, one at a time, which are within ΔE of the global minimum energy conformation of the CEP. The ΔE was set at 5 kcal/mol. The resulting low-energy conformations are individually evaluated using the correlation equation of the best 4D-QSAR model. The single conformation within ΔE that predicts the highest “activity” is selected as the active conformation of the ligand. The postulated active conformations can be used as structure design templates in other CAMD approaches.³⁵

The selected RI 4D-QSAR model should sterically fit and provide complementary ligand–receptor interaction sites. As already reported, the crystal structure of the complex containing the inhibitor (INH), the cofactor (NAD), and the enzyme (enoyl-*acp* reductase, *InhA*) is available (1zid),¹⁵ and it was used to explore the structural information within the selected RI 4D-QSAR model. It was done by docking the descriptors, GCODs, of the best model and the postulated active conformations of each ligand of the training set into the inhibitor binding site, considering the best alignment as a starting point. That is, the validation of an RI 4D-QSAR model was performed using the binding site geometry.^{16,35,38} In this study, comparisons were made of the distances of the relevant interactions between the amino acid residues of the active site and atoms of the inhibitor, INH, to the distances between the same amino acid residues and the GCODs plus the postulated active conformations to identify all possible complementary sets of interactions.

Results

Of the seven trial alignments reported in Table 3, alignment 4 provides the best 4D-QSAR models as defined by the highest cross-validated correlation coefficients. Moreover, alignment 4 provides 4D-QSAR models having the smallest values of the least-squares error, LSE, and LOF measures. The number of GCODs, statistical measures including the q^2 and r^2 values, and the number of outliers of the top 10 models are presented in Table 4 (Supporting Information) for each selected alignment when using a smoothing factor of 2.5 in the GFA optimization.

Table 5 (Supporting Information) shows the top 10 4D-QSAR models built from alignment 4. High values of both q^2 and r^2 for all models can be observed. However, models 2–5 and 8–10 have at least one outlier

and consequently were discarded from further analysis. Models 6 and 7 have no outliers, but their q^2 and r^2 values are lower and LSE values are higher than those of model 1.

To determine if the top 10 4D-QSAR models from alignment 4 are providing common, or distinct, structure–activity information, the correlation coefficients of the residuals of fit between pairs of models were computed and are reported in Table 6 (Supporting Information). The idea of determining the residual-pair correlations is that equivalent models will have near-identical residuals while distinct models should have noncorrelated residuals.^{16,30,35} Table 6 indicates that all of the top 10 4D-QSAR models have residuals of fit that are highly correlated to one another. Thus, there is a single unique 4D-QSAR model that is selected as the model with highest q^2 value (model 1).

The best 4D-QSAR model (model 1) is defined by

$$\text{pMIC} = -11.94 \text{ GC1(np)} + 11.37 \text{ GC2(any)} + \\ 23.07 \text{ GC3(any)} + 1.57 \text{ GC4(any)} + \\ 6.43 \text{ GC5(any)} + 30.76 \text{ GC6(np)} - 0.23 \quad (1)$$

$$N = 30; \quad r^2 = 0.88; \quad q^2 = 0.80; \quad \text{LSE} = 0.16$$

It is noteworthy that this model is composed of only two classes of IPEs: np and “any” atoms. Moreover, the GCOD (GC1) responsible for predicted decreases in biological activity corresponds to occupancy by a non-polar IPE-type (np).

A linear cross-correlation matrix of the GCODs for model 1 (eq 1) from alignment 4 was built and is reported in Table 7 (Supporting Information). This is done to verify if the independent variables, GCODs, contained in the selected 4D-QSAR model are correlated to one another.^{17,35,39} Table 7 shows that none of the GCODs of eq 1 are highly correlated to one another, since all pair correlations of GCODs are less than 0.5 (the absolute value). In other words, each of the GCODs provides independent information to the optimal 4D-QSAR model.

To ascertain the predictive power of model 1 based on screening a test set of compounds (Table 1), the pMIC value of each of the test set ligands (compound/NAD) was calculated using eq 1, as described in External Validation. In Table 8 (Supporting Information) the test set GCOD indices found for model 1 are presented. As already mentioned, $GCi(x,y,z)$ represents each descriptor, GCOD, of model 1 ($GCi = \text{GC1, GC2, GC3, GC4, GC5, and GC6}$), where x , y , and z are the Cartesian coordinates of the reference grid cell space. The test set predictions are given in Table 9 (Supporting Information). Four of the seven ligands of the test set have residuals whose absolute values ($\text{Exp_BA} - \text{Pred_BA} = \text{residuals}$) are less than or equal to 1.66, which is the SD value. This finding indicates that model 1 has a capacity of prediction of about 57%, which is a moderate predictive power.

As reported by Golbraikh and Tropsha (2002),⁴⁰ a high value of q^2 appears to be a necessary but not sufficient condition for the model to have a high predictive power. The external validation is consequently a better way to establish a reliable QSAR model. A reliable model should be also characterized by a high correlation coefficient R (R^2) between the predicted and observed

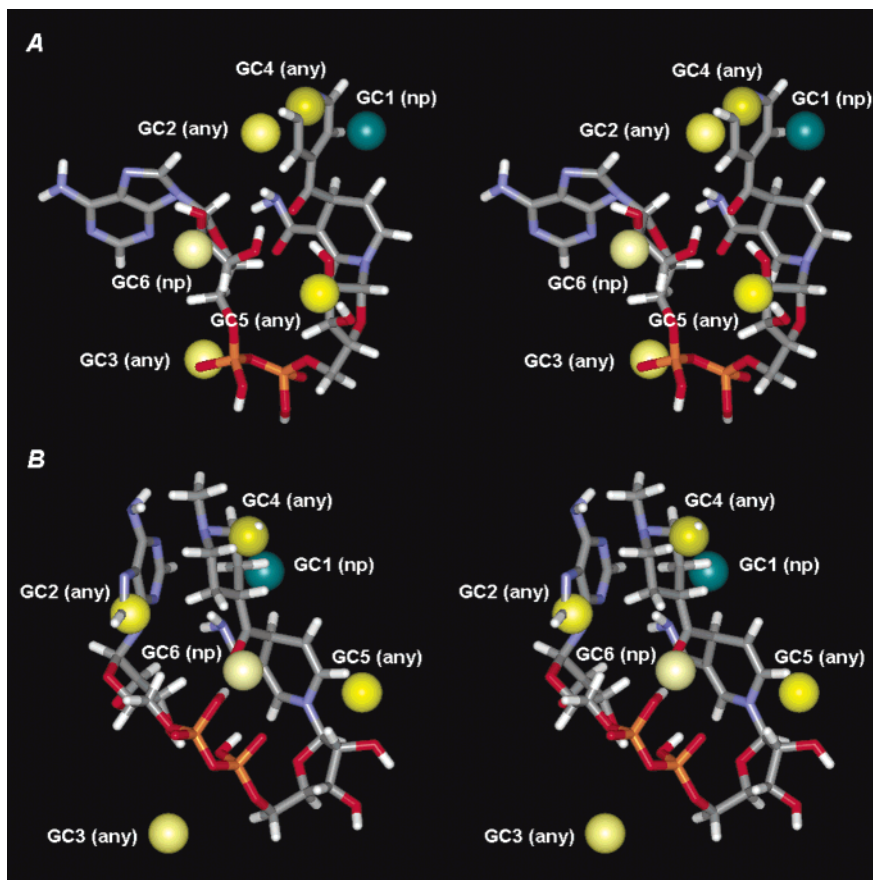


Figure 1. Stereoviews of the predicted active conformations for two ligands: INH1/NAD (active) (A) and INHd49/NAD (inactive) (B), using model 1/alignment 4. The nitrogen atoms are shown in blue, hydrogens are shown in white, oxygens are shown in red, phosphorus atoms are shown in orange, and carbon atoms are shown in gray. The GC1 and GC6 descriptors correspond to IPE type atom = np. The GC2, GC3, GC4, and GC5 descriptors have IPE type atom = any.

activities of compounds from a test set. In this study, the correlation coefficient between the predicted and observed activities of ligands from test set was 0.15.

To explore the possibility that the test set is not representative of the training set, ligands from these two sets were put together to form a single new training set. A 4D-QSAR analysis³⁵ was carried out for alignment 4 (the best alignment) for this new “training_test” set. Table 10 (Supporting Information) contains the statistical measures, the number of GCODs, and the number of outliers of the top 10 models from alignment 4 to the “training_test” set. The q^2 and r^2 values decreased relative to eq 1, and the best model in Table 10, model 1, has a high q^2 value but one outlier. The outlier is the ligand Idv90 that is part of the original test set.

The correlation coefficients of the residuals in error between pairs of models were computed and are reported in Table 11 (Supporting Information). Table 11 indicates that all of the top 10 4D-QSAR models resulting from the “training_test” set have residuals of fit that are highly correlated to one another. Thus, there is a single unique 4D-QSAR model, which is model 1 (highest q^2 value). After the outlier of model 1 (Idv90) was eliminated, 4D-QSAR models were rebuilt using the resulting training set. The q^2 and r^2 values of the top 10 models increased, and the number of outliers varied from zero to three. The model that has the highest q^2 value is model 1, as reported in Table 12 (Supporting Information). On the basis of the linear cross-correlation matrix of the residuals of fit, all models are highly

correlated to one another (see Table 13 in Supporting Information), which means all the models provide common structure–activity information. Considering all the data generated from the original and new training sets, the hypothesis that the test set was not representative of the original training set was discarded.

The test set predictions only marginally support the binding hypothesis used. Our future structure-based design study may help us understand what modifications need to be made to the hypothesis we have used.

The “active” conformation of each ligand in the training set was hypothesized using model 1 (eq 1) from alignment 4 by first identifying all conformer states sampled for each ligand within ΔE equal to 5 kcal/mol of the global minimum energy conformation of the CEP. The GCODs of each resulting set of low-energy conformations were employed to predict the activities for each ligand using eq 1, and the conformer with the highest predicted activity was selected as the “active” conformation of each ligand.

The predicted active conformations for two ligands are shown, in a ball-and-stick style, in Figure 1. The two ligands have been singled out from the training set based on their activities, ligand INH1/NAD being the most active and ligand INHd49/NAD being the least active analogue. These conformer states were developed using a grid cell resolution of 1 Å, which is the diameter of the spheres used to represent the GCODs of the 4D-QSAR equation. In Figure 1, the GCODs that increase the biological activity are shown as light-yellow spheres

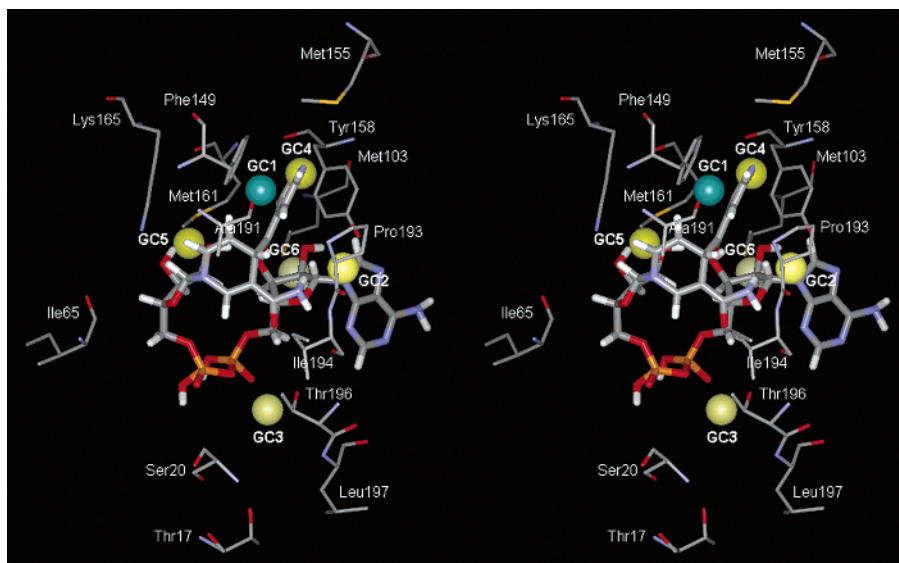


Figure 2. Stereoview of the predicted active conformation of ligand INH1/NAD (active) and its respective descriptors (GCODs, GC1–GC6) docked in the active site of InhA.

(GC2, GC3, GC4, GC5, and GC6), and the GCOD that decreases the biological activity is shown as a dark-green sphere (GC1). The yellow color intensity of the GCODs (spheres) is related to the regression coefficients in eq 1. The larger the absolute value of the regression coefficient, the more intense is the yellow color of its corresponding sphere, GCOD (see Figure 1).

To determine if model 1 from alignment 4 sterically fits in the active site and provides complementary ligand–receptor interaction sites to all compounds of the training set, the predicted active conformation of each ligand and its respective GCODs were docked into the active site of the crystallized structure of enoyl-acyl reductase from *M. tuberculosis*, InhA (1zid),¹⁵ using the binding alignment 4. In Figures 2, 3, and 4 this docking procedure can be visualized for three ligands of the training set: INH1/NAD (high activity), INHd18/NAD (medium activity), and INHd49/NAD (inactive), respectively. The distances between the amino acid residues of the active site (Met155, Phe149, Ala191, Ile194, Pro193, Thr196, Leu197, Thr17, Ser20, Ile95, Lys165, Met161, Tyr158, Met103, Met199) and the GCODs plus the predicted active conformations for the same three ligands of the training set were measured and compared to the corresponding distances in the crystallized complex (1zid).¹⁵ All possible complementary sets of ligand–receptor interactions, based on the selected model/alignment (model 1/alignment 4) of RI 4 D-QSAR analysis, can be identified from this process.

Discussion

Biological activity (pMIC) is predicted to increase with ligand atom occupancy of grid cells of GC2, GC3, GC4, GC5, and GC6 by the appropriate IPE types. The opposite is true for occupancy of GC1 (see eq 1). On the basis of the docking of ligand INH1/NAD (active) into the active site (Figure 2), the GC4 descriptor is located on the nitrogen atom of the pyridine ring (0.6 Å). The amino acid residues that could interact with GC4 are Met155 and Phe149. The IPE atom type of GC4 is any, which means all occupancy by any type of atom. This

nonspecificity in atom type for GC4 might be explained through the location of GC4 in the active site. GC4 can establish a hydrogen bond with a water molecule held by the side chain of Met155 (1zid) or have hydrophobic interactions with other amino acid residues of the active site. The inhibitor of the crystallized complex (1zid) has the pyridine ring of the isonicotinic acyl group surrounded by hydrophobic residues, including Phe149, Gly192, Pro193, Leu218, Tyr158, and Trp222. Furthermore, the side chain of Phe149 is located adjacent to the pyridine ring of the isonicotinic acyl group, allowing its participation in an aromatic ring stacking interaction.¹⁵ The distance between the two aromatic rings is the same for both the inhibitor of the crystallized complex (1zid) and the predicted active conformation of the active ligand (INH1/NAD), as determined from the 4D-QSAR analysis and docked into the active site based on alignment 4, also from the 4D-QSAR analysis.

The only descriptor responsible for a predicted decrease in pMIC is GC1, and it involves occupancy by nonpolar atom types (IPE). The distance between the GC1 descriptor and the pyridine ring is 1.6 Å, while GC1 is 2.3 Å from the side chain of Phe149. Phe is a nonpolar residue; thus, occupancy of GC1 by nonpolar ligand atoms is detrimental to biological activity. Ligand substituents that occupy GC1 could prevent rotation [movement] of the side chain of Phe149. Thus, the aromatic ring-stacking interaction between the Phe149 residue and the pyridine ring of the inhibitor might not be allowed. Moreover, rotation of the side chain of Phe149 away from the nicotinamide ring creates space for the isonicotinic acyl group. GC1 is also surrounded by the hydrophobic residues Phe149, Gly192, Pro193, Leu218, Tyr158, and Trp222 of the active site.

The descriptors GC2, GC3, and GC5 are predicted to increase pMIC in the following order: GC3 > GC2 > GC5, as specified by their regression coefficients in eq 1. GC2 is located near the amide nitrogen of the nicotinamide ring of the NAD portion (2.8 Å). The amino acid residues that could be participating in a ligand–receptor interaction identified by GC2 are Pro193 and Ile194. Thus, GC2 can reflect both hydrogen bond

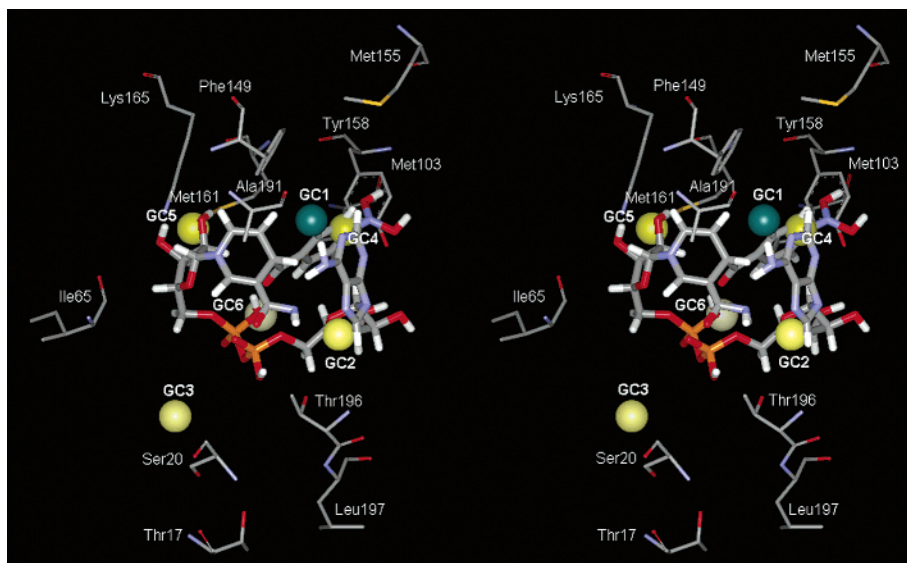


Figure 3. Stereoview of the predicted active conformation of ligand INHd18/NAD (medium activity) and its respective descriptors (GCODs, GC1–GC6) docked in the active site of InhA.

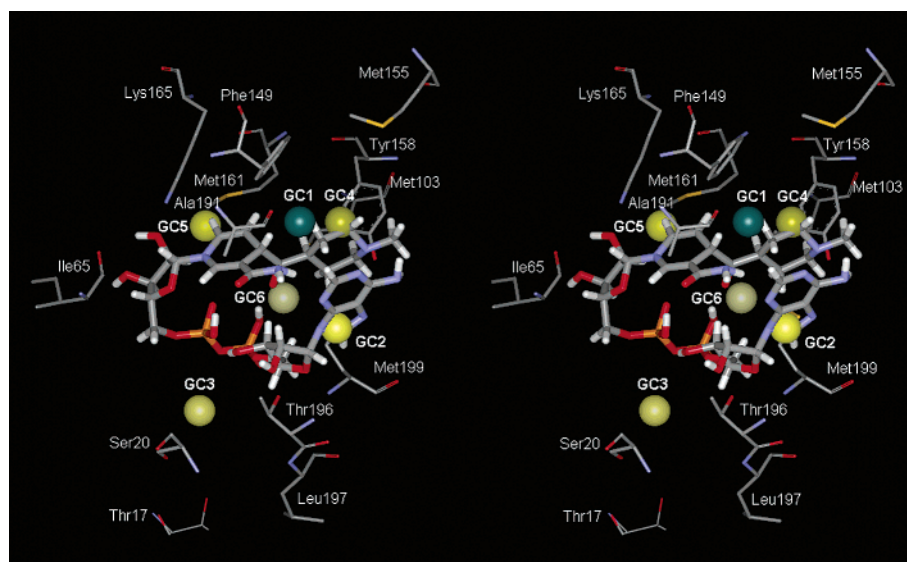


Figure 4. Stereoview of the predicted active conformation of ligand INHd49/NAD (inactive) and its respective descriptors (GCODs, GC1–GC6) docked in the active site of InhA.

acceptor and donor interactions. GC2 is located in the region of the active site surrounded by hydrophobic residues, suggesting it could also be identifying hydrophobic interactions. This variety of possible interaction types involving GC2 could be the reason for its IPE type of any. The GC3 descriptor is positioned about half-way between the two phosphate groups of the NAD structure. The distance between GC3 and the oxygen atom between the two phosphate groups is 2.6 Å. The amino acid residues that probably interact with ligand atoms captured by this descriptor are Thr196 (1.9 Å), Leu197 (3.7 Å), Thr17 (5.1 Å), and Ser20 (4.9 Å). In the crystallized complex (1zid),¹⁵ there is a water molecule involved in the interactions between the residues Thr17, Leu197, and the second phosphate group of the NAD portion of inhibitor. The IPE type of GC3 is any. This IPE might be explained by the participation of ligand atoms of GC3 as hydrogen bond acceptors and/or donors with amino acid residues (Thr196, Leu197, Thr17, and Ser20) of the active site. The GC5 descriptor is located

close to the hydrogen atom of the 2'-hydroxyl oxygen of the nicotinamide ribose ring of the NAD portion (1.1 Å). In the crystallized complex (1zid), this hydroxyl group interacts with the amino acid residues Lys165 and Ile95. There is also a water molecule interacting with Lys165 and Ile95 residues.¹⁵ Thus, the GC5 descriptor presents any as its IPE atom type because ligand atoms occupying it participate in interactions with the Lys165 and Ile95 residues as hydrogen bond acceptors and/or donors. In addition, GC5 could reflect hydrophobic interactions of the ligand with the Met161 residue of the active site.

The GC6 descriptor can be responsible for the largest increase in pMIC (regression coefficient = 30.76). It is located near the hydroxyl group of the second ribose ring of the NAD portion (1.5 Å). The IPE type of GC6 is nonpolar. In the crystallized complex (1zid), the second ribose ring is near the amino acid residues Gly14 and Ala22, and the isonicotinic acyl/NAD inhibitor is reported as an extended structure. The predicted active

conformation of the ligand INH1/NAD (Figure 2) resulting from model 1/alignment 4 is a bent ("U" shape) conformation, and it could be participating in interactions involving other amino acid residues of the active site including Tyr158 and Met103. Furthermore, it is important to mention that the predicted active conformations resulting from this RI 4D-QSAR analysis were presented, in general, as more bent structures compared with that if the cocrystallized inhibitor (1zid). The CEP of each ligand generated by the MDS procedure was carried out considering just the ligand, neglecting any interactions with amino acid residues of the active site. Moreover, the preferred conformations of a flexible isolated molecule, compared to the molecule interacting with other molecules, tend to be conformations in which the molecule collapses (folds) onto itself. This could explain why the bent conformations are obtained for the 4D-QSAR active conformation of each ligand of training set.

The predicted active conformation of ligand INHd18/NAD (medium activity) docked into the active site is shown in Figure 3. By use of the ligand INH1/NAD (the most active) as the reference, the GC1(np) and GC2(any) descriptors are not located in the region of the active site and surrounded by the hydrophobic residues (Phe149, Gly192, Pro193, Leu218, Tyr158, and Trp222), as previously seen. In addition, GC3(any) changes its position. It is located more distant from the amino acid residues Thr196 (6.5 Å) and Leu197 (7.7 Å). The change in positions of the GC1(np), GC2(any), and GC3(any) descriptors could be related to the decrease in biological activity of the ligand INHd18/NAD when compared to the most active ligand of the training set (INH1/NAD).

The predicted active conformation of ligand INHd49/NAD (inactive) (Figure 4) also undergoes changes in the locations of the GC1(np), GC2(any), and GC3(any) descriptors in the active site compared to the ligand INH1/NAD. Moreover, additional interactions involving ligand atoms of the GC2(any) and GC6(np) descriptors with Met199 of the active site are observed. Considering the regression coefficients of GC2 (11.37) and GC6 (30.76) in eq 1, it is clear that their interactions with Met199 would be detrimental to pMIC potency.

Model 1 of alignment 4 fits the active site well for 7 of the 10 active compounds of training set. The three ligands that did not fit into the active site using the 4D-QSAR model are INHd46/NAD, INHd20/NAD, and INHd23/NAD. The interactions expected to involve the GCODs of the predicted active conformations and the amino acid residues of the active site (1zid) do not occur like those observed for the ligand INH1/NAD. Thus, the 4D-QSAR model selected in this analysis does not provide pharmacophore sites, GCODs, that elucidate the biological activities of compounds INHd46, INHd20, and INHd23.

It is worthy of note that this is a receptor-independent 4D-QSAR analysis of a training set whose biological activities were considered to be the inhibition of the *M. tuberculosis* enoyl-acyl reductase (1zid). We are now in the process of expanding this investigation by doing a receptor-dependent 3D-QSAR analysis, using the free energy force field 3D-QSAR method,^{17,18} using the same target enzyme, InhA, and inhibitor training set and the

respective best ligand alignment as found and employed in this study.

Acknowledgment. K.F.M.P. is grateful to the CAPES Foundation, a federal scientific agency of Brazil, for scholarship support, to the Department of Medicinal Chemistry and Pharmacognosy of University of Illinois at Chicago for financial support, and to the Laboratory of Molecular Modeling and Design for providing the resources that were used in this study.

Supporting Information Available: Tables 4–13 listing some RI 4D-QSAR results. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Wolinsky, E. Tuberculosis. In *Cecil Textbook of Medicine*, 19th ed.; Wyngaarden, J. B., Smith, L. H., Jr., Bennett, J. C., Eds.; W. B. Saunders Company: Philadelphia, PA, 1992; Vol. 2, pp 1733–1742.
- (2) Sensi, P.; Grass, I. G. G. Antimycobacterial Agents. In *Burger's Medicinal Chemistry and Drug Discovery*, 5th ed.; Burger, A., Wolff, M. E., Eds.; John Wiley & Sons: New York, 1996; Vol. 2, pp 575–635.
- (3) *Country Profiles: Brazil*; Annual Report; World Health Organization, 1997 [www.who.int/gtb/publications/tbrep_97/countries/brazil.htm].
- (4) *Tuberculosis*; Fact Sheet No. 104 (revised in August 2002); World Health Organization, 2002 [www.who.int/mediacentre/factsheets/who104/en/index.html].
- (5) Barry, C. E., III. New horizons in the treatment of tuberculosis. *Biochem. Pharmacol.* **1997**, *54*, 1165–1172.
- (6) Pasqualoto, K. F. M.; Ferreira, E. I. An approach for the rational design of new antituberculosis agents. *Curr. Drug Targets* **2001**, *2*, 427–437.
- (7) Brennan, P. J.; Nikaido, H. The envelope of mycobacteria. *Annu. Rev. Biochem.* **1995**, *64*, 29–63.
- (8) Barry, C. E., III.; Lee, R. E.; Mdluli, K.; Sampson, A. E.; Schroeder, B. G.; Slayden, R. A.; Yuan, Y. Mycolic acids: structure, biosynthesis and physiological functions. *Prog. Lipid Res.* **1998**, *37*, 143–179.
- (9) Liu, J.; Barry, C. E., III.; Besra, G. S.; Nikaido, H. Mycolic acid structure determines the fluidity of the mycobacterial cell wall. *J. Biol. Chem.* **1996**, *271*, 29545–29551.
- (10) McCarthy, A. D.; Hardie, D. G. Fatty acid synthase—an example of protein evolution by gene fusion. *Trends Biochem. Sci.* **1984**, *9*, 60–63.
- (11) Magnuson, K.; Jackowski, S.; Rock, C. O.; Cronan, J. E., Jr. Regulation of fatty acid biosynthesis in *Escherichia coli*. *Microbiol. Rev.* **1993**, *57*, 522–542.
- (12) Bergler, H.; Fuchsichler, S.; Högenauer, G.; Turnowsky, F. The enoyl-[acyl-carrier-protein] reductase (FabI) of *Escherichia coli*, which catalyzes a key regulatory step in fatty acid biosynthesis, accepts NADH and NADPH as cofactors and is inhibited by palmitoyl-CoA. *Eur. J. Biochem.* **1996**, *242*, 689–694.
- (13) Stewart, M. J.; Parikh, S.; Xiao, G.; Tonge, P. J.; Kisker, C. Structural basis and mechanism of enoyl reductase inhibition by triclosan. *J. Mol. Biol.* **1999**, *290*, 859–865.
- (14) Rozwarski, D. A.; Vilchèze, C.; Sugantino, M.; Bittman, R.; Sacchettini, J. C. Crystal structure of the *Mycobacterium tuberculosis* enoyl-acyl reductase, InhA, in complex with NAD⁺ and a C16 fatty acyl substrate. *J. Biol. Chem.* **1999**, *274*, 15582–15589.
- (15) Rozwarski, D. A.; Grant, G. A.; Barton, D. H. R.; Jacobs, W. R., Jr.; Sacchettini, J. C. Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science* **1998**, *279*, 98–102.
- (16) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (17) Tokarski, J. S.; Hopfinger, A. J. Constructing protein models for ligand–receptor binding thermodynamic simulations: An application to a set of peptidomimetic rennin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 779–791.
- (18) Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand–receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: Application to a set of peptidomimetic rennin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792–811.
- (19) Bernstein, J.; Lott, W. A.; Steinberg, B. A.; Yale, H. L. Chemotherapy of experimental tuberculosis. V. Isonicotinic acid hydrazide (nydrazid) and related compounds. *Am. Rev. Tuberc.* **1952**, *65*, 357–364.

- (20) Bernstein, J.; Jambor, W. P.; Lott, W. A.; Pansy, F.; Steinberg, B. A.; Yale, H. L. Chemotherapy of experimental tuberculosis. VI. Derivatives of isoniazid. *Am. Rev. Tuberc.* **1953**, *67*, 354–365.
- (21) Bernstein, J.; Jambor, W. P.; Lott, W. A.; Pansy, F.; Steinberg, B. A.; Yale, H. L. Chemotherapy of experimental tuberculosis. VII. Heterocyclic acid hydrazides and derivatives. *Am. Rev. Tuberc.* **1953**, *67*, 366–375.
- (22) Klopman, G.; Fercu, D.; Jacob, J. Computer-aided study of the relationship between structure and antituberculosis activity of a series of isoniazid derivatives. *Chem. Phys.* **1996**, *204*, 181–193.
- (23) *HyperChem Program Release 6.0 for Windows*; Hypercube, Inc.: Gainesville, FL, 1996.
- (24) Doherty, D. *MOLSIM: Molecular Mechanics and Dynamics Simulation Software. User's Guide*, version 3.0; The Chem21 Group Inc. (1780 Wilson Drive, Lake Forest, IL 60045), 1997.
- (25) Dewar, M. J. S. E.; Zoebisch, G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (26) van Gunsteren, W. F.; Berendsen, H. J. C. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.
- (27) Rake, G.; Jambor, W.; McKee, C. M.; Pansy, F.; Wiselogle, F. Y.; Donovick, R. The use of mouse in a standardized test for antituberculous activity of compounds of natural or synthetic origin: III. The standardized test. *Am. Rev. Tuberc.* **1949**, *60*, 121.
- (28) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal components analysis and partial least-squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349–354.
- (29) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
- (30) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (31) Rogers, D. *WOLF Genetic Function Approximation. Reference Manual*, version 5.5; Molecular Simulation Inc.: Burlington, MA, 1994.
- (32) Dunn, W. J., III; Rogers, D. Genetic Partial Least Squares in QSAR. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic: London, 1996; pp 109–130.
- (33) Rogers, D. G/SPLINES: A hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm. In *The Proceedings of the Fourth International Conference on Genetic Algorithms*; Belew, R. K., Booker, L. B., Eds.; Morgan Kaufmann Publishers: San Francisco, 1991; pp 38–46.
- (34) Friedman, J. *Multivariate adaptive regression splines*; Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University: Stanford, CA, 1988.
- (35) Hopfinger, A. J. *4D-QSAR Software. User's Manual*, version 3.0; The Chem21 Group Inc. (1780 Wilson Drive, Lake Forest, IL 60045), 1999.
- (36) Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule approaches for calculating log P. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337–354.
- (37) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure–activity relationships III: Modeling hydrophobic interactions. *J. Comput. Chem.* **1988**, *9*, 80–90.
- (38) Venkatarangan, P.; Hopfinger, A. Prediction of ligand–receptor free energy by 4D-QSAR analysis: Application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141–1150.
- (39) Kubinyi, H. QSAR: Hansch analysis and related approaches. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Krosgaard-Larsen, P., Timmerman, H., Eds.; VHC: New York, 1993; pp 91–107.
- (40) Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

JM049913K