J Comput Aided Mol Des (2010) 24:157–172
DOI 10.1007/s10822-010-9323-y

ORIGINAL PAPER

3D-Pharmacophore mapping of thymidine-based inhibitors of TMPK as potential antituberculosis agents

Carolina Horta Andrade · Kerly F. M. Pasqualoto · Elizabeth I. Ferreira · Anton J. Hopfinger

Received: 6 October 2009 / Accepted: 22 February 2010 / Published online: 10 March 2010

Abstract Tuberculosis (TB) is the primary cause of mortality among infectious diseases. *Mycobacterium tuberculosis* monophosphate kinase (TMPKmt) is essential to DNA replication. Thus, this enzyme represents a promising target for developing new drugs against TB. In the present study, the receptor-independent, RI, 4D-QSAR method has been used to develop QSAR models and corresponding 3D-pharmacophores for a set of 81 thymidine analogues, and two corresponding subsets, reported as inhibitors of TMPKmt. The resulting optimized models are not only statistically significant with r^2 ranging from 0.83 to 0.92 and q^2 from 0.78 to 0.88, but also are robustly predictive based on test set predictions. The most and the least potent inhibitors in their respective postulated active conformations, derived from each of the models, were docked in the active site of the TMPKmt crystal structure. There is a solid consistency between the 3D-pharmacophore sites defined by the QSAR models and interactions with binding site residues. Moreover, the QSAR models provide insights regarding a probable mechanism of action of the analogues.

C. H. Andrade (✉) · K. F. M. Pasqualoto · E. I. Ferreira
Department of Pharmacy, Faculty of Pharmaceutical Sciences,
University of Sao Paulo, Av. Prof. Lineu Prestes, 580, Bloco 13,
Sao Paulo, SP 05586-900, Brazil
e-mail: carolhorta@usp.br; carolhandrade@gmail.com

C. H. Andrade · A. J. Hopfinger
College of Pharmacy, MSC09 5360, 1 University of New
Mexico, Albuquerque, NM 87131-0001, USA

A. J. Hopfinger
The Chem21 Group, Inc., 17870 Wilson Drive, Lake Forest,
IL 60045, USA

Keywords 4D-QSAR · 3D-Pharmacophore · Thymidine monophosphate kinase · TMPK inhibitors · Antituberculosis agents

Introduction

Tuberculosis (TB), second only to AIDS among infectious diseases, kills more than 2 million people a year worldwide [1]. The emergence of multiple-drug-resistant TB, coupled to its synergism with HIV, has stimulated the search for new intervention targets, and the development of new drugs to overcome this global threat [1–5].

Mycobacterium tuberculosis thymidine monophosphate kinase (TMPKmt) is one of the promising intervention targets in the treatment of tuberculosis [6]. TMPK is an enzyme that catalyses the conversion of deoxythymidine monophosphate (dTMP) to deoxythymidine diphosphate (dTDP) using ATP as a phosphoryl donor. This step lies at the junction of the de novo and salvage pathways of thymidine triphosphate (TTP) metabolism, and is the last specific enzyme for its synthesis [7].

The sequence of TMPKmt in comparison to its human isozyme shares only 22% sequence identity [8]. Recently, a X-ray crystal structure of TMPKmt in complex with dTMP was solved at a resolution of 1.95 Å (PDB entry code 1G3U) [8] allowing the structure-based design of TMPKmt inhibitors. Still, several TMPKmt inhibitors have been synthesized [9–12], but very few structure–activity relationship studies have been reported [13–15]. Overall, TMPKmt represents an attractive potential target for the rational design of new tuberculostatic drugs.

4D-QSAR analysis [16] has been used to develop 3D-pharmacophore models because of its capability of exploring large degrees of both conformational and alignment freedoms in the search for the active conformation and binding mode, respectively. In our previous 4D-QSAR study [15], we constructed 3D-pharmacophore models for a set of thirty-four 5′-thiourea-substituted α-thymidine analogues, reported as inhibitors of TMPKmt. The model identified new regions of the inhibitors that contain pharmacophore sites, such as the sugar-pyrimidine ring structure and the region of the 5′-arylthiourea moiety. In the present study, we report an extension of this investigation by doing a receptor-independent, RI, 4D-QSAR analysis of a larger and more chemically diverse set of TMPKmt inhibitors. Moreover, here we compare different subclasses of analog inhibitors, to map out subtle differences in binding pharmacophores for high potency inhibitors.

Although the crystal structure of the biomacromolecule target is available, the RI-4D-QSAR approach was selected in this study because of the uncertainty in the binding modes of the inhibitors. The hypothesized active conformations of the inhibitors resulting from RI-4D-QSAR analysis can be used as structural design templates, which include their deployment as the molecular geometries in structure-based ligand–receptor binding studies. In addition, the RI-4D-QSAR models can identify new regions of the inhibitors that contain pharmacophore sites which suggest insights into the mechanism of action of these inhibitors.

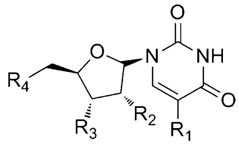
Overall, the two main objectives of this study have been to, (a) use the 4D-QSAR method to map the 3D-pharmacophore sites of the inhibitors developed by Van Calenbergh et al. [9–12], and, (b) to propose structural changes in these TMPKmt inhibitors that will make them more potent anti-tuberculosis agents.

Methods

TMPKmt inhibitors

The TMPKmt inhibitors used in this study have been gathered from a series of papers by Van Calenbergh et al. [9–12]. The experimental inhibitory activities were converted into their corresponding pK_i ($-\log K_i$) measures, where K_i is the equilibrium dissociation constant that quantitatively measures the affinity of the inhibitor to the enzyme. All the K_i values were obtained by the same assay method [17], and the pK_i values span a wide inhibitory range from 2 to 6. Tables 1, 2, 3 list the chemical structures, and the corresponding pK_i values, of the inhibitors divided into training and test sets. The inhibitors can also be characterized as three structural subclasses as follows:

Table 1 Definition of the chemical structures and pK_i values of compounds ddTMP1-ddTMP47



Cpd	R_1	R_2	R_3	R_4	pK_i
ddTMP1	CH_3	H	OH	$NHCOCH_3$	4.04
ddTMP2	CH_3	H	OH	N_3	5.15
ddTMP3	CH_3	H	OH	NH_2	4.92
ddTMP4	CH_3	OH	OH	OH	3.14
ddTMP5	CH_3	OH	OH	OH	3.62
ddTMP6*	CF_3	H	OH	OH	4.01
ddTMP7	C_2H_5	H	OH	OH	2.94
ddTMP8	F	OH	OH	OH	3.28
ddTMP9	F	OH	OH	H	3.25
ddTMP10	CH_3	H	NH_2	OH	3.64
ddTMP11	CH_3	H	F	OH	4.55
ddTMP12	CH_3	F	OH	OH	3.91
ddTMP13*	CH_3	OH	N_3	I	3.31
ddTMP14	CH_3	OH	NH_2	H	2.72
ddTMP15	CH_3	OH	$NHC(NH)NH_2$	OH	3.85
ddTMP16	CH_3	H	CH_2N_3	OPO_3^{2-}	4.92
ddTMP17	CH_3	H	CH_2NH_2	OPO_3^{2-}	4.98
ddTMP18	CH_3	H	CH_2F	OPO_3^{2-}	4.82
ddTMP19	CH_3	H	CH_2OH	OPO_3^{2-}	4.54
ddTMP20	CH_3	OH	CH_2N_3	OPO_3^{2-}	3.93
ddTMP21	CH_3	OH	CH_2NH_2	OPO_3^{2-}	3.50
ddTMP22	CH_3	H	CH_2N_3	OH	4.40
ddTMP23	CH_3	H	CH_2NH_2	OH	4.24
ddTMP24	CH_3	H	CH_2F	OH	4.35
ddTMP25	CH_3	H	CH_2OH	OH	4.39
ddTMP26	CH_3	OH	CH_2N_3	OH	3.11
ddTMP27	CH_3	OH	CH_2NH_2	OH	3.40
ddTMP28	CH_3	H	CH_2CH_2OH	OH	3.82
ddTMP29*	CH_3	H	NH_2	OPO_3^{2-}	3.63
ddTMP30	CH_3	OH	NH_2	OPO_3^{2-}	4.57
ddTMP31	CH_3	NH_2	OH	OPO_3^{2-}	4.26
ddTMP32	CH_3	Cl	OH	OPO_3^{2-}	4.72
ddTMP33*	CH_3	F	OH	OPO_3^{2-}	4.37
ddTMP34	$C_6H_5CH_2$	H	OH	OPO_3^{2-}	4.55
ddTMP35	CH_3	H	N_3	OPO_3^{2-}	5.00
ddTMP36*	CH_3	H	N_3	OH	4.55
ddTMP37	Br	H	N_3	OH	4.98
ddTMP38	Br	H	OH	OH	5.30
ddTMP39*	CH = CHBr	H	OH	OH	3.20
ddTMP40	CH2OH	H	OH	OH	3.09
ddTMP41	Cl	H	N_3	OH	4.79
ddTMP42*	CH_3	H	OH	OH	4.57

Table 1 continued

Cpd	R_1	R_2	R_3	R_4	pK_i
ddTMP43	H	H	OH	OH	2.99
ddTMP44	F	H	OH	OH	3.67
ddTMP45	I	H	OH	OH	4.48
ddTMP46	OH	H	OH	OH	3.57
ddTMP47	H	H	N_3	OH	3.09

* Test set compounds of Data Set I: dTMP6, dTMP13, dTMP29, dTMP33, dTMP36, dTMP39, dTMP42. In bold are highlighted the Data Set II compounds, and the test set compounds of Data Set II: dTMP6, dTMP13, dTMP36, dTMP39, dTMP42. *ddTMP* deoxythymidine monophosphate derivatives

Data Set I

Data Set I is composed of 47 thymidine analogues (compounds **ddTMP1–ddTMP47**, Table 1). The seven compounds with an asterisk (*) after their index number in Table 1 have been selected as test set compounds. The remaining 40 compounds form the training set.

Data Set II: non-phosphate (NP) Data Set

Thirty-four thymidine analogues that are highlighted in bold in Table 1, which do not have the phosphate group attached in R4 position, have been selected as a subset from Data Set I, and are called Data Set II or *non-phosphate* Data Set. The five compounds with an asterisk (*) in bold after their index number in Table 1 have been selected as test set compounds. The remaining 29 compounds form the training set of Data Set II.

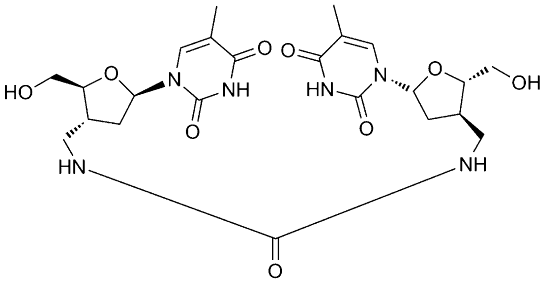
Overall Data Set

The Overall Data Set includes a total of 81 inhibitors (compounds **ddTMP1–ddTMP47**, Table 1; and compounds **ATT48–ATT81**, Tables 2, 3). The training set is composed by 70 compounds and the test set is composed by 11 compounds, as showed with an asterisk (*) after their index number in Tables 1, 2, 3.

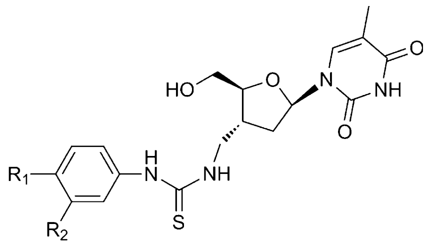
The generation of consistent statistical models depends, in part, upon the composition of the training and test sets. The compounds of the test sets were randomly selected subject to the constraints that, (a) the entire range of pK_i values was included, and, (b) a representative sampling of the chemical diversity of the Data Set was built into the test set. The best models were externally validated using compounds **dTMP6, dTMP13, dTMP29, dTMP33, dTMP36, dTMP39,** and **dTMP42** as test set compounds for Data set I; compounds **dTMP6, dTMP13, dTMP36, dTMP39,** and **dTMP42** as test set for Data set II; and compounds **dTMP6, dTMP13, dTMP29, dTMP33, dTMP36, dTMP39, dTMP42**, **ATT51**, **ATT57**, **ATT63**, and **ATT73** as the test set for the Overall Data Set.

All molecules in both the training and test sets were built in their neutral forms using the HyperChem 7.05

Table 2 Definition of the chemical structures and pK_i values of compounds ATT48–ATT55



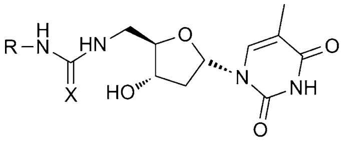
ATT48



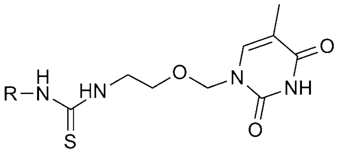
ATT49-ATT55

Cpd	R_1	R_2	pK_i
ATT48	–	–	4.43
ATT49	H	H	4.16
ATT50	Cl	H	4.68
ATT51*	OCH_3	H	4.34
ATT52	CH_3	H	4.44
ATT53	Cl	Cl	5.14
ATT54	OCH_2Ph	H	4.92
ATT55	Cl	CF_3	5.30

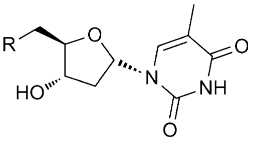
Table 3 Definition of the chemical structures and pK_i values of compounds ATT56–ATT81



ATT56-ATT75; ATT79



ATT77-ATT78



ATT76; ATT80-ATT81

Cpd	R	X	pK_i
ATT56	Ph	S	4.80
ATT57*	4-Cl-Ph	S	5.49
ATT58	4-MeO-Ph	S	5.00
ATT59	4-Me-Ph	S	5.11
ATT60	3,4-di-Cl-Ph	S	6.00
ATT61	3-CF_3-4-Cl-Ph	S	6.22
ATT62	4-Morpholinophenyl	S	4.72
ATT63*	1-Adamantyl	S	4.82
ATT64	3-Pyridil	S	4.74
ATT65	Fluoroesceinyl	S	5.27
ATT66	Phenylmethyl	S	5.27
ATT67	Benzhydryl	S	5.31
ATT68	Benzoyl	S	5.14
ATT69	3-CF_3-4-Cl-Phenylmethyl	S	5.58
ATT70	Phenylethyl	S	5.42
ATT71	3,4-di-Cl-Phenylethyl	S	5.66
ATT72	3,4-di-Cl-Ph	O	5.96
ATT73*	3-CF_3-4-Cl-Ph	O	5.72
ATT74[a]	3-CF_3-4-Cl-Ph	S	5.64
ATT75[b]	3-CF_3-4-Cl-Ph	S	5.42
ATT76	Benzamido	-	4.46
ATT77	Phenyl	S	3.58
ATT78	3-CF_3-4-Cl-Ph	S	4.43
ATT79	5′-Deoxy-*β*-D-adenosin-5′-yl	S	4.58
ATT80	N_3	-	4.58
ATT81	NH_2	-	4.80

[a] 3′-deoxyribonucleoside

[b] 3′-deoxy-2′-3′-didehydronucleoside

* The test set of the Overall Data Set comprises the compounds dTMP6, dTMP13, dTMP29, dTMP33, dTMP39, dTMP42, ATT51, ATT57, ATT63 and ATT73. *ddTMP* deoxythymidine monophosphate derivatives, *ATT* arylthiourea thymidine derivatives

software [18]. The crystallized structure of deoxythymidine monophosphate (dTMP) co-crystallized with TMPKmt (1G3U, resolution 1.95 Å) [8] was used as a reference structure in constructing the set of inhibitors. Each of the resultant initial 3D structures was subsequently energy-minimized and its partial atomic charges computed using the AM1 [16] semi-empirical method as implemented in the HyperChem program [18]. The energy-minimized structures were, in turn, used as the initial structures in each respective molecular dynamics simulation (MDS) employed to generate the conformational ensemble profile (CEP) of each inhibitor in the RI-4D-QSAR analyses described below.

The RI-4D-QSAR method

The operational formulation of the methodology for RI-4D-QSAR analysis, available in the 4D-QSAR software product, version 3.0 [20], consists of ten operational steps, which are only summarized here. The overall theory and formalism underlying 4D-QSAR can be found in Refs. [15, 16, 21, 22].

Basically, each molecule in a training set is placed in a grid cell lattice under some trial alignment. The ensemble of conformational states of each molecule are sampled, usually by doing molecular dynamics simulation, MDS, and the distribution of particular types of atoms occupying each of the grid cells of the lattice determined. The occupancy of each grid cell by each possible type of atom forms the basic trial descriptor pool in 4D-QSAR analysis. Additional descriptors, of any type, can be added to this trial descriptor pool. A QSAR model is then constructed by relating the key members of the trial descriptor pool to the observed activities of the training set.

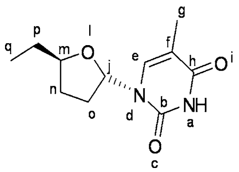
The 4D-QSAR methodology can be used in a receptor-dependent, RD, mode when the geometry of the receptor is available as is the case here. However, RD 4D-QSAR analysis requires a relatively large and chemically diverse training set, and also definitive information on binding alignment[s], in order to achieve a non-ambiguous QSAR model. Unfortunately, these requirements are not met for the present study and receptor independent, RI, 4D-QSAR analysis is carried out to maximize the extraction of structure–activity information. The benefits of doing the RI 4D-QSAR analysis performed as part of this study include;

1. providing a reliable predictive 3D-pharmacophore model for the limited range of substituent sites and substituent chemistry;
2. developing a rational basis of where substituent can, and cannot, be placed on the scaffold structures of the analogs;
3. the use of the 3D-pharmacophore model as a docking alignment for general ligand–receptor modeling including future RD-4D-QSAR studies; and
4. employing the 4D-QSAR model as an initial virtual screen for future studies that can be structure-based.

The molecules are divided into their "functional pieces", called interaction pharmacophore elements (IPEs). These IPEs correspond to the types of interactions that may occur upon ligand-receptor binding, and are used to ultimately aid in characterizing the 3D-pharmacophore of a RI-4D-QSAR model. Seven IPEs are currently employed and defined as follows: any type of atom (Any or A), nonpolar atom (NP), polar atom of positive partial charge (P+), polar atom of negative partial charge (P−), hydrogen bond acceptor (HA), hydrogen bond donor (HD) and atom in aromatic systems (Ar).

In this study, ten 3-ordered atom trial alignments were selected to systematically explore the common scaffold of the inhibitors. These ten trial alignments are listed in Table 4. Systematically spanning the scaffold using different trial alignments provides information regarding model sensitivity for a preferred alignment, and, likely, a corresponding binding mode, with respect to both the structure of the scaffold, as well as the substituents of the training set analogs.

Table 4 The set of trial alignments used in constructing the 4D-QSAR models



Align no.	First atom	Second atom	Third atom
1	a	b	d
2	i	d	m
3	a	c	l
4	g	j	p
5	l	m	p
6	a	g	q
7	a	i	q
8	a	d	q
9	a	c	q
10	a	l	q

Each conformational ensemble profile, CEP, of a given inhibitor was placed in a reference grid cell space according to the alignment under consideration. In this work, the selected size of the cubic grid cell was 1 Å, and the size of the overall grid cell lattice was chosen to enclose all inhibitors of the training set. The normalized absolute occupancy of each grid cell by each IPE atom type over the CEP for a given alignment forms a set of grid cell occupancy descriptors, or GCODs, which are unique to the training set. The GCODs were computed and used as the trial set of descriptors in this RI 4D-QSAR analysis [16].

An additional "traditional" QSAR descriptor was also calculated and added to the descriptor pool. Relative lipophilicity was considered as a trial descriptor by using ClogP values calculated for all inhibitors of the training set. The Ghose, Pritchett, and Crippen method [24] was used to compute the ClogP values using HyperChem 7.05 software [17]. ClogP along with the set of GCODs formed the complete trial descriptor pool used to build the RI 4D-QSAR models.

ClogP of a molecule is rigorously defined only for the molecule in its neutral and non-zwitterionic form. This, in turn, raises the question of what are the ionization states of many of the compounds, like those possessing an amine, in this Data Set at neutral pH of about 7 where the inhibition measurements were made. Moreover, the issue of properly selecting compound ionization state generally plagues QSAR analyses. One thing that needs to be kept in mind is that if a molecule is known to be ionized in bulk aqueous solvent at a given pH, that may not, effectively, be the case when the molecule is bound to a receptor in the same aqueous solvent at the same pH. The molecule may be effectively neutralized by taking up protons as part of the ligand–receptor binding process. Operationally, one way to deal with the uncertainty of ionization state in a QSAR analysis is to try different ionization states, build corresponding QSAR models and then let the quality of the best QSAR model decide what ionization states are most likely. That approach was begun in this study, but quickly concluded when the MDS found in sampling conformational states of ionized analogs were spurious, and markedly different from those of neutral analogs. Thus, all QSAR modeling in this study deals only with a neutral molecule which is considered the effective form of all members of the Data Set.

Partial least-squares (PLS) [25] regression was employed as a data reduction tool to identify the most highly weighted GCODs [and ClogP] from the entire set of trial descriptors for a training set. The 200 most highly weighted PLS GCODs were used to form the trial descriptor basis set for subsequent QSAR model fitting and optimization. Multiple linear regression and the genetic function approximation algorithm (GFA-MLR) [26, 27] was employed in combination for fitting and optimizing the RI 4D-QSAR models. A mutation frequency during the crossover optimization cycle was set to 10% for each new generation of RI-4D-QSAR models.

The ten best scored models found by GFA-MLR analysis [26, 27], according to their values of Friedman's LOF [28] measure, were tested for internal validation by the leave-one-out (LOO) cross-validation method available in the 4D-QSAR program [19]. In this study, the diagnostic measures used to characterize the RI-4D-QSAR models were: r^2 (linear correlation coefficient), q^2 (LOO cross-validation correlation coefficient), LSE (least squares error), LOF (Friedman's lack-of-fit score), residuals (difference between the observed and calculated pK_i values), and SD (standard deviation of the residuals). Compounds were considered outliers if the difference in the predicted and observed activities (the residual of fit) exceeded 2.0 standard deviations from the mean.

Results and discussion

Data Set I analysis

RI-4D-QSAR models were constructed for each of the ten trial alignments applied to Data Set I. The number of GCODs, statistical measures of fit and the number of outliers in each of the top 10 models are given in Table 5 for each of the selected alignments. Alignment 4 was judged to provide the best RI-4D-QSAR models because it yields models which collectively have the highest r^2 and q^2 values. Table 6 shows the top 10 RI-4D-QSAR models built from alignment 4. There are high values for both q^2 and r^2 for all ten models. However, models 1–3, 5–8 and 10 have at least one outlier and consequently were not considered further in the analysis. Model 4 was judged the best RI-4D-QSAR model and is highlighted in Table 6 in bold.

Table 5 Statistical measures of fit, number of GCODs and number of outliers for the top ten 4D-QSAR models for *Data Set I* using each trial alignment

Align	No. GCODs	r^2	q^2	Outliers	LSE
1	50–8	0.85–0.86	0.79–0.81	0–3	0.06–0.07
2	8–9	0.90–0.91	0.80–0.87	0–1	0.04–0.05
3	8–10	0.89–0.91	0.83–0.85	1–3	0.04–0.05
4	**8–10**	**0.92–0.93**	**0.85–0.88**	**0–2**	**0.03–0.04**
5	8–10	0.74–0.75	0.54–0.59	1–3	0.12–0.13
6	7–10	0.84–0.88	0.75–0.78	1–4	0.06–0.08
7	7–10	0.79–0.85	0.69–0.73	0–2	0.07–0.10
8	6–8	0.72–0.74	0.61–0.61	1–2	0.13–0.14
9	8–10	0.87–0.89	0.77–0.80	1–2	0.04–0.06
10	7–10	0.72–0.80	0.57–0.60	0–2	0.11–0.14

Table 6 Statistical measures of fit, number of GCODs and number of outliers for the top ten 4D-QSAR models for alignment 4

Models	No. GCODS	r^2	q^2	Outliers	LSE	LOF
1	10	0.93	0.88	1	0.03	0.09
2	10	0.93	0.88	1	0.03	0.10
3	9	0.93	0.88	1	0.03	0.09
4	**8**	**0.92**	**0.88**	**0**	**0.04**	**0.09**
5	10	0.93	0.87	1	0.03	0.10
6	9	0.93	0.87	1	0.03	0.09
7	10	0.93	0.86	2	0.03	0.10
8	9	0.92	0.86	1	0.03	0.10
9	8	0.92	0.85	0	0.04	0.09
10	10	0.93	0.85	2	0.03	0.10

A cross-correlation matrix of the residuals of fit between pairs of the top 10 RI- 4D-QSAR models from alignment 4 was constructed and is given in Table 7 [16, 20]. Table 7 was constructed to determine if the top 10 RI-4D-QSAR models are providing common, or distinct, structure–activity information. In other words, it is possible to identify the set of best and unique RI-4D-QSAR models. Pairs of models with highly correlated residuals of fit ($R^2 \approx 1$) are judged to be nearly the same model, while pairs of models with poorly correlated residuals ($R^2 < 0.5$) are considered to be distinct from one another. The cross-correlation values in Table 7 show that all of the models are highly correlated to one another. Thus, there is only one

Table 7 Linear cross-correlation matrix of the residuals of fit for the top ten 4D-QSAR models from model 4

Model	1	2	3	4	5	6	7	8	9	10
1	**1.00**									
2	0.98	**1.00**								
3	0.93	0.98	**1.00**							
4	0.89	0.88	0.89	**1.00**						
5	0.98	1.00	0.98	0.89	**1.00**					
6	1.00	0.98	0.94	0.90	0.98	**1.00**				
7	0.88	0.86	0.82	0.81	0.85	0.88	**1.00**			
8	0.90	0.87	0.87	0.92	0.88	0.91	0.79	**1.00**		
9	0.89	0.88	0.88	1.00	0.89	0.90	0.81	0.92	**1.00**	
10	0.98	1.00	0.98	0.87	1.00	0.98	0.85	0.87	0.87	**1.00**

unique RI-4D-QSAR model and it is best represented by model 4 highlighted in bold print.

Model 4 of Table 6, the best 4D-QSAR model, is expressly given by:

$$\begin{aligned} \mathrm{p}K_{\mathrm{i}} = & -3.03\,\mathrm{GC1(NP)} + 1.75\,\mathrm{GC2(Any)} \\ & - 8.48\,\mathrm{GC3(NP)} + 10.76\,\mathrm{GC4(HA)} \\ & + 8.18\,\mathrm{GC5(Any)} + 7.27\,\mathrm{GC6(Any)} \\ & + 3.74\,\mathrm{GC7(NP)} - 5.39\,\mathrm{GC8(NP)} + 3.31 \end{aligned} \quad (1)$$

$$N = 40;\ r^2 = 0.92;\ q^2 = 0.88$$

where pK_i is the molar TMPKmt inhibition constant and GCI (IPE) are the grid cell occupancy descriptors (GCODs), with IPE being the specific interaction of the GCOD. The statistical measures of model fit and significance are *N,* the number of sampled inhibitors, r^2 and q^2, the regression and leave-one-out cross-validation correlation coefficients, respectively.

Table 8 is the linear cross-correlation matrix for the GCODs in Eq. 1 [Model 4 of Table 6]. The linear cross-correlations in Table 8 are used to determine if the GCODs of the model are correlated to one another. As can be inferred from this table, three pairs GCODs are highly cross-correlated to one to another ($R > 0.9$) and their R values are given in bold print (see Table 8).

Table 8 Linear cross-correlation matrix of the GCODs for the optimal 4D-QSAR model (Eq. 1)

	GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC8
GC1	1.00							
GC2	−0.10	1.00						
GC3	0.32	−0.19	1.00					
GC4	−0.17	−0.08	−0.43	1.00				
GC5	0.30	−0.18	**0.96**	−0.46	1.00			
GC6	−0.09	−0.05	−0.13	−0.12	−0.06	1.00		
GC7	0.30	−0.19	**0.97**	−0.42	**0.96**	−0.12	1.00	
GC8	0.07	−0.26	0.30	0.39	0.35	0.49	0.33	1.00

Some cross-correlations of Table 8 are negative. One interpretation of such a negative relationship is that an increasing occupancy of one GCOD by the appropriate inhibitor atom type corresponds to a conformational change in the inhibitor that decreases the occupancy of the other GCOD. But it is important to note that the absolute values of correlation coefficients are all rather small other than for the three in bold.

In order to explore the roles of each of the three highly cross-correlated GCODs upon pK_i, each of these three GCODs was systematically excluded from the descriptor pool, one at a time, and 4D-QSAR models were rebuilt and optimized. Table 9 contains the resultant optimized 4D-QSAR models built with each of the descriptors GC3, GC5 and GC7 individually removed from the trial descriptor pool prior to model building. The deletion of any one of these highly correlated GCODs in Eq. 1 leads to marginal models, having lower values of q^2 and r^2, when compared with to Eq. 1, the best model. In addition, comparing to Eq. 1, the number of the descriptors increases in the optimized models generated without the availability of GC3, GC5, or GC7, and these models still have at least one outlier. Thus, each of the eight GCODs of the best RI-4D-QSAR model seems to provide independent and significant information of the variance in pK_i over the training set.

The postulated “bioactive” conformation of each inhibitor in the training set was constructed using model 4 (Eq. 1) from alignment 4 by first identifying all conformer states sampled for each inhibitor within a ΔE equal to 2 kcal/mol of the global minimum energy conformation of the CEP. It is important to remember that the ΔE equal to 2 kcal/mol involves only the ligand conformational energetics in the overall binding process. The GCODs of each resulting set of low-energy conformations were employed to predict the pK_i for each inhibitor using Eq. 1, and that particular conformer of the inhibitor with the highest predicted activity was selected as the “bioactive” conformation. A pictorial representation of the 3D-pharmacophore embedded in the 4D-QSAR model given by Eq. 1, is shown in Fig. 1. The reference compounds used in Fig. 1

Table 9 Statistical measures of fit, number of GCODs and number of outliers for the top ten 4D-QSAR models from alignment 4, rebuilt without the GCOD GC3, then GC5 and then GC7 in the trial descriptor pool

Model	No. GCODs	r^2	q^2	LSE	No. outliers
Without GC3	8–11	0.83–0.87	0.78–0.79	0.03–0.06	0–2
Without GC5	7–9	0.82–0.86	0.69–0.78	0.07–0.09	1–2
Without GC7	6–9	0.79–0.85	0.66–0.77	0.07–0.09	1–3

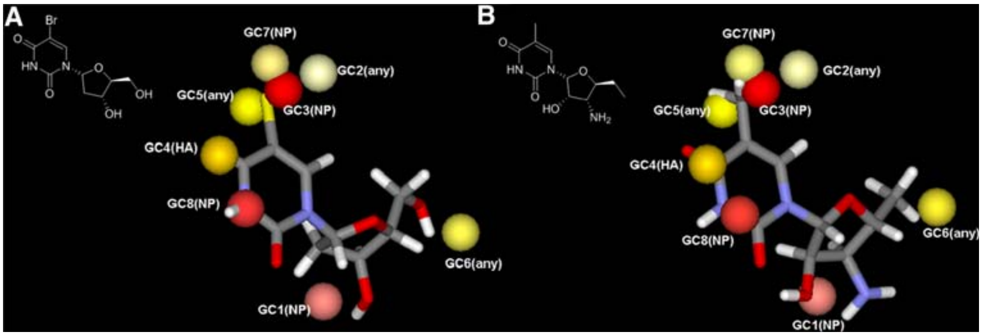


Fig. 1 Graphical representation of the predicted bioactive conformation found for the most active compound ddTMP38 (**A**), and the most inactive compound ddTMP14 (**B**), using Eq. 1 (Accelrys DS Visualizer [29]). Inhibition-enhancing and inhibition-diminishing grid cells are shown, respectively, as *yellow* and *red spheres*. The bioactive conformations are presented as stick models. Carbon atoms are in *gray*, oxygen in *red*, nitrogen in *blue*, bromine in *yellow*, and hydrogen atoms are in *white*

are the most active compound, **ddTMP38**, and the most inactive compound, **ddTMP14**, each in its postulated bioactive conformation.

In Fig. 1 the GCODs that increase pK_i are shown as yellow spheres (GC2, GC4, GC5, GC6 and GC7), and the GCODs that decrease the inhibition potency are shown as red spheres (GC1, GC3 and GC8). The color intensity of the GCODs (spheres) is proportional to the magnitude of the regression coefficients in Eq. 1. The larger the absolute value of the regression coefficient, the more intense is the color of its corresponding sphere, GCOD (see Fig. 1). The GCODs GC3, GC5, and GC7 which are quite close in space, are all highly cross-correlated GCODs to one another, and contribute to enhancing pK_i. Therefore, it is likely that these three GCODs are actually components of a common, single pharmacophore site.

The relative importance or weight of each GCOD in Eq. 1 is determined by the magnitude of the regression coefficients multiplied by the range of occupancy of the GCOD across the training set. The regression coefficient, range in occupancy and relative importance [weight] of each GCODs of Eq. 1 is given in Table 10. GC3 and GC4 descriptors are the two most highly weighted GCODs and they are located close to the R1 substituent and at C4 position of the pyrimidine ring, respectively.

Table 10 Relative measures of significance of the GCODs in the best 4D-QSAR model (Eq. 1)

GCOD	Regression coefficient	Range occupancy	Relative importance
GC1 (NP)	3.03	0.97	2.94
GC2 (Any)	1.75	0.71	1.24
GC3 (NP)	8.48	1.00	8.48
GC4 (HA)	10.76	1.00	10.76
GC5 (Any)	8.18	0.28	2.25
GC6 (Any)	7.27	1.00	7.27
GC7 (NP)	3.74	0.83	3.09
GC8 (NP)	5.39	1.00	5.39

To ascertain the predictive power of model 4, the pK_i value of each of the 7 test set inhibitors of Table 1 was calculated using Eq. 1. The GCOD values for the test set compounds are given in Table 11, and the pK_i predictions of the test set are given in Table 12. Six of the seven test set inhibitors had residuals whose absolute values (pK_i obs − pK_i pred = residual) were less than or equal to 0.52, which is the value of the standard deviation, SD, found for model 4 built from the training set. This finding indicates that model 4 has a substantial capacity for external prediction, approximately 85%. Test set compound **ddTMP29** was the only test set compound poorly predicted. The amino group at R3 on this inhibitor could be responsible for predicting this inhibitor to be much more active than observed. If the amino group was ionized, the corresponding CEP of **ddTMP29** would be quite different from that of its neutral form and, thusly, also its GCOD values. The resulting pK_i prediction would accordingly also be different from that of the neutral compound.

Data Set II: analysis of the non-phosphate containing inhibitors

It is reasonable to postulate that the "phosphate containing" compounds bind differently to the receptor than the non-phosphate containing inhibitors. In order to explore this possibility Data Set I was split up and a 4D-QSAR analysis was carried out for the non-phosphate containing inhibitors (Data Set II). RI-4D-QSAR models were constructed for each of the ten trial alignments with the goal of identifying the optimal alignment for Data Set II. Alignment 4 of Table 4 was found to be the best alignment for Data Set II in the same way as realized for Data Set I. The best RI-4D-QSAR model found for the Data Set II was the following:

Table 11 The GCOD coordinates and occupancy values for each test set inhibitor from *Data Set I*

Test set	GC1 (2, 0, 7)	GC2 (−3, 0, 0)	GC3 (−1, 0, 0)	GC4 (0, −3, 1)	GC5 (1, 0, 0)	GC6 1, 6, 7)	GC7 (−1, 0, −1)	GC8 (0, −3, 3)
ddTMP6	0.000	0.000	0.000	0.024	0.172	0.000	0.301	0.359
ddTMP13	0.252	0.000	0.229	0.024	0.320	0.000	0.485	0.373
ddTMP29	0.005	0.003	0.212	0.020	0.321	0.071	0.514	0.364
ddTMP33	0.000	0.000	0.223	0.022	0.289	0.051	0.426	0.326
ddTMP36	0.000	0.000	0.220	0.018	0.312	0.000	0.479	0.352
ddTMP39	0.002	0.000	0.000	0.001	0.001	0.000	0.001	0.016
ddTMP42	0.000	0.000	0.190	0.017	0.307	0.000	0.486	0.344

Table 12 Predicted pK_i using Eq. 1 for each test set inhibitor from *Data Set I*

Test set	pK_i Obs	pK_i Pred	ΔpK_i Residuals
ddTMP6	4.01	4.16	−0.15
ddTMP13	3.31	3.27	0.03
ddTMP29	3.63	4.82	**−1.09**
ddTMP33	4.37	4.23	0.14
ddTMP36	4.55	4.09	0.46
ddTMP39	3.20	3.25	−0.04
ddTMP42	4.57	4.36	0.21

$$
\begin{aligned}
\mathrm{p}K_i = & -3.07\,\mathrm{GC1(NP)} - 7.77\,\mathrm{GC2(Any)} - 5.46\,\mathrm{GC3(NP)} \\
& + 12.98\,\mathrm{GC4(Any)} - 1.99\,\mathrm{GC5(Any)} \\
& - 2.14\,\mathrm{GC6(NP)} + 5.30 \qquad (2)
\end{aligned}
$$

$N = 30$; $r^2 = 0.89$; $q^2 = 0.84$

Equation 2 derived from the non-phosphate Data Set is composed of only two types of IPEs: NP and Any. Moreover, only one descriptor, GC4, has a positive regression coefficient and, thus, can be responsible for an increase in potency. GC4 involves occupancy by *any* IPE type of atom. Figure 2 shows the graphical representation of the most active and most inactive compounds of the non-phosphate Data Set, **ddTMP38** and **ddTMP14**, in their respective bioactive conformations based upon Eq. 2.

Table 13 contains the linear cross-correlation matrix of the GCODs of Eq. 2, the best 4D-QSAR model of the non-phosphate Data Set. It is noteworthy that the same three pairs of GCODs highly cross-correlated to one another ($R > 0.9$) in Eq. 1 are also present in Eq. 2. These three GCODS are GC2, GC3, and GC4 (see Table 13), respectively. As seen in Fig. 2, also these three GCODs are close in space, surrounding the R1 substituent position.

Table 13 Linear cross-correlation matrix of the GCODs for the optimized non-phosphate 4D-QSAR model (Eq. 2)

	GC1	GC2	GC3	GC4	GC5	GC6
GC1	1.00					
GC2	−0.18	1.00				
GC3	−0.19	**0.85**	1.00			
GC4	−0.14	**0.89**	**0.96**	1.00		
GC5	0.12	−0.25	−0.17	−0.14	1.00	
GC6	−0.03	−0.30	−0.13	−0.08	0.49	1.00

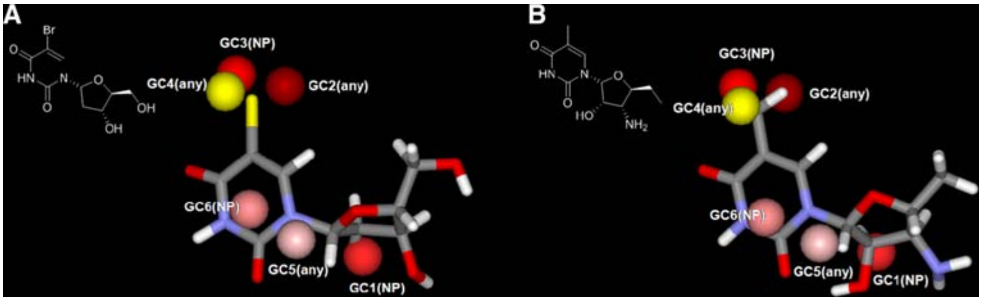


Fig. 2 Graphical representation of the predicted bioactive conformations found for the most active compound ddTMP38 (**A**), and the most inactive compound ddTMP14 (**B**), using Eq. 2 (Ayccelrys DS Visualizer [29]). Inhibition-enhancing and inhibition diminishing grid cells are shown, respectively, as *yellow* and *red spheres*. The bioactive conformations are presented as stick models. Carbon atoms are in *gray*, oxygen in *red*, nitrogen in *blue*, bromine in *yellow*, and hydrogen atoms are in *white*

Therefore, it is likely that these three GCODs are actually components of a common, single pharmacophore site.

In terms of specifying a 3D pharmacophore, the models numerically represented by Eqs. 1 and 2, are similar. In turn, this similarity suggests that, in general, non-polar atoms are detrimental to biological activity. In contrast, the model from Eq. 1 shows two additional descriptors that could increase the inhibition potency, namely GC4(HA), which is located near the oxygen atom of C4 pyrimidine ring, and GC6(Any), which includes the importance of the R4 substituent. Presumably GC4(HA) captures the significance of the oxygen atom of C4 pyrimidine ring in being able to form a ligand–receptor hydrogen bond. GC6 likely captures at least part of the interactions between the phosphate group on an inhibitor with the enzyme active site. Thus, the "phosphate containing" compounds are suggested to bind to the receptor in different fashion than the non-phosphate containing inhibitors, largely due to the interactions captured by GC6 position which represents, in part, the phosphate group.

Analysis of the Overall Data Set

As already described for the other Data Sets I and II, RI-4D-QSAR models were constructed and optimized for each of the 10 trial alignments listed in Table 4 for the Overall Data Set. Alignment 4 provided the best 4D-QSAR models for the 70 compounds of the training set as judged by the LOO cross-validation correlation coefficient. The number of GCODs, the r^2 and q^2 measures and the number of outliers of the top-ten models for the training set using alignment 4 are given in Table 14.

The cross-correlation coefficients of the residuals of fit for all pairs of top ten models were calculated. All pairs of the top ten 4D-QSAR models had residuals of fit which were highly correlated to one another ($R > 0.90$) (data not shown) indicating there is only a single RI-4D-QSAR model, which is best represented by the most significant fit of the data given by model 1 of Table 14. Model 1 is explicitly given by,

Table 14 Statistical measures of fit, number of GCODs and number of outliers for the top ten 4D-QSAR models from alignment 4 for the training set of the Overall Data Set ($N = 70$)

Model	No. GCODs	r^2	q^2	No. outliers	LSE
1	**11**	**0.83**	**0.78**	**0**	**0.10**
2	11	0.83	0.78	1	0.10
3	12	0.83	0.77	1	0.10
4	12	0.84	0.77	1	0.10
5	12	0.83	0.77	2	0.10
6	13	0.83	0.77	2	0.10
7	13	0.84	0.77	1	0.10
8	13	0.83	0.77	1	0.10
9	13	0.83	0.76	2	0.10
10	13	0.84	0.76	0	0.10

$$\begin{aligned} \mathrm{p}K_{\mathrm{i}} = & -7.29\,\mathrm{GC1(NP)} + 2.12\,\mathrm{GC2(Any)} - 1.63\,\mathrm{GC3(NP)} \\ & + 1.55\,\mathrm{GC4(Any)} - 1.91\mathrm{GC5(NP)} - 2.85\,\mathrm{GC6(NP)} \\ & - 0.75\,\mathrm{GC7(NP)} + 1.97\,\mathrm{GC8(Any)} + 0.18\,\mathrm{ClogP} \\ & + 0.04(\mathrm{ClogP})^2 + 4.66 \end{aligned} \quad (3)$$

$N = 70$; $r^2 = 0.83$; $q^2 = 0.78$

Equation 3 is composed of only two IPE types, namely NP and Any. Moreover, the best 4D-QSAR model includes in a parabolic dependence on ClogP, the descriptor which describes relative molecular lipophilicity. In Eq. 3, pK_{i} is minimized for ClogP $= -2.25$ and increases with increasingly positive ClogP values for the inhibitors. Thus, operationally, pK_{i} increases with ClogP for majority of inhibitors of this particular inhibitor Data Set according to Eq. 3.

Figure 3 is a plot of the ClogP values of the 81 inhibitors considered in this study partitioned across the three training sets. The average ClogP values across each of the respective three training sets are: −1.08, −1.46 and 0.05, for Data Set I, Data Set II and the Overall Data Set, respectively. As can be inferred from Fig. 3, the ClogP values for the Data Sets I and II are predominantly negative values, whereas the Overall Data Set is also composed of many inhibitors with positive ClogP values. This observation likely explains why ClogP is present only in Eq. 3, and why it contributes positively to the inhibitory potency of the ligands.

The predicted bioactive conformations for certain analogues, based upon Eq. 3 for the training set of the Overall Data Set, are shown in Fig. 4 along with the GCODs of Eq. 3. The linear cross-correlation matrix of the descriptors of Eq. 3 is given in Table 15. As can be inferred from Table 15, GC8 is reasonably correlated with GC1. As already demonstrated for the 4D-QSAR models of the other two Data Sets, multiple GCODs can define a single pharmacophore binding site. The correlated descriptors GC1 and GC8 (Table 15) can be seen in Fig. 4 to be close in space, and, therefore, likely define a single pharmacophore site.

From an inspection of Fig. 4, and Eq. 3, it is possible to gain a pharmacophoric interpretation of the GCODs in the model. The positive regression coefficients for GC2, GC4, and GC8 indicate that pK_{i} should increase with increasing appropriate inhibitor IPE occupancy, while the opposite is true for GC1, GC3, GC5, GC6 and GC7 which have negative regression coefficients. As already pointed out for the models of the two other Data Sets (Eqs. 1, 2), remarkably all GCODs responsible for a decrease in inhibition potency

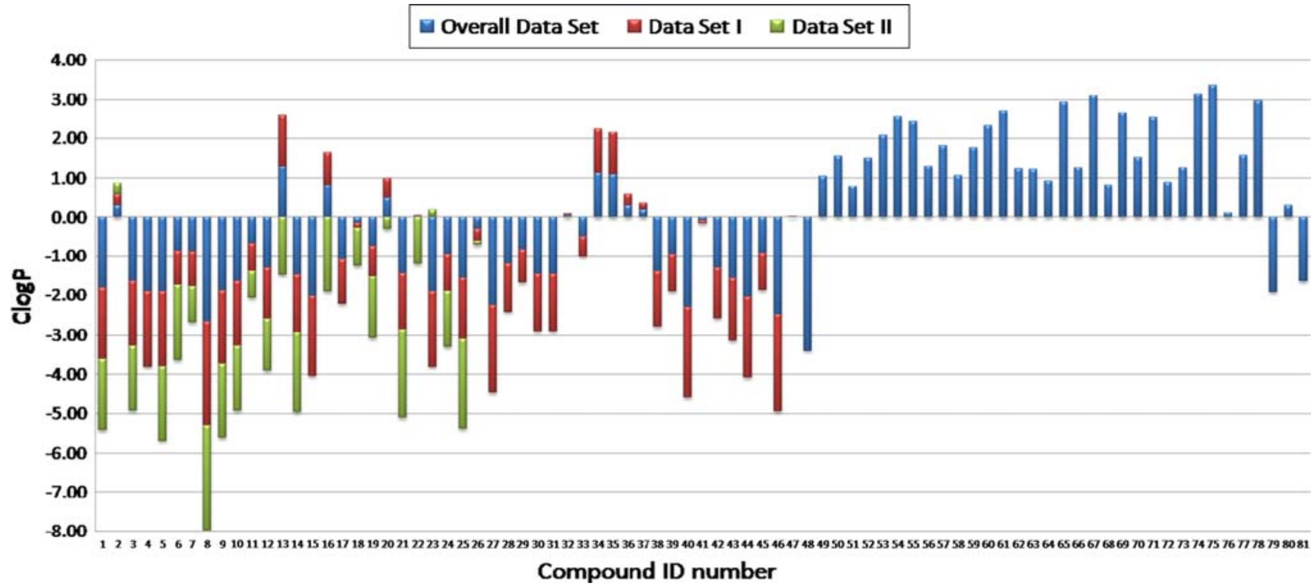


Fig. 3 A plot of the ClogP values of the 81 inhibitors of the study partitioned, by color-coding, across the three training sets

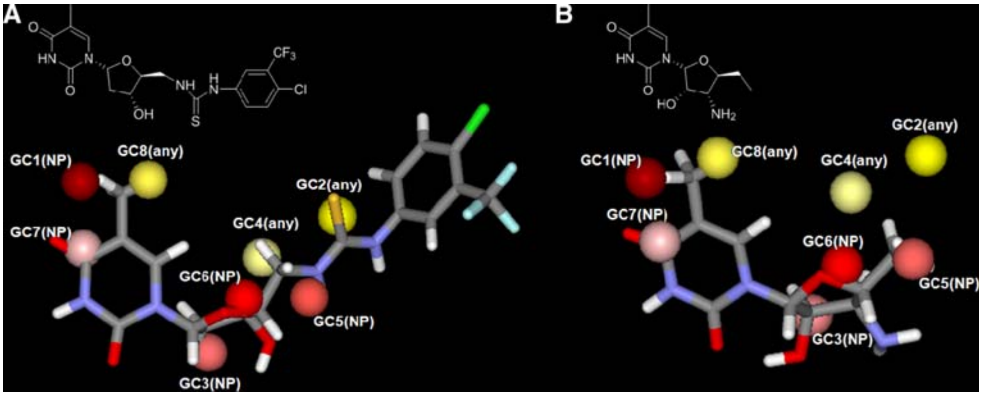


Fig. 4 Graphical representation of the predicted bioactive conformation found for the most active compound **ATT61** (**A**), and the most inactive compound **ddTMP14** (**B**) from the 70 compound training set, based upon Eq. 3 (Accelrys DS Visualizer [29]). Inhibition-enhancing and inhibition diminishing grid cells are shown, respectively, as *yellow* and *red spheres*. The bioactive conformations are presented as stick models. Carbon atoms are in *gray*, oxygen in *red*, nitrogen in *blue*, sulfur in *yellow*, chlorine in *green*, fluorine in *cyan*, and hydrogen atoms are in *white*

in Eq. 3 correspond to occupancy by nonpolar atom types. Therefore, it can be inferred that nonpolar atoms and/or groups of atoms of the inhibitors near these GCODs, as well as nonpolar substituents placed near these sites, are all detrimental to inhibition potency.

The test set predictions of the Overall Data Set are given in Table 16. Nine of the eleven test set inhibitors have residuals of fit to Eq. 3 whose absolute values (expressed as, [pK_i obs − pK_i pred = residual]) were less than or equal to 0.39 which is the standard deviation, SD value.

This finding indicates that model represented by Eq. 3 has a capacity of external prediction of approximately 82%, which is virtually identical to the variance explained by the model for the training set from which it was derived. Compound **ddTMP29** is one of the poorly predicted test set compounds. The amino group at R3 on this inhibitor may be responsible for predicting this inhibitor to be much more active than is observed. If the amino group was actually ionized, the corresponding CEP of **ddTMP29** would be quite different from that of its neutral form and, thusly, also its GCOD values. The resulting pK_i prediction would accordingly also be different from that of the neutral compound. No explanation can be put forth regarding why test set compound **ddTMP36**, is an outlier, but it is noted that it also is not a huge outlier.

Table 15 Linear cross-correlation matrix of the GCODs for the optimal 4D-QSAR model (Eq. 3) of the Overall Data Set

	GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC8	ClogP	$(\text{ClogP})^2$
GC1	1.00									
GC2	0.32	1.00								
GC3	−0.12	−0.14	1.00							
GC4	0.13	−0.19	−0.15	1.00						
GC5	0.09	−0.22	−0.01	0.38	1.00					
GC6	0.24	0.03	−0.27	0.29	0.24	1.00				
GC7	−0.07	−0.03	0.06	0.40	0.22	0.19	1.00			
GC8	**0.73**	0.33	−0.05	0.24	0.38	0.23	0.20	1.00		
ClogP	0.07	0.31	−0.38	0.30	0.16	0.11	−0.07	−0.15	1.00	
$(\text{ClogP})^2$	0.08	−0.01	−0.06	0.17	0.11	−0.07	0.06	−0.06	0.26	1.00

Table 16 The test set predictions for the Overall Data Set using Eq. 3

Test inhibitor	pK_i Obs	pK_i Pred	ΔpK_i Residuals
ddTMP6	4.01	4.25	−0.24
ddTMP13	3.31	3.29	0.02
ddTMP29	3.63	4.71	**−1.08**
ddTMP33	4.37	4.12	0.25
ddTMP36	4.55	4.15	**0.40**
ddTMP39	3.20	3.09	0.11
ddTMP42	4.57	4.66	−0.09
ATT51	4.33	4.22	0.11
ATT57	5.49	5.55	−0.06
ATT63	4.82	4.69	0.13
ATT73	5.72	5.43	0.29

Docking and coupled energy minimization and MDS

The predicted bioactive conformations for compounds **ddTMP38** and **ATT61**, the most potent members of the training sets, and for compound **ddTMP14**, the least potent inhibitor were adopted as representative trial bioactive conformations for the entire Data Set of 81 inhibitors. Each of these three inhibitors in its postulated bioactive conformations was docked into the active site of the crystal structure of TMPKmt (1G3U) [8]. The binding pose was correspondingly based upon alignment 4. MOLSIM 3.2 [23] was then used to perform coupled energy minimization and MDS relaxation studies on each of the docked complexes. The goal of these docking - relaxation studies were to elucidate the types of interactions occurring between the inhibitors and the surrounding amino acid residues lining the binding site. The results of these docking and relaxations studies are graphically illustrated in Fig. 5 for the most potent inhibitors of the training sets, namely;

1. **ddTMP38** using the predicted bioactive conformation from Eq. 1 (part A) and Eq. 2 (part B), and
2. **ATT61** using the predicted bioactive conformation from Eq. 3 (part C).

The least potent inhibitor of the training set, **ddTMP14**, is shown in part D of Fig. 5 using the predicted bioactive conformation from Eq. 3.

An inspection of Fig. 5 shows similarities among GCODs (pharmacophore sites) of these respective 3D-pharmacophores models. Remarkably, the highly cross-correlated GCODs are present in all three models. These highly cross-correlated GCODs are GC3, GC5, and GC7 (Eq. 1), which correspond to GC2, GC3, and GC4 (Eq. 2), which also correspond to GC1 and GC8 (Eq. 3). These GCODs, as discussed above, can be inferred from Fig. 5 to possibly define a single pharmacophore binding site. All of these descriptors are located near the C5 position of the thymine ring which corresponds to a binding site cavity formed by the residues Arg74, Phe36, Pro37, and Arg95. This cavity seemingly defines the volume, and possibly the shape, of the substituent at C5 of the inhibitor scaffold. Also, the orientation of the backbone carbonyl of Phe36 makes it available for an electrostatic interaction with a substituent at this position.

The descriptors defined as GC1 in both Eqs. 1 and 2 correspond to GC3 in Eq. 3, and appear able to be involved in interactions between the sugar motif of the inhibitor and residue Tyr103. These GCODs all correspond to occupancy by a nonpolar atom type (NP), and have negative regression coefficients in the three models. Thus, the 4D-QSAR models are suggesting that the occupation of the region defined by these GCODs by nonpolar atom/group IPE types is detrimental to inhibition potency. This loss in inhibition potency could be due to nonpolar groups being at locations to interfere with inhibitor-enzyme polar and/or hydrogen-bond interactions that stabilize the complex.

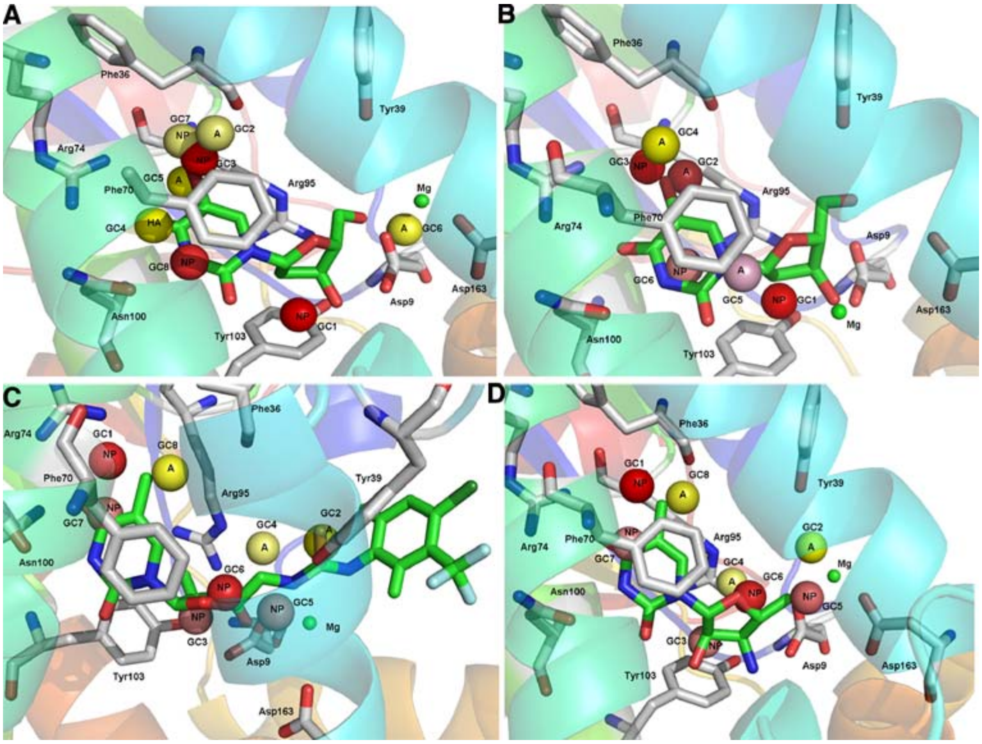


Fig. 5 The MDS relaxed conformations, starting with the RI-4D-QSAR postulated bioactive conformations, of the most potent inhibitors, **ddTMP38** from Eq. 1 (**A**) and from Eq. 2 (**B**), and **ATT61** from Eq. 3 (**C**), and the least potent inhibitor **ddTMP14** from Eq. 3 (**D**), each bound at the TMPKmt active site. The respective GCODs of each 4D-QSAR model are also shown in the active site of the crystal structure of TMPKmt. Only the main interacting residues in the pocket of the binding site are shown in these stick model representations. The carbon atoms are in *gray*, nitrogen in *blue*, oxygen in *red*, sulfur in *yellow*, fluorine in *cyan*, chlorine in *dark green*. The inhibitors **ddTMP38, ATT61** and **ddTMP14** are also presented as stick models, but the carbon atoms and bonds are in *green*. The GCODs are represented as spheres of 1 Å radius and its IPEs atom types are as follows: *A* any, *NP* nonpolar, *HA* hydrogen bond acceptor. The Pymol Viewer was used to make these renderings of the complexes [30]

GC8 in Eq. 1 and GC6 in Eq. 2 are also equivalent GCOD descriptors, and are both inhibition-potency-diminishing GCODs. These two GCODs also map a pharmacophore site that seem to capture the drawback of possibly having interfering nonpolar groups located in regions where polar or hydrogen-bond interactions can occur and stabilize the complex. GC8 in Eq. 1 and GC6 in Eq. 2 are both located near the nitrogen atom of the inhibitor pyrimidine ring. Thus, nonpolar IPE located near GC8 of Eq. 1, and near GC6 of Eq. 2, may compromise the ability of the nitrogen group to interact with the Asn100 residue of the enzyme, which is an essential interaction for the substrate dTMP in the active site of TMPKmt [8]. Moreover, the large size of a typical nonpolar group near these pharmacophore site descriptors could also lead to steric repulsions between the corresponding inhibitor and the local enzyme binding site wall.

In Eqs. 1 and 3, the descriptors GC4(HA) and GC7(NP), contribute positively and negatively, respectively, to the inhibitory potency of the inhibitors when occupied. The pharmacophore site defined by these descriptors seems to be nearly the same and located by the oxygen atom O4 of the pyrimidine ring (see Fig. 5a, c). As is clear in Table 10, GC4 of Eq. 1 can produce the largest increase in pK_i among the GCOD descriptors of Eq. 1 (its relative importance = 10.76). The key interaction involving O4 that can be postulated in the binding site of the docking complex is for one of two possible hydrogen bonds, or a combination of them, with the Arg74 residue. As pointed out by Sierra et al. [8], the X-ray structure of TMPKmt reveals that one of the main binding forces between dTMP and the enzyme is a hydrogen bond between O4 of thymine and the Arg74 side-chain, which results in a preference for thymine over cytosine. Thus, the presence in this part of an inhibitor by a hydrogen-bond acceptor inhibitor atom is favorable to inhibition potency. On the other hand, occupancy of this region of the inhibitor by nonpolar atoms is detrimental to binding potency owing to a loss in favorable hydrogen-bonding, and/or electrostatic interactions near the GC7 site and Arg74. The size of the nonpolar group near GC7 could also lead to steric repulsions between the corresponding inhibitor and enzyme binding site.

In summary, the pharmacophore sites defined by the GCODs of the models from Data Set I and II (Eqs. 1, 2) are very similar one to another, and are largely restricted to the region of the inhibitor between the pyrimidine and sugar rings. The main difference between the pharmacophores of these two models is the pharmacophore site of GC6 (Eq. 1), which is not present in Eq. 2, and probably captures at least part of the interactions between the phosphate group on an inhibitor with the active site. The interactions represented by GC6 enhance inhibition potency with increasing occupancy by the Any IPE type. This form of the GCOD most likely means that there are two, or more, specific IPE types that occupy GC6 and increase inhibition potency.

The RI-4D-QSAR model for the Overall Data Set (Eq. 3) distributes the GCODs more fully across the entire structure of the inhibitors than Eqs. 1 and 2. Eq. 3 provides a self-consistent 3D-pharmacophore explanation of inhibition potency as function of the three classes of chemical structures of the training set. Of the three models, only Eq. 3 contains ClogP, in quadratic form, as a descriptor. One interpretation of this finding is that to capture in a single 4D-QSAR model the inhibition potency across the Overall Data Set requires ClogP. Data Set I and the non-phosphate Data Set each, by themselves, do not exhibit sufficient variance in inhibition potency with inhibitor relative lipophilicity so as to have ClogP survive as a significant descriptor in each respective optimized RI-4D-QSAR model.

The descriptors GC2(Any), GC4(Any), GC5(NP) and GC6(NP) of the Overall Data Set model (Eq. 3) do not have any direct corresponding GCODs in the models given by Eqs. 1 and 2. The descriptors GC2 and GC4 of Eq. 3 are both inhibition-potency-enhancing GCODs. The IPE atom type of these both descriptors is Any, which means occupancy by any of the types of atoms sampled across training set I in building Eq. 3 increases pK_i. These two GCODs identify 3D-pharmacophore sites embedded in the 5′-arylthiourea moieties of the inhibitors. GC4 of Eq. 3 is positioned between the C5′ position of the sugar ring and the nitrogen of thiourea. The amino acids residues that can possibly interact with inhibitor atoms captured by this GCOD descriptor include Arg95, Asp9 and Tyr39. The non-specific, Any IPE type for GC4 may be explained through the participation of multiple IPE atoms types across the training set acting as hydrogen bond acceptors and donors, as well as polar + and polar − IPE types, with the amino acid residues of the active site near GC4. In addition, interactions from inhibitor substituents in the region corresponding to the location of GC2 of Eq. 3 with residues Tyr39, Asp9 and Asp163 (see Figs. 5c, d) may be a source of increasing inhibitory potency. Overall, molecular modification in this region is a good target strategy to develop better anti-tuberculosis agents, that is, more potent inhibitors of TMPKmt. Some examples of such modifications could be the replacement of the 5′-thiourea by carbamide, guanidine, propylguanidine or nitroguanidine functionalities in order to verify the importance of sulphur, as well as to explore the extent of steric tolerance at this position.

Increasing nonpolar occupancy of the descriptors GC5(NP) and GC6(NP) of Eq. 3 lead to a decrease of pK_i. GC6 is located near the oxygen atom of the sugar ring of the inhibitor scaffold. GC5 is positioned between the C5′ of the sugar ring and the nitrogen of thiourea moiety. The amino acids residues that may possibly interact with inhibitor atoms captured by this GCOD descriptor include Asp9 and Asp163. Moreover, the IPE atoms/groups which occupy GC5 could also interact with the partially solvated Mg^{2+} ion present near this position. Thus, nonpolar substituents at GC5 may displace water molecules that otherwise occupy the coordination sites of the metal atom and, thereby, decrease the inhibitory potency [14].

All GCODs of Eq. 3 that are responsible for decreasing inhibition potency have occupancy by the nonpolar IPE type. This study reveals a preponderance of nonpolar GCODs composing the RI-4D-QSAR models. Hence, it is reasonable to conclude that the variance in the inhibition potency across the training set of the different subclasses of TMPKmt inhibitors is largely governed by hydrophobic interactions. This interpretation of the current RI-4D-QSAR models is consistent with our previous RI-4D-QSAR investigation [15], showing that the 5′-arylthiourea moieties of these molecules are oriented toward the outside of the enzyme through a channel, which is surrounded by nonpolar and aromatic residues including Ala35, Phe36, Pro37, and Arg160. In addition, these results strongly suggest ClogP is needed for the Overall Data Set model, but not for the Data Set I and non-phosphate training set. This finding is in concordance with our previous results [15], showing that the presence of lipophilic substituents on the 5′-aryl moiety is an important inhibitory component of the dTMP analogues possibly representing an additional pharmacophore site responsible for the generally higher pK_i values of these inhibitors.

Conclusion

One of the two principal benefits of having done a 4D-QSAR analysis on this Data Set is in being able to identify and rank the relative importance of pharmacophore sites on the inhibitors that most influence the observed variance in inhibition potency. This, however, does not mean these pharmacophore sites contribute most to the absolute potency of the inhibitors, but rather only account for differences in potency among these inhibitors. The other major benefit of this study is in being able to define a receptor alignment, using the 3D-pharmacophore sites of the 4D-QSAR equations, which provide a rational basis for ligand-docking studies to be performed.

The drawbacks to doing this 4D-QSAR analysis are largely a consequence of the relatively limited size and structural diversity of the Data Set used in the study. 4D-QSAR studies require greater size and diversity in training sets than other QSAR methods, especially 2D-QSAR approaches. This is because so many 4D-QSAR descriptors are created, the grid cell occupancy descriptors, all of which need to be reasonably sampled to build up reliable conformational occupancy profiles. The lack of analogs that sample all of the spaces around the scaffold structure[s], also limits the generality of a 4D-QSAR model. The model can only make structure–activity inferences for those spaces sampled by, in this case, the limited number of substituents at a limited number of the possible substitution sites on the scaffolds.

The 4D-QSAR models of all three training sets, expressed by Eqs. 1–3, are largely composed of only Any and NP GCOD IPE types. Only Eq. 1 had a highly specific type of IPE, namely hydrogen bond acceptor (HA). The lack of more specific IPE types in the GCODs of the models is likely due to (a) inhibitor interactions with the large number of aliphatic and aromatic residues lining the active site which are captured by NP, and (b) the non-specific, Any IPE type representing participation of multiple types of atoms across an inhibitor training set, such as hydrogen bond acceptors and donors, as well as polar + and polar − IPE types, interacting with different groups of the amino acid residues of the active site. The preponderance of nonpolar GCODs in all of the optimized models is also in accordance with previous studies [12, 14, 15].

One other significant point to make is that the GCODs of the 4D-QSAR equations having negative regression coefficients indicate where inhibitor atoms/groups of the corresponding IPE type should NOT be located. That is, these GCOD terms in the 4D-QSAR models provide design constraints for maintaining inhibition potency while one is searching for substituents to enhance potency.

The specific intermolecular interactions that can be seen in the bound inhibitor complexes map nicely into the 3D-pharmacophore models generated by the three 4D-QSAR models and, in particular, Eq. 3. Thus, the bound inhibitor complexes, in conjunction with the 4D-QSAR models, in addition to supporting a basis for the limited number of IPE types found in the 4D-QSAR models, also present a self-consistent pharmacophore explanation of inhibition potency as function of the three classes of chemical structures of the training set.

Further validation, refinement and application of these RI-4D-QSAR models could be realized by doing complementary receptor-dependent, RD, 4D-QSAR analyses as well as using all 4D-QSAR models in virtual high-throughput screening (VHTS) in the search for new potent inhibitors.

Acknowledgments The authors are grateful to the CAPES Foundation, a federal scientific agency of Brazil, for scholarship support. This work was also funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 R21 GM075775. Information on Novel Preclinical Tools for Predictive ADME-Toxicology can be found at http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-023.html. Links to nine initiatives are found here http://nihroadmap.nih.gov/initiatives.asp. Resources of the Laboratory of Molecular Modeling and Design at UNM and The Chem21 Group, Inc. were used in performing this work.

References

1. Dye C, Floyd K, Uplekar M (2008) World Health Organization Document, WHO/HTM/TB/2008.393
2. Aziz MA, Wright A, Laszlo A, De Muynck A, Portaels F, Van Deun A, Wells C, Nunn P, Blanc L, Raviglione M (2006) Lancet 368:2142–2154
3. Dye C (2006) Lancet 367:938–940
4. Andrade CH, Salum LB, Pasqualoto KFM, Ferreira EI, Andricopulo AD (2008) Lett Drug Des Discov 5:377–387
5. Andrade CH, Pasqualoto KFM, Zaim MH, Ferreira EI (2008) Braz J Pharm Sci 44:167–179
6. Haouz A, Vanheusden V, Munier-Lehmann H, Froeyen M, Herdewijn P, Van Calenbergh S, Delarue M (2003) J Biol Chem 278:4963–4971
7. Munier-Lehmann H, Chafotte A, Pochet S, Labesse G (2001) Protein Sci 10:1195–1205
8. Li de la Sierra I, Munier-Lehmann H, Gilles AM, Bârzu O, Delarue M (2001) J Mol Biol 311:87–100
9. Vanheusden V, Munier-Lehmann H, Froeyen M, Busson R, Rozenski J, Herdewijn P, Van Calenbergh S (2004) J Med Chem 47:6187–6194
10. Vanheusden V, Munier-Lehmann H, Pochet S, Herdewijn P, Van Calenbergh S (2002) Bioorg Med Chem Lett 12:2695–2698
11. Vanheusden V, Van Rompaey P, Munier-Lehmann H, Pochet S, Herdewijn P, Van Calenbergh S (2003) Bioorg Med Chem Lett 13:3045–3048
12. Van Daele I, Munier-Lehmann H, Froeyen M, Balzarini J, Van Calenbergh S (2007) J Med Chem 50:5281–5292
13. Gopalakrishnan B, Aparna V, Jeevan J, Ravi M, Desiraju GR (2005) J Chem Inf Model 45:1101–1108
14. Aparna V, Jeevan J, Ravi M, Desiraju GR, Gopalakrishnan B (2006) Bioorg Med Chem Lett 16:1014–1020
15. Andrade CH, Pasqualoto KFM, Ferreira EI, Hopfinger AJ (2009) J Chem Inf Model 49:1070–1078
16. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) J Am Chem Soc 119:10509–10524
17. Blondin C, Serina L, Wiesmuller L, Gilles AM, Barzu (1994) Anal Biochem 220:219–222
18. HyperChem Program Release 7.05 for Windows (2005) Hybercube Inc. Gainesville, FL
19. Dewar MJSE, Zoebisch G, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
20. 4D-QSAR Package version 2.0 (1997) The Chem21 Group Inc. Lake Forest, IL
21. Pasqualoto KFM, Ferreira EI, Santos OAF, Hopfinger AJ (2004) J Med Chem 47:3755–3764

22. Romeiro NC, Albuquerque MG, Alencastro RB, Ravi M, Hopfinger AJ (2005) J Comput Aided Mol Des 19:385–400
23. Doherty DC (2001) MOLSIM Package version 3.2. The Chem21 Group Inc, Lake Forest, IL
24. Ghose AK, Pritchett A, Crippen GM (1988) J Comput Chem 9:80–90
25. Glen WG, Dunn WJ III, Scott DR (1989) Tetrahedron Comput Methodol 2:349–354
26. Rogers DG, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854–866
27. Dunn WJ III, Rogers D (1996) In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press, London
28. Friedman JH (1991) Ann Stat 19:1–141
29. Discovery Studio Visualizer version 2.0 (2007) Accelrys Software Inc. San Diego, CA. http://accelrys.com/
30. DeLano WL (2004) The Pymol Molecular Graphics System version 1.0. Delano Scientific LLC: Palo Alto, CA. http://www.pymol.org/