Eur J Med Chem (1997) 32, 279–299

A QSAR study confirming the heterogeneity of the HEPT derivative series regarding their interaction with HIV reverse transcriptase

JMJ Tronchet[1]*, M Grigorov[1,2], N Dolatshahi[1], F Moriaud[1], J Weber[2]

[1]*Department of Organic Pharmaceutical Chemistry, Sciences II;*
[2]*Department of Physical Chemistry, Sciences II, University of Geneva, CH-1211 Geneva 4, Switzerland*

(Received 29 April 1996; accepted 30 September 1996)

Summary — QSAR concerning the anti-HIV and cytotoxic activities of a series of HEPT analogues has been established using a Hansch-type approach (TSAR™), a neural network approach (TSAR) and a pharmacophore search method (CATALYST™). The techniques employed allowed reliable activity predictions and confirmed the heterogeneity of this series of compounds, which was previously established in biochemical experiments.

HEPT / HIV / TSAR™ / CATALYST™ / reverse transcriptase / neural network

Introduction

The reverse transcriptase (RT) plays a central role in the replication of HIV. A number of reverse transcriptase inhibitors active either against both HIV-1 and HIV-2 RT or only against HIV-1 RT have been described. Among the representatives of the latter type, the 1-[(2-hydroxyethoxy)methyl]-6-phenylthiothymine, known as the HEPT derivatives, first described in 1989 [1, 2], constitute an important series and have been considerably developed since that time [3–8]. In order to plan future synthetic work in this series, we submitted its already known representatives [3, 4] to a QSAR study to determine the salient chemical features of these compounds responsible for their biological activity.

Materials and methods

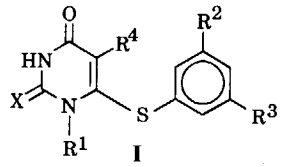
The study encompasses all the compounds described in the two earlier papers by Tanaka et al [3, 4], for which either EC_{50} or CC_{50} values (or most generally both), expressed in μM throughout, were known. The structures of these compounds are presented in table I. The 2D-QSAR results have been obtained using the TSAR™ [9] software. The core structure (**I**, $R^1 = R^2 = R^3 = R^4 = H$) of these molecules and each R^1 substituent have been generated and minimized with the PIMMS™ [10] software integrating the COSMIC™ force field until obtention of energy self consistency and energy gradient of 10^{-10} kcal•mol^{-1} and 10^{-5} kcal•mol^{-1}•Å^{-1}, respectively. Among the four substituents, the R^1 group is, in most cases, the only one to exhibit conformational flexibility. We have generated molecular structures which can be viewed as 'pseudoconformers', by combining the core moiety with the minimized substituents R^2-R^4 and one of the conformers of group R^1. This technique offers the possibility to take into account the effects of the conformational flexibility even at the 2D-QSAR level. For each R^1 group, three to ten different conformations in the 0–10 kcal/mol range have been generated using the COBRA™ [11] software. In figures 1–3 we have presented, in increasing energy order, the generated conformers of group R^1 when this group is a hydroxyethoxymethyl, an ethoxybenzyl and a benzoxymethyl substituent, respectively. For each substituent (each conformer for R^1) 60 electronic, shape and topological descriptors have been introduced in the TSAR structure–activity table. Some of the parameters used come from standard databases [12], the others have been calculated using the TSAR software.

The neural network analysis has been performed using the functionality offered by the TSAR software. Neural networks are parallel distributed computing devices simulating the brain at a first-order approximation level. They are composed of a large number of simple processing elements, called neurons, interconnected to form a highly parallel network. In such a network, the neurons can be viewed as being grouped in three layers. The input layer consists of neurons processing the incoming information flow, while the output layer consists of neurons providing the output flow. Between these two layers, other neurons are grouped in one or more hidden layers. A detailed discussion on the essential components of neural networks can be found elsewhere [13].

Within the TSAR software, we have used for QSAR purposes the multiple-layer feed forward neural network topology which undergoes a supervised training by back propagation of errors [13]. The input for such a neural network was

*Correspondence and reprints

Table I. Numerical data concerning in vitro anti-HIV-1 activities (EC_{50}/μM) of HEPT analogues: experimental and predicted values, obtained by TSAR and CATALYST analysis.



I

Compound	X	$R^1 = CH_2OR$ R	R^2	R^3	R^4	R^1 Conformer[a]		Experiment[b] EC_{50}	Experiment[b] $-LogEC_{50}$	TSAR $-LogEC_{50}$ fit[c]		TSAR Residuals fit[d]		TSAR Class[e]	TSAR $-LogEC_{50}$ fit[f]		CATALYST R^1 Conformer[g]	CATALYST EC_{50} fit[h]	CATALYST Error fit[i]	CATALYST Class[j]	CATALYST EC_{50} fit[k]
						Low	High			Low	High	Low	High		Low	High					
1	O	CH_2CH_2OH	Me	H	Me	(1a)	(1i)	2.6	−0.415	−0.386	−0.174	−0.029	−0.241	C	−0.474	−0.395	(1c)	6.50	+2.5	C	7.1
2	O	CH_2CH_2OH	Et	H	Me	(1a)	(1i)	2.7	−0.431	−0.308	−0.091	−0.123	−0.340	C	−0.329	−0.248	(1b)	0.86	−2.9	C	5.8
3	O	CH_2CH_2OH	*i*–Bu	H	Me	(1a)	(1i)	12.0	−1.079	−1.293	−1.087	−0.214	0.007	C	−1.118	−1.040	(1b)	11.0	−1.1	C	3.0
4	O	CH_2CH_2OH	CH_2OH	H	Me	(1a)	(1i)	>292	—	−0.903	−0.743	—	—	C	−0.958	−0.942	(1c)	5.3	—	C	6.4
5	O	CH_2CH_2OH	CF_3	H	Me	(1a)	(1i)	45.0	−1.653	−1.648	−1.437	−0.005	−0.216	C	−1.697	−1.609	(1a)	6.40	−7.0	A	10.0
6	O	CH_2CH_2OH	F	H	Me	(1a)	(1i)	3.3	−0.519	−1.311	−1.092	0.793	0.574	B	−0.531	0.506	(1i)	7.90	+2.4	C	8.5
7	O	CH_2CH_2OH	Cl	H	Me	(1a)	(1i)	13.0	−1.114	−0.915	−0.696	−0.199	−0.418	C	−1.034	−0.952	(1d)	4.50	−2.9	C	7.6
8	O	CH_2CH_2OH	Br	H	Me	(1a)	(1i)	5.7	−0.756	−0.960	−0.740	0.204	−0.016	C	−1.063	−0.981	(1a)	0.30	−19.0	A	4.9
9	O	CH_2CH_2OH	I	H	Me	(1a)	(1i)	10.0	−1.000	−0.875	−0.660	−0.125	−0.340	C	−0.965	−0.884	(1a)	9.60	−1.0	C	13.0
10	O	CH_2CH_2OH	NO_2	H	Me	(1a)	(1i)	34.0	−1.531	−1.483	−1.271	−0.049	−0.261	C	−1.425	−1.346	(1a)	12.0	−2.9	C	30.0
11	O	CH_2CH_2OH	OH	H	Me	(1a)	(1i)	82.0	−1.914	−1.687	−1.409	−0.227	−0.505	A	−1.844	−1.763	(1a)	6.7	−12.0	A	19.0
12	O	CH_2CH_2OH	OMe	H	Me	(1a)	(1i)	22.0	−1.342	−1.466	−1.246	0.123	−0.097	C	−1.520	−1.438	(1e)	6.0	−3.6	A	10.0
13	O	CH_2CH_2OH	Me	Me	Me	(1a)	(1i)	0.26	0.585	1.042	1.253	−0.457	−0.668	A	0.507	0.588	(1a)	0.009	−28.9	A	0.2
14	O	CH_2CH_2OH	Cl	Cl	Me	(1a)	(1i)	1.30	−0.114	0.738	0.962	−0.852	−1.076	A	−0.204	−0.024	(1a)	0.197	−6.6	A	0.25
15	S	CH_2CH_2OH	Me	Me	Me	(1a)	(1i)	0.22	0.658	1.042	1.253	−0.384	−0.596	A	0.507	0.588	(1a)	0.38	+1.7	C	0.27
16	O	CH_2CH_2OH	COOMe	H	Me	(1a)	(1i)	7.90	−0.898	−0.248	−1.029	0.350	0.131	C	−1.011	−0.929	(1f)	18.0	+2.3	C	3.5
17	O	CH_2CH_2OH	COMe	H	Me	(1a)	(1i)	7.30	−0.863	−0.945	−0.731	0.081	−0.132	C	−0.951	−0.872	(1e)	5.10	−1.4	C	3.7
18	O	CH_2CH_2OH	COOH	H	Me	(1a)	(1i)	>352	–	−1.200	−1.115	–	–	C	−1.318	−1.275	(1b)	6.7	–	C	5.3
19	O	CH_2CH_2OH	$COONH_2$	H	Me	(1a)	(1i)	>306	–	−1.124	−1.007	–	–	C	−1.196	−1.130	(1c)	9.5	–	C	10.0
20	O	CH_2CH_2OH	CN	H	Me	(1a)	(1i)	10.0	−1.000	−0.777	−0.566	−0.223	−0.434	C	−1.055	−0.976	(1a)	25.0	+2.5	C	13.0
21	O	CH_2CH_2OH	H	H	Allyl	–	–	2.5	–	–	–	–	–	–	–	–	(1c)	3.4	+1.3	C	2.1
22	O	CH_2CH_2OH	H	H	COOMe	–	–	>6.6	–	–	–	–	–	–	–	–	(1e)	1.7	–	C	0.42
23	O	CH_2CH_2OH	H	H	COONHPh	–	–	>18	–	–	–	–	–	–	–	–	(1a)	2.9	–	C	4.3
24	S	CH_2CH_2OH	H	H	Et	(1a)	(1i)	0.11	0.959	−0.010	0.212	0.969	0.747	B	0.929	0.950	(1g)	0.46	+4.2	B	0.11
25	S	CH_2CH_2OH	H	H	Pr	(1a)	(1i)	10.0	−1.000	0.157	0.378	−1.157	−1.378	A	−0.829	−0.670	(1f)	0.25	−40.0	A	5.0
26	S	CH_2CH_2OH	H	H	*i*–Pr	(1a)	(1i)	0.059	1.229	1.063	1.267	0.166	−0.038	C	1.311	1.388	(1c)	0.034	−1.7	C	0.039
27	S	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	0.008	2.097	1.313	1.524	0.784	0.573	B	2.016	2.048	(1f)	0.008	−1.0	C	0.008
28	S	CH_2CH_2OH	Me	Me	*i*–Pr	(1a)	(1i)	0.005	2.301	1.900	1.701	0.401	0.600	B	2.295	2.375	(1h)	0.005	+8.8	B	0.004
29	S	CH_2CH_2OH	Cl	Cl	Et	(1a)	(1i)	0.043	1.367	1.001	1.228	0.366	0.139	C	1.456	1.540	(1a)	0.019	−2.3	C	0.041
30	O	CH_2CH_2OH	H	H	Et	(1a)	(1i)	0.12	0.921	−0.010	0.212	0.931	0.709	B	0.929	0.950	(1h)	0.088	−1.4	C	0.56
31	O	CH_2CH_2OH	H	H	Pr	(1a)	(1i)	3.4	−0.531	0.157	0.378	−0.689	−0.910	A	−0.829	−0.670	(1a)	0.67	−5.1	A	2.3
32	O	CH_2CH_2OH	H	H	*i*–Pr	(1a)	(1i)	0.063	1.201	1.063	1.267	0.138	−0.066	C	1.311	1.388	(1c)	0.035	−1.8	C	0.040
33	O	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	0.013	1.886	1.313	1.524	0.574	0.362	B	2.016	1.939	(1h)	0.14	+11	B	0.020
34	O	CH_2CH_2OH	Me	Me	*i*–Pr	(1a)	(1i)	0.003	2.523	2.390	2.606	0.133	−0.083	C	2.295	2.375	(1b)	0.015	+5.7	B	0.001
35	O	CH_2CH_2OH	Cl	Cl	Et	(1a)	(1i)	0.014	1.854	1.001	1.228	0.853	0.626	B	1.846	1.862	(1i)	0.069	+4.9	B	0.031
36	O	CH_2CH_2OH	H	H	H	(1a)	(1i)	7.0	−0.845	−0.549	−0.341	−0.296	−0.505	A	−0.457	−0.843	(1b)	8.0	+1.1	C	20.0
37	S	CH_2CH_2OH	H	H	H	(1a)	(1i)	0.98	0.009	−0.449	−0.341	0.457	0.349	B	0.009	−0.379	(1f)	0.49	+2.0	C	0.90
38	O	CH_2CH_2OMe	H	H	Me	2.588	6.479	8.7	−0.940	−0.705	−0.170	−0.235	−0.770	A	−0.740	−0.774	3.729	7.3	−1.2	C	7.4
39	O	CH_2CH_2O–*n*–C_5H_{11}	H	H	Me	3.361	3.974	>55	–	−0.924	−0.899	–	–	C	−0.699	−0.679	3.429	4.5	–	C	5.5
40	O	$CH_2CH_2OCH_2Ph$	H	H	Me	6.169	9.177	>20	–	0.398	0.375	–	–	C	−0.041	−0.033	7.182	3.1	–	C	5.0
41	O	Me	H	H	Me	1.030	–	2.1	−0.322	0.431	–	−0.753	–	A	−0.257	–	1.030	4.8	+2.3	C	6.7

Table I. *Continued.*

42	O	Et	H	H	Me	(2a)	(2c)	0.33	0.481	0.541	0.518	–0.060	–0.037	C	0.520	0.504	(2a)	0.029	–11.0	A	0.25
43	O	Pr	H	H	Me	2.178	9.351	3.6	–0.556	–0.274	–0.013	–0.282	–0.569	A	–0.483	–0.673	3.417	0.41	–8.7	A	2.7
44	O	Bu	H	H	Me	2.401	5.862	4.7	–0.672	–0.755	–0.198	0.083	–0.474	C	–0.793	–0.731	4.810	3.4	–1.4	C	3.3
45	O	$CH_2CH_2SiMe_3$	H	H	Me	–	–	>32	–	–	–	–	–	—	–	–	–	5.2	–	C	1.3
46	O	CH_2Ph	H	H	Me	(3a)	(3e)	0.088	1.056	1.217	1.192	–0.161	–0.137	C	1.032	1.270	(3b)	0.2	+2.3	C	0.075
47	S	Et	H	H	Et	(2a)	(2c)	0.026	1.585	0.817	0.794	0.768	0.791	B	1.677	1.630	(2b)	0.039	+1.5	C	0.030
48	S	Et	Me	Me	Et	(2a)	(2c)	0.004	2.398	2.133	2.113	0.265	0.285	C	2.369	2.344	(2c)	0.021	+4.8	B	0.009
49	S	Et	Cl	Cl	Et	(2a)	(2c)	0.013	1.886	1.821	1.786	0.065	0.100	C	1.966	1.935	(2b)	0.052	+4.0	B	0.009
50	S	$R^1 = CH_2$- *i*-Pr	H	H	*i*–Pr	6.619	9.203	0.22	0.658	0.974	–	–0.316	–	C	0.676	–	7.821	0.11	–2.0	C	0.26
51	S	*c*–Hex	H	H	Et	8.096	9.639	1.6	–0.204	0.177	0.388	–0.381	–0.542	A	0.229	–0.338	8.096	0.11	–14.0	A	0.36
52	S	CH_2-*c*-Hex	H	H	Et	3.963	7.911	0.35	0.456	–0.124	–0.044	0.579	0.499	C	0.457	0.182	3.963	0.045	–7.7	A	0.32
53	S	CH_2Ph	H	H	Et	(3a)	(3e)	0.008	2.097	1.483	1.466	0.614	0.631	B	2.159	2.158	(3c)	0.28	+36	B	0.013
54	S	CH_2Ph	Me	Me	Et	(3a)	(3e)	0.007	2.155	2.813	2.788	–0.658	–0.623	A	2.212	2.166	(3d)	0.096	+14	B	0.007
55	S	$CH_2C_6H_4$(4-Me)H	H	H	Et	5.499	9.288	0.078	1.108	1.542	1.551	–0.434	–0.443	C	1.105	1.097	5.499	0.022	–3.6	A	0.085
56	S	$CH_2C_2H_4$(4-Cl)	H	H	Et	4.492	8.412	0.012	1.921	1.319	1.335	0.602	0.586	B	1.686	1.685	8.412	0.047	+3.8	B	0.011
57	S	CH_2CH_2Ph	H	H	Et	5.362	9.495	0.091	1.041	1.107	0.878	–0.066	0.163	C	1.175	0.870	5.362	0.21	+2.4	C	0.071
58	S	Et	H	H	*i*–Pr	(2a)	(2c)	0.014	1.854	1.868	1.843	–0.014	0.011	C	1.816	1.789	(2a)	0.005	–2.9	C	0.012
59	S	CH_2Ph	H	H	*i*–Pr	(3a)	(3e)	0.007	2.155	2.549	2.524	–0.394	–0.369	C	2.236	2.182	(3c)	0.007	+1.1	C	0.007
60	S	Et	H	H	*c*–Pr	(2a)	(2c)	0.095	1.022	0.459	0.427	0.563	0.595	B	0.433	0.995	(2c)	0.11	+1.2	C	0.081
61	O	Et	H	H	Et	(2a)	(2c)	0.019	1.721	0.817	0.794	0.905	0.927	B	1.677	1.630	(2c)	0.070	+3.7	B	0.026
62	O	Et	Me	Me	Et	(2a)	(2c)	0.005	2.301	2.133	2.113	0.168	0.188	C	2.369	2.095	(2a)	0.010	+2.0	C	0.006
63	O	Et	Cl	Cl	Et	(2a)	(2c)	0.007	2.155	1.821	1.786	0.334	0.369	C	1.966	1.935	(2a)	0.023	+2.9	C	0.005
64	O	*i*–Pr	H	H	Et	1.382	–	0.34	0.469	1.674	–	–1.205	–	A	0.470	–	1.382	0.051	–6.7	A	0.54
65	O	*c*–Hex	H	H	Et	8.096	9.639	4.0	–0.602	0.228	–0.346	–0.831	–0.256	A	–0.485	–0.425	8.096	0.67	–6.0	A	2.30
66	O	CH_2-*c*-Hex	H	H	Et	3.963	7.911	0.45	0.347	–0.124	–0.044	0.470	0.390	C	0.125	0.182	3.963	0.81	+1.8	C	0.23
67	O	CH_2Ph	H	H	Et	(3a)	(3e)	0.006	2.222	1.483	1.466	0.739	0.756	B	2.159	2.158	(3e)	0.035	+6.0	B	0.005
68	O	CH_2Ph	Me	Me	Et	(3a)	(3e)	0.003	2.523	2.813	2.788	–0.290	–0.255	C	2.789	2.731	(3d)	0.003	+1.2	C	0.004
69	O	CH_2CH_2Ph	H	H	Et	5.362	9.485	0.096	1.018	1.123	0.878	–0.105	0.140	C	1.184	0.870	5.362	0.036	–2.7	C	0.19
70	O	Et	H	H	*i*–Pr	(2a)	(2c)	0.012	1.921	1.868	1.843	0.053	0.078	C	1.816	1.789	(2a)	0.039	+2.9	C	0.015
71	O	CH_2Ph	H	H	*i*–Pr	(3a)	(3e)	0.003	2.523	2.549	2.524	–0.026	–0.001	C	2.520	2.526	(3e)	0.068	+21	B	0.004
72	O	Et	H	H	*c*–Pr	(2a)	(2c)	0.100	1.000	0.459	0.427	0.541	0.573	B	1.032	0.995	(2b)	0.04	–2.5	C	0.091
73	O	$R^1 = H$	H	H	Me	–	–	>250	–	–	–	–	–	—	–	–	–	4.1	–	C	4.9
74	O	$R^1 = Me$	H	H	Me	–	–	>150	–	–	–	–	–	—	–	–	–	1.7	–	C	1.4
75	O	$R^1 = Et$	H	H	Me	1.324	–	2.2	–0.342	–0.160	–	–0.182	–	C	–0.426	–	1.324	2.8	+1.3	C	7.7
76	O	$R^1 = Bu$	H	H	Me	1.787	2.730	1.2	–0.079	–0.231	–0.426	0.152	0.346	C	0.063	–0.138	1.787	2.3	+1.9	C	2.6

[a]Conformers of substituent R^1 used for the construction of the general linear regression model with non-linear corrections for anti-HIV-1 activity (equation (2)) and for the specific models for class **A** (equation (3)), class **B** (equation (4)) and class **C** (equation (6)). The term low corresponds to the lowest-energy conformers and the term high to the highest-energy conformers for each substituent R^1 in the 0–10 kcal/mol range. Labels (1a) through (1i) refer to figure 1, labels (2a) through (2c) to figure 2, and labels (3a) through (3e) to figure 3. For substituents other than those displayed in these figures, the energies in kcal/mol for the most and least stable conformers in the 0–10 kcal/mol range are provided. [b]Experimental EC_{50} and $-\log EC_{50}$ values. [c] $-\text{LogEC}_{50}$ values obtained with the general model with non-linear corrections (equation (2)) for the lowest-energy and the highest-energy conformers in the range 0–10 kcal/mol. [d]Residual errors on the $-\log EC_{50}$ values obtained with the general model with non-linear corrections (equation (2)) for the lowest-energy and the highest-energy conformers for each substituent R^1 in the range 0–10 kcal/mol. [e]Class to which each compound belongs following the TSAR treatment. [f] $-\text{LogEC}_{50}$ values obtained for the lowest-energy and the highest- energy conformers for each substituent R^1 in the range 0–10 kcal/mol using equation (3) (class **A**), equation (4) (class **B**), or equation (6) (class **C**). The particular model used for the fit is given in the previous column 'Class'. [g]Conformation of the R^1 substituent in the CATALYST-derived conformer leading to the best fit for the anti-HIV-1 activity of HEPT analogues. [h]Best fit for the anti-HIV-1 activity of HEPT analogues obtained by using the most statistically significant hypothesis, number one, in table III. [i]Error relative to the CATALYST fitted anti-HIV activity of HEPT analogues. An error of + δ means that the fitted anti-HIV activity is equal to the experimental activity times δ, while an error of – δ means that the fitted anti-HIV activity is equal to the experimental activity times 1/δ. [j]Class to which each compound belongs following the CATALYST treatment. [k]Best fit for the anti–HIV-1 activity of HEPT analogues obtained using the specific hypothesis derived for classes **A**, **B** and **C** respectively (table IV). The specific hypothesis being used in every particular case is indicated in the previous column 'Class'.

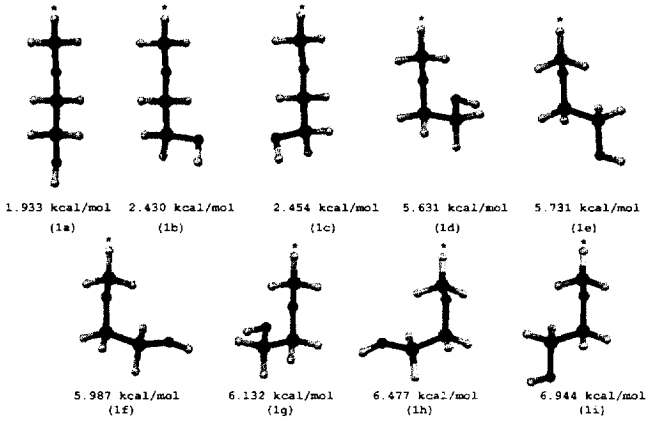


Fig 1. COBRA-generated conformers of the hydroxyethoxymethyl substituent classified by increasing energy in the 0–10 kcal/mol range. The star indicates the point of branching to the core structure.

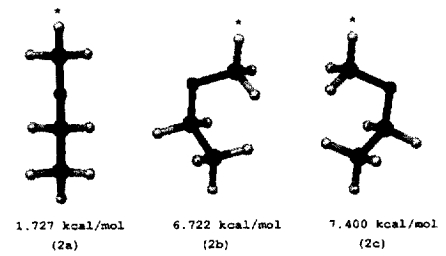


Fig 2. COBRA-generated conformers of the ethoxymethyl substituent classified by increasing energy in the 0–10 kcal/mol range. The star indicates the point of branching to the core structure.

in our case selected descriptors present in every row of the structure–activity table, while the output was the $-\log EC_{50}$ or $-\log CC_{50}$ values. The topology employed is highly flexible in terms of the number of input and output nodes, the number of layers and number of neurons within each layer. In most practical applications the number of input and output nodes are predetermined by the experimental data set, therefore the neurons which can be subject to adjustments in their number are the hidden neurons. It is believed that the ratio of the number of data points in the training set and the number of variables controlled by the network, ρ, is critical to the predictive power of the neural net. The range $1.8 < \rho < 2.2$ has been suggested as a guideline of acceptable ρ values. It is claimed that, for $\rho \ll 1.0$, the network simply memorizes the data, whereas for $\rho \gg 3.0$, the network loses its ability to generalize [14]. The neural network functionality within the TSAR software automatically computes the number of hidden neurons, as well as the number of training and test patterns in order to achieve the best ρ factor taking into account the number of substituent descriptors used.

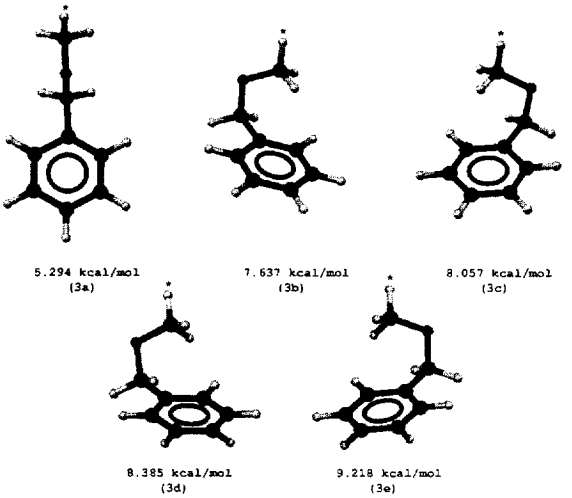


Fig 3. COBRA-generated conformers of the benzoxymethyl substituent classified by increasing energy in the 0.10 kcal/mol range. The star indicates the point of branching to the core structure.

The 3D-QSAR study was performed with CATALYST™ [15]. Each compound was minimized using the CHARMM-like [16] force field implemented in the software and a conformational sampling, applying the 'best searching procedure' option, selected representative conformers in the 0–30 kcal/mol range from the global minimum. As we shall see later, the conformation of the R^1 substituent does strongly influence the behavior of a HEPT analogue as an inhibitor of the HIV-1 reverse transcriptase. That is the reason why we looked carefully for some relationship between the global energy of the conformers generated with CATALYST and the energy of the R^1 substituent. Although we were not able to observe any direct relationship between the overall CATALYST-derived conformer energy and the particular conformation energy of the R^1 substituent for a given structure, for all of the molecules we found essentially the same conformers of substituent R^1 both in CATALYST-derived conformers and in structures generated by the COBRA software.

Hypotheses regarding the structure of the pharmacophore, described as a set of hydrophobic, hydrogen bond (HB) donor, HB acceptor, positively and negatively ionizable sites distributed in a 3D space, were generated using the CATALYST's CatHypo™ module [17]. The relevances of the different hypotheses obtained were submitted to statistical scrutinity based on their cost relative to the null hypothesis [17], their correlation coefficient r and their rms.

Results and discussion

Assessment of QSAR for antiviral activity

TSAR treatment

The number of conformers of substituent R^1 in the 0–10 kcal/mol range being different from one compound to the other, the statistical weight of each

structure in the final linear regression model will be different, thus leading to biased results. On the other hand, as revealed by a preliminary study, on a graph of predicted $-\log EC_{50}$ versus observed $-\log EC_{50}$, the representative points of all of the intermediate-energy conformers of substituent R^1, for every compound **1** through **76**, lie between the representative points of the most stable and the least stable conformers of the same substituent. These are the reasons why, in order to take into account in a statistically correct way the influence of the conformational flexibility of substituent R^1, we have chosen to represent each compound by two structures or 'pseudoconformers': one including the most stable conformer of substituent R^1, and a second using the least stable conformer, in the 0–10 kcal/mol range, for the same substituent. The same anti-HIV-1 activity was assigned to both structures owing to the lack of information regarding the bioactive form of the HEPT analogues.

Among all the substituent descriptors present in the TSAR structure–activity table, a small descriptor subset was selected using the TSAR multiple regression analysis functionality. We employed the multiple regression analysis technique instead of the common PLS method, in order to avoid handling PLS vectors which are linear combinations of descriptors with disparate physical meaning, ie, electronic, shape or topological descriptors. The descriptors in the selected subset are those which in the initial set exhibit correlation coefficients inferior to 0.5 and which lead to high partial F values during the automatic linear-regression model generation when processing the data corresponding to compounds **1–76**. After exclusion of the compounds having undefined activities, the best following equation (1) was obtained:

$$-\log EC_{50} = 0.098X_1 + 0.202X_2 + 0.426X_3 + 0.172X_4 + 2375.1X_5 - 15.98X_6 - 0.323X_7 + 0.971X_8 - 1.338X_9 + 1.424X_{10} + 0.171X_{11} - 0.004X_{12} - 1.858X_{13} - 11.06X_{14} - 7.735X_{15} + 6.060X_{16} + 0.845X_{17} - 2.186 \quad (1)$$

$n = 126$, $r = 0.919$, $s = 0.546$, $F = 35.25$, $r^2 = 0.845 > r^2_{cv} = 0.642$

where:

X_1 = calculated Verloop's L for R^1
X_2 = calculated Verloop's B_1 for R^1
X_3 = calculated Verloop's B_3 for R^1
X_4 = calculated Verloop's B_4 for R^1
X_5 = calculated total lipole for R^1
X_6 = calculated Kier ChiV6 (ring) index for R^1
X_7 = calculated flexibility index for R^1
X_8 = hydrogen bond acceptor index for R^1, which is set to one if the concerned substituent can act as a hydrogen bond acceptor and to zero otherwise
X_9 = Swain and Lupton F index from database [12] for R^2
X_{10} = Swain and Lupton R index from database [12] for R^2
X_{11} = π aromatic index from database [12] for R^2
X_{12} = calculated ellipsoidal volume for R^2
X_{13} = calculated Kier Chi3 (cluster) index for R^2
X_{14} = Swain and Lupton R index from database [12] for R^3
X_{15} = Swain and Lupton F index from database [12] for R^4
X_{16} = Swain and Lupton R index from database [12] for R^4
X_{17} = Verloop's B_3 from database [12] for R^4

The cross-validation method employed throughout this work for the validation of the multiple regression models corresponds to the fixed pattern deletion with three fixed groups. This means that every third row is deleted and the associated values predicted using the remaining two-thirds. This is repeated starting with the second row and then the third row. After the required group of data has been deleted, the remaining data are used to produce a new model. These values are compared to the exact values for the rows that have been held out. A model is produced for each group of three rows held out and values for the PRESS (predictive sum of squares) coefficient and the cross-validation correlation coefficient r_{cv} = 1 – PRESS/(total sum of squares) are derived [20].

From a graphical representation (fig 4) of the values of anti-HIV-1 activity estimated from equation (1) for every conformer versus the corresponding experi-

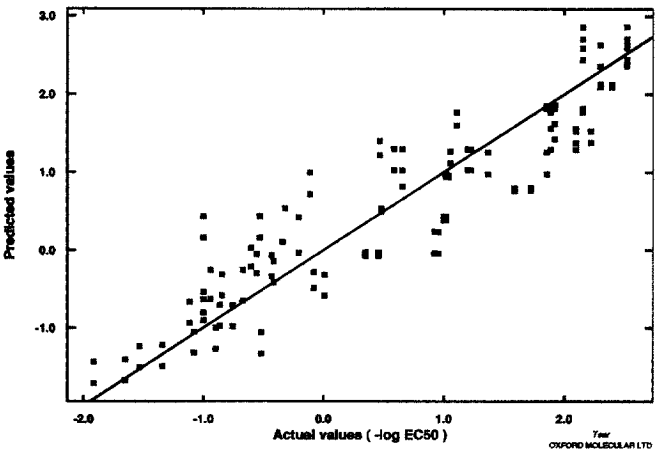


Fig 4. Estimated values of $-\log EC_{50}$ from equation (1) versus the corresponding experimental values for every most stable/least stable pair of conformers of compounds **1–76**. HEPT-O analogues are displayed in blue, while HEPT-S analogues are displayed in red.

mental values measured on a conformational equilibrium, it was apparent that the HEPT-O (**I**, X = O) and HEPT-S (**I**, X = S) were evenly distributed among each other, or approximately so, indicating that there was no need to separately treat these two types of HEPT analogues. On the other hand, considering the residual errors on $-\log EC_{50}$ (table I) for the same linear regression equation (1) reveals two subsets of outlying conformers:

Outliers with overestimated activities (class **A**). More strictly, we will define them as the 'pseudo-conformers' giving a residual error on $-\log EC_{50}$ inferior to –0.5 within model (1). The most salient representatives of this class are 'pseudoconformers' of the compounds **14**, **25**, **31** and **64**.

Outliers with underestimated activities (class **B**). More strictly, we will define them as the 'pseudo-conformers' giving a residual error on $-\log EC_{50}$ superior to 0.5 within model (1). The most salient representatives of this class are 'pseudoconformers' of compounds **24**, **30**, **35**, **47**, **53**, **56**, **61** and **67**.

Finally, class **C** includes all compounds having the two 'pseudoconformers' giving a residual error in the interval (–0.5, 0.5).

Multiple linear-regression techniques cannot directly provide information about nonlinear relationships between the dependent Y variable and the independent variables X_i. In the search of such dependencies, the selected parameters presented above have been used as input variables in a neural network approach to the 2D-QSAR problem concerning the HEPT analogues. As we have already mentioned, the topology we have used is the multiple layer feed forward network, here with a configuration 17–3–1 (17 input neurons, three hidden neurons and one output neuron). Exactly 20% of the rows in the TSAR structure–activity table have been excluded from the training set and used as a test set. The factor ρ was equal to 2.0 and the learning procedure was conducted until reaching the default 0.01 rms convergence on synaptic weights. Furthermore, we were able to determine some differences between the multiple linear regression QSAR model (1) and the QSAR model implemented in the trained neural network. These differences are due to nonlinear relationships, depicted in TSAR in the so-called dependence graphs, between the neural network output, the $-\log EC_{50}$ value, and every one of the input parameters. We found out from these graphs that the relationships between the output $-\log EC_{50}$ value on one hand and on the other hand X_3 (calculated Verloop's B_3 for R^1), X_4 (calculated Verloop's B_4 for R^1), X_{16} (Swain and Lupton R index for R^4) and X_{17} (Verloop's B_3 from database for R^4) can be accurately approximated by the sigmoidal function $sf(x) = 1/[1 + \exp(-x)]$. Moreover, it seems that X_{14} (Swain and Lupton R index for R^3) is related to $-\log EC_{50}$ by a function of the form $f(x) = \exp(-x)$. For all of the other relationships between the parameters X_i and the output $-\log EC_{50}$, we have adopted the linear approximation. Relying on these findings, we derived the following new multiple linear regression model (equation (2)):

$$-\log EC_{50} = 0.083X_1 - 0.083X_2 + 8.256 \cdot 1/[1 + \exp(-X_3)] + 17.88 \cdot 1/[1 + \exp(-X_4)] + 3896.3X_5 - 16.47X_6 - 0.346X_7 + 1.151X_8 - 1.343X_9 + 1.426X_{10} + 0.168X_{11} - 0.004X_{12} - 1.854X_{13} + 10.24 \cdot \exp(-X_{14}) + 2.902X_{15} + 38.87 \cdot 1/[1 + \exp(-X_{16})] + 11.92 \cdot 1/[1 + \exp(-X_{17})] - 62.43 \quad (2)$$

$n = 126$, $r = 0.920$, $s = 0.542$, $F = 35.84$,
$r^2 = 0.847 > r^2_{cv} = 0.811$

We can conclude that the predictive power of the model has been improved when including the non-linear effects as the value of the r^2_{cv} coefficient is now closer to the value of r^2. The graph plot of the predicted $-\log EC_{50}$ values versus the actual $-\log EC_{50}$ values can be found in figure 5. The color-coding represents the value of the residual errors on $-\log EC_{50}$. In table II we summarize the data characterizing the statistical significance of each of these parameters. We chose to estimate the statistical significance by the t-statistics, providing the t-value and the corresponding t-probability for each parameter. In analyzing these values, one should bear in mind that a t-probability of 0.05, for instance, indicates that a variable is significant at the 95% level. When considering table II,

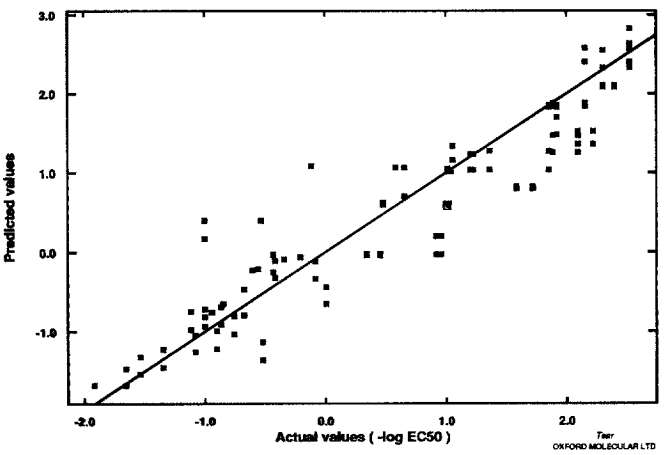


Fig 5. Estimated $-\log EC_{50}$ values derived from equation (2) versus the experimental $-\log EC_{50}$ values for all of the considered conformers. Red: –1.435 to –0.948; yellow: –0.947 to –0.461; green: –0.460 to 0.025; light blue: 0.026 to 0.512; dark blue: 0.513 to 0.999.

Table II. Statistical significance of parameters X_1 through X_{17} in the TSAR-derived model (2) describing the anti-HIV activity of HEPT analogues.

	Coefficient	*t-Value*	*t-Probability*
X_1	0.083	–0.797	0.284
X_2	–0.083	–0.766	0.905
X_3	8.256	7.949	7.588×10^{-6}
X_4	17.88	8.813	1.400×10^{-4}
X_5	3896.3	0.110	0.872
X_6	–16.47	–13.21	2.635×10^{-7}
X_7	–0.346	–6.717	4.462×10^{-9}
X_8	1.151	3.713	5.631×10^{-4}
X_9	–1.343	–8.948	7.297×10^{-5}
X_{10}	1.426	7.408	0.001
X_{11}	0.168	3.863	0.331
X_{12}	–0.004	–1.989	0.036
X_{13}	–1.854	–5.380	0.005
X_{14}	10.24	10.05	7.438×10^{-15}
X_{15}	2.902	–8.043	0.674
X_{16}	38.87	13.11	8.010×10^{-6}
X_{17}	11.92	4.385	4.850×10^{-10}

it becomes obvious that the parameters in model (2) which have the highest statistical significance are X_7 (calculated flexibility index for R^1), X_{14} (Swain and Lupton R index from database [12] for R^3) and X_{17} (Verloop's B_3 from database [12] for R^4). The other factors identified as strongly influencing the anti-HIV-1 activity of HEPT analogues are the shape, the lipophilicity and the hydrogen bonding properties of substituent R^1 as well as the electronic properties of the substituents R^2 and R^3 on the thiophenyl moiety, and the size and the electronic inductive properties of substituent R^4.

In order to investigate the particular behavior of each of the identified classes **A**, **B** and **C**, we first searched the 'pseudoconformers' falling in the subset **A** for similarities. The obvious similarity in the behavior of these molecular structures as inhibitors of the HIV-1 reverse transcriptase enzyme, is that for most of them, equation (2) gives overestimated activities especially when the R^1 substituent adopts an energetically unconstrained conformation as can be seen from figure 5 and table I. The chemical structures of these compounds are quite disparate and even after the computation of Carbo [18] and Hodgkin [18] similarity indices for the compounds falling in the subset **A** by using the ASP [18] software, we were not able to detect any general structural similarity.

As the number of 'pseudoconformers' belonging to the subset **A** was not large enough to achieve acceptable values for the factor ρ, we could not carry out a neural network analysis to search for nonlinear relationships among the variables. The linear multiple regression study revealed that the biological activity of 'pseudoconformers' belonging to the subset **A** is better described by the introduction in the model of some new descriptors different from those used in model (2). We found that some of the descriptors appearing in the previous general model (2) were no longer statistically significant and they were eliminated during the cross-validation procedure. We also found that the model was improved when hydrogen bond donor properties of substituent R^1 were taken as a descriptor instead of hydrogen bond acceptor properties. This indicates a possibly different mechanism of HIV-1 RT inhibition by the compounds of the subset **A**.

The linear regression model was derived by the procedure already described. Among all the substituent descriptors present in the TSAR structure–activity table, a small descriptor subset was selected using the TSAR multiple regression analysis functionality by sampling the descriptors exhibiting correlation coefficients less than 0.5 and which lead to high partial F values. Although not statistically very significant, because built on a small number of data points, we provide here the linear model obtained for the 'pseudoconformers' belonging to the subset **A**:

$$-\log EC_{50} = 0.116X_1 + 1.103X_2 + 0.281X_3 + 0.173X_4 + 27920X_5 - 12.03X_6 - 0.243X_7 - 0.184X_8 - 1.818X_9 - 0.002X_{10} - 10.13X_{11} - 2.684X_{12} + 1.179X_{13} - 2.601 \quad (3)$$

$n = 28$, $r = 0.992$, $s = 0.188$, $F = 51.88$, $r^2 = 0.984 > r^2_{cv} = 0.719$

where :

X_1 = calculated Verloop's L for R^1
X_2 = calculated Verloop's B_1 for R^1
X_3 = calculated Verloop's B_3 for R^1
X_4 = calculated Verloop's B_4 for R^1
X_5 = calculated total lipole for R^1
X_6 = calculated Kier ChiV6 (ring) index for R^1
X_7 = calculated flexibility index for R^1
X_8 = hydrogen bond donor index for R^1, which is set to one if the concerned substituent can act as a hydrogen bond donor and to zero otherwise
X_9 = Swain and Lupton F index for R^2
X_{10} = calculated ellipsoidal volume for R^2
X_{11} = Swain and Lupton R index from database [12] for R^3
X_{12} = Swain and Lupton F index from database [12] for R^4
X_{13} = Swain and Lupton R index from database [12] for R^4

Conformers of underestimated activity have been taken out of the general set and constitute the group **B**. According to figure 5 and table I, these molecules have in common the particularity of having their activities underestimated by the general model (2), especially when the R^1 substituent adopts a constrained, high-energy conformation. They are also characterized by the following structural features: $R^2 = R^3 = H$, $R^1 = CH_2OEt$ or CH_2OCH_2Ph, $R^4 = Me$ or Et.

As the number of compounds in subset **B** was not large enough to attain acceptable values for the factor ρ, we were not able to carry out a neural network analysis to search for a nonlinear relationship among the variables. Similarly to the previous case, the multiple regression linear study applied to the biological activity of confomers in subset **B** led to a model which was slightly different from model (2). Some of the descriptors appearing in the previous general model (2) were no longer statistically significant and they were eliminated during the cross-validation procedure. We again found that the model was improved when hydrogen bond donor properties of substituent R^1 were used as a descriptor instead of hydrogen bond acceptor properties. This indicates a possibly different mechanism of HIV-1 RT inhibition by the compounds of subset **B**. To obtain a linear model among all the substituent descriptors present in the TSAR structure–activity table a small descriptor subset was selected using the TSAR multiple regression analysis functionality by sampling the descriptors exhibiting correlation coefficients less than 0.5 and which led to high partial F values. We provide below (equation (4)) this linear model obtained for the conformers of subset **B**:

$$-\log EC_{50} = -0.923X_1 + 0.748X_2 + 0.260X_3 + 0.271X_4 - 4165.6X_5 + 0.093X_6 - 2.648X_7 - 0.039X_8 - 13.34X_9 - 20.83X_{10} + 3.790X_{11} + 0.286X_{12} + 2.733 \quad (4)$$

$n = 28$, $r = 0.998$, $s = 0.071$, $F = 296.9$, $r^2 = 0.996 > r^2_{cv} = 0.903$

where:

X_1 = calculated Verloop's L for R^1
X_2 = calculated Verloop's B_1 for R^1
X_3 = calculated Verloop's B_3 for R^1
X_4 = calculated Verloop's B_4 for R^1
X_5 = calculated total lipole for R^1
X_6 = hydrogen bond donor index for R^1, which is set to one if the concerned substituent can act as a hydrogen bond donor and to zero otherwise
X_7 = Swain and Lupton F index from database [12] for R^2
X_8 = calculated ellipsoidal volume for R^2
X_9 = Swain and Lupton R index from database [12] for R^3
X_{10} = Swain and Lupton F index from database [12] for R^4
X_{11} = Swain and Lupton R index from database [12] for R^4
X_{12} = Verloop's B_3 from database [12] for R^4

The rest of the conformers, lying close to the regression line in figure 5 (common behavior), constitutes the group **C**. For these compounds, a linear regression model has been obtained:

$$-\log EC_{50} = 0.009X_1 - 0.104X_2 + 0.353X_3 + 0.199X_4 - 2803.4X_5 - 17.06X_6 - 0.181X_7 + 0.405X_8 - 1.306X_9 + 1.230X_{10} + 0.265X_{11} - 0.001X_{12} - 1.736X_{13} - 7.980X_{14} - 63.91X_{15} + 22.65X_{16} + 0.389X_{17} - 1.195 \quad (5)$$

$n = 72$, $r = 0.994$, $s = 0.183$, $F = 243.2$, $r^2 = 0.987 > r^2_{cv} = 0.964$

where the X_i parameters are defined exactly as for models (1) and (2).

As regards the conformers in subset **C**, we were able to carry out a neural network analysis to reveal nonlinear relationships among the variables selected by the preliminary linear multiple regression study. The neural network topology used was the already mentioned multiple-layer feed forward neural network topology with a configuration 17–2–1. The value of the ρ factor was 1.8. The obtained dependence plots show the same functional relationship between the output variable $-\log EC_{50}$ and the input variables as established previously for the general model (1). A multiple linear-regression model has been obtained which improved the predictive qualities of model (5) as r^2 and r^2_{cv} are now much closer:

$$-\log EC_{50} = -0.029X_1 - 0.243X_2 + 7.215 \cdot 1/[1 + \exp(-X_3)] + 20.25 \cdot 1/[1 + \exp(-X_4)] + 1213.8X_5 - 18.22X_6 - 0.204X_7 + 0.629X_8 - 1.339X_9 + 1.274X_{10} + 0.250X_{11} - 0.002X_{12} - 1.723X_{13} + 7.054 \cdot \exp(-X_{14}) - 62.03X_{15} + 97.57 \cdot 1/[1 + \exp(-X_{16})] + 4.829 \cdot 1/[1 + \exp(-X_{17})] - 84.45 \quad (6)$$

$n = 72$, $r = 0.993$, $s = 0.174$, $F = 278.0$, $r^2 = 0.987 > r^2_{cv} = 0.981$

where the X_i parameters are defined exactly as for model (1). In figure 6 we present the graph plot of the predicted $-\log EC_{50}$ values versus the actual $-\log EC_{50}$ values color-coded according to the value of the residual errors on $-\log EC_{50}$. We can conclude that the overall quality of models (1) and (2) has been significantly improved as a general increase in the F, r^2 and r^2_{cv} parameters is observed.

CATALYST treatment

A pharmacophore search was undertaken using CATALYST with its default parameters (function weight 0.302, mapping coefficient 0, resolution 297 pm,

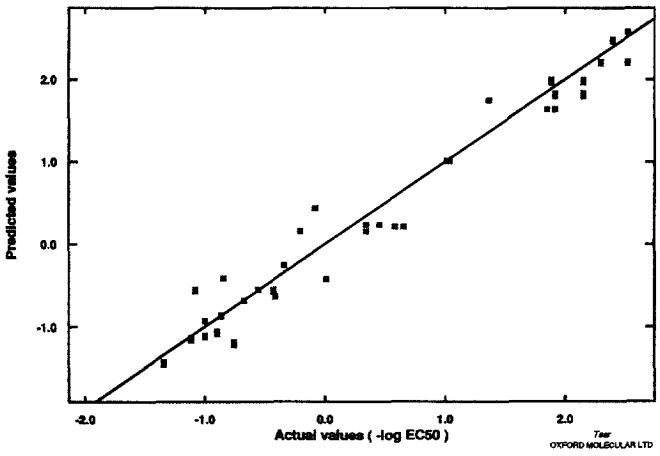


Fig 6. Estimated $-\log EC_{50}$ values from equation (6) versus the experimental $-\log EC_{50}$ values. Red: –0.244 to –0.121; yellow: –0.120 to –0.021; green: –0.020 to 0.091; light blue: 0.092 to 0.173; dark blue: 0.173 to 0.301.

activity uncertainty 3). Hypotheses were generated automatically taking into account hydrophobic sites (HPh), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD) positively ionizable sites (PosIon), and negatively ionizable sites (NegIon). Using the complete set of compounds (**1–76**), ten hypotheses were obtained. The interfeature distances for the most statistically significant hypothesis C are summarized in table III. The deviations of the fitted activities obtained with this hypothesis from the observed activities are presented in table I. Within the CATALYST paradigm a deviation δ of the biological activity means that the activity is situated somewhere in the interval: activity•1/δ...activity•δ. When comparing these deviations with the activity uncertainty 3, cited above, the compounds were grouped into three classes (table I):

Compounds belonging to class **A** with overestimated activities. More precisely, we define them as those compounds giving a predicted activity at least three times greater than the observed one.

Table III. Statistical significance criteria and geometrical parameters for the best CATALYST hypothesis describing the anti-HIV-1 activity of HEPT analogues.

Hyp	F_1	F_2	F_3	F_4	*Cost*	*Error*	*Weight*	*Config*	*Map*	*rms*	*r*
C	HPh1	HPh2	HBA1	HBA2	669.32	650.10	5.128	14.09	0	2.908	0.843
C null					1874.00	1874.00	0.000	0.000	0	8.276	0.000

	HPh1	*HPh2*	*HBA1 vector*		*HBA2 vector*	
			Origin	*End*	*Origin*	*End*
Weight	2.854 38	2.854 38	2.854 38		2.854 38	
Tolerance	1.60	1.60	1.60	2.20	1.60	2.20
HPh1	0.000					
HPh2	6.500	0.000				
HBA1						
Origin	7.200	4.300	0.000			
End	9.100	4.800	3.000	0.000		
HBA2						
Origin	3.900	3.700	4.600	5.600	0.000	
End	5.000	6.100	7.100	7.200	3.000	0.000

Interfeature distances and tolerances are in angstroms.

Compounds belonging to class **B** with underestimated activities. More precisely, we define them as those compounds giving a predicted activity at least three times smaller than the observed one.

All other compounds were grouped in class **C**.

We want to stress here that the tolerances on the fitted activities, in both TSAR and CATALYST treatments, for a compound to belong to the general class **C** are nearly equal because a deviation of ± 0.5 of the $-\log EC_{50}$ in TSAR corresponds to a deviation of 3 in CATALYST ($10^{0.5} \sim 3$). When referring to table I, the composition of the three classes is seen to be almost the same for the two techniques, confirming the relevance of this distribution into classes. In these conditions, three novel sets of hypotheses, each specific to one group, were generated and the most significant ones retained (table IV). The correlation coefficients are much better, particularly for compounds of groups **B** and **C**. The activities predicted using these three hypotheses are also presented in table I.

In figure 7 we present the best fits for $-\log EC_{50}$ values obtained with CATALYST versus $-\log EC_{50}$ values obtained with TSAR. As most of the points fall directly on a straight line, the cross-validation of the methods can be qualified as satisfactory. The statistical parameters relative to this relationship are: $r = 0.965$, $r^2 = 0.931$, $r^2_{cv} = 0.925$, $s = 0.341$, $F = 916.4$.

The stuctural features of compounds of type **B** are also encountered in a group of HEPT analogues which has been shown to interact with HIV reverse transcriptase in a peculiar way [19]. Whereas HEPT and its common congeners act as competitive inhibitors toward both deoxythymidine triphosphate (dTTP) and deoxyguanosine triphosphate (dGTP), this family of compounds behaves as competitive inhibitors in the presence of a dTTP and as non-competitive inhibitors toward dGTP. These distinct behaviors have been rationalized [19, 20] by the presence of two target sites: a high affinity site occupied by both natural substrates and common HEPT; and a low affinity allosteric site on which would attach compounds belonging to group **B**.

The substituent R^1 in a typical class **B** compound, **53**, adapts its geometry to the specific hypothesis B_1 when adopting a high-energy folded conformation (fig 8), whereas to adapt to the common hypothesis C_1, it should exist in a low-energy conformation (fig 9). In both cases, the 5-ethyl group acts as a hydrophobic site and the sulfur atom as a HB acceptor. Conversely, the second hydrophobic site corresponds either to the benzyloxy group (hypothesis B_1) or to the phenylthio group (hypothesis C_1). These observations confirm the results obtained by TSAR, namely that compounds in class **B** behave as powerful inhibitors of HIV-1 reverse transcriptase enzyme and probably bind to the allosteric site only when the R^1 substituent adopts a folded high-energy conformation. Conversely, they behave as competitive inhibitors toward dTTP and dGTP as do most of HEPT congeners when the R^1 substituent adopts an unfolded low-energy conformation.

For compounds of group **A**, which also constitute a homogeneous subset, no specific biochemical behavior has been discovered so far.

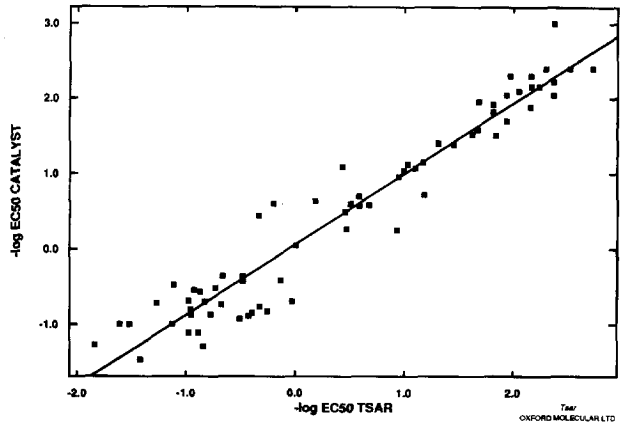


Fig 7. Cross-validation of the results obtained with TSAR and CATALYST software: $-\log EC_{50}$ values obtained with CATALYST versus $-\log EC_{50}$ values obtained with TSAR.

Table IV. CATALYST's best and null hypotheses for the anti-HIV activity of each of the three classes **A**, **B** and **C**.

Hyp	F_1	F_2	F_3	F_4	*Cost*	*Error*	*Weight*	*Config*	*Map*	*rms*	*r*
A1	HPh	HPh	HPh	HBA	96.00	79.46	1.439	15.110	0	1.327	0.885
A null					146.1	146.1	0.000	0.000	0	2.847	0
B1	HPh	HBA	HBA		44.86	29.74	1.283	13.840	0	0.376	0.987
B null					54.07	54.07	0.000	0.000	0	2.238	0
C1	HPh	HPh	HBA	HBA	144.8	127.9	2.152	14.760	0	1.311	0.924
C null					298.9	298.9	0.000	0.000	0	3.431	0

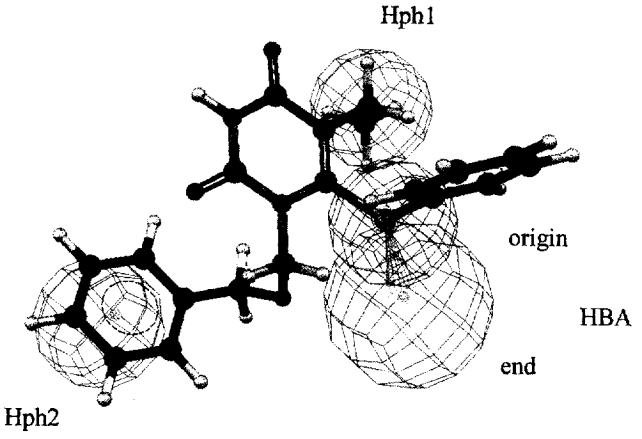


Fig 8. Superposition of compound **53** (class B) on CATALYST's B1 hypothesis (cf table IV).

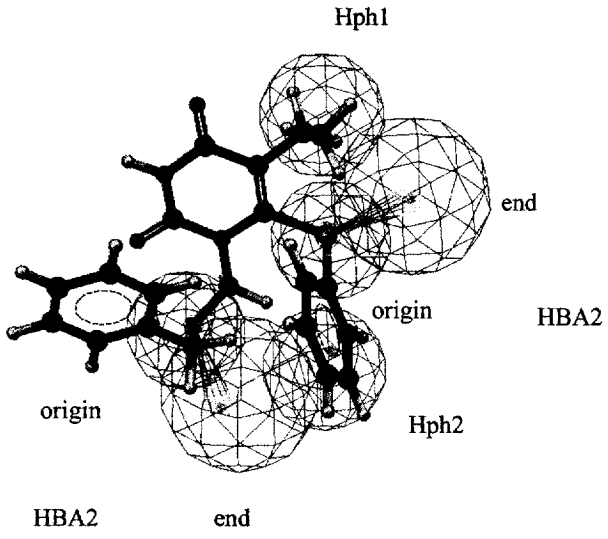


Fig 9. Superposition of compound **53** (class B) on CATALYST's C1 hypothesis (cf table IV).

Assessment of QSAR for cytotoxicity

Cytotoxicity constitutes the second component of the selectivity index; hence the interest in predicting its value. Experimental data (CC_{50}) are generally available but often as the superior limit of the non-cytotoxic concentration. Moreover, they encompass a narrower concentration range than the EC_{50} data (cf table V). Notwithstanding these limitations, the three former approaches (2D-QSAR, neural network analysis with TSAR, and CATALYST) have been applied to SAR cytotoxicity studies.

TSAR treatment

In view of obtention of a linear model, among all the substituent descriptors present in the TSAR structure–activity table, a small descriptor subset was selected using a regression analysis encompassing all compounds with known cytotoxic concentrations. The following descriptors exhibiting correlation coefficients less than 0.5 and leading to high partial F values during the automatic model generation were selected:

X_1 = calculated Verloop's B_1 for R^1
X_2 = calculated total dipole for R^1
X_3 = hydrogen bond donor index for R^1, which is set to one if the concerned substituent can act as a hydrogen bond donor and to zero otherwise
X_4 = calculated log$P$ for R^1
X_5 = Verloop's B_1 from database [12] for R^2
X_6 = calculated Wiener topological index with hydrogens included for R^2
X_7 = Verloop's B_1 from database [12] for R^3
X_8 = calculated Wiener topological index with hydrogens included for R^3
X_9 = Verloop's B_1 from database [12] for R^4
X_{10} = calculated log$P$ for R^4
X_{11} = calculated bond dipole for R^4

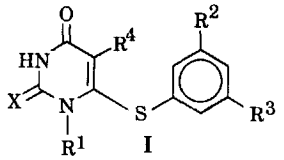
The following model describing the cytotoxicity of the HEPT analogues has been obtained:

$$-\log CC_{50} = 0.068X_1 - 0.005X_2 - 0.129X_3 + 0.369X_4 + 0.253X_5 + 0.0009X_6 + 0.492X_7 - 0.014X_8 + 0.347X_9 + 0.080X_{10} - 0.479X_{11} - 3.407 \quad (7)$$

$n = 118$, $r = 0.882$, $s = 0.206$, $F = 41.72$,
$r^2 = 0.778 > r^2_{cv} = 0.743$

Its statistical significance was not high, but the examination of figure 10 representing predicted $-\log CC_{50}$ versus experimental $-\log CC_{50}$ values led to the following conclusions. Whereas the HEPT-O compounds (**I**, X = O, colored in blue in figure 10) exhibited an acceptable correlation between predicted and experimental values, the values corresponding to the HEPT-S series (**I**, X = S, colored in red in figure 10) were considerably dispersed. On the other hand, contrary to our observations regarding the EC_{50} values, the conformation of the R^1 group appeared to play a minor role. These findings were confirmed by a complementary neural network analysis of the cytotoxic data available. The employed neural network topology was a multiple-layer feed forward neural

Table V. Numerical data concerning in vitro cytotoxicities (CC_{50}/μM) of HEPT analogues: experimental and estimated values obtained by TSAR and CATALYST analysis.



I

Compound	X	R^1 = CH_2OR R	R^2	R^3	R^4	R^1 Conformer[a] Low	R^1 Conformer[a] High	Experiment[b] CC_{50}	Experiment[b] $-LogCC_{50}$	TSAR $-LogCC_{50}$ fit[c] Low	TSAR $-LogCC_{50}$ fit[c] High	TSAR Residuals fit[d] Low	TSAR Residuals fit[d] High	TSAR $-LogCC_{50}$ fit[e] Low	TSAR $-LogCC_{50}$ fit[e] High	CATALYST CC_{50} fit[f]	CATALYST Fitted conformer[g]
1	O	CH_2CH_2OH	Me	H	Me	(1a)	(1i)	420	−2.623	−2.376	−2.387	−0.247	−0.236	−2.408	−2.433	190	8
2	O	CH_2CH_2OH	Et	H	Me	(1a)	(1i)	181	−2.258	−2.345	−2.355	0.087	0.098	−2.392	−2.417	210	4
3	O	CH_2CH_2OH	*t*-Bu	H	Me	(1a)	(1i)	75	−1.875	−1.908	−1.919	0.033	0.044	−1.881	−1.906	100	11
4	O	CH_2CH_2OH	CH_2OH	H	Me	(1a)	(1i)	292	−2.465	−2.368	−2.379	−0.098	−0.087	−2.404	−2.429	200	4
5	O	CH_2CH_2OH	CF_3	H	Me	(1a)	(1i)	196	−2.292	−2.261	−2.272	−0.031	−0.021	−2.224	−2.429	220	16
6	O	CH_2CH_2OH	F	H	Me	(1a)	(1i)	282	−2.450	−2.428	−2.439	−0.022	−0.012	−2.481	−2.506	230	29
7	O	CH_2CH_2OH	Cl	H	Me	(1a)	(1i)	210	−2.322	−2.315	−2.326	−0.007	0.003	−2.300	−2.325	140	13
8	O	CH_2CH_2OH	Br	H	Me	(1a)	(1i)	141	−2.149	−2.277	−2.288	0.128	0.139	−2.240	−2.265	170	9
9	O	CH_2CH_2OH	I	H	Me	(1a)	(1i)	106	−2.025	−2.227	−2.288	0.202 0	0.212	−2.160	−2.185	180	38
10	O	CH_2CH_2OH	NO_2	H	Me	(1a)	(1i)	170	−2.230	−2.366	−2.347	0.106	0.116	−2.338	−2.363	150	29
11	O	CH_2CH_2OH	OH	H	Me	(1a)	(1i)	446	−2.649	−2.427	−2.438	−0.222	−0.212	−2.480	−2.505	290	22
12	O	CH_2CH_2OH	OMe	H	Me	(1a)	(1i)	>250	–	2.398	−2.378	–	–	2.473	−2.498	110	14
13	O	CH_2CH_2OH	Me	Me	Me	(1a)	(1i)	243	−2.386	−2.243	−2.254	−0.143	−0.132	−2.307	−2.332	130	27
14	O	CH_2CH_2OH	Cl	Cl	Me	(1a)	(1i)	130	−2.114	−1.916	−1.927	−0.198	−0.187	−1.961	−1.986	100	17
15	S	CH_2CH_2OH	Me	Me	Me	(1a)	(1i)	172	−2.236	−2.243	−2.254	0.008	0.018	−2.307	−2.332	220	9
16	O	CH_2CH_2OH	COOMe	H	Me	(1a)	(1i)	221	−2.344	−2.245	−2.256	−0.099	−0.088	−2.238	−2.263	110	29
17	O	CH_2CH_2OH	COMe	H	Me	(1a)	(1i)	228	−2.358	−2.263	−2.273	−0.095	−0.084	−2.247	−2.272	200	43
18	O	CH_2CH_2OH	COOH	H	Me	(1a)	(1i)	352	−2.547	−2.355	−2.366	−0.191	−0.180	−2.376	−2.401	96	22
19	O	CH_2CH_2OH	$CONH_2$	H	Me	(1a)	(1i)	306	–	–	–	–	–	2.380	−2.405	250	23
20	O	CH_2CH_2OH	CN	H	Me	(1a)	(1i)	234	−2.369	−2.364	−2.375	−0.005	0.006	−2.457	−2.482	270	22
21	O	CH_2CH_2OH	H	H	Allyl	(1a)	(1i)	183	−2.262	−2.402	−2.413	0.140	0.150	−0.970	−0.995	110	33
22	O	CH_2CH_2OH	H	H	COOMe	(1a)	(1i)	6.6	−0.820	−0.924	−0.935	0.104	0.115	−0.840	−0.907	11	46
23	O	CH_2CH_2OH	H	H	COONHPh	(1a)	(1i)	18	−1.255	−0.989	−0.994	−0.266	−0.261	−0.993	−0.990	29	3
24	S	CH_2CH_2OH	H	H	Et	(1a)	(1i)	148	−2.170	−2.491	−2.502	0.321	0.332	−2.334	−2.252	200	14
25	S	CH_2CH_2OH	H	H	Pr	(1a)	(1i)	230	−2.362	−2.456	−2.466	0.094	0.105	−2.334	−2.252	170	22
26	S	CH_2CH_2OH	H	H	*i*-Pr	(1a)	(1i)	400	−2.602	−2.295	−2.306	−0.307	−0.296	−2.221	−2.139	200	30
27	S	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	277	−2.442	−2.218	−2.229	−0.224	−0.213	−2.156	−2.074	110	40
28	S	CH_2CH_2OH	Me	Me	*i*-Pr	(1a)	(1i)	52	−1.716	−2.022	−2.033	0.306	0.317	−2.043	−1.961	78	33
29	S	CH_2CH_2OH	Cl	Cl	Et	(1a)	(1i)	64	−1.806	−1.892	−1.902	0.086	0.096	−1.878	−1.796	88	10
30	O	CH_2CH_2OH	H	H	Et	(1a)	(1i)	400	−2.602	−2.491	−2.502	−0.111	−0.100	−2.564	−2.589	110	5
31	O	CH_2CH_2OH	H	H	Pr	(1a)	(1i)	244	−2.387	−2.456	−2.466	0.068	0.079	−2.496	−2.521	85	24
32	O	CH_2CH_2OH	H	H	*i*-Pr	(1a)	(1i)	231	−2.364	−2.295	−2.306	−0.068	−0.058	−2.338	−2.363	160	5
33	O	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	149	−2.173	−2.218	−2.229	0.045	0.056	−2.250	−2.275	98	41
34	O	CH_2CH_2OH	Me	Me	*i*-Pr	(1a)	(1i)	128	−2.107	−2.022	−2.033	−0.085	−0.074	−2.024	−2.049	130	63
35	O	CH_2CH_2OH	Cl	Cl	Et	(1a)	(1i)	51	−1.708	−1.892	−1.902	0.184	0.195	−1.904	−1.929	75	51
36	O	CH_2CH_2OH	H	H	H	(1a)	(1i)	743	−2.871	−2.672	−2.683	−0.199	−0.188	−2.813	−2.838	460	8
37	S	CH_2CH_2OH	H	H	H	(1a)	(1i)	123	−2.090	−2.672	−2.672	0.583	0.593	−2.447	−2.365	80	12
38	O	CH_2CH_2OMe	H	H	Me	2.588	6.479	299	−2.476	−2.413	−2.428	−0.062	−0.047	−2.502	−2.537	350	15
39	O	CH_2CH_2O-*n*-C_5H_{11}	H	H	Me	3.361	3.974	55	−1.740	−1.820	−1.830	0.080	0.090	−1.815	−1.839	88	19

Table V. *Continued.*

40	O	$CH_2CH_2OCH_2Ph$	H	H	Me	6.169	9.177	45	−1.653	−1.757	−1.746	0.103	0.092	−1.741	−1.799	98	7
41	O	Me	H	H	Me	1.030	–	244	−2.387	−2.360	–	0.028	–	2.449	–	300	8
42	O	Et	H	H	Me	(2a)	(2c)	231	−2.364	−2.233	−2.233	−0.131	−0.131	−2.302	−2.302	91	21
43	O	Pr	H	H	Me	2.178	9.351	147	−2.167	−2.059	−2.060	−0.108	−0.107	−2.099	−2.099	81	25
44	O	Bu	H	H	Me	2.401	5.862	83	−1.919	−1.912	−1.912	−0.007	−0.007	−1.928	−1.929	76	8
45	O	$CH_2CH_2SiMe_3$	H	H	Me	–	–	32	–	–	–	–	–	–	–	84	6
46	O	CH_2Ph	H	H	Me	(3a)	(3e)	95	−1.978	−1.702	−1.683	−0.275	−0.295	−1.686	−1.722	89	24
47	S	Et	H	H	Et	(2a)	(2c)	81	−1.908	−2.208	−2.208	0.300	0.300	−2.079	−2.078	92	20
48	S	Et	Me	Me	Et	(2a)	(2c)	>100	–	1.903	−1.897	–	–	−1.90	−1.90	93	37
49	S	Et	Cl	Cl	Et	(2a)	(2c)	45	−1.653	−1.609	−1.609	−0.044	−0.044	−1.622	−1.622	86	34
50	S	$R^1 = CH_2$-*i*-Pr	H	H	*i*-Pr	6.619	9.203	>100	–	2.079	−2.090	–	–	−2.00	–	240	24
51	S	*c*-Hex	H	H	Et	8.096	9.639	223	−2.348	−1.747	−1.748	−0.602	−0.601	−1.784	−1.778	91	10
52	S	CH_2-*c*-Hex	H	H	Et	3.963	7.911	>100	–	−2.316	−2.312	–	–	−1.699	−1.694	120	17
53	S	CH_2Ph	H	H	Et	(3a)	(3e)	>100	–	−2.068	−2.057	–	–	−1.722	−1.520	–	–
54	S	CH_2Ph	Me	Me	Et	(3a)	(3e)	>20	–	−1.208	−1.211	–	–	−1.544	−1.342	–	–
55	S	$CH_2C_6H_4$(4-Me)	H	H	Et	5.499	9.288	>20	–	−1.398	−1.407	–	–	−1.612	−1.404	110	32
56	S	$CH_2C_6H_4$(4-Cl)	H	H	Et	4.492	8.412	>20	–	−1.602	−1.607	–	–	−1.542	−1.387	110	27
57	S	CH_2CH_2Ph	H	H	Et	5.362	9.495	>20	–	−1.301	−1.307	–	–	−1.682	−1.463	160	4
58	S	Et	H	H	*i*-Pr	(2a)	(2c)	>100	–	−1.954	−1.934	–	–	−1.966	−1.965	110	19
59	S	CH_2Ph	H	H	*i*-Pr	(3a)	(3e)	>20	–	−1.205	−1.201	–	–	−1.609	−1.408	120	23
60	S	Et	H	H	*c*-Pr	(2a)	(2c)	46	−1.663	−2.058	−2.058	0.395	0.395	−1.979	−1.978	110	6
61	O	Et	H	H	Et	(2a)	(2c)	161	−2.207	−2.208	−2.208	0.002	0.002	−2.245	−2.246	74	19
62	O	Et	Me	Me	Et	(2a)	(2c)	>100	–	−1.863	−1.857	–	–	−1.931	−1.927	85	7
63	O	Et	Cl	Cl	Et	(2a)	(2c)	45	−1.653	−1.609	−1.609	−0.044	−0.044	−1.585	−1.585	123	28
64	O	*i*-Pr	H	H	Et	1.382	–	143	−2.155	−2.055	–	0.010	–	−2.067	–	79	16
65	O	*c*-Hex	H	H	Et	8.096	9.639	>100	–	−1.653	−1.657	–	–	−1.708	−1.709	110	36
66	O	CH_2-*c*-Hex	H	H	Et	3.963	7.911	17	−1.230	−1.632	−1.629	0.401	0.399	−1.575	−1.575	69	18
67	O	CH_2Ph	H	H	Et	(3a)	(3e)	34	−1.531	−1.678	−1.658	0.146	0.127	−1.630	−1.665	80	12
68	O	CH_2Ph	Me	Me	Et	(3a)	(3e)	>20	–	−1.176	−1.169	–	–	−1.316	−1.351	86	15
69	O	CH_2CH_2Ph	H	H	Et	5.362	9.485	38	−1.580	−1.582	−1.564	0.002	−0.016	−1.515	−1.555	93	19
70	O	Et	H	H	*i*-Pr	(2a)	(2c)	106	−2.025	−2.012	−2.012	−0.013	−0.013	−2.019	−2.019	82	10
71	O	CH_2Ph	H	H	*i*-Pr	(3a)	(3e)	>20	–	−1.544	−1.532	–	–	−1.404	−1.404	120	11
72	O	Et	H	H	*c*-Pr	(2a)	(2c)	224	−2.350	−2.058	−2.058	−0.292	−0.292	−2.105	−2.105	120	17
73	O	$R^1 = H$	H	H	Me	–	–	250	−2.398	−2.437	–	0.039	–	−2.322	–	1200	4
74	O	$R^1 = Me$	H	H	Me	–	–	150	−2.176	−2.301	–	0.125	–	−2.327	–	420	3
75	O	$R^1 = Et$	H	H	Me	1.324	–	94	−1.973	−2.174	–	0.201	–	−2.179	–	130	4
76	O	$R^1 = Bu$	H	H	Me	1.787	2.730	89	−1.949	−1.854	−1.855	−0.095	−0.095	−1.808	−1.810	82	17

[a]Conformers of substituent R^1 used for the construction of the general linear regression model with non-linear corrections for cytotoxicity (equation (8)) and for the specific models for HEPT–O (equation (10)) and HEPT-S analogues (equation (11)). The term low corresponds to the lowest-energy conformers and the term high to the highest-energy conformers for each substituent R^1 in the 0–10 kcal/mol range. Labels (1a) through (1i) refer to figure 1, labels (2a) through (2c) to figure 2 and labels (3a) through (3e) to figure 3. For substituents other than those displayed in these figures, the energies in kcal/mol for the most and least stable conformers in the 0–10 kcal/mol range are provided. [b]Experimental CC_{50} and $-\log CC_{50}$ values. [c]$-\text{LogCC}_{50}$ values obtained with the general model with non-linear corrections (equation (8)) for the lowest-energy and the highest-energy conformers in the range 0–10 kcal/mol. [d]Residual errors on the $-\log CC_{50}$ values obtained with the general model with non-linear corrections (equation (8)) for the lowest-energy and the highest-energy conformers for each substituent R^1 in the range 0–10 kcal/mol. [e]$-\text{LogCC}_{50}$ values obtained with the specific models for HEPT-O (equation (10)) and HEPT-S analogues (equation (11)) respectively for the lowest-energy and the highest-energy conformers for each substituent R^1 in the range 0–10 kcal/mol. [f]Values for CC_{50} obtained with CATALYST. [g]Conformers leading to the best mapping onto the CATALYST-derived cytotoxicity model. The conformers are labeled with numbers directly proportional to their energy in the 0–30 kcal/mol range.

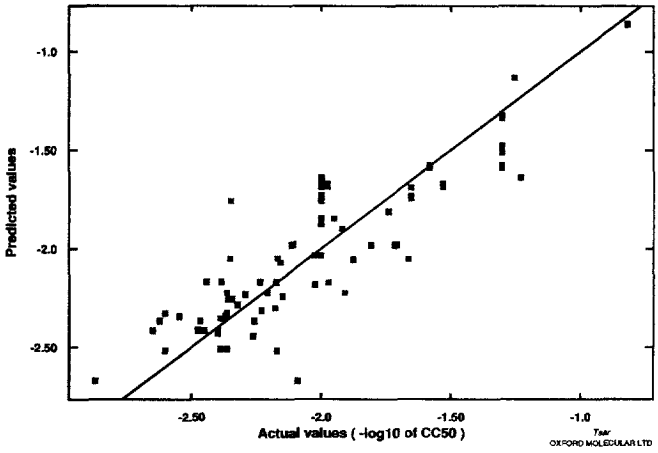


Fig 10. Estimated values of $-\log CC_{50}$ from equation (7) versus the corresponding experimental values for every most stable/least stable pair of conformers for compounds **1–76**: HEPT-O analogues are displayed in blue, while HEPT-S analogues are displayed in red.

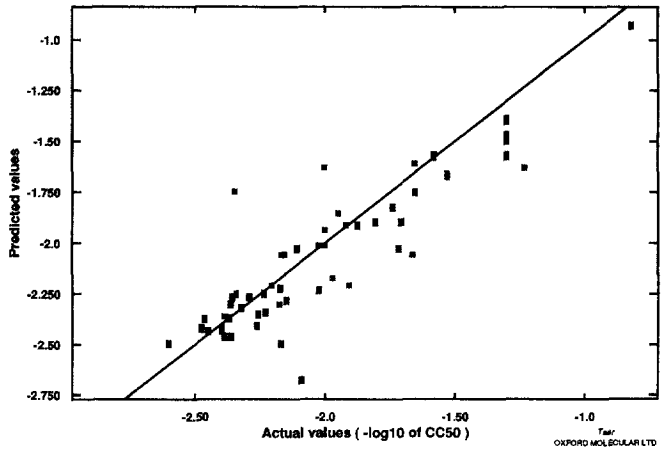


Fig 11. Estimated values of $-\log CC_{50}$ for HEPT analogues from equation (8) versus experimental $-\log CC_{50}$ values. Red: –0.314 to –0.191; yellow: –0.190 to –0.059; green: –0.058 to 0.084; light blue: 0.085 to 0.207; dark blue: 0.208 to 0.340.

network topology with a configuration 11–4–1. The value of the ρ factor for the problem was 2.1. Again an improvement of the predictive power of the model is observed when taking in account the nonlinear relationship between the output variable $-\log CC_{50}$ and only one parameters X_{11} (bond dipole for R^4). This relationship was established to be inverse sigmoidal when referring to the dependence plots between the output variable $-\log CC_{50}$ and the parameters X_i for the obtained model. This new model is given by:

$$-\log CC_{50} = 0.068X_1 - 0.006X_2 - 0.127X_3 + 0.371X_4 + 0.251X_5 + 0.000\,9643X_6 + 0.498X_7 - 0.014X_8 + 0.374X_9 + 0.064X_{10} + 2.575 \cdot 1/[1 + \exp(X_{11})] - 4.639 \quad (8)$$

$n = 118$, $r = 0.882$, $s = 0.207$, $F = 61.75$, $r^2 = 0.777 > r^2_{cv} = 0.759$

In figure 11, we present the graph plot of the predicted $-\log CC_{50}$ values versus the actual $-\log CC_{50}$ values, color-coded according to the value of the residual errors on $-\log CC_{50}$.

We divided the general set into two distinct subsets, the first containing the HEPT-O and the second the HEPT-S derivatives and for each compound, took into consideration two conformers of the R^1 group, the most and the least stable. The descriptors selected from the initial set by the multiple linear regression procedure already described were the same as those used for the general set.

For the HEPT-O subset, the following linear regression equation was established:

$$-\log CC_{50} = -0.174X_1 - 0.013X_2 - 0.197X_3 + 0.430X_4 + 0.405X_5 + 0.0005X_6 + 0.416X_7 - 0.013X_8 + 0.346X_9 + 0.169X_{10} - 0.518X_{11} - 3.131 \quad (9)$$

$n = 94$, $r = 0.948$, $s = 0.139$, $F = 72.02$, $r^2 = 0.898 > r^2_{cv} = 0.829$

An additional neural network analysis of the cytotoxic data for the HEPT-O analogues was performed in view of improving the initial linear model. The employed neural network topology was a multiple-layer feed forward neural network topology with a configuration 11–4–1. The value of the ρ factor for the problem was 2.0. The same nonlinear relationship between the parameters and the output value, $-\log CC_{50}$, were obtained as in the case of the general model for the cytotoxicities. An improvement of the predictive power of the model was observed when taking into account the nonlinear relationship between the output variable $-\log CC_{50}$ and parameter X_{11} (bond dipole for R^4):

$$-\log CC_{50} = -0.172X_1 - 0.013X_2 - 0.194X_3 + 0.431X_4 + 0.401X_5 + 0.0005X_6 + 0.425X_7 - 0.013X_8 + 0.384X_9 + 0.144X_{10} + 2.773 \cdot 1/[1 + \exp(X_{11})] - 4.472 \quad (10)$$

$n = 94$, $r = 0.949$, $s = 0.131$, $F = 69.27$, $r^2 = 0.897 > r^2_{cv} = 0.842$

In table VI we have presented the data characterizing the statistical significance of each of these parameters. The statistical significance is described by the

Table VI. Statistical significance of parameters X_1 through X_{11} in the TSAR-derived model (8) describing the cytotoxicity of HEPT analogues.

	Coefficient	*t-Value*	*t-Probability*
X_1	0.068	0.318	0.751
X_2	–0.006	–0.238	0.812
X_3	–0.127	–1.316	0.190
X_4	0.371	14.900	0.0
X_5	0.251	3.171	0.002
X_6	9.643×10^{-4}	1.134	0.259
X_7	0.498	4.700	6.504×10^{-6}
X_8	–0.014	–1.952	0.053
X_9	0.374	3.595	4.582×10^{-4}
X_{10}	0.064	0.851	0.396
X_{11}	2.575	11.160	0.0

t-value and the *t*-probability associated with each parameter. The most statistically significant parameters are X_4 (calculated log*P* for R[1]), X_7 (Verloop's B_1 from database [12] for R[3]) and X_{11} (calculated bond dipole for R[4]).

For the HEPT-S subset, by using the same descriptors as for the HEPT-O derivatives, equation [11] has been obtained:

$$-\log CC_{50} = 0.915X_1 + 0.043X_2 + 0.028X_3 + 0.243X_4 + 0.570X_5 - 0.013X_6 + 0.217X_9 - 4.710 \quad (11)$$

$n = 24$, $r = 0.745$, $s = 0.298$, $F = 5.873$, $r^2 = 0.555 > r^2_{cv} = 0.490$

CATALYST treatment

We have investigated all the HEPT analogues within CATALYST without distributing them into HEPT-O and HEPT-S subsets. An acceptable hypothesis was selected from the ten generated. Data relative to its geometrical properties and its statistical significance are summarized in table VII, while the superposition of HEPT onto this hypothesis is represented in figure 12.

Table VII. Statistical significance criteria and geometrical parameters for the best CATALYST hypothesis describing the cytotoxicity of HEPT analogues.

Hyp	F_1	F_2	F_3	F_4	*Cost*	*Error*	*Weight*	*Config*	*Map*	*rms*	*r*
CC	HPh 1	HPh 2	HBD	HBA	221.7	202.7	4.131	14.93	0	1.542	0.724
CC null					277.0	277.0	0.000	0.000	0	2.233	0

	HPh1	*HPh2*	*HBD vector*		*HBA vector*	
			Origin	*End*	*Origin*	*End*
Weight	0.88411	0.88411	0.88411		0.88411	
Tolerance	1.60	1.60	1.60	2.20	1.60	2.20
HPh1	0.000					
HPh2	7.400	0.000				
HBD						
Origin	3.500	9.900	0.000			
End	4.300	9.300	3.000	0.000		
HBA						
Origin	4.800	3.400	7.700	7.900	0.000	
End	6.600	5.300	9.800	10.300	3.000	0.000

Interfeature distances and tolerances are in angstroms.

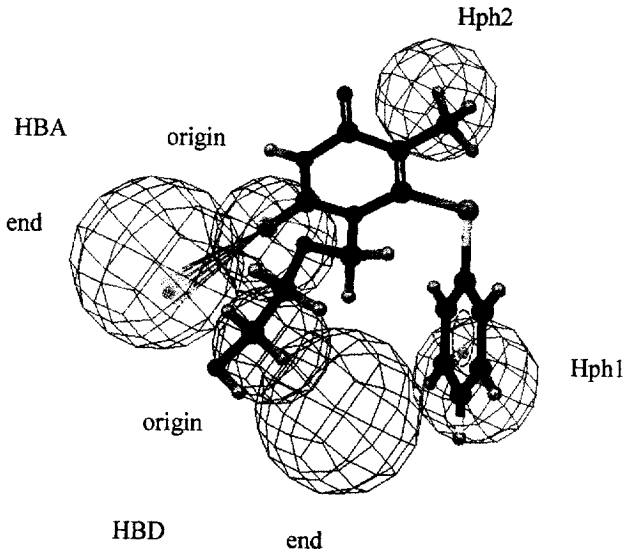


Fig 12. Superposition of HEPT on CATALYST's cytotoxicity hypothesis CC_1: red, HB donor sites, other color-coding as in figures 8 and 9.

Predicted and experimental cytotoxicity values are presented in table V.

The cross-validation of the results concerning the cytotoxicities and obtained by the TSAR and CATALYST software is presented in figure 13. The statistical parameters relative to this distribution are the following: n = 143, r = 0.622, r^2 = 0.387, r^2_{cv} = 0.348, s = 0.211, F = 44.28.

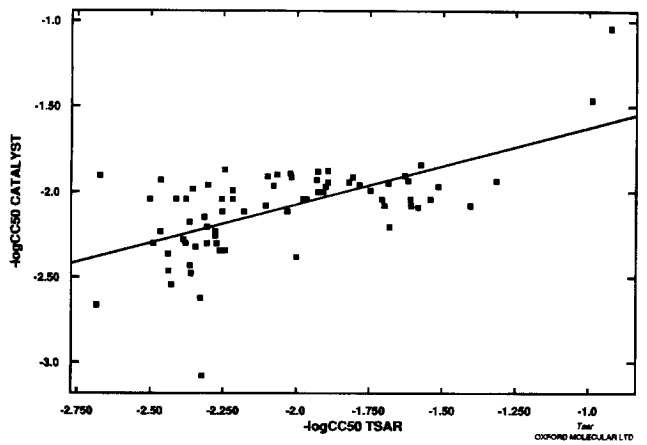


Fig 13. Cross-validation of the results obtained with TSAR and CATALYST software: $-\log CC_{50}$ values obtained with CATALYST versus $-\log CC_{50}$ values obtained with TSAR.

Testing the proposed models on compounds not included in the former QSAR analyses

Our models were tested on compounds considerably different structurally from the systems used in the assessment of the QSAR, ie, 6-phenylselenyl analogues of HEPT [5], analogues bearing a variety of substituents other than phenylthio or phenylselenyl at the 6 position [6, 7] and tricyclic analogues (**III** in table XII) of HEPT [8].

TSAR treatment

The TSAR statistical model (equation (2)) describing the activity of HEPT derivatives was slightly modified and tested against the structurally different 6-phenylselenyl analogues of HEPT [5] **77–82**. The only parameter which had to be changed was the intercept of the correlation line. This parameter takes into account the influence of the core structure and instead of the original value of –62.43 derived for the set of compounds bearing the thiophenyl fragment, we obtained a result for the intercept of –65.73 when a statistical model for the phenylselenyl analogues was built (r = 0.853, s = 0.157, F = 30.64). This model was obtained by the method previously described and by considering for the linear regression two conformers for the hydroxyethoxymethyl substituent – the most and the least stable ones. The results for the tests are presented in table VIII. The failure of our model to correctly predict the activities of the 5-halogenated 6-phenylselenyl compounds **79–81** can probably be explained by the fact that no 5-halogenated derivatives were considered when building the model. Conversely, the predicted relative activities for the two compounds bearing either a hydrogen atom or a methyl group at the 5 position are in good agreement with the observed values.

The capacity of the TSAR model to predict the activity of 6-benzyl analogues of HEPT **83–95** was also studied (table IX). A linear regression model keeping the original form of equation [2] was derived (r = 0.918, s = 0.211, F = 73.75). The obtained new intercept was –63.79. The activities of the 6-benzyl analogues of HEPT [7] are well reproduced and the two compounds which are close analogues of those belonging to the **B** group (**86** and **90**) have their activities again underestimated. This is additional proof of the existence of at least two general classes of HEPT analogues with different behavior toward the HIV-1 reverse transcriptase.

The availability of data concerning the cytotoxicity of this class of 6-benzyl analogues of the HEPT made possible the testing of our TSAR-derived model for the cytotoxicity. A linear regression model keeping the original form of equation (8) was obtained with only the intercept of the regression line changed to

Table VIII. Predicted and observed activities of 6-phenylselenyl HEPT analogues bearing a 1-[2-(hydroxyethoxy)methyl] group.

Compound	*Nucleobase*	*R^1*		*Experiment*[b]		*TSAR*		*CATALYST*	
		Conformer[a]		*EC_{50}*	*$-LogEC_{50}$*	*$-LogEC_{50}$ fit*[c]		*R^1 conformer*[d]	*EC_{50} fit*[e]
		Low	*High*			*Low*	*High*		
77	6-(Phenylselenyl)uracil	(1a)	(1i)	13.0	−1.139	0.719	0.783	(1c)	8.7
78	6-(Phenylselenyl)thymine H	(1a)	(1i)	0.96	0.018	0.972	0.984	(1a)	1.0
79	5-Fluoro-6-(phenylselenyl)uracil	(1a)	(1i)	2.0	−0.301	16	16	(1a)	0.031
80	5-Chloro-6-(phenylselenyl)uracil	(1a)	(1i)	3.1	−0.491	16	16	(1b)	0.12
81	5-Bromo-(phenylselenyl)uracil	(1a)	(1i)	3.7	−0.568	16	16	(1c)	0.053
82	6-(Phenylselenyl)-2-thiothymine	(1a)	(1i)	2.8	−0.447	16	16	(1a)	0.074

[a]Conformers of substituent R^1 used to predict the anti-HIV-1 activity of 6-phenylselenyl HEPT analogues with equation (2). The term low corresponds to the lowest-energy conformers and the term high the highest-energy conformers for each substituent R^1 in the 0–10 kcal/mol range. Labels (1a) through (1i) refer to figure 1, labels (2a) through (2c) to figure 2 and labels (3a) through (3e) to figure 3. For substituents other than those displayed in these figures the energies in kcal/mol for the most and least stable conformers in the 0–10 kcal/mol range are provided. [b]Experimental EC_{50} and $-\log EC_{50}$ values. [c]$-LogEC_{50}$ values predicted with the general model with non-linear corrections (equation (2)) for the lowest-energy and the highest-energy conformers in the range 0–10 kcal/mol. [d]Conformation of the R^1 substituent in the CATALYST-derived conformer leading to the best fit for the anti-HIV-1 activity of 6-benzyl HEPT analogues. [e]Best fit for the anti-HIV-1 activity of 6-benzyl HEPT analogues obtained using the most statistically significant hypothesis, C, in table III.

−6.380 (r = 0.811, s = 0.080, F = 12.85). The results are given in table X. The predicted values for the cytoxicities are in fair agreement with the experimental ones and our model has proved to be able to sort out the most and least cytotoxic compounds.

CATALYST treatment

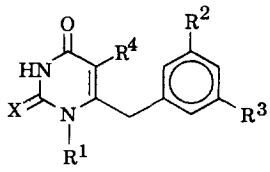
The best CATALYST hypothesis (table III), used to model the activity of HEPT derivatives was tested against the class of the 6-phenylselenyl analogues of HEPT. The results are presented in table VIII. For each molecule, the conformers were generated according to the methodology previously described. The CATALYST model did not fully succeed in predicting the activities of the set of the 5-halogenated compounds, but afforded better results than TSAR.

The same CATALYST model was also tested against the 6-benzyl HEPT analogues **II** and a reasonable agreement between the predicted activities and the experimentally observed ones was obtained. The results are presented in table IX. In table X we present the results obtained from the testing of the CATALYST-derived model for cytotoxicity against this class of 6-benzyl HEPT analogues.

Contrary to the TSAR treatment, restricted to compounds having a common core moiety, a CATALYST hypothesis can be used to locate a pharmacophore among any series of molecules. Our CATALYST model was thus tested against two additional classes of molecules. The first was taken from [6]. This series is a heterogeneous one (table XI), but it is characterized by the presence of some compounds bearing unsaturated hydrocarbon fragments at the 5 position. The CATALYST model has proved to be able to isolate the most active compound in the series and to reproduce reasonably well the experimentally observed activities. The same conclusion can be formulated concerning the performances of the cytotoxicity model against this same series of molecules. The results are summarized in table XI.

Finally, the CATALYST-derived model for the activity was tested against a series of thoroughly structurally different HEPT analogues, the tricyclic derivatives **III** [8] (table XII). While the numerical values obtained for the predicted activities are somewhat different from the experimentally obtained ones, the CATALYST model has successfully reproduced the relative order of activity, as can be seen from table XII. The difference in the numerical values of the predicted and the observed activities is in part due to the fact that the measured values are IC_{50}, while our model was obtained on the basis of EC_{50} values.

Table IX. Numerical data concerning in vitro anti-HIV-1 activities (EC_{50}/μM) of 6-benzyl HEPT analogues: experimental and estimated values, obtained by TSAR and CATALYST analysis.



II

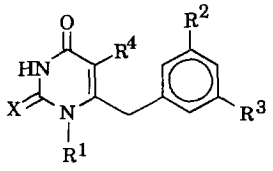
Compound	*X*	*R^1 = CH_2OR*	*R^2*	*R^3*	*R^4*	*R^1*		*Experiment*[b]		*TSAR*		*CATALYST*	
		R				*Conformer*[a]		*EC_{50}*	*$-LogEC_{50}$*	*$-LogEC_{50}$ fit*[c]		*R^1 conformer*[d]	*EC_{50} fit*[e]
						Low	*High*			*Low*	*High*		
83	O	CH_2CH_2OH	H	H	Me	(1a)	(1i)	23	−1.362	−1.386	−1.174	(1a)	8.7
84	O	CH_2CH_2OH	H	H	Et	(1a)	(1i)	0.35	0.456	0.624	0.440	(1b)	1.0
85	O	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	0.013	1.886	1.921	1.736	(1a)	0.031
86	O	Et	H	H	Et	(2a)	(2c)	0.041	1.387	1.280	1.534	(2b)	0.12
87	O	Et	Me	Me	Et	(2a)	(2c)	0.0016	2.796	2.576	2.831	(2c)	0.053
88	O	CH_2CH_2OH	H	H	*i*-Pr	(1a)	(1i)	0.063	1.201	1.328	1.143	(1c)	0.074
89	O	CH_2CH_2OH	Me	Me	*i*-Pr	(1a)	(1i)	0.0027	2.569	2.625	2.440	(1e)	0.089
90	O	Et	H	H	*i*-Pr	(2a)	(2c)	0.0042	2.377	1.984	2.238	(2c)	0.078
91	O	Et	Me	Me	*i*-Pr	(2a)	(2c)	0.0006	3.222	3.280	3.535	(2c)	0.0009
92	O	R^1 = Bu	H	H	Et	1.787	2.730	0.21	0.678	0.632	0.724	(1a)	0.089
93	O	R^1 = Bu	H	H	*i*-Pr	1.787	2.730	0.042	1.377	1.336	1.428	(1a)	0.01
94	O	R^1 = $MeOCH_2CH_2$	H	H	Et	(1a)	(1i)	0.25	0.602	0.739	0.573	(1e)	0.85
95	O	R^1 = $MeOCH_2CH_2$	H	H	i-Pr	(1a)	(1i)	0.052	1.284	1.443	1.277	(1a)	1.8

[a]Conformers of substituent R^1 used to predict the anti-HIV-1 activity of 6-benzyl HEPT analogues with equation (2). The term low corresponds to the lowest-energy conformers and the term high to the highest-energy conformers for each substituent R^1 in the 0–10 kcal/mol range. Labels (1a) through (1i) refer to figure 1, labels (2a) through (2c) to figure 2 and labels (3a) through (3e) to figure 3. For substituents other than those displayed in these figures the energies in kcal/mol for the most and least stable conformers in the 0–10 kcal/mol range are provided. [b]Experimental EC_{50} and $-logEC_{50}$ values. [c]$-LogEC_{50}$ values predicted with the general model with non-linear corrections (equation [2]) for the lowest-energy and the highest-energy conformers in the range 0–10 kcal/mol. [d] Conformation of the R^1 substituent in the CATALYST-derived conformer leading to the best fit for the anti-HIV-1 activity of 6-benzyl HEPT analogues. [e] Best fit for the anti-HIV-1 activity of 6-benzyl HEPT analogues obtained using the most statistically significant hypothesis, C, in table III.

Conclusion

In this series, two fundamentally different drug design approaches, a Hansch's type, connectivity oriented method (TSAR) and a pharmacophore search procedure based on 3D spatial relationships (CATALYST) led to convergent results. The models obtained by the connectivity oriented method (TSAR) were improved by inclusion of non-linear effects derived through a neural network approach. An acceptable predictive power was proved by these treatments when applied to compounds not included in the original QSAR

Table X. Numerical data concerning in vitro cytotoxicities (CC_{50}/μM) of 6-benzyl HEPT analogues: experimental and estimated values, obtained by TSAR and CATALYST analysis.

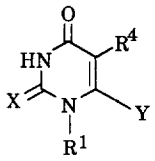


II

Compound	*X*	*R^1 = CH_2OR*	*R^2*	*R^3*	*R^4*	*R^1*		*Experiment*[b]		*TSAR*		*CATALYST*	
		R				*Conformer*[a]		*CC_{50}*	*$-LogCC_{50}$*	*$-LogCC_{50}$ fit*[c]		*CC_{50} fit*[d]	*Fitted conformer*[e]
						Low	*High*			*Low*	*High*		
83	O	CH_2CH_2OH	H	H	Me	(1a)	(1i)	352	–2.547	–2.695	–2.695	310	10
84	O	CH_2CH_2OH	H	H	Et	(1a)	(1i)	391	–2.592	–2.695	–2.695	110	15
85	O	CH_2CH_2OH	Me	Me	Et	(1a)	(1i)	281	–2.449	–2.484	–2.484	110	6
86	O	Et	H	H	Et	(2a)	(2c)	245	–2.389	–2.490	–2.399	100	26
87	O	Et	Me	Me	Et	(2a)	(2c)	207	–2.316	–2.280	–2.189	88	3
88	O	CH_2CH_2OH	H	H	*i*-Pr	(1a)	(1i)	295	–2.470	–2.347	–2.347	43	11
89	O	CH_2CH_2OH	Me	Me	*i*-Pr	(1a)	(1i)	221	–2.344	–2.137	–2.137	81	6
90	O	Et	H	H	*i*-Pr	(2a)	(2c)	186	–2.270	–2.143	–2.052	79	22
91	O	Et	Me	Me	*i*-Pr	(2a)	(2c)	43	–1.633	–1.933	–1.841	76	19
92	O	R^1 = Bu	H	H	Et	(1a)	(1i)	>500	–	–2.397	–2.392	65	14
93	O	R^1 = Bu	H	H	*i*-Pr	(1a)	(1i)	58	–1.763	–2.049	–2.045	120	7
94	O	R^1 = $MeOCH_2CH_2$	H	H	Et	(1a)	(1i)	362	–2.559	–2.603	–2.603	130	18
95	O	R^1 = $MeOCH_2CH_2$	H	H	*i*-Pr	(1a)	(1i)	195	–2.290	–2.255	–2.255	130	8

[a]Conformers of substituent R^1 used to predict the cytotoxicity of 6-benzyl HEPT analogues with equation (8). The term low corresponds to the lowest-energy conformers and the term high to the highest-energy conformers for each substituent R^1 in the 0–10 kcal/mol range. Labels (1a) through (1i) refer to figure 1, labels (2a) through (2c) to figure 2 and labels (3a) through (3e) to figure 3. For substituents other than those displayed in these figures, the energies in kcal/mol for the most and least stable conformers in the 0–10 kcal/mol range are provided. [b]Experimental CC_{50} and $-logCC_{50}$ values. [c] $-LogCC_{50}$ values predicted with the general model with non-linear corrections (equation (8)) for the lowest-energy and the highest-energy conformers in the range 0–10 kcal/mol. [d]Best fit for the cytotoxicities of 6-benzyl HEPT analogues obtained by using the most statistically significant hypothesis (table VII). [e]Conformation of the R^1 substituent in the CATALYST-derived conformer leading to the best fit for the cytotoxicities of 6-benzyl HEPT analogues.

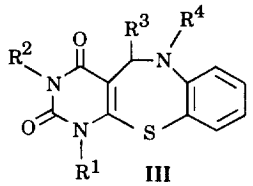
Table XI. Structures and in vitro anti-HIV-1 activities and cytotoxicities of the HEPT analogues described in [6] and used as tests.



Compound	*X*	*Y*	*R^1*	*R^4*	*Conformers*[a]	*Experiment*[b]		*CATALYST*			
						EC_{50}	*CC_{50}*	*EC_{50}*[c]	*Fitted conformer*[d]	*CC_{50}*[e]	*Fitted conformer*[f]
96	O	SBu	$CH_2OCH_2CH_2OH$	Me	41	130	>250	5.1	3	230	3
97	O	SC_6H_{11}	$CH_2OCH_2CH_2OH$	Me	41	8.2	664	3.7	9	120	2
98	O	OPh	$CH_2OCH_2CH_2OH$	Me	41	85	345	1.8	20	290	17
99	O	CH_2Ph	$CH_2OCH_2CH_2OH$	Me	41	23	352	8.7	9	310	9
100	O	SPh	$CH_2OCH_2CH_2OH$	I	41	3.6	20	2.4	3	23	11
101	O	SPh	$CH_2OCH_2CH_2OH$	$CH{=}CPh_2$	41	0.84	21	0.022	0	81	9
102	O	SPh	$CH_2OCH_2CH_2OH$	CH=CHPh (Z)	41	6.0	95	1.2	9	110	10
103	O	SPh	$CH_2OCH_2CH_2OH$	$CH{=}CH_2$	41	1.1	76	1.9	6	130	15

[a]Total number of conformers in the 0–30 kcal/mol range generated by CATALYST. [b]Experimental EC_{50} and CC_{50}. [c]CATALYST derived best fit for the anti-HIV-1 activity of HEPT analogues from the most statistically significant hypothesis, C, in table III. [d]Number of the conformer in the 0–30 kcal/mol range leading to the best mapping onto the CATALYST's activity model. [e]CATALYST derived best fit for the cytotoxicity of HEPT analogues using the most statistically significant hypothesis (table VII). [f]Numbers of conformers in the 0–30 kcal/mol range leading to the best mapping onto the CATALYST's cytotoxicity model. The conformers are labelled with numbers directly proportional to their energy in the 0–30 kcal/mol range.

Table XII. Structures and in vitro anti-HIV-1 activities of the pyrimidobenzothiazepine HEPT analogues **III** described in [8] and used as tests.



III

Compound	*R^1*	*R^2*	*R^3*	*R^4*	*Conformers*[a]	*Experiment IC_{50}*[b]	*CATALYST*	
							EC_{50}[c]	*Fitted conformer*[d]
104	$AcO(CH_2)_2OCH_2$	H	H	Me	41	14.7	8.5	6
105	$AcO(CH_2)_2OCH_2$	H	H	H	41	11.5	4.2	18
106	$CH_3CH_2OCH_2$	H	H	H	41	50.2	83.0	3
107	$EtO(CH_2)_2OCH_2$	H	H	H	41	0.64	1.0	8

[a]Total number of conformers in the 0–30 kcal/mol range generated by CATALYST. [b]Experimental IC_{50} values. [c]CATALYST derived fit for anti-HIV-1 activity of pyrimidobenzothiazepine HEPT analogues from the most statistically significant hypothesis, C, in table III. [d]Number of the conformer in the 0–30 kcal/mol range leading to the best mapping onto the activity model derived from CATALYST.

study and structurally different from the original set. Both methods demonstrated the chemotherapeutical heterogeneity of the HEPT family, thus confirming earlier biochemical results. The basic factor determining this heterogeneity of HEPT analogues as inhibitors of the HIV-1 reverse transcriptase enzyme was identified to be the folding features of the substituent in position 1 of the nucleobase.

Acknowledgments

This work was generously supported by the Swiss Federal Office for Public Health and the Swiss National Research Foundation (Grants # 3139-037156, 20-37626.93, 20-43552.95 and 20-41830.94). The CATALYST software was made available through a University Research partnership.

Supporting information available

Several additional graphs, correlation matrices for the decriptors used in equations [2] and [8] as well as the dependence plots depicting the relationships between the neural network outputs, $-\log EC_{50}$ and $-\log CC_{50}$, and the input parameters used in equations (2) and (8) respectively are available upon request from the Editor-in-Chief.

References

1 Miyasaka T, Tanaka H, Baba M et al (1989) *J Med Chem* 32, 2507–2509
2 Baba M, Tanaka H, De Clercq E et al (1989) *Biochem Biophys Res Commun* 165, 1375–1381
3 Tanaka H, Takashima H, Ubasawa M et al (1992) *J Med Chem* 35, 227–245
4 Tanaka H, Takashima H, Ubasawa M et al (1992) *J Med Chem* 35, 4713–4719
5 Goudgaon NM, Schinazi RF (1991) *J Med Chem* 34, 3305–3309
6 Tanaka H, Baba M, Hayakawa H et al (1991) *J Med Chem* 34, 349–357
7 Tanaka H, Takashima H, Ubasawa M et al (1995) *J Med Chem* 38, 2860–2865
8 Maruenda H, Johnson F (1995) *J Med Chem* 38, 2145–2151
9 *TSAR Version 2.02 for Silicon Graphics Platforms*, Oxford Molecular Ltd, 1993, and references included in the documentation
10 *PIMMS Version 2.02 for Silicon Graphics Platforms*, Oxford Molecular Ltd, 1993, and references included in the documentation
11 *COBRA Version 2.02 for Silicon Graphics Platforms*, Oxford Molecular Ltd, 1993, and references included in the documentation
12 Hansch C, Leo A, Hoekman D (1995) *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, ACS Professional Reference Book, American Chemical Society, Washington, DC 1995
13 Zupan J, Gasteiger J (1993) *Neural Networks for Chemists: An Introduction*, VCH, Weinheim
14 So SS, Richards WG (1992) *J Med Chem* 35, 3201–3206
15 *CATALYST™ Version 2.2 for Silicon Graphics Platforms*, Molecular Simulations Inc, 1994, and references included in the documentation
16 Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) *J Comp Chem* 4, 187–217
17 *CatHypo Version 2.2 for Silicon Graphics Platforms*, Molecular Simulations Inc
18 *ASP Version 2.02 for Silicon Graphics Platforms*, Oxford Molecular Ltd, 1993, and references included in the documentation
19 Debyser Z, Pauwels R, Baba M, Desmyter J, De Clercq E (1991) *Mol Pharmacol* 41, 963–968
20 Merluzzi VJ, Hargrave KD, Labadia M et al (1990) *Science* 250, 1411–1413