

Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks

Michael Fernández and Julio Caballero*

Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba

Received 2 June 2005; revised 4 August 2005; accepted 5 August 2005

Available online 3 October 2005

Abstract—Artificial neural networks (ANNs) were used to model both inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC_{90}) for 55 cyclic urea derivatives using constitutional and 2D descriptors. As a preliminary step, linear dependences were established by multiple linear regression (MLR) approaches, selecting the relevant descriptors by genetic algorithm (GA) feature selection. For ANN models non-linear GA feature selection was also applied. Non-linear modeling of K_i overcame the results of the linear one using four properties, keeping in mind standard Pearson R correlation coefficients (0.931 vs. 0.862) and leave one out (LOO) cross-validation analysis ($Q^2_{LOO} = 0.703$ vs. 0.510). On the other hand, IC_{90} modeling was insoluble by a linear approach: no predictive model was achieved; however, a non-linear relation was encountered according to statistic results ($R = 0.891$; $Q^2_{LOO} = 0.568$). The best non-linear models suggested the influence of the presence of nitrogen atoms and the molecular volume distribution in the inhibitor structures on the HIV-1 protease inhibition as well as that the inhibition of HIV replication was dependent on the occurrence of five-member rings. Finally, inhibitors were well distributed regarding its activity levels in a Kohonen self-organizing map built using the input variables of the best non-linear models.
© 2005 Elsevier Ltd. All rights reserved.

1. Introduction

The human immunodeficiency virus (HIV), which has been identified as the causative agent of acquired immune-deficiency syndrome (AIDS), has promoted an unequalled scientific effort to understand and control this disease. The life cycle of the human HIV-1 provides a number of targets for potential chemotherapeutic intervention. Enzymes related to HIV-1 replication have been identified as therapeutic targets; in consequence, a wide set of inhibitors of these enzymes have been discovered. As a matter of fact, current therapies utilize a combination of reverse transcriptase and protease inhibitors.¹ Furthermore other targets have been identified as therapeutic agents that block viral entry, fundamentally the chemokine receptor CCR5.²

One of the crucial stages in the HIV life cycle is the protease-mediated transformation from the immature, non-dangerous virion, to the mature, infective virus.

HIV-1 protease inhibitors have thus become a major target for anti-AIDS drug design,^{3,4} its inhibition has been shown to extend the length and improve the quality of life of AIDS patients. A large number of inhibitors have been designed, synthesized, and assayed, and several HIV-1 protease inhibitors are now utilized in the treatment of AIDS.^{5–7} Unfortunately, not all potent inhibitors of the enzyme make good drugs. The peptide-like nature and size of many HIV-1 protease inhibitors limit their oral bioavailability and half-life in humans, making high blood levels difficult to achieve and sustain.⁸ Moreover, many peptide-like agents bind to plasma proteins, limiting the effective concentration of free drug available to interact with the intracellular target sites.⁹ For this, the search continues for new inhibitors that exhibit increased potency, lower toxicity, and critically, effectiveness against the growing population of mutants.

The availability of computational techniques based on structure–activity relationships has accelerated the drug design process. Large databases of candidate inhibitors exist that have yet to be evaluated against HIV-1 protease or its resistant variants. This backlog has exerted pressure to develop faster and more effective strategies for the ‘virtual screening’ of candidate inhibitors. In sil-

Keywords: QSAR analysis; Backpropagation neural network analysis; Bayesian regularization; Self-organizing maps; HIV-1 protease inhibitors; Genetic algorithm.

*Corresponding author. Tel.: +53 45 26 1251; fax: +53 45 25 3101; e-mail: jmcr77@yahoo.com

ico models that are able to predict the biological activity of compounds by its structural properties are powerful tools to design highly active molecules. In this sense, several theoretical tools have been applied for the sketch of new anti-HIV candidates analyzing enzyme–inhibitor interactions^{10,11} and scanning the contribution of molecular structures to biological activity by mathematical relations.

Quantitative structure–activity relationship (QSAR) studies have been successfully applied for modeling activities of several kinds of anti-HIV agents.^{12–19} Due to the wide set of compounds reported, the inhibitors of HIV-1 replication have been of interest for modelers. Among anti-HIV QSAR models, the broad majority has been accomplished for HIV-1 reverse transcriptase inhibitors^{12–14} and HIV-1 protease inhibitors.^{15–19}

In this work, constitutional and 2D descriptors were used for encoding structural information from cyclic

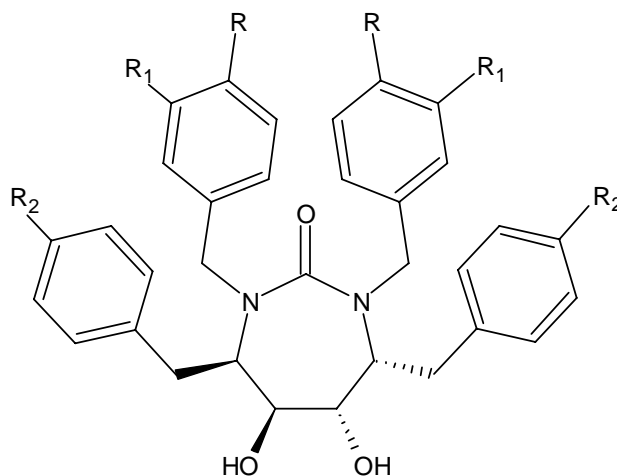
urea derivatives, and linear and non-linear models for their inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC_{90}) were built using multivariate linear regression analysis (MLR) and artificial neural networks (ANNs). A comparative study was developed according to the results of data fitting and the predictive power of the models measured by cross-validation technique. The versatility of ANNs was also used for mapping both anti-HIV activities on topological maps using competitive neural networks.

2. Materials and methods

2.1. Data sets: source and prior preparation

Inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC_{90}) for 55 cyclic urea derivatives were taken from the literature.^{20–24} The chemical structures and logarithmic experimental activities are shown in Tables

Table 1. Structural features of cyclic urea derivatives (compounds 1–17)

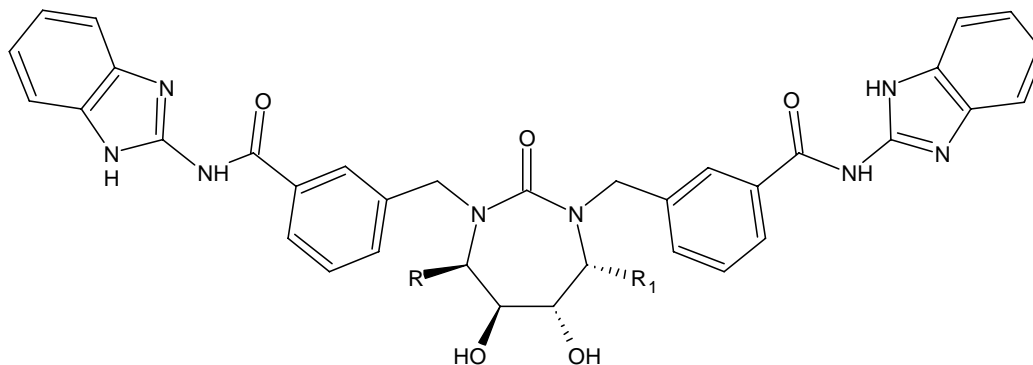


Experimental and predicted K_i and IC_{90} by multilinear regression analysis and artificial neural network models

Compound ^a		log(10 ⁶ /K _i (nM))			log(10 ⁶ /IC ₉₀ (nM))		
		Experimental	MLR	ANN2	Experimental	MLR	ANN2 ^b
1 (DMP323)	R = CH ₂ OH; R ₁ = R ₂ = H	6.54	6.70	6.18	4.04	3.92	4.13
2 (DMP450)	R = R ₂ = H; R ₁ = NH ₂	6.55	6.20	6.93	3.90	3.83	3.33
3 (XL075)	R = R ₁ = R ₂ = H	5.47	6.00	5.63	3.10	3.68	3.69
4 (XN975)	R = R ₂ = H; R ₁ = 3(1H)pyrazole	7.57	7.34	7.47	4.25	3.63	3.84
5	R = R ₁ = H; R ₂ = OH	5.96	6.11	5.84	3.74	3.63	3.44
6	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = OH	7.79	7.45	7.53	2.87	3.80	3.57
7	R = R ₁ = H; R ₂ = OCH ₃	5.07	5.46	5.47	3.22	3.56	3.14
8	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = OCH ₃	7.11	6.88	7.26	5.00	4.57	4.37
9	R = R ₁ = H; R ₂ = OCH ₂ CH ₂ OH	6.12	5.81	5.96	4.11	3.76	4.15
10	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = OCH ₂ CH ₂ OH	7.23	7.17	6.93	3.34	3.75	3.56
11	R = R ₁ = H; R ₂ = 4-pyridinylmethoxy	6.08	5.81	6.21	4.00	3.96	3.99
12	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = 4-pyridinylmethoxy	7.15	7.19	7.26	4.74	4.13	4.50
13	R = R ₁ = H; R ₂ = 2-(4-morpholinyl)ethoxy	5.23	4.97	5.20	3.64	3.12	3.83
14	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = 2-(4-morpholinyl)ethoxy	6.89	6.47	6.83	2.99	3.16	2.81
15	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = 3-pyridinylmethoxy	7.23	7.23	7.30	4.49	4.21	4.72
16	R = R ₁ = H; R ₂ = OCH ₂ CH ₂ NHCH ₃	5.80	6.22	6.04	3.10	2.97	3.04
17	R = H; R ₁ = 1H-pyrazol-3-yl; R ₂ = OCH ₂ CH ₂ N(CH ₃) ₂	7.21	7.18	7.61	2.27	3.03	2.72

^a Compounds 1–17 are from Ref. 21.

^b Model considering compound 19 as an outlier.

Table 2. Structural features of cyclic urea derivatives (compounds **18–30**)Experimental and predicted K_i and IC_{90} by multilinear regression analysis and artificial neural network models

Compound ^a		$\log(10^6/K_i \text{ (nM)})$			$\log(10^6/IC_{90} \text{ (nM)})$		
		Experimental	MLR	ANN2	Experimental	MLR	ANN2 ^b
18 (SD146)	R = R ₁ = benzyl	7.62	7.27	7.48	5.30	4.62	4.86
19	R = R ₁ = ethyl	5.35	6.31	5.92	3.24	3.72	—
20	R = R ₁ = isobutyl	6.52	6.64	6.31	4.23	4.13	4.32
21	R = R ₁ = hexyl	6.46	7.02	6.71	4.07	4.16	3.87
22	R = R ₁ = cyclohexyl	6.54	6.85	6.95	4.85	4.76	5.12
23	R = R ₁ = 4-aminobenzyl	7.80	8.31	7.62	5.00	4.94	4.85
24	R = R ₁ = 4-(dimethylamino)benzyl	7.42	7.32	7.11	5.10	5.33	4.96
25	R = R ₁ = 4-hydroxybenzyl	7.46	7.37	7.56	4.46	4.84	4.83
26	R = R ₁ = 4-(benzyloxy)benzyl	5.43	6.18	5.57	3.90	4.35	3.97
27	R = R ₁ = 4-methoxybenzyl	7.59	6.87	7.30	4.72	5.26	4.60
28	R = ethyl; R ₁ = benzyl	7.16	6.88	7.12	4.59	4.48	4.98
29	R = isobutyl; R ₁ = benzyl	7.42	6.90	7.16	5.00	4.51	5.00
30	R = hexyl; R ₁ = benzyl	7.34	7.14	7.32	5.05	4.46	5.08

^a Compounds **18–30** are from Ref. 22.^b Model considering compound **19** as an outlier.

1–3. K_i (nM) values represent the enzyme inhibition constants, meanwhile the activity parameters IC_{90} are measures of antiviral potency and refer to the nanomolar concentration of each compound required to reduce the concentration of HIV viral RNA by 90% from the level measured in an infected culture.

Prior to molecular descriptor calculations, 3D structures of the studied compounds were geometrically optimized using semi-empirical quantum-chemical method PM3²⁵ implemented in MOPAC 6.0²⁶ computer software.

2.2. Molecular descriptors

The simplest information that can be extracted from a molecule is the amount of atoms or congregations of atoms, according to its chemical nature and chemical environment, defining molecular sub-fragments that can be considered as ‘structure making’ factors.²⁷ The constitutional descriptors used in this work are averaged physical properties, number of atoms, groups of atoms, and functional groups (Table 4). In addition, some atom centered fragments (Table 5) were accounted for taking a more precise description of molecules.

Contrary to the constitutional descriptors, the 2D descriptors integrate the molecular information.²⁸ They are graph theoretical indexes that are based on the topological properties of a molecule viewed as a graph²⁹

without taking into account the conformations. In this work two kinds of 2D descriptors were used: autocorrelation vectors and BCUT descriptors.

The autocorrelation vectors represent the degree of similarity between molecules. H-depleted molecular structure is represented as a graph G and physico-chemical properties of atoms as real values assigned to the vertices of G (Table 6). These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in G . Three spatial autocorrelation vectors were employed for modeling the inhibitory activity:

Moran’s indexes³⁰

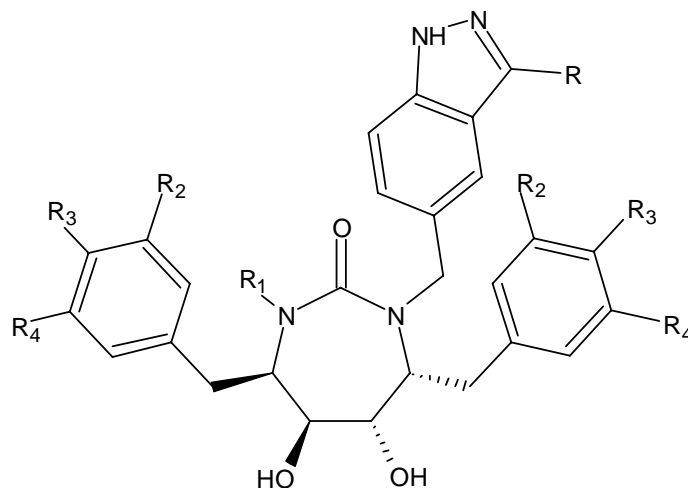
$$I(p_k, l) = \frac{N}{2L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)}, \quad (1)$$

Geary’s coefficients³¹

$$c(p_k, l) = \frac{(N-1)}{4L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)}, \quad (2)$$

and Broto-Moreau’s autocorrelation coefficients³²

$$A(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj}, \quad (3)$$

Table 3. Structural features of cyclic urea derivatives (compounds **31–55**)Experimental and predicted K_i and IC_{90} by multilinear regression analysis and artificial neural network models

Compound ^a		$\log(10^6/K_i \text{ (nM)})$			$\log(10^6/IC_{90} \text{ (nM)})$		
		Experimental	MLR	ANN2	Experimental	MLR	ANN2 ^b
31 (DMP850)	R = NH ₂ ; R ₁ = benzyl; R ₂ = R ₃ = R ₄ = H	7.51	7.12	7.46	4.21	4.22	4.42
32 (DMP851)	R = NH ₂ ; R ₁ = <i>n</i> -butyl; R ₂ = R ₃ = R ₄ = H	7.68	7.67	7.47	4.25	3.83	4.04
33 (SE063)	R = R ₂ = R ₃ = R ₄ = H; R ₁ = 3-aminobenzyl	7.52	7.00	7.33	4.96	4.24	4.40
34	R = NH ₂ ; R ₁ = cyclopropylmethyl; R ₂ = R ₃ = R ₄ = H	7.70	7.37	7.75	3.94	3.94	4.14
35	R = NH ₂ ; R ₁ = cyclobutylmethyl; R ₂ = R ₃ = R ₄ = H	7.80	7.41	7.73	4.74	4.31	4.45
36	R = NH ₂ ; R ₁ = <i>n</i> -pentyl; R ₂ = R ₃ = R ₄ = H	7.39	7.18	7.23	3.69	3.85	3.82
37	R = NH ₂ ; R ₁ = <i>n</i> -hexyl; R ₂ = R ₃ = R ₄ = H	6.89	7.47	7.36	3.12	4.02	3.82
38	R = NH ₂ ; R ₁ = naphthylmethyl; R ₂ = R ₃ = R ₄ = H	7.64	7.15	6.98	4.64	5.22	4.57
39	R = 2-thienyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.74	6.88	6.91	4.51	4.38	4.41
40	R = 3-thienyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.89	6.85	6.99	4.49	4.22	4.39
41	R = 3-pyrazole; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	7.24	7.62	7.63	3.97	4.06	3.73
42	R = 4-fluorophenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.52	6.69	6.58	4.41	4.33	4.47
43	R = 3-trifluoromethylphenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.21	6.39	6.44	4.41	4.15	4.10
44	R = 4-trifluoromethylphenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.60	6.70	6.46	3.59	3.96	4.01
45	R = 3-aminophenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	7.05	6.93	7.44	4.52	4.72	4.64
46	R = 3-methoxyphenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.62	6.77	6.43	4.48	4.66	4.64
47	R = 4-methoxyphenyl; R ₁ = 3-aminobenzyl; R ₂ = R ₃ = R ₄ = H	6.59	6.53	6.44	4.85	5.11	4.66
48	R = NH ₂ ; R ₁ = benzyl; R ₂ = R ₄ = H; R ₃ = CH ₃	7.33	7.47	7.22	4.92	4.99	4.50
49	R = NH ₂ ; R ₁ = benzyl; R ₃ = R ₄ = H; R ₂ = CH ₃	7.21	6.97	7.14	4.60	4.48	4.72
50	R = NH ₂ ; R ₁ = benzyl; R ₂ = R ₄ = H; R ₃ = CH ₂ CH ₃	7.14	6.84	6.98	4.72	4.93	4.74
51	R = NH ₂ ; R ₁ = benzyl; R ₂ = R ₄ = CH ₃ ; R ₃ = H	6.70	6.49	6.55	4.30	4.32	4.31
52	R = NH ₂ ; R ₁ = <i>n</i> -butyl; R ₂ = R ₄ = H; R ₃ = CH ₃	7.43	8.00	7.59	4.92	4.79	4.62
53	R = NH ₂ ; R ₁ = <i>n</i> -butyl; R ₃ = R ₄ = H; R ₂ = CH ₃	7.35	7.48	7.06	4.51	4.28	4.80
54	R = NH ₂ ; R ₁ = <i>n</i> -butyl; R ₂ = R ₄ = H; R ₃ = CH ₂ CH ₃	7.03	7.35	7.29	5.00	4.78	4.85
55	R = NH ₂ ; R ₁ = <i>n</i> -butyl; R ₂ = R ₄ = CH ₃ ; R ₃ = H	6.72	6.97	6.29	4.18	4.16	4.54

^a Compounds **31**, **32**, and **34–38** are from Ref. 20; **33**, **39–47** are from Ref. 23 and **48–55** are from reference 24.^b Model considering compound **19** as an outlier.

where $I(p_k, l)$, $c(p_k, l)$, and $A(p_k, l)$ are Moran's index, Geary's coefficient, and Broto-Moreau's autocorrelation coefficient at spatial lag l , respectively; p_{ki} and p_{kj} are the values of property k of atom i and j , respectively; \bar{p}_k is the average value of property k and $\delta(l, dij)$ is a Dirac-delta function defined as

$$\delta(l, dij) = \begin{cases} 1 & \text{if } dij = l \\ 0 & \text{if } dij \neq l \end{cases}, \quad (4)$$

where d_{ij} is the topological distance or spatial lag between atoms i and j .

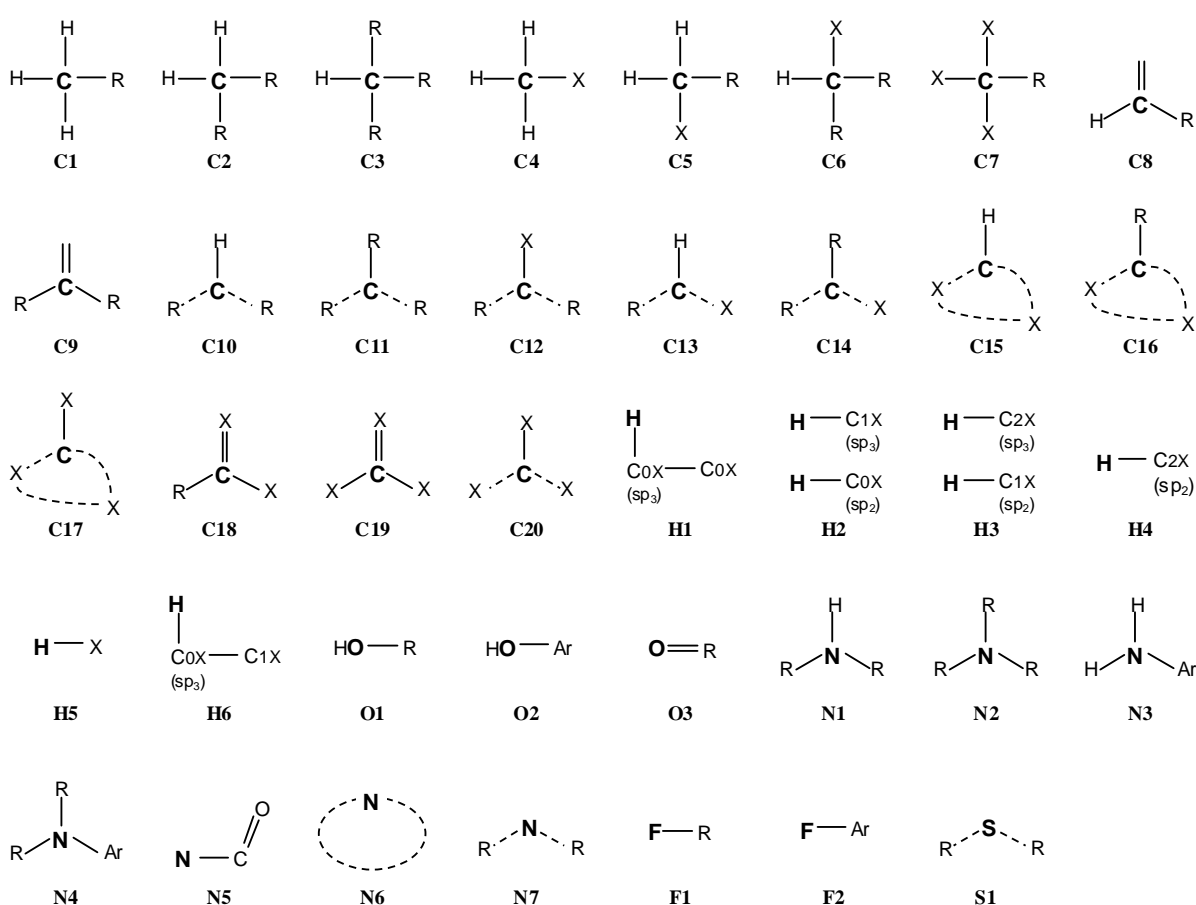
BCUT descriptors are derived from Burden's matrix, which represents the hydrogen-suppressed connection table of a molecule as a symmetrical $N \times N$ matrix with atomic numbers along the diagonal and bonding information in the off-diagonal elements.³³

BCUTs are highly compact indexes that keep the molecular information by solving the eigenvalue equation

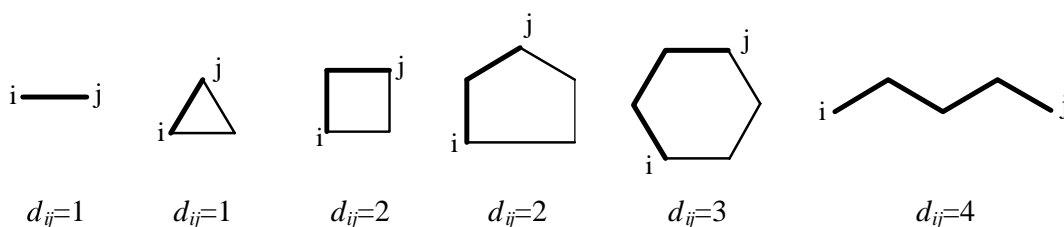
$$[B][V] = [V][e], \quad (5)$$

Table 4. Constitutional descriptors

<i>Averaged physical properties</i>	
MW	Molecular weight
AMW	Average molecular weight
Sv	Sum of atomic van der Waals volumes (scaled on carbon atoms)
Se	Sum of atomic Sanderson electronegativities (scaled on carbon atoms)
Sp	Sum of atomic polarizabilities (scaled on carbon atoms)
Ss	Sum of Kier–Hall electrotopological states
Mv	Mean atomic van der Waals volume (scaled on carbon atoms)
Me	Mean atomic Sanderson electronegativity (scaled on carbon atoms)
Mp	Mean atomic polarizability (scaled on carbon atoms)
Ms	Mean electrotopological state
<i>Atoms and groups of atoms</i>	
nAT	Number of atoms
nSK	Number of non-H atoms
nBT	Number of bonds
nBO	Number of non-H bonds
nBM	Number of multiple bonds
SCBO	Sum of conventional bond orders (H depleted)
nCIC	Number of rings
nCIR	Number of circuits
RBN	Number of rotatable bonds
RBF	Rotatable bond fraction
nDB	Number of double bonds
nAB	Number of aromatic bonds
nH	Number of Hydrogen atoms
nC	Number of Carbon atoms
nN	Number of Nitrogen atoms
nO	Number of Oxygen atoms
nS	Number of Sulphur atoms
nF	Number of Fluorine atoms
nR03	Number of 3-member rings
nR04	Number of 4-member rings
nR05	Number of 5-member rings
nR06	Number of 6-member rings
nR07	Number of 7-member rings
nR08	Number of 8-member rings
nR09	Number of 9-member rings
nR10	Number of 10-member rings
nBnz	Number of benzene-like rings
<i>Functional groups</i>	
nCp	Number of total primary C(sp ₃)
nCs	Number of total secondary C(sp ₃)
nCt	Number of total tertiary C(sp ₃)
nCq	Number of total quaternary C(sp ₃)
nCrH2	Number of ring secondary C(sp ₃)
nCrHR	Number of ring tertiary C(sp ₃)
nCaH	Number of unsubstituted aromatic C(sp ₂)
nCaR	Number of substituted aromatic C(sp ₂)
nCONHRPh	Number of secondary amides (aromatic)
nCONN	Number of urea derivatives
nNH2Ph	Number of primary amines (aromatic)
nNHR	Number of secondary amines (aliphatic)
nNR2	Number of tertiary amines (aliphatic)
nNR2Ph	Number of tertiary amines (aromatic)
nN = N	Number of N azo (aliphatic)
nOH	Number of total hydroxyl groups
nOHPh	Number of phenols
nOHp	Number of primary alcohols (aliphatic)
nOHs	Number of secondary alcohols (aliphatic)
nROR	Number of ethers (aliphatic)
nRORPh	Number of ethers (aromatic)
nRSR	Number of sulphurs
nCF3	Number of CF ₃ groups
nPhHal	Number of halogens on aromatic rings
nHDon	Number of donor atoms for H-bonds (with N and O)
nHAcc	Number of acceptor atoms for H-bonds (N O F)

Table 5. Atom centered fragments^a

^a Accounted atoms are in bold. Dashed lines represent bonds in aromatic rings. R represent aliphatic residues. Ar represent aromatic residues. X represent no-C atoms. COX are only C or H substituted carbons, C1X are one X substituted carbons, and C2X are two X substituted carbons.

Table 6. Representation of different molecular graphs G and topological distances or spatial lags d_{ij} 

where $[B]$ is a symmetric connectivity matrix, $[V]$ is a matrix of eigenvectors, and $[e]$ is a diagonal matrix of eigenvalues. BCUT descriptors are the smallest and highest eigenvalues taken to the matrix $[e]$. The eigenvectors have coefficients belonging to every atom, so eigenvalues reflect the topology of the whole molecule. For the above process several $[B]$ matrices are defined, which contain significant atomic properties on their diagonals.

As can be seen, 2D descriptors are 2D-dimensional chemistry spaces with the ability of integrating molecular information in defined arrays. In our work atomic

masses, polarizabilities, van der Waals volumes, and Sanderson electronegativities are the relevant atomic properties related to 2D descriptors (k terms in autocorrelation vectors and Burden's matrix diagonal terms). Dragon³⁴ computer software was used for calculating all descriptors used. Autocorrelation vectors were calculated at spatial lags l ranging from 1 up to 8 and BCUT descriptors were considered for the eight lowest and highest eigenvalues.

The total number of computed descriptors was 268. Descriptors that stayed constant or almost constant were eliminated and pairs of variables with a correlation

coefficient greater than 0.95 were classified as intercorrelated, and only one of these was included in the model. Finally, 108 descriptors were obtained.

2.3. Variable selections

Since 108 molecular descriptors are available for QSAR analysis and only a subset of them is statistically significant in terms of correlation with biological activities, deriving an optimal QSAR model through variable selection needs to be addressed. Following the principle of parsimony,³⁵ we selected just the variables that contain the information that is necessary for the modeling but nothing more. In this sense, linear and non-linear GA searches have been carried out in order to build the linear and non-linear models.

GAs are governed by biological evolution rules.³⁶ They are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space.³⁷ The first step is to create a population of N individuals. Each individual encodes the same number of randomly chosen descriptors. The fitness of each individual in this generation is determined. In the second step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from two or one of its parents.

GA implemented in this paper is a version of the So and Karplus report³⁸ and was programmed within the Matlab environment using genetic algorithm and neural network toolboxes.³⁹ We also included elitism which protects the fittest individual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. This selection, crossover, and mutation processes are repeated until all of the N parents in the population are replaced by their children. The fitness score of each member of this new generation is again evaluated and the reproductive cycle is continued until a 90% of the generations showed the same target fitness score.

Linear GA search was carried out exploring MLR models. In turn, neural network feature selection procedures that extract non-linear information from the data set were employed for data dimensionality reduction before network training.⁴⁰ Linear and non-linear models were generated between the activities ($\log(10^6/K_i)$ and $\log(10^6/IC_{90})$) and the respective selected molecular descriptors. The quality of each model was proven by the square multiple correlation coefficient (R^2) and the standard deviation (S). Models with R -values higher than 0.80 were selected and they were tested in cross-validation experiments.

2.4. Regularized-artificial neural networks

In contrast to common statistical methods, artificial neural networks (ANNs) are not restricted to linear

correlations or linear subspaces.⁴¹ They can take into account non-linear structures and structures of arbitrarily shaped clusters or curved manifolds. As biological phenomena are considered non-linear by nature, ANN technique was applied in order to discover the possible existence of non-linear relationships between activity and molecular descriptors that are ignored for the linear approach.⁴²

ANNs are computer-based models in which a number of processing elements, also called neurons, units, or nodes are interconnected by links in a netlike structure forming “layers.” A variable value is assigned to every neuron. The neurons can be one of three different kinds. The input neurons receive their values from independent variables and they constitute the input layer. The hidden neurons collect values from other neurons, giving a result that is passed to a successor neuron. The output neurons take values from other units and correspond to different dependent variables, forming the output layer. In this sense, network architecture is commonly represented as I–H–O, where I, H, and O are the number of neurons in the input, hidden, and output layers, respectively.

The links between units have associated values, named weights, that condition the values assigned to the neurons. There exist additional weights assigned to bias values that act as neuron value offsets. The weights are adjusted through a training process in order to minimize network error. For this, a non-linear transfer function relates the input parameters with the outputs. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

The characteristics of the ANNs have been found to be suitable for data processing, in which the functional relationship between the input and the output is not previously defined. This is due to the fact that structure–activity relationships are often non-linear and very complex, and neural networks are able to approximate any kind of analytical continuous function, according to Kolmogorov’s theorem.⁴³

When parameters (weights and biases) increase, network loses its ability to generalize. Error on the training set is driven to a very small value, but when new data are presented to the network the error is large. Network has memorized the training examples, but it has not learned to generalize to new situations, it means network overfits the data.

Typically, training aims to reduce the sum of squared errors $F = E_D$. Regularization involves modifying the performance function (F). It is possible to improve generalization if an additional term is adding

$$F = \beta E_D + \alpha E_W, \quad (6)$$

where E_W is the sum of squares of the network weights and biases, and α and β are objective function parameters. The relative size of the objective function parameters dictates the emphasis for training getting a smoother network response. MacKay’s Bayesian regularization automatically sets the correct values for the

objective function parameters,⁴⁴ in this sense the regularization is optimized. In the Bayesian framework, the weights of the network are considered random variables. After the data are taken, the density function for the weights can be updated according to Bayes' rule

$$P(w|D, \alpha, \beta, M) = \frac{P(D|w, \beta, M) \times P(w|\alpha, M)}{P(D|\alpha, \beta, M)}, \quad (7)$$

where D represents the data set, M is the particular neural network model used, and w is the vector of network weights. $P(w|D, \alpha, \beta, M)$ is the posterior probability, that is the plausibility of a weight distribution considering the information of the data set in the model used. $P(w|\alpha, M)$ is the prior density, which represents our knowledge of the weights before any data are collected. $P(D|w, \beta, M)$ is the likelihood function, which is the probability of the data occurring, given the weights. $P(D|\alpha, \beta, M)$ is a normalization factor, which guarantees that the total probability is 1.

Considering that the noise in the training set data is Gaussian and that the prior distribution for the weights is Gaussian, the posterior probability fulfills the relation

$$P(w|D, \alpha, \beta, M) = \frac{1}{Z_F} \exp(-F), \quad (8)$$

where Z_F depends on objective function parameters. So under this framework, minimization of F is identical to find the (locally) most probable parameters.

Bayesian regularization overcomes the remaining deficiencies of neural networks.⁴⁵ Bayesian methods produce models that are robust and able to comprise complex relations. No test or validation sets are involved so that all available training data can be devoted to the model and the potentially lengthy validation process can be avoided. At the end of training, ANN has optimal generalization qualities. The Bayesian neural net has the potential to give models which are relatively independent of neural network architecture, above a minimum architecture, and the Bayesian regularization method estimates the number of effective parameters.

Our Bayesian regularized ANN are classical back-propagation neural nets that incorporate the Bayesian regularization algorithm for finding the optimum weights. The Bayesian regularization takes place within the Levenberg–Marquardt algorithm⁴⁶ implemented in Matlab environment.³⁹ The initial value for μ was 0.005 with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when μ became larger than 10^{10} .

We used the following architecture:

- Input layer included the selected descriptors (four descriptors).
- One hidden layer with sigmoid transfer function was included. Hidden layer's architecture was varied from 2 to 4 neurons.

- Output layer had a linear transfer function and one neuron, representing the modeled activity (K_i or IC_{90}).

The input and output values were normalized.

2.5. Self-organizing maps

Often, cluster analysis is used to justify a chemistry space: if compounds with similar biological behavior cluster together in the proposed space, then it seems reasonable to conclude that the chemistry space is good. In order to settle structural similarities among the cyclic urea derivatives, self-organizing maps (SOM) were built for both activities. The selected descriptors in each model were used for unsupervised training of 9×9 neuron maps. Kohonen⁴⁷ introduced a neural network model that generates a SOM. Neurons are arranged in a 2D network. Molecules characterized by m descriptors are projected into this network. With $m > n$ a Kohonen network can be used to project a higher-dimensional space into a lower dimensional space.⁴⁸ Such maps of surface properties have been used for comparing a wide variety of biologically active compounds.⁴⁹

$$\text{out}_{c_s} \leftarrow \min \left[\sum_{i=1}^m (x_{si} - w_{ji})^2 \right] \quad (9)$$

Kohonen network is training using an unsupervised and competitive learning process. In our case a molecule s , characterized by m descriptors, x_{si} , will be projected into that (central) neuron, c_s , that has weights, w_{ji} , most similar to the input variables (Eq. 9). During the learning process, weights of the neurons in the network are changed to make them even more similar to the input variables. The weights of all neurons are adjusted but to an extent that decreases with increasing distance from the central, winning neuron, c_s . Finally, a molecule is projected into that neuron of the network with weights that come closest to the description of the molecule by the autocorrelation vector.

It should be noticed that the criterion embedded in Eq. 9 for determining the winning neuron for a molecule basically constitutes the measure determining the similarity of molecular structures. Molecules with similar autocorrelation vectors, Xs , are projected into the same or closely adjacent neurons. SOM were implemented in Matlab environment,³⁹ neurons were initially located at a grid topology. The ordering phase was developed in 1000 steps with 0.9 learning rate until tuning neighborhood distance (1.0) was achieved. The tuning phase learning rate was 0.02. Training was performed for a period of 2000 epochs in an unsupervised manner.

2.6. Model validation

The reliability of the models was indicated by cross-validation experiments quantified with predictive Q^2 . For leave one out (LOO) cross-validation a data point is removed (left-out) from the set, and the model refitted; the predicted value for that point is then compared to its actual value. This is repeated until each datum has

been omitted once; the sum of squares of these deletion residuals can then be used to calculate Q^2 , an equivalent statistic to R^2 .

$$Q^2 = 1 - \frac{\sum_{i=1}^N (Y_i - A_i)^2}{\sum_{i=1}^N (Y_i - \bar{A}_i)^2}, \quad (10)$$

where N is the number of compounds, Y_i and A_i are the predicted and experimental biological activities of i left-out compound, respectively, and \bar{A}_i is the average experimental activity of left-in compounds different to i .

In addition to the traditional LOO cross-validation, leave-five-out (L5O) cross-validations were also performed, where in each experiment the objects were left out randomly. In this case, the results were reported as the averaged Q^2 of 20 replies.

The Q^2 values can be considered a measure of the predictive power of a model: whereas r^2 can always be increased artificially by adding more parameters (descriptors or neurons), Q^2 decreases if a model is over-parameterized,³⁵ and is therefore a more meaningful summary statistic for predictive models.

3. Results and discussion

3.1. Multiple linear regression approach

Although inhibition of HIV-1 protease is generally paralleled by a reduced rate of HIV replication in cell culture, the correlation between inhibition constants (K_i) and inhibition of HIV replication is in fact rather poor (Fig. 1).

In a first approach, a MLR model for inhibitory activities of cyclic urea derivatives against HIV-1 protease was carried out with acceptable statistic significance and predictive power (Eq. 11). We chose the best four-

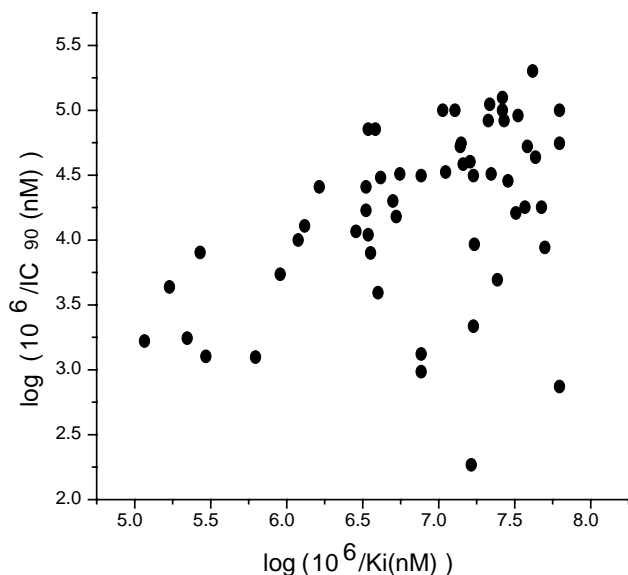


Figure 1. Correlation between inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC_{90}).

variable model as the most reliable linear relationship between the calculated descriptors and the HIV-1 protease inhibitory activity. The best three-variable model has a poor predictive power and the best five-variable model kept the precedent variables and introduced a variable which does not contribute to increase the predictive power of the model (Table 7).

$$\begin{aligned} \log(10^6/K_i) = & -0.055 \times nC + 0.413 \times nN + 7.114 \\ & \times \text{GATS6v} - 8.493 \times \text{GATS7v} \\ & + 8.653, \end{aligned} \quad (11)$$

$$N = 55 \quad R = 0.862 \quad S = 0.373 \quad p < 10^{-5},$$

$$\begin{aligned} Q^2_{\text{LOO}} &= 0.510 \quad S_{\text{CV LOO}} = 0.443 \\ Q^2_{\text{L5O}} &= 0.521 \quad S_{\text{CV L5O}} = 0.423. \end{aligned}$$

In Eq. 11, N is the number of compounds included in the model, R is the correlation coefficient, S is the standard deviation of the regression, p is the significance of the variables in the model, Q^2_{LOO} and $S_{\text{CV LOO}}$ are the correlation coefficient and standard deviation of the LOO cross-validation, respectively, and Q^2_{L5O} and $S_{\text{CV L5O}}$ are the correlation coefficient and standard deviation of the L5O cross-validation, respectively. There is no significant intercorrelation between the linear GA selected descriptors, as it is seen in Table 8.

Inhibitory activities (K_i) of the cyclic urea derivatives predicted by the linear model appear in Tables 1–3. This model is able to explain about 74% data variance and more important it is quite stable to the inclusion–exclusion of compounds as measured by LOO and L5O correlation coefficients ($Q^2 > 0.5$).⁵⁰ Two constitutional descriptors appear in the model corresponding to the number of carbons and nitrogens present in the molecular structure. In addition, two Geary's spatial autocorrelation coefficients weighted by atomic van der Waals volumes complete the model. These autocorrelation descriptors represent the degree of similarity between inhibitor molecules based on this atomic property at spatial lags 6 and 7, respectively.

On the other hand, the best linear model of the inhibition of HIV replication (IC_{90}) is unable to predict the activity of unknown compounds according to its Q^2 values lower than 0.5 (Table 7; Eq. 12).

$$\begin{aligned} \log(10^6/IC_{90}) = & -2.717 \times \text{BELm8} + 42.584 \\ & \times \text{BELv3} + 12.712 \times \text{MATS5e} \\ & + 3.599 \times \text{GATS6v} - 78.640, \end{aligned} \quad (12)$$

$$N = 55 \quad R = 0.817 \quad S = 0.406 \quad p < 10^{-5},$$

$$\begin{aligned} Q^2_{\text{LOO}} &= 0.392 \quad S_{\text{CV LOO}} = 0.449 \\ Q^2_{\text{L5O}} &= 0.405 \quad S_{\text{CV L5O}} = 0.427. \end{aligned}$$

Table 7. Descriptors and statistics for MLR and ANN models^a

Models	Variables ^b	<i>n</i>	<i>R</i>	<i>S</i>	<i>p</i>	LOO		L50 ^c	
						<i>Q</i> ²	<i>S</i> _{CV}	<i>Q</i> ²	<i>S</i> _{CV}
log(10 ⁶ / <i>K</i> _i)									
MLR3	nC, nN, MATS6v	55	0.795	0.441	<10 ^{−5}	0.249	0.487	—	—
MLR4	nC, nN, GATS6v, GATS7v	55	0.862	0.373	<10 ^{−5}	0.510	0.443	0.521	0.423
MLR5	nC, nN, GATS6v, GATS7v, H1	55	0.877	0.357	<10 ^{−5}	0.503	0.449	—	—
ANN1-Ki	nC, nN, GATS6v, GATS7v	55	0.917	0.280	<10 ^{−5}	0.592	0.398	0.568	0.431
ANN2-Ki	nC, nN, BEHv5, GATS6v	55	0.932	0.254	<10^{−5}	0.703	0.343	0.702	0.345
ANN2-Ki	nBnz, BEHm2, MATS6v, MATS1e, GATS5e	55	0.957	0.243	<10 ^{−5}	0.701	0.336	—	—
log(10 ⁶ /IC ₉₀)									
MLR3	BELm8, BELv3, MATS5e	55	0.792	0.426	<10 ^{−5}	0.303	0.445	—	—
MLR4	BELm8, BELv3, MATS5e, GATS6v	55	0.817	0.406	<10 ^{−5}	0.392	0.449	0.405	0.427
MLR5	BELm8, BELv3, MATS5e, GATS6v, C4	55	0.832	0.394	<10 ^{−5}	0.389	0.458	—	—
ANN1-IC90	BELm8, BELv3, MATS5e, GATS6v	55	0.864	0.339	<10 ^{−5}	0.406	0.420	0.430	0.423
ANN2-IC90	nR05, C11, BELm3, MATS7m	54	0.891	0.301	<10^{−5}	0.568	0.377	0.548	0.385
ANN2-IC90	BELm3, MATS3v, MATS1e, GATS2e, nNR2	54	0.903	0.297	<10 ^{−5}	0.566	0.407	—	—

^a The best *K*_i and IC₉₀ predictors appear in boldface.^b Variable definitions: nC, number of carbon atoms; nN, number of nitrogen atoms; nBnz, number of benzene-like rings; nR05, number of 5-member rings; nNR2, number of tertiary amines (aliphatic); C4, C11, H1, atom centered fragment (Table 5); MATS7m, Moran autocorrelation of lag 7 weighted by atomic masses; MATS3v and MATS6v, Moran autocorrelation of lag 3 and 6 weighted by atomic van der Waals volumes; MATS1e and MATS5e, Moran autocorrelation of lag 1 and 5 weighted by atomic Sanderson electronegativities; GATS6v and GATS7v, Geary autocorrelation of lag 6 and 7 weighted by atomic van der Waals volumes; GATS2e and GATS5e, Geary autocorrelation of lag 2 and 5 weighted by atomic Sanderson electronegativities; BEHm2, highest eigenvalue n. 2 of Burden matrix weighted by atomic masses; BEHv5, highest eigenvalue n. 5 of Burden matrix weighted by atomic van der Waals volumes; BELm3 and BELm8, lowest eigenvalue n. 3 and 8 of Burden matrix weighted by atomic masses; BELv3, lowest eigenvalue n. 3 of Burden matrix weighted by atomic van der Waals volumes.^c Average from 20 cross-validation experiments.**Table 8.** Correlation matrices of the descriptors selected by linear and non-linear GA

	nC	nN	GATS6v	GATS7v
<i>Linear GA K_i</i>				
nC	1			
nN	0.539	1		
GATS6v	0.159	0.215	1	
GATS7v	0.163	0.091	0.022	1
	nC	nN	BEHv5	GATS6v
<i>Non-linear GA K_i</i>				
nC	1			
nN	0.539	1		
BEHv5	0.559	0.303	1	
GATS6v	0.159	0.215	0.188	1
	BELm8	BELv3	MATS5e	GATS6v
<i>Linear GA IC₉₀</i>				
BELm8	1			
BELv3	0.386	1		
MATS5e	0.131	0.444	1	
GATS6v	0.097	0.396	0.338	1
	nR05	BELm3	MATS7m	C11
<i>Non-linear GA IC₉₀</i>				
nR05	1			
BELm3	0.278	1		
MATS7m	0.281	0.038	1	
C11	0.067	0.099	0.000	1

3.2. Artificial neural network approach

Since biological interactions are non-linear by nature; the main goal of this work was to train ANNs for modeling anti-HIV activities of cyclic urea derivatives. In our approach ANNs architecture was varied testing different

quantities of hidden layers. For both activities studied here, results were stable because the Bayesian regularization avoids overfitting, so the 4-2-1 architecture was chosen.

Two non-linear models were built for both *K*_i and IC₉₀ anti-HIV activities. ANN1 models were generated using

the descriptors that appeared in the MLR model as network inputs. On the other hand, ANN2 models were generated by non-linear GA (see Section 2).

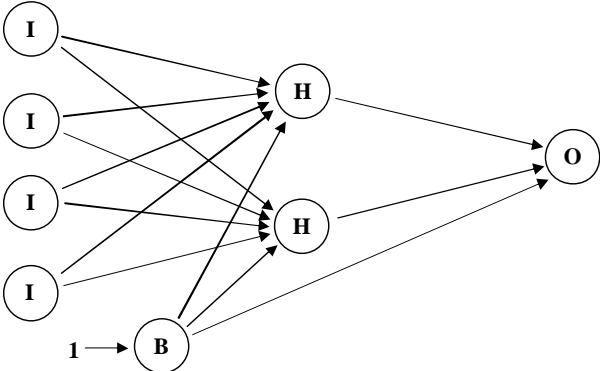
The descriptors and statistics for ANN1 and ANN2 models are depicted in Table 7; in turn, the adjusted parameters for each model are reported in Table 9. Both ANN1 statistics reveal that neural network approaches improve the results for linear models. Both ANN1 models are better in terms of fitness and predictive capacity; however, non-linear GA allows finding better predictors; in fact, ANN2 models show the best results in both activities examining the predictions in Tables 1–3 and the statistics in Table 7. Similar to the variables selected by linear GA, there is no significant intercorrelation between the non-linear selected descriptors for K_i and IC₉₀ modeling, as it is seen in Table 8.

For K_i modeling ANN2-Ki have nearly the same variables that are in the linear model: nC , nN and GATS6v

are common, but the Geary atomic van der Waals volume term GATS7v is substituted for the BCUT descriptor BEHv5. An outstanding aspect is that both descriptors encode the same atomic properties in two different 2D-spaces. Furthermore, the changed descriptor fits the activity in a non-linear form, but the same replacement in the linear model thoroughly deteriorate the correlation and predictive capacity ($R = 0.797$; $Q^2_{\text{LOO}} = 0.266$) of the predictor. The biggest success of model ANN2-Ki is that it reaches the highest predictive power ($Q^2_{\text{LOO}} = 0.703$; $Q^2_{\text{L50}} = 0.702$).

ANN2-IC90 model for IC₉₀ is wholly different to the linear analogue (Table 7). The four variables included are the number of five-member rings (nR05), the number of carbon substituted aromatic carbons (C11), and two atomic mass weighted 2D descriptors (BELm3 and MATS7m). When the 55 cyclic urea derivatives were included in the model, Q^2_{LOO} value was 0.481. ANN2-IC90 was more predictive, but not enough according

Table 9. Adjusted artificial neural network parameters^a



	Weights ^b		$j = 1$	$j = 2$	$j = 3$	$j = 4$
ANN1-Ki	$w(i,j)$ I–H	$i = 1$	–1.0387	1.4634	–0.5002	–0.1488
		$i = 2$	0.1703	–0.5508	–0.6341	0.2241
	$w(i,j)$ H–O	$i = 1$	0.8541	–1.4059		
	$w(i,j)$ B–H	$i = 1$	0.5754			
ANN2-Ki		$i = 2$	–0.5054			
	$w(i,j)$ B–O	$i = 1$	–0.8112			
	$w(i,j)$ I–H	$i = 1$	–0.2352	0.5103	–0.3265	–0.4868
		$i = 2$	–1.5720	2.4143	–0.6392	–0.9946
ANN1-IC90	$w(i,j)$ H–O	$i = 1$	–2.9390	1.6974		
	$w(i,j)$ B–H	$i = 1$	0.1392			
		$i = 2$	0.1891			
	$w(i,j)$ B–O	$i = 1$	0.0962			
ANN2-IC90	$w(i,j)$ I–H	$i = 1$	0.7041	–1.4024	–0.5427	–0.3896
		$i = 2$	–0.4976	–0.5110	–0.5069	0.2666
	$w(i,j)$ H–O	$i = 1$	–1.4439	–0.4336		
	$w(i,j)$ B–H	$i = 1$	–0.4687			
ANN2-IC90		$i = 2$	–0.1112			
	$w(i,j)$ B–O	$i = 1$	–0.4878			
	$w(i,j)$ I–H	$i = 1$	0.6646	–1.5015	–0.4003	–0.7477
		$i = 2$	–0.6788	0.6667	0.4895	0.5550
ANN2-IC90	$w(i,j)$ H–O	$i = 1$	–1.5813	–0.9370		
	$w(i,j)$ B–H	$i = 1$	–1.7032			
		$i = 2$	–0.1266			
	$w(i,j)$ B–O	$i = 1$	–1.0584			

I, neurons in the input layer; H, neurons in hidden layer; O, neurons in output layer; B, bias.

^a Weights derived from training process. Inputs and outputs were normalized before training.

^b Weights are represented by $w(i,j)$ where i is the posterior neuron and j is the prior neuron.

to the extended criteria ($Q^2 > 0.5$).⁵⁰ The five-variable model was searched, but an additional descriptor did not introduce a significant improvement in the predictive capacity. The analysis of possible outliers showed that the compound **19** had a large residual ($>3S$), when was predicted through ANN2-IC₉₀ model. Finally, the ANN2-IC₉₀ model without the compound **19** was the best approximation to the prediction of IC₉₀ values. It was able to describe about 79% of data variance. Moreover, the ANN-IC₉₀ was able to predict the inhibitory activity of unknown compounds with acceptable accuracy ($Q^2_{\text{LOO}} = 0.568$; $Q^2_{\text{L50}} = 0.548$).

According to results reported here, the quality of our linear and non-linear K_i models was superior to IC₉₀ models. The attempt to find a predictive QSAR model including all data set for predicting the antiviral activity was possible excluding an outlier and the consequent ANN model is less predictive than models developed for K_i prediction. We ascribe this difference to the complexity of both activities. Protease inhibition assays bring the measure of enzyme–inhibitor interactions; meanwhile, the antiviral assays also measure the permeation of the compounds through cell membranes. Both processes are independent and can implicate conformational changes in cyclic urea derivatives. Since the antiviral activity encompasses a more complex process, it is expected that the establishment of a relationship among molecular information and IC₉₀ must be more difficult. Since constitutional and 2D descriptors ignore the conformational flexibility of the molecular structures, we think that they do not contain sufficing information for describing the antiviral activity.

3.3. Kohonen self-organizing map

Variables selected in ANN2 models were used to obtain SOMs of both inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC₉₀) by cyclic urea derivatives. In a self-organizing neural network, if two input data vectors are similar, they will be mapped into the same neuron or into neurons close together in the 2D map. We built 9 × 9 Kohonen SOMs.

Figure 2a depicts SOM map for HIV-1 protease inhibition (K_i). 44 of a total of 81 neurons were occupied. Nine neurons were shared for two or more compounds at the same time and five neurons are classified as conflictive taking into account the selected boundary conditions. As it is observed, compounds with a similar range of activities were grouped into neighboring areas. Noteworthy, cyclic urea derivatives with high inhibitory activities were placed in adjacent highly active neurons at the upper-right zone and outstretched across the center of the map. On the other hand, the less active derivatives were placed at adjacent lowly active neurons through the whole left and the lower right area.

The SOM for HIV replications (IC₉₀) is showed in Figure 2b. 36 of a total of 81 neurons were occupied. Sixteen neurons were occupied by two or more compounds at the same time and ten neurons are classified as conflictive taking into account the selected

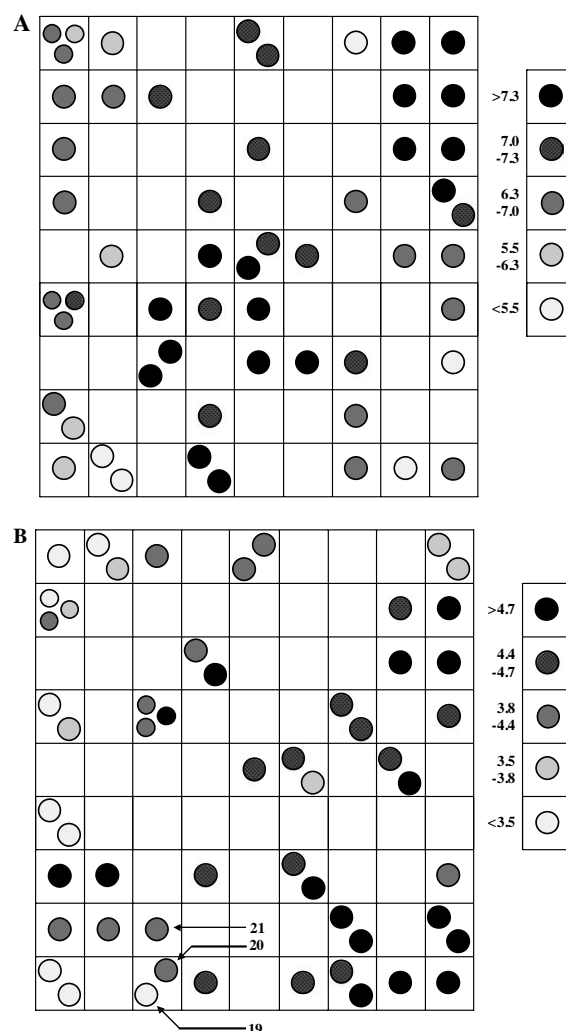


Figure 2. Kohonen SOM for the data set using descriptors from ANN2 models. Squares and circles represent neurons and compounds respectively. Circles at right decode the ranges of inhibitory activities ($\log 10^6/\text{IC}_{50}$). (a) K_i ; (b) IC₉₀.

boundary conditions. The set of variables included in the IC₉₀ model is less effective in the non-linear Kohonen's space. However, several clusters are well defined. The more active compounds are grouped fundamentally at the right and the less active ones are at the upper left.

Since the SOM can be used in order to display the non-linear relationship among data, visual inspection of the map allowed us to identify outlier's neighbourhood. In this sense, we analyzed the location of the compound **19**, which was defined as outlier in previous IC₉₀ ANN modeling. Compound **19** is structurally related to inhibitors represented in Table 2 (18–30), but its antiviral activity is extremely low. The SOM for HIV replications (IC₉₀) (Fig. 2b) shows that compound **19** ($\log(10^6/\text{IC}_{90}) = 3.24$) is narrowly connected to most active compounds **20** ($\log(10^6/\text{IC}_{90}) = 4.23$) and **21** ($\log(10^6/\text{IC}_{90}) = 4.07$), which provide an explanation of why this outlier is not predicted well.

Complex relations are established in the whole data set considering both activities. Actual clustering analysis

shows that there are several groups among the most and less active compounds. Pointing in that way, a QSAR model is an attempt to enclose this complex nature in a simple predictive mathematic model.

3.4. Models' interpretation

In a recent work, Garg and Bhatarai¹⁷ studied the effect of hydrophobicity in K_i and IC_{90} activities of 3-aminoindazole cyclic urea HIV-1 protease inhibitors (analogues of compounds in Table 3) using a linear comparative QSAR. This study evidenced linear dependences between hydrophobicity and the biological activities of the studied compounds, although only statistical significant models could be developed for small data sets with high degree of structural similarity. In addition, they built a linear model for IC_{90} comprising the non-linear effect of $ClogP$ that included eight compounds with a sufficient wide range of this physical–chemical parameter (Eq. 28, Ref. 17). In this connection, a parabolic relationship between IC_{90} and the calculated partition coefficient in octanol/water ($ClogP$) and, consequently, the existence of an interesting optimum $ClogP$ value were reported. The most comprehensive models were able to include only about 27 and 28 compounds (Eqs. 25 and 26, Ref. 17), but without considering non-linear effects of the descriptors. However, in a more recent comparative QSAR report of such authors on HIV-1 protease inhibition by 4-hydroxy-5,6-dihydropyran-2-ones,⁵¹ they were able to report a comprehensive linear model for IC_{50} including 57 compounds and considering the parabolic influence of $ClogP$. From our point of view, these facts strongly suggest that the inhibition of HIV-1 protease by 3-aminoindazole cyclic urea derivatives is rather more complex than inhibition by other chemicals. In this sense, obtaining a statistically significant predictive model, as comprehensive as possible, could be only accomplished inside the ANN framework and searching the relevant structural information among a sufficient varied pool of molecular descriptors.

In our work, the relevant structural information was extracted by means of an exhaustive GA-based search on a pool of 108 descriptors and is contained in the most relevant selected descriptors. This led to a gatherer model, in comparison with the previously reported for Garg and Bhatarai,¹⁷ able to integrate all the compounds in a unique relationship. In our case, if the $ClogP$ term or other molecular property is explicitly used, the QSAR models cannot comprise the whole data set and many compounds should be left out of the analysis. In fact, Garg and Bhatarai studied the antiviral activity of 35 3-aminoindazole cyclic urea derivatives by means of a comparative QSAR, when they attempt to create a comprehensive model, found 8 compounds as outliers. However, by means of GA search, we found one model able to find a relation that includes 54 inhibitors. Considering the prior works, modeling inhibition of HIV-1 protease (K_i) and inhibition of HIV replication (IC_{90}) of cyclic urea derivatives (Refs. 15,17), our report is the first including more than 30 inhibitors in a uniquely predictive model.

Our results corroborated that the employment of 2D descriptors is extremely useful in QSAR studies giving simple correlations between the molecular structures and biological activities.^{52,53} Multiple regression analysis is often employed in such studies in the hope that it might point to structural factors that influence a particular property. They may be viewed in terms of association of activity information content with the structural fragments. However, the interpretation of the information content of these descriptors is very complex as their computations involve integration of the structural fragments and due to this it is impossible to traverse backward from a higher state to a lower one.

As it is indicated in the achieved non-linear model ANN2- K_i , the number of carbons and nitrogens on the inhibitor molecule as well as the spatial distribution of atomic van der Waals volumes are the most relevant factors for the inhibitory activity of the cyclic urea derivatives. The HIV-1 protease has a homodimeric symmetric structure and each monomer contributes one catalytic aspartic residue and flexible flap, which is able to bind the substrates and inhibitors.⁵⁴ Cyclic ureas incorporate the hydrogen-bonding equivalents of an enzyme-bound water molecule into a low-molecular weight, conformationally rigid, seven-member ring system. This structure permits optimal interaction of substituents with corresponding S1, S1', S2, and S2' pockets in the active site of the HIV-1 protease dimer. The urea oxygen forms two long hydrogen bonds with flap residues Ile50 and Ile50', and diol oxygens interact with Asp25 and Asp25'.⁵⁵ The groups vicinal to the diol oxygens occupy the S1/S1' pockets and the urea substituents must occupy the S2/S2' pockets. In addition to the steric fit of molecular frames into the active site, strong hydrogen-bonding interactions provide a remarkable electrostatic complementarity that anchors and perhaps guides the inhibitor into place. In this sense, the number of nitrogens is related to the number of hydrogen-bonding interactions. Furthermore, the number of carbons is closely related to the hydrophobic extension of the inhibitors. On the other hand, the 2D descriptors encode the information about the shape of the compounds. The appearance of atomic van der Waals volumes in our model is justified for the requirements of the cyclic urea derivatives for occupying the HIV-1 protease hydrophobic pockets. At this level, it is suitable to clarify that this analysis cannot offer the specific positions of the atoms since 2D descriptors encode a global and dimension-limited information.

Many highly potent inhibitors of the HIV-1 protease enzyme show modest translation in a whole cell antiviral assay (IC_{90}). For anti-HIV agents an increment in cellular potency has been obtained by modifications designed to improve cell penetration. In this sense, too polar cyclic urea derivatives have a limited cellular activity, while increasing the lipophilicity of the substituents resulted in improved antiviral activity.²⁰ However, with either approach there was a limit on the size of the alkyl groups tolerated. Larger branched substituents significantly decreased binding due to unfavorable steric interactions

and limited further improvements in whole cell antiviral activity.

Model ANN2-IC90 suggests that the presence of five-member rings (nR05) influences the antiviral activity. In this sense, many authors report the best antiviral activities for five-member rings containing HIV-1 protease inhibitors^{56,57} including the clinically approved ritonavir,⁵ amprenavir,⁶ and indinavir.⁷ This positive effect is reflected in our data: all compounds with $\log(10^6/\text{IC}_{90}) > 4.2$ have five-member rings. On the other hand, the presence of the descriptor C11 is a measure of the presence of aromatic rings or ramifications in them, while the selected 2D descriptors (BELm3 and MATS7m) encode the distributions of atomic masses. In general, this set of descriptors must be oriented to the capacity of the drug for cell penetration process. The constitutional descriptors depict the relevance of ring topologies and the 2D descriptors suggest the importance of atomic masses, in such a way that this subset of descriptors encodes the adequate molecular sizes and shapes for transfixing cell membrane.

4. Conclusions

In the last decades, QSAR methodology has demonstrated that the biological activities of drug molecules can be correlated by a linear combination of the chemical and structural information of the corresponding drugs. Neural networks are often superior to the traditional linear QSAR. Their key strength is that with the presence of hidden layers, neural networks are able to perform non-linear mapping of the parameters to the corresponding biological activity implicitly; the results are superior to linear regression analysis when judged in statistical terms and they also provide accurate predictions of activities of the compounds. That is not surprising since QSAR surfaces often have many kinks and wrinkles that cannot be modeled by linear hypersurfaces.

In this work, series of models were developed using QSAR. The models clearly demonstrate a connection between structure and anti-HIV activities of HIV-1 protease inhibitors. The structural information was numerically encoded as molecular descriptors. Objective feature selection and vector space analysis led to development of linear and non-linear models. A non-linear feature selection routine, which combined the genetic algorithm with a neural network fitness evaluator, was used to develop ANN models. These models overcome the linear results according to LOO cross-validation.

This study confirms that K_i and IC_{90} values can be predicted on the basis of simplest molecular structure information. The ANN models can be applied to prediction of inhibitory activities of compounds that are not present in data set used in this study as long as they are structurally similar. The development of non-linear QSAR combined with GA search can lead to more predictive models; therefore, a more accurate structure–ac-

tivity relation can be obtained. The possibility to extend the number of compounds included in a model has been interpreted like a more precise approximation to establish a connection between biological assays and the chemical structure.

Acknowledgments

Authors would like to acknowledge to the anonymous referees for their useful comments that helped to improve the quality of the manuscript. We also greatly acknowledge Professor Rajni Garg for sending valuable information for the development of this work.

References and notes

- Havliř, D. V.; Richman, D. D. *Ann. Intern. Med.* **1996**, *24*, 984.
- Samson, M.; Labbe, O.; Mollereau, C.; Vassart, G.; Parmentier, M. *Biochemistry* **1996**, *35*, 3362.
- Debouck, C. *AIDS Res. Hum. Retrov.* **1992**, *8*, 153.
- Katz, R. A.; Skalka, A. M. *Annu. Rev. Biochem.* **1994**, *63*, 133.
- Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X.-P.; Wideburg, N. E.; Saldivar, A.; Ruiz, L.; Kati, W. M.; Sham, H. L.; Robins, T.; Stewart, K. D.; Hsu, A.; Plattner, J. J.; Leonard, J. M.; Norbeck, D. W. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 2484.
- Reddy, P.; Ross, J. *Formulary* **1999**, *34*, 567.
- (a) Vacca, J. P.; Dorsey, B. D.; Schleif, W. A.; Levin, R. B.; McDaniel, S. L.; Darke, P. L.; Zugay, J.; Quintero, J. C.; Blahy, O. M.; Roth, E.; Sardana, V. V.; Schlabach, A. J.; Graham, P. I.; Condra, J. H.; Gotlib, L.; Holloway, M. K.; Lin, J.; Chen, I.-W.; Vastag, K.; Ostovic, D.; Anderson, P. S.; Emini, E. A.; Huff, J. R. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 4096; (b) Dorsey, B. D.; Levin, R. B.; McDaniel, S. L.; Vacca, J. P.; Guare, J. P.; Darke, P. L.; Zugay, J. A.; Emini, E. A.; Schleif, W. A.; Quintero, J. C.; Lin, J. H.; Chen, I.-W.; Holloway, M. K.; Fitzgerald, P. M. D.; Axel, M. G.; Ostovic, D.; Anderson, P. S.; Huff, J. R. *J. Med. Chem.* **1994**, *37*, 3443.
- Danner, S. A.; Carr, A.; Leonard, J. M.; Lehman, L. M.; Gudiol, F.; Gonzales, J.; Raventos, A.; Rubio, R.; Bouza, E.; Pintado, V.; Gil Aguado, A.; García de Lomas, J.; Delgado, R.; Borleffs, J. C. C.; Hsu, A.; Valdes, J. M.; Boucher, C. A. B.; Cooper, D. A. *N. Eng. J. Med.* **1995**, *333*, 1528.
- Livingston, D. J.; Pazhanisamy, S. D.; Porter, J. T.; Partaledis, J. A.; Tung, R. D.; Painter, G. J. *Infect Diseases* **1995**, *172*, 1238.
- Jayatilake, P. R.; Nair, A. C.; Zauhar, R.; Welsh, W. J. *J. Med. Chem.* **2000**, *43*, 4446.
- Avram, S.; Svab, I.; Bologa, C.; Flonta, M. L. *J. Cell. Mol. Med.* **2003**, *7*, 287.
- Jalali-Heravi, M.; Parastar, F. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147.
- Douali, L.; Villemin, D.; Cherqaoui, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1200.
- Douali, L.; Villemin, D.; Cherqaoui, D. *Curr. Pharm. Des.* **2003**, *9*, 1817.
- Gupta, S. P.; Babu, M. S.; Garg, R.; Sowmya, S. *J. Enzym. Inhib.* **1998**, *13*, 399.

16. Gupta, S. P.; Babu, M. S. *Bioorg. Med. Chem.* **1999**, *7*, 2549.
17. Garg, R.; Bhattacharai, B. *Bioorg. Med. Chem.* **2004**, *12*, 5819.
18. Katritzky, A. R.; Oliferenko, A.; Lomaka, A.; Karelson, M. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3453.
19. Gayathri, P.; Pande, V.; Sivakumar, R.; Gupta, S. P. *Bioorg. Med. Chem.* **2001**, *9*, 3059.
20. Rodgers, J. D.; Lam, P. Y. S.; Johnson, B. L.; Wang, H.; Ko, S. S.; Seitz, S. P.; Trainor, G. L.; Anderson, P. S.; Klabe, R. M.; Bacheler, L. T.; Cordova, B.; Garber, S.; Reid, C.; Wright, M. R.; Chang, C.-H.; Erickson-Viitanen, S. *Chem. Biol.* **1998**, *5*, 597.
21. Patel, M.; Bacheler, L. T.; Rayner, M. M.; Cordova, B. C.; Klabe, R. M.; Erickson-Viitanen, S.; Seitz, S. P. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 823.
22. Patel, M.; Kaltenbach, R. F., III; Nugiel, D. A.; McHugh, R. J., Jr.; Jadhav, P. K.; Bacheler, L. T.; Cordova, B. C.; Klabe, R. M.; Erickson-Viitanen, S.; Garber, S.; Reid, C.; Seitz, S. P. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 1077.
23. Patel, M.; Rodgers, J. D.; McHugh, R. J., Jr.; Johnson, B. L.; Cordova, B. C.; Klabe, R. M.; Bacheler, L. T.; Erickson-Viitanen, S.; Ko, S. S. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 3217.
24. Kaltenbach, R. F., III; Patel, M.; Waltermire, R. E.; Harris, G. D.; Stone, B. R. P.; Klabe, R. M.; Garber, S.; Bacheler, L. T.; Cordova, B. C.; Logue, K.; Wright, M. R.; Erickson-Viitanen, S.; Trainor, G. L. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 605.
25. Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 210.
26. MOPAC version 6.0. Frank, J.; Seiler Research Laboratory, U.S. Air Force academy Colorado Springs, CO, 1993.
27. Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, 2000.
28. Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205.
29. Pogliani, L. *Chem. Rev.* **2000**, *100*, 3827.
30. Moran, P. A. P. *Biometrika* **1950**, *37*, 17.
31. Geary, R. F. *Incorporated Statistician* **1954**, *5*, 115.
32. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
33. Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225.
34. DRAGON Software version 3.0, Milano Chemometrics, 2003.
35. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
36. Holland, H. *Adaption in natural and artificial systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
37. Cartwright, H. M. *Applications of artificial intelligence in chemistry*; Oxford University Press: Oxford, 1993.
38. So, S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
39. MATLAB version 7.0. The MathWorks, Inc. 2004. WEB: www.mathworks.com.
40. Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218.
41. Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.
42. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 905.
43. Kolmogorov, A. N. *Dokl Akad Nauk SSSR* **1957**, *114*, 953.
44. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
45. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.
46. Foresee, F. D.; Hagan, M. T. Gauss-Newton approximation to Bayesian regularization. Proceedings of the 1997 International Joint Conference on Neural Networks, 1997; pp 1930–1935.
47. Kohonen, T. *Biol. Cybern* **1982**, *43*, 59.
48. Gasteiger, J.; Zupan, J. *Angew. Chem. Int. Ed. Engl.* **1995**, *32*, 503.
49. Gasteiger, J.; Li, X. *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 643.
50. Wold, S. *Quant. Struct. Act. Relat.* **1991**, *10*, 191.
51. Bhattacharai, B.; Garg, R. *Bioorg. Med. Chem.* **2005**, *13*, 4078.
52. Fernández, M.; Caballero, J.; Helguera, A. M.; Castro, E. A.; González, M. P. *Bioorg. Med. Chem.* **2005**, *13*, 3269.
53. Gupta, M. K.; Sagar, R.; Shaw, A. K.; Prabhakar, Y. S. *Bioorg. Med. Chem.* **2005**, *13*, 343.
54. Navie, M. A.; Fitzgerald, P. M. D.; Mc Keever, B. M.; Leu, C. T.; Heimbach, J. C.; Herber, W. K.; Sigal, I. S.; Darke, P. L.; Spimge, J. P. *Nature* **1989**, *337*, 615.
55. Hodge, C. N.; Aldrich, P. E.; Bacheler, L. T.; Chang, C.-H.; Eyermann, C. J.; Garber, S.; Grubb, M.; Jackson, D. A.; Jadhav, P. K.; Korant, B.; Lam, P. Y. S.; Maurin, M. B.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Reid, C.; Sharpe, T. R.; Shum, L.; Winslow, D. L.; Erickson-Viitanen, S. *Chem. Biol.* **1996**, *3*, 301.
56. Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; Desolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. *J. Med. Chem.* **1995**, *38*, 305.
57. Pyring, D.; Lindberg, J.; Rosenquist, A.; Zuccarello, G.; Kvarnstrom, I.; Zhang, H.; Vrang, L.; Unge, T.; Classon, B.; Hallberg, A.; Samuelsson, B. *J. Med. Chem.* **2001**, *44*, 3083.