

Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure–Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method

V. P. Solov'ev[†] and A. Varnek^{*,‡}

Institute of Physiologically Active Compounds, Russian Academy of Sciences,
142432, Chernogolovka, Moscow Region, Russia, and Laboratoire d'Infochimie, UMR 7551 CNRS,
Université Louis Pasteur, 4, rue B. Pascal, Strasbourg, 67000, France

Received December 10, 2002

Substructural molecular fragments (SMF) method [Solov'ev, V. P.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858] was applied to assess anti-HIV activity for large data sets for three families of compounds: 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives, tetrahydroimida-zobenzodiazepinone (TIBO) derivatives, and cyclic urea (CU) derivatives. The SMF method uses 49 types of topological descriptors (atom/bond sequences and “augmented atoms”) which, being coupled with 3 linear and nonlinear fitting equations, allows the user to generate up to 147 structure–property models. For each family of compounds, the modeling was performed on several training sets followed by the validation calculations where three best fit models were applied. Calculated activities well reproduce available experimental data. On the basis of the “optimal” molecular fragments, the focused combinatorial library containing 252 virtual HEPT derivatives has been generated. Its filtering led to several hits potentially possessing anti-HIV activity.

1. INTRODUCTION

Since 1991, compounds possessing anti-HIV activity have been the subject of numerous QSAR studies. A search in the Chemical Abstracts database¹ for the “QSAR and HIV” query led to about 200 references describing structure–property relationships established using partial least-squares (PLS), artificial neural network (ANN), and multiple linear regression (MLR) methods involving 1D and/or 2D descriptors^{2–7} or 3D descriptors,^{8,9} the 4D-QSAR technique,¹⁰ comparative molecular field analysis (CoMFA),^{11–15} and electrostatic potential distribution.^{16,17} These studies concerned four main classes of anti-HIV compounds: the reverse transcriptase inhibitors, the protease inhibitors, the virus uncoating inhibitors, and the integrase inhibitors.¹⁸ Several types of activity were modeled: (i) the HIV-1 protease inhibition constant K_i ,^{2,19,20} (ii) the concentration of the compound required to reduce by 50% the number of mock-infected MT-4 or CEM cells (CC_{50}),² and (iii) the concentration of the compound leading to protection of 50% of the number of MT-4 or CEM cells against cytopathic effect of the virus (EC_{50}),^{3,12,21–23} or to 50% of the HIV-1 reverse transcriptase enzyme inhibition (IC_{50}).^{6,24–26} Usually, the logarithm of the inverse of these parameters ($\log(1/K_i)$, $\log(1/IC_{50})$, or $\log(1/EC_{50})$) is used in QSAR studies.

The key problem of any SAR study is related to selection of pertinent descriptors to model *different types of activities for different classes of compounds*. To our knowledge, only Garg et al² have performed MLR calculations using the same set of “classical” physicochemical descriptors ($\log P$, molec-

ular volume, molecular refraction, Taft and Hammett parameters) to model different types of anti-HIV activities of various types of compounds. In other QSAR studies, a particular set of descriptors was used for a given class of compounds.

This article is devoted to structure–property modeling of anti-HIV activities for two families of HIV-1 reverse transcriptase inhibitors (1-[2-hydroxyethoxy)methyl]-6-(phenylthio)-thymine (HEPT) derivatives and tetrahydroimida-zobenzodiazepinone (TIBO) derivatives) and one family of HIV-1 protease inhibitors (cyclic urea (CU) derivatives) using the recently developed substructural molecular fragment (SMF) method.²⁷ This method is based on the splitting of a molecular graph into fragments, and on calculations of their contributions to a given property. It uses two types of topological descriptors (fragments): atom/bond sequences, and “augmented atoms” (atoms with their nearest neighbors). The SMF approach was tested on physical properties of C2–C9 alkanes (boiling point, molar volume, molar refraction, heat of vaporization, surface tension, melting point, critical temperature, and critical pressures) and on octanol/water partition coefficients.²⁷ Then, it has been successfully applied to the assessment of complexation stability constants of the complexes of macrocyclic and open-chain ligands with cations and neutral guests in solution, and to model thermodynamic parameters (equilibrium constants and distribution coefficients of metals) of solvent extraction of metals by phosphoryl-containing ligands and amides.^{27–29}

Here, we apply the SMF method to model three types of anti-HIV activity: $\log(1/IC_{50})$ for TIBO, $\log(1/EC_{50})$ for HEPT, and $\log(1/K_i)$ for CU using available experimental data. It will be shown that statistical criteria of linear correlation between experimental data and those calculated

* Corresponding author phone: +33-3-90241549; fax: +33-3-90241589; e-mail: varnek@chimie.u-strasbg.fr.

[†] Russian Academy of Sciences.

[‡] Université Louis Pasteur.

by SMF activities both for training and for test sets are at least as good as those obtained in previous QSAR studies for TIBO,^{6,26} HEPT,^{3,12,23} and CU¹⁴ derivatives.

Below, we compare the performance of SMF calculations on HEPT, TIBO, and CU compounds to that of CoMFA analysis reported in the literature.^{14,26} Although COMFA, which is one the most popular 3D QSAR methods, represents a powerful tool for structure–properties studies, it does not always lead to reliable predictions because of several internal problems:³⁰ (i) identification of the active conformations/molecular shapes of flexible compounds in the training set, (ii) specification of the basis for molecular alignment, and (iii) partitioning of molecules in the training set with respect to intermolecular (receptor) interactions. Apparently, the 4D-QSAR method³⁰ is quite useful to overcome these difficulties, as it has been recently shown by Santos-Filhoa and Hopfinger¹⁰ in structure–property studies of new nonpeptidic HIV protease inhibitors.

2D QSAR methods, which use descriptors based on molecular graphs, provide an appealing alternative to 3D QSAR because they do not require extensive conformational analysis and spatial alignment of molecules. They are faster and easier to implement in an automated fashion and are typically characterized by the same or better statistics compared to 3D QSAR methods.^{31,32} Results reported by Brown and Martin^{33,34} show that 2D fingerprints could be more efficient in clustering and QSAR studies than 3D fingerprints.

The SMF method can be considered as an extension of the Free–Wilson approach³⁵ applying molecular fragments as variables in a multiple regression analysis. Computation of fragmental descriptors does not require the knowledge of the geometry and electronic structure of molecules; structural fragments are more easily interpretable than topological indices. Molecular fragments are successfully used in diversity analysis of large databases^{36,37} and in structure–property studies.^{38–43} The recently developed PASS method,^{39,44} used for estimation of a wide spectrum of biological activities, is also based on molecular fragments (augmented atoms). The success of the fragmental approach in QSAR/QSPR studies depends on the diversity of structural fragments as well as on the flexibility of atomic classification. In this sense, the SMF method represents a flexible structure–property tool because it generates 49 different types of fragments (atom/bond sequences and augmented atoms), then builds QSAR models involving linear and nonlinear fitting equations (see Section 2). As atom/bond sequence, for a given pair of atoms, TRAIL uses the shortest topological path. Earlier, for similarity studies, Carhart⁴⁵ suggested using sequences (shortest paths) represented by the first and the last atoms and the number of atoms in a connecting “spacer”. In contrast to such implicitly defined fragments, the SMF method accounts explicitly for all atoms and bonds in a sequence.

A significant advantage of the SMF method is a possibility to select during the training stage several best fit models (instead of a single QSPR model) related to different fragmentation schemes in combination with three fitting equations. Using selected QSAR models, one can calculate average activities of the compounds from the test set, which smoothes inaccuracies of particular models, thus improving a robustness of predictions.²⁹

An important aspect of any structure–property study is the theoretical design of new active compounds. The recent

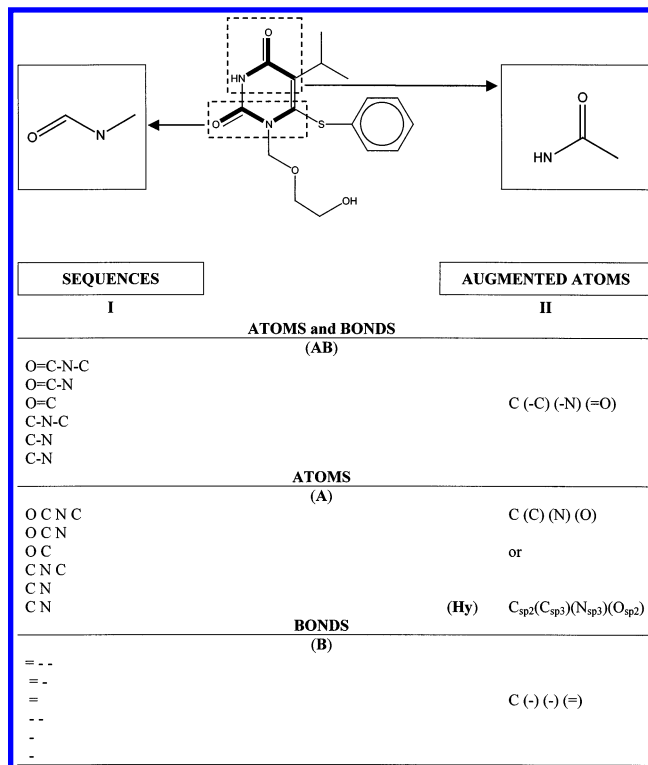


Figure 1. Substructural molecular fragments: atom/bond sequences and augmented atoms. Shortest paths sequences (**I**) and augmented atoms (**II**) including atoms and bonds (**AB**), only atoms (**A**), or only bonds (**B**). From top to bottom: the sequences (**I**) correspond to the **I(AB, 2–4)**, **I(A, 2–4)**, and **I(B, 2–4)** types involving paths between each pair of atoms. The **II(Hy)** augmented atoms correspond to the **II(A)** type, where hybridization of atom is taken into account.

paper of Pasqual et al.⁴⁶ showed that a combinatorial space of HEPT compounds is insufficiently investigated and new active compounds could be discovered. Attempts to interpret the anti-HIV activity of HEPT, TIBO, and CU derivatives as a function of substituents were made in previous QSAR studies,^{11,12,22} but no new compounds were proposed. Here we show how the SMF method can be used to generate and screen small focused virtual combinatorial libraries using the fragment contributions obtained in the training stage (see Sections 4.2 and 5.3). Several hypothetic HEPT derivatives were suggested as efficient anti-HIV active compounds.

2. METHOD

The substructural molecular fragments (SMF) method is based on the splitting of a molecular graph into fragments, and on the calculation of their contributions to a given property **Y**. Two different types of fragments are used (Figure 1): “sequences” (**I**) and “augmented atoms” (**II**). Three subtypes, **AB**, **A**, and **B**, are defined for each type of fragment (Figure 1). For the fragments **I**, they represent sequences of atoms and bonds (**AB**), of atoms only (**A**), or of bonds only (**B**). The number of atoms in these sequences varies from 2 to 6. Only shortest paths from one atom to the other are used, as shown in Figure 1. An “augmented atom” represents a selected atom with its environment including either neighboring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). Atomic hybridization (**Hy**) can be taken into account for augmented atoms of the **A** type.

Once a given molecular structure is split into constitutive fragments, its quantitative physical or chemical property Y is calculated from the fragments contributions using linear (1) or nonlinear (2) and (3) fitting equations.

$$Y = a_o + \sum_i a_i N_i \quad (1)$$

$$Y = a_o + \sum_i a_i N_i + \sum_i b_i (2N_i^2 - 1) \quad (2)$$

$$Y = a_o + \sum_i a_i N_i + \sum_{i,k} b_{ik} N_i N_k \quad (3)$$

where a_i and b_i (b_{ik}) are fragment contributions, N_i is the number of fragments of i type. The a_o term is fragment independent; it is fitted by default, but optionally it can be omitted. Linear eq 1 represents a molecular property as a sum of fragment contributions. Equation 2, representing the three first terms of Chebyshev polynomial,⁴⁷ accounts for nonadditive effects related to individual fragments, whereas eq 3 involves a cross term $N_i N_k$ which accounts for the nonadditivity effects of two different fragments.

The TRAIL program has been developed to calculate structure–property relationships based on the SMF partitioning. At the first step of the calculations, TRAIL generates up to 147 structure–property models involving 49 types of fragments coupled with 3 fitting equations, and uses all of them at the training and test stages. Molecules containing fragments of “rare” occurrence (i.e., found in less than 2 molecules) are excluded from the training set. If some fragments are linearly dependent, they are treated as one extended fragment. To fit the a_i and b_i terms in eqs 1–3, TRAIL uses the singular value decomposition method,⁴⁸ which allows one to exclude descriptors with negligible contributions to a given property and to calculate a matrix of pair correlations for the terms a_i and b_i (covariation matrix). Then the program calculates statistical characteristics of models (correlation coefficient (R), standard deviation (s), Fischer’s criterion (F), Kubyni’s criterion (FIT), cross-validation correlation coefficient (Q), and R_H -factor of Hamilton), and performs statistical tests⁴⁹ to select the best models. At the test stage, TRAIL calculates the “predicted” values using the fitted fragments contributions obtained at the training stage.

The *CombiLib* program has been prepared to generate virtual combinatorial libraries using Markush structures.⁵⁰ It uses its own editor of 2D structures allowing the user to prepare the molecular core, to define the type of the attachment “reaction”, to select the attachment entries (atoms, bonds) and to prepare collections of substituents. Attachment of substituents to the molecular core can be performed by either connecting two atoms belonging to the two fragments (“atom–atom attachment”), by overlapping two bonds of these fragments (“bond–bond attachment”), or by inclusion of an atom of the core into a cyclic substituent (“spiro attachment”). Activities of generated compounds are estimated using QSAR models selected on the training stage.

3. DATA SETS PREPARATION

To check the robustness of the SMF method, for all studied families, three different training/test sets were selected from

the initial parent set. To prepare the test sets, we followed the recommendations of Oprea et al.:⁵¹ (i) experimental methods for determination of activities in the training and test sets should be similar; (ii) the activity values should span several orders of magnitude, but should not exceed activity values in the training set by more than 10%; and (iii) the balance between active and inactive compounds should be respected for uniform sampling of the data. Each test set contained about 10% of the compounds from the corresponding parent set: 7 (TIBO), 8 (HEPT), and 9 (CU). For a given family of compounds, the molecules in a particular test set differ completely from other sets (Tables 3, 6, and 9). Other data from the initial parent set were included in the corresponding training sets (Tables 1, 4, and 7).

3.1. TIBO Derivatives. The parent data set of 84 TIBO compounds has been composed from the experimental IC_{50} values given in publications of Kukla et al.,⁵² Ho et al.,⁵³ and Breslin et al.⁵⁴ Because the current version of the SMF method does not recognize the difference between optical isomers or between E and Z isomers, for several compounds the data for only one isomer were used in our studies. 77 inactive compounds, or those for which only the lower limit of activity (IC_{50}) rather than its accurate value were given,^{52–54} were not included in this parent set. Modeling was performed for the $\log(1/IC_{50})$ values which vary from 3.06 to 8.52 for IC_{50} expressed in mol/L.

3.2. HEPT Derivatives. Experimental effective concentrations of compounds required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (EC_{50}) for 93 molecules have been critically selected from the literature.^{55–62} Compounds for which only the upper limit of EC_{50} values was reported were not included in the parent set. Modeling was performed for the $\log(1/EC_{50})$ values which vary from 3.85 to 9.22 for EC_{50} expressed in mol/L. Experimental error in EC_{50} varies from 1.1 to 45%,⁵⁷ which corresponds to inaccuracy in $\log(1/EC_{50})$ from 0.004 to 0.20.

3.3. CU Derivatives. Experimental HIV-1 protease inhibition constants K_i for 118 compounds were selected from the literature.^{63–65} Corresponding $\log(1/K_i)$ values vary in the range 5.11–10.96. As estimated by Wilkerson et al.,⁶³ the standard deviation in obtaining K_i is about 40%, i.e. about 0.2 in $\log(1/K_i)$ units.

4. RESULTS

4.1. Structure–Property Modeling. 4.1.1. TIBO Derivatives. The initial data set containing 84 compounds was split into training (77 compounds) and test (7 compounds) sets, as described in Section 3. At the training stage, TRAIL excluded 11 compounds containing unique fragments (occurs only in one compound), thus reducing the three training sets to 66 compounds (Table 1). Three best fit models were selected: **I(AB, 5–5)**, **I(AB, 3–4)**, and **I(AB, 2–4)** coupled with fitting eq 1 for which $R^2 = 0.85–0.92$, $s = 0.61–0.89$, $F = 3–7$, $Q^2 = 0.50–0.74$, and $FIT = 0.1–0.2$ (Table 2). It should be noted that poor fitting results were obtained for compounds 31 and 44 (residuals are about 1.3–1.7). Their exclusion from the training sets led to substantial improvement of the statistical criteria ($R^2 > 0.90$; $s < 0.6$, $F > 8$, $Q^2 > 0.58$, and $FIT > 0.3$).

Validation calculations were performed on three test sets containing compounds 2, 9, 19, 20, 27, 50, and 70 (*Set 1*),

Table 1. Initial Data Set of TIBO Derivatives: Experimental and Calculated Activities $\log(1/IC_{50})^a$

no.	ref. ^b	X	R ₁	R ₂	R ₃	exp.	Log(1/IC ₅₀)			mean ^c
							I(AB, 5–5)	I(AB, 3–4)	I(AB, 2–4)	
1	1a ⁵⁴	S	H	DMA ^e	8-Cl	7.34	7.79	8.26	8.25	8.10 (0.27)
2	1b ⁵⁴	S	H	DMA ^e	9-Cl	6.80	7.74 ^d	8.01 ^d	7.96 ^d	7.90 (0.14) ^d
3	10d ⁵³	O	5-Me (S)	DMA ^e	8-C(O)NH ₂	5.20	5.20	5.20	5.20	5.20 (0.00)
4	1r ⁵⁴	O	5,5-Me ₂	2-MA ^f	H	4.64	4.94	5.02	4.91	4.96 (0.06)
5	1u ⁵⁴	O	4-Me	2-MA ^f	H	4.49	4.47	4.52	4.60	4.53 (0.06)
6	1v ⁵⁴	S	4-Me	2-MA ^f	9-Cl	6.17	5.70	6.63	6.60	6.31 (0.53)
7	1w ⁵⁴	S	4-Me	CH ₂ -c-Pr	9-Cl	5.66	5.80	6.84	6.91	6.52 (0.62)
8	1y ⁵⁴	O	4- <i>i</i> -Pr	<i>n</i> -Pr	H	4.13	4.15	4.39	4.06	4.20 (0.17)
9	1z ⁵⁴	O	4- <i>i</i> -Pr	2-MA ^f	H	4.90	4.60 ^d	4.62 ^d	4.31 ^d	4.51 (0.17) ^d
10	13 ⁵³	O	5-Me (S)	CH ₂ -c-Pr	9-NO ₂	4.48	4.34	4.39	4.42	4.38 (0.04)
11	1ac ⁵⁴	O	4- <i>n</i> -Pr	2-MA ^f	H	4.32	4.32	4.18	4.27	4.26 (0.07)
12	14a ⁵³	O	5-Me (S)	CH ₂ -c-Pr	8-NH ₂	3.07	3.07	3.36	3.36	3.26 (0.17)
13	1ai ⁵⁴	O	7-Me	DMA ^e	H	4.92	5.27	5.48	5.54	5.43 (0.14)
14	1aj ⁵⁴	O	7-Me	DMA ^e	8-Cl	6.84	6.67	6.53	6.58	6.59 (0.07)
15	1ak ⁵⁴	O	7-Me	DMA ^e	9-Cl	6.80	6.32	6.28	6.29	6.29 (0.02)
16	1al ⁵⁴	S	7-Me	<i>n</i> -Pr	H	5.61	5.77	5.70	5.69	5.72 (0.05)
17	1am ⁵⁴	S	7-Me	DMA ^e	H	7.11	6.69	6.80	6.80	6.76 (0.07)
18	1an ⁵⁴	S	7-Me	DMA ^e	8-Cl	7.92	8.09	7.85	7.83	7.92 (0.14)
19	1ao ⁵⁴	S	7-Me	DMA ^e	9-Cl	7.64	7.74 ^d	7.59 ^d	7.54 ^d	7.62 (0.10) ^d
20	1as ⁵⁴	O	4,5-Me ₂ (cis)	DMA ^e	H	4.25	5.00 ^d	4.02 ^d	4.10 ^d	4.37 (0.54) ^d
21	1at ⁵⁴	S	4,5-Me ₂ (cis)	DMA ^e	H	5.65	5.18	5.33	5.36	5.29 (0.10)
22	1au ⁵⁴	S	4,5-Me ₂ (trans)	CH ₂ -c-Pr	H	4.87	4.81	4.67	4.81	4.76 (0.08)
23	1av ⁵⁴	S	4,5-Me ₂ (trans)	DMA ^e	H	4.84	5.18	5.33	5.36	5.29 (0.10)
24	14b ⁵³	O	5-Me (S)	CH ₂ -c-Pr	8-NMe ₂	5.18	5.18	4.89	4.89	4.99 (0.17)
25	15a ⁵³	O	5-Me (S)	CH ₂ -c-Pr	9-NH ₂	4.22	4.22	3.93	3.93	4.03 (0.17)
26	1az ⁵⁴	S	5,7-Me ₂ (trans)	DMA ^e	H	7.38	6.51	6.31	6.29	6.37 (0.12)
27	1ba ⁵⁴	S	5,7-Me ₂ (cis)	DMA ^e	H	5.94	6.51 ^d	6.31 ^d	6.29 ^d	6.37 (0.12) ^d
28	1bb ⁵⁴	O	5,7-Me ₂ (<i>R,R</i> ; trans)	DMA ^e	9-Cl	6.64	6.14	5.78	5.78	5.90 (0.21)
29	1bc ⁵⁴	S	5,7-Me ₂ (<i>R,R</i> ; trans)	DMA ^e	9-Cl	6.32	7.56	7.10	7.04	7.23 (0.29)
30	15b ⁵³	O	5-Me (S)	CH ₂ -c-Pr	9-NMe ₂	5.18	5.18	5.47	5.47	5.37 (0.17)
31	1bg ⁵⁴	S	4,7-Me ₂ (trans)	DMA ^e	H	4.59	5.11	6.31	6.29	5.90 (0.69)
32	1bl ⁵⁴	S	5,6-CH ₂ C(=CHCH ₃)CH ₂ (S)		9-Cl	5.42	^g	5.42	5.42	5.42 (0.00)
33	15c ⁵³	O	5-Me (S)	CH ₂ -c-Pr	9-NHC(O)Me	3.80	3.80	3.80	3.80	3.80 (0.00)
34	1bq ⁵⁴	S	5-Me (S)	DMA ^e	8-Cl	8.30	7.85	7.76	7.74	7.79 (0.06)
35	1bs ⁵⁴	O	5-Me (S)	DMA ^e	9-Cl	6.74	6.39	6.19	6.20	6.26 (0.11)
36	1bt ⁵⁴	S	5-Me (S)	DMA ^e	9-Cl	7.37	7.81	7.51	7.45	7.59 (0.19)
37	1bu ⁵⁴	S	5-Me (S)	CH ₂ -c-Pr	9-Cl	7.47	7.45	6.84	6.91	7.06 (0.33)
38	1bv ⁵⁴	S	5-Me (S)	CH ₂ -c-Pr	H	7.22	6.39	6.05	6.16	6.20 (0.17)
39	1bx ⁵⁴	O	5-Me	<i>n</i> -Pr	H	4.22	4.43	4.29	4.34	4.35 (0.07)
40	1by ⁵⁴	S	5-Me	<i>n</i> -Pr	H	5.78	5.85	5.61	5.60	5.69 (0.14)
41	1bz ⁵⁴	O	5-Me	2-MA ^f	H	4.46	4.88	4.52	4.60	4.67 (0.19)
42	1ca ⁵⁴	S	5-Me	DMA ^e	H	7.01	6.76	6.72	6.71	6.73 (0.03)
43	1cb ⁵⁴	O	5-Me (S)	DMA ^e	H	5.48	5.34	5.40	5.45	5.40 (0.06)
44	1cc ⁵⁴	S	5-Me (S)	2-MA ^f	H	7.58	6.30	5.84	5.86	6.00 (0.26)
45	9tt ⁵²	O	5-Me	CH ₂ C(CH=CH ₂)=CH ₂	H	4.15	4.45	4.12	4.00	4.19 (0.23)
46	9ss ⁵²	O	5-Me	CH ₂ CH=CHPh (<i>Z</i>)	H	3.91	^g	3.92	3.98	3.95 (0.04)
47	9hh ⁵²	O	5-Me	CH ₂ C(Me)=CHMe (<i>E</i>)	H	4.54	5.38	4.85	4.68	4.97 (0.36)
48	9nn ⁵²	O	5-Me	CH ₂ C(Et)=CH ₂	H	4.43	4.39	5.17	5.33	4.96 (0.50)
49	9a ⁵²	O	5-Me	CH ₂ CH=CH ₂	H	4.15	4.33	4.42	4.29	4.35 (0.06)
50	9aa ⁵²	O	5-Me	<i>n</i> -Bu	H	4.00	3.94 ^d	4.40 ^d	4.57 ^d	4.30 (0.32) ^d
51	10e ⁵³	O	5-Me (S)	DMA ^e	8-Br	7.32	7.21	7.26	7.29	7.25 (0.04)
52	10f ⁵³	S	5-Me (S)	DMA ^e	8-Br	8.52	8.63	8.58	8.55	8.59 (0.04)
53	10m ⁵³	S	5-Me (S)	DMA ^e	8-Me	7.86	7.64	7.59	7.56	7.60 (0.04)
54	10a ⁵³	O	5-Me (S)	DMA ^e	8-CN	5.94	5.89	5.94	5.97	5.93 (0.04)
55	10b ⁵³	S	5-Me (S)	DMA ^e	8-CN	7.25	7.30	7.25	7.22	7.26 (0.04)
56	10l ⁵³	O	5-Me (S)	DMA ^e	8-Me	6.00	6.22	6.27	6.30	6.26 (0.04)
57	19b ⁵³	S	5-Me (S)	DMA ^e	10-OMe	5.33	5.96	5.91	5.88	5.92 (0.04)
58	19a ⁵³	O	5-Me (S)	DMA ^e	10-OMe	5.18	4.55	4.60	4.63	4.59 (0.04)
59	10c ⁵³	S	5-Me (S)	DMA ^e	8-CHO	6.73	6.73	6.73	6.73	6.73 (0.00)
60	10g ⁵³	O	5-Me (S)	DMA ^e	8-I	7.06	6.48	6.53	6.56	6.52 (0.04)
61	10h ⁵³	S	5-Me (S)	DMA ^e	8-I	7.32	7.90	7.85	7.82	7.86 (0.04)
62	10i ⁵³	O	5-Me (S)	DMA ^e	8-C≡CH	6.36	6.24	6.29	6.32	6.28 (0.04)
63	10j ⁵³	S	5-Me (S)	DMA ^e	8-C≡CH	7.53	7.65	7.60	7.57	7.61 (0.04)
64	16 ⁵³	S	5-Me (S)	CH ₂ -c-Pr	9-NO ₂	5.61	5.75	5.70	5.67	5.71 (0.04)
65	18a ⁵³	O	5-Me (S)	DMA ^e	9-CF ₃	5.23	5.06	5.11	5.14	5.10 (0.04)
66	18b ⁵³	S	5-Me (S)	DMA ^e	9-CF ₃	6.31	6.48	6.43	6.40	6.44 (0.04)

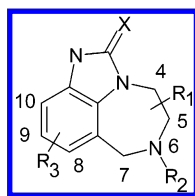


Table 1. Continued

no.	ref. ^b	X	R ₁	R ₂	R ₃	Log(1/IC ₅₀)				
						exp.	I(AB, 5-5)	I(AB, 3-4)	I(AB, 2-4)	mean ^c
67	18c ⁵³	O	5-Me (S)	CH ₂ CH=CEt ₂	9-Me	6.50	6.50	6.50	6.50	6.50 (0.00)
68	27a ⁵³	O	5-Me	DMA ^e	8-Me ^h	3.06	3.06	3.53	3.47	3.35 (0.26)
69	27c ⁵³	S	5-Me	CH ₂ CH=CEt ₂	8-Me ^h	6.61	6.61	6.14	6.20	6.32 (0.26)
70	27d ⁵³	O	5-Me	CH ₂ -c-Pr	8-Me ^h	3.22	2.70 ^d	2.86 ^d	2.93 ^d	2.83 (0.12) ^d
71	9o ⁵²	O	5-Me	CH ₂ CH ₂ CH=CH ₂	H	4.30	4.00	4.50	4.23	4.24 (0.25)
72	9s ⁵²	O	5-Me	CH ₂ -c-Pr	H	4.36	4.97	4.73	4.91	4.87 (0.12)
73	9y ⁵²	O	5-Me	CH ₂ CH=CHCH ₃ (Z)	H	4.46	4.83	4.41	4.47	4.57 (0.23)

^a Structures and experimental data for the inhibition of HIV-1 replication in MT-4 cells (IC₅₀, mol/L) for TIBO derivatives were selected from the references.⁵²⁻⁵⁴ The best fit models **I(AB, 5-5)**, **I(AB, 3-4)**, and **I(AB, 2-4)**/eq 1 obtained for *Training Set 1* were used in calculations. ^b Compound number in cited references. ^c Average log(1/IC₅₀) value calculated from the three best fit models; the standard deviation is given in parentheses. ^d "Predicted" log(1/IC₅₀) value from *Test Set 1*. ^e DMA = 3,3-dimethylallyl. ^f 2-MA = 2-methylallyl. ^g Compound was excluded on the training stage because it contains unique molecular fragment(s). ^h Pyridyl analogue of TIBO: the N atom occupies ninth position.

Table 2. Modeling of Anti-HIV Activity (log(1/IC₅₀)) for TIBO Derivatives: Statistical Criteria for Three Best Models Selected at the Training Stage for the Sets 1-3^a

no.	fragment set	<i>n</i>	<i>k</i>	<i>R</i> ²	<i>F</i>	FIT	<i>s</i>	<i>R</i> _H , %	<i>Q</i> ²
<i>Training Set 1^b</i>									
1	I(AB, 5-5)	64	40	0.908	6.09	0.15	0.66	6.8	0.671
2	I(AB, 3-4)	66	39	0.850	4.01	0.10	0.82	8.8	0.531
3	I(AB, 2-4)	66	44	0.854	2.97	0.07	0.89	8.7	0.510
<i>Training Set 2^b</i>									
1	I(AB, 5-5)	64	40	0.906	5.93	0.14	0.67	6.9	0.638
2	I(AB, 3-4)	66	39	0.854	4.14	0.10	0.80	8.6	0.536
3	I(AB, 2-4)	66	44	0.854	3.00	0.07	0.88	8.6	0.501
<i>Training Set 3^b</i>									
1	I(AB, 5-5)	64	40	0.920	7.11	0.18	0.61	6.4	0.736
2	I(AB, 3-4)	66	38	0.850	4.32	0.11	0.79	8.7	0.564
3	I(AB, 2-4)	66	43	0.854	3.18	0.07	0.86	8.7	0.536

^a Molecules are represented without hydrogen atoms. See text for the fragments definition. Fitting eq 1 was used. Statistical parameters for the training set: the number of compounds (*n*), the number of fitted coefficients (*k*), correlation coefficient (*R*), Fisher's criterion (*F*), fitness function (FIT), standard deviation (*s*), factor of Hamilton (*R*_H), and cross-validation correlation coefficient (*Q*). ^b The compounds in the training and test sets are specified in Table 1.

4, 6, 8, 16, 17, 34, and 68 (*Set 2*), and 1, 5, 21, 29, 33, 53, and 70 (*Set 3*) for which activity varies from 3.06 to 8.30 (Table 1). Results of "predictions" given in Table 3 show that the three best fit models reproduce well the experimental data for all three test sets (*R*² = 0.79–0.96; *s* = 0.3–1.0 for the linear regression $Y_{calc} = a + b \cdot Y_{exp}$). The average activities calculated for each compound as an arithmetic mean of "predicted" (with the three abovementioned individual QSAR models) values led to better statistical criteria: *R*² = 0.88–0.94; *s* = 0.4–0.9 (Table 3).

4.1.2. HEPT Derivatives. The initial data set containing 93 compounds was split into training (84 compounds) and test (8 compounds) sets. Elimination of compounds with unique fragments led to some reduction of the training sets to 65–76 compounds (Table 4). The three most adequate models selected at the training stage (**I(AB, 2-5)**, **II(Hy)**, and **I(AB, 2-4)**) coupled with fitting eq 1 correspond to good correlations between experimental and calculated activities: *R*² = 0.90–0.98; *s* = 0.55–0.66, *F* = 4–13, *Q*² = 0.37–0.87, and FIT = 0.1–0.4 (Table 5).

Test sets contained compounds 9, 12, 15, 28, 34, 40, 70, and 71 (*Set 1*), 10, 17, 35, 46, 57, 59, 67, and 73 (*Set 2*), and 42, 49, 64, 68, 69, 70, 72, and 74 (*Set 3*) for which activity varied from 4.89 to 8.27 (Table 6). Results of "predictions" given in Table 6 shows that for *Set 1* and *Set 2*, the three best fit models reproduce well the experimental data (*R*² = 0.79–0.89; *s* = 0.3–0.5), whereas for *Set 3* the agreement with the experiment is less good (*R*² = 0.39–0.78; *s* = 0.7–0.8). Nevertheless, the mean predicted

activities correspond to reasonable statistical criteria: *R*² = 0.87–0.89; *s* = 0.3–0.4 for *Set 1* and *Set 2* and *R*² = 0.84; *s* = 0.4 for *Set 3* (Table 6).

4.1.3. CU Derivatives. The initial data set containing 118 compounds was split into training (109 compounds) and test (9 compounds) sets. Having excluded 25 compounds with unique fragments, TRAIL reduced the three training sets to 84 compounds (Table 7). The three best fit models correspond to **I(AB, 2-4)**, **I(AB, 3-4)**, and **I(AB, 4-4)** fragments coupled with fitting eq 1 for which *R*² = 0.82–0.90; *s* = 0.63–0.76, *F* = 4.0–6.5, *Q*² = 0.29–0.57, and FIT = 0.1–0.2 (Table 8).

Test sets contained compounds 5, 7, 9, 12, 13, 16, 42, 72, and 85 (*Set 1*), 8, 11, 15, 19, 21, 56, 70, 82, and 87 (*Set 2*), and 14, 28, 37, 48, 51, 60, 62, 69, and 71 (*Set 3*) for which log(1/*K_i*) varied from 5.96 to 10.96 (Table 9). Table 9 shows that for *Set 1*, *Set 2*, and *Set 3* the three best fit models reproduce experimental data (*R*² = 0.83–0.95; *s* = 0.3–0.8). Calculations for the mean "predicted" activities led to reasonable statistical criteria: *R*² = 0.84; *s* = 0.6 (*Set 1*), *R*² = 0.85; *s* = 0.4 (*Set 2*) and *R*² = 0.92; *s* = 0.3 for *Set 3* (Table 9).

4.2. Design of Potential Anti-HIV Actives. Generation of new potential actives can be performed in three main steps: (i) choice of the molecular core and preparation of its "optimal" substituents; (ii) generation of combinatorial library; and (iii) evaluation of activity of virtual compounds and hits selection. In this section we demonstrate an application of this strategy to design the hypothetical HEPT

Table 3. Experimental and "Predicted" Activities ($\log(1/IC_{50})$) for TIBO Derivatives from the Test Sets 1–3^a

compound. no.	experimental	predicted			mean ^b
		I(AB, 5–5)	I(AB, 3–4)	I(AB, 2–4)	
Test Set 1					
2	6.80	7.74	8.01	7.96	7.90 (0.14)
9	4.90	4.60	4.62	4.31	4.51 (0.17)
19	7.64	7.74	7.59	7.54	7.62 (0.10)
20	4.25	5.00	4.02	4.10	4.37 (0.54)
27	5.94	6.51	6.31	6.29	6.37 (0.12)
50	4.00	3.94	4.40	4.57	4.30 (0.32)
70	3.22	2.70	2.86	2.93	2.83 (0.12)
R^2	s	0.937	0.935	0.914	0.943
		0.53	0.54	0.61	0.50
Test Set 2					
4	4.64	5.15	5.91	5.82	5.63 (0.42)
6	6.17	5.54	6.90	6.94	6.46 (0.80)
8	4.13	4.28	4.24	4.17	4.23 (0.06)
16	5.61	5.72	5.27	5.35	5.44 (0.24)
17	7.11	6.51	6.42	6.43	6.45 (0.05)
34	8.30	7.53	7.83	7.80	7.72 (0.17)
68	3.06	3.32	4.27	4.23	3.94 (0.54)
R^2	s	0.964	0.823	0.835	0.918
		0.29	0.62	0.60	0.41
Test Set 3					
1	7.34	8.05	8.38	8.37	8.27 (0.19)
5	4.49	3.94	4.56	4.66	4.39 (0.39)
21	5.65	4.88	5.30	5.29	5.15 (0.24)
29	6.32	7.87	7.17	7.11	7.38 (0.42)
33	3.80	2.84	1.74	2.08	2.22 (0.57)
53	7.86	7.38	7.27	7.23	7.29 (0.08)
70	3.22	2.97	3.03	3.09	3.03 (0.06)
R^2	s	0.859	0.792	0.869	0.880
		1.00	0.70	0.92	0.89

^a Compound numbers are given according to Table 1. Fitting eq 1 was used. ^b Average $\log(1/IC_{50})$ value calculated from the three best fit models; the standard deviation is given in parentheses. ^c Statistical characteristics (R and s) for the correlation between experimental and predicted (by given model) activities.

derivatives. The combinatorial library has been generated for the Markush structure given in Figure 2. This structure has a symmetrically disubstituted aromatic fragment at the C_S atom, as experimentally disubstituted derivatives are more active than monosubstituted ones (Table 4).

Preparation of optimal substituents R₁, R₃, and R₄ was performed using fragment contributions a_i calculated by TRAIL for the three best fit models on the training stage. To build R₁, R₃, and R₄, we have selected molecular fragments which correspond to the large possible positive a_i . For example, taking R₄ = Me instead of R₄ = H (Figure 3), for the I(AB, 2–4) (eq 1) model TRAIL generates five new atom/bond sequences: one C–C, two C–C_{ar}–C_{ar}, and two C–C_{ar}–C_{ar}–C_{ar} fragments. Because C–C_{ar}–C_{ar} and C–C_{ar}–C_{ar}–C_{ar} fragments were not linearly independent for a given model, on the training stage TRAIL has calculated their overall contribution (0.820 $\log(1/EC_{50})$ units); the contribution of the C–C fragment is negative (–1.006 $\log(1/EC_{50})$ units). Thus, the overall effect of replacement of H with Me is $2 \times 0.820 - 1.006 = 0.634 \log(1/EC_{50})$ units, which corresponds to the increase of the activity (Figure 3). In such a way, 7 candidates for R₁ and 6 candidates for R₃ and R₄ were suggested (Figure 2).

Generation of the virtual combinatorial library containing 252 compounds has been performed by the *CombiLib* program applying systematic attachment of substituents R₁, R₃, and R₄ to the molecular core. Then, activity of each compound in the library was evaluated by TRAIL using fragment contributions for the three best fit models selected on the training stage followed by calculations of the average

data set for these models. Several hits selected in Table 10 correspond to the compounds whose hypothetical activities are larger than the largest activity ($\log(1/EC_{50}) = 8.57$, Table 4) observed experimentally for this type of compounds. The best potential anti-HIV actives correspond to R₁ = *i*-Pr, *t*-Bu; R₃ = CH₂OCH₂Ph, CH₂OC₂H₅; and R₄ = Me, Et, *i*-Pr (Table 10).

5. DISCUSSION

5.1. Substructural Molecular Fragment Method as a Flexible and Reliable Structure–Property Tool. Empirical methods of calculation of a given molecular property from the contributions of molecular fragments are widely used in chemistry. Thus, octanol–water partition coefficients, oil–gas partition coefficients,^{43,66} formation enthalpies and heat capacities,⁶⁷ and aqueous solubilities⁴⁰ can be calculated using tabulated contributions of a priori defined fragments (atoms, bonds, or functional groups increments). In contrast to those approaches, the SMF method is not restricted to any particular decomposition scheme, but generates a large number of models involving 49 types of fragments used in combination with three fitting equations. Simultaneous utilization of several models (instead of a single one) at the test stage improves the predictability of property calculations. Indeed, any particular model involving a relatively large number of variables may lead to excellent statistical parameters for the training set, but poor correlation with experimental data at the test stage. An average data set which involves data calculated with several best SMF models,

Table 4. Initial Data Set of HEPT Derivatives: Experimental and Calculated Activities $\log(1/EC_{50})^a$

no.	ref. ^b	R ₁	R ₂	R ₃	X	Log(1/EC ₅₀)				
						exp.	I(AB, 2–5)	II(Hy)	I(AB, 2–4)	mean ^c
1	8 ⁵⁹	Me	S(<i>c</i> -Hex)	CH ₂ OCH ₂ CH ₂ OH	O	5.09	5.09	e	5.09	5.09 (0.00)
2	2 ⁶² ; 27 ⁵⁹	Me	CH ₂ Ph	CH ₂ OCH ₂ CH ₂ OH	O	4.64	4.86	5.21	5.30	5.12 (0.23)
3	24 ⁵⁶	CH ₂ Ph	SPh	CH ₂ OCH ₂ CH ₂ OH	O	4.37	e	4.37	5.57	4.97 (0.85)
4	49 ⁵⁹	CH=CPh ₂	SPh	CH ₂ OCH ₂ CH ₂ OH	O	6.08	e	e	6.16	6.16
5	54 ⁵⁹	CH=CHPh (cis)	SPh	CH ₂ OCH ₂ CH ₂ OH	O	5.22	e	e	5.06	5.06
6	55 ⁵⁹	CH=CH ₂	SPh	CH ₂ OCH ₂ CH ₂ OH	O	5.96	e	e	6.04	6.04
7	5 ⁶¹ ; 12 ⁵⁶	Me	SPh(2-Me)	CH ₂ OCH ₂ CH ₂ OH	O	4.15	e	5.69	e	5.69
8	8 ⁶¹	Me	SPh(2-OMe)	CH ₂ OCH ₂ CH ₂ OH	O	4.72	e	4.69	e	4.69
9	9 ⁶¹ ; 13 ⁵⁶	Me	SPh(3-Me)	CH ₂ OCH ₂ CH ₂ OH	O	5.59	5.44 ^d	5.69 ^d	5.86 ^d	5.66 (0.21) ^d
10	10 ⁶¹	Me	SPh(3-Et)	CH ₂ OCH ₂ CH ₂ OH	O	5.57	6.12	5.64	5.76	5.84 (0.25)
11	11 ⁶¹	Me	SPh(3- <i>t</i> -Bu)	CH ₂ OCH ₂ CH ₂ OH	O	4.92	4.74	e	4.80	4.77 (0.05)
12	15 ⁶¹	Me	SPh(3-Cl)	CH ₂ OCH ₂ CH ₂ OH	O	4.89	5.82 ^d	5.59 ^d	5.73 ^d	5.71 (0.12) ^d
13	18 ⁶¹	Me	SPh(3-NO ₂)	CH ₂ OCH ₂ CH ₂ OH	O	4.47	e	4.16	e	4.16
14	20 ⁶¹	Me	SPh(3-OMe)	CH ₂ OCH ₂ CH ₂ OH	O	4.66	e	4.69	e	4.69
15	28 ⁶¹ ; 15 ⁵⁶	Me	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	O	6.59	6.39 ^d	6.28 ^d	6.49 ^d	6.39 (0.11) ^d
16	29 ⁶¹	Me	SPh(3,5-Cl ₂)	CH ₂ OCH ₂ CH ₂ OH	O	5.89	6.09	6.08	6.24	6.13 (0.09)
17	30 ⁶¹	Me	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	S	6.66	6.38	6.19	6.44	6.34 (0.13)
18	33 ⁶¹	Me	SPh(3-C(O)OMe)	CH ₂ OCH ₂ CH ₂ OH	O	5.10	5.10	5.03	5.10	5.08 (0.04)
19	34 ⁶¹	Me	SPh(3-C(O)Me)	CH ₂ OCH ₂ CH ₂ OH	O	5.14	5.14	e	5.14	5.14 (0.00)
20	42 ⁶¹ ; 23 ⁵⁶	CH ₂ CH=CH ₂	SPh	CH ₂ OCH ₂ CH ₂ OH	O	5.60	e	e	5.60	5.60
21	48 ⁶¹	Et	SPh	CH ₂ OCH ₂ CH ₂ OH	S	6.96	6.93	6.36	6.31	6.53 (0.35)
22	49 ⁶¹	Pr	SPh	CH ₂ OCH ₂ CH ₂ OH	S	5.00	5.23	6.19	5.88	5.77 (0.49)
23	51 ⁶¹	Et	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	S	8.11	7.76	7.53	7.57	7.62 (0.12)
24	52 ⁶¹	<i>i</i> -Pr	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	S	8.30	8.23	8.28	8.45	8.32 (0.11)
25	53 ⁶¹	Et	SPh(3,5-Cl ₂)	CH ₂ OCH ₂ CH ₂ OH	S	7.37	7.46	7.33	7.32	7.37 (0.08)
26	54 ⁶¹ ; 13 ⁵⁷	Et	SPh	CH ₂ OCH ₂ CH ₂ OH	O	6.92	6.93	6.45	6.35	6.58 (0.31)
27	55 ⁶¹	Pr	SPh	CH ₂ OCH ₂ CH ₂ OH	O	5.47	5.24	6.28	5.92	5.82 (0.53)
28	56 ⁶¹	<i>i</i> -Pr	SPh	CH ₂ OCH ₂ CH ₂ OH	O	7.20	7.40 ^d	7.20 ^d	7.23 ^d	7.28 (0.11) ^d
29	57 ⁶¹	Et	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	O	7.89	7.77	7.62	7.62	7.67 (0.09)
30	58 ⁶¹	<i>i</i> -Pr	SPh(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	O	8.57	8.24	8.37	8.50	8.37 (0.13)
31	59 ⁶¹	Et	SPh(3,5-Cl ₂)	CH ₂ OCH ₂ CH ₂ OH	O	7.85	7.47	7.42	7.37	7.42 (0.05)
32	26 ⁶⁰	Me	SPh	CH ₂ OMe	O	5.68	5.98	5.75	6.61	6.11 (0.45)
33	28 ⁶⁰	Me	SPh	CH ₂ OPr	O	5.44	5.44	5.72	6.12	5.76 (0.34)
34	29 ⁶⁰	Me	SPh	CH ₂ OBu	O	5.33	4.62 ^d	5.56 ^d	5.69 ^d	5.29 (0.58) ^d
35	31 ⁶⁰ ; 12 ⁵⁷	Me	SPh	CH ₂ OCH ₂ Ph	O	7.06	6.55	6.22	6.32	6.36 (0.17)
36	33 ⁶⁰	Et	SPh(3,5-Me ₂)	CH ₂ OEt	S	8.36	8.35	8.31	8.34	8.33 (0.02)
37	34 ⁶⁰	Et	SPh(3,5-Cl ₂)	CH ₂ OEt	S	7.89	8.05	8.10	8.08	8.08 (0.03)
38	35 ⁶⁰	Et	SPh	CH ₂ - <i>i</i> -Pr	S	6.66	e	7.07	6.81	6.94 (0.18)
39	37 ⁶⁰	Et	SPh	CH ₂ OCH ₂ - <i>c</i> -Hex	S	6.46	6.40	6.16	5.93	6.16 (0.24)
40	38 ⁶⁰	Et	SPh	CH ₂ OCH ₂ Ph	S	8.11	7.92 ^d	7.47 ^d	7.40 ^d	7.60 (0.28) ^d
41	39 ⁶⁰	Et	SPh(3,5-Me ₂)	CH ₂ OCH ₂ Ph	S	8.16	8.76	8.65	8.67	8.69 (0.06)
42	40 ⁶⁰	Et	SPh	CH ₂ OCH ₂ C ₆ H ₄ (4-Me)	S	7.11	6.99	8.06	8.04	7.69 (0.61)
43	41 ⁶⁰	Et	SPh	CH ₂ OCH ₂ C ₆ H ₄ (4-Cl)	S	7.92	7.92	7.96	7.91	7.93 (0.02)
44	42 ⁶⁰	Et	SPh	CH ₂ OCH ₂ CH ₂ Ph	S	7.04	7.03	6.95	6.41	6.80 (0.34)
45	43 ⁶⁰	<i>i</i> -Pr	SPh	CH ₂ OEt	S	7.85	7.98	7.88	7.95	7.94 (0.05)
46	44 ⁶⁰	<i>i</i> -Pr	SPh	CH ₂ OCH ₂ Ph	S	8.17	8.39	8.22	8.28	8.30 (0.09)
47	45 ⁶⁰	<i>c</i> -Pr	SPh	CH ₂ OEt	S	7.02	7.01	7.26	7.19	7.15 (0.13)
48	48 ⁶⁰	Et	SPh(3,5-Cl ₂)	CH ₂ OEt	O	8.13	8.06	8.19	8.13	8.13 (0.07)
49	49 ⁶⁰	Et	SPh	CH ₂ O- <i>i</i> -Pr	O	6.47	6.77	6.66	6.24	6.55 (0.28)
50	50 ⁶⁰	Et	SPh	CH ₂ O- <i>c</i> -Hex	O	5.40	5.60	5.55	6.21	5.79 (0.36)
51	51 ⁶⁰	Et	SPh	CH ₂ OCH ₂ - <i>c</i> -Hex	O	6.35	6.41	6.25	5.98	6.21 (0.22)
52	52 ⁶⁰ ; 15 ⁵⁷	Et	SPh	CH ₂ OCH ₂ Ph	O	8.23	7.93	7.56	7.45	7.65 (0.25)
53	54 ⁶⁰	Et	SPh	CH ₂ OCH ₂ CH ₂ Ph	O	7.02	7.03	7.04	6.46	6.84 (0.34)
54	56 ⁶⁰	<i>i</i> -Pr	SPh	CH ₂ OCH ₂ Ph	O	8.57	8.40	8.31	8.33	8.35 (0.05)
55	60 ⁶⁰	Me	SPh	Et	O	5.66	5.66	5.68	5.81	5.72 (0.08)
56	61 ⁶⁰	Me	SPh	Bu	O	5.92	5.92	5.36	5.46	5.58 (0.30)
57	27 ⁶²	Et	CH ₂ Ph	CH ₂ OCH ₂ CH ₂ OH	O	6.46	6.59	6.55	6.43	6.52 (0.08)
58	28 ⁶²	Et	CH ₂ Ph(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	O	7.89	7.94	7.73	7.70	7.79 (0.13)
59	29 ⁶²	Et	CH ₂ Ph	CH ₂ OEt	O	7.39	7.17	7.32	7.20	7.23 (0.08)
60	31 ⁶²	<i>i</i> -Pr	CH ₂ Ph	CH ₂ OCH ₂ CH ₂ OH	O	7.20	7.40	7.30	7.31	7.34 (0.06)
61	32 ⁶²	<i>i</i> -Pr	CH ₂ Ph(3,5-Me ₂)	CH ₂ OCH ₂ CH ₂ OH	O	8.57	8.75	8.48	8.58	8.60 (0.14)
62	33 ⁶²	<i>i</i> -Pr	CH ₂ Ph	CH ₂ OEt	O	8.38	7.99	8.08	8.07	8.04 (0.05)
63	34 ⁶²	<i>i</i> -Pr	CH ₂ Ph(3,5-Me ₂)	CH ₂ OEt	O	9.22	9.34	9.25	9.34	9.31 (0.05)
64	38 ⁶²	<i>i</i> -Pr	CH ₂ Ph	Bu	O	7.38	7.44	7.56	7.55	7.51 (0.07)
65	39 ⁶²	Et	CH ₂ Ph	CH ₂ CH ₂ OMe	O	6.60	6.53	6.48	6.50	6.50 (0.03)
66	40 ⁶²	<i>i</i> -Pr	CH ₂ Ph	CH ₂ CH ₂ OMe	O	7.28	7.35	7.23	7.38	7.32 (0.08)

Table 4. Continued

no.	ref. ^b	R ₁	R ₂	R ₃	X	Log(1/EC ₅₀)				mean ^c
						exp.	I(AB, 2-5)	II(Hy)	I(AB, 2-4)	
67	1 ⁶⁰ , 5e ⁵⁵	Me	SPh	CH ₂ OCH ₂ CH ₂ OH	O	5.15	5.55	5.11	5.22	5.29 (0.23)
68	50 ⁶¹	<i>i</i> -Pr	SPh	CH ₂ OCH ₂ CH ₂ OH	S	7.23	7.40	7.11	7.18	7.23 (0.15)
69	2 ⁶⁰	Me	SPh	CH ₂ OCH ₂ CH ₂ OMe	O	5.06	5.06	5.23	5.06	5.12 (0.10)
70	27 ⁶⁰ ; 8 ⁵⁷	Me	SPh	CH ₂ OEt	O	6.48	6.14 ^d	5.88 ^d	5.99 ^d	6.00 (0.13) ^d
71	32 ⁶⁰	Et	SPh	CH ₂ OEt	S	7.59	7.51 ^d	7.13 ^d	7.07 ^d	7.24 (0.24) ^d
72	36 ⁶⁰	Et	SPh	CH ₂ O- <i>c</i> -Hex	S	5.80	5.60	5.46	6.16	5.74 (0.37)
73	14 ⁵⁷	Et	SPh	CH ₂ OEt	O	7.72	7.52	7.22	7.12	7.29 (0.21)
74	47 ⁶⁰	Et	SPh(3,5-Me ₂)	CH ₂ OEt	O	8.27	8.36	8.40	8.39	8.38 (0.02)
75	53 ⁶⁰	Et	SPh(3,5-Me ₂)	CH ₂ OCH ₂ Ph	O	8.49	8.77	8.74	8.72	8.74 (0.02)
76	55 ⁶⁰	<i>i</i> -Pr	SPh	CH ₂ OEt	O	7.92	7.99	7.97	7.99	7.99 (0.01)
77	57 ⁶⁰	<i>c</i> -Pr	SPh	CH ₂ OEt	O	7.00	7.01	7.35	7.24	7.20 (0.17)
78	30 ⁶²	Et	CH ₂ Ph(3,5-Me ₂)	CH ₂ OEt	O	8.80	8.52	8.50	8.46	8.50 (0.03)
79	37 ⁶²	Et	CH ₂ Ph	Bu	O	6.68	6.62	6.81	6.67	6.70 (0.09)
80	11 ⁵⁸	Me	SPh	CH ₂ OCH ₂ CH ₂ OH	S	6.01	5.55	5.02	5.18	5.25 (0.27)
81	7 ⁵⁹	Me	SBu- <i>n</i>	CH ₂ OCH ₂ CH ₂ OH	O	3.89	3.89	e	3.89	3.89 (0.00)
82	7 ⁶¹	Me	SPh(2-NO ₂)	CH ₂ OCH ₂ CH ₂ OH	O	3.85	e	4.16	e	4.16
83	28 ⁵⁶	Me	SPh	CH ₂ OCH ₂ CH ₂ OC(O)Me	O	5.12	5.12	e	5.12	5.12 (0.00)
84	29 ⁵⁶	Me	SPh	CH ₂ OCH ₂ CH ₂ OC(O)Ph	O	5.17	5.17	5.24	5.17	5.19 (0.04)

^a Structures and experimental data for the inhibition of HIV-1 replication in MT-4 Cells (EC₅₀, mol/L) for HEPT derivatives were selected from the references.⁵⁵⁻⁶² The best fit models I(AB, 2-5), II(Hy), and I(AB, 2-4)/eq 1 obtained for the training set 1 were used in calculations. ^b Compound number in cited references. ^c Average log(1/EC₅₀) value calculated from the three best fit models; the standard deviation is given in parentheses. ^d "Predicted" log(1/EC₅₀) value from *Test Set 1*. ^e Compound was excluded on the training stage because it contains unique molecular fragment(s).

Table 5. Modeling of Anti-HIV Activity (log(1/EC₅₀)) for HEPT Derivatives: Statistical Criteria for Three Best Models Selected at the Training Stage for the Sets 1-3^a

no.	fragment set	<i>n</i>	<i>k</i>	<i>R</i> ²	<i>F</i>	FIT	<i>s</i>	<i>R</i> _H , %	<i>Q</i> ²
<i>Training Set 1^b</i>									
1	I(AB, 2-5)	65	58	0.970	4.08	0.07	0.66	3.1	0.483
2	II(Hy)	67	30	0.904	11.99	0.38	0.56	6.0	0.759
3	I(AB, 2-4)	71	37	0.897	8.18	0.22	0.59	6.0	0.613
<i>Training Set 2^b</i>									
1	I(AB, 2-5)	65	59	0.980	5.26	0.09	0.59	2.6	0.870
2	II(Hy)	67	30	0.951	12.07	0.38	0.56	6.1	0.762
3	I(AB, 2-4)	71	37	0.904	7.98	0.21	0.61	6.1	0.372
<i>Training Set 3^b</i>									
1	I(AB, 2-5)	65	59	0.976	4.27	0.07	0.65	2.8	0.767
2	II(Hy)	66	29	0.908	13.18	0.43	0.55	5.9	0.724
3	I(AB, 2-4)	71	37	0.901	8.64	0.23	0.59	5.9	0.473

^a Molecules are represented without hydrogen atoms. Fitting eq 1 was used in all cases. See Table 2 footnotes for definition of statistical parameters.

^b The compounds in the training and test sets are specified in Table 4.

smoothes inaccuracies of individual "predicted" data sets. Here, such averaging always led to good correlation between experimental and calculated activities (Figure 4), although some particular models were not predictive enough (Tables 3, 6, and 9). Because of its ability to adapt to any particular set of compounds and properties, the SMF method is able to assess three different types of anti-HIV activity: log(1/IC₅₀), log(1/EC₅₀), and (log(1/*K*_i)) for a large variety of TIBO, HEPT and CU derivatives, respectively.

Another important aspect of application of the SMF method is related to generation and filtering of focused virtual combinatorial libraries using the molecular fragment contributions obtained at the training stage. In Sections 4.2 and 5.3 we have demonstrated that potential actives can be selected from relatively small libraries, prepared from selected "optimized" substituents.

5.2. Modeling of Anti-HIV Activity: SMF Method vs Other QSPR Techniques. In this section we compare the performance of the SMF calculations on HEPT, TIBO, and CU derivatives with those of previous QSAR studies.^{3,6,14,23,26} Because statistical criteria of fitting models obtained for different data sets using different techniques (partial least-squares (PLS), multiple linear regression (MLR), artificial

neural networks (ANN), and comparative molecular field analysis (CoMFA)) can hardly be compared, we have performed a linear regression analysis for calculated (*Y*_{calc}) in this work or in the literature^{3,6,14,23,26} for anti-HIV activities vs experimental data (*Y*_{exp}). As *Y*_{calc} activities in this work, we used the average values for the data sets calculated using the three best fit models for *Set 1* (see Section 4).

5.2.1. TIBO Derivatives. Hannongbua et al¹² used CoMFA to model log1/IC₅₀ activity for 46 TIBO derivatives. Later, this data set was used at the training stage by Huuskonen⁶ who performed MLR studies using atom level E-state indices and calculated molecular properties (logP, MR). One may notice that the data set in reference 12 and the training set used in reference 6 contained 9 compounds (of 46) for which only the lower limit of activity (IC₅₀) rather than its accurate value were given in the experimental paper of Breslin et al.⁵⁴

Statistical criteria for the *Y*_{calc} vs *Y*_{exp} correlations obtained in this work for "average" data sets (*R*² = 0.88; *s* = 0.43 for the training stage and *R*² = 0.94; *s* = 0.50 for the test stage (Figure 4 and Table 3)) are similar to those reported in Tables 1 and 5 in reference 6 (*R*² = 0.74; *s* = 0.56 for the training set and *R*² = 0.79; *s* = 0.55 for the test set) or to those obtained by Hannongbua et al¹² (*R*² = 0.86; *s* =

Table 6. Experimental and “Predicted” Activities ($\log(1/EC_{50})$) for HEPT Derivatives from the Test Sets 1–3^a

compound no.	experimental	predicted			
		I(AB, 2–5)	II(Hy)	I(AB, 2–4)	mean ^b
Test Set 1					
9	5.59	5.44	5.69	5.86	5.66 (0.21)
12	4.89	5.82	5.59	5.73	5.71 (0.12)
15	6.59	6.39	6.28	6.49	6.39 (0.11)
28	7.20	7.40	7.20	7.23	7.28 (0.11)
34	5.33	4.62	5.56	5.69	5.29 (0.58)
40	8.11	7.92	7.47	7.40	7.60 (0.28)
70	6.48	6.14	5.88	5.99	6.00 (0.13)
71	7.59	7.51	7.13	7.07	7.24 (0.24)
R^2	s	0.832	0.893	0.887	0.887
	c	0.50	0.28	0.26	0.32
Test Set 2					
10	5.57	6.56	5.65	5.73	5.98 (0.50)
17	6.66	6.38	6.09	6.41	6.29 (0.18)
35	7.06	6.50	6.06	6.19	6.25 (0.23)
46	8.17	8.20	8.20	8.36	8.25 (0.09)
57	6.46	6.71	6.58	6.37	6.55 (0.17)
59	7.39	7.24	7.42	7.18	7.28 (0.12)
67	5.15	5.68	5.00	5.08	5.25 (0.37)
73	7.72	7.68	7.31	7.11	7.37 (0.29)
R^2	s	0.792	0.861	0.878	0.867
	c	0.39	0.42	0.38	0.37
Test Set 3					
42	7.11	6.03	8.33	8.35	7.57 (1.33)
49	6.47	7.65	d	6.07	6.86 (1.12)
64	7.38	7.53	7.42	7.63	7.53 (0.11)
68	7.23	7.35	7.09	7.26	7.23 (0.13)
69	5.06	5.77	5.35	7.02	6.05 (0.87)
70	6.48	6.45	5.92	5.87	6.08 (0.32)
72	5.80	5.33	d	6.35	5.84 (0.72)
74	8.27	8.52	8.54	8.36	8.47 (0.10)
R^2	s	0.624	0.785	0.392	0.835
	c	0.73	0.66	0.82	0.40

^a Compound numbers are given according to Table 4. Fitting eq 1 was used. ^b Average $\log(1/EC_{50})$ value calculated from the three best fit models; the standard deviation is given in parentheses. ^c Statistical characteristics (R and s) for the correlation between experimental and predicted (by given model) activities. ^d Fragment contribution(s) for this compound was not available on the training stage.

0.44 for the training set and $R^2 = 0.86$; $s = 0.44$ for the test set).

5.2.2. HEPT Derivatives. Luco and Ferretti²³ used partial least-squares (PLS) and multiple linear regression (MLR) methods to model $\log 1/EC_{50}$ value for 107 HEPT derivatives (80 and 27 compounds in the training and test sets, respectively). Later, for the same as in reference 23 data sets, Jalali-Heravi and Parastar³ applied MLR and artificial neural network (ANN) methods using 59 topological, 3D geometrical and electronic descriptors. Hannongbua et al¹² has performed a CoMFA study to model $\log 1/EC_{50}$ for 101 HEPT compounds (82 and 19 compounds in the training and test sets, respectively).

Statistical criteria for \mathbf{Y}_{exp} vs \mathbf{Y}_{calc} correlations calculated for “average” data sets are $R^2 = 0.94$; $s = 0.32$ for the training stage and $R^2 = 0.89$; $s = 0.32$ for the test stage (Table 6 and Figure 4). These results are comparable with those obtained for the training sets by ANN ($R^2 = 0.92$, $s = 0.38$),³ MLR ($R^2 = 0.90$, $s = 0.40$ ²³ and $R^2 = 0.81$, $s = 0.53$ ³), and PLS ($R^2 = 0.88$, $s = 0.41$)²³ methods, and calculated for the training and test sets using CoMFA ($R^2 = 0.83$, $s = 0.50$ and $R^2 = 0.69$, $s = 0.80$, respectively).¹² Because for the compounds selected in references 3 and 23 only the upper limit of experimental activities ($\log 1/EC_{50}$) rather than their exact values were available, we have not

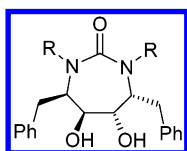
performed calculations of \mathbf{Y}_{exp} vs \mathbf{Y}_{calc} correlations for these test sets.

5.2.3. CU Derivatives. Debnath¹⁴ performed CoMFA studies to model the activity ($\log 1/K_i$) of 118 cyclic urea derivatives. Although \mathbf{Y}_{exp} vs \mathbf{Y}_{calc} correlations performed for 3 training sets containing 93 compounds in each one led to excellent results ($R^2 = 0.96$ – 0.98 and $s = 0.23$ – 0.25), statistical criteria obtained for the test sets are less good ($R^2 = 0.55$ – 0.64 and $s = 0.59$ – 0.87). Here, using the SMF method we have obtained similar statistical criteria for the training and test sets ($R^2 = 0.88$; $s = 0.39$ and $R^2 = 0.84$; $s = 0.64$, respectively, Figure 4).

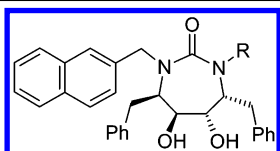
5.3. Are Predictions of Potential Actives Reasonable?

Recently, Pasqual et al.⁴⁶ reported the design of a large combinatorial library of HEPT analogues containing 125 396 molecules, where only 180 (i.e. 0.15%) of which were actually synthesized and tested. This study clearly shows that a huge combinatorial space of HEPT compounds remains to be investigated and new active compounds could be discovered.

Here, we have generated a small virtual library of 252 compounds focusing the cells in the combinatorial space near the lead. Having estimated by TRAIL the activity values of generated compounds, we suggested that several molecules with $R_1 = i\text{-Pr}$, $t\text{-Bu}$, $R_3 = \text{CH}_2\text{OCH}_2\text{Ph}$, $\text{CH}_2\text{OC}_2\text{H}_5$, and

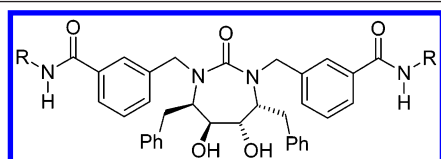
Table 7. Initial Data Set of Cyclic Ureas: Experimental and Calculated Activities $\log(1/K_i)^a$ 

no.	R	Log(1/ K_i)				
		exp.	I(AB, 2–4)	I(AB, 3–4)	I(AB, 4–4)	mean ^b
1	methyl	5.24	5.24	5.24	6.95	5.81 (0.99)
2	ethyl	7.00	8.11	8.09	7.43	7.88 (0.39)
3	<i>n</i> -propyl	8.10	8.09	8.20	7.90	8.06 (0.15)
4	<i>n</i> -butyl	8.85	8.00	8.04	7.81	7.95 (0.12)
5	<i>n</i> -pentyl	8.80	7.91 ^c	7.88 ^c	7.73 ^c	7.84 (0.10) ^c
6	<i>n</i> -hexyl	8.34	7.82	7.72	7.65	7.73 (0.09)
7	<i>n</i> -heptyl	6.59	7.73 ^c	7.56 ^c	7.56 ^c	7.62 (0.10) ^c
8	CH ₂ CH ₂ OCH ₃	6.10	6.10	6.10	6.10	6.10 (0.00)
9	CH ₂ CH ₂ OCH ₂ CH ₃	5.96	6.63 ^c	5.97 ^c	5.05 ^c	5.88 (0.79) ^c
10	CH ₂ CH ₂ OCH ₂ CH ₂ OCH ₃	5.11	5.11	5.11	5.11	5.11 (0.00)
11	<i>i</i> -butyl	7.31	8.10	8.43	8.37	8.30 (0.17)
12	<i>i</i> -pentyl	7.92	7.95 ^c	8.00 ^c	7.73 ^c	7.89 (0.14) ^c
13	<i>i</i> -hexyl	8.15	7.86 ^c	7.83 ^c	7.65 ^c	7.78 (0.11) ^c
14	<i>i</i> -heptyl	7.52	7.78	7.67	7.56	7.67 (0.11)
15	<i>i</i> -octyl	6.96	7.69	7.51	7.48	7.56 (0.11)
16	neohexyl	7.44	7.95 ^c	8.07 ^c	7.65 ^c	7.89 (0.22) ^c
17	allyl	8.28	8.30	8.30	8.06	8.22 (0.14)
18	2-methylpropen-3-yl	8.14	8.14	8.14	8.53	8.27 (0.23)
19	cyclopropylmethyl	8.68	8.48	8.31	8.37	8.39 (0.09)
20	cyclobutylmethyl	8.89	8.44	8.26	8.28	8.33 (0.09)
21	cyclopentylmethyl	8.37	8.43	8.34	8.20	8.32 (0.12)
22	cyclohexylmethyl	7.43	7.24	7.62	7.95	7.61 (0.36)
23	benzyl	8.52	8.53	8.53	8.76	8.61 (0.13)
24	3-picolyl	8.01	7.86	7.86	8.00	7.91 (0.08)
25	4-picolyl	7.05	7.57	7.57	7.62	7.59 (0.03)
26	α -naphthylmethyl	7.07	7.07	7.07	8.56	7.57 (0.86)
27	β -naphthylmethyl	9.51	9.30	9.30	8.94	9.18 (0.21)
28	<i>m</i> -fluorobenzyl	8.52	8.56	8.56	8.60	8.58 (0.02)
29	<i>p</i> -fluorobenzyl	8.85	8.56	8.56	8.60	8.58 (0.02)
30	<i>m</i> -chlorobenzyl	9.05	8.67	8.67	8.67	8.67 (0.00)
31	<i>p</i> -chlorobenzyl	8.28	8.67	8.67	8.67	8.67 (0.00)
32	<i>m</i> -bromobenzyl	8.85	8.21	8.21	8.21	8.21 (0.00)
33	<i>p</i> -bromobenzyl	7.57	8.21	8.21	8.21	8.21 (0.00)
34	<i>m</i> -methylbenzyl	8.15	8.17	8.17	8.01	8.11 (0.09)
35	<i>p</i> -methylbenzyl	8.24	8.17	8.17	8.01	8.11 (0.09)
36	<i>m</i> -(trifluoromethyl)benzyl	7.66	7.49	7.49	7.45	7.48 (0.02)
37	<i>p</i> -(trifluoromethyl)benzyl	7.29	7.49	7.49	7.45	7.48 (0.02)
38	<i>m</i> -methoxybenzyl	8.80	7.73	7.73	7.78	7.75 (0.03)
39	<i>p</i> -methoxybenzyl	6.80	7.73	7.73	7.78	7.75 (0.03)
40	<i>m</i> -nitrobenzyl	8.55	8.56	8.56	8.54	8.55 (0.01)
41	<i>p</i> -(hydroxymethyl)benzyl	9.47	9.33	9.33	9.40	9.35 (0.04)
42	<i>m</i> -(hydroxymethyl)benzyl	9.85	9.33 ^c	9.33 ^c	9.40 ^c	9.35 (0.04) ^c
43	<i>p</i> -hydroxybenzyl	9.92	9.89	9.89	9.90	9.90 (0.02)
44	<i>m</i> -hydroxybenzyl	9.92	9.89	9.89	9.93	9.90 (0.02)
45	<i>m</i> -aminobenzyl	9.55	9.75	9.75	9.72	9.74 (0.02)

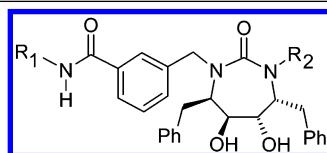


no.	R	Log(1/ K_i)				
		exp.	I(AB, 2–4)	I(AB, 3–4)	I(AB, 4–4)	mean ^b
46	<i>n</i> -propyl	8.96	8.69	8.75	8.42	8.62 (0.18)
47	<i>n</i> -butyl	9.22	8.65	8.67	8.38	8.57 (0.16)
48	allyl	8.85	8.80	8.80	8.50	8.70 (0.17)
49	cyclopropylmethyl	8.82	8.89	8.81	8.65	8.78 (0.12)
50	cyclopentylmethyl	9.55	8.86	8.82	8.57	8.75 (0.16)
51	benzyl	8.64	8.92	8.92	8.85	8.90 (0.04)
52	3-picolyl	8.28	8.25	8.25	8.21	8.24 (0.02)
53	4-picolyl	8.16	8.43	8.44	8.28	8.38 (0.09)
54	<i>p</i> -fluorobenzyl	8.44	8.93	8.93	8.77	8.88 (0.09)
55	<i>p</i> -(hydroxymethyl)benzyl	9.03	9.31	9.32	9.17	9.27 (0.08)
56	<i>m</i> -aminobenzyl	9.00	9.52	9.53	9.33	9.46 (0.11)
57	<i>m</i> -hydroxylbenzyl	9.48	9.59	9.60	9.44	9.54 (0.09)

Table 7. Continued



no.	R	Log(1/ K_i)				
		exp.	I(AB, 2-4)	I(AB, 3-4)	I(AB, 4-4)	mean ^b
58	H	10.41	10.41	10.41	10.41	10.41 (0.00)
59	CH ₃	10.18	9.92	9.98	8.80	9.57 (0.66)
60	CH ₂ CH ₃	9.68	9.47	9.44	9.04	9.31 (0.24)
61	CH(CH ₃) ₂	9.24	9.05	9.02	9.28	9.11 (0.14)
62	CH ₂ CH ₂ CH ₃	9.44	9.44	9.55	9.51	9.50 (0.05)
63	CH ₂ CH ₂ CH ₂ CH ₃	9.37	9.35	9.39	9.43	9.39 (0.04)
64	C(CH ₃) ₃	8.62	8.67	8.71	9.51	8.97 (0.47)
65	CH ₂ -C ₆ H ₅	9.13	9.84	9.66	9.98	9.83 (0.16)
66	phenyl	9.37	9.86	9.86	9.95	9.89 (0.05)
67	4-pyridinyl	9.39	8.90	8.90	8.81	8.87 (0.05)
68	3-pyridinyl	9.54	9.54	9.54	9.54	9.54 (0.00)
69	2-pyridinyl	10.37	10.79	10.79	11.04	10.87 (0.14)
70	2-(4-CH ₃ -pyridinyl)	10.57	10.43	10.43	10.28	10.38 (0.08)
71	2-(5-CH ₃ -pyridinyl)	10.96	10.72	10.72	10.66	10.70 (0.03)
72	2-(6-CH ₃ -pyridinyl)	10.70	11.17 ^c	11.17 ^c	11.47 ^c	11.27 (0.17) ^c
73	2-(4,6-di-CH ₃ -pyridinyl)	10.80	10.81	10.81	10.72	10.78 (0.05)
74	2-(4-CH ₃ -pyrimidinyl)	9.94	10.07	10.07	10.10	10.08 (0.02)
75	2-(5-CF ₃ -pyridinyl)	10.07	10.04	10.04	10.11	10.06 (0.04)
76	2-pirazinyl	10.74	11.17	11.17	11.18	11.17 (0.01)
77	2-pyrimidinyl	9.82	9.69	9.69	9.66	9.68 (0.02)
78	2-imidazolyl	10.85	11.10	11.10	11.10	11.10 (0.00)
79	2-benzimidazolyl	10.62	10.80	10.80	10.81	10.80 (0.00)



no.	R_1	R_2	Log(1/ K_i)				
			exp.	I(AB, 2–4)	I(AB, 3–4)	I(AB, 4–4)	mean ^b
80	2-pyrazinyl	3,5-dimethoxybenzyl	8.60	9.05	9.05	9.00	9.03 (0.03)
81	2-(5-CH ₃ -pyridinyl)	3,5-dimethoxybenzyl	8.72	9.15	9.15	8.99	9.10 (0.09)
82	2-(6-CH ₃ -pyridinyl)	3,5-dimethoxybenzyl	9.15	9.38	9.38	9.40	9.39 (0.01)
83	2-pyridinyl	3,5-dimethoxybenzyl	9.07	9.19	9.19	9.18	9.19 (0.01)
84	2-pyrazinyl	3-methoxybenzyl	10.42	9.45	9.45	9.48	9.46 (0.02)
85	2-(5-CH ₃ -pyridinyl)	3-methoxybenzyl	10.16	9.55 ^c	9.55 ^c	9.48 ^c	9.53 (0.04) ^c
86	2-(6-CH ₃ -pyridinyl)	3-methoxybenzyl	10.28	9.78	9.78	9.89	9.82 (0.06)
87	2-pyridinyl	3-methoxybenzyl	10.33	9.59	9.59	9.67	9.62 (0.04)
88	2-(5-CH ₃ -pyridinyl)	3-aminobenzyl	10.12	10.56	10.56	10.45	10.52 (0.06)
89	2-pyridinyl	3-nitrobenzyl	10.02	10.00	10.00	10.05	10.02 (0.02)
90	t-Bu	3-aminobenzyl	9.39	9.21	9.23	9.62	9.35 (0.23)
91	2-pyrazinyl	3-aminobenzyl	10.80	10.46	10.46	10.45	10.46 (0.00)
92	2-benzimidazolyl	3-aminobenzyl	10.64	10.28	10.28	10.26	10.27 (0.01)
93	2-imidazolyl	3-aminobenzyl	10.92	10.43	10.42	10.41	10.42 (0.01)

^a Structures 1-57,⁶⁵ 58-79,⁶³ and 80-93⁶⁴ and experimental HIV-1 protease inhibitory constants K_i (mol/L) for cyclic ureas were selected from the references.⁶³⁻⁶⁵ The best fit models I(AB, 2-4), I(AB, 3-4), and I(AB, 4-4)/eq 1 obtained for *Training Set 1* were used in calculations.

^b Average log(1/ K_i) value calculated from the three best fit models; the standard deviation is given in parentheses. ^c "Predicted" log(1/ K_i) value from *Test Set 1*.

R₄ = Me, Et, *i*-Pr (Table 10) were potential anti-HIV actives. Since there is still no experimental proof of the efficiency of hypothetical compounds selected from the combinatorial library (Table 10), we cannot claim that this work results in new actives. There are, however, some indications that the results obtained here look reasonable.

The parent set of the HEPT derivatives from which compounds with unique fragments have been excluded (Table 4) contains 84 molecules, 22 of which have X = S (*S* subset) and the others 62 compounds have X = O (*O*

subset). With respect to the substituent R₂, the *O* subset can be divided by three groups: that of 47 compounds containing a S-Ph fragment, the second one of 13 compounds containing a CH₂Ph fragment, and the third group of two compounds with S-Bu and S-cyclohexane fragments. Analysis of the largest sub-group containing S-Ph fragment and X = O allows one to select the "best" substituents R₁, R₃, and R₄ (Figure 2) which correspond to large log(1/EC₅₀) values. Notice, that for that sub-group, R₂ in Table 4 is equal to S-(3, 5-di-R₄-C₆H₃) in Figure 2.

Table 8. Modeling of Anti-HIV Activity ($\log(1/K_i)$) for Cyclic Ureas: Statistical Criteria for Three Best Models Selected at the Training Stage for Sets 1–3^a

no.	fragment set	<i>N</i>	<i>k</i>	<i>R</i> ²	<i>F</i>	FIT	<i>s</i>	<i>R</i> _H , %	<i>Q</i> ²
<i>Training Set 1^b</i>									
1	I(AB, 2–4)	84	56	0.895	4.30	0.08	0.70	4.5	0.540
2	I(AB, 3–4)	84	49	0.891	5.95	0.12	0.64	4.5	0.573
3	I(AB, 4–4)	84	35	0.817	6.42	0.18	0.70	5.9	0.352
<i>Training Set 2^b</i>									
1	I(AB, 2–4)	84	56	0.895	4.36	0.08	0.70	4.5	0.542
2	I(AB, 3–4)	84	49	0.895	6.24	0.12	0.63	4.5	0.567
3	I(AB, 4–4)	84	35	0.819	6.50	0.18	0.70	6.0	0.342
<i>Training Set 3^b</i>									
1	I(AB, 2–4)	84	56	0.885	3.96	0.07	0.76	4.9	0.461
2	I(AB, 3–4)	84	49	0.884	5.55	0.11	0.68	4.9	0.501
3	I(AB, 4–4)	84	35	0.817	6.48	0.18	0.72	6.2	0.288

^a Molecules are represented without hydrogen atoms. Fitting eq 1 was used in all cases. See Table 2 footnotes for definition of statistical parameters.^b The compounds in the training and test sets are specified in Table 7.**Table 9.** Experimental and “Predicted” Activities ($\log(1/K_i)$) for Cyclic Ureas from the Test Sets 1–3^a

compound no.	experimental	predicted			mean ^b
		I(AB, 2–4)	I(AB, 3–4)	I(AB, 4–4)	
Test Set 1					
5	8.80	7.91	7.88	7.73	7.84 (0.10)
7	6.59	7.73	7.56	7.56	7.62 (0.10)
9	5.96	6.63	5.97	5.05	5.88 (0.79)
12	7.92	7.95	8.00	7.73	7.89 (0.14)
13	8.15	7.86	7.83	7.65	7.78 (0.11)
16	7.44	7.95	8.07	7.65	7.89 (0.22)
42	9.85	9.33	9.33	9.40	9.35 (0.04)
72	10.70	11.17	11.17	11.47	11.27 (0.17)
85	10.16	9.55	9.55	9.48	9.53 (0.04)
R^2_c s^c		0.832	0.850	0.835	0.845
		0.59	0.61	0.77	0.64
Test Set 2					
8	6.10	6.70	6.69	7.20	6.86 (0.29)
11	7.31	8.77	8.78	8.73	8.76 (0.03)
15	6.96	7.33	7.33	7.41	7.36 (0.05)
19	8.68	8.69	8.68	8.73	8.70 (0.03)
21	8.37	8.48	8.48	8.43	8.46 (0.03)
56	9.00	9.59	9.59	9.34	9.51 (0.14)
70	10.57	10.26	10.26	10.01	10.18 (0.14)
82	9.15	9.37	9.37	9.38	9.37 (0.01)
87	10.33	9.49	9.49	9.58	9.52 (0.05)
R^2_c s^c		0.843	0.841	0.869	0.852
		0.48	0.48	0.37	0.44
Test Set 3					
14	7.52	7.65	7.57	7.53	7.58 (0.06)
28	8.52	8.59	8.59	8.68	8.62 (0.05)
37	7.29	7.64	7.64	7.57	7.62 (0.04)
48	8.85	8.80	8.80	8.43	8.68 (0.21)
51	8.64	8.97	8.98	8.88	8.94 (0.06)
60	9.68	9.41	9.41	8.89	9.24 (0.30)
62	9.44	9.45	9.57	9.40	9.47 (0.09)
69	10.37	10.89	10.89	11.20	10.99 (0.18)
71	10.96	10.67	10.67	10.65	10.66 (0.01)
R^2_c s^c		0.949	0.949	0.861	0.925
		0.28	0.28	0.50	0.34

^a Compound numbers are given according to Table 7. Fitting eq 1 was used. ^b Average $\log(1/K_i)$ value calculated from the three best fit models; the standard deviation is given in parentheses. ^c Statistical characteristics (*R* and *s*) for the correlation between experimental and predicted (by given model) activities.

Experimental data (Table 4) show that for the compounds with similar *R*₃ and *R*₄, *R*₁ = *i*-Pr is the most efficient substituent. Indeed, for *R*₄ = H and *R*₃ = CH₂OC₂H₄OH, the $\log(1/EC_{50})$ activity varies as a function of *R*₁ as follows: 5.15 (*R*₁ = Me, no. 67 in Table 4), 6.92 (Et, no. 26), 5.47 (*n*-Pr, no. 27), 7.20 (*i*-Pr, no. 28), 5.96 (CH=CH₂,

no. 6), and 5.22 (*cis*-CH=CH–Ph, no. 5). For *R*₄ = H and *R*₃ = CH₂OCH₂Ph, the $\log(1/EC_{50})$ activities are 7.06, 8.23, and 8.57 for *R*₁ = Me (no. 35), Et (no. 52), and *i*-Pr (no. 54), respectively.

Taking similar *R*₁ and *R*₃, one can see that a substitution of the hydrogen atoms by Me groups in the 3,5 position of

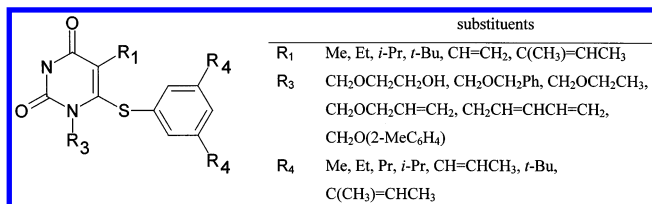


Figure 2. Molecular core and the set of “optimal” substituents for which the combinatorial library of virtual 252 HEPT derivatives was generated.

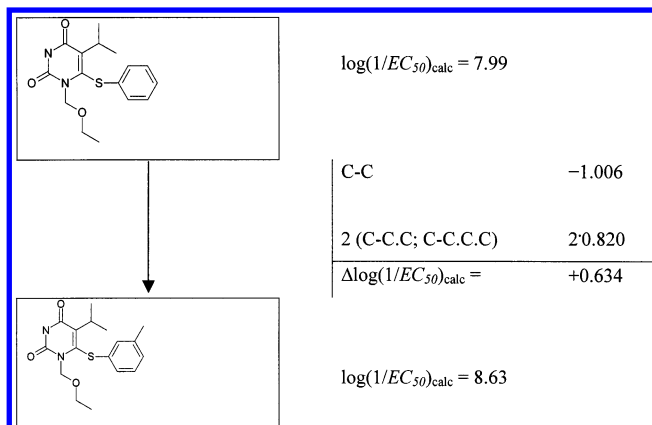


Figure 3. Preparation of the “optimal” substituents for hypothetical HEPT derivatives using the fragment contributions calculated for the I(AB, 2–4)/eq 1 model at the training stage.

the phenylthio group (i.e., R₄ = H by R₄ = Me, Figure 2) always leads to the increase of the activity. Thus, $\log(1/EC_{50})$ increases from 5.15 to 6.59 (R₁ = Me; R₃ = CH₂OC₂H₄OH (nos. 67 and 15, respectively) and from 6.92 to 7.89 (R₁ = Et; R₃ = CH₂OC₂H₄OH (nos. 26 and 29, respectively).

When R₁ = Et and R₄ = Me, the activity increases in the order 7.89, 8.27, and 8.49 for R₃ = CH₂OC₂H₄OH (no. 29), CH₂OC₂H₅ (no. 74), and CH₂OCH₂Ph (no. 75). This means that the latter is the most efficient R₃ substituent.

It follows from the above analysis of the experimental data that the compound with the “best” substituents R₁ = *i*-Pr, R₃ = CH₂OCH₂Ph, and R₄ = Me should display the highest activity. We have not found in the literature any information concerning synthesis and properties of that compound. Its calculated activity is 9.44 $\log(1/EC_{50})$ units (Table 10) which is larger than that for the most active compound with phenylthio group studied experimentally ($\log(1/EC_{50}) = 8.57$, no. 54).

Evaluation of $\log(1/EC_{50})$ for the generated virtual compounds suggests that R₁ = *t*-Bu and R₄ = Et, *i*-Pr substituents could be more efficient than those selected from the experimental data, R₁ = *i*-Pr and R₄ = Me. How reasonable are these suggestions?

The CoMFA calculations on HEPT derivatives^{12,68} as well as crystal structure studies of RT enzyme complexes with HEPT derivatives⁶⁹ reveal the role of the R₁ substituent: the moderately sized substituent pushes the amino acid Tyr181 into a position which renders the protein nonfunctional. This explains why the compounds with R₁ = *i*-Pr possess larger activities than small Me or Et groups (which have insufficient steric requirements) or big CH₂Ph and C(O)Ph groups (which are too large for the protein's cavity). Anyway, the experimental data reveal only the fact that the *i*-Pr group is the best among other studied substituents (Table 4), but they do

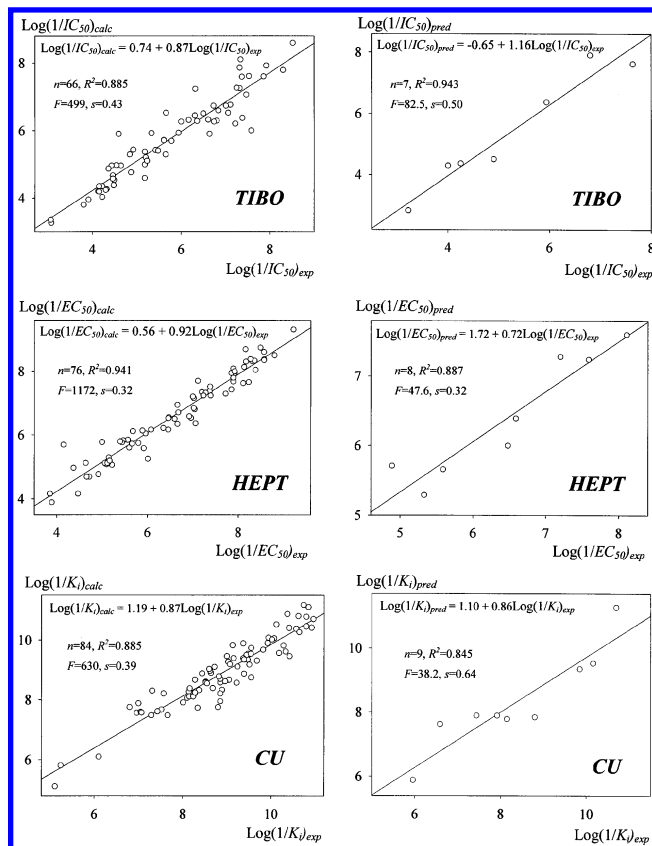


Figure 4. Modeling of anti-HIV activity by the SMF method. Linear correlation between calculated and experimental anti-HIV activities obtained with the averages for three “best fit” models for the training (left) and test (right) sets of TIBO, HEPT, and CU derivatives. Calculations were done for the Training and Test Sets 1 for each family of the compounds (see Section 4.1).

not exclude the possibility for a group of slightly different size to be an efficient substituent. One may assume that the *t*-Bu group, whose size is between that of *i*-Pr and large CH₂-Ph and C(O)Ph groups, could also fit the protein's cavity and efficiently interact with the Tyr181 amino acid.

According to the crystal structure studies by Hopkins et al.,⁶⁹ the phenylthio rings of the HEPT derivatives could favorably interact with the side chain of Tyr181. Although no structural information on the RT enzyme complexes with the compounds containing the substituted phenylthio group is available, the activity measurements show that the replacement of H by Me groups in the 3,5 positions may increase an inhibition of the enzyme (Table 4). One may assume that the larger Et and *i*-Pr groups could also fit the protein's cavity and efficiently interact with Tyr181.

The results obtained here show that the SMF method provides an efficient way of lead optimization due to utilization of fragments contributions calculated at the training stage. As far as the potential HEPT analogues are concerned, a generation of large combinatorial libraries might lead to better candidates than those given in Table 10 for the Markush structure presented in Figure 2. Thus, our calculations show that replacing the phenylthio fragment with the benzyl fragment may lead to the increase of the compounds activities (Table 10). Further experimental studies are needed to check whether our predictions are realistic.

Table 10. Selected Hypothetical HEPT Derivatives for Which High HIV-1 Inhibitory Activities ($\log(1/EC_{50})$) Were Predicted by the SMF Method^a

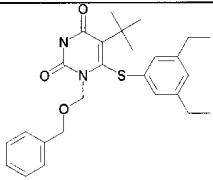
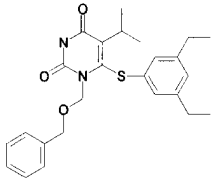
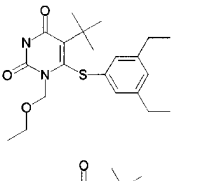
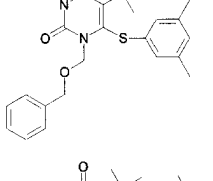
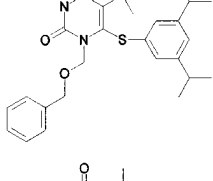
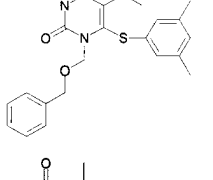
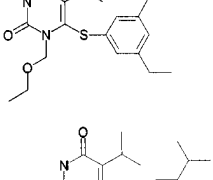
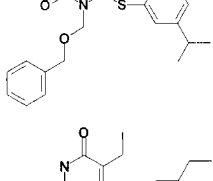
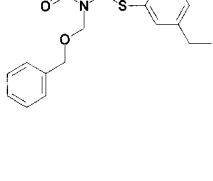
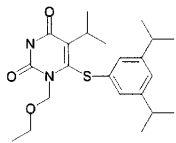
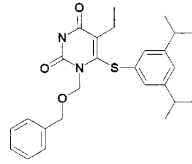
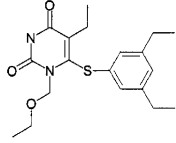
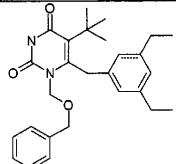
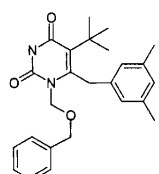
no.	compound	predicted $\log(1/EC_{50})$				
		I(AB, 2-5)	II(Hy)	I(AB, 2-4)	mean ^b	s ^c
1		10.15	<i>d</i>	10.02	10.09	0.09
2		10.59	9.37	9.40	9.79	0.70
3		9.74	<i>d</i>	9.69	9.72	0.04
4		8.79	<i>d</i>	10.22	9.51	1.01
5		9.68	<i>d</i>	9.32	9.50	0.25
6		9.24	9.49	9.60	9.44	0.18
7		10.18	9.03	9.07	9.43	0.65
8		10.13	<i>d</i>	8.70	9.42	1.01
9		10.12	8.62	8.53	9.09	0.89

Table 10 (Continued)

no.	compound	predicted $\log(1/EC_{50})$				
		I(AB, 2-5)	II(Hy)	I(AB, 2-4)	mean ^b	s ^c
10		9.72	^d	8.37	9.05	0.95
11		9.66	^d	7.82	8.74	1.30
12		9.72	8.28	8.19	8.73	0.86
13		11.00	^d	10.10	10.55	0.64
14		9.65	^d	10.29	9.97	0.45

^a The three best models obtained for *Training Set 1* were used in the calculations. ^b Average for three best fit models $\log(1/EC_{50})$ value. ^c Standard deviation for average activities. ^d Fragment contribution(s) for this compound was not available on the training stage.

6. CONCLUSION

The substructural molecular fragments (SMF) method²⁷ incorporated into the TRAIL program was applied to model different types of anti-HIV activity ($\log(1/IC_{50})$, $\log(1/EC_{50})$, and $\log(1/K_i)$) for 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives, tetrahydroimidazobenzodiazepinone (TIBO) derivatives (HIV-1 reverse transcriptase inhibitors), and cyclic urea (CU) derivatives (HIV-1 protease inhibitors). As variables in a multiple regression analysis, the SMF method uses molecular fragments (atom/bond sequences and augmented atoms). The TRAIL program generates 49 different types of fragments and uses them together with three linear and nonlinear fitting equations, thus allowing the user to build up to 147 QSAR models. A significant advantage of the SMF method is the possibility to select during the training stage several best fit models (instead of a single QSPR model) related to different fragmentation schemes in combination with three fitting equations. Using selected QSAR models, one can calculate average activities of the compounds from the test set, which smoothes inaccuracies of particular models, thus improving the robustness of predictions.²⁹

During the training stage, TRAIL has selected three best fit models for each family of studied anti-HIV compounds

and then applied them for test calculations. Calculations show that the sets of selected fragmental descriptors vary as a function of the class of compounds. Thus, three selected QSAR models for TIBO compounds involve atom/bond sequences containing from 2 to 4, from 3 to 4, and exactly 5 atoms, for CU compounds these are atom/bond sequences from 2 to 4, from 3 to 4, and exactly 4 atoms, whereas for HEPT compounds, atom/bond sequences from 2 to 4, from 2 to 5, and augmented atoms (with hybridization) were found the most pertinent descriptors. Calculated for the best models, "average" data sets reproduce available experimental data at least as good as those from earlier QSAR studies.

It has been demonstrated that the SMF method gives an interesting opportunity to build the "optimal" substituents whose attachment to the molecular core could result in increased activity. Based on such "optimal" substituents, the focused combinatorial library containing 252 virtual HEPT derivatives has been generated. Its filtering led to several hits potentially possessing activities larger than that of the "best" experimentally studied compounds.

ACKNOWLEDGMENT

V.P.S. gratefully acknowledges A.V. for the opportunity to have spent a term in 2001 as an Invited Professor at the

ULP, Strasbourg. We thank C. Schwab for help with the data preparation and Dr. R. Stote for linguistic help.

REFERENCES AND NOTES

- (1) *SciFinder Scholar, version 2000.1*. American Chemical Society: Washington, DC, 2000; <http://www.cas.org/SCIFINDER/SCHOLAR/index.html>.
- (2) Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S.; Debnath, A. K.; Hansch, C. Comparative Quantitative Structure–Activity Relationship Studies on Anti-HIV Drugs. *Chem. Rev.* **1999**, *99*, 3525–3602.
- (3) Jalali-Heravi, M.; Parastar, F. Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147–154.
- (4) Knaggs, M. H.; McGuigan, C.; Harris, S. A.; Heshmati, P.; Cahard, D.; Gilbert, I. H.; Balzarini, J. A QSAR study investigating the effect of L-alanine ester variation on the anti-HIV activity of some phosphoramidate derivatives of d4T. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2075–2078.
- (5) Tronchet, J. M. J.; Grigorov, M.; Dolatshahi, N.; Moriaud, F.; Weber, J. A QSAR study confirming the heterogeneity of the HEPT derivative series regarding their interaction with HIV reverse transcriptase. *Eur. J. Med. Chem.* **1997**, *32*, 279–299.
- (6) Huuskonen, J. QSAR modeling with the electrotopological state: TIBO derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425–429.
- (7) Maw, H. H.; Hall, L. H. E-State Modeling of HIV-1 Protease Inhibitor Binding Independent of 3D Information. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 290–298.
- (8) Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR methods: MS-WHIM and autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 293–306.
- (9) Klein, C. T.; Lawtrakul, L.; Hannongbua, S.; Wolschann, P. Accessible charges in structure–activity relationships. A study on HEPT-based HIV-1 RT inhibitors. *Sci. Pharm.* **2000**, *68*, 25–40.
- (10) Santos-Filho, O. A.; Hopfinger, A. J. The 4D-QSAR paradigm: application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct.–Act. Relat.* **2002**, *21*, 369–381.
- (11) Kireev, D. B.; Chretien, J. R.; Raevsky, O. A. Molecular modeling and quantitative structure–activity studies of anti-HIV-1,2-heteroarylquinoline-4-amines. *Eur. J. Med. Chem.* **1995**, *30*, 395–402.
- (12) Hannongbua, S.; Nivesanon, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative Structure–Activity Relationships of HEPT Derivatives as HIV-1 Reverse Transcriptase Inhibitors, Based on Ab Initio Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848–855.
- (13) Jayatilake, P. R. N.; Nair, A. C.; Zauhar, R.; Welsh, W. J. Computational Studies on HIV-1 Protease Inhibitors: Influence of Calculated Inhibitor–Enzyme Binding Affinities on the Statistical Quality of 3D-QSAR CoMFA Models. *J. Med. Chem.* **2000**, *43*, 4446–4451.
- (14) Debnath, A. K. Three-Dimensional Quantitative Structure–Activity Relationship Study on Cyclic Urea Derivatives as HIV-1 Protease Inhibitors: Application of Comparative Molecular Field Analysis. *J. Med. Chem.* **1999**, *42*, 249–259.
- (15) Buolamwini, J. K.; Assefa, H. CoMFA and CoMSIA 3D QSAR and Docking Studies on Conformationally-Restrained Cinnamoyl HIV-1 Integrase Inhibitors: Exploration of a Binding Mode at the Active Site. *J. Med. Chem.* **2002**, *45*, 841–852.
- (16) Mickle, T.; Nair, V. Anti-human immunodeficiency virus activities of nucleosides and nucleotides: correlation with molecular electrostatic potential data. *Antimicrob. Agents Chemother.* **2000**, *44*, 2939–2947.
- (17) Mickle, T.; Nair, V. Predictive QSAR analysis of anti-HIV agents. *Drugs Future* **2000**, *25*, 393–400.
- (18) De-Clercq, E. Toward improved anti-HIV chemotherapy: therapeutic strategies for intervention with HIV infections. *J. Med. Chem.* **1995**, *38*, 2491–2517.
- (19) Debnath, A. K. Comparative Molecular Field Analysis (CoMFA) of a Series of Symmetrical Bis-Benzamide Cyclic Urea Derivatives as HIV-1 Protease Inhibitors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 761–767.
- (20) Wilkerson, W. W. Anti-HIV activity of the C2-symmetric cyclic urea carboxamides: A QSAR study. *Book of Abstracts, 211th ACS National Meeting, New Orleans, LA, March 1996*; American Chemical Society: Washington, DC, 1996; MEDI-012.
- (21) Garg, R.; Kurup, A.; Gupta, S. P. Quantitative structure–activity relationship studies on some acyclovir derivatives acting as anti-HIV-1 drugs. *Quant. Struct. Act. Relat.* **1997**, *16*, 20–24.
- (22) Garg, R.; Hansch, C. Comparative QSAR studies on anti-HIV HEPT derivatives. *Book of Abstracts, 217th ACS National Meeting, Anaheim, CA, March 1999*; American Chemical Society: Washington, DC, 1999; COMP-110.
- (23) Luco, J. M.; Ferretti, F. H. QSAR Based on Multiple Linear Regression and PLS Methods for the Anti-HIV Activity of a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392–401.
- (24) Gupta, S. P.; Babu, M. S.; Kaw, N. Quantitative structure–activity relationships of some HIV-protease inhibitors. *J. Enzyme Inhib.* **1999**, *14*, 109–123.
- (25) Gupta, S. P.; Garg, R. Quantitative structure–activity relationship studies on anti-HIV-1 TIBO derivatives as inhibitors of viral reverse transcriptase. *J. Enzyme Inhib.* **1996**, *11*, 23–32.
- (26) Hannongbua, S.; Pungpo, P.; Limtrakul, J.; Wolschann, P. Quantitative structure–activity relationships and comparative molecular field analysis of TIBO derivatised HIV-1 reverse transcriptase inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 563–577.
- (27) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.
- (28) Varnek, A.; Wipff, G.; Solov'ev, V. P. Towards an Information System on Solvent Extraction. *Solvent Extr. Ion Exch.* **2001**, *19*, 797–837.
- (29) Varnek, A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F. Assessment of the macrocyclic effect for the complexation of crown-ethers with alkali cations using the substructural molecular fragments method. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 812–829.
- (30) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (31) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 144–154.
- (32) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure–Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
- (33) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (34) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand–Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (35) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure–Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
- (36) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249–258.
- (37) Klopman, G.; Tu, M. Diversity analysis of 14 156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure–activity relational expert system MCASE. *J. Med. Chem.* **1999**, *42*, 992–998.
- (38) Zefirov, N. S.; Palyulin, V. A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112–1122.
- (39) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432–2437.
- (40) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.
- (41) Avidon, V. V. The Criteria of Chemical Structures Similarity and the Principles for Design of Description Language for Chemical Information Processing of Biologically Active Compounds. *Chim. Pharm. J. (Russ.)* **1974**, *8*, 22–25.
- (42) Bawden, D. Computerized Chemical Structure-Handling Techniques in Structure–Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14–22.
- (43) Hansch, C.; Leo, A.; Hoekman, D. H. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1995; 580.
- (44) Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos, A. Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349–1355.
- (45) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
- (46) Pascual, R.; Mateu, M.; Gasteiger, J.; Borrell, J. I.; Teixido, J. Design and Analysis of a Combinatorial Library of HEPT Analogues: Comparison of Selection Methodologies and Inspection of the Actually Covered Chemical Space. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 199–207.
- (47) Korn, G. A.; Korn, T. M. *Mathematical Handbook for Scientists and Engineers*, 2nd ed.; McGraw-Hill Book Company: New York 1968.

- (48) Forsythe, G. E.; Malcolm, M. A.; Moler, C. B. *Computer Methods for Mathematical Computations*; Prentice Hall, Inc.: Englewood Cliffs, NJ, 1977; 259.
- (49) Kendall, M. G.; Stuart, A. *The Advanced Theory of Statistics*; Griffin: London, 1966.
- (50) Barnard, J. M.; Downs, J. M.; von-Scholley-Pfab, A.; Brown, R. D. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 452–463.
- (51) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure–activity relationship of human immunodeficiency virus (I) protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37*, 2206–2215.
- (52) Kukla, M. J.; Breslin, H. J.; Pauwels, R.; Fedde, C. L.; Miranda, M.; Scott, M. K.; Sherrill, R. G.; Raeymaekers, A.; van Gelder, J.; Andries, K.; et al. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo[4,5,1-jk][1,4]benzodiazepin-2(1H)-one (TIBO) derivatives. *J. Med. Chem.* **1991**, *34*, 746–751.
- (53) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; et al. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo-[4,5,1-jk][1,4]benzodiazepin-2(1H)-one (TIBO) derivatives. 4. *J. Med. Chem.* **1995**, *38*, 794–802.
- (54) Breslin, H. J.; Kukla, M. J.; Ludovici, D. W.; Mohrbacher, R.; Ho, W.; Miranda, M.; Rodgers, J. D.; Hitchens, T. K.; Leo, G.; et al. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo [4,5,1-jk][1,4]benzodiazepin-2(1H)-one (TIBO) derivatives. 3. *J. Med. Chem.* **1995**, *38*, 771–793.
- (55) 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.
- (56) Tanaka, H.; Baba, M.; Hayakawa, H.; Haraguchi, K.; Miyasaka, T.; et al. Lithiation of uracil nucleosides and its application to the synthesis of a new class of anti-HIV-1 acyclonucleosides. *Nucleosides Nucleotides* **1991**, *10*, 397–400.
- (57) Tanaka, H.; Baba, M.; Saito, S.; Miyasaka, T.; Takashima, H.; et al. Specific anti-HIV-1 acyclonucleosides which cannot be phosphorylated: synthesis of some deoxy analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1991**, *34*, 1508–1511.
- (58) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; et al. Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 1394–1399.
- (59) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; et al. A New Class of HIV-1 Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity of 5- or 6-Substituted Analogs of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349–357.
- (60) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; et al. 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.
- (61) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; et al. Structure–activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine analogs: effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity. *J. Med. Chem.* **1992**, *35*, 337–345.
- (62) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; et al. Synthesis and Antiviral Activity of 6-Benzyl Analogs of 1-[(2-Hydroxyethoxy)methyl]-5-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1995**, *38*, 2860–2865.
- (63) Wilkerson, W. W.; Akamike, E.; Cheatham, W. W.; Hollis, A. Y.; Collins, R. D.; et al. HIV Protease Inhibitory Bis-benzamide Cyclic Ureas: A Quantitative Structure–Activity Relationship Analysis. *J. Med. Chem.* **1996**, *39*, 4299–4312.
- (64) Wilkerson, W. W.; Dax, S.; Cheatham, W. W. Nonsymmetrically Substituted Cyclic Urea HIV Protease Inhibitors. *J. Med. Chem.* **1997**, *40*, 4079–4088.
- (65) Lam, P. Y.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLucca, G. V.; et al. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2' structure–activity relationship, and molecular recognition of cyclic ureas. *J. Med. Chem.* **1996**, *39*, 3514–3525.
- (66) Klopman, G.; Ding, C.; Macina, O. T. Computer Aided Olive Oil–Gas Partition Coefficient Calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 569–575.
- (67) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*; McGraw-Hill Book Co.: New York, 1977.
- (68) Pungpo, P.; Wolschann, P.; Hannongbua, S. Quantitative structure–activity relationships of HIV-1 reverse transcriptase inhibitors, using hologram QSAR. *Rational Approaches to Drug Design, Proceedings of the European Symposium on Quantitative Structure Activity Relationships, 13th, Duesseldorf, Germany, Aug 2001*; p 206–210.
- (69) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; et al. Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent nonnucleoside inhibitors. *J. Med. Chem.* **1996**, *39*, 1589–1600.

CI020388C