

# Classical and Hologram QSAR Studies on a Series of Inhibitors of Trypanosomatid Glyceraldehyde-3-Phosphate Dehydrogenase

Rafael V. C. Guido<sup>a</sup>, Marcelo S. Castilho<sup>b</sup>, Sabrina G. R. Mota<sup>b</sup>, Glaucius Oliva<sup>a</sup> and Adriano D. Andricopulo<sup>a\*</sup>

<sup>a</sup> Laboratório de Química Medicinal e Computacional, Centro de Biotecnologia Molecular Estrutural, Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador São-carlense 400, 13560-970 São Carlos, SP, Brazil, E-mail: aandrico@if.sc.usp.br

<sup>b</sup> Laboratório de Bioinformática e Modelagem Molecular, Faculdade de Farmácia, Universidade Federal da Bahia, Campus Universitário de Ondina, 40170-290 Salvador, BA, Brazil

**Keywords:** Enzyme inhibitors, GAPDH, QSAR, Tropical diseases, Trypanosomatids

Received: September 26, 2007; Accepted: February 02, 2008

DOI: 10.1002/qsar.200710139

## Abstract

Leishmaniasis and trypanosomiasis are major causes of morbidity and mortality in both tropical and subtropical regions of the world. The current available drugs are limited, ineffective, and require long treatment regimens. Due to the high dependence of trypanosomatids on glycolysis as a source of energy, some glycolytic enzymes have been identified as attractive targets for drug design. In the present work, classical Two-Dimensional Quantitative Structure–Activity Relationships (2D QSAR) and Hologram QSAR (HQSAR) studies were performed on a series of adenosine derivatives as inhibitors of *Leishmania mexicana* Glyceraldehyde-3-Phosphate Dehydrogenase (*Lm*GAPDH). Significant correlation coefficients (classical QSAR,  $r^2=0.83$  and  $q^2=0.81$ ; HQSAR,  $r^2=0.91$  and  $q^2=0.86$ ) were obtained for the 56 training set compounds, indicating the potential of the models for untested compounds. The models were then externally validated using a test set of 14 structurally related compounds and the predicted values were in good agreement with the experimental results (classical QSAR,  $r_{pred}^2=0.94$ ; HQSAR,  $r_{pred}^2=0.92$ ).

## 1 Introduction

Parasitic diseases are the foremost threat to human health and welfare around the world. In tropical and subtropical regions of the world, the consequences of parasitic infection are devastating both in terms of human morbidity and mortality, reaching dangerous levels which threaten future social stability and economic development of these areas. The currently available drugs for most of the parasitic diseases are inadequate, poorly tolerated, and ineffective [1–3]. A group of tropical diseases, including leishmaniasis and trypanosomiasis, is neglected and lie outside of the world pharmaceutical market. On one hand, it is estimated that 12 million people are currently affected by leishmaniasis in its different forms, with about 350 million living in risk areas. The most severe form, visceral leishmaniasis or

kala azar, is a serious disease that affects 500 000 people resulting in the death of over 65 000 every year [4]. On the other hand, Chagas' disease (also called American trypanosomiasis) is a tropical parasitic disease which occurs in Latin America, particularly in South America, where approximately 16–18 million people are infected causing over 45 000 deaths each year [5, 6].

Since safe, effective, and affordable chemotherapeutic agents against trypanosomatids are urgently needed, the identification of new macromolecular targets and small-molecule modulators is of utmost importance. The bloodstream form of trypanosomatids has no functional tricarboxylic acid cycle and is highly dependent on glycolysis for ATP production [7, 8]. The crucial dependence on glycolysis as a source of energy makes the glycolytic parasite enzymes promising targets for drug design. Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH, Enzyme Commission 1.2.1.12) is a key enzyme in the glycolytic cascade, which catalyzes the reversible oxidative phosphorylation of the substrate Glyceraldehyde-3-Phosphate (GAP) to 1,3-diphosphoglycerate in the presence of the cofactor Nicotinamide Adenine Dinucleotide (NAD<sup>+</sup>) and inor-

**Abbreviations:** GAPDH, glyceraldehyde-3-phosphate dehydrogenase; HQSAR, hologram quantitative structure–activity relationships; LMO, leave-many-out; LOO, leave-one-out; PLS, partial least squares; QSAR, quantitative structure–activity relationships; SDEP, standard error of prediction

ganic phosphate. The glycolytic GAPDH enzymes from the pathogenic protozoan parasites *Trypanosoma brucei* [9], *T. cruzi* [10], and *Leishmania mexicana* [11] are quite similar to each other (about 90% sequence identity) and have important structural differences when compared to their human counterpart [12] (about 45% sequence identity). Comparisons between parasite and human GAPDH crystal structures revealed significant differences in the neighboring of the NAD<sup>+</sup> (adenosine moiety) binding site, which have been explored for the development of new selective inhibitors of the trypanosomatid enzymes [13–17].

Quantitative Structure–Activity Relationships (QSAR) have been successfully employed as a valuable tool to assist the design of compounds of pharmaceutical interest [18–20]. As part of a research program aimed at the design of selective GAPDH inhibitors [21–24], and in order to evaluate the quantitative relationships of a large series of adenosine derivatives, we have created predictive 2D QSAR models employing both classical and hologram QSAR (HQSAR) methods [25–27].

## 2 Experimental Section

### 2.1 Data Sets

The data set used for the QSAR studies contains 70 inhibitors of GAPDH (adenosine derivatives), which were selected from the literature [13–17]. The structures and corresponding pIC<sub>50</sub> (–log IC<sub>50</sub>; where IC<sub>50</sub> is the concentration required for 50% inhibition of GAPDH) values for the whole set of inhibitors are included in Table 1. The complete data set was divided into training (compounds 1–56, Table 1) and test (compounds 57–70, Table 1) sets in the ratio of 5:1 (20%). A statistical cluster analysis was carried out with Tsar 3D version 3.3 (Accelrys, San Diego, USA) using the complete linkage clustering method (Euclidean distances) with no data standardization. The cluster analysis grouped the compounds into two major clusters, with a relatively well-balanced distribution of training and test set compounds. Considering that the structural diversity of the data set is mostly represented by adenine analogues of moderate structural diversity, the composition of the QSAR training and test sets is necessarily representative of the whole data set in terms of chemical similarity.

### 2.2 Descriptors Calculation and Selection

Classical 2D QSAR studies require the calculation of a variety of molecular descriptors that are used as independent variables in QSAR modeling. In the present work, these studies were carried out using the DRAGON 5.4 (Taletto SRL, Milan, Italy), BUILDQSAR [28], and PIROUETTE 3.11 software (Infometrix, Washington, USA). The 2D

molecular descriptors, including topological descriptors, connectivity indices, 2D autocorrelation descriptors, Burden eigenvalues indices, among others, were computed using the software DRAGON 5.4 and used as independent variables in the QSAR analyses. Approximately 930 molecular descriptors were subjected to the following selection criteria. Descriptors with constant values or found to have poor correlation to biological property were discarded ( $r^2 < 0.10$ ). This strategy afforded 800 descriptors. Then, the BUILDQSAR software was employed to systematically search for models of up to 4 variables that give rise to MLR models with  $r^2 > 0.70$ . All descriptors present in the MLR models were pooled together, autoscaled, and used for the Partial Least Square (PLS) analysis performed with the PIROUETTE software.

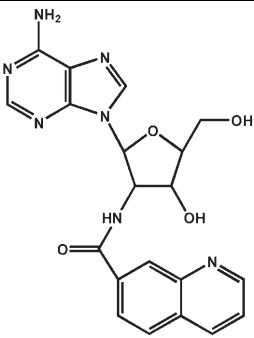
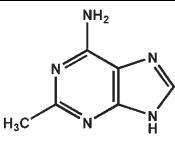
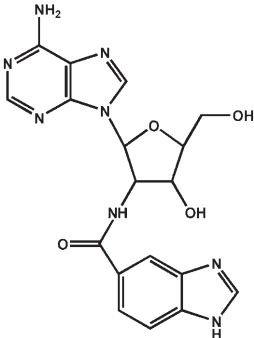
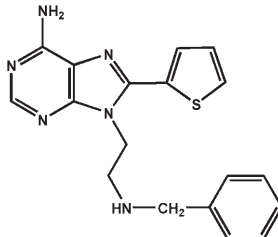
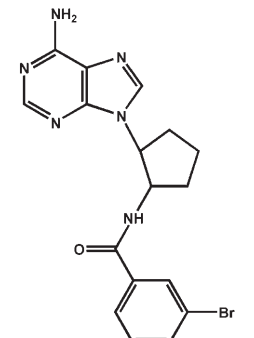
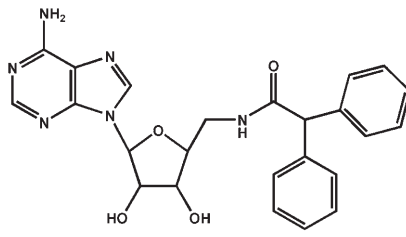
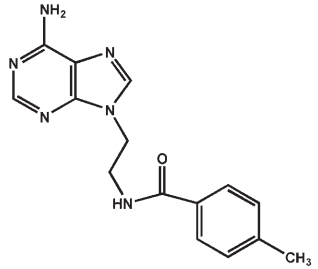
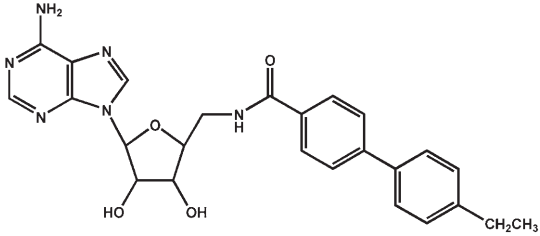
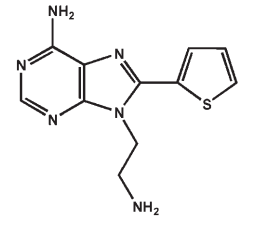
### 2.3 HQSAR Analysis

The HQSAR modeling analyses, calculations, and visualizations were performed using the SYBYL 7.3 package (Tripos Inc., St. Louis, USA) running on Red Hat Linux 7.3 workstations, as previously described [25, 27]. This method requires selecting values for parameters that specify the length of the hologram as well as the size and type of the fragments that are to be encoded [25, 27]. The generation of the molecular holograms was carried out using several combinations of the following fragment distinction: atoms (A), bonds (B), connections (C), hydrogen atoms (H), chirality (Ch), and donor and acceptor (DA). The fragment size default (4–7) was used to generate the QSAR models. HQSAR analysis was performed over the 12 default series of hologram lengths of 53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, and 401 bins. The influence of fragment size, which controls the minimum and maximum length of fragments to be included in the hologram, was further investigated by using 6 distinct fragment sizes over the 12 default series of hologram lengths values ranging from 53 to 401 bins. The patterns of fragment counts from the training set inhibitors were then related to the experimental biological data using the PLS analysis.

### 2.4 QSAR Model Validation

All QSAR models were investigated using full cross-validated  $r^2$  ( $q^2$ ) PLS. Leave-One-Out (LOO) cross-validation was applied to determine the number of components that yield optimally predictive models. Leave-Many-Out (LMO) with 10 (LMO<sub>10</sub>) and 5 (LMO<sub>5</sub>) randomly selected groups was used as a more rigorous test to assess model stability and statistical significance. Each random cross-validation run was repeated 25 times to obtain mean values for  $q^2$  and the corresponding Standard Error of Prediction (SDEP). External validation was performed with a test set of 14 compounds which were not considered for QSAR model generation.

**Table 1.** Chemical structures and corresponding  $pIC_{50}$  values for a series of inhibitors of *Lm*GAPDH.

Training set compounds					
Compound	Structure	$pIC_{50}$	Compound	Structure	$pIC_{50}$
1		3.30	2		2.52
3		3.15	4		2.40
5		2.80	6		2.22
7		3.40	8		3.60
9		3.12			

**Table 1.** (continued)

Training set compounds

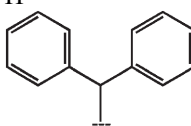
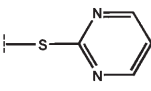
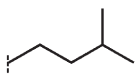
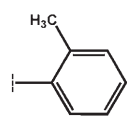
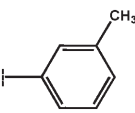
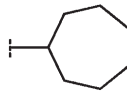
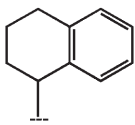
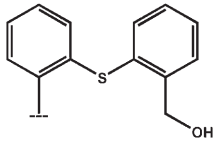
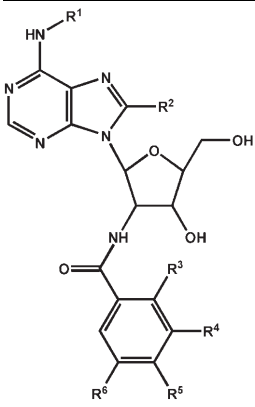
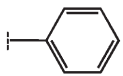
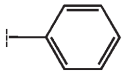
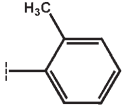
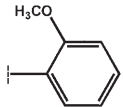
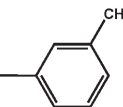
Compound	$R^1$	$R^2$	$R^3$	$pIC_{50}$
<b>10</b>	CH <sub>3</sub>	H	H	2.22
<b>11</b>	H	H		2.40
<b>12</b>	H		H	2.43
<b>13</b>	H		H	3.15
<b>14</b>	H		H	3.15
<b>15</b>	H		H	2.74
<b>16</b>	H		H	3.44
<b>17</b>	H		H	3.47
<b>18</b>	H		H	3.62

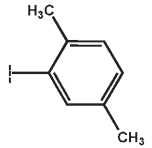
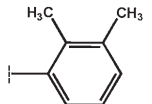
Table 1. (cont.)

## Training set compounds



Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	$pIC_{50}$
19	H	H	H	H	H	H	2.48
20	H	H	H	H	C <sub>6</sub> H <sub>5</sub>	H	3.70
21	H	H	H	OCH <sub>2</sub> CH <sub>3</sub>	H	H	3.40
22	H	H	H	OCH <sub>3</sub>	OH	H	3.30
23	H	Br	H	H	H	H	2.52
24	H	H	H	OH	OH	OH	3.60
25	H	H	H	OH	H	OH	3.19
26		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	4.70
27		H	H	OCH <sub>3</sub>	H	H	4.80
28		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	4.60
29		H	H	OCH <sub>3</sub>	H	H	4.60
30		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	4.60

Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	$pIC_{50}$
31		H	H	OCH <sub>3</sub>	H	H	4.60
32		H	H	OCH <sub>3</sub>	H	H	5.26

**Table 1.** (cont.)

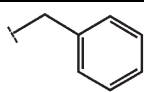
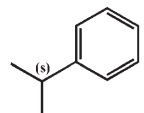
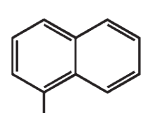
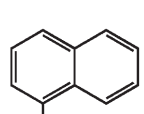
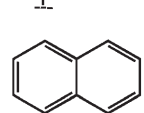
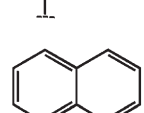
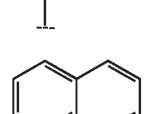
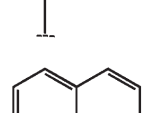
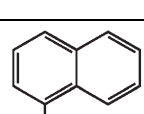
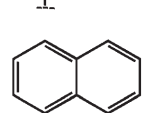
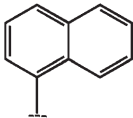
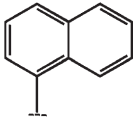
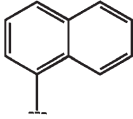
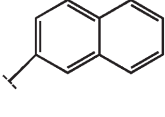
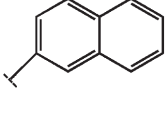
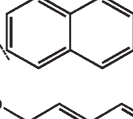
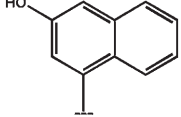
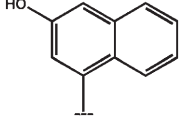
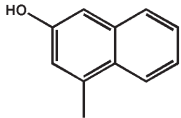
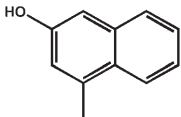
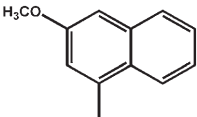
Training set compounds							
Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	$pIC_{50}$
33		H	H	OCH <sub>3</sub>	H	H	3.82
34		H	H	OCH <sub>3</sub>	H	H	4.10
35		H	H	Cl	H	H	5.00
36		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	5.70
37		H	H	Cl	H	Cl	4.92
38		H	H	OAc	H	OAc	5.00
39		H	H	OH	H	OH	5.30
40		H	H	H	OAc	H	4.00
Training set compounds							
Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	$pIC_{50}$
41		H	H	H	OH	H	4.10
42		H	H	OAc	OAc	H	4.60

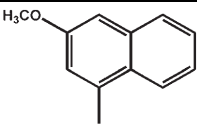
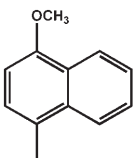
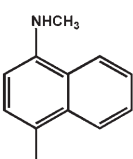
Table 1. (cont.)

Training set compounds							
Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	pIC <sub>50</sub>
43		H	H	OH	OCH <sub>3</sub>	H	4.43
44		H	H	OCH <sub>3</sub>	OH	H	5.00
45		H	H	N(CH <sub>3</sub> ) <sub>2</sub>	H	H	4.60
46		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	5.70
47		H	H	Cl	H	Cl	4.60
48		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	4.22
49		H	H	OCH <sub>3</sub>	H	H	5.22
50		H	H	Cl	H	H	5.40

## Training set compounds

Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	pIC <sub>50</sub>
51		H	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	5.70
52		H	H	Cl	H	Cl	4.60
53		H	H	OCH <sub>3</sub>	H	H	5.70

**Table 1.** (cont.)

Training set compounds							
Compound	$R^1$	$R^2$	$R^3$	$R^4$	$R^5$	$R^6$	$pIC_{50}$
54		H	H	Cl	H	Cl	4.60
55		H	H	OCH <sub>3</sub>	H	H	5.30
56		H	H	OCH <sub>3</sub>	H	H	4.08

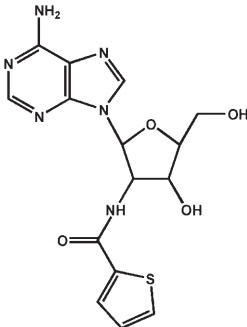
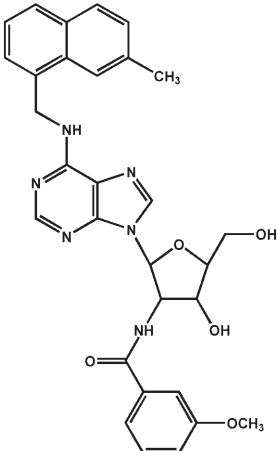
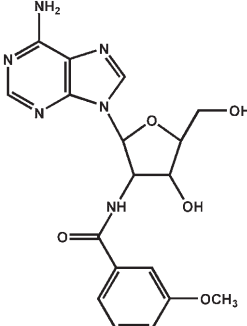
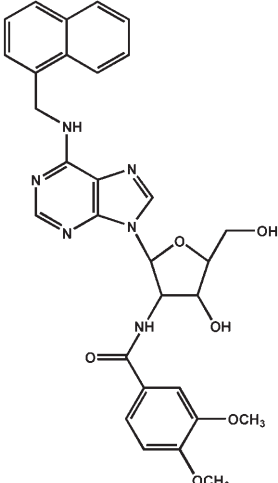
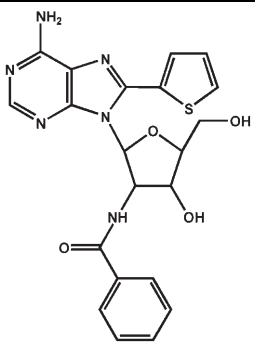
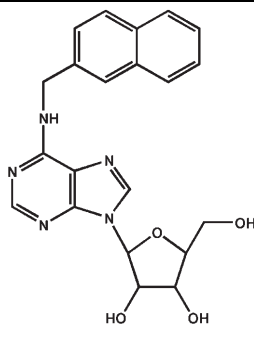
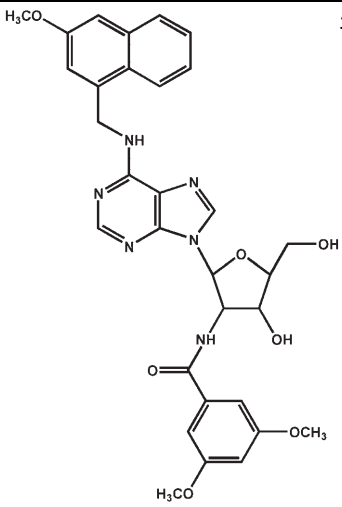
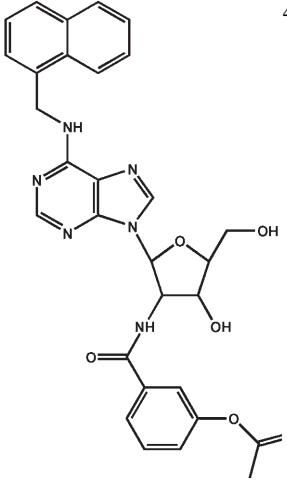
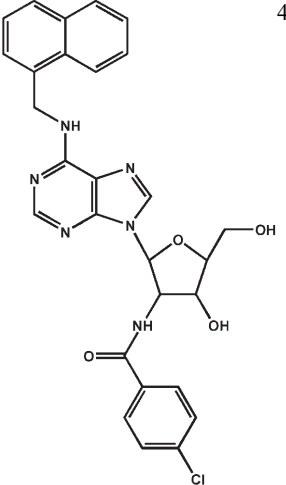
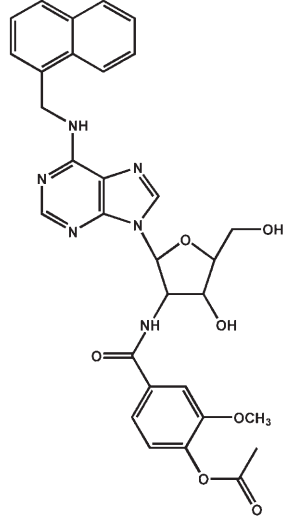
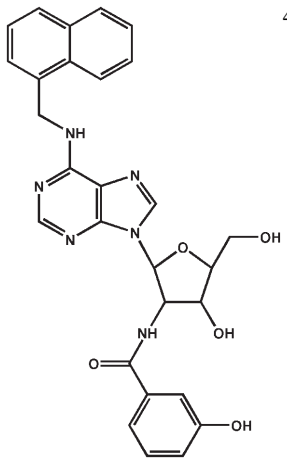
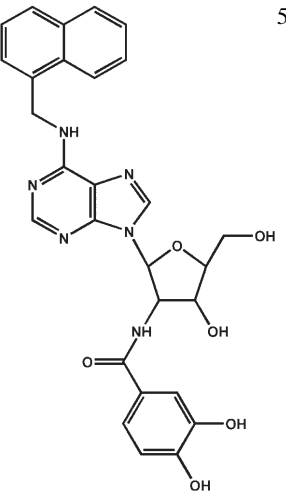
CompoundStructure	$pIC_{50}$ CompoundStructure	$pIC_{50}$ CompoundStructure	$pIC_{50}$ CompoundStructure
57 	2.62 <b>58</b>	3.52 <b>59</b>	3.30 
60 	3.22 <b>61</b>	3.07 <b>62</b>	3.82 



Table . (cont.)

CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub>	
<b>63</b> 	5.30	<b>64</b> 	5.40	<b>65</b> 	5.70
CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub> CompoundStructure	pIC <sub>50</sub>	
<b>66</b> 	4.30	<b>67</b> 	4.60	<b>68</b> 	4.00
<b>69</b> 	4.74	<b>70</b> 	5.00		

### 3 Results and Discussion

#### 3.1 Chemical and Biological Data

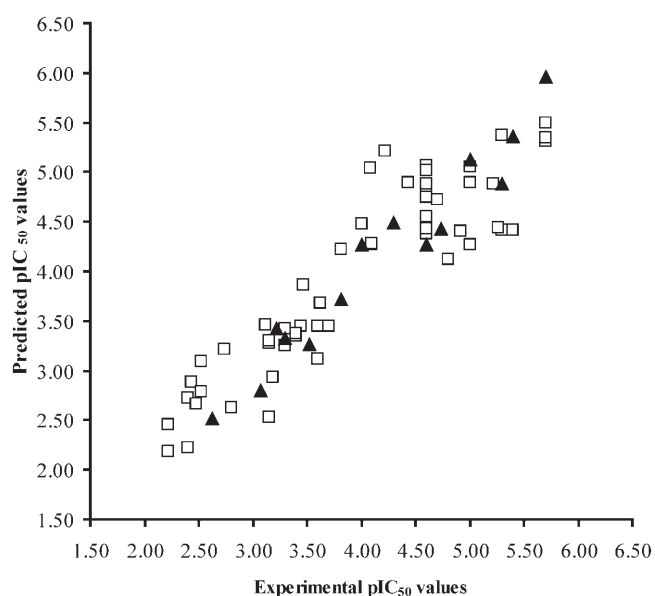
Classical QSAR and HQSAR models were derived for a series of 70 inhibitors of trypanosomatid GAPDHs listed in Table 1 [13–17]. The family of adenosine derivatives examined has moderate structural diversity. The bulk of the structural diversity lies in the nature of the substituent linked to the N<sup>6</sup>- and 9-position of the adenine system. The *in vitro* IC<sub>50</sub> values of *Lm*GAPDH employed in this work were measured under the same experimental conditions [13–17], a fundamental requirement for QSAR studies [29, 30]. The generation of consistent statistical models is dependent on the adequacy of the training and test sets. Therefore, from the original data set, 56 inhibitors (compounds **1–56**, Table 1) were selected as members of the training set for model generation, whereas the other 14 inhibitors (compounds **57–70**, Table 1) were held as members of the test set for external validation. A statistical cluster analysis confirmed that structurally diverse molecules possessing activities of wide range were included in both training and test sets. The pIC<sub>50</sub> values used as dependent variables in the QSAR analyses span about three and a half orders of magnitude and are acceptably distributed across the range of values. Thus, the data set is appropriate for the purpose of QSAR model development.

#### 3.2 Classical QSAR Analysis

The strategy used in this work with the program DRAGON 5.4 resulted in the selection of 12 descriptors (Table 2) that were then explored using more robust statistical methods such as Principal Component Analysis (PCA) and PLS, as implemented in the PIROUETTE software.

The PCA results indicated that three principal components accounted for 95% of total variance, while other additional components did not significantly contribute to these models. The first PC broadly accounts for potency,

while PC2 roughly describes molecular steric properties. The less potent inhibitors have negative PC1 values whereas the most potent ones display positive values. These preliminary results prompted us to use the selected descriptors for QSAR modeling studies. PLS QSAR models were created using the LOO cross-validation procedure, and the results are shown in Figure 1. As it can be seen, the best statistical model has substantial correlation within the training set as shown by the high values of  $r^2$  (0.83) and  $q^2$  (0.81). The use of the LMO cross-validation protocols did not lead to considerable quantitative changes in the parameters of the final models (LMO<sub>5</sub>  $q^2$  = 0.81, and LMO<sub>10</sub>  $q^2$  = 0.80), thus confirming the stability and statistical significance of the model generated with LOO.



**Figure 1.** Plot of predicted vs. experimental values of pIC<sub>50</sub> for the training (open squares) and test (black triangle) set GAPDH inhibitors for the best classical QSAR model.

**Table 2.** Descriptors selected for QSAR model development<sup>a</sup>.

Symbol	Definition and description
<b>Har2</b>	Square reciprocal distance sum index
dZ	Pogliani index
<b>SRW01</b>	Self-returning walk count of order 01
RHyDp	Reciprocal hyper-distance-path index
HaR	Harary H index
CID	Randic ID number
<b>Pol</b>	Polarity number
EPS0	Edge connectivity index of order 0
<b>EEig02d</b>	Eigenvalue 02 from edge adjacency matrix weighted by dipole moments
<b>O-058</b>	Atom centered fragments corresponding to carbonyl
<b>O-060</b>	Atom centered fragment corresponding to Al–O–Ar/Ar–O–Ar–R–O–R/R–O–C=X
TPSA(NO)	Topological polar surface area using Nitrogen and oxygen polar contributions

<sup>a</sup>Descriptors considered in the final model are highlighted in Table, and were selected according to the regression vector score.

**Table 3.** Experimental and predicted activities ( $pIC_{50}$ ) with residual values for the 14 test set compounds.

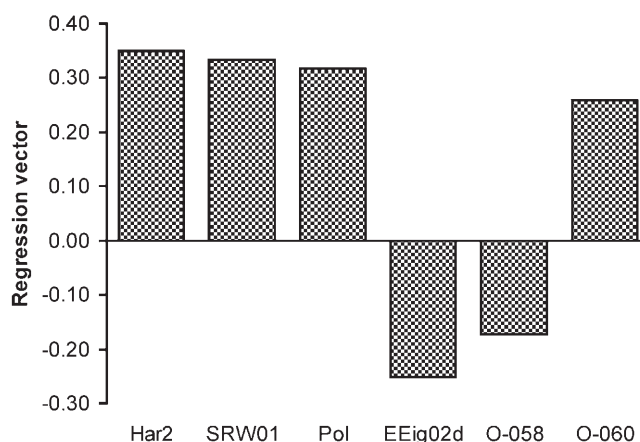
Compound	Experimental	Classical QSAR		HQ SAR	
		Predicted	Residual <sup>a</sup>	Predicted	Residual <sup>a</sup>
Test set					
57	2.62	2.51	0.11	2.63	-0.01
58	3.52	3.26	0.26	3.35	0.17
59	3.30	3.33	-0.03	3.32	-0.02
60	3.22	3.42	-0.20	3.27	-0.05
61	3.07	2.80	0.27	3.17	-0.10
62	3.82	3.72	0.10	3.71	0.11
63	5.30	4.88	0.42	5.31	-0.01
64	5.40	5.36	0.04	5.12	0.28
65	5.70	5.96	-0.26	5.75	-0.05
66	4.30	4.50	-0.20	4.91	-0.61
67	4.60	4.27	0.33	4.87	-0.27
68	4.00	4.27	-0.27	4.66	-0.66
69	4.74	4.43	0.31	4.72	0.02
70	5.00	5.12	-0.12	4.99	0.01

<sup>a</sup>The difference between experimental and predicted values.

Although cross-validation protocols give a suitable representation of the internal consistency and predictive power of QSAR models, they may lead to very optimistic statistics [27–29]. The predictive power of the best QSAR model derived using the 56 training set molecules (compounds 1–56, Table 1) was assessed by predicting the biological property for 14 test set compounds (57–70, Table 1). The results of the external validation process are listed in Table 3. The predicted values fall close to the experimental  $pIC_{50}$  values, not deviating by more than 0.42 log units. No outliers were detected in this series of GAPDH inhibitors. The good agreement between actual and predicted  $pIC_{50}$  values for the test set compounds suggests that the constructed model is reliable and has predictive power for new inhibitors within this structural class.

In addition to their ability to predict the biological property of interest, QSAR models should also provide useful chemical insights for drug design. In this context, Figure 2 illustrates the influence of the descriptors over the regression vector, which can be regarded as a weighted sum of the loadings included in the final model. This information was used to select the six most relevant descriptors during the development of the QSAR models, highlighted in Table 2.

It can be noted that Har2 and EEig2d represent, respectively, the most important positive and negative contributions to the final QSAR model. Descriptors with small coefficients did not significantly affect the QSAR model. Har2 [31, 32] is a topological descriptor derived from distance matrix which accounts for through-bond interactions in molecular systems. Due to its reciprocal nature, central atoms have larger influence than the external ones on the final summed value. This is clearly observed in our data set once the structurally related and nearly equipotent compounds 46, 51, and 53 show different Har2 values (192,

**Figure 2.** Descriptors score to regression vector.

201, and 207, respectively). This result indicates that structural modifications that do not significantly alter the main scaffold have minor effects on the inhibitory potency. Additional insight into the influence of molecular shape on the inhibitory potency comes from the interpretation of the EEig02 values [33, 34]. While the most potent inhibitors have the same EEig02d index (3.49), decreases in the inhibitory potency were observed for larger or smaller values. Although the classical QSAR analysis has extracted valuable SAR data, the topological descriptors (*e.g.*, Har2, EEig02d) provide only information about the character of the whole molecule. Accordingly, the information collected does not differentiate major and minor accounts from individual molecular fragments to the inhibitory potency. In the present work, in order to gain further insight into the fragment-based structure–activity relationships for this series of GAPDH inhibitors, and also for the better understanding of the synergies and challenges between these approaches, we have employed the HQ SAR method [25, 27].

### 3.3 HQ SAR Analysis

The specialized fragment fingerprint (called molecular hologram) consists of structural fragments of a molecule that are counted in bins of a fixed length array. The bins represent all of the unique fragments included within a particular molecule and are assigned by a cyclic redundancy check (CRC) algorithm. The bin occupancies of the molecular hologram are structural descriptors (independent variables) encoding compositional and topological molecular information. Accordingly, the hologram includes information on the quantity and type of each fragment in each molecule. HQ SAR encodes all possible molecular fragments (*i.e.*, linear, branched, and overlapping) without requiring any explicit 3D information of the ligands (*e.g.*, determination of 3D structure, putative binding conformations, and molecular alignment). The generation of molecular holograms was carried out using the fragment size de-

**Table 4.** HQSAR analysis for various fragment distinctions on the key statistical parameters using fragment size default (4–7).

Model	Fragment distinction	$q^2$	SDEP	$r^2$	SEE	HL	PC
1	A/B	0.831	0.439	0.885	0.355	257	3
2	A/B/C	0.835	0.436	0.888	0.350	353	3
3	A/B/C/H	0.836	0.429	0.884	0.362	307	3
4	A/B/C/H/Ch	0.835	0.432	0.887	0.357	353	3
5	A/B/C/H/Ch/DA	0.823	0.447	0.886	0.359	353	3
6	<b>A/B/H</b>	<b>0.851</b>	<b>0.409</b>	<b>0.895</b>	<b>0.343</b>	<b>151</b>	<b>3</b>
7	A/B/C/Ch	0.832	0.437	0.884	0.356	353	3
8	A/B/DA	0.834	0.435	0.894	0.345	97	3
9	A/B/C/DA	0.827	0.442	0.886	0.359	199	3
10	<b>A/B/H/DA</b>	<b>0.848</b>	<b>0.414</b>	<b>0.896</b>	<b>0.342</b>	<b>151</b>	<b>3</b>
11	A/B/C/H/DA	0.821	0.449	0.876	0.374	151	3
12	<b>A/B/H/Ch/DA</b>	<b>0.855</b>	<b>0.405</b>	<b>0.903</b>	<b>0.331</b>	<b>307</b>	<b>3</b>

Fragment distinction: A, atoms; B, bonds; C, connections; H, hydrogen atoms; Ch, chirality; DA, donor and acceptor.  $q^2$ , cross-validated correlation coefficient; SDEP, standard error of prediction;  $r^2$ , noncross-validated correlation coefficient; SEE, noncross-validated standard error; HL, hologram length; PC, optimal number of components.

fault (4–7), the 12 default series of hologram lengths (53–401), and the following combinations of fragment distinction: AB, ABC, ABCH, ABCHCh, ABCHChDA, ABH, ABCCh, ABDA, ABCDA, ABHDA, ABCHDA, and ABHChDA. The statistical results from the PLS analyses for the 56 training set inhibitors using several fragment distinction combinations are summarized in Table 4.

Statistical models of very good quality were obtained using different combinations of fragment distinction as revealed by their high  $q^2$  and  $r^2$  values. The best statistical results were obtained for models **6**, **10**, and **12** (Table 4), which were derived using A/B/H, A/B/H/DA, and A/B/H/Ch/DA, respectively. The influence of different fragment sizes, which control the minimum and maximum length of

fragments to be included in the hologram, was further investigated for the three best HQSAR models as shown in Table 5. As it can be seen, the variation of the fragment size led to the generation of slightly better HQSAR models (models **16**, **22**, and **28**, generated with the fragment size 5–8).

In order to investigate the statistical significance and stability of the models, a more consistent cross-validation test was carried out for the three best HQSAR models using 10 and 5 groups (LMO<sub>10</sub> and LMO<sub>5</sub>, respectively). Each model was evaluated 25 times by measuring its accuracy in predicting the activity of the remaining about 10 and 20% dataset compounds. The results of the LMO analyses are presented in Table 6, confirming the stability and robustness of the models generated with LOO.

The predictive ability of the models was assessed using the same test set compounds employed in the classical 2D QSAR studies (compounds **57–70**, Table 1). In order to select the best predictive statistical model, values of  $r^2_{\text{pred}}$  (predictive- $r^2$ ) were calculated for the three models **16**, **22**,

**Table 6.** Mean values of  $q^2$  and SDEP obtained after 25 independent runs of LMO cross-validation protocol.

Fragment distinction	Model	LMO <sub>10</sub> <sup>a</sup>		LMO <sub>5</sub> <sup>b</sup>	
		$q^2$	SDEP	$q^2$	SDEP
A/B/H	<b>16</b>	0.861	0.396	0.858	0.399
A/B/H/DA	<b>22</b>	0.839	0.425	0.840	0.424
A/B/H/Ch/DA	<b>28</b>	0.838	0.427	0.841	0.422

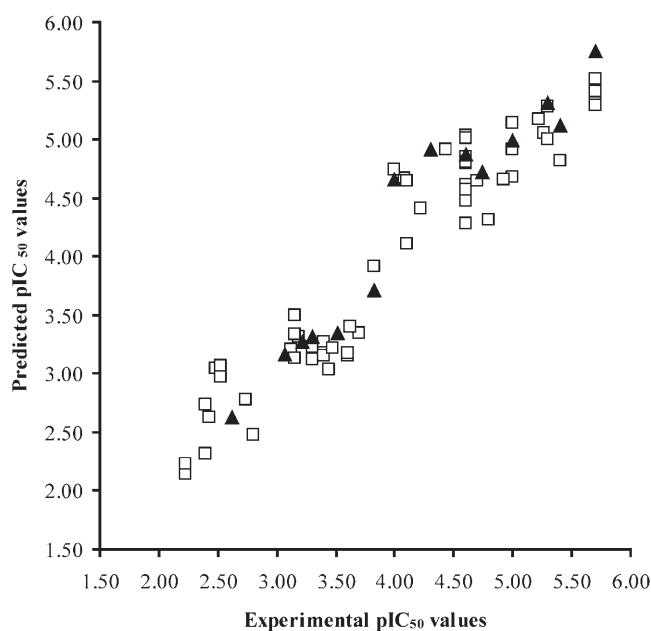
Fragment distinction: A, atoms; B, bonds; H, hydrogen atoms; Ch, chirality; DA, donor and acceptor.

a Mean value of  $q^2$  and SDEP obtained after 25 independent runs of cross-validation protocol with 10% of compounds left out.

b Mean value of  $q^2$  and SDEP obtained after 25 independent runs of cross-validation protocol with 20% of compounds left out.

**Table 5.** HQSAR analysis for the influence of different fragment sizes on the statistical parameters.

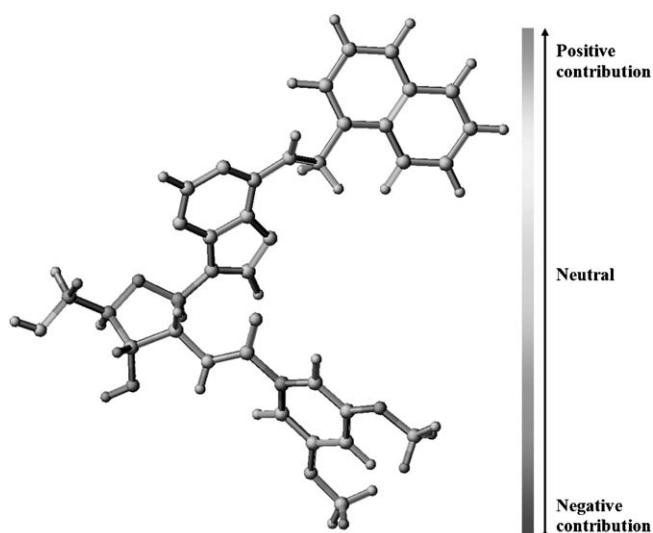
Fragment distinction	Model	Fragment size	$q^2$	SDEP	$r^2$	SEE	HL	PC
A/B/H	<b>13</b>	2–5	0.837	0.428	0.880	0.368	61	3
	<b>14</b>	3–6	0.845	0.418	0.888	0.355	257	3
	<b>15</b>	4–7	0.851	0.409	0.895	0.343	151	3
	<b>16</b>	<b>5–8</b>	<b>0.864</b>	<b>0.391</b>	<b>0.905</b>	<b>0.326</b>	<b>59</b>	<b>3</b>
	<b>17</b>	6–9	0.840	0.422	0.892	0.348	353	3
	<b>18</b>	7–10	0.835	0.431	0.883	0.362	353	3
A/B/H/DA	<b>19</b>	2–5	0.835	0.431	0.887	0.357	71	3
	<b>20</b>	3–6	0.837	0.428	0.889	0.354	257	3
	<b>21</b>	4–7	0.848	0.414	0.896	0.342	151	3
	<b>22</b>	<b>5–8</b>	<b>0.857</b>	<b>0.399</b>	<b>0.895</b>	<b>0.344</b>	<b>307</b>	<b>3</b>
	<b>23</b>	6–9	0.836	0.429	0.888	0.355	97	3
	<b>24</b>	7–10	0.838	0.427	0.884	0.360	257	3
A/B/H/Ch/DA	<b>25</b>	2–5	0.839	0.426	0.889	0.354	71	3
	<b>26</b>	3–6	0.843	0.420	0.894	0.345	97	3
	<b>27</b>	4–7	0.855	0.405	0.903	0.331	307	3
	<b>28</b>	<b>5–8</b>	<b>0.860</b>	<b>0.397</b>	<b>0.897</b>	<b>0.340</b>	<b>97</b>	<b>3</b>
	<b>29</b>	6–9	0.821	0.449	0.870	0.383	71	3
	<b>30</b>	7–10	0.808	0.465	0.858	0.400	71	3



**Figure 3.** Plot of predicted vs. experimental values of  $pIC_{50}$  for the training (open squares) and test (black triangle) set GAPDH inhibitors for the best HQSAR model.

and **28**, which presented  $r^2_{pred}$  of 0.92, 0.83, and 0.85, respectively. The results of the external validation for the best predictive model (**16**,  $r^2_{pred} = 0.92$ ) are shown in Table 3, and the graphic results for the experimental versus predicted for both training and test sets are shown in Figure 3. As can be seen, the test set compounds are well predicted within the range of values, not deviating by more than 0.66 log units. Therefore, the model presents good correlative and predictive abilities.

Besides predicting the property value of interest of untested molecules, HQSAR models should also provide hints about the relationships of different molecular fragments to biological activity [25, 27]. HQSAR models can be graphically represented in the form of contribution maps where the color of each molecular fragment reflects the contribution of an atom or a small number of atoms to the activity of the molecule under study. The colors at the red end of the spectrum (e.g., red, red–orange, and orange) reflect poor (or negative) contributions, while colors at the green end (e.g., yellow, green–blue, and green) reflect favorable (positive) contributions. Atoms with intermediate contributions are colored white. For example, the most important fragments of compound **36** (one of the most potent inhibitor of the data set) are shown in Figure 4. The HQSAR contribution map shows that the substituted benzamide ring has a favorable contribution to the inhibitory activity. This observation is in good agreement with previously reported results [14], and it is further supported by the fact that when the benzamide ring is replaced by the quinoline ring, biphenyl rings (**1**, **2**) or is ab-



**Figure 4.** HQSAR contribution map for the GAPDH inhibitor **36**.

sent (**8–18**), it causes a significant decrease in potency. Furthermore, the contribution maps of compounds (**19–56**) indicate that a different substitution pattern around the benzamide moiety is tolerable by the receptor (maps not shown) as previously described [17], and simultaneously, suggest a positive influence of *meta* substituents for the inhibitory potency. This can be seen, for instance, in the analysis of compound **36** ( $IC_{50} = 2 \mu M$ ) for which the incorporation of a methoxy substituent at the *meta*-position of the benzamide ring results in 50- and 40-fold increase in potency when compared to the *para*-substituted compounds **40** ( $IC_{50} = 100 \mu M$ ) and **41** ( $IC_{50} = 80 \mu M$ ), respectively.

HQSAR is an important drug design technique which encodes useful information that combines the fragment-based and topological description of molecular structures. Nevertheless, for a more robust QSAR analysis, it is important to incorporate additional detailed description of molecular properties that are relevant for activity (e.g., physicochemical parameters). For this reason, the integration of classical and HQSAR methods are complementary in nature, especially because vital components such as steric or electrostatic play a major role towards the inhibitory potency of series of compounds. In this context, it can be demonstrated that compound **36** has higher Har2 index (59.51) than compounds **40** (50.86) and **41** (52.81), whereas the EEig02d value is optimal for all the three compounds (3.49). These results are confirmed by the HQSAR contribution maps, and also suggest that *meta* substituted benzamide adenosine analogs have a steric component that is favorable for potency and hence show increased inhibitory potency, confirming previously reported findings [35].

## 4 Conclusion

The classical and hologram QSAR models described in this work exhibit both good internal and external consistency, with substantial predictive power. It is worth noting that we have employed the same combination of training and test sets for both QSAR analyses. The highly consistent results confirmed that investigations can be carried out concomitantly to search for synergies between 2D QSAR technologies, revealing a potential of shedding some light on the effects of the substitution pattern over molecular fragments within the data set compounds. Finally, as the models reported herein have been internally and externally validated employing different 2D QSAR approaches, they appear to be valuable tools for ligand-based enzyme inhibitor design.

## Acknowledgements

We gratefully acknowledge financial support from the State of São Paulo Research Foundation (FAPESP), the State of Bahia Research Foundation (FAPESB), and the National Council for Scientific and Technological Development (CNPq), Brazil.

## References

- [1] B. M. Watkins, *Trends Parasitol.* **2003**, *19*, 477–478.
- [2] B. D. Nicholson, J. D. Walley, D. S. Baguley, *Trop. Med. Int. Health* **2006**, *11*, 1339–1340.
- [3] J. H. McKerrow, C. Caffrey, B. Kelly, P. Loke, M. Sajid, *Annu. Rev. Pathol. Mech. Dis.* **2006**, *1*, 497–536.
- [4] S. L. Croft, G. H. Coombs, *Trends Parasitol.* **2003**, *19*, 502–508.
- [5] World Health Organization, *Tech. Rep. Ser.* **2002**, *905*, 1–109.
- [6] C. E. Sanz-Rodriguez, J. L. Concepcion, S. Pekerar, E. Oldfield, J. A. Urbina, *J. Biol. Chem.* **2007**, *282*, 12377–12387.
- [7] K. J. Kennedy, J. C. Bressi, M. H. Gelb, *Bioorg. Med. Chem. Lett.* **2001**, *11*, 95–98.
- [8] G. C. Guerra, L. Quijada, C. E. Clayton, *J. Cell Sci.* **2002**, *115*, 2651–2658.
- [9] F. M. Vellieux, J. Hajdu, C. L. M. J. Verlinde, H. Groendijk, R. J. Read, T. J. Greenhough, J. W. Campbell, K. H. Kalk, J. A. Littlechild, H. C. Watson, W. G. J. Hol, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 2355–2359.
- [10] D. H. F. Souza, R. C. Garratt, A. P. Araujo, B. G. Guimarães, W. D. Jesus, P. A. M. Michels, V. Hannaert, G. Oliva, *FEBS Lett.* **1998**, *424*, 131–135.
- [11] H. Kim, I. K. Feil, C. L. M. J. Verlinde, P. H. Petra, W. G. Hol, *Biochemistry* **1995**, *34*, 14975–14986.
- [12] J. L. Jenkins, J. J. Tanner, *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 290–301.
- [13] C. L. M. J. Verlinde, M. Callens, S. Van Calenbergh, A. Van Aerschot, P. Herdewijn, V. Hannaert, P. A. M. Michels, F. R. Opperdoes, W. G. Hol, *J. Med. Chem.* **1994**, *37*, 3605–3613.
- [14] S. Van Calenbergh, C. L. M. J. Verlinde, J. Soenens, A. De Bruyn, M. Callens, N. M. Blaton, O. M. Peeters, J. Rozen-skik, W. G. Hol, P. Herdewijn *J. Med. Chem.* **1995**, *38*, 3838–3849.
- [15] A. M. Aronov, C. L. M. J. Verlinde, W. G. Hol, M. H. Gelb, *J. Med. Chem.* **1998**, *41*, 4790–4799.
- [16] A. M. Aronov, S. Suresh, F. S. Buckner, W. C. Van Voorhis, C. L. M. J. Verlinde, F. R. Opperdoes, W. G. Hol, M. H. Gelb, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 4273–4278.
- [17] J. C. Bressi, C. L. M. J. Verlinde, A. M. Aronov, M. L. Shaw, S. S. Shin, L. N. Nguyen, S. Suresh, F. S. Buckner, W. C. Van Voorhis, I. D. Kuntz, W. G. Hol, M. H. Gelb, *J. Med. Chem.* **2001**, *44*, 2080–2093.
- [18] V. Farutin, L. Masterson, A. D. Andricopulo, J. Cheng, R. Riley, R. Hakimi, J. Frazer, E. H. Cordes, *J. Med. Chem.* **1999**, *42*, 2422–2431.
- [19] A. D. Andricopulo, R. A. Yunes, *Chem. Pharm. Bull.* **2001**, *49*, 10–17.
- [20] A. D. Andricopulo, A. Willian, R. Correa, A. R. S. Santos, R. J. Nunes, R. A. Yunes, V. Cechinel, *Pharmazie* **1998**, *53*, 493–498.
- [21] M. S. Castilho, F. Pavao, G. Oliva, S. Ladame, M. Willson, J. Perie, *Biochemistry* **2003**, *42*, 7143–7151.
- [22] S. Ladame, M. S. Castilho, C. H. Silva, C. Denier, V. Hannaert, J. Perie, G. Oliva, M. Willson, *Eur. J. Biochem.* **2003**, *270*, 4574–4586.
- [23] A. Leitao, A. D. Andricopulo, G. Oliva, M. T. Pupo, A. A. de Marchi, P. C. Vieira, M. F. G. F. da Silva, V. F. Ferreira, M. C. B. V. de Souza., M. M. Sa, V. R. S. Moraes, C. A. Montanari, *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2199–2204.
- [24] A. A. de Marchi, M. S. Castilho, P. G. Nascimento, F. C. Archanjo, G. del Ponte, G. Oliva, M. T. Pupo, *Bioorg. Med. Chem.* **2004**, *12*, 4823–4833.
- [25] M. S. Castilho, R. V. C. Guido, A. D. Andricopulo, *Bioorg. Med. Chem.* **2007**, *15*, 6242–6252.
- [26] M. S. Castilho, M. P. Postigo, C. B. de Paula, C. A. Montanari, G. Oliva, A. D. Andricopulo, *Bioorg. Med. Chem.* **2006**, *14*, 516–527.
- [27] K. M. Honorio, R. C. Garratt, A. D. Andricopulo, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3119–3125.
- [28] D. B. Oliveira, A. C. Gaudio, *Quant. Struct. Act. Relat.* **2003**, *19*, 599–601.
- [29] K. M. Honorio, R. C. Garratt, I. Polikarpov, A. D. Andricopulo, *J. Mol. Graph. Model.* **2007**, *25*, 921–927.
- [30] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
- [31] M. Randic, A. F. Kleiner, L. M. DeAlba, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277–286.
- [32] M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 292–299.
- [33] E. Estrada, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.
- [34] E. Estrada, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 701–707.
- [35] S. Suresh, J. C. Bressi, K. J. Kennedy, C. L. M. J. Verlinde, M. H. Gelb, W. G. Hol, *J. Mol. Biol.* **2001**, *309*, 423–435.