

2D and 3D QSAR studies of diarylpyrimidine HIV-1 reverse transcriptase inhibitors

Joseph Rebehmed · Florent Barbault ·
Cátia Teixeira · François Maurel

Received: 8 October 2007 / Accepted: 20 April 2008 / Published online: 28 May 2008
© Springer Science+Business Media B.V. 2008

Abstract 2D and 3D QSAR studies were applied on a set of 28 diarylpyrimidine derivatives to model and understand their HIV-1 reverse transcriptase (RT) inhibitory activities. Special cares were taken to build our set of molecules according to their bioactive conformations which is crucial to elaborate good QSAR models. 2D QSAR was performed using the heuristic method in CODESSA which had led to a linear model ($R^2 = 0.928$ and $s^2 = 0.015$) between the inhibitory activity and five descriptors. CoMFA and CoMSIA models were established using SYBYL package of programs. The better predictive ability of the CoMSIA model ($q^2 = 0.730$) over the CoMFA model ($q^2 = 0.597$) was assigned to the large contribution of hydrogen-bonding interactions to the inhibitory activity. CoMSIA physico-chemical properties are in agreement with the 2D QSAR descriptors. The CoMSIA PLS contour surfaces were mapped to the binding pocket of the RT and showed that the results obtained by the 2D and 3D models are in respect with the protein environment. This link permitted us to validate our model and give important insights for the structure activity interpretations. These results will guide further structural modification and prediction of new HIV-1 RT inhibitors.

Keywords HIV-1 · Reverse transcriptase ·
Diarylpyrimidine · QSAR · CoMFA · CoMSIA

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9217-4) contains supplementary material, which is available to authorized users.

J. Rebehmed · F. Barbault (✉) · C. Teixeira · F. Maurel
Université Paris Diderot (Paris 7), ITODYS (Interfaces,
Traitements, Organisation et Dynamique des Systèmes), CNRS
UMR 7086, 1 rue Guy de la Brosse, Paris 75005, France
e-mail: florent.barbault@univ-paris-diderot.fr

Introduction

AIDS, or acquired immunodeficiency syndrome is caused by the human immunodeficiency virus (HIV). AIDS has become a major worldwide epidemic with more than 40 million persons infected by 2005. The infection begins with the attachment of the virus to the cell surface of CD4 positive T lymphocytes. After binding, viral and cellular membranes fuse and the viral core is released into the cytoplasm of the cell. After this step, the retrovirus uses three main enzymes to spread its life cycle: (i) the reverse transcriptase (RT), a RNA-dependent DNA polymerase necessary to transcribe the viral genomic RNA into a proviral DNA [1]; (ii) the integrase, which is the enzyme responsible for insertion of the retrotranscribed DNA into the host cell genome [2]; and (iii) the protease, the enzyme necessary for the processing of new viral particles [3, 4].

Theoretically, all steps of the viral life cycle represent targets for antiretroviral therapy. However, Anti-AIDS therapy is actually based on three major groups of drugs: the nucleoside/nucleotide reverse transcriptase (NRTIs), the non-nucleoside reverse transcriptase (NNRTIs) and the protease inhibitors (PIs); these inhibitors mixed together make a highly active antiretroviral therapy. Recently, a fourth class of antiretrovirals started to be used clinically, with the introduction of enfuvirtide (36 amino acid residues peptide), the first fusion inhibitor [5, 6].

Reverse transcriptase inhibitors come out to be the first drug class with potent activity against HIV, inhibiting one the earliest steps in the viral life cycle. The use of the protease protein as a target is not beneficial in preventive therapies because PIs block a post-integration step of the viral cycle [7]. The reverse transcriptase has become an excellent target for attempts to stop the HIV proliferation

for several reasons: (i) it is a crucial enzyme in the viral replication cycle; (ii) its properties are quite different from those of the other cellular DNA polymerases; (iii) it is active in the cytoplasmic compartment of the infected cell, separate from the nuclear and mitochondrial DNA polymerases [8].

NRTIs possess limited therapeutic index and commonly more severe side effects in humans than NNRTIs [9]. The latter, a structurally diverse group of compounds [10, 11], have the advantages of high potency, low toxicity and excellent selectivity (they are highly selective for HIV-1 and do not inhibit HIV-2 or any other retrovirus). They inhibit the RT protein by binding to an allosteric binding site, a lipophilic cavity situated at 10 Å from the catalytic site [12]. The Diarylpyrimidine (DAPY) analogs consist on a class of NNRTIs [7, 13–15]; and like all NNRTIs, these compounds induce viral resistance [9]; also the similarities in chemical structures lead to the emergence of cross-resistance among members of the same class, where a real medical need to develop new generation of NNRTIs which do not give rise to cross-resistance and are effective against clinically relevant mutant strains [12].

In this article, we report 2D and 3D QSAR molecular modeling studies on this family of HIV-1 non nucleoside reverse transcriptase inhibitors that were based on structures and activities data of DAPY compounds reported in the literature [13]. Until now only one QSAR analysis was published by Thakur et al. [16]; their study was limited to a 2D QSAR analysis and they didn't take into consideration the bioactive conformation of these molecules that adopt a butterfly-like conformation at the binding site of the protein [17]. So their obtained model doesn't reflect the conformational space of the inhibitor in the protein and it might not be suitable to elaborate new inhibitors with improved activity.

In this work, one 2D QSAR model was elaborated at the beginning with Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) software package [18]. While 2D methods are especially assigned by the topological connectivity of the molecular structure thus neglecting the conformational space of the molecule [19], 3D QSAR model such as Comparative Molecular Field Analysis (CoMFA) [20, 21] were also made to complement the 2D QSAR model since the interaction between a ligand and a receptor is a 3D phenomenon. Comparative Molecular Similarity Indices Analysis (CoMSIA) [22, 23] were also computed to gain insight into how hydrophobic and hydrogen-bonding, in addition to steric and electrostatic interactions determined by the CoMFA model, influence the activity [3]. Our resulting 2D and 3D model will guide further structural modification and predict the potency and physicochemical properties of clinical drug candidates.

Methods and materials

Data set preparation

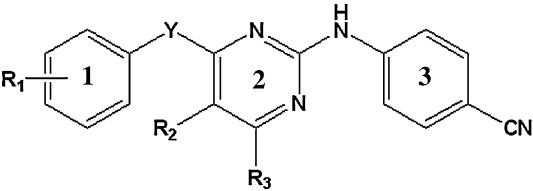
In preparing the data set, we selected data that had been determined by the same experimental conditions. 28 Structures and activities of DAPY (Diarylpyrimidine) analogs were extracted from the literature [13]. All compounds were tested for potency (IC₅₀, μM) to achieve 50% protection of MT-4 cells from the HIV-1 cytopathicity as determined by the MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] method [24]. It is a rapid and sensitive assay procedure to evaluate anti-HIV agents by assessing spectrophotometrically the viability of HIV virus *via* in situ reduction of the MTT molecule. The LAI strain of HIV-1 was the infecting virus. These values are presented in Table 1.

Tridimensional structure building

3D-QSAR analysis may be mainly influenced by a number of factors, such as conformers, alignment of the compounds and their orientation in the lattice [25, 26]. The X-ray structure of the HIV-1 RT protein in complex with DAPY 13A determined with a resolution of 3.00 Å was extracted from the Brookhaven Protein Databank (PDB code: 1S6Q) [27]. The conformation of the inhibitor (DAPY 13A) in the X-ray complex, assumed to be the bioactive conformation, was conserved and the structures of the entire diarylpyrimidine analogs set were built based on this geometry using the SYBYL version 7.3 molecular modeling package [28]. Energy minimization was carried out using the Tripos force field with 20 iterations of Simplex followed by Powell [29] minimization algorithm with a 0.05 kcal/mol energy gradient convergence criterion; partial atomic charges were attributed using the Gasteiger–Marsili method [28].

2D-QSAR Analysis

To obtain the 2D QSAR model, CODESSA software was used in this study [18]. Before calculation of the descriptors, a final geometry optimization of the molecules in the bioactive conformation was performed using the AM1 semi-empirical method [30] implemented in MOPAC program [31]. The resulting output files exported from MOPAC containing the refined geometry were served as input for CODESSA software to calculate the molecular descriptors [32]. A large number of molecular descriptors divided into five categories (constitutional, topological, geometrical, electrostatic and quantum-chemical) were generated. Only some of them are significantly correlated with the activity of inhibiting the HIV-1 RT enzyme. In

Table 1 Diarylpyrimidine scaffold with the different chemical substitutions and their respective HIV-1 RT IC50 values


Compound	Substituent				IC50 (μM)
	R ₁	Y	R ₂	R ₃	
13A	2,4,6-triMe	NH	–	–	0.0010
13B	2,6-diMe-4-CN	O	–	–	0.0011
13C	2,6-diMe-4-CN	NH	–	–	0.0004
13D	2,6-diMe-4-Br	O	–	–	0.0029
13E	2,6-diMe-4-Br	S	–	–	0.0057
13F	2,6-diMe-4-(C≡CH)	O	–	–	0.0055
13G	2,4,6-triMe	S	–	–	0.0036
13H	2,4,6-triMe	O	–	–	0.0029
13I	2,4-diBr-6-F	NH	–	–	0.0006
13J	2,4,6-triCl	NH	–	–	0.0007
13K	2,6-diMe	NH	–	–	0.0007
13L	2,4-diCl-6-Me	NH	–	–	0.0010
13N	2,6-diBr-4-Me	NH	–	–	0.0007
15A	2,6-diMe-4-CN	NH	Br	–	0.0004
15B	2,6-diMe-4-CN	O	Br	–	0.0014
15C	2,4,6-triMe	NH	Br	–	0.0055
16	2,4,6-triMe	NH	(C≡CH)	–	0.0042
17	2,4,6-triMe	NH	Vinyl	–	0.0025
18	2,4,6-triMe	NH	Ph	–	0.0240
19A	2,6-diMe-4-CN	NH	CN	–	0.0005
19B	2,4,6-triMe	NH	CN	–	0.0010
20A	2,6-diMe-4-CN	NH	Cl	–	0.0012
20C	2,4,6-triMe	NH	Cl	–	0.0027
21A	2,6-diMe-4-CN	NH	Me	–	0.0008
21B	2,4,6-triMe	NH	Me	–	0.0017
24	2,6-diMe-4-CN	O	NHAc	–	0.0019
25	2,6-diMe-4-CN	O	Br	NH ₂	0.0014
2A	2,6-diCl	C	–	NH ₂	0.0010

addition, many of the descriptors are highly intercorrelated. The use of a method to reduce the number of descriptors to an appropriate size is crucial to obtain a QSAR model with a good predictivity.

The heuristic method is a very useful tool for searching the best pool of descriptors. It is a quick method and presents no restrictions on the size of the data set. First preselection of descriptors is executed. All descriptors are checked to ensure that values of each descriptor are available for each structure and that there

is enough variation in these values. Descriptors that do not fulfill these conditions are eliminated. To reduce further the number, the following criteria are implemented and a descriptor is eliminated if the *F*-test value is less than 1 in the one-parameter regression and the *t*-value are less than the user-specified values [32, 33]. All highly intercorrelated descriptors ($R^2 > 0.8$) are also eliminated.

From the remaining list of descriptors, the best multi-parameter regression model with the optimum values of statistical criteria (the square correlation R^2 , the cross-validated q^2 and the *F*-test values) was determined [33, 34]. A major point in developing QSAR model is the number of descriptors used to elaborate the equation. The lack of an adequate control leads to over-correlated equations which contains an excess of descriptors and are difficult to interpret [35]. A simple technique to control the model expansion and to choose the smallest optimum number of descriptors with a significant quality of the regression equation is the 'breaking point' rule. It considers the improvement of the R^2 by addition of further descriptor to the model. If the improvement between the models *n* and *n* + 1 descriptors is negligible, then the optimum model will be the one with *n* descriptors [35–37]. Considering the optimum number of descriptors to use in the calculation, the QSAR equation, showing the contribution of each descriptor in the inhibitory activity, was determined. AMPAC software [38] is then used to analyze some descriptors obtained in the optimum model.

3D QSAR model

A quantitative structure-activity relationship (QSAR) relates numerical properties of the molecular structure to its biological activity by a mathematical model. The starting point for a 3D QSAR analysis is a set of conformations, one for each molecule in the set. All optimized structures were aligned by fitting them on the pyrimidine heterocycle as a common substructure. The molecular field surrounding each molecule is then calculated by placing the overlapped molecules in a 3D grid and using appropriate placed probe at each point on the lattice.

Comparative molecular field (CoMFA) model

The CoMFA methodology is a 3D QSAR technique which ultimately allows one to design and predict activities of molecules. After aligning all the molecules within a 3D cubic lattice with a grid spacing of 2.0 Å, steric and electrostatic fields are calculated by interacting a probe with each molecule at a series of grid points surrounding the aligned database in 3D space; a sp^3 -hybridized carbon with +1.0 net charge was used as a probe atom. The steric and

electrostatic fields were calculated at each lattice point using 6–12 Lennard–Jones and Coulombic potentials. A distance dependent dielectric constant was used (energy falls as $1/r^2$). To avoid too high and unrealistic energy values inside the molecule, a 30 kcal/mol energy cutoff was specified and the electrostatic fields were rejected at the lattice points with maximal steric interactions [20, 39].

The regression analysis was carried out using the partial least-squares (PLS) to derive a linear correlation between the CoMFA fields and the activity values of the reverse transcriptase inhibitors; the R^2 and q^2 values were measured. The R^2 is the Pearson correlation coefficient which is the correlation between the experimental activities and the predicted ones.

$$R^2 = \left(\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N} \right) \left(\sum Y^2 - \frac{(\sum Y)^2}{N} \right)}} \right)^2$$

where X is the experimental values, Y the predicted ones and N is the number of compounds.

The q^2 is the calculated value based on the leave-one-out cross-validation method [40]. This technique, recommended for checking the quality of regression model, consists of removing one of the values from the data set, deriving a regression model for the remainder and then predicting the values for the data left out. A value can then be predicted for the data left out and compared with the true observed value. This is repeated for every data point in the set and permits the calculation of a cross-validated q^2 value.

$$q^2 = 1 - \frac{\sum_{i=1}^N (y_{i,\text{observed}} - y_{i,\text{predicted}})^2}{\sum_{i=1}^N (y_{i,\text{observed}} - \bar{y}_{i,\text{observed}})^2}$$

During the cross-validation analysis, column filtering (σ) was set to 2.0 kcal/mol to speed up the calculation and to reduce the noise.

Comparative molecular similarity indices (CoMSIA) model

CoMSIA, developed by Klebe et al. [22, 23], is known as one of the newer 3D QSAR methods considering five physicochemical properties: steric, electrostatic, hydrogen bond donor, hydrogen bond acceptor and hydrophobic fields [41]. Taking the same aligned molecules and the same lattice box that were used for CoMFA, CoMSIA similarity indices [23], between the compounds of interest and the probe atom, were calculated at each lattice intersection by using the following formula:

$$A_{f,k}^q = - \sum_i \omega_{\text{probe},k} \omega_{ik} e^{\alpha r_{iq}^2}$$

where A is the similarity index at grid point q; i the summation index over all atoms of the molecule j under investigation; ω_{ik} the actual value of the physicochemical property k of atom i; $\omega_{\text{probe},k}$ is the probe atom with charge +1, radius 1 Å and hydrophobicity +1; and r_{iq} is the mutual distance between probe atom at grid point q and atom i of the test molecule. The attenuation factor α was fixed at 0.3. Like CoMFA analysis, the correlation between biological activities and descriptor values was evaluated by the method of partial least-square; R^2 and q^2 values were calculated [22].

Results and discussion

2D QSAR

By using the heuristic method, equation for DAPY analogs was constructed with 5 descriptors (see Table 2). This optimum number of descriptors was determined by using a simple “breaking point” rule [35–37]. Considering the plot of number of descriptors involved versus the squared correlation coefficient (R^2), it seems that the statistical amelioration of the model is higher at lower number of descriptors until one point (the breaking point) and after that the improvement is trifling. Considering this, the

Table 2 The multilinear QSAR model obtained with the heuristic method for the DAPY compounds with the CODESSA software ($R^2 = 0.928$, $q^2 = 0.844$, $F = 56.880$ and $s^2 = 0.015$) X, ΔX and

t-test are the regression coefficient of the linear model, standard errors of the regression coefficient and the t significance coefficient of the determination respectively

No	X	ΔX	t-test	Name of the descriptor
0	−2.98E+02	5.52E+01	−5.3889	Intercept
1	5.18E−02	4.32E−03	11.9953	ZX Shadow
2	−3.67E−01	3.51E−02	−10.4655	Number of N atoms
3	2.06E+01	3.83E+00	5.3736	Max Coulombic interaction for a C–N bond
4	1.05E+02	2.48E+01	4.2361	Min 1-electron react. index for a N atom
5	−5.92E−01	1.59E−01	−3.7265	LUMO+1 energy

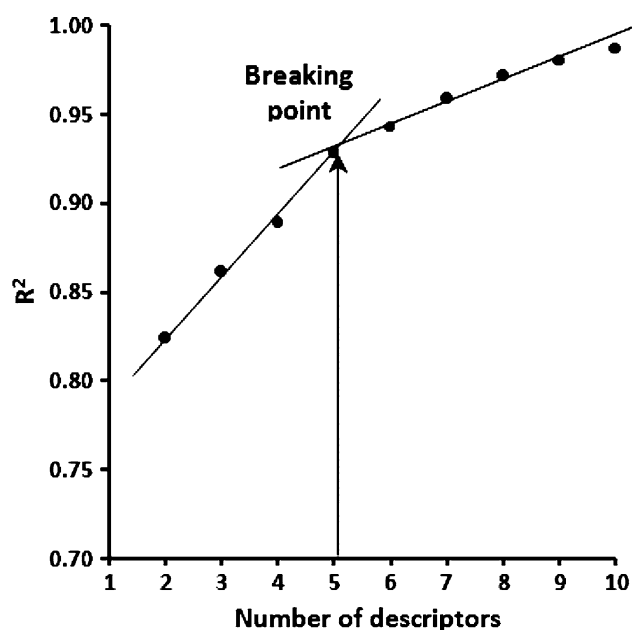


Fig. 1 The influence of number of descriptors on heuristic correlation R^2 coefficient

breaking point, shown in Fig. 1, corresponds to the optimum number of descriptors to be used in modeling the activity of interest.

The QSAR equation for DAPY analogs is characterized by the statistical parameters presented in Table 2. The values of these parameters are calculated for each descriptor. R^2 is the squared correlation coefficient; q^2 is the squared cross-validation coefficient; F is the Fisher's criterion and s^2 is the squared standard error. The value of each descriptor and the predicted log (IC₅₀) obtained with the QSAR model presented in Table 2, for all the DAPY compounds are respectively listed in Table 4 and Table 5 in the supplementary data. The predicted log (IC₅₀) values with the experimental ones are plotted with the linear regression trendline in Fig. 2.

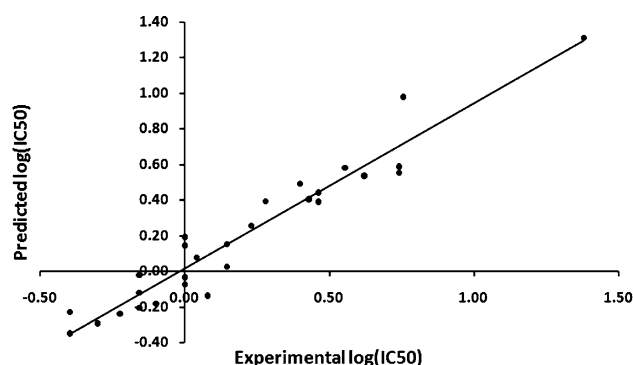


Fig. 2 The experimental and predicted values of log(IC₅₀) for DAPY analogs according to the model in Table 2

Discussion of the descriptors

The five types of descriptors involved in the equation of HIV-1 RT DAPY analogs inhibitors can be organized in: topological (D2), geometrical (D1), and quantum-chemical (D3, D4, D5). The t -test indicated the following order of significance for the descriptors included in the equation (Table 2): $D1 > D2 > D3 > D4 > D5$. The complex nature of the RT inhibition process had made the direct interpretation of the descriptors more difficult, where the needs to bond (or link) those descriptors to the physico-chemical properties behind the inhibitors interaction with the protein.

ZX Shadow. This geometrical descriptor is the most significant one. By the orientation of the molecule in the space along the axes of inertia (X coordinate is along the main axis of inertia and so on) the areas of the shadows S1, S2 and S3 of the molecule as projected on the XY , YZ and XZ planes, are calculated [42]. The geometrical indices ZX shadow reflects the overall shape of the molecule projected onto the plane ZX oriented with respect to its moments of inertia [43].

Axes of inertia of a molecule depend on its substituent. For example, DAPY 18 inhibitor contains a phenyl substituent on the middle ring and presents a different orientation from DAPY 13N (see Fig. 3) that can be reflected by their respective ZX shadows.

The size, the shape and the orientation of the inhibitor in the allosteric binding pocket of the reverse transcriptase are in part crucial for a good inhibition activity. This idea is confirmed by the X-ray structure of the complex RT with the experimental inhibitor DAPY 13A, where we can notice the tightness of the protein NNRTI binding site (see Fig. 7 right). This observation and the crucial role of the bioactive conformation of the inhibitory will be lately validated with the 3D QSAR fields.

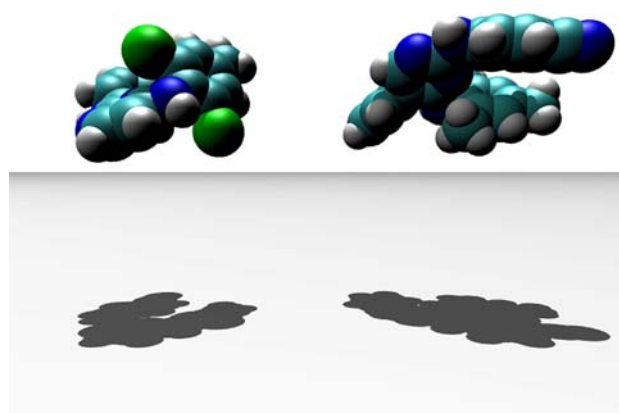


Fig. 3 The structure of DAPY 13N (left) and DAPY 18 (right) inhibitors with the presentation of their respective shadow on the ZX plane

Nitrogen atom related descriptors. According to the *t*-test values, the second important descriptor is the number of nitrogen atoms. It could be seen that the increase of the number of nitrogen atoms in the molecule leads to a better HIV-1 RT inhibitor activity. DAPY 13H, with four nitrogen atoms, has an activity of 2.9 nM; otherwise DAPY 13A that has one more nitrogen atom instead of the oxygen atom presents an activity of 1 nM. Nitrogen plays an essential role as hydrogen bond donor/acceptor in the interaction with the protein, thus increasing the inhibitory activity. However, increasing the number of nitrogen atoms in the molecule to more than six shows no significant difference in the inhibition activity.

The maximum Coulomb interaction for a C–N bond is a descriptor related to the electrostatic and Van der Waals interactions and which is correlated with the polarity of the carbon–nitrogen bond. In the equation, it has small positive regression coefficient and is calculated using the following formula:

$$CI_{\max, C-N} = E_{ee}(CN) + E_{ne}(CN) + E_{nn}(CN)$$

where $E_{ee}(CN)$ corresponds to the electronic repulsion between the two atoms, $E_{ne}(CN)$ is the nuclear-electron attraction energy between the 2 atomic species and $E_{nn}(CN)$ the nuclear repulsion energy. This descriptor is lightly intercorrelated with the number of N atom descriptor; it will permit to discriminate between two molecules having the same number of N but different C–N bond order.

Minimum one-electron reactivity index for an N atom is a quantum-chemical descriptor defined as followed:

$$RI_{\min, N}^{1e} = \sum_{i \in A} \sum_{j \in A} \frac{C_{iHOMO} C_{jLUMO}}{(\varepsilon_{LUMO} - \varepsilon_{HOMO})}$$

where the summations are performed over all atomic orbitals i, j at the given atom, C_{iHOMO} and C_{jLUMO} denote the i -th and j -th atomic orbital coefficients on the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), respectively, and ε_{LUMO} and ε_{HOMO} are the energies of these orbitals [44], which probably reflect involvement of the nitrogen function in the activity against the HIV-1 reverse transcriptase. The reactivity indices estimate the relative reactivity of the atoms in the molecule for a given series of compounds and are related to the activation energy of the corresponding chemical reaction.

LUMO+1. The final descriptor is the LUMO+1. All non-occupied orbitals, especially those of low energy i.e. LUMO and LUMO+1, characterize the ability of the molecule to interact with a nucleophilic reactant through a two electron stabilizing interaction. The visualization of these molecular orbitals on the molecules with AMPAC

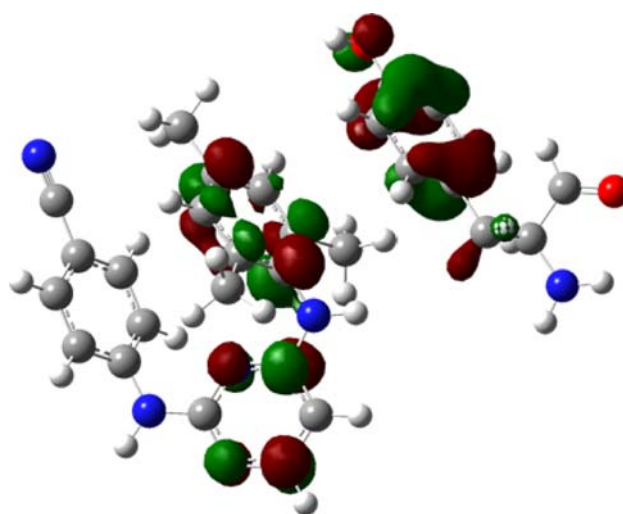


Fig. 4 π - π staking interaction between the inhibitor and one amino acid of the binding pocket. Molecular orbitals were calculated by AMPAC: (left) LUMO+1 of the DAPY 13A; (right) the HOMO of the tyrosine 181 amino acid

software was essential to understand the role of this descriptor in the inhibitory activity of DAPY analogs.

As it is shown in Fig. 4, the LUMO+1 molecular orbital of DAPY 13A (the experimental ligand) is localized on the ring 1 where most of the chemical modifications happen, however the LUMO is situated on the other two rings of the ligand. In the crystal structure of the complex, the ring 1 forms a stacking interaction with the aromatic ring of the tyrosine 181 amino acid that acts as a nucleophilic group. This two electrons interaction plays an essential role in stabilizing the complex. According to the chemical substituent of the ring 1 the localization of the LUMO+1 differs affecting thus the interaction with the protein.

Validation of the QSAR model

An important side of any QSAR calculation is the validation of the model. After the first cross-validation method based on the leave-one-out algorithm implemented in the CODESSA package, another internal validation was executed. It consists of predicting the property values for each one-third of the compounds with the model obtained for the remaining two-third of the compounds. This validation technique has been used by many research groups [37, 45, 46]. The entire parent set is divided into three subsets: the first, fourth, seventh, etc., molecules form the first subset; the second, the fifth, the eighth, etc., entries for, the second subset; and the third, sixth, ninth, etc form the third subset. Three training sets (S1, S2 and S3) were prepared by combining two subsets and the third one constitutes the test set. The correlation equation was derived for each training set with the same descriptors. Then the equation obtained

Table 3 Internal validation of the 2D QSAR model with the statistical characteristics

Training set	Number of molecules	R^2	s^2	Test set	Number of molecules	$R^2_{\text{predicted}}$	$s^2_{\text{predicted}}$
S1 + S2	19	0.932	0.019	S3	9	0.889	0.013
S1 + S3	19	0.959	0.010	S2	9	0.834	0.046
S2 + S3	18	0.950	0.015	S1	10	0.920	0.019
Average		0.947	0.015			0.881	0.026

for the training set was used to predict the $\log(\text{IC}_{50})$ values of the test set data. The results of the internal validation applied to our data are presented in Table 3

The internal validation shows that our model obtained by CODESSA is able to accurately predict the activity of the molecules in the test set; indeed all R^2 values are superior than 0.834.

CoMFA and CoMSIA analysis

One of the most essential points in a molecular field analysis is to have the bioactive conformation of the molecules known as the butterfly conformation for the DAPY analogs [27]. This information was given by the X-ray structure of the RT/DAPY 13A complex extracted from the PDB. This point was confirmed by parallel 3D QSAR studies that were realized starting with an extended conformation of the molecules. The resulting models do not possess a good predictive ability with correlation coefficients in the vicinity of zero.

Two 3D QSAR methods were applied to derive a statistically significant and highly predictive QSAR model: CoMFA and CoMSIA. For several reasons, only the CoMSIA model results will be presented and analyzed in the following. The CoMSIA approach considers three more similarities fields than the CoMFA: hydrophobic, hydrogen bond donor and hydrogen bond acceptor. The purpose of using more descriptors is not to increase the significance and the predictive power of the model by overcorrelating it. The goal is to partition the various properties into spatial locations where they play a decisive role in determining biological activity. The advantages of CoMSIA are the better ability to visualize and link the obtained correlation to the field contributions because it provides smoother and more interpretable contour maps as a results of using Gaussian type distance dependence with the molecular similarities indices it calculates [22]. In addition, many studies had shown that CoMSIA results are independent of the grid spacing and of the translations and rotations of the superimposed molecules with respect to the lattice [23, 25].

In CoMSIA, five physicochemical properties are taken into consideration: steric and electrostatic, hydrogen bond donor, hydrogen bond acceptor and hydrophobic fields. In former studies, it has been examined whether the five

different descriptors are totally independent of each other [23, 47, 48]; dependencies of the individual fields may reduce the statistical significance of the results and the predictivity of QSAR model [23, 49]. Taking this in consideration, all possible combinations of CoMSIA fields were calculated to elaborate a predictive model. Finally, 31 possibilities with their respective q^2 values were determined (see Fig. 5).

All these reasons were proved by the results of the best CoMSIA model (SEDA) with $q^2 = 0.730$ and $R^2 = 0.988$ (see Fig. 6) as compared to a $q^2 = 0.597$ and an $R^2 = 0.948$ for CoMFA.

The results were visualized by colored 3D contour plots. The CoMSIA steric contours are shown in Fig. 7. Green contours indicate regions where steric bulk is favored to increase activity. Yellow contours, presenting a surface enclosing almost all the inhibitor (not shown on the Fig. 7), indicate regions where steric bulk is disfavored. This result confirms the previous result obtained with the ZX shadow descriptor of the 2D QSAR model and both are in agreement with the protein environment (Fig. 7 and 10 left) where the size of the ligand and its conformation in the allosteric binding site of the protein are crucial for its inhibitory activity [27].

A substitution of the methyl (DAPY 13A, $\text{IC}_{50} = 1 \text{ nM}$) by a nitrile group (DAPY 13C, $\text{IC}_{50} = 0.4 \text{ nM}$) on the *para*-position of R1 increase the inhibitory activity of the molecule. The CN being longer than the CH_3 will allow the ligand having this group at the *para*-position of the ring 1 to have one extra H-bond interaction with the TYR188 amino acid (Fig. 10); it should be noted that we have just showed the amino acids at the allosteric binding pocket of the protein that are essential to analyze the different CoMSIA fields. The size of the group at this position can not be important because the hydrophobic pocket is limited. At the R2 position (see Table 1), the substitution is more critical because green and yellow contours overlap. A bromine can be added at this position while reserving the activity (DAPY 15a, $\text{IC}_{50} = 0.4 \text{ nM}$), but a bigger group as a phenyl ring (DAPY 18, $\text{IC}_{50} = 24 \text{ nM}$) induce a quick decreasing of the inhibitory activity of the ligand caused by a strong repulsive steric interaction with some amino acids of the protein including GLU138 and THR139 (Fig. 10). In fact,

Fig. 5 Results of the 31 possible CoMSIA field combinations (S = steric, E = electrostatic, H = hydrophobic, D = H-bond donor, A = H-bond acceptor) with their respective q^2 values (LOO cross-validation using the PLS method)

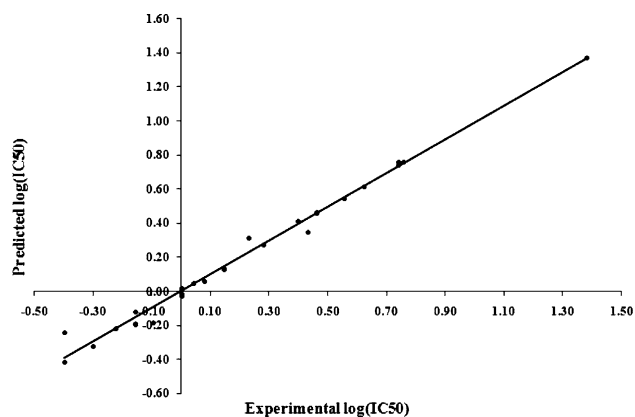
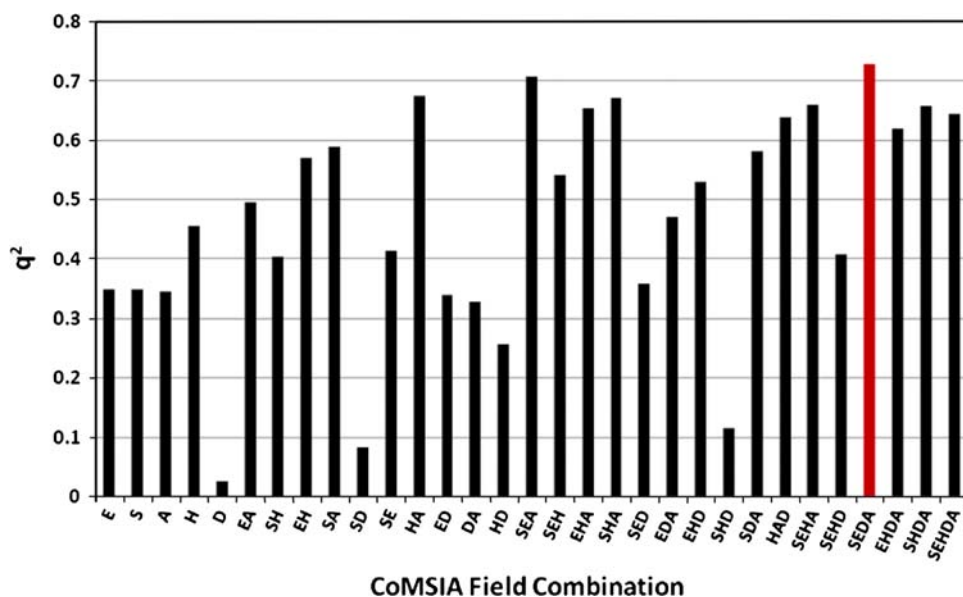


Fig. 6 The experimental and predicted values of $\log(\text{IC}_{50})$ for DAPY analogs according to the SEDA combination field CoMSIA model ($R^2 = 0.988$)

this part is located at the entrance of the binding pocket which is a bottleneck. A modification at this position is interesting since it adds an additional interaction that could stabilize the ligand into the binding pocket. However, any modifications must be carefully considered. An important bulked group may prevent the penetration of the ligand into the pocket.

The Fig. 8 shows the electrostatic maps contribution of the best CoMSIA model. The activity of a molecule with a more negative group at the R2 position like a nitrile (DAPY 19B, $\text{IC}_{50} = 1$ nM) is better than the activity of a molecule with an alkyne (ethynyl) group (DAPY 16, $\text{IC}_{50} = 4.2$ nM) at the same position. This property permits to discriminate between two groups of identical size but with different charges. In parallel, we deduced from the blue contour near the aromatic ring 1 that a positive charge

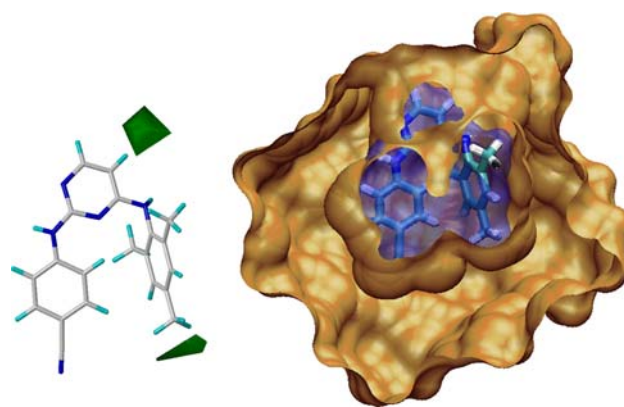


Fig. 7 (left) CoMSIA steric contours; green contours indicate regions where steric bulk is favored to increase activity. (right) The X-ray structure of RT protein allosteric binding pocket in complex with DAPY 13A inhibitor (Protein is shown in surface presentation and the ligand in licorice). Z-clipping was made to the structure to be able to see the ligand in the protein cavity

at this position increases the activity. The molecule DAPY 13J ($\text{IC}_{50} = 0.7$ nM) with a 2,4,6-tri chloro substitution has a better activity than DAPY 13A ($\text{IC}_{50} = 1$ nM) with 2,4,6-triMe. We can deduce that electron withdrawing groups at this position increase the inhibitory activity by turning the phenyl ring more electropositive. In the X-ray structure of the protein complexed with DAPY 13A, the orientation of the ring 1 and the aromatic ring of the TYR181 as well as the mean distance between these groupment suggests a π - π staking interaction (Fig. 10). This result is complementary to the analysis shown by the quantum LUMO+1 descriptor in the 2D QSAR studies.

H-bond acceptor field may be implicated in the anti-RT activity of the molecules. Fig. 9 (left) shows the compound DAPY 13A with the R1 *para*-position surrounded by the

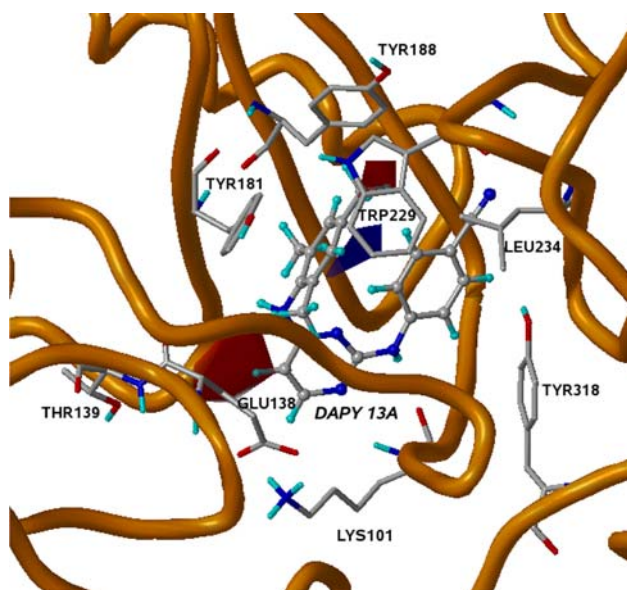


Fig. 8 CoMSIA electrostatic properties for the set of DAPY analogs. Contour plots are projected into the binding site of the protein and the amino acids involved are presented: blue surface encapsulate region where a more positively charged group will improve the activity and red contours where a more negatively charged group enhance activity

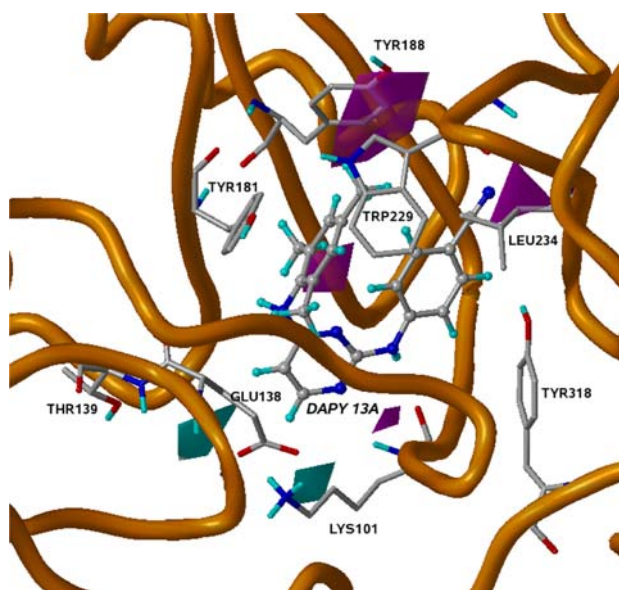


Fig. 9 CoMSIA H-bond contour plots are projected into the binding site of the protein and the amino acids involved are presented: Areas where H-bond acceptor groups on the ligand are favored are colored in magenta (left). Contours where H-bond donor groups are favored are colored in cyan and disfavored in purple (right)

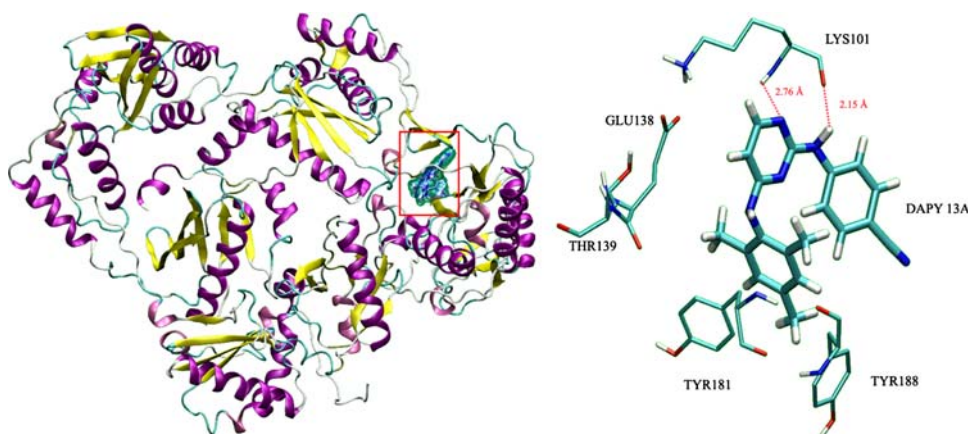
H-bond acceptor group favored contour plot. Compound 13C with a nitrile group at the *para*-position of R1 is more active than compound 13A with a methyl group at the same position. This information reinforces the results obtained from the steric and electrostatic fields where we found that a longer and more electronegative substituent at this position will increase the activity of the inhibitor that will gain one more H-bond interaction with the hydroxyl group of the TYR188 side chain.

The region at the amino group that links the pyrimidine ring to the ring 3 is crucial in the H-bond interactions between the ligand and the protein. We have a cyan area that means that an H-bond donor groups are favored and a magenta contour where an H-bond acceptor groups on the

ligand are favored to have better activity. The structure–activity relationship studies of the complex show two H-bond interactions between the inhibitor and the LYS101 of the protein (Fig. 10). We have one interaction between the NH of the ligand (that plays the role of an H-bond donor group) with the carboxylate group of the LYS101 main chain, and another H-bond interaction between the N atom of the pyrimidine ring (in this case the inhibitor plays the role of an H-bond acceptor group) and the amino group of the same amino acid main chain.

The superior performance of CoMSIA 3D QSAR model was assigned to the large contribution of the hydrogen-bonding interactions (44%) between the inhibitor and the receptor.

Fig. 10 Global overview of the complex structure showing the NNIPB with DAPY 13A (left). Zooming into the binding site (right) to put into evidence the inhibitor and its protein environment



Conclusion

A set of DAPY analogs was investigated to relate IC₅₀ values against the reverse transcriptase to the molecular structure. The correlation obtained from the 2D and 3D QSAR models show that the inhibitory activity of these compounds can be modeled and determined. The good predictivity of these models was verified by leave-one-out and internal validation methods. These results demonstrated that steric, electrostatic and hydrogen bonding influence significantly the inhibitory activity. The analysis of the 3D contour maps allowed us to mark areas of known inhibitors that require a particular physicochemical property to increase activity. For example, substitution at the *para*-position of ring 1 is crucial in affecting the HIV-1 RT inhibitory activity of the ligand. This result was confirmed by the different CoMSIA maps. The steric contours are in agreement with the ZX shadow geometrical descriptor from the 2D QSAR analysis. The electrostatic maps complement the LUMO+1 quantum descriptor in the 2D QSAR studies.

The knowledge of the protein structure is not a prerequisite to perform QSAR analysis; however the presence of the X-ray structure of RT/DAPY 13A complex had provided us the biological conformation of the inhibitor and the opportunity to interpret the results obtained by the 2D and 3D models with respect to the protein environment. This link permitted us to validate our model and give precious insights for the structure activity interpretations.

We hope that these results will give new insights on chemical modifications that can be realized with the aim of designing new inhibitors with improved pharmacological properties.

Acknowledgments This work was supported by various fellowships: Joseph Rebehmed from the French Ministry of Research and Technology, Catia Teixeira from the “Fundação para a Ciência e a Tecnologia” of Portugal.

References

- Rodgers DW, Gamblin SJ, Harris BA, Ray S, Culp JS, Hellmig B, Woolf DJ, Debouck C, Harrison SC (1995) *Proc Natl Acad Sci USA* 92:1222
- Pommier Y, Marchand C, Neamati N (2000) *Antiviral Res* 47:139
- Buolamwini JK, Assefa H (2002) *J Med Chem* 45:841
- Kovalevsky AY, Liu F, Leshchenko S, Ghosh AK, Louis JM, Harrison RW, Weber IT (2006) *J Mol Biol* 363:161
- Sierra S, Kupfer B, Kaiser R (2005) *J Clin Virol* 34:233
- Ragno R, Artico M, De Martino G, La Regina G, Coluccia A, Di Pasquali A, Silvestri R (2005) *J Med Chem* 48:213
- Van Herrewege Y, Vanham G, Michiels J, Franssen K, Kestens L, Andries K, Janssen P, Lewi P (2004) *Antimicrob Agents Chemother* 48:3684
- Tarrago-Litvak L, Andreola ML, Nevinsky GA, Sarih-Cottin L, Litvak S (1994) *Faseb J* 8:497
- Wainberg MA (2003) *J Acquir Immune Defic Syndr* 34(Suppl 1):S2
- De Clercq E (2004) *Chem Biodivers* 1:44
- Sapre NS, Gupta S, Pancholi N, Sapre N (2008) *J Comput Aided Mol Des* 22:69
- Guillemont J, Pasquier E, Palandjian P, Vernier D, Gaurrand S, Lewi PJ, Heeres J, de Jonge MR, Koymans LM, Daeyaert FF, Vinkers MH, Arnold E, Das K, Pauwels R, Andries K, de Bethune MP, Bettens E, Hertogs K, Wigerinckx P, Timmerman P, Janssen PA (2005) *J Med Chem* 48:2072
- Ludovici DW, De Corte BL, Kukla MJ, Ye H, Ho CY, Lichtenstein MA, Kavash RW, Andries K, de Bethune MP, Azijn H, Pauwels R, Lewi PJ, Heeres J, Koymans LM, de Jonge MR, Van Aken KJ, Daeyaert FF, Das K, Arnold E, Janssen PA (2001) *Bioorg Med Chem Lett* 11:2235
- Drake SM (2000) *J Antimicrob Chemother* 45:417
- Campiani G, Ramunno A, Maga G, Nacci V, Fattorusso C, Catalano B, Morelli E, Novellino E (2002) *Curr Pharm Des* 8:615
- Thakur A, Thakur M, Bharadwaj A, Thakur S (2007) *Eur J Med Chem* 43:471
- Das K, Lewi PJ, Hughes SH, Arnold E (2005) *Prog Biophys Mol Biol* 88:209
- CODESSA software version 2.63, University of Florida 2002
- Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M (2006) *J Med Chem* 49:6802
- Cramer RD, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959
- Cramer RD 3rd, Patterson DE, Bunce JD (1989) *Prog Clin Biol Res* 291:161
- Klebe G, Abraham U, Mietzner T (1994) *J Med Chem* 37:4130
- Bohm M, St rzebecher J, Klebe G (1999) *J Med Chem* 42:458
- Pauwels R, Balzarini J, Baba M, Snoeck R, Schols D, Herdewijn P, Desmyter J, De Clercq E (1988) *J Virol Methods* 20:309
- Cho SJ, Tropsha A (1995) *J Med Chem* 38:1060
- Chen H, Li Q, Yao X, Fan B, Yuan S, Panaye A, Doucet JP (2004) *QSAR Combinatorial Sci* 23:36
- Das K, Clark AD Jr, Lewi PJ, Heeres J, De Jonge MR, Koymans LM, Vinkers HM, Daeyaert F, Ludovici DW, Kukla MJ, De Corte B, Kavash RW, Ho CY, Ye H, Lichtenstein MA, Andries K, Pauwels R, De Bethune MP, Boyer PL, Clark P, Hughes SH, Janssen PA, Arnold E (2004) *J Med Chem* 47:2550
- SYBYL 7.3, Tripos Inc
- Powell MJD (1977) *Math Progr* 12:241
- Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) *J Am Chem Soc* 107:3902
- Stewart JJ (1990) *J Comput Aided Mol Des* 4:1
- Katritzky AR, Perumal S, Petrukhin R, Kleinpeter E (2001) *J Chem Inf Comput Sci* 41:569
- Ren Y, Liu H, Yao X, Liu M (2007) *J Chromatogr A* 1155:105
- Bauvais C, Barbault F, Zhu Y, Petitjean M, Fan BT (2006) *SAR QSAR Environ Res* 17:253
- Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M (2006) *Bioorg Med Chem* 14:2333
- Katritzky AR, Dobchev DA, Tulp I, Karelson M, Carlson DA (2006) *Bioorg Med Chem Lett* 16:2306
- Katritzky AR, Pacureanu LM, Dobchev DA, Fara DC, Duchowicz PR, Karelson M (2006) *Bioorg Med Chem* 14:4987
- AMPAC 8.15, © 1992–2004 Semichem, Inc. PO Box 1649, Shawnee, KS 66222
- Srivani P, Srinivas E, Raghu R, Sastry GN (2007) *J Mol Graph Model* 26:378
- So SS, Karplus M (1997) *J Med Chem* 40:4360

41. Suh M-E, Park S-Y, Lee H-J (2002) Bull Korean Chem Soc 23:417
42. Rohrbaugh RH, Jurs PC (1987) Anal Chem 59:1048
43. Rohrbaugh RH, Jurs PC (1987) Anal Chimica Acta 199:99
44. Katritsky AR, Petrukhin R, Perumal S, Karelson M, Prakash I, Desai N (2002) Croat Chem Acta 75:475
45. Coi A, Massarelli I, Murgia L, Saraceno M, Calderone V, Bianucci AM (2006) Bioorg Med Chem 14:3153
46. Eroglu E, Turkmen H (2007) J Mol Graph Model 26:701
47. Norinder U (1998) Perspect Drug Discovery Des 12/13/14:25
48. Kim KHG, Novellino E (1998) Perspect Drug Discovery Des 12:257
49. Bringmann G, Rummey C (2003) J Chem Inf Comput Sci 43:304