

Molecular Modeling and Receptor-Dependent (RD) 3D-QSAR Approach to a Set of Antituberculosis Derivatives

Kerly Fernanda Mesquita Pasqualoto,* Márcia Miguel Castro Ferreira

Laboratory for Theoretical and Applied Chemometrics, Department of Physical Chemistry, Institute of Chemistry, The State University of Campinas – UNICAMP, Campinas, SP 13084-971, P. O. Box 6154, Brazil

*E-mail: kerly@netpoint.com.br; kerly@usp.br; Phone: +55 19 3521 3102; Fax: +55 19 3521 3023

Keywords: Molecular modeling; Enoyl-ACP reductase; Molecular dynamics simulation; Structure-based design; Tuberculosis, Drug design

Received: April 2, 2009; Accepted: November 10, 2009

DOI: 10.1002/qsar.200960035

Abstract

In this study, *receptor-dependent* (RD) 3D-QSAR models were built for a set of thirty-seven isoniazid derivatives bound to the enoyl-ACP reductase from *M. tuberculosis*, called InhA (PDB entry code 1zid). Ligand-receptor (L-R) molecular dynamics (MD) simulations [500 000 steps; the step size was 0.001 ps (1 fs)] were carried out at 310 K (biological assay temperature). The hypothesized active conformations resulting from a previously reported *receptor-independent* (IR) 4D-QSAR analysis were used as the molecular geometries of each ligand in this structure-based L-R binding research. The dependent variable is the reported MIC values against *M. tuberculosis* var. *bovis*. The independent variables (descriptors) are energy terms of a modified first-generation AMBER force field combined with a hydration shell aqueous solvation model. Genetic function approximation (GFA) formalism and partial least squares (PLS) regression were employed as the fitting functions to develop 3D-QSAR models. The bound ligand solvation energy, the sum of electrostatic and hydrogen bonding energies of the unbound ligand, the bending energy of the unbound ligand, the electrostatic intermolecular L-R energy, and the change in hydrogen bonding energy upon binding were found as important energy contributions to the binding process. The 3D-QSAR model at 310 K has good internal and external predictability and may be regarded as representative of the binding process of ligands to InhA.

1 Introduction


Enoyl-ACP reductase (ENR) 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 [1–3]. Enzymes that form the biosynthetic apparatus for fatty acid production, the fatty acid synthase (FAS), are considered ideal targets for designing new antibacterial and antimycobacterial agents due to the difference between the molecular organization of FAS found in most bacteria/mycobacteria and mammals [4–6].

Biochemical evidence has suggested that isoniazid (INH), a first-line drug to treat tuberculosis disease, blocks the mycolic acids biosynthesis in *M. tuberculosis*. Those acids constitute the major components of mycobacterial cell wall [4, 7, 8]. The mycolic acids as well as the key enzyme responsible for their elongation are considered at-

tractive targets for the rational design of new antituberculosis agents.

The crystal structure of the *M. tuberculosis* enoyl-ACP reductase, named InhA, in complex with cofactor nicotinamide adenine dinucleotide (NAD) and the inhibitor INH was isolated by Rozwarski and co-workers (1998) (PDB entry code 1zid) [9]. 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 NAD bound within the active site of InhA, resulting in the formation of an acylpyridine/NAD adduct [9].

Previous *receptor-independent* (RI) 4D-QSAR analysis of a set of hydrazides (INH analogues) was carried out to

 Supporting information for this article is available on the WWW under www.qcs.wiley-vch.de

determine the optimum model and alignment for those compounds. It was assumed that all hydrazides would act like INH, forming an adduct with cofactor NAD in the active site of InhA [10].

The hypothesized active conformations resulting from a *RI* 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 [11].

In this work, a set of thirty-seven hydrazides (including INH) are explored in terms of ligand-receptor MD simulations to generate the thermodynamic descriptors regarding both ligand-receptor states, bound (L-R) and unbound (L and R, respectively). These independent variables were used to construct *receptor-dependent* (RD) 3D-QSAR models employing a genetic function approximation (GFA) [12] formalism and partial least squares (PLS) [13] regression as the fitting functions. Both GFA and PLS are valuable analytical tools for datasets that have more descriptors than samples, where GFA selects appropriate basis functions to be used in a model of the data and PLS regression is the fitting technique to weigh the basis functions' relative contributions in the final model.

2 Methodology

2.1 Biological Data and Starting Geometries for Ligands and Receptor (InhA)

The same set used in [10] was investigated here: 37 hydrazides, including INH, which were evaluated with the same biological assay. Biological activities were reported as the minimum inhibitory concentration, *MIC* ($\mu\text{g}/\text{mL}$), against strains of *M. tuberculosis* var. *bovis* at 310 K [14–17]. The minimum inhibitory concentrations were converted to molar units and then expressed in negative logarithmic units, *pMIC* ($-\log \text{MIC}$). The range in activity for the analogues is about five *pMIC* units (0.22–4.70). Additionally, six compounds were selected as an external validation set, using the Hierarchical Cluster Analysis (HCA) (see Table 1).

It was assumed that all compounds would act like the lead drug INH, forming an adduct with cofactor NAD in the active site of InhA, as reported by Rozwarski and co-workers [9], and the hypothesized active conformations from a previous *RI* 4D-QSAR analysis [10] were used as the ligands starting geometries. Each structure was energy-minimized using the HyperChem 7.51 [18] MM+ force field without any restriction. The Molsim 3.2 program [18] was also used for the optimization of each structure investigated. Partial atomic charges were computed using the AM1 [20] semiempirical method, also implemented in the HyperChem program [18]. The charges were calculated using the electrostatic potential [18].

The X-ray structure of the complex InhA-NAD-INH (PDB entry code 1zid, 2.7 Å of resolution) was selected as

starting model for the receptor geometry. The 1zid structure has one polypeptide chain or subunit containing 268 amino acid residues. The N-terminus and C-terminus were both modeled as neutral and the CH_3 groups were used as the block groups. AMBER [21] partial charges were assigned to all atoms of the enzyme structure, except to the block groups, using the HyperChem 7.51 program [18]. The charge state of ionizable residues was modeled at neutral pH. Lone pair electrons were not modeled explicitly. Only four water solvent molecules, which participate in ligand-receptor (L-R) interaction [9], were maintained in the InhA active site model. The MOLSIM 3.2 program [19] was used to perform the energy minimization of the modeled InhA-NAD-INH complex. The energy-minimized structure of the complex was used as initial structure in the MD calculations (item 2).

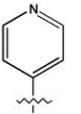
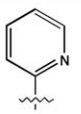
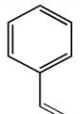
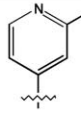
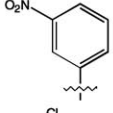
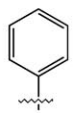
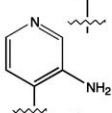
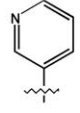
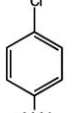
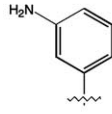
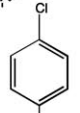
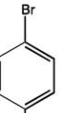
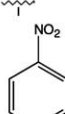
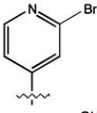
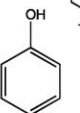
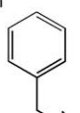
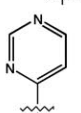
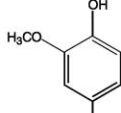
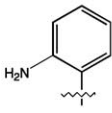
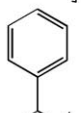
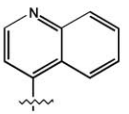
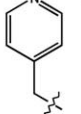
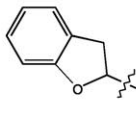
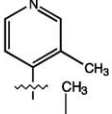
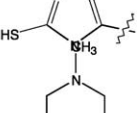
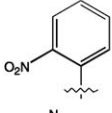
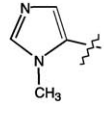
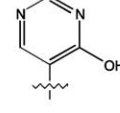
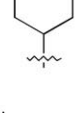
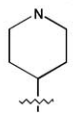
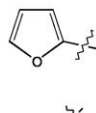
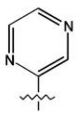
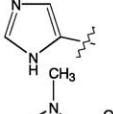
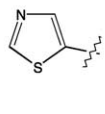
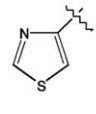
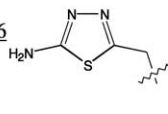
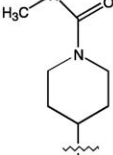
2.2 Molecular Dynamics Procedure and QSAR Models

Energy minimization and MD calculations were performed using the MOLSIM program, version 3.2 [19]. The hydration shell model proposed by Hopfinger [22] was included in the force field representation to estimate aqueous solvation energies. Solvation energy and hydrogen bonding energy contributions were only evaluated for the lowest energy structures. The dielectric constant was set to a value of 3.5. The simulation temperature was 310 K, the same used in the biological assay [14–17]. It was held constant during the simulation by coupling the system to a temperature external bath with a relaxation time of 0.01 ps [23].

The energy-minimized structure of the complex InhA-NAD-INH was used as initial structure in MD calculations. The MD simulations²⁴ protocol employed 500 000 steps with a step size of 0.001 ps (1 fs) at 310 K. An output trajectory file was saved every 20 simulation steps resulting 25 000 conformations. The solvation energy of the lowest energy conformation obtained from MD simulations was calculated using the hydration shell model [22]. The lowest energy conformation of the InhA-NAD-INH model was used to dock the energy-optimized structures of all ligands (adducts) employing the optimum model/alignment selected in [12] (HyperChem 7.51) [18]. The energy-minimized structure of each complex InhA-NAD-analogue was used to perform MD simulations following the same protocol mentioned before, and an output trajectory file was recorded every 20 simulation steps. The solvation energy and hydrogen bonding energy contributions of the lowest energy conformation from MD simulations of each InhA-NAD-analogue model (L-R bound state) were calculated. At this point, the L-R bound state thermodynamic descriptors were generated.

The INH/NAD adduct was extracted from the lowest energy conformation of the InhA-NAD-INH complex (HyperChem 7.51) [18] and the InhA model without the INH/NAD adduct was employed to obtain the thermody-

Table 1. Structures and Biological Activities of the 37 Hydrazides [a].

R—CONHNH₂											
Compound	R	pMIC	Compound	R	pMIC	Compound	R	pMIC	Compound	R	pMIC
INH1		4.70	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 [19–22]). The test set comprises the compounds Idv90, Idv128, Idv131, Idv132, and INHd51 (underlined letters). INH = isoniazid; INHd = aromatic, heteroaromatic, and ring substituted hydrazides, isoniazid derivatives. Idv = heterocyclic acid hydrazides and derivatives.

namic descriptors of the receptor unbound state (R unbound state). The energy-minimized structure of the InhA model without the INH/NAD adduct was used as initial structure to perform the MD simulations, as already described. The solvation energy and hydrogen bonding energy contributions of the R lowest energy conformation ob-

tained from MD simulations were calculated, and the R unbound state thermodynamic descriptors were generated.

Likewise, the thermodynamic descriptors of each ligand (L) in its unbound state were generated. The lowest energy conformation of each InhA-NAD-analogue model

Table 2. Thermodynamic descriptors from MD simulations and their definitions [9].

Descriptors	Definitions of the thermodynamics descriptors
$\Delta E_{\text{stre}} = ELR_{\text{stre}} - EL_{\text{stre}} - ER_{\text{stre}}$	Change in stretching energy upon binding
$\Delta E_{\text{bend}} = ELR_{\text{bend}} - EL_{\text{bend}} - ER_{\text{bend}}$	Change in bending energy upon binding
$\Delta E_{\text{tors}} = ELR_{\text{tors}} - EL_{\text{tors}} - ER_{\text{tors}}$	Change in torsion energy upon binding
$\Delta E_{\text{vdW}} = ELR_{\text{vdW}} - EL_{\text{vdW}} - ER_{\text{vdW}}$	Change in van der Waals energy upon binding
$\Delta E_{\text{el}} = ELR_{\text{el}} - EL_{\text{el}} - ER_{\text{el}}$	Change in electrostatic energy upon binding
$\Delta E_{E1,4} = ELR_{E1,4} - EL_{E1,4} - ER_{E1,4}$	Change in 1,4 interaction energy upon binding
$\Delta E_{\text{Hb}} = ELR_{\text{Hb}} - EL_{\text{Hb}} - ER_{\text{Hb}}$	Change in hydrogen bonding energy upon binding
$\Delta E_{\text{solv}} = ELR_{\text{solv}} - EL_{\text{solv}} - ER_{\text{solv}}$	Change in solvation energy upon binding
$\Delta E_{\text{stre+bend}} = ELR_{\text{stre+bend}} - EL_{\text{stre+bend}} - ER_{\text{stre+bend}}$	Sum of changes in stretching and bending energies
$\Delta E_{\text{stre+bend+tors}} = ELR_{\text{stre+bend+tors}} - EL_{\text{stre+bend+tors}} - ER_{\text{stre+bend+tors}}$	Sum of changes in stretching, bending and torsion energies
$\Delta E_{\text{el+Hb}} = ELR_{\text{el+Hb}} - EL_{\text{el+Hb}} - ER_{\text{el+Hb}}$	Sum of changes in electrostatic and hydrogen bonding energies
$\Delta E_{\text{el+Hb+E1,4}} = ELR_{\text{el+Hb+E1,4}} - EL_{\text{el+Hb+E1,4}} - ER_{\text{el+Hb+E1,4}}$	Sum of changes in electrostatic, hydrogen bonding and 1,4 interaction energies
$E_{LR}(\text{LL,RR,LR})$	Ligand-receptor complex energy
$E_{LR}(\text{LR})$	Intermolecular ligand-receptor energy
$E_{LR,\text{vdW}}$	Van der Waals intermolecular ligand-receptor energy
$E_{LR,\text{el}}$	Electrostatic intermolecular ligand-receptor energy
$E_{LR,\text{Hb}}$	Hydrogen bonding intermolecular ligand-receptor energy
$E_{LR,\text{el+Hb}}$	Sum of electrostatic and hydrogen bonding intermolecular ligand-receptor energies
$E_{LR,\text{el+Hb+vdW}}$	Sum of electrostatic, hydrogen bonding and van der Waals intermolecular ligand-receptor energies
$\Delta E_L(\text{LL}) = E_{LR}(\text{LL}) - E_L(\text{LL})$	Change in intramolecular ligand energy upon binding
$E_{LR}(\text{LL})$	Intramolecular energy of bound ligand
$\Delta E_R(\text{RR}) = E_{LR}(\text{RR}) - E_R(\text{RR})$	Change in intramolecular receptor energy upon binding
$E_{LR}(\text{RR})$	Intramolecular energy of bound receptor
$E_R(\text{RR})$	Intramolecular energy of unbound receptor
$E_{LR}(\text{LRM}) = ELR_{\text{solv}}$	Ligand-receptor complex solvation energy
$\Delta E_L(\text{LM}) = E_{LR}(\text{LM}) - E_L(\text{LM})$	Change in ligand solvation energy upon binding
$E_{LR}(\text{LM})$	Bound ligand solvation energy
$E_L(\text{LM}) = EL_{\text{solv}}$	Unbound ligand solvation energy
$\Delta E_R(\text{RM}) = E_{LR}(\text{RM}) - E_R(\text{RM})$	Change in receptor solvation energy upon binding
$E_{LR}(\text{RM})$	Bound receptor solvation energy
$E_R(\text{RM}) = ER_{\text{solv}}$	Unbound receptor solvation energy

from MD simulations was used to extract the adduct, analogue/NAD (HyperChem 7.51) [18]. The energy-minimized structure of each adduct model was employed as initial structure to perform the same MD simulations protocol. The solvation energy and hydrogen bonding energy contributions of each L lowest energy conformation from MD simulations were calculated.

The thermodynamic descriptors from MD calculations and their respective definitions [25] are presented in Table 2. The energy terms (52 descriptors) were used as independent variables to built QSAR models employing PLS regression and GFA algorithm, which are fitting functions available in the WOLF 5.5 program [26].

The GFA algorithm uses a population of many models and tests only the final, fully-constructed model. Improved models are constructed by performing the genetic crossover operation to recombine the terms of the better-performing models. The initial models are generated by randomly selecting some number of features from the training data set, building basis functions from these features using the user-specified basis functions types, and then con-

structing the genetic models from random sequences of that basis functions. GFA can build models using not only linear polynomials but also higher-order polynomials, splines, and other nonlinear functions [12]. In this study, the top eight QSAR models were selected by the WOLF 5.5 program [26]. Linear and second-degree polynomials were the functions tested.

Statistical measures of significance including the correlation coefficient (r^2), leave-one-out (LOO) cross-validation coefficient (q^2), least squares error (LSE), and lack-of-fit measure (LOF) developed by Friedman, were calculated to test the robustness of the models. The cross-correlation descriptor matrix was examined to eliminate trial QSARs in which pairs of energy terms have cross-correlation coefficients greater than 0.70 (GFA-PLS). Also, the cross-correlation matrix of residuals of fit between pairs of models was computed to determine if the top eight QSAR models provide common or distinct information. 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. The descriptor usage in a GFA analysis as a function of the number of crossovers was also monitored as an indication of statistical significance [12, 25]. The mutation probability over the crossover optimization cycle was set at 10%. We tested a number of genetic operations or crossovers of 50000 to 100000. The models are scored using Friedman's LOF measure, which is a penalized least-squares measure. The smoothing factor or parameter (d), which is part of the LOF definition, is the only parameter adjustable by the user [12], and it alters the balance between the number of independent variables (energy terms) in the models and the reduction in LSE measure, and it controls *overfitting*. The default value of smoothing factor is 1.0. In this study, smoothing factor values of 1.0 to 0.1 were tested for generating the RD 3D-QSAR models.

Here, the ligands of the training set whose differences in experimental and predicted activities exceeded 2.0 standard deviation, SD, from the mean of a model were considered as outliers.

Approaches to QSAR model validation, including y -randomization or y -scrambling, robust internal validation strategies such as multiple leave-many-out (LMO) cross-validations, and external validation were applied in this study whereas only validated QSAR models can offer a meaningful mechanistic interpretation, especially in the context of design or discovery of novel chemical agents with desired properties [27–29].

The model internal validation by LMO procedure employs smaller training sets than LOO procedure and can be repeated many more times due to possibility of larger combinations in leaving many compounds out from the training set. Here, LMO procedure was repeated up to ten compounds were left out from the training set. Ideal expectation is high average q^2 . In other words, if a QSAR model has a high average q^2 in LMO validation, it can be reasonably concluded that the obtained model is robust [28].

The y -randomization test is a widely used technique to ensure the robustness of a QSAR/QSPR model [27]. In this test, the dependent-variable vector, y -vector, is randomly shuffled and a new QSAR model is developed using the original independent-variable matrix. The process is repeated several times [28]. In the present study, that procedure was repeated ten times. It is expected that the resulting QSAR models should generally have low r^2 and low LOO q^2 values. Otherwise, if all QSAR models obtained in the y -randomization test have relatively high r^2 and LOO q^2 , it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modeling method, probably due to a chance correlation or structural redundancy of the training set.

2.3 External Validation

The six compounds of the test set were not included in the building of the 3D-QSAR models, but they were used to

validate the best QSAR model constructed from the training set and to evaluate its prediction capacity. It is recommended [29] that the external test set must contain at least five compounds, representing the whole range of both structure and activity of compounds included into the training set. The predicted activity value ($pMIC$) of each ligand in the test set was calculated using the equation of the best model by substitution of the energy values or thermodynamic descriptors from MD simulations at 310 K, which were selected as the most relevant to the biological activity.

3 Results and Discussion

The top eight models ($N=31$) selected by the WOLF 5.5 program, using a smoothing factor of 0.7; 10% probability of mutation for each crossover; and 70000 genetic operations or crossovers, presented two functions type (linear and quadratic or simply second-degree polynomial terms) suggesting differences in the mechanism of action of the investigated hydrazides set. Five of eight models had one outlier, the INHd41 compound, which is inactive. The atypical behavior of INHd41 is probably because its higher L-R complex energy value [$E_{LR}(LL,RR,LR) = -25799.13$ kcal/mol] when compared to the other investigated compounds (Table 3 – Supporting Information). Then, a new analysis was performed ($N=30$), using the supra-mentioned conditions.

The top eight models had an increment in their statistical measures (see Table 4 – Supporting Information), indicating better models. The r^2 and q^2 values ranged from 0.77 to 0.81, and from 0.61 to 0.68, respectively, whereas the LOF and LSE values varied from 0.56 to 0.58, and from 0.24 to 0.30, respectively. The number of energy descriptors changed from 5 to 6 and the number of outliers ranged from 0 to 2.

The cross-correlation matrix of residuals of fit between pairs of models was computed and they were highly correlated to one another ($R=0.84$ to 1.00) (Table 5 – Supporting Information). Thus, the models were judged to be nearly the same model meaning that there is a single unique model. Model 8 (Eq. 1) at 310 K was selected as the best model because it did not have any outliers (see Table 5 and Figure 1).

Model 8

$$pMIC = 3.05 - 0.0077449 [E_{LR}(LM) + 3.62]^2 - 0.000142 [EL_{el+Hb} + 118.13]^2 - 0.003716 [EL_{bend} - 63.86]^2 + 0.000008 [E_{LR,el} - 95.22]^2 - 0.000001 [\Delta E_{Hb} - 719.26]^2 \quad (1)$$

$N=30$, $r^2=0.77$, $q^2=0.61$, $LOF=0.58$, $LSE=0.30$, outliers=0

Where: $E_{LR}(LM)$ is the bound ligand solvation energy; EL_{el+Hb} is the sum of electrostatic and hydrogen bonding

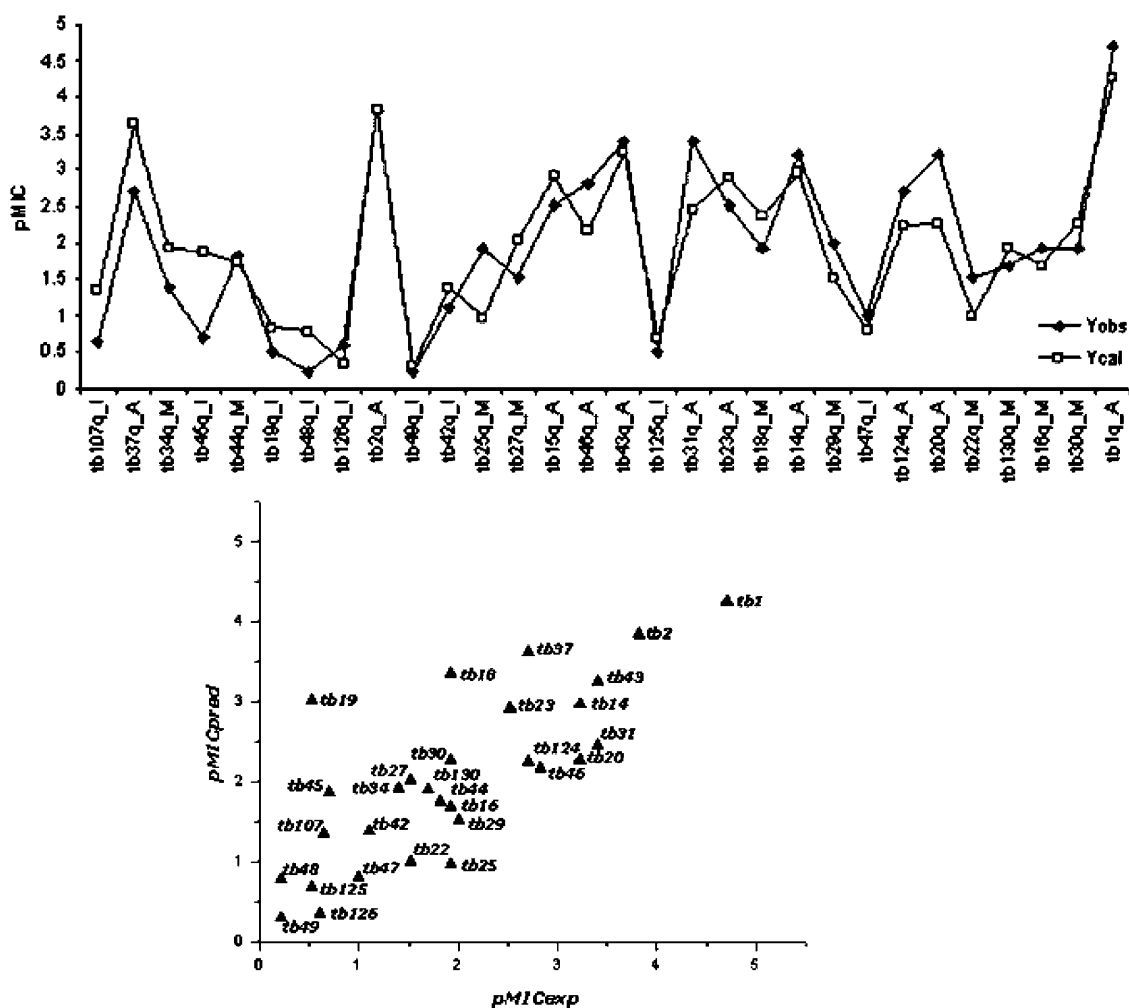


Figure 1. Predicted or calculated (Y_{cal} , white squares) and observed or experimental (Y_{obs} , black losanges) activity values found for the training set ($N=30$), considering model 8.

energies of unbound ligand; EL_{bend} is the bending energy of unbound ligand; $E_{\text{LR,el}}$ is the electrostatic intermolecular ligand-receptor energy; and ΔE_{Hb} is the change in hydrogen bonding energy upon binding.

The model 8 is a nonlinear model, hence as high is the value of sum or difference of the terms between [] as more significant the (favorable or unfavorable) contribution to biological activity. It is noteworthy that this model is composed of energy contributions of unbound ligand [$EL_{\text{el+Hb}}$ and EL_{bend}], bound ligand or L-R complex [E_{LR} (LM), $E_{\text{LR,el}}$], and changes of energy upon binding [ΔE_{Hb}], which considers the both states, bound and unbound. Moreover, the electrostatic intermolecular ligand-receptor energy is the only favorable contribution (positive regression coefficient) to the biological activity (pMIC).

Those energy descriptors can be interpreted in terms of specific L-R binding and they incorporate the dynamic features of the chemical system where the molecular flexibility is extensively embedded in performing a RD 3D-QSAR analysis.

The ΔE_{Hb} term, for example, is the total change in hydrogen bonding of the ligand and the receptor upon L-R binding. This term is a measure of how much hydrogen bonding energy the isolated ligand and the isolated receptor sacrifice to achieve the bound L-R state. The hydrogen bonding contribution has also a polar or electronic character and must be compensated, in some way, by the $E_{\text{LR,el}}$ term, that is the electrostatic binding energy, according the Equation 1 (the regression coefficients present opposite signals).

The E_{LR} (LM) descriptor corresponds to the bound ligand aqueous solvation energy and its related to the water interactions between the ligand and the amino acid residues into the active site of the InhA enzyme. That term becomes more negative as the ligand becomes more soluble in the L-R complex. Thus, there is an intermediate, but optimum, aqueous solvation energy with respect to minimizing the unfavorable contribution to the desirable activity.

The $E_{\text{LR,el}}$ and $EL_{\text{el+Hb}}$ terms can be considered together because they work against one another in the binding pro-

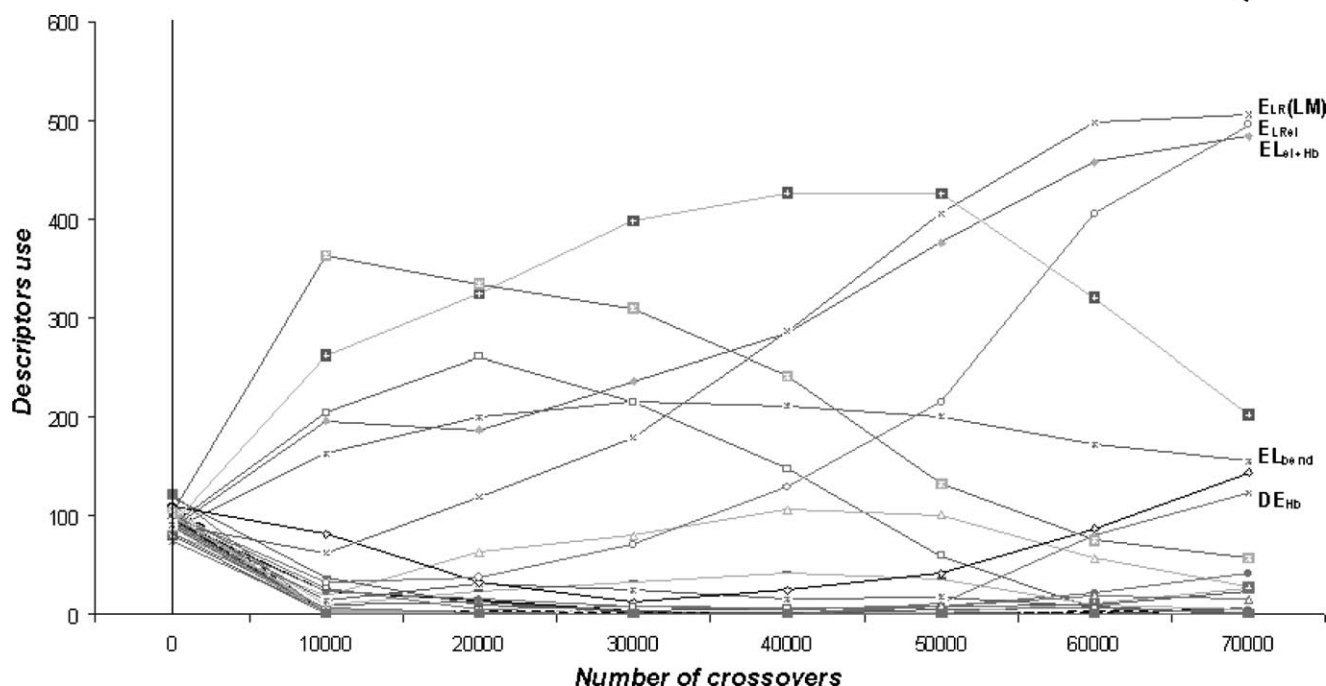


Figure 2. The descriptor usage of the energy terms is plotted as a function of the number of crossovers in the GFA analysis. The thermodynamic descriptors in this figure are defined in Table 2 and were calculated at 310 K. Not all terms are shown but rather the most often used terms in model evolution, particularly those that appeared in model 8. A smoothing factor of 0.7 was used.

cess (see Eq. 1). EL_{el+Hb} is the sum of electrostatic and hydrogen bonding energies of the isolated ligand. If an unbound ligand presents a very lower value of EL_{el+Hb} it will have more difficulty to form the L-R complex, contributing more negatively to the inhibitory activity.

The bound state energy descriptors [$E_{LR}(LM)$, $E_{LR,el}$] and the energy term upon binding [ΔE_{Hb}] in model 8 can be related to those grid cell occupancy descriptors (GCODs) selected in the best *RI* 4D-QSAR model [10], particularly the GCODs or GCs whose the interaction pharmacophore elements (IPEs) were any atom group types (GC2, GC3, GC4, and GC5). The GC2, GC3, GC4, and GC5 were identified as the appropriate occupancies by groups that reflect both hydrogen bond acceptor and donor interactions considering the amino acid residues or/and water molecules in the InhA active site [10]. Moreover, the GC4 was located on the nitrogen atom of the pyridine ring (0.6 Å) of the INH1/NAD adduct [10], and could also be associated with the unbound energy term [EL_{el+Hb}].

The EL_{bend} descriptor is the bending energy of the isolated ligand and how much bent is the ligand, higher is the EL_{bend} value and more negative is the contribution to the activity. This energy term represents the intramolecular ligand interactions, which are undesirable and are responsible for impairing the L-R binding process [10].

A linear cross-correlation matrix of the energy descriptors found for model 8 (Eq. 1) was built and none of them were highly correlated to one another, since all pair corre-

lations of energy terms were lesser than 0.7 ($R=0.01$ to 0.27) (see Table 6 – Supporting Information) meaning that each of the energy terms provides independent information to the optimal 3D-QSAR model.

A crossover versus descriptors usage plot reveals the relative significance of the independent variables. The more times a descriptor is used in generating new models, the greater its relative role in explaining variance in the biological activity [12, 26]. The results of a GFA optimization analysis for predicting *pMIC* for the thermodynamic descriptors calculated at 310 K are presented in Figure 2. The energy terms $E_{LR}(LM)$, $E_{LR,el}$ and EL_{el+Hb} are more often used in building *RD* 3D-QSAR models as the number of crossovers increases in the evolution of the GFA optimization.

The internal validation LMO procedure (internal prediction power) and γ -randomization technique (check for chance of correlations) were carried out to verify the robustness of model 8. In Table 7 (Supporting Information) are presented the obtained q^2_{LMO} values when up to ten compounds from the training set were left out (m is the number of objects excluded in the internal validation process which varied from 2 to 10).

Good QSAR models must have q^2_{LMO} values closest to the q^2_{LOO} value of the selected best model. Furthermore, the q^2_{LMO} values must be closest to the average, $\langle q^2_{LMO} \rangle$, and the oscillation range accepted is 0.1. All q^2_{LMO} values are closest to the q^2_{LOO} value (0.61) (Table 7 – Supporting Information). Also, the q^2_{LMO} values oscillated from 0.01

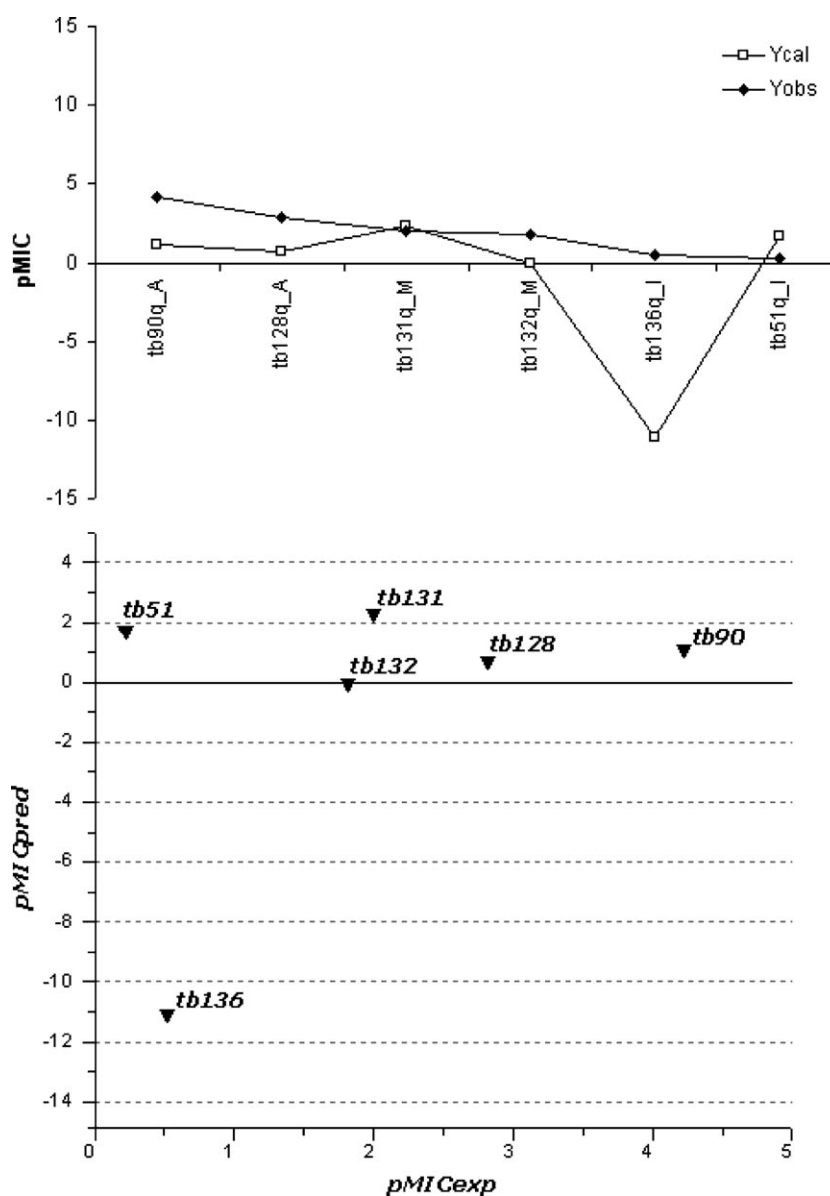


Figure 3. External validation – predicted or calculated (Y_{cal} , white squares) and observed or experimental (Y_{obs} , black losanges) activity values found for the test set ($N=6$), considering Model 8.

to 0.07 in comparison to the average value [$\langle q^2_{\text{LMO}} \rangle = 0.60$], indicating a good internal predictability.

In Table 8 (Supporting Information) are shown the resulting LOO q^2 and r^2 values when the \mathbf{y} -vector was randomly shuffled ten times and new ten QSAR models were developed for the same data set, using the original independent-variable matrix and the same conditions employed in the building of the selected best QSAR model (model 8). All QSAR models obtained in the \mathbf{y} -randomization test have low LOO q^2 and r^2 values, it implies that an acceptable QSAR model can be obtained for the given data set by the current modeling method, which is Model 8.

The pMIC value of each of the test set ligands (adducts) was calculated using Equation 1, as described in Sec. 2.3.

Five of the six ligands of the test set had residuals whose absolute values were lesser than or equal to the standard deviation value from the mean of the model (Table 9 – Supporting Information, and Figure 3), indicating that model 8 has a good external predictive power (83.33%). Additionally, a random sampling scheme similar to the one used in bootstrapping was applied to truly assess the external predictability. The data set ($N=36$) was split ten times (training set $N=30$; test set $N=6$) and the GFA-PLS was running on the existing descriptor again. The prediction power of test set remained as 83.33% (see Table 10 – Supporting Information).

The test set compound Idiv136 did not have its activity well predicted by model 8 probably because its behavior in

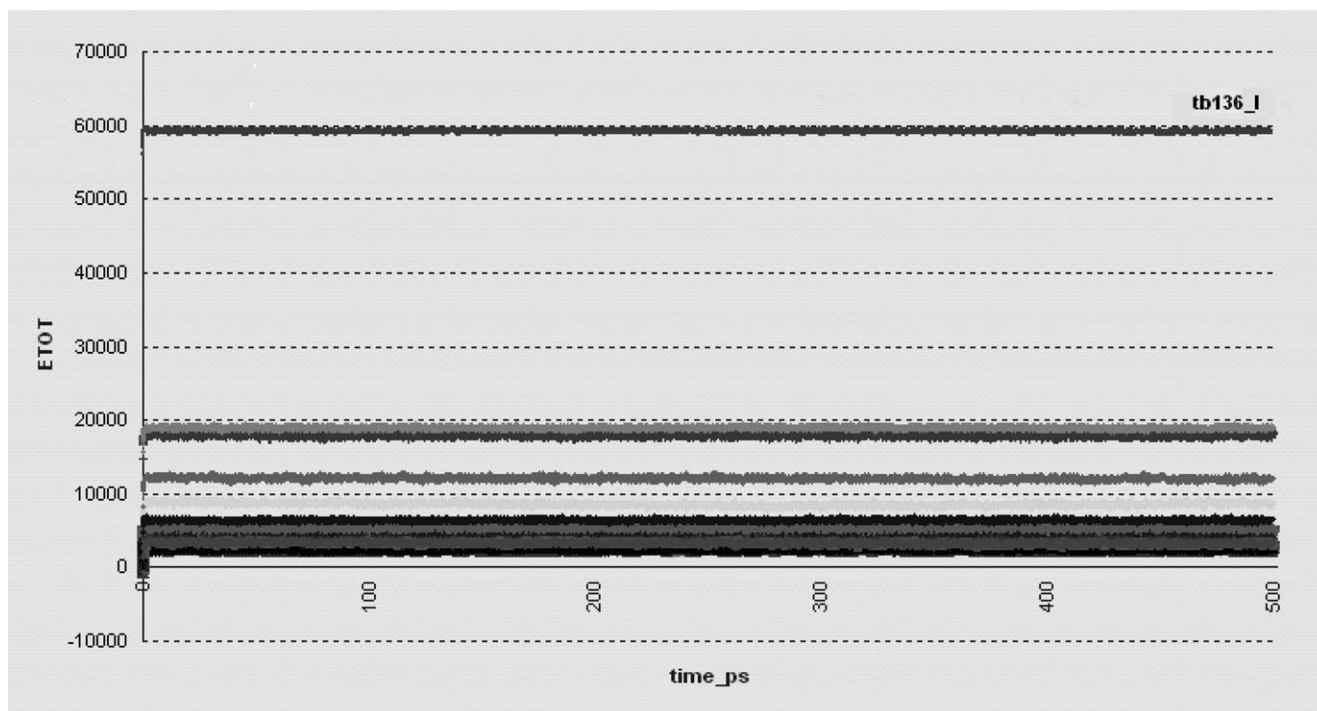


Figure 4. Plot of total energy (E_{tot} , kcal/mol) of L-R complexes versus time (ps) from MD simulations at 310 K. $E_{\text{total}} = ELR_{\text{stre}} + ELR_{\text{bend}} + ELR_{\text{tors}} + ELR_{E_{1,4}} + ELR_{\text{vdW}} + ELR_{\text{el}} + ELR_{\text{vdW+el}}$. The solvation and hydrogen bonding energies are not computed in this diagram.

the bound state during the MD simulations at 310 K. The L-R conformational ensemble profile of the complex tb136q_I (tb136q refers to the adduct Idv136/NAD in complex with InhA; I = inactive) presented very distinct energy contributions from the rest of the L-R complexes investigated (see Fig. 4). The total energy (E_{total}) corresponds to the summation of the following L-R complex energy (ELR) contributions during the MD simulations at 310 K: stretching energy (ELR_{stre}), bending energy (ELR_{bend}), torsion energy (ELR_{tors}), Lennard–Jones or 1,4 interactions energy ($ELR_{E_{1,4}}$), intramolecular van der Waals energy (ELR_{vdW}), intramolecular electrostatic energy (ELR_{el}), and sum of intermolecular van der Waals and electrostatic energies ($ELR_{\text{vdW+el}}$).

A comparison of the partitioned energy contributions values from all lowest energy minimum L-R conformations investigated (bound state) at 310 K was carried out. The highest intramolecular van der Waals energy value found for the tb136q_I complex seems to be the responsible for the significant increase in its total energy. Also, that energy contribution in the bound state affects directly the sum of intermolecular van der Waals and electrostatic energies contribution.

4 Conclusions

Considering the conditions adopted in this study, the *RD* 3D-QSAR model at 310 K has good internal and external

predictability and can be taken into account in the binding process of ligands to InhA. A larger set of compounds is already being tested to verify the reliability of the model generated. However, if the biological data were expressed as binding constants and/or inhibition in vitro constants, such as K_i or IC_{50} , the findings could be even better, since those constants provide a more effective estimative of the binding free energies (ΔG).

5 Acknowledgement

The authors are grateful to FAPESP for financial support and to the Chem21 Group, Inc., for providing the license of the MOLSIM 3.2 and WOLF 5.5 programs used in this study.

6 References

- [1] H. Bergler, S. Fuchsbichler, G. Högenauer, F. Turnowsky, *Eur. J. Biochem.* **1996**, *242*, 689–694.
- [2] M. Stewart, S. Parikh, G. Xiao, P. J. Tonge, C. Kisker, *J. Mol. Biol.* **1999**, *290*, 859–865.
- [3] D. A. Rozwarski, C. Vilchèze, M. Sugantino, R. Bittman, J. C. Sacchettini, *J. Biol. Chem.* **1999**, *274*, 15582–15589.
- [4] C. E. Barry, III, R. E. Lee, K. Mdluli, A. E. Sampson, B. G. Schroeder, R. A. Slayden, Y. Yuan, *Prog. Lipid Res.* **1998**, *37*, 143–179.

- [5] A. D. McCarthy, D. G. Hardie, *Trends Biochem.* **1984**, *9*, 60–63.
- [6] K. Magnuson, S. Jackowski, C. O. Rock, J. E. Cronan Jr., *Microbiol. Rev.* **1993**, *57*, 522–542.
- [7] K. F. M. Pasqualoto, E. I. Ferreira, *Curr. Drug Targets* **2001**, *2*, 427–437.
- [8] P. J. Brennan, H. Nikaido, *Ann. Rev. Biochem.* **1995**, *64*, 29–63.
- [9] D. A. Rozwarski, G. A. Grant, D. H. R. Barton, W. R. Jacobs, Jr., J. C. Sacchettini, *Science* **1998**, *279*, 98–102.
- [10] K. F. M. Pasqualoto, E. I. Ferreira, O. A. Santos-Filho, A. J. Hopfinger, *J. Med. Chem.* **2004**, *47*, 3755–3764.
- [11] A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Jin, M. G. Albuquerque, P. J. Madhav, C. Duraiswami, *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- [12] D. Rogers, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- [13] W. G. Glen, W. J. Dunn, III, D. R. Scott, *Tetrahedron Comput. Methodol.* **1989**, *2*, 349–354.
- [14] J. Bernstein, W. A. Lott, B. A. Steinberg, H. L. Yale, *Am. Rev. Tuberc.* **1952**, *65*, 357–364.
- [15] J. Bernstein, W. P. Jambor, W. A. Lott, F. Pansy, B. A. Steinberg, H. L. Yale, *Am. Rev. Tuberc.* **1953**, *67*, 354–365.
- [16] J. Bernstein, W. P. Jambor, W. A. Lott, F. Pansy, B. A. Steinberg, H. L. Yale, *Am. Rev. Tuberc.* **1953**, *67*, 366–375.
- [17] G. Klopman, D. Fercu, J. Jacob, *Chem. Phys.* **1996**, *204*, 181–193.
- [18] *HyperChem Program Release 7.51 for Windows*; Hypercube, Inc., Gainesville, FL **2002**.
- [19] D. Doherty, MOLSIM: *Molecular Mechanics and Dynamics Simulation Software*, User's Guide, Version 3.2, The Chem21 Group Inc., Chicago, IL **1997**.
- [20] M. J. S. E. Dewar, G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- [21] S. J. Weiner, P. A. Kollman, D. T. Nguyen, D. A. Case, *J. Comput. Chem.* **1986**, *7*, 230–252.
- [22] A. J. Hopfinger, *Conformational Properties of Macromolecules*, Academic Press, New York **1973**.
- [23] H. J. C. Berendsen, J. P. M. Postman, W. F. van Gunsteren, A. di Nola, J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- [24] W. F. van Gunsteren, H. J. C. Berendsen, *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.
- [25] J. S. Tokarski, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792–811.
- [26] D. Rogers, *WOLF Reference Manual Version 5.5*, The Chem21 Group Inc., Chicago, IL **1994**.
- [27] S. Wold, L. Eriksson, in *Chemometric Methods in Molecular Design* (Ed: H. van de Waterbeemd), VCH, Weinheim **1995**, 309–318.
- [28] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69–77.
- [29] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Mod.* **2002**, *20*, 269–276.