QSAR Based on Multiple Linear Regression and PLS Methods for the Anti-HIV Activity of a Large Group of HEPT Derivatives

Juan M. Luco*,† and Ferdinando H. Ferretti‡

Laboratorio de Alimentos and Departamento de Química, Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis, Chacabuco y Pedernera 5700 San Luis, Argentina

Received December 11, 1996⊗

Quantitative structure−activity relationships have been developed for a set of 107 inhibitors of the HIV-1 reverse transcriptase, derivatives of a recently reported HIV-1 specific lead: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). The activity of these compounds was investigated by means of multiple linear regression (MLR) and PLS regression techniques and topological indexes as well as several tabulated physicochemical substituent constants were used as predictor variables. The results obtained indicate that the anti-HIV activity of the HEPT derivatives is strongly dependent on hydrophobic factors as expressed by the Hansch constant ($\sum\pi$(**R1**+**R2**)), and especially dependent on the geometric factors mainly accounted for by the $^{1}\chi^{N}$(**R2**) and $^{4}\chi_{p}^{N}$ molecular connectivity indexes and also for the molecular volume (Vx), the Taft steric constant (Es(**2R1**)), and the Verloop parameter for the smallest width value (B1(**3R1**)). Besides, for this data set, comparison of the quality of MLR and PLS models show that PLS is a better approach to MLR for improving the interpretability of the data and also to exhibit models with a better predictive quality.

1. INTRODUCTION

The reverse transcriptase (RT) of human immunodeficiency virus 1 (HIV-1) is an essential enzyme required to catalyze the conversion of the viral RNA into proviral DNA and is thus a prime target for antiviral therapy against acquired immune deficiency syndrome (AIDS). In fact, a large number of compounds have been synthesized to target various active sites on this enzyme.[1−5] Among them, several 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives have proved to be potent and selective inhibitors of HIV- RT.[6,7]

Extensive synthetic and structure−activity studies have been carried out on these compounds,[8−14] and recently, Hansch and Zhang[15] have derived a quantitative structure−activity relationship (QSAR) study for a set of 33 HEPT analogues. Furthermore, very recently the X-ray crystal structure of the enzyme and enzyme-inhibitor complex was reported,[16,17] and conventional QSAR and 3D-QSAR models for the same set of compounds have been derived.[18,19] The overall picture which emerges from these studies shows that the hydrophobic and principally the steric characteristics of substituents have a predominant role in the anti-HIV-1 activity of these compounds.

Among different methods for quantifying the size and the shape characteristics of molecules, those based on the chemical graph theory such as molecular connectivity have been found useful in establishing quantitative structure−activity and structure−property relationships.[20−23]

There are, however, difficulties in the use of these approaches when the QSARs are derived by means of multiple linear regression (MRL) techniques. Commonly, the topological descriptors are mutually interrelated by simple or multiple correlations, and, therefore, fortuitous or artifactual MLR models may be obtained. The disadvantage of the MLR method has recently been overcome through the development of a partial least squares (PLS) method (see ref 24 and references therein) which is insensitive to the collinearity among the variables and also offers the advantage of handling data sets where the number of independent variables is greater than the number of observations. Furthermore, because the complexity of a PLS model is determined by cross-validation, the probability of obtaining chance correlations is kept to a minimum. In this report, both PLS and MLR techniques are used for modeling the observed anti-HIV-1 activity of 80 HEPT derivatives. These represent most of the compounds of this class for which precise activity data are available. The adequacy of the developed QSAR models was examined by means of cross-validation and also by calculating a "total" QSAR ($n = 107$) based on the 80 compounds originally considered as well as 27 HEPT derivatives for whose activity imprecise data have been reported. This is the largest group of HEPT analogues to be included in a single QSAR to date. General structure of compounds under study is given in Figure 1. Among the several descriptors used in the characterization of the compounds such as tabulated physicochemical substituent constants, it was of interest, as already mentioned, to evaluate whether the molecular connectivity-type topological indexes, applied to whole molecules or to structural fragments, can account for the role of geometric factors in influencing the anti-HIV-1 activity of the compounds under study. Finally, it seemed interesting to compare the quality of QSARs derived by means of PLS and MLR analysis.

2. MATERIALS AND METHODS

2.1. Biological Data. The chemical structures along with observed and calculated activity data of the compounds used in this study are shown in Table 1. The activity data were taken from various studies by H. Tanaka and co-workers.[7−14] The log 1/*C* values were used as dependent variable in which *C* represents the molar concentration of drug required to

* To receive all correspondence.
† Laboratorio de Alimentos.
‡ Departamento de Quimica.
⊗ Abstract published in *Advance ACS Abstracts,* February 15, 1997.

S0095-2338(96)00487-8 CCC: $14.00

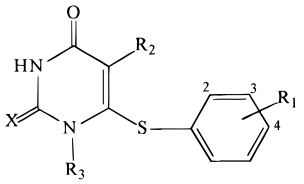

Figure 1. General structure of HEPT derivatives.

achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (HTLV-III**B** strain).

2.2. Structural Descriptors. Several structural descriptors and physicochemical variables were used to characterize the HEPT derivatives under study.

In describing the chemical and structural variation of the **R1** and **R2** substituents, the following physicochemical descriptors[25] were considered: the Hansch constant, π(**R1**) and $\sum\pi(\pi(\mathbf{R1})+\pi(\mathbf{R2}))$, the Hammett constant, σ(**R1**), the Taft steric constant for ortho substituents, Es(**2R1**), and molar refractivity, MR(**R1**). In addition, to account for the total H-bond capacity of the **R1** substituents, the parameter Λ_{alk} as defined by Van de Waterbeemd and Kansy was used.[26]

The next and large group of nonempirical structural descriptors considered were topological indexes. Calculations were performed with the Molconn-X program, which was obtained from L. H. Hall, Eastern, Nazarene College, Quincy, MA. The following topological indexes calculated on whole molecules were considered: the Wiener index,[27] the valence and connectivity molecular indexes,[28] $^{m}\chi$'s, including path, path/cluster, and chain types to the fifth order, and the differential molecular connectivity indexes of zeroth and first order ($^{0}\Delta$X and $^{1}\Delta$X) selected as electronic parameters based on graph theory.[29] To quantify the structural variation in the acyclic structure (**R3**) of HEPT, the following molecular connectivity indexes to the second order were considered: $^{0}\chi$(**R3**), $^{0}\chi^{v}$(**R3**), $^{1}\chi$(**R3**), $^{1}\chi^{v}$(**R3**), $^{2}\chi$-(**R3**), $^{2}\chi v$(**R3**), $^{0}\Delta$X(**R3**), $^{1}\Delta$X(**R3**). On the other hand, the structural variation at the C-5 position (**R2**) was described by the index of connectivity level, $^{1}\chi^{N}$(**R2**), which was calculated by dividing the value of the $^{1}\chi$(**R2**) index by the number of atoms involved in their calculus in order to minimize the role of substituent size. The last group of structural descriptors considered in this study included: molecular volume, molecular weight, the Verloop steric parameters[30] such as L**R2** (length of **R2**) and B1−B4(**3R1**) (width parameters of **R1** at the 3-position), and the indicator variable, Ind-S, that takes the value zero for X = oxygen and the unity for X = sulfur (see Table 1). Molecular volume was approximated by using the McGowan characteristic volume (denoted by Vx), which can easily be calculated for any solute simply from a knowledge of its molecular structure.[31,32]

2.3. Statistical Methods. The partial least squares projections in latent variables (PLS) together with multiple regression analysis were the methods here used to search for relationships between the biological activity data and the structural descriptors. Since detailed descriptions of the PLS method have been published, a brief outline is given here. PLS is based on the projection of the original multivariate data matrices down onto smaller matrices (**T,U**) with orthogonal columns, which relates the information in the response matrix **Y** to the systematic variance in the descriptor matrix **X**, as shown below

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{TP}' + \mathbf{E}$$

$$\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{UC}' + \mathbf{F}$$

$$\mathbf{U} = \mathbf{T} + \mathbf{H} \text{ (the inner relation)}$$

where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the corresponding mean value matrices, **T** and **U** are the matrices of scores that summarize the *x* and *y* variables respectively, **P** is the matrix of loadings showing the influence of the *x* variables in each component, **C** is the matrix of weights expressing the correlation between **Y** and **T**(**X**) and **E**, **F**, and **H** are the corresponding residuals matrices. The PLS calculations also give an auxiliary matrix **W** (PLS weights), which expresses the correlation between **U** and **X** and is used to calculate **T**. Determination of the significant number of model dimensions (A) was made by cross-validation.[33]

PLS analysis was carried out using the SIMCA-S 5.1a software package obtained from Umetri AB, Box 7960, 907 19 Umea, Sweden. Regression analysis was performed by using the 7.0 version of Statgraphics Plus. All statistical and Molconn-X calculations were carried out on a PC-IBM Computer.

3. RESULTS AND DISCUSSION

3.1. Multiple Regression Analysis. Table 2 lists the physicochemical constants and structural descriptors included in the analysis. It seems appropriate to show the equation derived by Hansch and Zhang[15] for comparison with the results presented in this study. The reported correlation equation, which was checked from original paper, was as follows:

$$\log 1/C = 0.88(0.19)\sum\pi(\mathbf{R1}+\mathbf{R2}) + 12.1(1.85)\mathrm{L}(\mathbf{R2}) - 1.59(0.24)(\mathrm{L}(\mathbf{R2}))^2 - 1.17(0.42)\mathrm{B1}\text{-}\mathbf{3R1} + 1.53(0.40)\mathrm{Es}\text{-}\mathbf{2R1} - 15.30(3.61)$$

$$r = 0.944 \quad Q = 0.824 \quad s = 0.50 \quad n = 33 \quad F = 42.14 \qquad (1)$$

Note that the coefficient of B1-**3R1** is negative, in contrast to the positive sign which was reported. Equation 1 was derived for compounds **1**−**33** shown in Table 1. Compound **34** was excluded by the authors, because it fit poorly into the equation. In this and the following correlation equations, *n* is the number of compounds, *s* is the standard deviation, *r* is the correlation coefficient, *F* is the Fisher test for significance of the equation and the figures in parentheses are the standard deviations of the coefficients. The *Q* coefficient is the cross-validated *r* which describes the predictive power of the model. This was calculated with the SIMCA-S software by the leave-one-out procedure. It should be noted that, for the case in which one dependent variable is analyzed and the number of extracted PLS components is equal to the number of independent variables, the results from regression and PLS analysis are numerically identical. MLR was performed on compounds **1**−**80** described in Table 1. Because of the large number of descriptors considered, a stepwise multiple regression procedure based on the forward-selection and backward-elimination methods was used for inclusion or rejection of descriptors in the screened models. In order to avoid overestimations or difficulties in interpretation of the resulting

Table 1. Chemical Structures and Observed and Calculated Anti-HIV-Activities of HEPT Derivatives

						PLS		MRL	
No.	R_1	R_2	R_3	X	obsd[a]	calcd[b]	calcd[b]	calcd[c]	calcd[c]
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	3.82	3.79	3.84	3.72
2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	4.18	4.03	4.10	3.97
3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	4.89	4.79	4.94	4.90
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.43	5.22	5.44	5.30
5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.52	5.47	5.65	5.70
6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	4.98	4.79	4.93	4.77
7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	4.74	4.41	4.62	4.33
8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	5.33	4.95	5.29	4.94
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.34	5.07	5.26	5.04
10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.34	5.11	5.24	5.08
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.37	5.19	5.26	5.16
12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.76	4.31	4.57	4.21
13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	5.09	4.59	4.93	4.51
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	5.23	4.98	5.23	5.08
15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	6.26	6.35	6.42	6.64
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	6.20	6.26	6.28	6.45
17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.31	6.43	6.50	6.77
18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	4.80	4.44	4.63	4.39
19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	4.64	4.15	4.36	4.01
20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	4.93	4.46	4.72	4.36
21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.96	5.72	5.68	5.44
22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	6.88	6.73	6.74	6.56
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	6.17	6.07	6.01	5.93
24	H	*i*-Pr	$CH_2OCH_2CH_2OH$	S	7.23	7.33	7.37	7.32	7.35
25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.82	7.92	7.76	7.85
26	3,5-Me_2	*i*-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.23	8.39	8.30	8.42
27	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	7.77	7.86	7.64	7.71
28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	6.84	6.61	6.66	6.37
29	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	6.12	6.00	5.93	5.80
30	H	*i*-Pr	$CH_2OCH_2CH_2OH$	O	7.20	7.28	7.29	7.24	7.22
31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.77	7.88	7.68	7.78
32	3,5-Me_2	*i*-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.18	8.39	8.22	8.41
33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	7.72	7.81	7.56	7.62
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66		4.01		4.04
35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	5.23	4.87	5.30	4.94
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	5.28	5.03	5.38	5.20
37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	5.46	5.34	5.66	5.59
38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	6.75	6.45	6.53	6.14
39	H	$CH{=}CHPh$	$CH_2OCH_2CH_2OH$	O	5.22	4.85	4.91	4.75	5.02
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	5.04	5.04	4.81	5.07
41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	5.70	5.29	6.06	5.48
42	H	Me	$CH_2OCH_2CH_2OMe$	O	5.06	5.18	4.95	5.35	5.21
43	H	Me	$CH_2OCH_2CH_2OAc$	O	5.17	4.56	4.30	4.62	4.43
44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	5.59	5.30	5.66	5.41
45	H	Me	CH_2OCH_2Me	O	6.48	5.58	5.21	5.75	5.41
46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.49	5.26	5.73	5.60
47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	4.74	4.43	4.74	4.46
48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	5.19	4.77	5.23	4.78
49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.45	5.23	5.67	5.55
50	H	Me	CH_2OCH_2Ph	O	7.06	6.25	6.27	6.42	6.65
51	H	Et	CH_2OCH_2Me	O	7.72	7.20	7.01	7.14	6.89
52	H	Et	CH_2OCH_2Me	S	7.58	7.25	7.15	7.22	7.11
53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.16	8.34	8.17	8.40
54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.20	8.39	8.26	8.49
55	H	Et	CH_2OCH_2Ph	O	8.23	7.71	7.70	7.64	7.66
56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	8.50	8.46	8.51	8.40
57	H	Et	CH_2OCH_2Ph	S	8.09	7.76	7.70	7.72	7.67
58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	8.55	8.37	8.59	8.28
59	H	*i*-Pr	CH_2OCH_2Me	O	7.99	7.65	7.71	7.72	7.79
60	H	*i*-Pr	CH_2OCH_2Ph	O	8.51	8.10	8.14	8.16	8.19
61	H	*i*-Pr	CH_2OCH_2Me	S	7.89	7.70	7.81	7.80	7.94
62	H	*i*-Pr	CH_2OCH_2Ph	S	8.14	8.15	8.10	8.24	8.14
63	H	Me	CH_2OMe	O	5.68	5.95	5.43	6.08	5.49
64	H	Me	CH_2OBu	O	5.33	5.31	5.17	5.58	5.58
65	H	Me	Et	O	5.66	6.45	5.91	6.64	6.02
66	H	Me	Bu	O	5.92	5.71	5.48	5.98	5.85
67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.15	8.33	8.13	8.34
68	H	Et	CH_2O-*i*-Pr	S	6.66	6.86	6.76	6.87	6.77
69	H	Et	CH_2O-c-Hex	S	5.79	5.98	6.05	6.06	6.11
70	H	Et	CH_2OCH_2-c-Hex	S	6.45	5.89	5.82	6.01	5.87
71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.56	7.34	7.57	7.30
72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.59	7.40	7.63	7.41
73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	7.59	7.38	7.60	7.33
74	3,5-Cl_2	Et	CH_2OCH_2Me	O	8.13	8.10	8.27	8.05	8.23
75	H	Et	CH_2O-*i*-Pr	O	6.47	6.81	6.66	6.78	6.62
76	H	Et	CH_2O-c-Hex	O	5.40	5.94	6.05	5.98	6.10
77	H	Et	CH_2OCH_2-c-Hex	O	6.35	5.84	5.86	5.93	5.91
78	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	7.55	7.42	7.52	7.38
79	H	c-Pr	CH_2OCH_2Me	S	7.02	6.79	6.73	6.61	6.65
80	H	c-Pr	CH_2OCH_2Me	O	7.00	6.74	6.60	6.53	6.45

Table 1 (Continued)

						PLS		MRL	
No.	R_1	R_2	R_3	X	obsd[a]	calcd[b]	calcd[b]	calcd[c]	calcd[c]
81	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-*n*	O	<4.46		4.47		4.89
82	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	<3.89		4.20		4.21
83	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	<3.53		4.33		4.24
84	4-F	Me	$CH_2OCH_2CH_2OH$	O	<3.60		3.61		3.48
85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	<3.60		4.06		4.09
86	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	<3.72		3.64		3.59
87	4-CN	Me	$CH_2OCH_2CH_2OH$	O	<3.60		3.58		3.48
88	4-OH	Me	$CH_2OCH_2CH_2OH$	O	<3.56		3.26		3.06
89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	<3.60		3.92		3.92
90	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	<3.96		3.63		3.62
91	3-COOH	Me	$CH_2OCH_2CH_2OH$	O	<3.45		4.41		4.37
92	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	<3.51		3.85		3.68
93	H	COOMe	$CH_2OCH_2CH_2OH$	O	<5.18		5.30		4.87
94	H	CONHPh	$CH_2OCH_2CH_2OH$	O	<4.74		4.72		4.58
95	H	SPh	$CH_2OCH_2CH_2OH$	O	<4.68		5.20		5.27
96	H	C≡CH	$CH_2OCH_2CH_2OH$	O	<4.74		6.17		5.77
97	H	C≡C-Ph	$CH_2OCH_2CH_2OH$	O	<5.47		4.90		5.01
98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	<3.60		4.26		4.10
99	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	<4.92		5.23		4.91
100	H	COPh	$CH_2OCH_2CH_2OH$	O	<4.89		5.37		5.31
101	H	C≡CMe	$CH_2OCH_2CH_2OH$	O	<4.72		5.60		5.27
102	H	F	$CH_2OCH_2CH_2OH$	O	<4.00		4.42		4.31
103	H	Cl	$CH_2OCH_2CH_2OH$	O	<4.52		4.91		4.99
104	H	Br	$CH_2OCH_2CH_2OH$	O	<4.70		5.08		5.23
105	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	<4.70		5.49		5.79
106	H	Me	H	O	<3.60		4.22		3.83
107	H	Me	Me	O	<3.82		4.66		4.50

[a] Logarithm of reciprocal molar concentration required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1, from refs 6−13. [b] From the three-dimensional PLS models derived with 79 (PLS-1) and 107 (PLS-2) compounds. [c] From the regression eqs 4 and 5, respectively.

models, pairs of variables with an $r \geq 0.70$ were classified as intercorrelating ones, and only one of these was included in the screened model.[34] After some consideration, the following equation was selected:

$$\log 1/C = 0.63(0.14)\sum\pi(\mathbf{R1}+\mathbf{R2}) + 209.5(16.3)^1\chi^{\mathbf{N}}(\mathbf{R2}) - 308.1(23.7)\,(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2 - 1.53(0.23)\text{B1-}\mathbf{3R1} + 1.91(0.33)\text{Es-}\mathbf{2R1} + 1.14(0.16)^4\chi_p - 26.97(3.80)^6\chi^v_{ch} - 1.21(0.23)^0\Delta X(\mathbf{R3}) - 32.46(2.76)$$

$$r = 0.929 \quad Q = 0.826 \quad s = 0.509 \quad n = 79 \quad F = 55.78 \qquad (2)$$

The statistical quality of eq 2 is good and accounts for 86.4% of the variance in log 1/*C*. Compound **34** was not included here, nor in the next two presented equations. The reason for this is that none of the descriptors used in this study can account for the complete loss of activity that results from the introduction of substituents at the 4-position of the 6-(phenylthio) moiety of HEPT. On the other hand, it is not possible to use an indicator variable since, for the data set considered ($n = 80$), compound **34** is the only one with this substitution pattern. Another interesting point to highlight is the lower level of statistical significance that is obtained when L(**R2**) is used to derive eq 2 instead of $^1X^{\mathbf{N}}(\mathbf{R2})$ ($r = 0.877$, $s = 0.67$, $F = 28.43$).

The evaluation of the descriptor weights for this equation (see Table 5) shows that primarily the two terms in $^1\chi^{\mathbf{N}}(\mathbf{R2})$ followed by the $^4\chi_p$ index in second place make the highest contributions to the variation in log 1/*C*. The $^1\chi^{\mathbf{N}}(\mathbf{R2})$ index is a bond additive descriptor in which the influence of substituent size has been minimized, and their presence in the developed equation shows that the activity is strongly dependent on the structural variation of **R2** substituents (see below in section 3.4).

On the other hand, the $^4\chi_p$ index describes a weighted count of all fragments or subgraphs consisting of four bonds joined as a path. As a result of their mathematical definition, this and almost all molecular connectivity indexes are size-dependent quantities. The high correlation between molecular volume (Vx) and the $^4\chi_p$ index provides a good example of this fact (see Table 3). However, a careful examination of the $^4\chi_p$ values shows that the discrimination among compounds produced by this index is not based entirely on size but also on other structural factors such as branching, cyclization, and substitution pattern. Hence, an attempt to treat separately the information encoded by $^4\chi_p$ was carried out. Consequently, both the $^4\chi_p^{\mathbf{N}}$ index (normalized to the compound's size and calculated by dividing the $^4\chi_p$ value by the number of atoms involved in their calculus) and Vx were used as independent variables instead of $^4\chi_p$. Supporting the above hypothesis, the equation of minimal complexity in which all terms are of high significance was as follows:

$$\log 1/C = 99.7(18.3)^1\chi^{\mathbf{N}}(\mathbf{R2}) - 146.7(25.2)(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2 - 0.85(0.19)\text{B1-}\mathbf{3R1} + 1.74(0.31)\text{Es}-\mathbf{2R1} + 73.81(7.60)^4\chi_p^{\mathbf{N}} - 25.01(3.13)^6\chi^v_{ch} + 1.40(0.33)\text{Vx}/100 - 31.78(2.63)$$

$$r = 0.934 \quad Q = 0.844 \quad s = 0.493 \quad n = 79 \quad F = 68.81 \qquad (3)$$

Equation 3 clearly shows the positive dependence of log 1/*C* on the molecular size as reflected by Vx and also the large influence of the differents aspects of the molecular shape

Table 2. Molecular Descriptors[a] of HEPT and Its Analogues

no.	$\sum\pi$(**R1**+**R2**)	$^{1}\chi^{N}$(**R2**)	$^{4}\chi_{p}^{N}$	Vx	B1-**3R1**	Es-**2R1**	$^{0}\Delta\chi$(**R3**)	$^{6}\chi_{ch}^{\nu}$	Ich-**R3**	I-**4R1**
1	1.12	0.289	0.2646	234.5	1.00	−1.24	0.852	0.042	0	0
2	0.28	0.289	0.2702	237.8	1.00	−1.01	0.852	0.042	0	0
3	0.54	0.289	0.2700	240.4	1.00	−0.55	0.852	0.042	0	0
4	1.12	0.289	0.2676	234.5	1.52	0.00	0.852	0.042	0	0
5	1.58	0.289	0.2652	248.6	1.52	0.00	0.852	0.042	0	0
6	2.54	0.289	0.2562	276.7	2.60	0.00	0.852	0.042	0	0
7	1.40	0.289	0.2562	239.8	1.99	0.00	0.852	0.042	0	0
8	0.70	0.289	0.2676	222.2	1.35	0.00	0.852	0.042	0	0
9	1.27	0.289	0.2676	232.6	1.80	0.00	0.852	0.042	0	0
10	1.42	0.289	0.2676	237.9	1.95	0.00	0.852	0.042	0	0
11	1.68	0.289	0.2676	246.2	2.15	0.00	0.852	0.042	0	0
12	0.28	0.289	0.2610	237.8	1.70	0.00	0.852	0.042	0	0
13	−0.11	0.289	0.2676	226.3	1.35	0.00	0.852	0.042	0	0
14	0.54	0.289	0.2652	240.3	1.35	0.00	0.852	0.042	0	0
15	1.68	0.289	0.2791	248.6	1.52	0.00	0.852	0.038	0	0
16	1.96	0.289	0.2791	244.9	1.80	0.00	0.852	0.038	0	0
17	1.68	0.289	0.2791	259.1	1.52	0.00	0.852	0.038	0	0
18	0.55	0.289	0.2604	256.0	1.90	0.00	0.852	0.042	0	0
19	0.01	0.289	0.2610	250.1	1.90	0.00	0.852	0.042	0	0
20	−0.01	0.289	0.2652	235.9	1.60	0.00	0.852	0.042	0	0
21	1.10	0.404	0.2702	244.3	1.00	0.00	0.852	0.046	0	0
22	1.02	0.372	0.2768	244.9	1.00	0.00	0.852	0.046	0	0
23	1.55	0.404	0.2702	259.1	1.00	0.00	0.852	0.046	0	0
24	1.53	0.372	0.2812	259.1	1.00	0.00	0.852	0.046	0	0
25	2.14	0.372	0.2905	273.1	1.52	0.00	0.852	0.038	0	0
26	2.65	0.372	0.2940	287.2	1.52	0.00	0.852	0.038	0	0
27	2.44	0.372	0.2905	269.4	1.80	0.00	0.852	0.038	0	0
28	1.02	0.372	0.2768	234.5	1.00	0.00	0.852	0.046	0	0
29	1.55	0.404	0.2702	248.6	1.00	0.00	0.852	0.046	0	0
30	1.53	0.372	0.2812	248.6	1.00	0.00	0.852	0.046	0	0
31	2.14	0.372	0.2905	262.7	1.52	0.00	0.852	0.038	0	0
32	2.65	0.372	0.2940	276.7	1.52	0.00	0.852	0.038	0	0
33	2.44	0.372	0.2905	258.9	1.80	0.00	0.852	0.038	0	0
34	1.12	0.289	0.2581	234.5	1.00	0.00	0.852	0.042	0	1
35	0.56	0.289	0.2636	220.4	1.00	0.00	0.852	0.046	0	0
36	0.56	0.289	0.2636	230.9	1.00	0.00	0.852	0.046	0	0
37	1.12	0.289	0.2636	232.1	1.00	0.00	0.852	0.046	0	0
38	0.82	0.372	0.2768	230.2	1.00	0.00	0.852	0.046	0	0
39	2.68	0.459	0.2781	290.9	1.00	0.00	0.852	0.078	0	0
40	2.01	0.454	0.2814	281.2	1.00	0.00	0.852	0.078	0	0
41	4.64	0.474	0.2941	351.7	1.00	0.00	0.852	0.110	0	0
42	0.56	0.289	0.2597	234.5	1.00	0.00	0.598	0.046	0	0
43	0.56	0.289	0.2506	250.1	1.00	0.00	1.267	0.046	0	0
44	0.56	0.289	0.2713	296.8	1.00	0.00	1.993	0.078	0	0
45	0.56	0.289	0.2673	214.5	1.00	0.00	0.299	0.046	0	0
46	0.56	0.289	0.2636	226.8	1.00	0.00	0.165	0.046	0	0
47	0.56	0.289	0.2561	235.9	1.00	0.00	1.371	0.046	0	0
48	0.56	0.289	0.2636	216.3	1.00	0.00	0.921	0.046	0	0
49	0.56	0.289	0.2636	228.6	1.00	0.00	0.299	0.046	0	0
50	0.56	0.289	0.2809	261.2	1.00	0.00	1.025	0.078	0	0
51	1.02	0.372	0.2809	228.6	1.00	0.00	0.299	0.046	0	0
52	1.02	0.372	0.2809	239.1	1.00	0.00	0.299	0.046	0	0
53	2.14	0.372	0.2949	256.8	1.52	0.00	0.299	0.038	0	0
54	2.14	0.372	0.2949	267.3	1.52	0.00	0.299	0.038	0	0
55	1.02	0.372	0.2914	275.3	1.00	0.00	1.025	0.078	0	0
56	2.14	0.372	0.3021	303.5	1.52	0.00	1.025	0.070	0	0
57	1.02	0.372	0.2914	285.8	1.00	0.00	1.025	0.078	0	0
58	2.14	0.372	0.3021	314.0	1.52	0.00	1.025	0.070	0	0
59	1.53	0.372	0.2853	242.7	1.00	0.00	0.299	0.046	0	0
60	1.53	0.372	0.2947	289.4	1.00	0.00	1.025	0.078	0	0
61	1.53	0.372	0.2853	253.2	1.00	0.00	0.299	0.046	0	0
62	1.53	0.372	0.2947	299.9	1.00	0.00	1.025	0.078	0	0
63	0.56	0.289	0.2757	200.4	1.00	0.00	0.299	0.046	0	0
64	0.56	0.289	0.2597	242.7	1.00	0.00	0.299	0.046	0	0
65	0.56	0.289	0.2840	194.6	1.00	0.00	0.000	0.046	0	0
66	0.56	0.289	0.2673	222.7	1.00	0.00	0.000	0.046	0	0
67	2.44	0.372	0.2949	263.6	1.80	0.00	0.299	0.038	0	0
68	1.02	0.372	0.2720	253.2	1.00	0.00	0.299	0.046	0	0
69	1.02	0.372	0.2945	284.6	1.00	0.00	0.299	0.148	1	0
70	1.02	0.372	0.2914	298.7	1.00	0.00	0.299	0.148	1	0
71	1.02	0.372	0.2859	299.9	1.00	0.00	0.973	0.074	0	0
72	1.02	0.372	0.2859	298.0	1.00	0.00	0.839	0.074	0	0
73	1.02	0.372	0.2869	299.9	1.00	0.00	1.025	0.078	0	0
74	2.44	0.372	0.2949	253.1	1.80	0.00	0.299	0.038	0	0
75	1.02	0.372	0.2720	242.7	1.00	0.00	0.299	0.046	0	0
76	1.02	0.372	0.2945	274.1	1.00	0.00	0.299	0.148	1	0
77	1.02	0.372	0.2914	288.2	1.00	0.00	0.299	0.148	1	0
78	1.02	0.372	0.2869	289.4	1.00	0.00	1.025	0.078	0	0
79	1.14	0.412	0.2875	242.3	1.00	0.00	0.299	0.046	0	0
80	1.14	0.412	0.2875	231.8	1.00	0.00	0.299	0.046	0	0

Table 2 (Continued)

no.	$\Sigma\pi$(**R1**+**R2**)	$^1\chi^{\mathbf{N}}$(**R2**)	$^4\chi_p^{\mathbf{N}}$	Vx	B1-**3R1**	Es-**2R1**	$^0\Delta\chi$(**R3**)	$^6\chi^{v}_{ch}$	Ich-**R3**	I-**4R1**
81	0.56	0.289	0.2469	290.8	1.00	0.00	0.598	0.046	0	0
82	1.27	0.289	0.2646	232.6	1.00	−0.97	0.852	0.042	0	0
83	−0.47	0.289	0.2652	240.3	1.52	0.00	0.852	0.042	0	0
84	0.70	0.289	0.2581	222.2	1.00	0.00	0.852	0.042	0	1
85	1.27	0.289	0.2581	232.6	1.00	0.00	0.852	0.042	0	1
86	0.28	0.289	0.2585	237.8	1.00	0.00	0.852	0.042	0	1
87	−0.01	0.289	0.2604	235.9	1.00	0.00	0.852	0.042	0	1
88	−0.11	0.289	0.2581	226.3	1.00	0.00	0.852	0.042	0	1
89	0.54	0.289	0.2604	240.3	1.00	0.00	0.852	0.042	0	1
90	0.01	0.289	0.2585	250.1	1.00	0.00	0.852	0.042	0	1
91	0.24	0.289	0.2610	241.9	1.60	0.00	0.852	0.042	0	0
92	−0.93	0.289	0.2610	246.0	1.60	0.00	0.852	0.042	0	0
93	−0.01	0.405	0.2738	241.9	1.00	0.00	0.852	0.046	0	0
94	0.49	0.454	0.2898	292.7	1.00	0.00	0.852	0.078	0	0
95	2.32	0.454	0.2814	283.4	1.00	0.00	0.852	0.078	0	0
96	0.40	0.372	0.2768	225.9	1.00	0.00	0.852	0.046	0	0
97	2.65	0.459	0.2781	288.7	1.00	0.00	0.852	0.078	0	0
98	−0.67	0.289	0.2676	230.4	1.50	0.00	0.852	0.042	0	0
99	0.06	0.400	0.2667	264.2	1.00	0.00	0.852	0.046	0	0
100	1.05	0.451	0.2928	282.7	1.00	0.00	0.852	0.078	0	0
101	0.96	0.404	0.2702	240.0	1.00	0.00	0.852	0.046	0	0
102	0.14	0.289	0.2636	208.1	1.00	0.00	0.852	0.046	0	0
103	0.71	0.289	0.2636	218.5	1.00	0.00	0.852	0.046	0	0
104	0.86	0.289	0.2636	223.8	1.00	0.00	0.852	0.046	0	0
105	0.56	0.289	0.2696	295.3	1.00	0.00	1.324	0.078	0	0
106	0.56	0.289	0.2690	166.4	1.00	0.00	0.000	0.048	0	0
107	0.56	0.289	0.2692	180.5	1.00	0.00	0.000	0.046	0	0

[a] For explanation of the symbols of molecular descriptors see text.

Table 3. Correlation Matrix for the Molecular Descriptors Applied in This Study[a,b]

	$\Sigma\pi$(**R1**+**R2**)	$^1\chi^{N}$**R2**	$^4\chi^{\mathbf{N}}_p$	$^4\chi_p$	Vx	B1-**3R1**	Es-**2R1**	$^0\Delta$X(**R3**)	$^6\chi^{v}_{ch}$	Ich-**R3**
$\Sigma\pi$(**R1**+**R2**)	1.00	0.55	0.58	0.52	0.52	0.26	0.05	−0.02	0.15	−0.01
	(1.00)	(0.59)	(0.56)	(0.57)	(0.58)	(0.34)	(0.13)	(0.03)	(0.05)	(−0.07)
$^1\chi^{\mathbf{N}}$(**R2**)	0.55	1.00	0.69	0.66	0.58	−0.22	0.17	−0.06	0.41	0.13
	(0.59)	(1.00)	(0.69)	(0.62)	(0.59)	(−0.27)	(0.19)	(−0.11)	(0.37)	(0.15)
$^4\chi^{\mathbf{N}}_p$	0.58	0.69	1.00	0.68	0.56	0.00	0.11	−0.15	0.47	0.27
	(0.56)	(0.69)	(1.00)	(0.70)	(0.61)	(−0.08)	(0.14)	(−0.16)	(0.43)	(0.28)
$^4\chi_p$	0.52	0.66	0.68	1.00	**0.95**	0.02	0.09	0.36	0.62	0.21
	(0.57)	(0.62)	(0.70)	(1.00)	**(0.96)**	(−0.01)	(0.09)	(0.34)	(0.60)	(0.23)
Vx	0.52	0.58	0.56	**0.95**	1.00	0.06	0.10	0.36	0.59	0.23
	(0.58)	(0.59)	(0.61)	**(0.96)**	(1.00)	(0.03)	(0.12)	(0.34)	(0.60)	(0.26)
B1-**3R1**	0.26	−0.22	0.00	0.02	0.06	1.00	0.12	0.09	−0.31	−0.12
	(0.34)	(−0.27)	(−0.08)	(−0.01)	(0.03)	(1.00)	(0.13)	(0.12)	(−0.35)	(−0.16)
Es-**2R1**	0.05	0.17	0.11	0.09	0.10	0.12	1.00	−0.06	0.10	0.04
	(0.13)	(0.19)	(0.14)	(0.09)	(0.12)	(0.13)	(1.00)	(−0.07)	(0.10)	(0.04)
$^0\Delta$X(**R3**)	−0.02	−0.06	−0.15	0.36	0.36	0.09	−0.06	1.00	−0.02	−0.28
	(0.03)	(−0.11)	(−0.16)	(0.34)	(0.34)	(0.12)	(−0.07)	(1.00)	(−0.04)	(−0.29)
$^6\chi^{v}_{ch}$	0.15	0.41	0.47	0.62	0.59	−0.31	0.10	−0.02	1.00	**0.79**
	(0.05)	(0.37)	(0.43)	(0.60)	(0.60)	(−0.35)	(0.10)	(−0.04)	(1.00)	**(0.83)**
Ich-**R3**	−0.01	0.13	0.27	0.21	0.23	−0.12	0.04	−0.28	**0.79**	1.00
	(−0.07)	(0.15)	(0.28)	(0.23)	(0.26)	(−0.16)	(0.04)	(−0.29)	**(0.83)**	(1.00)
I-**4R1**	−0.20	−0.25	−0.34	−0.18	−0.16	−0.17	0.05	0.10	−0.14	−0.14

[a] Full data set, 107 compounds. [b] The figures in parentheses are the correlations for the reduced data set (79 compounds).

encoded by $^4\chi_p^{\mathbf{N}}$ (see section 3.4). Furthermore, this equation provides evidence on the dominant role of steric effects for the anti-HIV-1 activity of this series of HEPT analogues.

The chain terms ($^6\chi^{v}_{ch}$) included in the above developed equations encodes information about the number and type of six-membered rings present in the molecule. For this data set, however, this index primarily indicates the presence of a saturated or aromatic ring in the acyclic structure (**R3**) of HEPT. Hence, it is possible to use an indicator variable to account for these structural features which has the advantage of simplicity over $^6\chi^{v}_{ch}$. Consequently, the Ich-**R3** variable which takes the value of 1 or 0 for the presence or absence of a six-membered saturated ring in **R3** was used instead of $^6\chi^{v}_{ch}$. The next best regression equation found including all descriptors as yet considered, was as follows:

$$\log 1/C = 0.48(0.12)\sum\pi(\mathbf{R1}+\mathbf{R2}) + 162.5(17.6)^1\chi^{\mathbf{N}}(\mathbf{R2}) - 239.8(25.1)(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2 - 0.85(0.19)\text{B1}-\mathbf{3R1} + 1.52(0.28)\text{Es-}\mathbf{2R1} + 52.06(7.54)^4\chi_p^{\mathbf{N}} + 0.78(0.34)\text{Vx}/100 - 2.22(0.28)\text{Ich-}\mathbf{R3} - 0.56(0.21)^0\Delta\text{X}(\mathbf{R3}) - 35.99(2.41)$$

$$r = 0.949 \quad Q = 0.863 \quad s = 0.439 \quad n = 79 \quad F = 69.45 \tag{4}$$

There is some general improvement in the statistics of eq 4,

as compared to eqs 1−3. The standard deviation as well as the predictive and fitting capabilities expressed by Q and r, respectively, are clearly better, and the agreement between the observed and calculated values is satisfactory (see Table 1).

Finally, the best regression equation found for all compounds described in Table 1 was as follows:

$$\log 1/C = 0.65(0.09)\sum\pi(\mathbf{R1}+\mathbf{R2}) + 130(17.05)^1\chi^{\mathbf{N}}(\mathbf{R2}) - 197.65(23.2)(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2 - 1.14(0.19)\text{B1}-\mathbf{3R1} + 1.60(0.28)\text{Es-}\mathbf{2R1} + 63.31(7.23)^4\chi_p^{\mathbf{N}} + 12.04(2.3)\text{Vx}/100 - 2.13(0.43)(\text{Vx}/100)^2 - 2.27(0.30)\text{Ich-}\mathbf{R3} - 1.25(0.20)\text{I-}\mathbf{4R1} - 0.72(0.21)^0\Delta\text{X}(\mathbf{R3}) - 47.7(2.96)$$

$$r = 0.951 \quad Q = 0.82 \quad s = 0.49 \quad n = 107 \quad F = 82.08 \tag{5}$$

The presence of the quadratic term in Vx/100, as much as of the I-4R1 variable in this equation, is related to the expansion of the original series of compounds. I-**4R1** is an indicator variable that takes the value of 1 or 0 for the presence or absence of a substituent at the 4-position of the C-6 aromatic ring. The calculated values given in Table 1 show the quality of the derived model. The optimum values for Vx and $^1\chi^{\mathbf{N}}(\mathbf{R2})$ are 283.05 cm^3/mol and 0.329, respectively, which are in the range of the values for the more potent compounds. However, it should be noted that eq 5, although highly significant statistically, is not as good a predictor as eq 4 (see Q values). This fact may be due to inaccuracies in the biological testing of compounds incorporated in the analysis. Nonetheless, as will be shown afterwards, the PLS analysis carried out with the same variables used in eqs 4 and 5 yields two models with a high predictive capability but only three PLS components are necessary to describe the observed activity. This is evidence of the great influence that the intercorrelation has among the variables when MLR is applied to construct a QSAR model. The correlation matrix for the descriptors used in this study is given in Table 3. Inspection of these results shows that the source of the collinearity (the highest intercorrelation allowed was 0.70) lies among the descriptors that make the principal contribution to the activity.

3.2. PLS Analysis. All variables used in the PLS calculations were initially autoscaled to zero mean and unit variance. The statistical significance of the obtained models was judged by the parameters already mentioned (r, Q, s, and F). The Q cross-validated coefficient was calculated using different sizes of cross-validation groups.

Preliminary analysis of the different PLS models performed on both the original and whole series of compounds showed that the most important variables were those which had been useful in the regression analysis and led to the selection of two PLS models.

The first model (PLS-1) was performed on the 79 compounds originally considered, and the descriptor matrix consisted of the same variables as used in eq 4 (included the quadratic term of $^1\chi^{\mathbf{N}}(\mathbf{R2})$). The PLS analysis resulted in a significant three-component model with the following statistics: $\boldsymbol{r = 0.943}$, $\boldsymbol{Q = 0.927}$, $\boldsymbol{s = 0.44}$, and $\boldsymbol{F = 202.27}$. The model accounted for 89% (64.6%, 22.2%, and 2.2%, respectively) of the variance in log 1/C.

Table 4. PLS Weight Vectors of the First, Second, and Third Dimension[a]

	PLS-1 ($n = 79$)			PLS-2 ($n = 107$)		
descriptor	WC[1]	WC[2]	WC[3]	WC[1]	WC[2]	WC[3]
$\sum\pi(\mathbf{R1}+\mathbf{R2})$	**0.37**	−0.23	0.24	**0.44**	−0.08	0.09
$^1\chi^{\mathbf{N}}(\mathbf{R2})$	**0.42**	−0.28	−0.03	**0.37**	−0.29	0.06
$(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2$	−0.29	**−0.70**	0.33	−0.17	**−0.77**	0.06
$^4\chi_p^{\mathbf{N}}$	**0.64**	0.09	0.01	**0.62**	0.12	0.07
Vx	**0.33**	−0.34	0.13	**0.34**	−0.23	0.14
B1-**3R1**	−0.04	−0.08	**−0.44**	0.07	−0.04	**−0.76**
Es-**2R1**	0.26	0.08	−0.20	0.17	0.07	0.01
$^0\Delta$X(**R3**)	−0.11	−0.09	0.28	−0.11	−0.10	0.17
Ich-**R3**	−0.05	**−0.47**	**−0.71**	0.02	**−0.30**	**−0.51**
I-**4R1**				**0.31**	0.03	−0.23
(Vx)^2				0.01	**−0.36**	0.19

[a] Descriptors weights having a large contribution on PLS components have been highlighted for clarity.

The second model (PLS-2), which was performed on all compounds considered in this study ($n = 107$), was based on the same variables as used in eq 5 including the squares of the $^1\chi^{\mathbf{N}}(\mathbf{R2})$ and Vx variables. The PLS analysis resulted in a significant three-dimensional model with the following statistics: $\boldsymbol{r = 0.944}$, $\boldsymbol{Q = 0.931}$, $\boldsymbol{s = 0.503}$, and $\boldsymbol{F = 283.57}$. This model accounted for 89.2% (62.4%, 21.5%, and 5.3%, respectively) of the variance in log 1/C. The Q values above cited were calculated using three cross-validation groups and represent the highest obtained values. The quality of both models may be demonstrated by the direct comparison between the observed and calculated activities given in Table 1.

In order to interpret the obtained models, the weights of the descriptor variables on the three PLS components of both models were analyzed. From these values it can be seen how much a single variable contributes in each PLS component to the modeling of the observed activity. The PLS weights for the established three-dimensional PLS models are listed in Table 4. As seen from this Table, both models are similarly affected by the same type of descriptors. The first PLS components of both models are mainly related to the size, shape, and hydrophobicity of the compounds, since Vx, $^4\chi_p^{\mathbf{N}}$, $^1\chi^{\mathbf{N}}(\mathbf{R2})$, and $\sum\pi(\mathbf{R1}+\mathbf{R2})$ have the largest contribution. The second PLS components are strongly dominated by the squared term in $^1\chi^{\mathbf{N}}(\mathbf{R2})$ indicating that there is a significant nonlinear relationship between log 1/C and $^1\chi^{\mathbf{N}}(\mathbf{R2})$. These components are also influenced by the Ich-**R3** variable and by the quadratic term in Vx (this term appears only in the PLS-2 model). Finally, the third PLS components describe a minor part of the variance in the data set and are therefore rather difficult to interpret. However, due to large contribution of B1-**3R1**, these components may be related to the molecular shape.

3.3. Comparison of the Quality of MLR and PLS Models. Based on the results obtained, two important general observations can be made. Firstly, the PLS models developed show the true dimensionality ($A = 3$) of dataspace considered. Secondly, a comparison of the quality of the MLR and PLS models shows that the PLS models have substantially better predictive capabilities than the MLR models (higher Q values). However, as expressed by the correlation coefficients, the data-fitting abilities for both type of models are similar.

Table 5. Comparison of Regression Coefficients[a] for the MLR and PLS Models

	MLR models				PLS models		
descriptor	eq 2 ($n = 79$)	eq 3 ($n = 79$)	eq 4 ($n = 79$)	eq 5 ($n = 107$)	PLS-1 ($n = 79$(	PLS-2 ($n = 107$)	PLS-3 ($n = 103$)
$\sum\pi$(**R1**+**R2**)	2.97		2.27	3.61	0.25	0.44	0.38
$^1\chi^{\mathbf{N}}$(**R2**)	38.93	18.52	30.19	24.15	0.21	0.18	0.17
$(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2$	−43.67	−20.79	−33.99	−28.02	−0.61	−0.69	−0.68
$^4\chi_p^{\mathbf{N}}$		3.80	2.68	3.50	0.67	0.78	0.81
Vx		2.20	1.22	**22.30**	0.13	0.23	0.29
B1-**3R1**	−2.44	−1.37	−1.37	−1.82	−0.17	−0.25	−0.25
Es-**2R1**	2.37	2.16	1.89	1.98	0.26	0.24	0.24
$^0\Delta$X(**R3**)	−2.40		−1.11	−1.43	−0.11	−0.13	−0.18
Ich-**R3**			−2.22	−2.27	−0.48	−0.41	
I-**4R1**				−1.25		−0.27	−0.26
$(Vx)^2$				−20.40		−0.12	−0.13
$^4\chi_p$	5.57						
$^6\chi_{ch}^{v}$	−2.97						

[a] All coefficients were calculated by scaling the variables to zero mean and unit variance.

Another way of comparing the quality of the MLR and PLS models, is to examine the stability of their coefficients when different subsets of data are used to derive these models. Table 5 shows the PLS models expressed as pseudoregression solutions together with the regression coefficients of eqs 2−5. Inspection of the data in this table reveals that the coefficients of the PLS models are reasonably stable. Note that the coefficients of the PLS-1 model are only slightly different from those of the PLS-2 or PLS-3 models despite the inclusion of 28 or 24 more data points, respectively. On the other hand, comparison of the MLR models also shows some agreement among the coefficients except for the volume term which shows a considerable change between eqs 4 and 5. These results support the conclusion that the intercorrelation among the variables affects not only the predictive capability of the MLR models but also makes more difficult a clear interpretation of the data. For example, eq 4 indicates that $^1\chi^{\mathbf{N}}$(**R2**)/$(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2$ and $^4\chi_p^{\mathbf{N}}$ are the most important contributors to the activity, but due to the instability of the Vx coefficient, eq 5 does not lead to the same conclusion. These results are in agreement with the findings by other authors[35,36] who state that PLS is a better approach to MLR for improving interpretability of the data and also to exhibit models with a better predictive quality.

3.4. Structural Interpretation of the Models. Analysis of the PLS coefficients given in Table 5 shows that $^4\chi_p^{\mathbf{N}}$ and $^1\chi^{\mathbf{N}}$(**R2**) are the most important variables to explain the observed activity. Thus, it is important to offer some insight into the significance of these indexes in the obtained models.

The $^1\chi^{\mathbf{N}}$(**R2**) index encodes information about the degree of branching or connectivity of the **R2** substituents, and their high contribution in all models shows that the shape of **R2** substituents has a strong effect on the activity. This is in agreement with the findings by other authors[17,37] who have shown that the differences of activity among the HEPT derivatives are strongly dependent on the positioning of the 6-phenylthio group, which is quite sensitive to the steric effects of the C-5 substituents. Thus, the $^1\chi^{\mathbf{N}}$(**R2**) index quantifies the effect of these substituents and shows that according to the $^1\chi^{\mathbf{N}}$(**R2**)/$(^1\chi^{\mathbf{N}}(\mathbf{R2}))^2$ relationship a maximum activity can be encountered when an optimum $^1\chi^{\mathbf{N}}$(**R2**) value is reached. Hence, the substitution of a methyl group at the C-5 position of HEPT by either ethyl or isopropyl (but not *n*-propyl or cyclopropyl) brought about a marked increase of the activity, which is consistent with the $^1\chi^{\mathbf{N}}$(**R2**) values for the more potent compounds. The log 1/*C* and $^1\chi^{\mathbf{N}}$(**R2**) values of the compounds **22**, **23**, **24**, and **36**, or **28**, **29**, **30**, and **35** or **45**, **51**, **59**, and **80** provide a good example of this.

A separate discussion requires the $^4\chi_p^{\mathbf{N}}$ index. In the most general sense, this index may be considered as conveying information about the shape characteristics of whole molecule. Analyzing the $^4\chi_p^{\mathbf{N}}$ coefficients given in Table 5, it is seen that an increase of the $^4\chi_p^{\mathbf{N}}$ value will generally result in a greater anti-HIV-1 activity. From a structural point of view, this fact can be achieved in different ways. Thus, increasing the number of **R1** substituents will result in an increase of $^4\chi_p^{\mathbf{N}}$ value and, therefore, in a greater activity. For example, compare the log 1/*C* and $^4\chi_p^{\mathbf{N}}$ values of compounds **4** and **9** with those of **15** and **16** or, likewise, compare **22**−**24** with **25**−**27**. On the other hand, this index encodes information about the degree of branching of the **R2** substituents. Thus, an increase of branching of **R2** results in a greater value of $^4\chi_p^{\mathbf{N}}$ index and, consequently, in a greater activity. The log 1/*C* and $^4\chi_p^{\mathbf{N}}$ values of **22**, **23**, **24**, and **36** or **28**, **29**, **30**, and **35** provide a good example of this. Further, the combined effects of branching and disubstitution results in even greater values of $^4\chi_p^{\mathbf{N}}$ index and consequently of the activity (see values of **25**−**27** or **31**−**33**). Finally, the $^4\chi_p^{\mathbf{N}}$ index also encodes information about the presence of cyclic fragments in the molecule. For example, the presence of a phenyl ring in **R3** results in a strong increase of $^4\chi_p^{\mathbf{N}}$ value, which is generally consistent with the corresponding enhancement of the activity. The log 1/*C* and $^4\chi_p^{\mathbf{N}}$ values of compounds **42**−**50** illustrate this situation. However, in the case of the compounds **69**, **70**, **76**, and **77**, the replacement of the phenyl ring by a cyclohexyl group in **R3** yields compounds less active than one would expect from its $^4\chi_p^{\mathbf{N}}$ values. Thus, this structural feature is encoded by the Ich-**R3** indicator variable. It should be noted that the quality of the obtained models is not affected when these compound data and the corresponding Ich-**R3** variable are excluded of the PLS models. Comparison of coefficients between the PLS-2 and PLS-3 models illustrate this fact, as shown in Table 5.

With respect to the other descriptors incorporated in the models, it is clear from the coefficients given in Table 5 that hydrophobicity contributes strongly to the activity while

that other parameters such as Vx and B1-**3R1** make lesser but significant contributions to the activity. Finally, the presence of $^0\Delta X$(**R3**) index in all models is related to the polarity of **R3** because this index encodes information about the presence of π and lone-pair electrons in a molecule. Thus, when the value of $^0\Delta X$(**R3**) is higher, the **R3** substituent is richer in non-Csp^3 atoms. However, the coefficient for $^0\Delta X$-(**R3**) is small, so this parameter makes only small contributions to the predicted activities.

4. CONCLUSION

Two important consequences emerge from the present report.

Firstly, taking into account the complex nature of modeled biological phenomena, on the one hand, and the large number of analyzed compounds, on the other hand, our results clearly indicate that the hydrophobic and especially geometric factors are of prime importance for the anti-HIV-1 activity of the HEPT derivatives under study. These results are in complete agreement with, and extend the implications of, those reported by Hansch and Zhang.[15] In addition, it is important to highlight that the PLS models presented in this study have substantially better predictive capabilities than the QSAR models recently published for a smaller group of HEPT analogs.[18,19]

Secondly, this study provides evidence for the great potential of the topological approach for the development of QSAR models.

Further, addressing to the problem of steric relationships to biological activity, the present report combines the statistical advantages of the principal component-like methods, such as PLS regression, with the power of the topological indexes to account for the geometric aspects of the drug-receptor interaction.

ACKNOWLEDGMENT

This work was supported by a grant of University National of San Luis (Project No 219300). We wish to thank Prof. John A. Luco for linguistic advice.

REFERENCES AND NOTES

(1) Mitsuya, H.; Yarchoan, R.; Broder, S. Molecular Targets for AIDS Therapy. *Science* **1990**, *249*, 1533−1544.

(2) Baba, M.; Debyser, Z.; Shigeta, S.; De Clercq, E. Highly Potent and Selective Inhibition of Human Immunodeficiency Virus Type 1 (HIV-1) by the HIV-1-Specific Reverse Transcriptase Inhibitors. *Drugs of the Future* **1992**, *17*, 891−897.

(3) De Clercq, E. Trends in Drug Development for the Treatment of AIDS Compounds Interfering with the Initial Stages of the HIV Replicative Cycle. *Eur. J. Pharmaceutical Sci.* **1994**, *2*, 4−6.

(4) Sandberg, J. A.; Slikker, W., Jr. Development Pharmacology and Toxicology of Anti-HIV Therapeutic Agents: Dideoxynucleosides. *FASEB J.* **1995**, *9*, 1157−1163.

(5) De Clercq, E. Toward Improved Anti-HIV Chemotherapy: Therapeutic Strategies for Intervention with HIV Infections. *J. Med. Chem.* **1995**, *38*, 2491−2517.

(6) Baba, M.; Tanaka, H.; De Clercq, E.; Pauwels, R.; Balzarini, J.; Schols, D.; Nakashima, H.; Perno, C. F.; Walker, R. T.; Miyasaka, T. Highly Specific Inhibition of Human Immunodeficiency Virus Type 1 by a Novel 6-Substituted Acyclouridine Derivative. *Biochem. Biophys. Res. Commun.* **1989**, *165*, 1375−1381.

(7) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1 Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1989**, *32*, 2507−2509.

(8) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Miyasaka, T.; Nitta, I.; Walker, R. T.; De Clercq, E. The Synthesis of Novel 6-Substituted Acyclouridine Derivatives which Show Specific Inhibition of HIV-1. *Collect. Czech. Chem. Commun.* **1990**, *55*, 89−92.

(9) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV-1-Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity of 5- or 6-Substituted Analogues of 1-[(2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.

(10) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Anti-HIV Activity of 2-, 3-, and 4-Substituted Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 1394−1399.

(11) Tanaka, H.; Baba, M.; Saito, S.; Miyasaka, T.; Takashima, H.; Sekiya, K.; Ubasawa, M.; Nitta, I.; Walker, R. T.; Nakashima, H.; De Clercq, E. Specific Anti-HIV-1 "Acyclonucleosides" Which Cannot Be Phosphorylated: Synthesis of Some Deoxy Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1991**, *34*, 1508−1511.

(12) Tanaka, H.; Baba, M.; Hayakawa H.; Haraguchi, K.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Walker, R. T.; De Clercq, E. Lithiation of Uracilnucleosides and its Application to the Synthesis of a New Class of Anti-HIV-1 Acyclonucleosides. *Nucleosides & Nucleotides* **1991**, *10*, 397−400.

(13) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure−Activity Relationships of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)-thymine Analogues: Effect of Substitutions at the C-6 Phenyl Ring and the C-5 Position on Anti-HIV-1 Activity. *J. Med. Chem.* **1992**, *35*, 337−345.

(14) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of Deoxy Analogs of 1-[(2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1992**, *35*, 4713−4719.

(15) Hansch, C.; Zhang, L. QSAR of HIV Inhibitors. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 1165−1169.

(16) Kelly, T. A.; Proudfoot, J. R.; McNeil, D. W.; Patel, U. R.; David, E.; Hargrave, K. D.; Grob, P. M.; Cardozo, M.; Agarwal, A.; Adams, J. Novel Non-nucleoside Inhibitors of Human Immunodeficiency Virus Type 1 Reverse Transcriptase. 5. 4-Substituted and 2,4-Disubstituted Analogs of Nevirapine. *J. Med. Chem.* **1995**, *38*, 4839−4847.

(17) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; Ross, C.; Miyasaka, T.; Walker, R. T.; Tanaka, H.; Stammers, D. K.; Stuart, D. I. Complexes of HIV-1 Reverse Transcriptase with Inhibitors of the HEPT Series Reveal Conformational Changes Relevant to the Design of Potent Non-Nucleoside Inhibitors. *J. Med. Chem.* **1996**, *39*, 1589−1600.

(18) Hannongbua, S.; Lawtrakul, L.; Limtrakul, J. Structure−Activity Correlation Study of HIV-1 Inhibitors. Electron and Molecular Parameters. *J. Comput.-Aided Mol. Design* **1996**, *10*, 145−152.

(19) Hannongbua, S.; Lawtrakul, L.; Sotriffer, C. A.; Rode, B. M. Comparative Molecular Field Analysis of HIV-1 Reverse Transcriptase Inhibitors in the Class of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)-thymine. *Quant. Struct.−Act. Relat.* **1996**, *15*, 389−394.

(20) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Chemistry and Drug Research*; de Stevens, G., Ed.; Academic Press: New York, 1976.

(21) Luco, J. M.; Sosa, M. E.; Cesco, J. C.; Tonn, C. E.; Giordano, O. S. Molecular Connectivity and Hydrophobicity in the Study of Anti-feedant Activity of Clerodane Diterpenoids. *Pestic. Sci.* **1994**, *41*, 1−6.

(22) Luco, J. M.; Yamin, L. J.; Ferretti, H. F. Molecular Topology and Quantum Chemical Descriptors in the Study of Reversed-Phase Liquid Chromatography. Hydrogen-Bonding Behavior of Chalcones and Flavanones. *J. Pharm. Sci.* **1995**, *84*, 903−908.

(23) Santagata, L. N.; Luco, J. M.; Enriz, R. D. Aproximación Analítica de la Superficie Molecular de Van Der Waals Aplicada en un Estudio de QSAR. *An. Asoc. Qum. Argentina* in press.

(24) Wold, S. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; Vol. 2, Chapter 4, p 198.

(25) Hansch, C.; Leo, A.; Hoekman, D. In *Hydrophobic, Electronic and Steric Constants, Vol. 2 of Exploring QSAR*; Heller, S. R., Ed.; American Chemical Society: Washington DC, 1995; pp 217−304.

(26) van de Waterbeemd, H.; Kansy, M. Hydrogen-Bonding Capacity and Brain Penetration. *Chimia* **1992**, *46*, 299−303.

(27) Kaliszan, R. In *Quantitative Structure-Chromatographic Retention Relationships*; Winefordner, J. D., Ed.; John Wiley & Sons: New York, 1987; Chapter 8, p 138.

(28) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Chemistry and Drug Research*; Bawden, D., Ed.; Research Studies Press LTD: Letchworth, Hertfordshire, England, 1986; Chapter 3, p 43.

(29) Kier, L. B.; Hall, L. H. A Differential Molecular Connectivity Index. *Quant. Struct.−Act. Relat.* **1991**, *10*, 134−140.

(30) Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. VCH: Weinheim, 1993; Chapter 3, p 21.

(31) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid Chromatography. *Chromatographia* **1987**, *23*, 243−246.

(32) Abraham, M. H. Scales of Solute Hydrogen-Bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Soc. Rev.* **1993**, *22*, 73−83.

(33) Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397−405.

(34) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(35) Clark, M.; Cramer III, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.−Act. Relat.* **1993**, *12*, 137−145.

(36) Verhaar, Henk, J. M.; Eriksson, L.; Sjostrom, M.; Schuurmann, G.; Willem, S.; Hermens, Joop L. M. Modelling the Toxicity of Organophosphates: a Comparison of the Multiple Linear Regression and PLS Regression Methods. *Quant. Struct.−Act. Relat.* **1994**, *13*, 133−143.

(37) Tanaka, H.; Baba, M.; Takahashi, E.; Matsumoto, K.; Kittaka, A.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Design and Synthesis of Regioisomeric Analogues of a Specific Anti-HIV-1 Agent 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *Nucleosides & Nucleotides* **1994**, *13*, 155−162.

CI960487O