J Mol Model (2002) 8:1–7
DOI 10.1007/s00894-001-0054-9

ORIGINAL PAPER

Halima Bazoui · Mohamed Zahouily · Saïd Sebti
Saïd Boulajaaj · Driss Zakarya

Structure–cytotoxicity relationships for a series of HEPT derivatives

Received: 6 April 2001 / Accepted: 14 September 2001 / Published online: 15 January 2002

Abstract Structure–cytotoxicity relationships were studied for a series of 90 HEPT derivatives by means of multiple linear regression (MLR) and artificial neural network (ANN) techniques. The values of $\log(1/CC_{50})$ (CC_{50}=cytotoxic dose of compound required to reduce the proliferation of normal uninfected MT-4 cells by 50%) of the studied compounds were correlated with the descriptors encoding the chemical structures. Using the pertinent descriptors revealed by the regression analysis, a correlation coefficient of 0.935 (s=0.149) for the training set (n=81) was obtained for the ANN model with a 5–6–1 configuration. The results obtained from this study indicate that the cytotoxicity of HEPT derivatives is strongly dependent on hydrophobic factors, mainly $\log P(R_1)$, and dependent on the steric factors, especially $\Sigma MW(R_3+R_4)$. Comparison of the descriptors' contribution obtained in MLR and ANN analysis shows that the contribution of some of the descriptors to cytotoxicity may be non-linear.

Keywords Artificial neural network · Descriptors' contribution · HEPT derivatives · Multiple linear regression · Structure–cytotoxicity relationships

Electronic supplementary material to this paper can be obtained by using the Springer LINK server located at http://dx.doi.org/10.1007/ s00894-001-0054-9

H. Bazoui · M. Zahouily (✉) · D. Zakarya
UFR Chimie Appliquée,
Laboratoire de Synthèse Organique
et Traitement de l'Information Chimique,
Département de Chimie, Faculté des Sciences et Techniques,
B.P. 146, Mohammedia 20650, Morocco
e-mail: zahouily@voila.fr
Tel.: +212-23-314705/08, Fax: +212-3-315353

H. Bazoui · S. Sebti · S. Boulajaaj
Laboratoire de Chimie Organique Appliquée et Catalyse (LCOAC),
Faculté des Sciences Ben M'Sik, B.P. 7955, Sidi Othmane,
Casablanca, Morocco

Introduction

Infection with human immunodeficiency virus type-1 (HIV-1) causes progressive destruction of the immune system, which ultimately results in acquired immunodeficiency syndrome (AIDS). An essential step in the life cycle of HIV-1 is reverse transcription of the viral RNA genome to produce a double-stranded DNA copy. This process is mediated by the virally encoded reverse transcriptase (RT). [1]

Since RT is essential for virus replication and has no closely related identified cellular homologue, it has been the prime target for antiviral therapy against AIDS. [2] Various compounds have been reported as potent and selective inhibitors of HIV.

Currently, dideoxynucleoside including 3'-deoxy-3'-azidothymidine (Zidovidine or AZT), [3] 2',3'-dideoxyinosine (ddI), 2',3'-dideoxycytidine (ddC); and, most recently, 2',3'-didehydro-3'-deoxythymine(d4T) are approved for use in the treatment of HIV infection; these prodrugs are thought to produce their antiviral effects through the inhibition of HIV reverse transcriptase and viral DNA chain termination via their triphosphate metabolites. [4] However, it is found that the treatment of some of these nucleoside inhibitors such as AZT is sometimes associated with considerable site effects such as bone marrow suppression.

Another class of HIV-RT inhibitors, non-nucleoside inhibitors (NNRTIs), have been identified as potent and highly specific inhibitors of HIV-1 replication. They interact with HIV-1 RT at a nonsubstrate binding site. Examples are HEPT, TIBO, nevirapine, pyridinone, TSAO, α-APA, and quinoxaline. [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]

Among these NNRTIs, HEPT has proved to be a potent and selective inhibitor of HIV-1. Other animal retroviruses and even HIV-2 are totally unaffected by this compound. Design of new HEPT derivatives requires a more detailed knowledge of the mechanism of RT inhibition by this class of compounds and their cytotoxicity.

Assessment of QSAR for cytotoxicity, which constitutes the second component of the selectivity index, rep-

Table 1 Chemical structure of HEPT derivatives and experimental cytotoxicity values

Molecule No.	R_1	R_2	R_3	R_4	X	Y	CC_{50} (μM)[a]
Training set							
01	$CH_2OCH_2CH_2OH$	Me	H	Me	O	S	420
02	$CH_2OCH_2CH_2OH$	Et	H	Me	O	S	181
03	$CH_2OCH_2CH_2OH$	*t*-Bu	H	Me	O	S	75
04	$CH_2OCH_2CH_2OH$	CH_2OH	H	Me	O	S	292
05	$CH_2OCH_2CH_2OH$	CF_3	H	Me	O	S	196
06	$CH_2OCH_2CH_2OH$	F	H	Me	O	S	282
07	$CH_2OCH_2CH_2OH$	Cl	H	Me	O	S	210
08	$CH_2OCH_2CH_2OH$	Br	H	Me	O	S	141
09	$CH_2OCH_2CH_2OH$	I	H	Me	O	S	106
10	$CH_2OCH_2CH_2OH$	NO_2	H	Me	O	S	170
11	$CH_2OCH_2CH_2OH$	OH	H	Me	O	S	446
12	$CH_2OCH_2CH_2OH$	Me	Me	Me	O	S	243
13	$CH_2OCH_2CH_2OH$	Cl	Cl	Me	O	S	130
14	$CH_2OCH_2CH_2OH$	Me	Me	Me	S	S	172
15	$CH_2OCH_2CH_2OH$	COOMe	H	Me	O	S	221
16	$CH_2OCH_2CH_2OH$	COMe	H	Me	O	S	228
17	$CH_2OCH_2CH_2OH$	COOH	H	Me	O	S	352
18	$CH_2OCH_2CH_2OH$	$CONH_2$	H	Me	O	S	306
19	$CH_2OCH_2CH_2OH$	CN	H	Me	O	S	234
20	$CH_2OCH_2CH_2OH$	H	H	CH_2=CH–CH_2	O	S	183
21	$CH_2OCH_2CH_2OH$	H	H	COOMe	O	S	6,6
22	$CH_2OCH_2CH_2OH$	H	H	COOHNPh	O	S	18
23	$CH_2OCH_2CH_2OH$	H	H	Et	S	S	148
24	$CH_2OCH_2CH_2OH$	H	H	Pr	S	S	230
25	$CH_2OCH_2CH_2OH$	H	H	*i*-Pr	S	S	400
26	$CH_2OCH_2CH_2OH$	Me	Me	*i*-Pr	S	S	52
27	$CH_2OCH_2CH_2OH$	Cl	Cl	Et	S	S	64
28	$CH_2OCH_2CH_2OH$	H	H	Et	O	S	40
29	$CH_2OCH_2CH_2OH$	H	H	Pr	O	S	244
30	$CH_2OCH_2CH_2OH$	H	H	*i*-Pr	O	S	231
31	$CH_2OCH_2CH_2OH$	Me	Me	Et	O	S	149
32	$CH_2OCH_2CH_2OH$	Me	Me	*i*-Pr	O	S	128
33	$CH_2OCH_2CH_2OH$	Cl	Cl	Et	O	S	51
34	$CH_2OCH_2CH_2OH$	H	H	Me	S	S	123
35	$CH_2OCH_2CH_2OMe$	H	H	Me	O	S	299
36	$CH_2OCH_2CH_2O$–*n*-C_5H_{11}	H	H	Me	O	S	55
37	$CH_2OCH_2CH_2OCH_2Ph$	H	H	Me	O	S	45
38	CH_2OMe	H	H	Me	O	S	244
39	CH_2OEt	H	H	Me	O	S	231
40	CH_2OPr	H	H	Me	O	S	147
41	CH_2OBu	H	H	Me	O	S	83
42	$CH_2OCH_2CH_2SiMe$	H	H	Me	O	S	32
43	CH_2OCH_2Ph	H	H	Me	O	S	95
44	CH_2OEt	H	H	Et	S	S	81
45	CH_2OEt	Cl	Cl	Et	S	S	45
46	CH_2OEt	H	H	*c*-Pr	S	S	46
47	CH_2OEt	H	H	Et	O	S	161
48	CH_2OEt	Cl	Cl	Et	O	S	45
49	CH_2O–*i*-Pr	H	H	Et	O	S	143
50	CH_2OCH_2-*c*-Hex	H	H	Et	O	S	17
51	CH_2OCH_2Ph	H	H	Et	O	S	34
52	$CH_2OCH_2CH_2Ph$	H	H	Et	O	S	38
53	CH_2OEt	H	H	*i*-Pr	O	S	106
54	H	H	H	Me	O	S	250
55	Me	H	H	Me	O	S	150
56	Et	H	H	Me	O	S	94
57	Bu	H	H	Me	O	S	89
58	$CH_2OCH_2CH_2OH$	H	H	Me	O	CH_2	352
59	$CH_2OCH_2CH_2OH$	H	H	Et	O	CH_2	391
60	$CH_2OCH_2CH_2OH$	Me	Me	Et	O	CH_2	281
61	CH_2OEt	Me	Me	Et	O	CH_2	245
62	CH_2OEt	H	H	Et	O	CH_2	207
63	$CH_2OCH_2CH_2OH$	H	H	*i*-Pr	O	CH_2	295
64	$CH_2OCH_2CH_2OH$	Me	Me	*i*-Pr	O	CH_2	221
65	CH_2OEt	H	H	*i*-Pr	O	CH_2	186

Table 1 (continued)

Molecule No.	R_1	R_2	R_3	R_4	X	Y	CC_{50} (μM)[a]
66	CH_2OEt	Me	Me	*i*-Pr	O	CH_2	43
67	Bu	H	H	*i*-Pr	O	CH_2	58
68	CH_2 CH_2OMe	H	H	Et	O	CH_2	362
69	CH_2CH_2OMe	H	H	*i*-Pr	O	CH_2	195
70	$CH_2OCH_2CH_2OH$	H	H	Me	O	O	345
71	$CH_2OCH_2CH_2OH$	H	H	I	O	S	20
72	$CH_2OCH_2CH_2OH$	H	H	$CH{=}CPh_2$	O	S	21
73	$CH_2OCH_2CH_2OH$	H	H	CH=CHPh	O	S	95
74	CH_2SCH_3	H	H	Et	O	CH_2	32
75	$CH_2SCH_2CH_3$	H	H	Et	O	CH_2	37
76	CH_2SCH_3	Me	Me	Et	O	CH_2	52
77	$CH_2SCH_2CH_3$	Me	Me	Et	O	CH_2	68
78	CH_2SCH_3	H	H	*i*-Pr	O	CH_2	37
79	$CH_2SCH_2CH_3$	H	H	*i*-Pr	O	CH_2	37
80	$CH_2OCH_2CH_2OH$	H	H	$COCH(Me)_2$	O	S	12
81	$CH_2OCH_2CH_2OH$	H	H	COPh	O	S	13
Prediction set							
82	$CH_2OCH_2CH_2OH$	Me	Me	Et	S	S	277
83	$CH_2OCH_2CH_2OH$	H	H	Me	O	S	743
84	CH_2O–*c*-Hex	H	H	Et	S	S	223
85	CH2OEt	H	H	*c*-Pr	O	S	224
86	$CH_2OCH_2CH_2OH$	H	H	$CH{=}CH_2$	O	S	76
87	$CH_2OCH_2CH_2OH$	H	H	CH_2Ph	O	S	23
88	$CH_2OCH_2CH_2OH$	H	H	C≡CMe	O	S	19
89	$CH_2OCH_2CH_2OH$	H	H	C≡CPh	O	S	3,4
90	$CH_2OCH_2CH_2OH$	H	H	C≡CH	O	S	18

[a] CC_{50}: cytotoxic dose of compound required to reduce the proliferation of normal uninfected MT-4 cells by 50%

resents one of the most effective computational approaches for inspection of this component.

Experimental data (CC_{50}) are generally available only as the upper limit of the non-cytotoxic concentration. Moreover, they encompass a narrower concentration range than the EC_{50} data. Notwithstanding these limitations, both earlier approaches [multiple linear regression (MLR) and an artificial neural network (ANN)analysis] have been applied in this work to SAR cytotoxicity studies.

Both ANN and MLR techniques were used for modeling the observed cytotoxicity of HEPT derivatives. The adequacy of the models developed was examined by means of the prediction of the cytotoxicity of nine HEPT derivatives, which represent 10% of our total subsets (90 molecules).

The results obtained by 2D-QSAR approaches will be analyzed in order to find the best parameters responsible for the interactions between active molecules and the possible interaction site.

Material and methods

Biological data

All compounds with known cytotoxic concentrations were taken from various published studies. [15, 16, 17, 18, 19] The data set consists of 90 compounds, for which the in vitro cytotoxicity was measured in MT-4 cell cultures.

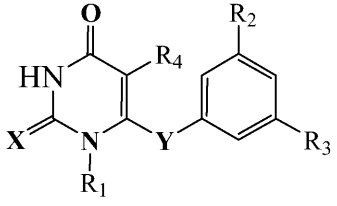

Fig. 1 General structure of HEPT derivatives

The $\log(1/CC_{50})$ values were used as dependent variable. CC_{50} (μM) represents the molar concentrations of drug required to reduce the viability of 50% of the mock-infected MT-4 cells.

The chemical structures considered in the training and prediction sets are given in Table 1 (see Fig. 1 for the general structure).

Molecular descriptors

A set of common molecular descriptors related to physicochemical, electronic and geometric properties of the molecules was used for this study. As all the compounds studied have a common skeleton, we found it judicious to describe the molecule by means of properties of the substituents (R_1, R_2, R_3 and R_4) attached to the basic skeleton. Determination of the pertinent properties for a given substituent may be useful for evaluating local interactions between the molecule and the receptor site.

Moreover, we tried to take into account properties of the molecule such as its molecular weight, size, height etc. This is justified by the fact that, before their possible interaction with a given receptor site, the molecules must be transported through many liquid layers and have correct general dimensions for site access.

Molecular properties used for each substituent and the whole molecule, were:

- Size and shape described by means of van der Waals volume (*V*) and surface (*S*). [20]
- Molecular dimensions (length, width and height). Length (*L*) is the distance along the screen *x*-axis between the left- and rightmost atoms plus their van der Waals radii. Width (*W*) is the distance along the screen *y*-axis between the top and bottommost atoms plus their van der Waals radii. Height (*H*) is the distance along the screen *z*-axis between the nearest and farthest atoms plus their van der Waals radii. [20]
- The ratios *V*/*L*, *V*/*W*, *W*/*H* were also calculated.
- log *P*, the partition coefficient between octanol and water. [20]
- Molar refractivity (MR). [20]
- Molecular weight (MW). [20]
- Ovality estimation (*O*), for each substituent was that given by Bodor [21]

$$O = S/(4\pi K) \tag{1a}$$

$$K = (3V/4\pi)^{2/3} \tag{1b}$$

59 parameters were calculated for each compound.

Statistical methods

Multiple linear regression

This method [22] was used to generate linear models between the cytotoxicity and the molecular descriptors.

Because of the large number of descriptors considered, a stepwise MLR procedure based on the forward-selection and backward-elimination methods was used to select the powerful variable descriptors.

In order to avoid all difficulties in the interpretation of the resulting models, pairs of variables with a correlation coefficient greater than 0.7 were classified as intercorrelated, and only one of these was included in the screened model.

The validity of the model was proven by the multiple correlation coefficient (*r*), the standard deviation (*s*) and the *F*-test value. The reliability of the model was indicated in terms of predictive q^2.

Artificial neural network

As biological phenomena are considered non-linear by nature, it therefore appears very interesting to study the present series of compounds with the ANN technique, [23] in order to discover the possible existence of non-linear relationships between cytotoxicity and molecular descriptors that appeared pertinent for the linear model.

The ANN was trained by the back-propagation (BP) of errors algorithm [24] and had the following architecture:

- An input layer including the pertinent descriptors of MLR.
- A hidden layer for which the ratio of the number of data points in the training set to the number of variables controlled by the network, ρ, is critical to the predictive power of the neural net. The range & 1.8lt;ρ<2.2 [ρ=(number of data points in the training set)/(number of adjustable weights controlled by the network)], [25] was used as a guideline of the acceptable number of neurons in the hidden layer. It is claimed that, for ρ≤1.0, the network simply memorizes the data, whereas for ρ≥3.0, the network loses its ability to generalize.
- An output layer of one neuron, representing the cytotoxic concentration. The input and output values were normalized.

After this step, the learning rate was varied from 0.01 to 0.9 and for each learning rate the momentum was examined from 0.1 to 0.9.

Finally, the number of the neurons at the hidden layer with the use of optimized momentum and learning rate was determined.

Results and discussion

Multiple linear regression analysis

MLR analysis was performed on the compounds described in Table 1; we have included all 81 molecules of the training set for the model generation.

After collecting the data, we submitted all parameters to regression; a few suitable models were obtained. The best model is shown in Eq. (2):

$$\begin{aligned}\log(I/\mathrm{CC}_{50}) \;=\; & (2.592 \pm 0.319) \\ & + (0.249 \pm 0.026)\log P(\mathrm{R}_1) \\ & + (0.024 \pm 0.007)\,\mathrm{MR}(\mathrm{R}_2) \\ & + (0.014 \pm 0.001)\,\Sigma\mathrm{MW}(\mathrm{R}_3 + \mathrm{R}_4) \\ & - (0.482 \pm 0.066)\log P(\mathrm{R}_4) \\ & + (0.241 \pm 0.055)\,H(z)\end{aligned} \tag{2}$$

$$n = 81 \quad s = 0.204 \quad r = 0.884 \quad F(5,75) = 53.839$$

This equation shows that five descriptors appear in the model. These descriptors consist of log *P*, which is a hydrophobic descriptor for substituents R_1 and R_4; MR(R_2), which can be considered as both an electronic and a steric descriptor, and the ΣMW(R_3+R_4) parameter, which is a steric descriptor. The parameter *H*(*z*) is a descriptor that expresses the dimension of molecule along the *z*-axis.

It seems appropriate to compare these results with those presented previously by Tronchet et al. [26] The statistical quality of our equation is slightly superior to their results, with a lower number of descriptors (five parameters compared with 11 parameters) and a larger data set.

According to the results above, log(1/CC_{50}) depends positively on log *P*(R_1), as reflected in molecules **38**, **39**–**40**, **41**–**42**, **43**–**33**, **48**.

The molecules **1**, **2**, **3** and **15**, **16**, **17**, **18** unambiguously show that log(1/CC_{50}) increases with the molar refraction of substituent R_2.

We also observed the positive influence of ΣMW(R_3 + R_4) as seen in the case of the set of compounds **1**, **12**–**31**, **32**–**21**, **22**–**44**, **46**–**68**, **69**–**61**, **66**–**72**, **73**. In the same way, increasing *H*(*z*) leads to an increase in log (1/CC_{50}).

Table 2 Correlation matrix

	log 1/(CC_{50})	log $P(R_1)$	RM(R_2)	ΣMW(R_3+R_4)	log $P(R_4)$	$H(z)$
log 1/(CC_{50})	1.000					
log $P(R_1)$	0.364	1.000				
RM(R_2)	–0.148	–0.360	1.000			
ΣMW(R_3+R_4)	0.564	–0.163	–0.235	1.000		
log$P(R_4)$	0.054	–0.009	–0.172	0.587	1.000	
$H(z)$	0.296	0.082	0.006	0.016	0.015	1.000

The cytotoxicity of HEPT derivatives increases with decreasing log $P(R_4)$ values, as shown if one analyzes the cytotoxicity of molecules **24**, **25** and **29**, **30**.

It is noteworthy that there is no significant intercorrelation between descriptors appearing in the selected model, as seen in Table 2.

We have used two strategies for testing the validity of the selected MLR model. In the first strategy, a cross-validation method [27] was used for which the Q coefficient is the cross-validated q^2 [28] that describes the predictive power of the model.

$$Q = q^2 = 1 - (\text{PRESS}/\text{Var}) \quad (3)$$

- PRESS: predictive residual sum of squares
- Var: variance of the observed values around the mean value

In the present work, nine molecules have been removed randomly from the training set each time and a model was developed with the remaining molecules. At each step, the cytotoxicity values of the nine molecules were predicted by the model obtained. This process was repeated until each molecule had a chance to be predicted once. For this procedure, the Q value was found to be 0.730, which is close to the value (0.743) obtained by Tronchet et al. [26]

The model obtained was considered to be a good predictive one, according to Wold [27] (Q>0.6).

As a second strategy, the cytotoxicity of nine HEPT derivatives was predicted by using the best MLR model (Eq. 2).

An appropriate measure of the model's predictive ability is the PRESS/SSY ratio, [27] where PRESS is predictive residual sum of squares and SSY is the sum of the squares of the experimental values. The PRESS/SSY ratio for the test set was 0.023, and a value of this ratio smaller than 0.1, indicating an excellent predictive quality of the model.

Artificial neural network analysis

The ANN was generated by using the pertinent descriptors that appeared in the MLR model as input. A 5–6–1 neural network architecture was developed with optimum momentum and learning rate of 0.9 and 0.02, respectively, and with 30,000 iterations. The six hidden neurons were chosen to maintain ρ [25] between 1.8 and 2.2. To verify this condition we have also performed a trial by taking three to nine neurons in the hidden layer and it was found that the six hidden neurons give the best result for the training and test sets, as given in Table 3.

Table 3 Variation of r and s with number of hidden neurons

Hidden neurons	r(training)	s(training)	r(test)	s(test)
3	0.8988	0.1851	0.5116	0.6324
4	0.8970	0.1867	0.5102	0.6333
5	0.8978	0.1860	0.4981	0.6382
6	0.8990	0.1850	0.5117	0.6322
7	0.9000	0.1839	0.5040	0.6356
8	0.8999	0.1849	0.4967	0.6383
6	0.8978	0.1860	0.4982	0.6379

To evaluate the neural network, the correlation coefficient r of its results is compared with the r for the regression model developed in this work. The r values were 0.884 and 0.935 for the training set in the present MLR and the ANN, respectively. The corresponding standard error s for the two analyses was 0.204 and 0.149 respectively. This reveals the improvement of the MLR model.

We used the same procedure as for the MLR analysis for testing the validity of the ANN model selected. The corresponding PRESS/SSY for the prediction set is 0.023 for both MLR and ANN. This reveals an excellent predictive quality for both methods used in this work.

The corresponding Q values in cross-validation methods were 0.730 and 0.70 in MLR and ANN, respectively. These results indicate that both MLR and ANN have good predictive ability. [27]

The plot in Fig. 2a, b indicates that there is a significant correlation between actual values and calculated values of log(1/CC_{50}) from MLR and ANN for the training and test sets. An examination of the possible outliers showed that three compounds (**21**, **84** and **89**) have residuals higher than 3s (Fig. 2a). However, the ANN indicated four compounds (**72**, **84**, **88** and **89**) with residuals larger than 3s (Fig. 2b).

Analysis of descriptors' contributions in ANN and MLR models

The evaluation of the relevance of descriptors proved quite interesting and useful, so we chose to estimate the relative contributions of descriptors in two different ways:

Table 4 Evaluating the impact of each descriptor in ANN and MLR

Removed descriptor	*C*% ANN[a]	*C*% MLR[a]	*r* ANN[b]	s ANN[b]	*r* MLR[b]	s MLR[b]	PRESS/SSY[c] ANN (test)	PRESS/SSY[c] MLR (test)
log $P(R_1)$	25.532	23.021	0.759	0.277	0.721	0.302	0.016	0.016
$RM(R_2)$	17.312	16.498	0.914	0.170	0.865	0.218	0.023	0.021
$\Sigma MW(R_3+R_4)$	22.695	26.612	0.769	0.264	0.455	0.388	0.029	0.031
log $P(R_4)$	18.076	17.404	0.852	0.220	0.791	0.266	0.025	0.026
$H(z)$	16.384	16.463	0.913	0.172	0.851	0.228	0.024	0.026
Any one	100	100	0.935	0.149	0.884	0.204	0.023	0.023

[a] The contribution (*C*%) of descriptor given by the first method (i) described in the text

[b] Given by the second method (ii) described in the text

[c] PRESS: predictive residual sum of squares; SSY: sum of square of the observed values

i. The contribution of descriptors *i* (*i*=1–5) was estimated from the trained 5–6–1 configuration network. The descriptor under study was removed from the 5–6–1 ANN together with its corresponding weights. Then the network (4–6–1) calculated the output of each molecule as usual. The mean of the absolute deviation values Δm_i between the observed value and the estimated value for all compounds was calculated. This process was reiterated for each descriptor. Finally, the contribution C_i [29] of descriptor *i* is given by:

$$C_i = 100\Delta m_i \Big/ \sum_{j=1}^{5} \Delta m_i \qquad (4)$$

ii. We elaborated a method, [30] which consists of removing a descriptor and analyzing the statistical coefficient between observed and calculated using ANN and MLR. Comparison between these statistics and those calculated by ANN or MLR when no descriptor was removed gave an idea of the importance of the descriptor removed. The results of this section are given in Table 4.

According to the results above, it appears that log $P(R_1)$, $\Sigma MW(R_3+R_4)$ and log $P(R_4)$ have the same classification in the two methods (i and ii) used, and that these parameters contribute greatly to $\log(1/CC_{50})$ for the training and test sets. These results also indicate that the relative importance of the descriptors log $P(R_1)$ and $\Sigma MW(R_3+R_4)$ have changed order in the two techniques used (MLR and ANN).

From the results obtained, one may conclude that of some of the parameters contributing to the RT cytotoxicity property can be non-linear. This conclusion arises from the fact that the same descriptors have been used for the development of the MLR model and the ANN.

To ensure that the results obtained were not due to chance and to lend credence to our results, we have run a scrambling experiment. The dependent variable $[\log(1/CC_{50})]$ was randomly scrambled and then the same algorithms used in MLR and ANN run once again. The statistical results (the correlation coefficient *r* and the standard deviation of the results) were compared with the *r* and *s* of the MLR and ANN models developed in this work. The *r* values were 0.170 and 0.460 compared with 0.884 and 0.935; for the *s* values we have obtained 0.432 and 0.402 compared with 0.204 and 0.149 for the training set in MLR and ANN, respectively. This test confirms and clearly shows that the descriptors selected in this study describe the activity studied very well.

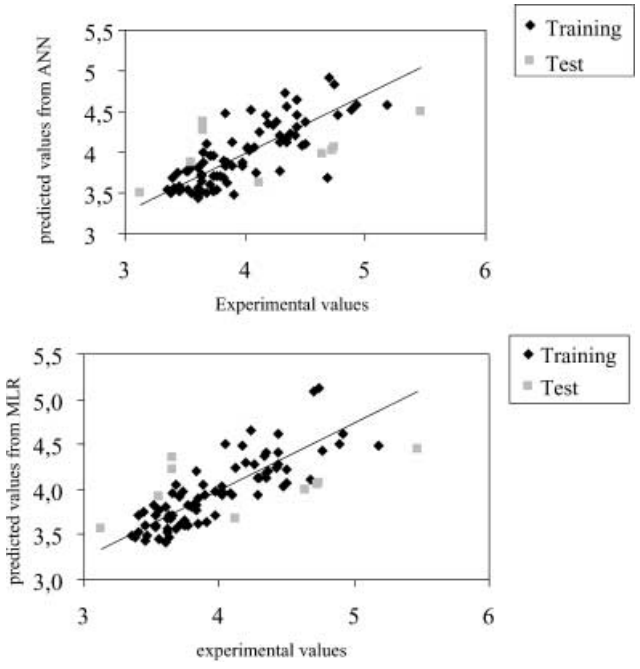

Fig. 2 Experimental and predicted values from MLR (*below*) and ANN (*above*) for the training and test sets

Conclusions

In this series, the cytotoxicity of the above compounds was investigated by means of MLR and ANN techniques. The results of the QSAR study obtained in this work indicate that the cytotoxicity of HEPT derivatives depends strongly on the hydrophobic factors as expressed by log $P(R_1)$ and log $P(R_4)$ together, and is specially dependent on the steric factors mainly accounted for by $\Sigma MW(R_3+R_4)$ and the molar refraction $MR(R_2)$ and also by $H(z)$. In addition this study reveals the non-linear effects in HEPT analogues.

Supporting information available The pertinent descriptor values and the log($1/CC_{50}$) used in this work, with the descriptive statistics, are available in the supporting material.

References

1. Goldman ME, Nunberg JH, O'Brien JA, Quintero JC, Schleif WA, Freund KF, Lee Gaul S, Saari WS, Wai JS, Hoffman JM, Anderson PS, Hupe DJ, Emin EA (1991) Proc Natl Acad Sci USA 88:6863–6867
2. Larder BA, Kemp SD, Purifoy DJM (1989) Proc Natl Acad Sci USA 86:4803–4807
3. Mitsuya H, Weinhold KJ, Furman PA, Clair MHS, Lehrman SN, Gallo RC, Bolognesi D, Barry DW, Broder S (1985) Proc Natl Acad Sci USA 82:7096–7100
4. Sandberg JA, Slikker WJR (1995) Review 9:1157–1163
5. Baba M, Tanaka H, De Clercq E, Pauwels R, Balzarini J, Schols D, Nakashima H, Perno CF, Walker RT, Miyasaka T (1989) Biochem Biophys Res Commun 165:1375–1381
6. Miyasaka T, Tanaka H, Baba M, Hayakawa H, Walker RT, Balzarini J, De Clercq E (1989) J Med Chem 32:2507–2509
7. Pauwels R, Andries K, Desmyter J, Schols D, Kukla MJ, Breslin HJ, Raeymaeckers A, Van Gelder J, Woestenborghs R, Heykants J, Schellekens K, Janssen MAC, De Clercq E, Janssen PAJ (1990) Nature 343:470–474
8. Merluzzi VJ, Hargrave KD, Labadia M, Grozinger K, Skoog M, Wu JC, Faanes R, Eckner RJ, Koup RA, Sullivan JL (1990) Science 250:1411–1413
9. Koup RA, Merluzzi VJ, Hargrave KD, Adams J, Grozinger K, Eckner RJ, Sullivan JL (1991) J Infect Dis 163:966–970
10. Romero DL, Busso M, Tan CK, Reusser F, Palmer JP, Poppe SM, Aristoff PA, Downey KM, So AG, Resnick L, Tarpley WG (1991) Proc Natl Acad Sci USA 88:8806–8810
11. Balzarini J, Pérez-Pérez MJ, Sn-Félix A, Schols D, Perno CF, Vandame AM, Camarasa MJ, De Clercq E (1992) Proc Natl Acad Sci USA 89:4392–4396
12. Kleim JP, Bender R, Billhardt UM, Meichsner C, Rieb G, Rosner M, Winkler I, Paessens A (1993) Antimicrob Agents Chemother 37:1659–1664
13. Pauwels R, Andries K, Debyser Z, Van Daele P, Schols D, Stoffels P, De Vreese K, Woestenborgh G, Desmyter J, Heykants J, Janssen MAC, De Clercq E, Janssen PAJ (1993) Proc Natl Acad Sci USA 90:1711–1715
14. Zhang H, Vrang L, Bäckbro K, Unge T, Engelhardt P, Högberg M, Kangasmestä J, Lind P, Noréen R, Sahlberg C, Zhou XX, Johansson NG, Cantrell AS, Jaskunas R, Morin JM, Ternansky RJ, Bel FW, Jordan CL, Kinnick MD, Palkowitz JA, Parrish CA, Pranc P, Vasileff RT, West SJ, Oberg B (1993) Abst 2nd International Workshop on HIV Drug-Resistance, Noordwijk, 3–5 June
15. Tanaka H, Baba M, Hayakawa H, Sakamaki T, Miyasaka T, Ubasawa M, Takashima H, Sekiya K, Nitta I, Shigeta S, Walker RT, Balzarini J, De Clercq E (1991) J Med Chem 34:349–357
16. Tanaka H, Takashima H, Ubasawa M, Sekiya K, Nitta I, Baba M, Shigeta S, Walker RT, De Clercq E, Miyasaka T (1992) J Med Chem 35:337–345
17. Tanaka H, Takashima H, Ubasawa M, Sekiya K, Nitta I, Baba M, Shigeta S, Walker RT, De Clercq E, Miyasaka T (1992) J Med Chem 35:4713–4719
18. Tanaka H, Takashima H, Ubasawa M, Sekiya K, Nitta I, Baba M, Shigeta S, Walker RT, De Clercq E, Miyasaka T (1992) J Med Chem 35:4713–4719
19. Danel K, Larsen E, Pederson EB, Vestergaard BF, Neilsen C (1996) J Med Chem 39:2427–2431
20. MMP, molecular modelling pro-Demo (TM) Revision 301 demo. ChemSW Software (TM)
21. Bodor N, Harget A, Huang MJ (1991) J Am Chem Soc 113:9480–9483
22. Unistat statistical package, version 4.0 for Excel. UNISTAT Ltd. London http://www.unistat.com/
23. Data pro Qnet2000 for Windows V2 K build 721 neural network modelling
24. Rumhelart DE, Hinton CE, Williams RJ (1986) Nature 323:533–536
25. So S, Richards WG (1992) J Med Chem 35:3201–3207
26. Tronchet JMJ, Grigorov M, Dolatshahi, N, Moriaud F, Weber J (1997) Eur J Med Chem 32:279–299
27. Wold S (1991) Quant Struct–Act Relat 10:191–193
28. Tetko IV, Villa AEP, Livingstone DJ (1996) J Chem Inf Comput Sci 36:794–803
29. Cherquaoui D, Esseffar M, Villemin D, Cence JM, Chastrette M, Zakarya D (1998) New J Chem 839–843
30. Chastrette M, Zakarya, D, Peyraud JF (1994) Eur J Med Chem 29:343–348