

## LQTA-QSAR: A New 4D-QSAR Methodology

João Paulo A. Martins, Euzébio G. Barbosa, Kerly F. M. Pasqualoto, and Márcia M. C. Ferreira\*

Laboratory for Theoretical and Applied Chemometrics, Department of Physical Chemistry, Institute of Chemistry, The State University of Campinas - UNICAMP, Campinas, SP 13084-971, POB 6154, Brazil

Received January 13, 2009

A novel 4D-QSAR approach which makes use of the molecular dynamics (MD) trajectories and topology information retrieved from the GROMACS package is presented in this study. This new methodology, named LQTA-QSAR (LQTA, *Laboratório de Quimiometria Teórica e Aplicada*), has a module (LQTAgrid) that calculates intermolecular interaction energies at each grid point considering probes and all aligned conformations resulting from MD simulations. These interaction energies are the independent variables or descriptors employed in a QSAR analysis. The comparison of the proposed methodology to other 4D-QSAR and CoMFA formalisms was performed using a set of forty-seven glycogen phosphorylase b inhibitors (data set 1) and a set of forty-four MAP p38 kinase inhibitors (data set 2). The QSAR models for both data sets were built using the ordered predictor selection (OPS) algorithm for variable selection. Model validation was carried out applying *y*-randomization and leave-*N*-out cross-validation in addition to the external validation. PLS models for data set 1 and 2 provided the following statistics:  $q^2 = 0.72$ ,  $r^2 = 0.81$  for 12 variables selected and 2 latent variables and  $q^2 = 0.82$ ,  $r^2 = 0.90$  for 10 variables selected and 5 latent variables, respectively. Visualization of the descriptors in 3D space was successfully interpreted from the chemical point of view, supporting the applicability of this new approach in rational drug design.

### INTRODUCTION

The quantitative structure–activity relationship (QSAR) is an important research field in theoretical medicinal chemistry, which deals with the prediction of the biological activities of new compounds using mathematical relationships based on structural, physicochemical, and conformational properties of previously tested potential agents. QSAR relationships are helpful in understanding and explaining the mechanism of drug action at the molecular level and allow the design and development of new compounds presenting desirable biological properties.<sup>1</sup>

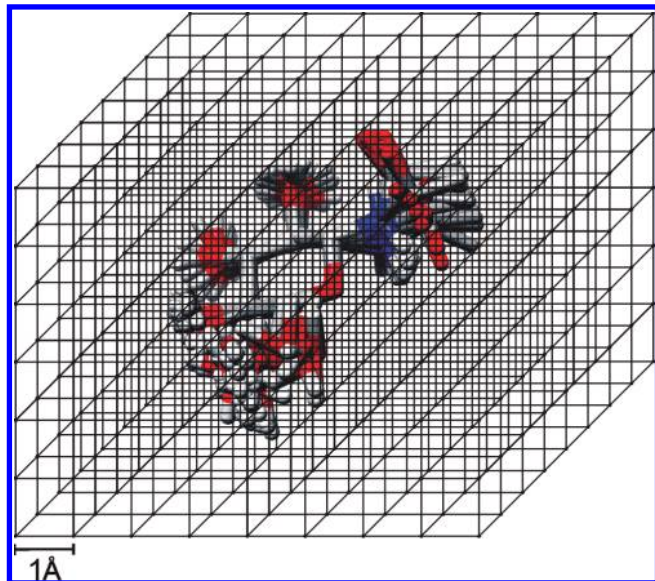
After Cramer and co-workers<sup>2</sup> had proposed the Comparative Molecular Field Analysis (CoMFA) in 1988, such methodology diffused quickly in medicinal chemistry and related fields, becoming a cornerstone for 3D-QSAR studies.<sup>3,4</sup> In CoMFA formalism, field descriptors or three-dimensional properties (electronic, steric, hydrophobic, and hydrogen bond) are determined in a 3D virtual lattice. The grid corresponds to a rigid hypothetical receptor and must be large enough to contain all aligned molecules. At each grid point, energies of interaction (descriptors) between a probe and all the atoms of each molecule of the investigated set are computed. In such an approach, partial least-squares (PLS) regression<sup>5–8</sup> is employed to model the relationships between the biological activity of a set of aligned compounds and their calculated 3D descriptors. The 3D-QSAR analysis using CoMFA can be divided, basically into three steps: alignment of the molecules, calculation of 3D descriptors, and mathematical modeling by PLS.<sup>9</sup>

The 4D-QSAR analysis, originally proposed by Hopfinger and co-workers in 1997,<sup>10</sup> incorporates conformational and

alignment freedom to the development of 3D-QSAR models by performing molecular state ensemble averaging, *i.e.*, the fourth “dimension”. In this approach, the descriptors values at each cell of the cubic grid are the occupancy measures for the atoms making up the molecules of the investigated set from the sampling of conformation and alignment spaces. The grid cell occupancy descriptors, GCODs, are generated for a number of different atom types (polar positive, polar negative, aromatic, hydrogen bond acceptor, hydrogen bond donor), called interaction pharmacophore elements, IPE. In a 4D-QSAR analysis each compound of the investigated set can be partitioned into classes (IPE), which are chosen regarding possible interactions with a common receptor. Thus, IPE are related to the descriptors’ nature in 4D-QSAR analysis, while GCOD are related to the coordinates of IPE mapped in a common grid. The idea underlying a 4D-QSAR analysis is that variations in biological responses are related to differences in the Boltzmann average spatial distribution of molecular shape with respect to the IPE.<sup>10</sup>

A new 4D-QSAR approach introduced in the present work and named LQTA-QSAR (LQTA, *Laboratório de Quimiometria Teórica e Aplicada*), is based on the generation of a conformational ensemble profile, CEP, for each compound instead of only one conformation, followed by the calculation of 3D descriptors for a set of compounds. This methodology explores jointly the main features of CoMFA and 4D-QSAR paradigms. LQTA-QSAR makes use of the GROMACS free package<sup>11</sup> to run the molecular dynamics, MD, simulations and estimate the CEP generated for each compound or ligand. The MD simulations can be performed considering explicit solvent molecules, which is a better approximation of the biological environment. The ordered predictor selection, OPS, algorithm,<sup>12</sup> recently developed by our research group,

\* Corresponding author e-mail: marcia@iqm.unicamp.br.



**Figure 1.** 3D virtual lattice or grid representation generated by the LQTAgrid module. The recommended distance between the CEP coordinates and the 3D lattice border is at least 5 Å. The grid distance between each adjacent point is 1 Å.

**Table 1.** Probes Available in the LQTAgrid Module

probes
COO <sup>-</sup> , C=O, NH <sub>3</sub> <sup>+</sup> , SH, CH <sub>3</sub> , NH <sub>2</sub> (arginine), C–H (aromatic), OH (H <sub>2</sub> O), OH, Zn <sup>2+</sup> , NH <sub>2</sub> (amide), Cl <sup>-</sup> , N–H (aromatic), Na <sup>+</sup>

was applied as the variable selection method in the construction of the PLS models. The LQTA-QSAR is available on the Internet at <http://lqta.iqm.unicamp.br>.

## METHODOLOGY

Prior to 4D-QSAR analysis, MD simulations of the molecules under study are carried out employing the GROMACS software. The coordinates in the GROMACS trajectory output files are stored in “gro” file format. Charges and atom types for calculating the Coulombic and van der Waals energies are retrieved from a gromos96 topology file (“top” or “itp”)<sup>13</sup> created at PRODRG server.<sup>14</sup> These two files are used as input files to the LQTAgrid module, which generates the 3D-interaction energy descriptors.

In the LQTAgrid program, the user can define the initial coordinates and the size of the 3D virtual lattice with defined grid, considering the coordinates from the “gro” files. It is recommended that one use a grid size sufficient to contain all conformers of the investigated set. A grid spacing of 1 Å is selected to generate several thousand points at the intersections of a regular 3D lattice (Figure 1).

Different types of atoms, ions, or functional groups, called probes (e.g., an NH<sub>3</sub> group positively charged; carbonyl and carboxyl groups; cations; and, anions), are used to compute the energy values for the interactions that the selected probe experiences in a respective position of the regular 3D lattice. The probes available in LQTAgrid are defined based on the ff43a1 force field parametrization<sup>11</sup> for atoms or molecular fragments, and they are presented in Table 1.

Each probe selected by the user runs over the grid, and the electrostatic and steric 3D properties are computed for

each individual grid point, based on the Coulombic (eq 1) and Lennard-Jones potential functions (eq 2), respectively.

$$E_{ele} = \frac{1}{n} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (1)$$

$$E_{vdW} = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^6} \rightarrow \begin{aligned} C_{ij}^{(12)} &= \left( \frac{1}{n} C_{ii}^{(12)} \times C_{jj}^{(12)} \right)^{1/2} \\ C_{ij}^{(6)} &= \left( \frac{1}{n} C_{ii}^{(6)} \times C_{jj}^{(6)} \right)^{1/2} \end{aligned} \quad (2)$$

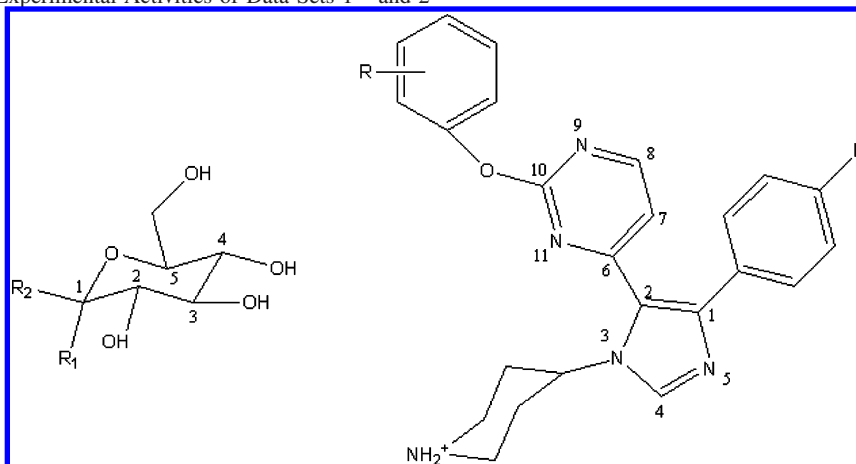
where  $q_i$  is the charge of the  $i$ th probe;  $q_j$  is the charge of the  $j$ th atom from CEP;  $\epsilon_0$  is vacuum permittivity;  $C_{ii}^{(12)}$ ,  $C_{ii}^{(6)}$ , and  $C_{jj}^{(6)}$  are parameters adapted from the ffG43a1 Gromos force field<sup>11</sup> for probes and atoms in CEP, respectively;  $n$  indicates the number of frames aligned in CEP; and  $r_{ij}$  represents the distances between the  $i$ th probe and the  $j$ th atom of CEP. Note that in both equations the energies are divided by  $n$  in order to take an average of the energies calculated for all copies of the ligands (CEP) in each grid point.

The output of a LQTAgrid analysis is a matrix whose columns contain the descriptors, which are the energies calculated for each grid point (according to eqs 1 and 2), and the rows represent the molecules of the investigated set. This matrix is used in a multivariate regression, e.g., multiple linear regression (MLR), principal components regression (PCR), or PLS regression, with the biological activity as the dependent variable, to construct the QSAR model.

**Data Sets Investigated - Comparison of Methodologies.** Considering that the approach presented in this study intends to incorporate the main advantages of 4D-QSAR and CoMFA methodologies, two reported data sets applying *receptor-independent* (RI) 4D-QSAR (set 1)<sup>15</sup> and CoMFA (set 2)<sup>16</sup> formalisms were used to evaluate the LQTA-QSAR paradigm. Set 1 consisted of forty-seven glycogen phosphorylase b inhibitors. Glycogen phosphorylase may help shift the balance between glycogen synthesis and its degradation favoring the glycogen synthesis in both muscle and liver, and such inhibitors may be useful therapeutic agents for the treatment of diabetes. Thus, glucose analogue inhibitors of glycogen phosphorylase may be of clinical interest in the regulation of glycogen metabolism in diabetes. The biological activities were expressed as the free binding energies ( $\Delta G$ , kcal/mol)<sup>15</sup> calculated from the inhibitory binding constant ( $K_i$ , mM) values employing eq 3 where  $T$  is temperature and  $R$  is the gas constant.

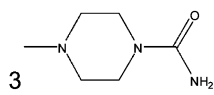
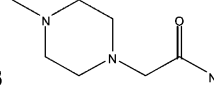
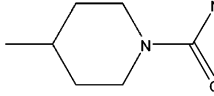
$$\Delta G = -RT \ln K_i \quad (3)$$

Set 2 was composed of forty-four p38 kinase inhibitors. p38 Kinase plays a vital role in inflammation mediated by tumor necrosis factor- $\alpha$  (TNF $\alpha$ ) and interleukin-1 $\beta$  (IL-1 $\beta$ ) pathways, and inhibitors of p38 kinase provide an effective approach for the treatment of inflammatory diseases. Pyridinyl and pyrimidinyl imidazoles, selectively inhibit p38a MAP kinase, are useful in the treatment of inflammatory diseases like rheumatoid arthritis. The biological activities were expressed as pIC<sub>50</sub>.<sup>16</sup> Seven compounds from set 1 [3, 8, 11, 13, 20, 30, and 38] and seven from set 2 [4, 10, 13, 17, 23, 30, and 38] formed the external validation sets and, subsequently, were used to test the predictability of the selected QSAR model. The structures and biological responses of the two data sets are presented in Table 2.

**Table 2.** Structures and Experimental Activities of Data Sets 1<sup>15</sup> and 2<sup>16a</sup>

data set 1			data set 2		
	$R_1$	$R_2$	$\Delta G$	$R$	$pIC_{50}$
1	H	NHC(=O)CH <sub>3</sub>	6.23	2,4 CH <sub>3</sub>	8.22
2	H	NHC(=O)CH <sub>2</sub> CH <sub>3</sub>	6.11	2 HO	8.18
3	H	NHC(=O)CH <sub>2</sub> Br	6.04	4 CH <sub>3</sub> CH <sub>2</sub>	8.13
4	H	NHC(=O)CH <sub>2</sub> Cl	6.03	2,5 CH <sub>3</sub>	7.97
5	H	NHC(=O)C <sub>6</sub> H <sub>5</sub>	5.67	2 F	7.89
6	H	NHC(=O)CH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	5.58	4 HO	7.85
7	H	NHC(=O)NH <sub>2</sub>	5.34	2 CH <sub>3</sub>	7.82
8	H	C(=O)NHCH <sub>3</sub>	5.26	2,3 CH <sub>3</sub>	7.82
9	H	NHC(=O)CH <sub>2</sub> NH <sub>2</sub>	4.76	4 CH <sub>3</sub>	7.72
10	C(=O)NH <sub>2</sub>	H	4.76	3 CH <sub>3</sub> O	7.70
11	H	C(=O)NH <sub>2</sub>	4.65	2,4 CH <sub>3</sub>	7.66
12	H	C(=O)NHNH <sub>2</sub>	4.17	3,4 -OCH <sub>2</sub> O-	7.64
13	H	SH	4.16	3 CH <sub>3</sub> O	7.60
14	CH <sub>2</sub> OH	H	3.92	2,6 CH <sub>3</sub>	7.46
15	OH	H	3.84	4 (CH <sub>3</sub> ) <sub>2</sub> CH	7.39
16	H	C(=O)NHC <sub>6</sub> H <sub>5</sub>	3.14	2,5 CH <sub>3</sub>	7.39
17	H	OH	2.95	3 NH <sub>2</sub> C=O	7.25
18	H	CH <sub>2</sub> CN	2.84	4 C <sub>6</sub> H <sub>5</sub>	7.22
19	OH	CH <sub>2</sub> OH	2.50	3 CH <sub>3</sub> NHC=O	7.12
20	H	OCH <sub>3</sub>	2.23	4 (CH <sub>3</sub> ) <sub>3</sub> C	7.10
21	CH <sub>2</sub> NH <sub>2</sub>	H	2.03	4 COOH	7.09
22	C(=O)NHCH <sub>3</sub>	H	1.99	4 CH <sub>3</sub> CH <sub>2</sub> O(C=O)	7.07
23	CH <sub>3</sub>	H	1.77	4 NH <sub>2</sub> C=O	7.05
24	C(=O)NH <sub>2</sub>	NHCOOCH <sub>3</sub>	6.65	3 F	7.02
25	H	NHCOOCH <sub>2</sub> Ph	4.79	4 Cl	6.94
26	H	NHC(=O)CH <sub>2</sub> NHCOCH <sub>3</sub>	4.17	4 CH <sub>3</sub> C=O	6.92
27	H	C(=O)NHNHCH <sub>3</sub>	3.81	4 C <sub>6</sub> H <sub>5</sub> O	6.89
28	OH*	H	3.74	3 CF <sub>3</sub>	6.88
29	H	C(=O)NHCH <sub>2</sub> CH <sub>2</sub> OH	3.58	2 CH <sub>3</sub> C=ONH	6.87
30	H	COOCH <sub>3</sub>	3.54	2 CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> C=ONH	6.82
31	C(=O)NHNH <sub>2</sub>	H	3.50	3,4 Cl	6.78
32	H	SCH <sub>2</sub> C(=O)NHPh	3.39	4 CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> C=O	6.76
33	H	C(=O)NH-4-OHPh	3.27	4 CN	6.75
34	H	CH <sub>2</sub> CH <sub>2</sub> NH <sub>2</sub>	3.25	3 (CH <sub>3</sub> ) <sub>2</sub> CHNHC=O	6.72
35	C(=O)NH-4-OHPh	H	3.12	3,4 F	6.68
36	OH	CH <sub>2</sub> N <sub>3</sub>	2.95	4 CF <sub>3</sub>	6.67

Table 2. Continued

data set 1			data set 2		
	$R_1$	$R_2$	$\Delta G$	$R$	$pIC_{50}$
37	OH	CH <sub>2</sub> CN	2.94	4 F	6.52
38	H	C(=O)NHCH <sub>2</sub> CF <sub>3</sub>	2.90	3 (CH <sub>3</sub> ) <sub>2</sub> NC=O	6.51
39	C(=O)NHPh	H	2.63	4 C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub> O	6.50
40	COOH	H	2.52		6.39
41	H	CH <sub>2</sub> NH <sub>2</sub>	2.46		6.34
42	C(=O)NHCH <sub>2</sub> CH <sub>2</sub> OH	H	2.46		6.22
43	H	SCH <sub>2</sub> C(=O)NH-2.4-F <sub>2</sub> Ph	2.39	4 CH <sub>3</sub> SO <sub>2</sub>	6.19
44	H	SCH <sub>2</sub> C(=O)NH <sub>2</sub>	2.32	H	7.72
45	CH <sub>2</sub> N <sub>3</sub>	H	2.29		
46	COOCH <sub>3</sub>	H	2.24		
47	C(=O)NHCH <sub>2</sub> -2.4-F <sub>2</sub> Ph	H	2.17		

<sup>a</sup> The atoms numbered were used for aligning the CEPs of all ligands. The  $\Delta G$  values are expressed in kcal/mol.

In this study, the starting geometries used to build the 3D models of each ligand were retrieved from the Brookhaven Protein Data Bank, and the entry codes are the following: 2gpb (2.20 Å resolution)<sup>17</sup> and 1bl7 (2.50 Å resolution),<sup>18</sup> for data sets 1 and 2, respectively. Although the 3D structures of the biomacromolecules were available, they were not considered in the construction of the QSAR models because the approach applied here is *RI* 4D-QSAR as already mentioned. The 3D models of all ligands [data set 1 and 2] were energy-minimized applying the DFT/B3LYP level<sup>19</sup> using the cc-pVDZ basis set (Gaussian' 03 program).<sup>20</sup> The electrostatic partial atomic charges (CHELPG)<sup>21</sup> were used in the calculation of the Coulombic interaction energy descriptors by the LQTAgrid program. The energy-optimized structures were submitted to the PRODRG<sup>13</sup> server for generating the GROMACS topology and Cartesian coordinate formats. Gasteiger<sup>22</sup> partial atomic charges schemes, calculated by AutoDockTool<sup>23</sup> and adapted for the ffG43a1 force field, were used for performing MD simulations.

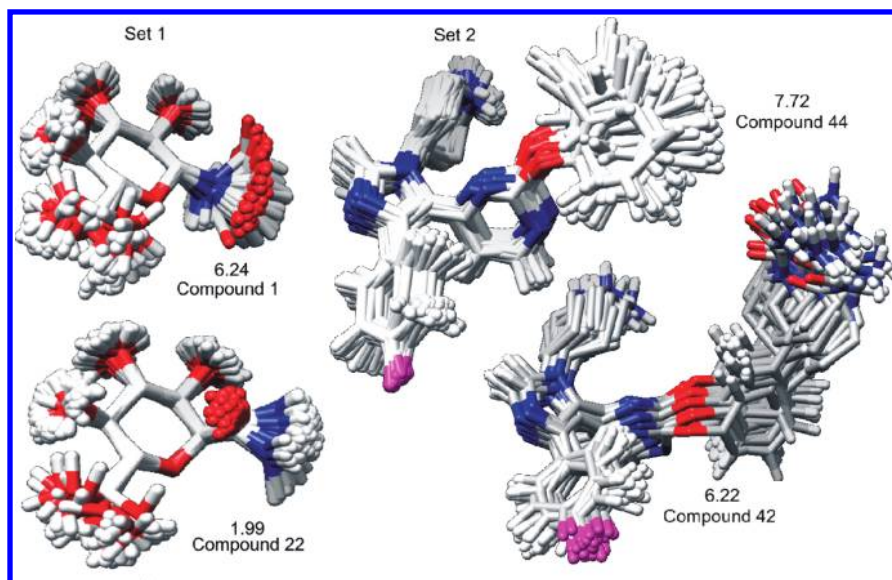
**Molecular Dynamics Simulations.** The MD simulations of all unbound ligands considering an explicit aqueous medium (extended single point charge (SPC/E)<sup>24</sup> water models) were carried out. Counterions were added to satisfy the electroneutrality condition, when necessary. Periodic boundary cubic boxes were built large enough, with the distance of 10 Å between the solute (ligands models) and water solvent molecules. The Particle Mesh Ewald (PME)<sup>25</sup> method was used for computing long-range electrostatic and van der Waals interaction energies, with a cutoff radius of 10 Å. All chemical bonds were constrained to their nominal values using the linear constraint solver (LINCS)<sup>26</sup> algorithm. Each component (ions, solute and solvent) was separately coupled in the NPT (constant particle number, pressure, and temperature) ensemble. The system pressure was controlled

by Parrinello-Rahman coupling,<sup>27</sup> and the temperature was managed by Berendsen thermostat.<sup>28</sup>

Atomic positions were optimized using the steepest descent and conjugated gradient algorithm. The energy minimization convergence criterion was 50 N of maximum force applied to atoms in the investigated systems where the volume was balanced using a stepwise heating of the system. The heating or warming up scheme was the following: 50, 100, 200, and 350 K for 20 ps simulation time performed in 1 fs step size. After that, the system was cooled down to 300 K, and then a MD simulation of 500 ps was carried out. The trajectory file was recorded every 1000 simulation steps. The CEP of all ligands were assembled in the same file considering the ligands conformations recorded from 50 to 500 ps, and these data were used for building the QSAR models. In Figure 2 the CEPs for one of the most and least active compounds from each training set are presented. It can be noticed that the conformers from the second data set did not sweep out a very large 3D conformational space, even without the presence of the biomacromolecule. It was verified in preliminary tests that longer simulations would not be necessary to obtain reliable PLS models.

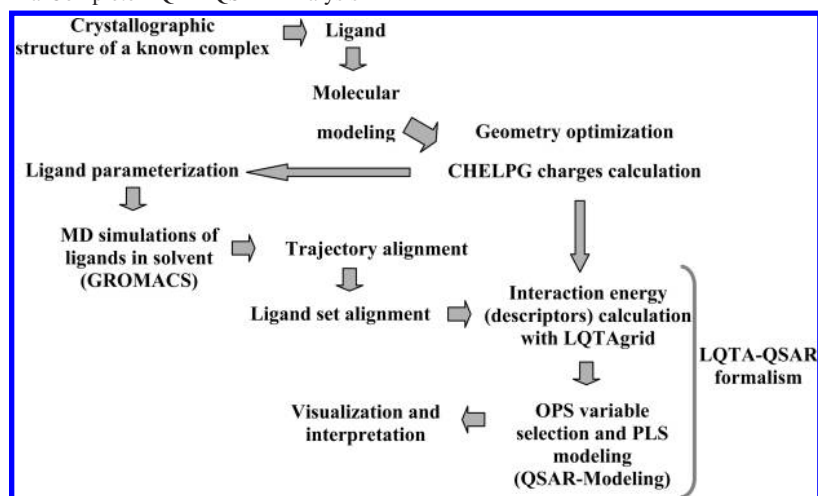
**LQTAgrid Analysis.** The CEPs, resulting from GROMACS MD simulations, were aligned using conditions similar to those from the literature<sup>15,16</sup> (numbered atoms in Table 1) and were used to generate the energy descriptors of intermolecular interaction. As already mentioned, the probes created for the LQTAgrid program are based on the ff43a1 force field parametrization to simulate atoms or molecular fragments as NH<sub>3</sub><sup>+</sup>, for example, which corresponds to the amino-terminal portion of peptides. The probes explore every point of a 1 Å resolution grid. The grid size was 24 × 22 × 20 for set 1, and 38 × 28 × 28 Å for set 2, and only the NH<sub>3</sub><sup>+</sup> probe was used to calculate the 3D descriptors. Preliminary tests indicated that good results could be





**Figure 2.** Comparison of the CEPs resulting from MD simulations to one of the most active and inactive compounds of each investigated data set. The biological data of set 1 and set 2 are expressed as  $\Delta G$  (kcal/mol) and  $\text{pIC}_{50}$ , respectively.

**Scheme 1.** Steps Involved in a Complete LQTA-QSAR Analysis



accomplished even employing one probe to generate the energy descriptors. The generated 3D-energy interaction descriptors were exploited in the variable selection procedure. Scheme 1 illustrates the steps involved in a complete LQTA-QSAR analysis.

**Variable Selection and Model Validation.** Descriptor matrices generated by the LQTAgrid module (21,120 variables for data set 1 and 59,584 for data set 2) were previously autoscaled to perform the variable selection and model building procedures. The absolute values of the correlation coefficients between each descriptor and the biological activity were calculated, and those with coefficients lower than 0.2 were eliminated from the analysis. At this point, 2449 independent variables remained for set 1 and 19,924 for set 2. In addition, descriptors whose plots versus the dependent variable showed nonuniform distribution or dispersion were also eliminated. The initial sets of descriptors used to carry out the variable selection using the ordered predictors selection (OPS)<sup>12</sup> algorithm were 1570 descriptors (set 1) and 8265 (set 2), respectively. The basic idea of this algorithm is to attribute an importance to each descriptor based on an informative vector. The columns of the matrix are rearranged in such a way that the most important

descriptors are presented in the first columns. Then, successive PLS regressions are performed with an increasing number of descriptors in order to find the best PLS model. In this analysis, the regression vector was used as an informative vector and the correlation coefficient of cross-validation,  $q^2$ , as a criterion to select the best models.

Regression models were validated applying the leave- $N$ -out (LNO) cross-validation and  $y$ -randomization.<sup>29–32</sup> In the LNO cross-validation procedure,  $N$  compounds ( $N = 1, 2, \dots, 10$ ) were left out from the training set. For a particular  $N$ , the data were randomized 20 times, and the average and standard deviation values for  $q^2$  were used. In the  $y$ -randomization, the dependent variable-vector was randomly shuffled 50 times for the two investigated sets.

## RESULTS AND DISCUSSION

PLS regression models were built after the OPS variable selection, which resulted in good statistics (Table 3). For set 1, with 40 compounds in the training set and 12 variables selected, the model with two latent variables (LV) was indicated as the best model by the leave-one-out (LOO) cross-validation. The  $q^2$  and  $r^2$  values for this model are 0.72

**Table 3.** Statistical Parameters Found for the OPS-PLS Models and Literature Models<sup>15,16a</sup>

	$q^2$	$r^2$	RMSECV	RMSEC
set 1 (2 LVs)	0.72	0.81	0.70	0.60
ref 15 (MLR)	0.80	0.87	—	—
set 2 (4 LVs)	0.82	0.90	0.23	0.21
ref 16 (5 LVs)	0.55	0.91	0.41	0.19

<sup>a</sup> The values in parentheses correspond to the number of latent variables used in the PLS models.

and 0.81, respectively (see Table 3). The residuals [experimental activity ( $y_{\text{exp}}$ ) – calculated or estimated activity ( $y_{\text{cal}}$ )] for each compound of set 1 hardly exceed 1 kcal/mol in  $\Delta G$  predictions. For set 2, with 37 compounds in the training set, the best model was constructed with 10 variables (OPS-PLS) and 5 LV using LOO cross-validation, which resulted in  $q^2 = 0.82$  and  $r^2 = 0.90$ . The statistical parameters of the resulting OPS-PLS models for both data sets are close to those found in the literature<sup>15,16</sup> (Table 3).

The models obtained in this work were also validated applying the  $y$ -randomization and LNO cross-validation in order to evaluate their reliability and robustness. Good QSAR models must have an average value of  $q_{\text{LNO}}^2$ ,  $q_{\text{LNO}}^2$ , close to the  $q_{\text{LOO}}^2$  and standard deviation for each  $N$  should not exceed 0.1. It is recommended that  $N$  represents a significant fraction of samples (like leave-30%-out) in a satisfactory LNO test.<sup>32</sup>

The model for data set 1 is robust for at least  $N = 10$ , where  $q_{\text{LNO}}^2$  value was 0.71 being close to model  $q_{\text{LOO}}^2$  (0.72). Deviations from  $q_{\text{LNO}}^2$  for each  $N$  oscillate from 0.017 to 0.036. Regarding set 2, for up to  $N = 9$  the model presented  $q_{\text{LNO}}^2$  value (0.78) close to  $q_{\text{LOO}}^2$  value (0.82) and deviation did not exceed 0.08. However, for  $N = 10$ , deviation from  $q_{\text{LNO}}^2$  (0.11) exceeded the 0.1 limit. In both cases the parameters indicated a satisfactory robustness (see Figure 3).

*Overfitting* and chance correlation between the dependent variable and the descriptors were checked employing the  $y$ -randomization validation. Poor regression models, with low  $r^2$  and LOO  $q^2$  values, are expected when  $y$  values (dependent variables) are scrambled. Otherwise, if good regression models are obtained in the  $y$ -randomization test (relatively high  $r^2$  and LOO  $q^2$ ), it implies that

the QSAR model proposed is not acceptable, probably due to a chance correlation or structural redundancy of the training set.<sup>28,29</sup>

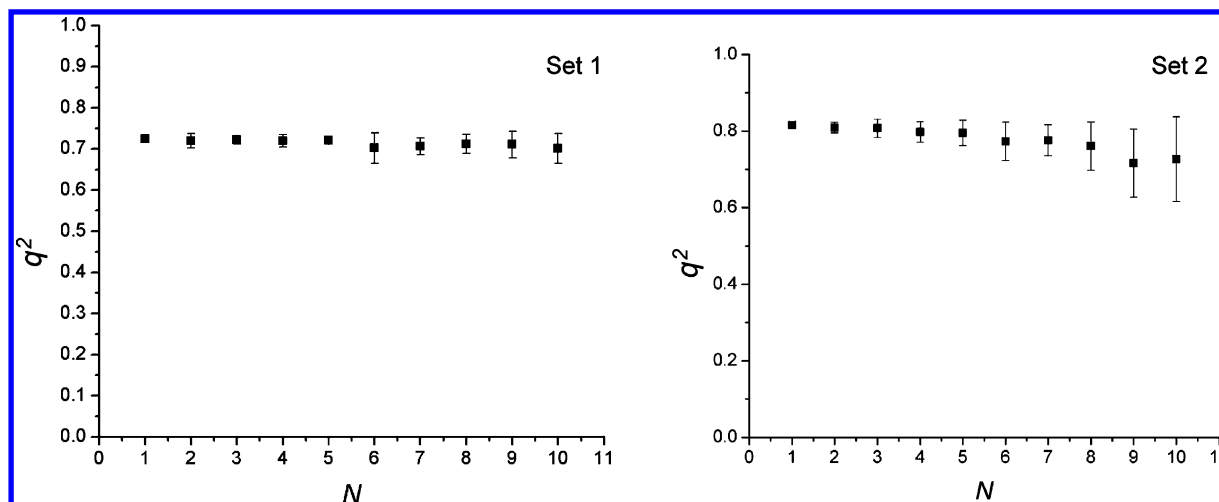
The LOO  $q^2$  and  $r^2$  values resulting from  $y$ -randomization for the set 1 were  $-0.32 \pm 0.20$  and  $0.18 \pm 0.06$ , respectively. Additionally, the LOO  $q^2$  and  $r^2$  values found for set 2 were  $-0.94 \pm 0.64$  and  $0.21 \pm 0.06$ , respectively. The  $y$ -randomizations performed imply that acceptable QSAR models were obtained for the given data sets by the current modeling method. The results of these internal validation methods are presented in Figure 4.

Unfortunately, the literature models<sup>15,16</sup> were not thoroughly validated. However, nowadays such procedures are highly recommended, particularly in the case of the literature model<sup>16</sup> for which the difference between  $q^2$  and  $r^2$  (0.36) is higher than 0.2, suggesting that the model was *overfitted*.

To ascertain the predictive power of the selected OPS-PLS models, two test sets were used containing seven (set 1) and seven (set 2) ligands, respectively. The external validation statistics ( $q_{\text{ext}}^2$ ) found for sets 1 and 2 were 0.60 and 0.69, respectively, demonstrating good external predictability. The individual residuals values [experimental activity ( $y_{\text{exp}}$ ) – predicted activity ( $y_{\text{pred}}$ )] are presented in Table 4, and the plots between experimental and predicted activities found for training and test sets 1 and 2 are shown in Figure 5.

**Descriptors Interpretation.** The descriptors selected by OPS are visualized in Figures 6 and 7 as solvent accessibility surfaces (ViewerLite 5.0, Accelrys, Inc., 2002). Light blue regions denote steric interactions corresponding to positive PLS regression coefficients, while pink regions represent steric regions related to negative regression coefficients. Likewise, dark blue color and red regions denote electrostatic descriptors with positive and negative regression coefficients, respectively. A conformation of the most active compound for each investigated set and its relation to the binding site interactions are shown.

Descriptors unveiled by the variable selection can be related to the interactions found in the binding pocket, mainly regarding the most active molecules. For data set 1, the **LJ+** descriptor region can be associated with the hydrophobic interactions involving the amino acid residue HIS377, which contributes to the ligand stabilization at the binding site by

**Figure 3.** Plots of the LNO results found for sets 1 and 2, respectively.

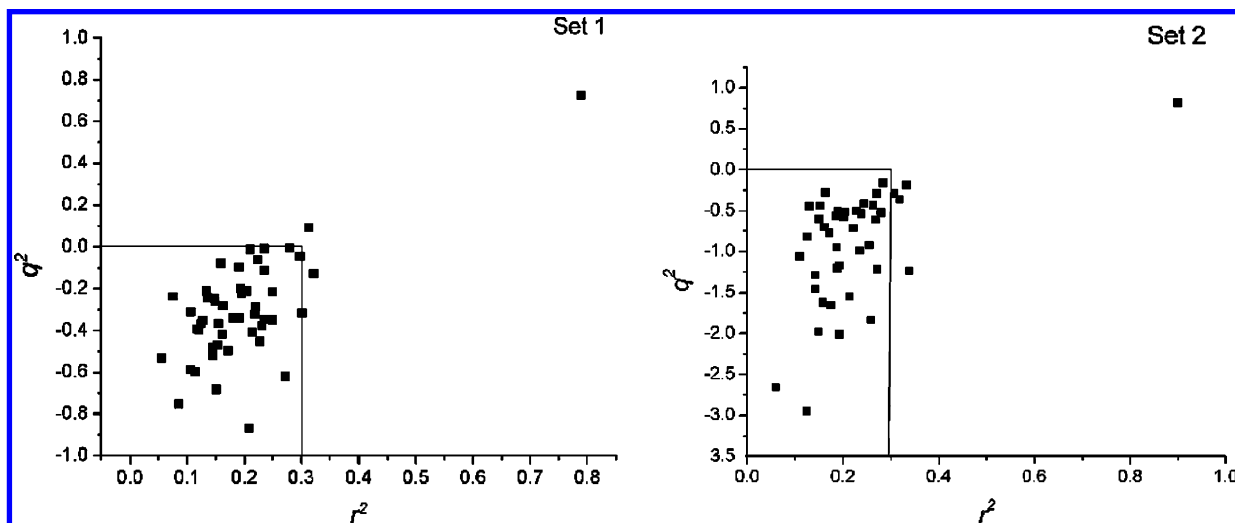


Figure 4. Plots of  $q^2$  versus  $r^2$  found for 50 y-randomizations.

Table 4. Residuals Values Obtained for the Test Sets Using the OPS-PLS Models

set 1	$y_{exp}$	$y_{pred}$	residuals (kcal/mol)	set 2	$y_{exp}$	$y_{pred}$	% residuals
3	6.04	4.80	1.24	4	7.97	7.50	5.9%
8	5.26	4.17	1.09	10	7.7	7.65	0.6%
11	4.65	3.93	0.72	13	7.6	7.41	2.6%
13	3.81	3.31	0.50	17	7.25	6.97	3.9%
30	3.39	3.66	-0.27	23	7.05	7.10	-0.7%
38	2.90	3.52	-0.62	30	6.82	7.17	-5.1%
20	2.32	3.31	-0.99	38	6.51	6.77	-3.9%

establishing hydrogen bonds interactions. The **2LJ-** descriptors are inversely related to the biological activities. They might be correlated to the ligand affinity for water molecules in the binding site, suggesting an unfavorable orientation in the active site. Another interesting descriptor is the **1LJ-**, which is related to the carbonyl group directly attached to the glucosyl ring, responsible for reducing 10-fold the ligand affinity. Considering the electrostatic descriptor **C-**, it can be related to the profile of the carbonyl group, which is far

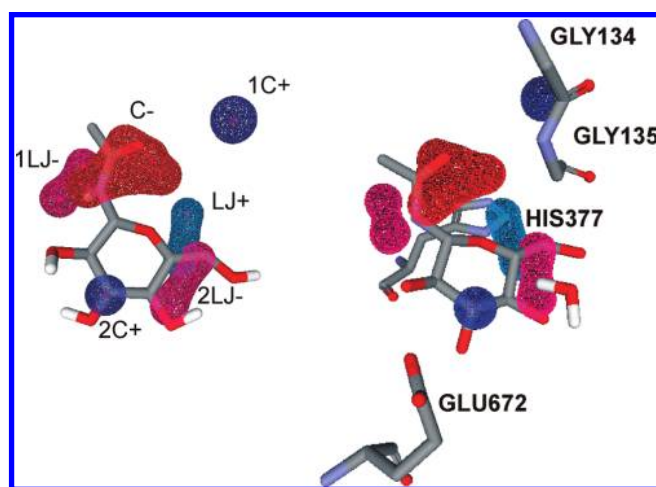


Figure 6. Visualization of the LQTAgrid descriptors found for the most active molecule of set 1 (ViewerLite 5.0, Accelrys, Inc., 2002).

from the glucosyl ring by just one atom. The **2C+** descriptor is possibly related to the hydroxyl groups of the glucosyl

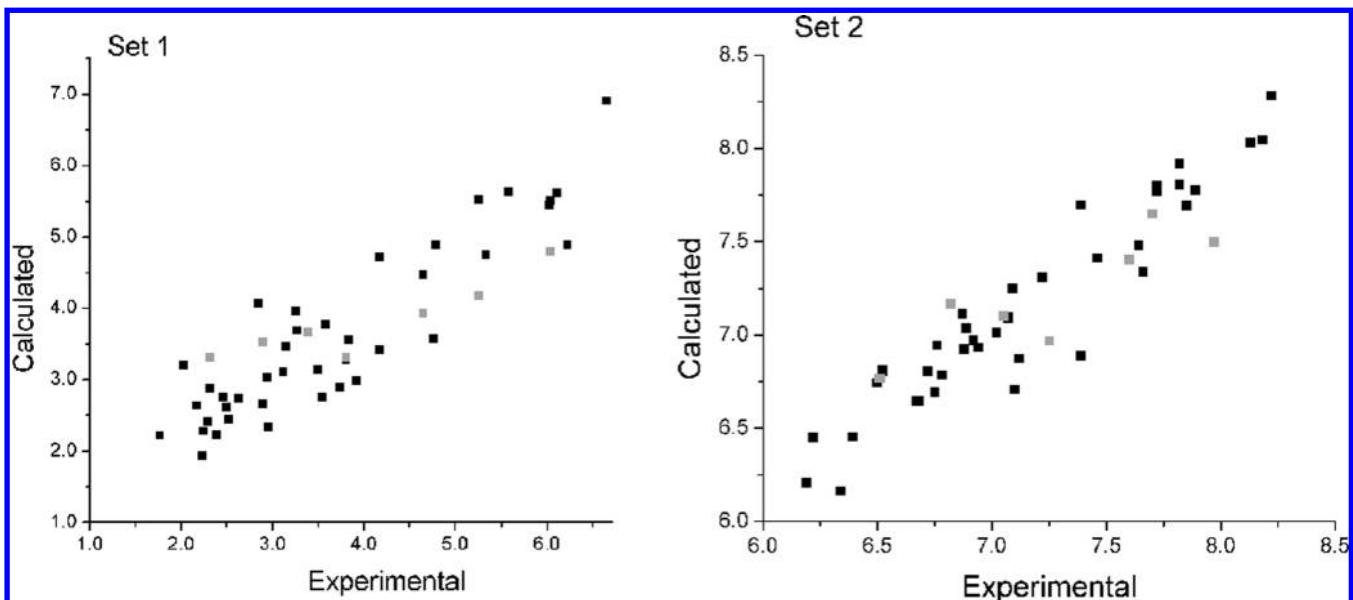
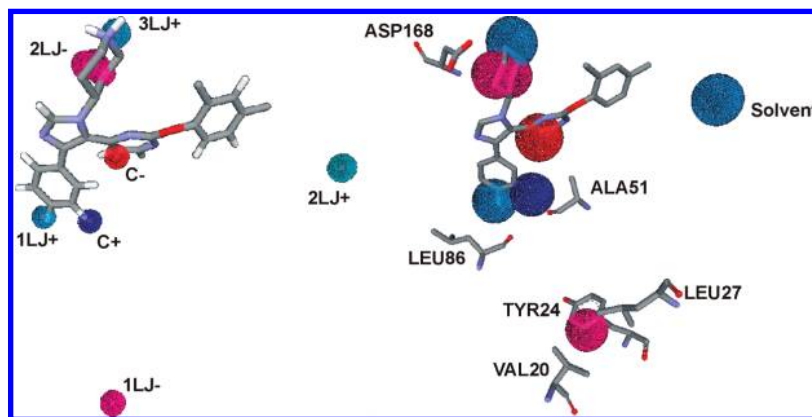


Figure 5. Plot of observed (experimental) versus predicted (calculated) activities found for training (black) and test (light gray) sets (set 1 and 2).



**Figure 7.** Visualization of the LQTAgrid descriptors found for the most active molecule of set 2 (ViewerLite 5.0, Accelrys, Inc., 2002).

ring, pointing out their relevance in the interaction with the GLU672 residue at the binding site. The **1C+** descriptor is probably related to the ligand's hydrophilic interaction with the backbone of the GLY134 and GLY135 residues.

The descriptors selected for data set 2 can be also interpreted based on the key interactions occurring at the binding site, including the long-range interactions. The **1LJ+** variable is positioned in a hydrophobic region at the binding site. The higher the frequency of fluorine atoms on the **1LJ+** region, the stronger is the hydrophobic interaction around the LEU86 residue (Figure 7). The **C+** descriptor probably describes the electron density of the ring containing a fluorine atom as substituent. The **2LJ+** descriptor is positively related to  $pIC_{50}$ , suggesting that more electropositive rings interact stronger with the ALA51 residue backbone. The **2LJ-** and **3LJ+** descriptors can be mostly related to the interaction with the ASP168 residue of the binding consistent with a solvent region next to the protein surface. The **C-** descriptor can be interpreted as a grid point proximity occupied by a richer electron region in the pyrimidine ring. Even though the **1LJ-** descriptor is 11 Å far from the ligand, it is probably related to the TYR24 residue region at the binding site. It is always recommended to test other pretreatments to avoid those descriptors far from the CEP especially when the structure of the receptor is not available. In this work, blockscaling<sup>33</sup> taking into account the two blocks (Coulomb and LJ) was applied, but the models obtained could not be well validated by methodologies described earlier. Thus, the autoscaling pretreatment was kept for this data set.

When the literature model for set 1<sup>15</sup> is compared to the OPS-PLS model, it can be seen that the descriptors are very similar concerning the **C-** and **1LJ-** regions. Both models provide quite the same interpretations except that the OPS-PLS model does not include descriptors for hydrogen bonding interactions. However, differences between the models appear in descriptors found at the glycoside ring portion, which were not reported in the literature.<sup>15</sup> Thus, the approach presented in this study provides descriptors for a more extended region of the system under investigation.

The descriptors selected for set 2 in the final OPS-PLS model were not well related to the CoMFA surfaces reported by Ravindra and co-workers.<sup>16</sup> The calculated LQTAgrid descriptors were quite distinct from those reported in the literature,<sup>16</sup> impairing any kind of comparison.

## CONCLUSIONS

A new formalism that takes advantage of GROMACS MD frames to build interaction energy models was presented in this study. The LQTA-QSAR formalism can be adapted to reach the user needs on building 4D-QSAR models, using a recent algorithm for variable selection, OPS, which has proved to be fast and capable of providing suitable variables for a PLS multivariate analysis.

The statistical parameters found for the LOO cross-validation procedure and external validation presented similar values to those obtained in the refs 15 and 16. However, the LQTA-QSAR models were thoroughly validated applying the LNO internal cross-validation and **y**-randomization methods, which were not employed in the original studies. Thus, the best OPS-PLS models have demonstrated robustness and a good predictability for both investigated sets, using unbound ligands in a solvent medium.

As the CEPs are calculated using the GROMACS program, the users have freedom to create and align the ligands' profiles using conformers from a more realistic representation of the investigated system (explicit solvent medium, ligand–receptor complexes, etc.). In this sense, the LQTA-QSAR formalism is a promising tool for the ligand- and structure-based drug design strategies.

It is noteworthy that the LQTA-QSAR paradigm is also a quite user-friendly computational method, the calculations are not time-consuming, and the calculation options can be adapted to better describe each investigated system. This methodology can be used employing only open source software, which guarantees free access to explore the available tools, and monitoring all steps involved in the construction of 4D-QSAR models. As already mentioned, it is available for evaluation by the scientific community at <http://lqta.iqm.unicamp.br>.

## ACKNOWLEDGMENT

The authors thank FAPESP, CAPES, and CNPq for the financial support.

## REFERENCES AND NOTES

- (1) Ferreira, M. M. C. Multivariate QSAR. *J. Braz. Chem. Soc.* **2002**, *13* (6), 742–753.
- (2) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967.



- (3) Martinez-Merino, V.; Cerecetto, H. CoMFA-SIMCA model for antichagasic nitrofurazone derivatives. *Bioorg. Med. Chem.* **2001**, *9*, 1025–1030.
- (4) Wen-Na, Z.; Qing-Sen, Y.; Jian-Wei, Z.; Ma, M.; Ke-Wen, Z. Three-dimensional quantitative structure-activity relationship study for analogues of TQXs using CoMFA and CoMSIA. *J. Mol. Struct. (Theochem)* **2005**, *723* (1–3), 69–78.
- (5) Martens, H.; Naes, T. *Multivariate Calibration*; Ed. Chichester: Wiley, 1989.
- (6) Manne, R. Analysis of two partial least squares algorithms for multivariate calibration. *Chemom. Intell. Lab. Syst.* **1987**, *1*, 187–197.
- (7) Agnar, H. PLS regression methods. *J. Chemom.* **1988**, *2* (3), 211–228.
- (8) de Jong, S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.
- (9) Nilsson, J.; de Jong, S.; Smilde, A. K. Multiway calibration in 3D QSAR. *J. Chemom.* **1997**, *11* (6), 511–524.
- (10) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524.
- (11) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7* (8), 306–317.
- (12) Teófilo, R. F.; Martins, J. P.; Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemometr.* **2009**, *23* (1), 32–48.
- (13) Spoel, D. v. d.; Lindahl, E.; Hess, B.; Buuren, A. R. v.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; Drunen, R. v.; Berendsen, H. J. C. *Gromacs User Manual version 3.3*; 2005. GROMACS: Fast, Free and Flexible MD-Paper Manuals. <http://www.gromacs.org/content/view/27/42/> (accessed Apr 07, 2006).
- (14) Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–1363.
- (15) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand-receptor binding free energy by 4D-QSAR analysis: Application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1141–1150.
- (16) Ravindra, G. K.; Achaiah, G.; Sastry, G. N. Molecular modeling studies of phenoxypyrimidinyl imidazoles as p38 kinase inhibitors using QSAR and docking. *Eur. J. Med. Chem.* **2008**, *43* (4), 830–838.
- (17) Watson, K. A.; Chrysina, E. D.; Tsitsanou, K. E.; Zographos, S. E.; Archontis, G.; Fleet, G. W. J.; Oikonomakos, N. G. Kinetic and crystallographic studies of glucopyranose spirohydantoin and glucopyranosylamine analogs inhibitors of glycogen phosphorylase. *Proteins: Struct., Funct., Bioinf.* **2005**, *61* (4), 966–983.
- (18) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in MAP kinases. *Structure* **1998**, *6* (9), 1117–1128.
- (19) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37* (2), 785.
- (20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Laham, A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *GAUSSIAN03, revision C.02*; Department of Chemistry, Carnegie Mellon University: Pittsburgh, PA, 2003.
- (21) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11* (3), 361–373.
- (22) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36* (22), 3219–3228.
- (23) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662.
- (24) Kusalik, P. G.; Svishchev, I. M. The spatial structure in liquid water. *Science* **1994**, *265* (5176), 1219–1221.
- (25) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.
- (26) Hess, B.; Henk, B.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.
- (27) Parrinello, M.; Rahman, A. Crystal structure and pair potentials: A molecular dynamics study. *Phys. Rev. Lett.* **1980**, *45* (14), 1196.
- (28) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
- (29) Bratchell, N. Chemometric methods in molecular design. *J. Chemom.* **1997**, *11* (1), 93–94.
- (30) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22* (1), 69–77.
- (31) Golbraikh, A.; Tropsha, A. Beware of q<sup>2</sup>! *J. Mol. Graphics Modell.* **2002**, *20* (4), 269–276.
- (32) Kiralj, R.; Ferreira, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and applications. *J. Braz. Chem. Soc.*, in press.
- (33) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of Comparative Molecular Field Analysis Models: Effects of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospholipase A<sub>2</sub> Inhibitors. *J. Med. Chem.* **1997**, *40*, 1136–1148.

CI900014F