JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Four-Dimensional Structure−Activity Relationship Model to Predict HIV-1 Integrase Strand Transfer Inhibition using LQTA-QSAR Methodology

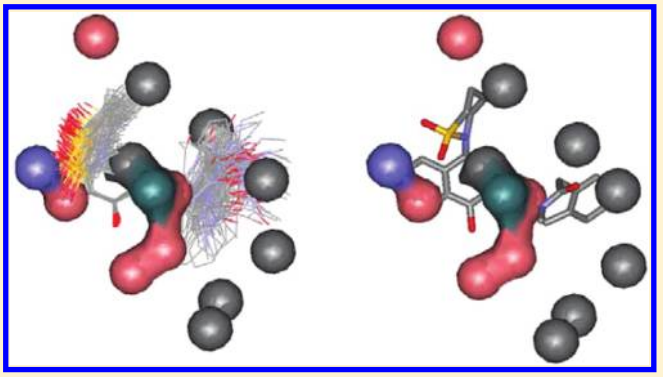
Eduardo B. de Melo[†] and Márcia M. C. Ferreira*[,‡]

[†]Theoretical Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Parana State University−Unioeste, 2069 Universitaria St, Cascavel, PR, 85819-110, Brazil

[‡]Laboratory for Theoretical and Applied Chemometrics (LQTA), Institute of Chemistry, University of Campinas−UNICAMP, P.O. Box 6154, Campinas, SP, 13084-971, Brazil

Supporting Information

ABSTRACT: Despite highly active antiretroviral therapy (HAART) implementation, there is a continuous need to search for new anti-HIV agents. HIV-1 integrase (HIV-1 IN) is a recently validated biological target for AIDS therapy. In this work, a four-dimensional quantitative structure−activity relationship (4D-QSAR) study using the new methodology named LQTA-QSAR approach with a training set of 85 HIV-1 IN strand transfer inhibitors (INSTI), containing the β-diketo acid (DKA) substructure, was carried out. The GROMACS molecular dynamic package was used to obtain a conformational ensemble profile (CEP) and LQTA-QSAR was employed to calculate Coulomb and Lennard−Jones potentials and to generate the field descriptors. The partial least-squares (PLS) regression model using 14 field descriptors and 8 latent variables (LV) yielded satisfactory statistics (R^2 = 0.897, SEC = 0.270, and F = 72.827), good performance in internal (Q_{LOO}^2 = 0.842 and SEV = 0.314) and external prediction (R_{pred}^2 = 0.839, SEP = 0.384, ARE_{pred} = 4.942%, k = 0.981, k' = 1.016, and $|R_0^2 - R_0'^2|$ = 0.0257). The QSAR model was shown to be robust (leave-N-out cross validation; average Q_{LNO}^2 = 0.834) and was not built by chance (**y**-randomization test; R^2 intercept = 0.109; Q^2 intercept = −0.398). Fair chemical interpretation of the model could be traced, including descriptors related to interaction with the metallic cofactors and the hydrophobic loop. The model obtained has a good potential for aid in the design of new INSTI, and it is a successful example of application of LQTA-QSAR as an useful tool to be used in computer-aided drug design (CADD).



INTRODUCTION

Human immunodeficiency virus (HIV), a retrovirus, is the primary cause of acquired immunodeficiency syndrome (AIDS) and one of the principal medical and social problems currently. However, despite the considerable progress in antiretroviral therapy, the eradication of HIV-1 remains unfeasible and, due to the development of resistant strains, side effects, and the establishment of viral reservoirs in memory T cells, there is a continuous need for new anti-HIV agents.[1−3] Data from the UNAIDS 2010 report on the global AIDS epidemic estimated that between 31.4 and 35.3 million people were living with HIV at the end of 2009.[4]

HIV type 1 integrase (HIV-1 IN) is one of the three viral enzymes essential for viral replication. In two reactions that HIV-1 IN catalyzes, 3P′-processing and strand transfer, the last two nucleotides of the viral DNA 3P′-end are cleaved and the remaining viral fragment is inserted into the host DNA.[5] HIV-1 IN strand transfer inhibitors (INSTI) are one of the most recent breakthroughs for AIDS pharmacological treatment. The first representative of this class of drugs is raltegravir (Figure 1), an N-Me pyrimidone derivative of the 4,5-dihydroxypyrimidine carboxamides, known to be well-tolerated and which has no serious drug-related adverse events.[3,6]

Computational chemistry is currently an important contributor to rational drug design.[7] Quantitative structure−activity relationships (QSAR) describe how a given biological activity or a physical−chemical property can vary as a function of molecular structures of a set of chemical compounds. The chemical structures are characterized by molecular descriptors, such as those obtained by a well-specified algorithm applied to a defined molecular representation (such as those based on graph theory or derived from quantum chemical calculations) or from experimental procedures (physical−chemical properties such as 1-octanol/water partition coefficient or water solubility at 25 °C).[8]

Received: January 20, 2012

   dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

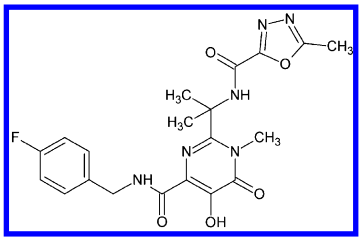


Figure 1. Chemical structure of raltegravir (Isentress; Merck Co.).

In the literature, there are several QSAR studies with HIV-IN inhibitors, as revised by Nunthaboot et al.[9] Recently, Melo and Ferreira[10] published a 2D-QSAR model built with electronic descriptors obtained through molecular modeling (E_{HOMO}, α_{yy}, E_T) and based on the atomic intrinsic state theory (*SeaC2C2aa*). The selected descriptors were well-related to the most accepted inhibition mechanism.[11] However, although providing good results, the model was based on a set of only 33 4,5-dihydroxypyrimidine carboxamides,[12] structurally homogeneous and related to raltegravir.

The 4D-QSAR formalism is a grid-based technique and was originally proposed by Hopfinger et al.[13] As compared to traditional 3D-QSAR methods,[14] this approach inserts a "fourth dimension" in the study: conformational flexibility.[15] This is obtained with a conformational ensemble profile (CEP) for each compound generated by molecular dynamics (MD) simulation, followed by alignments of the different conformations obtained.[14−16] In this paper, a new QSAR study is presented for a data set built with 85 INSTI and performed by means of the new 4D-QSAR approach named LQTA-QSAR.[15] This paper should be considered one of the first applications using this new methodology. The findings can be helpful for designing new active derivatives and providing a better understanding of the inhibition of the strand transfer (ST) reaction.

MATERIALS AND METHODS

1. Training Set. A training set containing 85 INSTI was selected from the literature[12,17−22] and included the compounds used previously by Melo and Ferreira.[10] All compounds contain the pharmacophore of the INSTI, the DKA substructure (Figure 2). In accordance with the information available in each paper,

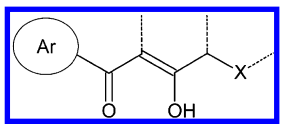


Figure 2. β-Diketo acid (DKA) substructure.

the biological activity is the concentration required for 50% inhibition of the ST reaction (IC_{50}, in moles per liter) and was determined by the methodology described by Hazuda et al.[23] The experimental IC_{50} values were converted into their corresponding pIC_{50} ($-\log IC_{50}$) values, obtaining a range from 4.30 to 8.15 logarithmic units. The structures and pIC_{50} of all compounds are available in the Supporting Information.

2. LQTA-QSAR Study. The LQTA-QSAR approach explores both the main features of comparative molecular field analysis (CoMFA) and of 4D-QSAR paradigms. This method is based on the generation of a CEP for each compound, instead of only one conformation, followed by the calculation of 3D descriptors, using the Coulomb and Lennard-Jones potentials.[24] The main points of the whole process are summarized in Table 1 and

Table 1. Operational Steps Performed in this LQTA-QSAR Study

step	description of the step
1	geometry optimization of each compound from data set using DFT methodology, B3LYP/6-31G(d,p)*
2	calculation of ChelpG charges for predominant microspecies of biological test
3	conversion of output files (*.out) into mol2 input files
4	generation of *.gro and *.top input files
5	calculation of Gasteiger and Marsili charges for molecular dynamics (MD) simulations and parametrization
6	MD simulations of the molecules
7	alignment of all conformations obtained for each molecule based on the pharmacophore INSTI inhibitors to obtain the CEP
8	use of CEP to obtain a virtual grid
9	inclusion of the input files in the LQTA-QSAR (*.gro file) with CEP and the *. top files with CHELPG charges
10	generation of Coulomb and Lennard-Jones descriptors
11	elimination of descriptors with $r < \|0.2\|$
12	elimination of distant descriptors (cut off by variance: 0.01)
13	elimination of poorly distributed descriptors
14	variable selection with OPS method
15	building of QSAR models
16	QSAR model validation
17	mechanistic interpretation

detailed in the subsequent sections. LQTA methodology was totally developed as open access software and is implemented in the JAVA environment, which allows its use on any O.S. compatible with this computational language.

3. Molecular Modeling. The molecular set was built by means of the HyperChem 7 software,[25] based on crystallographic structures obtained from the Cambridge Structural Database[26] (see Supporting Information). Geometry optimizations were performed in the following sequence: MM+ (using HyperChem 7) → HF/6-31G(d) (using Gaussian 03)[27] → B3LYP/6-31G(d,p) (also using Gaussian 03 program). The DFT/B3LYP functional was chosen because comparative studies between B3LYP, ab initio, and semiempirical theories reported that this method leads to better QSAR models when molecular geometries and energies are considered.[28−33]

The LQTA-QSAR descriptors are obtained through the CEP, which in turn are based on MD simulations. These simulations were performed through the GROMACS[34] package, another open access software, with explicit water molecules, so that steric and electrostatic forces that result from the solvent on the various conformations were taken into account. Each molecule was simulated as its corresponding predominant microspecies at pH 7.5, according to predictions made by Marvin 4.1.8.[35] Thus, 75 structures were used in the ionized form and 10 as neutral ones.

ChelpG atomic charges (charges from electrostatic potentials using a grid based method) were calculated for each microspecies in Gaussian 03.[27] This method is based on the electrostatic potencial (ESP) and is considered to be quite appropriate for a 4D-QSAR study, because it is able to provide a more accurate reproduction of the molecular ESP,[36] favoring the description of Coulomb potentials between the probes and the molecules.

From the output files (*.out), the *.mol2 files were obtained through Open Babel 2.1.1,[37] which were used on the server PRODRG 2.5 (http://davapc1.bioch.dundee.ac.uk/prodrg)[38] to build the geometry (*.gro) and topology (*.top) files used as input data in GROMACS. Because this program uses a force field parametrized for empirical atomic charges (ffG43a1), Gasteiger

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

and Marsili atomic charges[39] were derived using the software HyperChem 7.[25]

The MD simulations were carried out in a virtual cubic box with periodic properties specific to each molecule. The only parameter in common was the minimum distance of 10 Å between the molecule and the edge of the box. In the simulations of ionized molecules, one charged atom of Na^+ or Cl^- was used as counterion to meet the condition of system neutrality. Each molecule was optimized using the steepest descent and conjugate gradient[40] algorithms with convergence in 50 N of maximum force applied to the atoms. Then, the volume of the system was balanced using a heating process divided into steps of 50, 100, 200, and 350 K for 20 ps in each step. The system was then cooled to 300 K and simulated for 500 ps. A trajectory file was saved every 1000 steps of simulation. The conformations obtained from each derivative were organized in the same *.gro file. For the construction of the CEP, the rotational and translational motions of each conformation were eliminated. All conformations obtained were aligned having the DKA substructure (Figure 2) as a basis, without considering the aromatic side chains (Ar).

4. LQTA-QSAR Descriptors. The Coulomb and Lennard-Jones descriptors were generated using a virtual cubic grid of 28 × 28 × 28 $Å^3$ (center in $x = 6$ Å, $y = 4$ Å, and $z = 3$ Å) built by Auto Docking.[41] Next, each point of this grid was explored by probes selected based on the mechanism of inhibition of INSTI: Ar(C−H) and Zn^{2+}. Thus, for the Zn^{2+} probe (used to represent the metallic cofactor Mg^{2+}), a total of 21 952 Coulomb descriptors were obtained and the same number of Lennard-Jones descriptors. On the other hand, only 21 952 Lennard-Jones descriptors were generated by the Ar(C−H) probe (representing the atoms of the aromatic side chain) totaling 65 856 descriptors.

5. Variable Reduction and Preprocessing. The number of descriptors was reduced to 3056 (1440 obtained with the Zn^{2+} probe and 1616 with the Ar(C−H) probe), with the elimination of those with absolute values of the Pearson correlation coefficient ($|r|$) of the pIC_{50} less than 0.2. It was considered that below this threshold no useful statistical information would be provided to the model.[8]

The order of magnitude of the Lennard-Jones descriptors is higher than that of the Coulomb descriptors. Thus, it was necessary to choose the appropriate method of data preprocessing to avoid that all descriptors within the virtual cubic box had the same importance, because the more important ones should be those close to the molecules. In this situation, the interpretation of the model with regard to the mechanism of inhibition can be compromised. Considering the type of behavior and descriptors, the most appropriate procedure for this study is blockscaling. In this approach, the data are preprocessed separately (in blocks). This helps in the selection of descriptors with maximized variance with respect to biological activity.[42,43]

The matrix containing the descriptors with $|r| > 0.2$ was split into two: one with the Lennard-Jones descriptors and another with the Coulomb descriptors. Then, a cutoff was applied to the Lennard-Jones descriptor values, using the limit of 30.00 kcal·mol^{-1} (125.5 kJ·mol^{-1}). This cutoff was used in order to adjust the order of magnitude, but without affecting the relative importance between descriptors of the same class. Then, a digital filter[44] was applied to identify and eliminate those descriptors with abnormal dispersion in relation to biological activity. The digital filter works by dividing the range of descriptors values into small intervals. The lower the f_{max} (maximum frequency) calculated for each interval, the more dispersed the descriptor with regard to the range of variation of biological activity and the better its performance. Figure 3 illustrates the procedure.

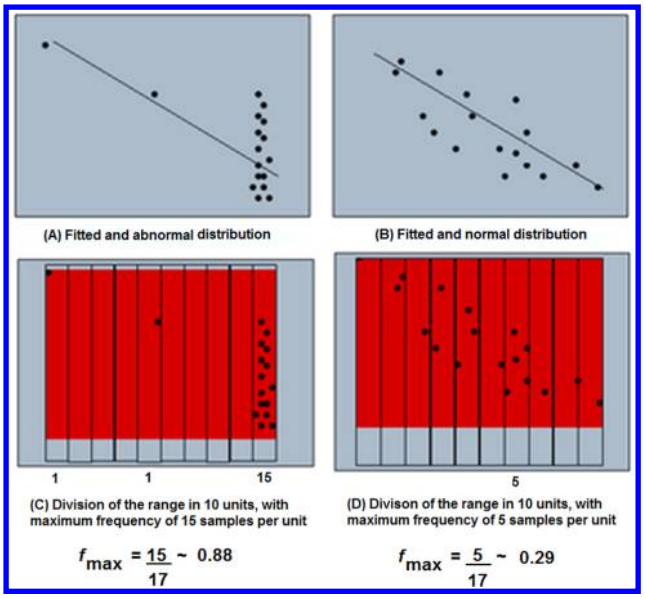


Figure 3. Schematic representation of the digital filter operation, applied to select descriptors with linear trend and normal distribution in relation to biological activity.

6. Variable Selection and Building of the Model. Initially, variable selection was carried out for the two separate matrices, one for Lennard-Jones and one for Coulomb descriptors, using mean centered data (the mean of each column was subtracted from all values in that column). After the selection of relevant variables, they were merged into a single array. From this point on, the data were autoscaled, i.e., columnwise, mean-centered, and scaled to unity variance.

The remaining descriptors (575) were further analyzed employing the ordered predictor selection (OPS) algorithm. This variable selection method, developed by Teófilo et al.,[45] has presented good results for QSAR studies where applied,[10,16,46−50] and is able to build models by rearranging the columns of the matrix in such a way that the most important descriptors, classified according to an informative vector, are placed in the first columns. Then, successive partial least-squares (PLS) regressions[51−53] were built with increasing number of descriptors, in order to find the best model. In this work, three available informative vectors were used: correlation vector, whose elements are the correlation coefficients between each descriptor and the biological activity; the regression vector; and an elementwise product between both vectors. The best models were classified in descending order of statistical quality according to their coefficient of determination of leave-one-out cross validation (Q_{LOO}^2) or standard error of cross validation (SEV) values.

The best reduced combination of descriptors was refined through the software Pirouette 4,[51] removing the descriptors that were less relevant to the model and, if necessary, outliers, seeking to obtain a statistically significant, robust, and interpretable model. The outliers were identified by plotting the studentized residuals (σ) versus the sample leverage generated through Pirouette. The compounds removed were those presenting high residuals and leverage or very high residuals (higher than 2σ). The outliers were parsimoniously removed and only if required.

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Contribution of Descriptors from the Model to Each LV, its Correlation Coefficient to Biological Activity, and Autoscaled Coefficients

descriptor	LV1*	LV2	LV3	LV4	LV5	LV6	LV7	LV8	r	autoscaled coefficients
24.13.12 Ar(C−H).LJ	0.196	0.226	0.103	0.298	0.118	0.186	−0.727	−0.349	0.236	0.251
18.18.19 Ar(C−H).LJ	0.227	−0.126	0.394	0.182	−0.457	0.204	0.046	0.304	0.272	0.188
21.15.16 Ar(C−H).LJ	−0.357	0.028	−0.263	−0.096	−0.645	−0.024	−0.299	−0.030	−0.429	−0.324
26.18.14 Ar(C−H).LJ	0.211	0.638	0.038	−0.054	−0.071	0.082	−0.150	0.411	0.254	0.349
19.21.24 Ar(C−H).LJ	0.211	0.005	−0.264	0.296	−0.106	0.318	0.347	−0.241	0.253	0.148
20.13.16 Ar(C−H).LJ	−0.258	0.258	−0.068	0.444	0.160	0.263	0.195	0.378	−0.310	−0.081
14.23.23 Ar(C−H).LJ	−0.249	0.154	0.541	−0.034	−0.292	0.108	0.188	−0.086	−0.299	−0.077
18.24.13 Ar(C−H).LJ	0.236	0.402	0.052	−0.498	−0.114	0.211	0.232	−0.404	0.283	0.266
20.17.19 Zn^{+2}.C	−0.208	0.286	−0.500	−0.053	−0.007	0.199	0.129	−0.007	−0.250	−0.138
22.16.9 Ar(C−H).LJ	0.243	−0.304	−0.184	−0.436	−0.056	0.303	−0.137	0.360	0.291	0.037
20.18.17 Ar(C−H).LJ	−0.210	−0.190	−0.169	0.097	−0.238	0.326	−0.179	−0.074	−0.252	−0.233
15.15.18 Ar(C−H).LJ	−0.220	−0.156	0.206	−0.232	0.372	0.625	−0.111	0.047	−0.264	−0.179
26.20.19 Ar(C−H).LJ	0.416	0.001	−0.189	−0.003	−0.045	−0.101	−0.081	0.258	0.499	0.283
14.15.21 Zn^{+2}.LJ	0.376	−0.193	−0.058	0.276	−0.144	0.227	0.155	−0.205	0.451	0.235
cumulated information by LV	12.4%	7.6%	10.7%	7.6%	6.7%	8.0%	10.4%	5.9%		

*LV = latent variable

7. Model Validation. The use of validation tools for QSAR models is an essential process to ensure the quality of a prediction model.[10,54−60] In this study, the most useful procedures known for quality certification of a model were used[10] (see Supporting Information for equations).

The model was internally validated in a comprehensive manner using a set of procedures suggested in the literature: the coefficient of multiple determination (R^2), the standard error of calibration (SEC), the F-ratio test (F), the Q_{LOO}^2, and the SEV. The robustness of the model was examined by a leave-N-out cross-validation[61] (LNO, N = 1, ..., 17, where "N" was repeated three times) procedure. The absence of chance correlation was checked using **y**-randomization analysis.[54,59] In this test, the **y** vector is randomized a certain number of times (50 times in this work), and the randomized models should not present good explained and predicted variances.

For external validation, the data set was split into training and test sets with the aid of hierarchical cluster analysis (HCA).[62] The results of external validation were evaluated by means of coefficient of determination of external validation (R_{pred}^2), the standard error of external prediction (SEP), and the average relative error (ARE_{pred}). However, to assess if the model obtained has the conditions that may be considered really predictive, Golbraikh−Tropsha statistics was also applied[10,58,63] by calculating the slopes of the regression lines of the external validation (k and k'), and the absolute value of the difference between the coefficients of multiple determination ($|R_0^2 - R_0'^2|$).

Another factor that has to be evaluated is the coincidence agreement between the signals of r for each descriptor with pIC_{50} and the corresponding signals of regression coefficients in the model. The mismatch between the contributions of these two factor signs is an indication of lack of self-consistency of the model.[65]

RESULTS

The therapeutic activity of a drug is usually caused by interactions with specific sites of proteic structures. These interactions may be described by molecular properties. Although it is possible to use structurally heterogeneous data sets, descriptors derived from these types of sets cannot adequately explain the structure−activity under study.[65,66] One of the advantages of the nD-QSAR (n = 3−6) approaches is the possibility of obtaining good results, since the selected descriptors can be translated into pharmacophoric models.[67−72]

The OPS algorithm generated a model with 34 descriptors. A step of refinement was performed through the program Pirouette 4, and the number of descriptors was reduced to 14: 12 derived from the Ar(C−H) probe and 2 from the Zn^{2+} probe. Nine compounds (10.6% of the training set) were identified as outliers (see Supporting Information). Thus, the final model was obtained based on 76 compounds, 14 descriptors, and 8 latent variables (LV). These LV accumulate 69.2% of the information (Table 2), which is enough to explain 89.7% of the variance and produce a model with a low standard error of calibration (SEC = 0.270). The results of the F-ratio test (n = 76; p = 8; $n - p - 1$ =

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

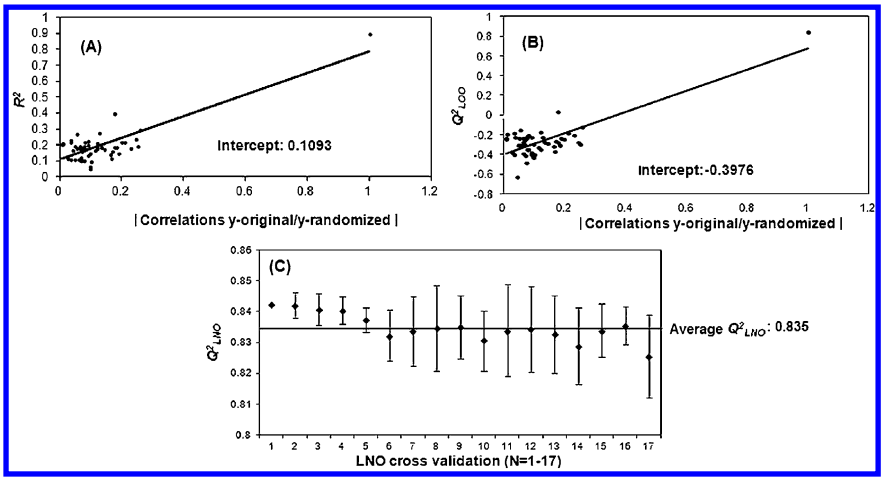


Figure 4. Results of the y-randomization test (A and B) and LNO cross-validation (C).

67; $\alpha = 0.05$–95% of confidence) was much larger than its critical value ($F = 72.8 > 2.08$). This amount of information was also enough to predict 84.2% of variance, with a low standard error of validation (SEV = 0.314). The difference between R^2 and Q_{LOO}^2 is 0.055 units, indicating the absence of overfiting.[55,73] The result of predictive residual sum of squares of the cross-validation procedure ($PRESS_{Val} = 7.487$) is smaller than the sum of squares of the experimental pIC_{50} (SSy = 47.47), which can be considered a first indication that the variance predicted by the model is real and not due to chance.[56] Finally, the model also proved to be self-consistent: the sign of the regression coefficient for each descriptor in the model and the sign of correlation coefficient between the respective descriptor and the biological activity are equal.[64]

$$\begin{aligned} pIC_{50} = {} & 6.948 + 0.201[24.13.12.Ar(C{-}H).LJ] \\ & + 0.147[18.18.19.Ar(C{-}H).LJ] \\ & - 0.260[21.15.16.Ar(C{-}H).LJ] \\ & + 0.287[26.18.14.Ar(C{-}H).LJ] \\ & + 0.118[19.21.24.Ar(C{-}H).LJ] \\ & - 0.064[20.13.16.Ar(C{-}H).LJ] \\ & - 0.062[14.23.23.Ar(C{-}H).LJ] \\ & + 0.213[18.24.13.Ar(C{-}H).LJ] \\ & - 0.106[20.17.19.Zn2{+}.C] \\ & + 0.030[22.16.9.Ar(C{-}H).LJ] \\ & - 0.211[20.18.17.Ar(C{-}H).LJ] \\ & - 0.156[15.15.18.Ar(C{-}H).LJ] \\ & + 0.268[26.20.19.Ar(C{-}H).LJ] \\ & + 0.185[14.15.21.Zn2{+}.LJ] \end{aligned}$$

$n = 76$; $R^2 = 0.897$; SEC = 0.270; $PRESS_{cal} = 4.896$; $F = 72.827$ ($cF = 2.082$); $Q_{LOO}^2 = 0.842$; SEV = 0.314; $PRESS_{val} = 7.484$ (SSy = 47.469).

Figure 4 presents the results of the y-randomization test and of LNO cross-validation. The result for the y-randomization shows the absence of chance correlation: the intercept $r(y,y_{rand})R^2 = 0.109 < 0.300$ and the intercept $r(y,y_{rand})Q_{LOO}^2 = -0.398 < 0.05$.[54] In LNO cross-validation, the difference between Q_{LOO}^2 and the average Q_{LNO}^2 is only 0.007 units (0.842 and 0.834, respectively), and they may be considered virtually identical. The greatest oscillation was observed at $N = 11$ ($Q_{L11O}^2 = 0.834 \pm 0.014$). Thus, the model may be considered robust.

Using HCA, the data set was divided into training and test sets.[54] The dendrogram (Supporting Information) shows that the test set is representative of the data set. The real model, obtained after the selection of the test set, may be considered virtually identical to the original model, also called auxiliary ($n = 61$; cumulated information: 69.8%; $R^2 = 0.905$; SEC = 0.251; $PRESS_{cal} = 3.270$; $F = 62.0$; $Q_{LOO}^2 = 0.832$; SEV = 0.308; $PRESS_{val} = 5.79$). The external validation results are presented in Table 3. The result for R_{pred}^2 is higher than the minimum threshold of 0.5, and only 0.003 units lower than Q_{LOO}^2, with a

Table 3. Results from External Validation[a] Performed with the Real Model

compound	pIC_{50} obs	pIC_{50} pred	residues
E7	8.000	8.373	−0.373
E8B	7.350	7.626	−0.276
G12	7.400	6.930	0.470
S9	7.300	6.835	0.465
S12	6.700	7.059	−0.359
W7	6.260	6.054	0.206
W11	6.150	6.393	−0.243
Z5	6.430	7.344	−0.914
Z7	8.000	8.005	−0.005
P10	6.000	6.272	−0.272
P16	6.210	6.421	−0.211
P19	4.770	5.169	−0.399
P32	6.890	6.650	0.240
L731927	6.300	6.388	−0.088
R_{pred}^2		0.839	
SEP		0.384	
ARE_{pred}		4.942%	
k		0.981	
k'		1.016	
$\lvert R_0^2 - R_0'^2 \rvert$		0.0257	

[a] R_{pred}^2: coefficient of multiple determination of external prediction. SEP: standard error of external prediction. ARE_{pred}: average relative error of external prediction. k and k': Golbraikh–Tropsha's slopes.

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

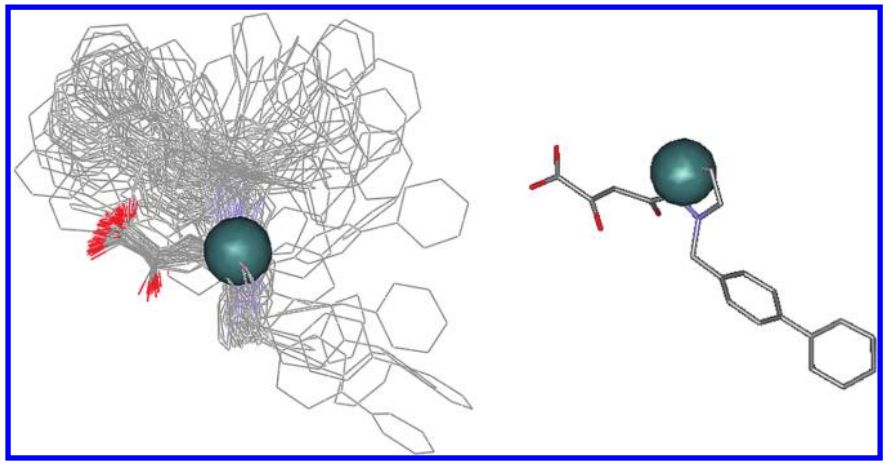


Figure 5. Descriptor 20.17.19.Zn^{2+}.C (represented in green) in three-dimensional space on the CEP (wire format) of L731942. Due to the large conformational flexibility of the inhibitor, a single conformation (stick format) is shown for clarificaton.

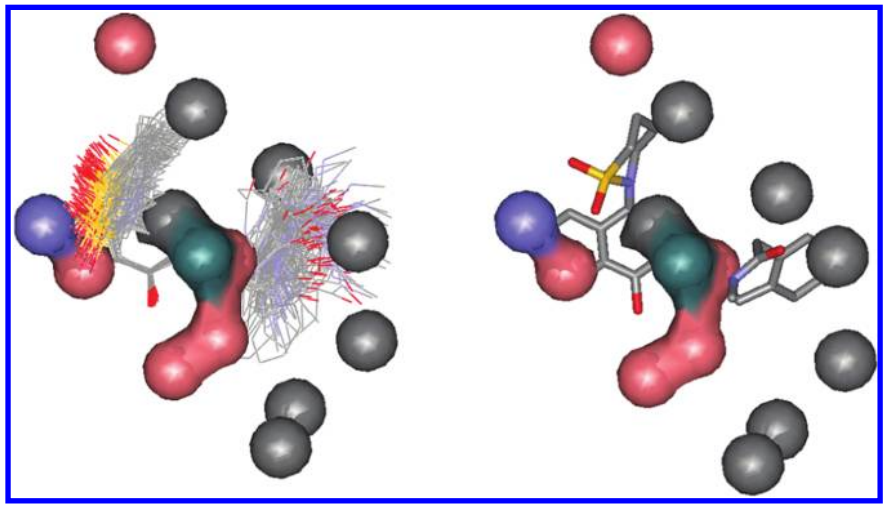


Figure 6. Full model in three-dimensional space surrounding the CEP (wire format) and a single conformation (sticks format) of inhibitor **E5b**: (blue) positive Zn^{2+}.C descriptor; (green) negative Zn^{2+}.C descriptor; (gray) positive descriptor Ar(C−H).LJ; (pink) negative descriptor Ar(C−H).LJ.

low associated SEP. The results indicate that the ability for external and internal prediction are also equivalent. The result of ARE_{pred} shows a low error potential. The results for the Golbraikh−Tropsha's slopes k and k' are located within the proposed range (0.85 at 1.15), and the parameter ($|R_0^2 - R_0'^2|$) is below 0.3, as proposed by Tropsha, Gramatica, and Gombar.[63]

DISCUSSION

In a QSAR study, it is always desirable to obtain an interpretable model that is able to relate the characteristics represented by the selected molecular descriptors to the end point under study.[74] One of the advantages of the 4D-QSAR model, compared to other grid-based methods, is the number of generated descriptors: the 3D-QSAR approach normally gives origin to models with hundreds of descriptors. Thus, the models must necessarily be built using the PLS regression method.[24] Moreover, the 4D-QSAR approach, including LQTA-QSAR, can generate models with much smaller amounts of descriptors.[13,15,75] This characteristic considerably facilitates the interpretation of results, including the possibility of individual interpretation for each descriptor, and also allows the construction of models with unbiased regression methods, such as multiple linear regression (MLR).

In the previous 2D-QSAR study,[10] it was possible to obtain a mechanistic interpretation relating the selected descriptors with a nucleophilic attack on metallic cofactors located in the HIV-1 IN catalytic triad (E_{HOMO} and α_{yy}), through interaction with the disordered hydrophobic loop (*SeaC2C2aa*), and with the importance of conformational stability for the binding of the inhibitor in the HIV-1 IN (E_T). All these interpretations could be supported by the literature.[76−83] Obviously, a 2D-QSAR study has major differences when compared to 4D-QSAR (2D vs 4D). In this case, a major difference is that the 2D data set (a subset of the 4D) has a low structural variability, unlike the current one. Another important difference to note is that the molecular descriptors dependent on the 3D optimized structures (E_{HOMO}, α_{yy}, and E_T) were obtained based on a single low energy conformation, as well as the 3D-QSAR descriptors.[24] On the other hand, in 4D-QSAR studies, the descriptors are based on the CEP obtained by molecular dynamics and, therefore, they represent not only the possibility of a specific conformation interaction with a binding site, but also the possibility to estimate how the spatial features arising from conformational flexibility may be important for a specific biological activity.[13−16] The 4D-QSAR methods are also capable of incorporating ligand conformational flexibility, exploring multiple alignments, evaluating ligand-embedded pharmacophore groups as part of QSAR

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

model building and optimization processes and proposing an active ligand conformation.[84] Besides, since it was possible to obtain a mechanistic interpretation in the first study,[10] it is expected that the same happens for the 4D-QSAR model.

The field descriptors are obtained by measuring the energy of interaction between a probe and all atoms of each molecule of the investigated set in each point of a three-dimensional grid that corresponds to a rigid hypothetical receptor. Thus, it can be related to possible drug-receptor interactions.[15] In the model and in Table 2, it is possible to see a greater number of descriptors Ar(C−H) (12) than descriptors Zn^{2+} (2). In three-dimensional space, the Zn^{2+} descriptors represent the metal cofactors, and they are positioned close to each other (6.60 Å). Interestingly, this result is equivalent to others available in the literature. An electron spin resonance (ESR) spectroscopy study using benzopiruvic acid in the presence of Mg^{2+} led to the proposal of an intermetallic distance between 4.00 and 6.00 Å.[85] In the theoretical study of Pais et al.,[86] a distance of 7.00 Å between them was proposed.

Despite this good relationship with independent results from other authors,[85,66] it is possible that the Coulomb descriptors do not necessarily encode information about the metal−inhibitor interactions. The negative sign (−0.138) of the autoscaled coefficient (last column in Table 2) for the 20.17.19.Zn^{2+} descriptor indicates that, despite the proximity to the keto-enol carbonyl (1.00 Å), this descriptor is detrimental to the interaction. The CEP for **L731942** (pIC_{50} = 5.12) (Figure 5), a low potency compound, shows that the side chain is quite flexible and can occupy the point x = 20, y = 17, and z = 19 of the grid. Thus, the information encoded by this descriptor may be an unfavorable inhibitor−metal interaction. Since the descriptor 14.15.21.Zn^{+2}.LJ, with positive sign (0.235), is located 3.62 Å from the front of the **E5b** inhibitor (Figure 6, blue descriptor) (pIC_{50} = 8.150) and 3.66 Å from the **L731942** inhibitor and in both profiles this descriptor is close to the hydrophobic groups (tetrahydro-thiopyrane in **E5b** and the aromatic side chain in **L731942**), it is feasible to propose that the occupation of this coordinate is related to the increase of potency. Figure 5 and all other figures that follows were built by using the software Accelrys Viewer Lite 4.2.[87]

20.17.19.Zn^{2+}.C and 14.15.21.Zn^{2+}.LJ may also be related to ligand−receptor interactions, unrelated to interactions with cofactors or to the disorderly loop. Some studies have suggested that other interactions can occur exactly in the "front" region of the inhibitors. For instance, Healy et al.[88] proposed for a naphthyridine structurally similar to **E5b** the occurrence of van der Waals interactions with residues His67 and Glu92. Figures 5 and 6 are also useful for comparing the degree of flexibility between a compound with a high value of pIC_{50} (i.e., high inhibitory potency) and with a low value of pIC_{50}. Conformationally less stable compounds (such as **L731942**) are probably less prone to the formation of stable interactions with the enzyme, which would make them less powerful. In the 2D-QSAR study, similar information was given by the descriptor E_T. Conformational flexibility may also be considered a feasible explanation for the descriptors scattered around the inhibitors, and not concentrated only in regions close to the aromatic side chain and the DKA substructure.

The flexibility of the side chain is probably the reason for the number of descriptors Ar(C−H) selected (twelve). Even in the less flexible compounds, such as **E5b**, this region has considerable conformational variation. Eight descriptors Ar(C−H) were concentrated in this region (Figure 7). The most important

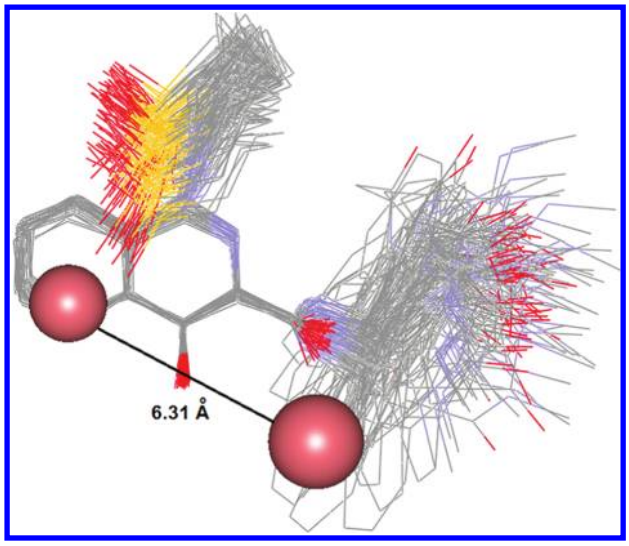


Figure 7. Descriptors 21.15.16.Ar(C−H).LJ and 15.15.18.Ar(C−H).LJ in three-dimensional space around the inhibitor E5b, and the distance between them. See the legend from Figure 6 for the color convention.

descriptor of the model, 26.18.14.Ar(C−H).LJ lies in this region, something which may be due to the occurrence of hydrophobic aromatic or π-stacking interactions with the disordered loop. In the 2D-QSAR model previously published,[10] the descriptor *SeaC2C2aa* was related to interactions of this nature.

A more interesting interpretation may be proposed for the descriptors Ar(C−H), specifically 21.15.16.Ar(C−H).LJ. This descriptor is positioned very close to the keto-enol group (1.88 Å of carbonyl and 3.61 Å of hydroxyl). Its autoscaled coefficient (−0.324) indicates that the inhibitory potency is favored when this coordinate is occupied. Besides its importance to the model (it has the second highest coefficient, in absolute value) and its proximity to the keto-enol group, this descriptor is located 6.32 Å away from descriptor 15.15.18.Ar(C−H).LJ, near the atom X (X = N, O) of the DKA substructure, which also has a negative coefficient (−0.179). Such a distance is in the range proposed by Maurin et al.[85] as a possible distance between the metal cofactors (4.00−6.00 Å). Thus, it is possible to propose that these two descriptors are really describing the interactions of the inhibitors with the metallic cofactors.

Although the model is the result of a receptor independent 4D-QSAR (RI-4D-QSAR), Martins et al.[15] have shown that a simple overlap of the selected descriptors with the binding site under study can aid in the interpretation of results. In Figure 8, one can observe that the descriptors override the secondary structure of HIV-1 IN (right) and possibly with only the most important amino acids for the inhibition (left).[11,88−93] The descriptors are mainly located close to amino acid residues responsible for catalytic activity (Asp64, Asp116, and Glu152) and the corresponding region in the disordered loop. This strengthens the mechanistic interpretation proposed for the model and shows that it is possible to generate a RI-4D-QSAR model directly related to a mechanism of enzymatic inhibition without the need for MD studies of the protein, which would be required in case of a Receptor Dependent 4D-QSAR (RD-4D-QSAR) study. Considering the computational cost and time it would take for a large set as used in this study, this indicates that the approach LQTA-QSAR can be of great help as a support tool in CADD studies.

dx.doi.org/10.1021/ci300039a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

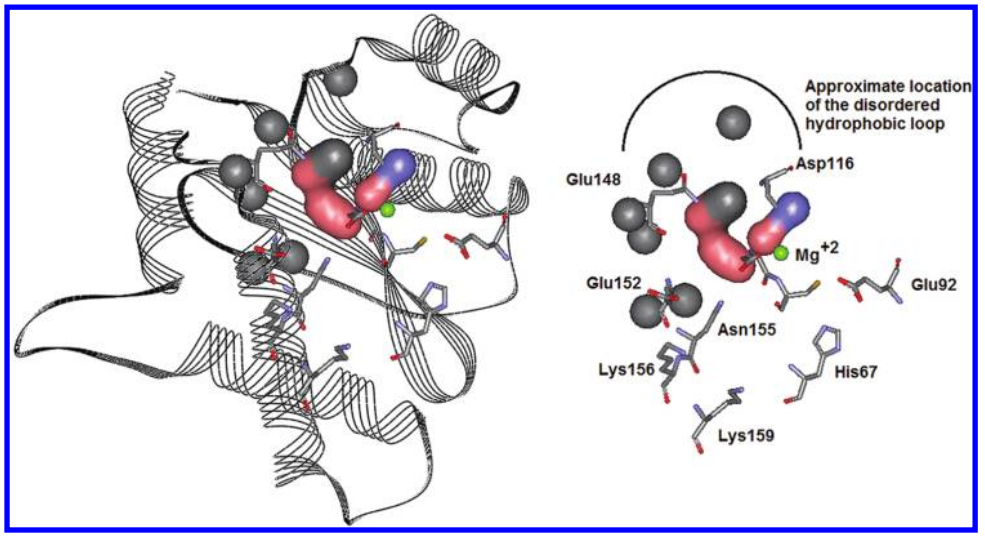


Figure 8. Plot of the descriptors on the binding site of HIV-1 IN (PDB 1QS4; 5CITEP were removed). The hydrophobic loop is represented only as a region (right), because it has not been solved in the crystal structure available. The overlay, based on the position of the ion Mg^{2+}, was performed using HyperChem 7.1.[25] See the legend of Figure 6 for the color convention.

Whereas the LQTA-QSAR uses the Coulomb potential and Lenard-Jones,[15] as it is done in 3D-QSAR studies (more specifically in CoMFA),[24] it is possible to compare the results to those obtained by Raghavan et al.[70] and Kuo et al.,[94] who also used INSTI to build models. In the first study, the descriptors from the 3D model obtained for a set of flavonoids is related to the importance of positive charge located in an area equivalent to the DKA substructure. In the second model, developed for a set of thiazolothiazepines, the selected electrostatic and steric fields correspond to regions filled by the amino acids Asp64 and Asp116 in the complex crystallographic 1QS4, to which the Mg^{2+} ion is coordinated. The results of these two studies corroborate the proposals made by the model in the present study. Both described regions of high electron density, near the binding region with the metallic cofactors that tend to favor the power of inhibitors, may be compared to the intended meaning of the descriptors 15.15.18.Ar(C−H).LJ, 20.17.19.Zn^{2+}.C, and 21.15.16.Ar(C−H).LJ. The regions close to the aromatic ring of the model from Kuo et al. [94] may be compared to the descriptors 26.18.14.Ar(C−H).LJ, 26.20.19.Ar(C−H).LJ, and 24.13.12.Ar(C−H).LJ, since these are also positive, form a fairly large surface (Figure 8), and are also close to the aromatic side chain.

CONCLUSION

The study has built and interpreted, from the mechanistic point of view, a 4D-QSAR model for a data set with 85 compounds described as INSTI that have in common the DKA substructure. For this, a new methodology named LQTA-QSAR was used. The study resulted in a model with good internal and external statistical quality, including robustness and absence of chance correlation. But the highlight is the mechanistic interpretation, because it has shown a direct relationship to the most accepted mechanism of action for INSTI, confirming that the descriptors calculated based on CEP generated by MD are sensitive enough to accumulate the information that is relevant to an action mechanism, even without the use of a biological target. This interpretation is corroborated by other studies, including the 2D-QSAR study previously published by the present authors, but mainly by the 3D-QSAR studies which also used descriptors based on Coulomb and Lennard-Jones potentials.

The results presented here encourage the continued use of this approach in the evaluation of other data sets, aimed at developing new drugs and at providing a greater understanding of similar mechanisms of action. The software LQTA-QSAR is freely available for evaluation by the scientific community at http://lqta.iqm.unicamp.br.

ASSOCIATED CONTENT

Supporting Information

Structures of data set and the used validation tools, the structures of the outliers, and the dendrogram used to select the test set. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: marcia@iqm.unicamp.br

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

E.B.d.M. acknowledges the MCT/CNPq/Fundacao Araucaria (www.fundacaoaraucaria.org.br) for financial support (under Protocol 2010/7354). M.M.C.F. acknowledges the Sao Paulo Research Foundation (Fapesp, www.fapesp.br/en) for financial support (under Protocol 2004/04686-5). The authors thank Prof. Dr. Carol H. Collins for English revision.

REFERENCES

(1) Hammer, S. M.; Eron, J. J., Jr.; Reiss, P.; Schooley, R. T.; Thompson, M. A.; Walmsley, S.; Cahn, P.; Fischl, M. A.; Gatell, J. M.; Hirsch, M. S.; Jacobsen, D. M.; Montaner, J. S.; Richman, D. D.; Yeni, P. G.; Volberding, P. A. Antiretroviral treatment of adult HIV infection—2008 Recommendations of the International AIDS Society USA Panel. *J. Am. Med. Assoc.* **2009**, *300*, 555−570.

(2) Nikolopoulos, G.; Bonovas, S.; Tsantes, A.; Sitaras, N. M. HIV/AIDS: Recent advances in antiretroviral agents. *Mini-Rev. Med. Chem.* **2009**, *9*, 900−910.

(3) Serrao, E.; Odde, S.; Ramkumar, K.; Neamati, N. Raltegravir, Elvitegravir, and Metoogravir: The birth of "me-too" HIV-1 integrase inhibitors. *Retrovirology* **2009**, *6* (25).
(4) UNAIDS/WHO. Epidemic update. In *Global Report: UNAIDS Report on the Global AIDS Epidemic 2010*; UNAIDS: Geneva, 2011; pp 16−61.
(5) Delelis, O.; Carayon, K.; Saïb, A.; Deprez, E.; Mouscadet, J.-F. Integrase and integration: Biochemical activities of HIV-1 integrase. *Retrovirology* **2009**, *5* (114).
(6) Cocohoba, J.; Dong, B. J. Raltegravir: The first HIV integrase inhibitor. *Clin. Ther.* **2009**, *30*, 1747−1765.
(7) Csizmadia, I. G.; Enriz, R. D. The role of computational medicinal chemistry in the drug discovery process. *J. Mol. Struct.: THEOCHEM* **2000**, *504*, ix−x.
(8) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, 2009; Vol. *1*, 967 pp.
(9) Nunthaboot, N.; Pianwanit, S.; Parasuk, V.; Kokpol, S.; Briggs, J. M. Computational studies of HIV-1 integrase and its inhibitors. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 160−190.
(10) Melo, E. B.; Ferreira, M. M. C. Multivariate QSAR Study of 4,5-dihydroxypyr-imidine carboxamides as HIV-1 Integrase inhibitors. *Eur. J. Med. Chem.* **2009**, *44*, 3577−3583.
(11) Barreca, M. L.; De Luca, L.; Ferro, S.; Rao, A.; Monforte, A.; Chimirri, A. Computational and synthetic approaches for the discovery of HIV-1 integrase inhibitors. *ARKIVOC* **2006**, *7*, 224−244.
(12) Petrocchi, A.; Koch, U.; Matassa, V. G.; Pacini, B.; Stillmock, K. A.; Summa, V. From dihydroxypyrimidine carboxylic acids to carboxamide HIV-1 integrase inhibitors: SAR around the amide moiety. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 350−353.
(13) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(14) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in drug design. *Molecules* **2010**, *15*, 3281−3294.
(15) Martins, J. P. A.; Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C. LQTA-QSAR: A new 4D-QSAR methodology. *J. Chem. Inf. Model.* **2009**, *49*, 1428−1436.
(16) Vedani, A.; Dobler, M. Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct.-Act. Relat.* **2002**, *21*, 382−390.
(17) Summa, V.; Petrocchi, A.; Matassa, V. G.; Gardelli, C.; Muraglia, E.; Rowley, M.; Paz, O. G.; Laufer, R.; Monteagudo, E.; Pace, P. 4,5-Dihydroxypyrimidine carboxamides and N-alkyl - 5-hydroxypyrimidinone carboxamides are potent, selective HIV integrase inhibitors with good pharmacokinetic profiles in preclinical species. *J. Med. Chem.* **2006**, *49*, 6646−6649.
(18) Guare, J. P.; Wai, J. S.; Gomez, R. P.; Anthony, N. J.; Jolly, S. M.; Cortes, A. R.; Vacca, J. P.; Felock, P. J.; Stillmock, K. A.; Schleif, W. A.; Moyer, G.; Gabryelski, L. J.; Jin, L.; Chen, I. W.; Hazuda, D. J.; Young, S. D. A series of 5-aminosubstituted 4-fluorobenzyl-8-hydroxy-[1,6]-naphthyridine-7-carboxamide HIV-1 integrase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2900−2904.
(19) Egbertson, M. S.; Moritz, M.; Melamed, J. Y.; Han, W.; Perlow, D. S.; Kuo, M. S.; Embrey, M.; Vacca, J. P.; Zrada, M. M.; Cortes, A. R.; Wallace, A.; Leonard, Y.; Hazuda, D. J.; Miller, M. D.; Felock, P. J.; Stillmock, K. A.; Witmer, M. V.; Schleif, W.; Gabryelski, L. J.; Moyer, G.; Ellis, J. D.; Jin, L.; Xu, W.; Braun, M. P.; Kassahun, K.; Tsou, N. N.; Young, S. D. A potent and orally active HIV-1 integrase term inhibitor. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1392−1398.
(20) Hazuda, D. J.; Witmer, P. F. M.; Wolfe, A.; Stillmock, K.; Grobler, J. A.; Espeseth, A.; Gabryelski, L.; Schleif, W.; Blau, C.; Miller, M. D. Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science* **2000**, *287*, 646−650.
(21) Zhuang, L.; Wai, J. S.; Embrey, M. W.; Fisher, T. E.; Egbertson, M. S.; Payne, L. S.; Guare, J. P., Jr; Vacca, J. P.; Hazuda, D. J.; Felock, P. J.; Wolfe, A. L.; Stillmock, K. A.; Witmer, M. V.; Moyer, G.; Schleif, W. A.; Gabryelski, L. J.; Leonard, Y. M.; Lynch, J. J., Jr; Michelson, S. R.; Young, S. D. Design and synthesis of 8-hydroxy-[1,6]-naphthyridines as novel inhibitors of HIV-1 integrase in vitro and in infected cells. *J. Med. Chem.* **2003**, *46*, 453−456.
(22) Wai, J. S.; Egbertson, M. S.; Payne, L. S.; Fisher, T. E.; Embrey, M. V.; Tran, L. O.; Melamed, J. Y.; Langford, H. M.; Guare, J. P., Jr; Zhuang, L.; Grey, V. E.; Vacca, J. P.; Holloway, M. K.; Naylor-Olsen, A. M.; Hazuda, D. J.; Felock, P. J.; Wolfe, A. L.; Stillmock, K. A.; Schleif, W. A.; Gabryelski, L. J.; Young, S. D. 4-Aryl-2,4-dioxobutanoic acid inhibitors of HIV-1 integrase and viral replication in cells. *J. Med. Chem.* **2000**, *43*, 4923−4926.
(23) Hazuda, D. J.; Felock, P. J.; Hastings, J. C.; Pramanik, B.; Wolfe, A. L. Differential divalent cation requirements uncouple the assembly and catalytic reactions of human immunodeficiency virus type 1 integrase. *J. Virol.* **1997**, *71*, 7005−7011.
(24) 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.
(25) *HyperChem 7*, version 7.1; HyperCube Inc.: Gainesville, USA.
(26) *Cambridge Structural Database*, version 5.29−2007 + 1 update; Cambridge Crystalographic Data Centre: Cambridge, UK.
(27) *Gaussian03*, version 6.0; Gaussian Inc.: Wallingford, USA.
(28) Molfetta, F. A.; Bruni, A. T.; Rosseli, F. P.; Silva, A. B. F. A partial least squares and principal component regression study of quinone compounds with trypanocidal activity. *Struct. Chem.* **2007**, *18*, 49−57.
(29) Lameira, J.; Medeiros, I. G.; Reis, M.; Santos, A. S.; Alves, C. N. Structure activity relationship study of flavone compounds with anti-HIV-1 integrase activity. *Bioorg. Med. Chem.* **2006**, *14*, 7105−7112.
(30) Wan, J.; Zhang, L.; Yang, G.; Zhan, C. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2099−2105.
(31) Yan, X.; Xiao, H.; Gong, X.; Ju, X. A comparison of semiempirical and first principle methods for establishing toxicological QSARs of nitroaromatics. *J. Mol. Struct.: THEOCHEM* **2006**, *764*, 141−148.
(32) Basak, S. C.; Balasubramanian, K.; Gute, B. D.; Mills, D.; Gorczynska, A.; Roszak, S. Prediction of cellular toxicity of halocarbons from computed chemodescriptors: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1103−1109.
(33) Zhang, J.; Xiao, J. J. Theoretical studies on heats of formation for cubylnitrates using density functional theory B3LYP method and semiempirical MO methods. *Int. J. Quantum Chem.* **2002**, *86*, 305−312.
(34) Van der Spoel, 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.; van Drunen, R.; Berendsen, H.J.C. *Gromacs User Manual Version 3.3*. ftp://ftp.gromacs.org/pub/manual/manual-3.3.pdf (accessed May 23, 2012).
(35) *Marvin*, version 4.1.8; Chemaxom Inc.: Budapest, HU.
(36) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems*; Wiley-Interscience: New York, 2001.
(37) Morley, C. *Open Babel*, version 2.1.1, 2006.
(38) Schüttelkopf, A. W.; Van Aalten, D. M. F. PRODRG: A tool for highthroughput crystallography of protein ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355−1363.
(39) Gasteiger, G.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(40) *HyperChem Computational Chemistry - Part 1: Pratical Guide*; HyperCube, Inc: Gainesville, 2002; pp 59−70.
(41) Sanner, M. F. Python: A programming language for software integration and development. *J. Mol. Graph. Model.* **1999**, *17*, 57−61.
(42) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of comparative molecular force field analysis models: Effects of data scaling and variable selection using a set of human synovial fluid phospholipase A2 inhibitors. *J. Med. Chem.* **1997**, *40*, 1136−1148.

(43) Knekta, E.; Andersson, P. L.; Johansson, M.; Tysklind, M. An overview of OSPAR priority compounds and selection of a representative training set. *Chemosphere* **2004**, *57*, 1495−1503.
(44) Barbosa, E. G.; Ferreira, M. M. C. Digital filters for molecular interaction field descriptors. *Mol. Inf.* **2012**, *31*, 75−84.
(45) 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. Chemom.* **2009**, *23* (1), 32−48.
(46) Hernández, N.; Kiralj, R.; Ferreira, M. M. C.; Talavera, I. Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 65−77.
(47) Melo, E. B. Multivariate SAR/QSAR of 3-aryl-4-hydroxyquinolin-2(1H)-one derivatives as type I fatty acid synthase (FAS) inhibitors. *Eur. J. Med. Chem.* **2010**, *45*, 5817−5826.
(48) Melo, E. B.; Martins, J. P. A.; Jorge, T. C. M.; Friozi, M. C.; Ferreira, M. M. C. Multivariate QSAR Study on the antimutagenic activity of flavonoids against 3-NFA on salmonella typhimurium TA98. *Eur. J. Med. Chem.* **2010**, *45*, 4562−4596.
(49) Melo, E. B. A new quantitative structure-property relationship model to predict bioconcentration factors of polychlorinated biphenyls (PCBs) in fishes using E-state index and topological descriptors. *Ecotoxicol. Environ. Saf.* **2011**, *75*, 213−222.
(50) Silla, J. M.; Nunes, C. A.; Cormanich, R. A.; Guerreiro, M. C.; Ramalho, T. C.; Freitas, M. P. MIA-QSPR and effect of variable selection on the modeling of kinetic parameters related to activities of modified peptides against dengue type 2. *Chemom. Intell. Lab. Syst.* **2011**, *108*, 146−149.
(51) *Pirouette*, version 4.0; Infometrix Inc.: Bothel, USA.
(52) Ferreira, M. M. C. Multivariate QSAR. *J. Braz. Chem. Soc.* **2002**, *13*, 742−745.
(53) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(54) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1374.
(55) Kiralj, R.; Ferreira, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and applications. *J. Braz. Chem. Soc.* **2009**, *20*, 770−787.
(56) Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; Van de Waterbeemd, H., Ed.; Wiley-VCH: Weinheim, 1998; pp 309−318.
(57) Gáudio, A. C.; Zandonade, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. *Quím. Nova* **2001**, *24*, 658−671.
(58) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(59) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its Variants in QSAR/QSPR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.
(60) Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(61) Melagraki, G.; Afantitis, A.; Sarimveis, H.; Koutentis, P. A.; Markopolus, J.; Igglessi-Markopoulou, O. Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening. *J. Comput. Aided Mol. Des.* **2007**, *21*, 251−267.
(62) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998.
(63) 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*, 69−77.
(64) Kiralj, R.; Ferreira, M. M. C. Is your QSAR/QSPR descriptor real or trash? *J. Chemom.* **2010**, *24*, 681−693.
(65) Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.
(66) Van Drie, J. H. Monty Kier and the origin of the pharmacophore concept. *Int. Electron. J. Mol. Des.* **2007**, *6*, 271−279.
(67) Masunari, A.; Tavares, L. C. A new class of nifuroxazide analogs: Synthesis of 5-nitrothiophene derivatives with antimicrobial activity against multidrug-resistant staphylococcus aureus. *Bioorg. Med. Chem.* **2007**, *15*, 4229−4236.
(68) Vedani, A.; Dobler, M. Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct.-Act. Relat.* **2002**, *21*, 382−390.
(69) Sant'Anna, C. M. R. Glossário de termos usados no planejamento de fármacos (Recomendações da IUPAC para 1997). *Quím. Nova* **2002**, *25*, 505−512.
(70) Raghavan, K.; Buolamwini, J. K.; Fesen, M. R.; Pommier, Y.; Kohn, K. W.; Weinstein, J. N. Three-dimensional quantitative structure-activity relationship (QSAR) of HIV integrase inhibitors: A comparative molecular field analysis (CoMFA) study. *J. Med. Chem.* **1995**, *38*, 890−897.
(71) Debnath, A. K. Application of 3D-QSAR techniques in anti-HIV drug design − An overview. *Curr. Pharm. Des.* **2005**, *11*, 3091−3110.
(72) Kubinyi, H. QSAR and 3D QSAR in drug design. part 1: Methodology. *Drug Discov. Today* **1997**, *2*, 457−467.
(73) Roy, P. P.; Leonard, J. T.; Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 31−42.
(74) OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD: Paris, 2007.
(75) Ghasemi, J. B.; Safavi-Sohi, R.; Barbosa, E. G. 4D-LQTA-QSAR and docking study on potent gram-negative specific LpxC inhibitors: a comparison to CoMFA modeling. *Mol. Divers.* **2012**, *16*, 203−213.
(76) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.
(77) Miller, K. J.; Savchik, J. A. A new empirical method to calculate average molecular polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(78) Marvin User's Guide. Calculator Plugins, Charge Plugin (2008). http://www.chemaxon.com/marvin/help/calculations/chargegroup.html (accessed September 2008).
(79) Morley, J. O.; Matthews, T. P. Structure-activity relationships in nitrothiophenes. *Bioorg. Med. Chem.* **2006**, *14*, 8099−8108.
(80) Silakari, P.; Shrivastava, S. D.; Silakari, G.; Kohli, D. V.; Rambabu, G.; Srivastava, S.; Silakari, O. QSAR Analysis of 1,3-diaryl-4,5,6,7-tetrahydro-2H-isoindole derivatives as selective COX-2 inhibitors. *Eur. J. Med. Chem.* **2008**, *43*, 1559−1569.
(81) Philips, O. A.; Udo, E. E.; Samuel, S. M. Synthesis and structure-antibacterial activity of triazolyl oxazolidinones containing long chain acyl moiety. *Eur. J. Med. Chem.* **2008**, *43*, 1095−1104.
(82) Lohray, B. B.; Gandhi, N.; Srivastava, B. K.; Lohray, V. B. 3D QSAR studies of N-4-arylacryloylpiperazin-1-yl-phenyl-oxazolidinones: A novel class of antibacterial agents. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3817−3823.
(83) Toit, K.; Elgorashi, E. E.; Malan, S. F.; Drewes, S. E.; Van Staden, J.; Croueh, N. R.; Mulholland, D. A. Anti-inflammatory activity and QSAR studies of compounds isolated from hyacinthaceae species and tachiadenus longiflorus griseb. (Gentianaceae). *Bioorg. Med. Chem.* **2005**, *13*, 2561−2568.
(84) Pita, S. S. R.; Albuquerque, M. G.; Rodrigues, C. R.; Castro, H. C.; Hopfinger, A. J. Receptor-dependent 4D-QSAR analysis of peptidemimetic inhibitors of *Trypanosoma cruzi* trypanothione reductase with receptor-based alignment. *Chem. Biol. Drug. Des.* **2012**, *79*, 740−748.
(85) Maurin, C.; Bailly, F.; Buisine, E.; Vezin, H.; Mbemba, G.; Mouscadet, J. F.; Cotelle, P. Spectroscopic studies of diketoacids-metal interactions. A probing tool for the pharmacophoric intermetallic distance in the HIV-1 integrase active site. *J. Med. Chem.* **2004**, *47*, 5583−5586.
(86) Pais, G. C. G.; Zhang, X.; Marchand, C.; Neamati, N.; Cowansage, K.; Svarovskaia, E. S.; Pathak, V. K.; Tang, Y.; Nicklaus, M.; Pommier, Y.;

Burke, T. R., Jr Structure activity of 3-aryl-1,3-diketo-containing compounds as HIV-1 integrase inhibitors. *J. Med. Chem.* **2002**, *45*, 3184−3194.
(87) *ViewerLite*, version 4.2; Accelrys Inc.: San Diego, USA.
(88) Healy, E. F.; Sanders, J.; King, P. J.; Robinson, W. E., Jr. A docking study of L-chicoric acid with HIV-1 integrase. *J. Mol. Graph. Model.* **2009**, *27*, 584−589.
(89) Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: A platform for antiviral drug design. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13040−13043.
(90) Sotriffer, C. A.; Ni, H.; McCammon, J. A. HIV-1 Integrase inhibitors interactions at the active site: Prediction of binding modes unaffected by crystal packing. *J. Am. Chem. Soc.* **2000**, *122*, 6136−6137.
(91) Ni, H.; Sotriffer, C. A.; McCammon, J. A. Ordered water and ligand mobility in the HIV-1 integrase-5CITEP complex: A molecular dynamics study. *J. Med. Chem.* **2001**, *44*, 3043−3047.
(92) Keserû, G. M.; Kolossvary, I. Fully flexible low-mode docking: application to the induced fit in HIV-integrase. *J. Am. Chem. Soc.* **2001**, *123*, 12708−12709.
(93) Chen, X.; Tsiang, M.; Yu, F.; Hung, M.; Jones, G. S.; Zeynalzadegan, A.; Qi, X.; Jin, H.; Kim, C. U.; Swaminathan, S.; Chen, J. M. Modeling, analysis, and validation of a novel HIV integrase structure provide insights into the binding modes of potent integrase inhibitors. *J. Mol. Biol.* **2008**, *380*, 504−519.
(94) Kuo, C. L.; Assefa, H.; Kamath, S.; Brzozowski, Z.; Slawinski, J.; Saczewski, F.; Buolamwini, J. K.; Neamati, N. Application of CoMFA and CoMSIA 3D-QSAR and docking studies in optimization of mercaptobenzenesulfonamides as HIV-1 integrase inhibitors. *J. Med. Chem.* **2004**, *47*, 385−399.