Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 14 (2006) 5876–5889

Bioorganic & Medicinal Chemistry

QSAR for non-nucleoside inhibitors of HIV-1 reverse transcriptase

Pablo R. Duchowicz,[a,*] Michael Fernández,[b] Julio Caballero,[b]
Eduardo A. Castro[a] and Francisco M. Fernández[a]

[a] *INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, (1900) La Plata, Argentina*

[b] *Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, (44740) Matanzas, Cuba*

Received 26 March 2006; revised 6 May 2006; accepted 15 May 2006
Available online 12 June 2006

Abstract—By means of QSAR algorithms we model the potency pIC_{90} [mM] of 154 non-nucleoside reverse transcriptase inhibitors (NNRTI) of the wild-type HIV-1 virus, considered as the second generation analogues of Efavirenz. In addition, 56 inhibitors of the K-103N viral mutant form are also investigated. A pool of 1494 theoretical molecular descriptors provided mainly by the Dragon 5 software is explored by several methods of variable selection: forward stepwise regression, the replacement method, and the genetic algorithm approach. The optimal models found include up to seven parameters: $R = 0.7991$, $R_{l\text{-}20\%\text{-}o} = 0.7233$ for the case of wild-type, and $R = 0.9261$, $R_{l\text{-}5\%\text{-}o} = 0.8802$ for the K-103N mutation.

1. Introduction

Increasing efforts have been carried out during the last years for developing effective clinical treatments that interfere with the replication cycle of the human immunodeficiency virus type-1 (HIV-1). The basic strategy for combating the disease consists in either inhibiting the HIV protease or the reverse transcriptase (RT) site. All the inhibitors of RT proposed up to now may or may not compete with the nucleoside triphosphate binding site on RT, thus behaving as nucleoside (NRTI) or non-nucleoside (NNRTI) RT inhibitors.[1,2] It was shown both in adult and paediatric patients that a high rate of virological control can be achieved using a triple combination therapy such as two NRTI plus an NNRTI or a protease inhibitor.[3]

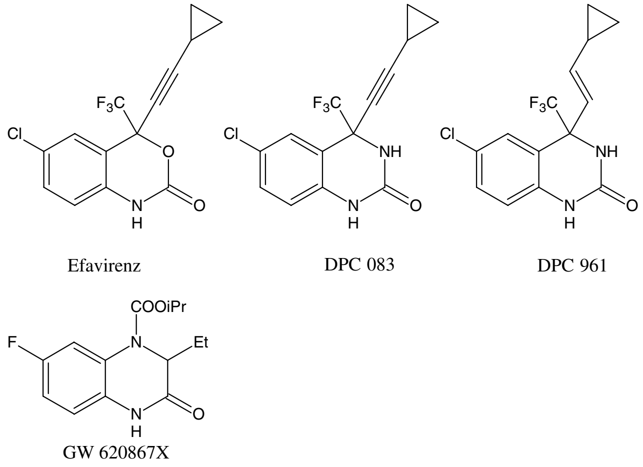

Figure 1. The structures of the most popular NNRTI.

Figure 1 shows the structure of Efavirenz,[4] the first NNRTI to be approved for a joint once a day dosing. In previous studies it was demonstrated that this type of drug is able to penetrate into the cerebrospinal fluid, a common viral environment.[5] Despite this, a prolonged therapy of the agent leads to mutations of RT including different forms, such as K103N, L100I, Y188L, or others. This fact motivated many researchers to develop hybrid analogues of Efavirenz with better biological activities, by optimization of the structure–activity relationship (SAR) of the functional groups present in the leading structure in terms of the inhibitory activity on RT. In this way it is expected to identify an NNRTI with a broad spectrum of activities against the various viral forms mentioned, and having pharmacokinetics consistent with once-daily dosing. Other quinazolinone drugs which are currently under clinical investigations are DPC 961, DPC 083, and GW 620867X.

Keywords: QSPR–QSAR; Molecular descriptors; Reverse transcriptase inhibitor; Anti-HIV potency.

* Corresponding author. Fax: +54 221 425 4642; e-mail: pabloducho@yahoo.com.ar

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.05.027

The purpose of the present study was to model both the wild-type and mutant K-103N forms of HIV-1 by quantitative structure–activity relationships (QSAR). To this end, we use three different computational procedures explained in the next section: the forward stepwise regression method (SR), the recently proposed replacement method (RM), and the genetic algorithm (GA).

2. Results and discussion

2.1. Wild-type HIV-1 study

We first applied the forward stepwise method to the total set of molecular descriptors, which in present analysis included $D = 1494$ descriptors. The best relationship achieved for modeling the inhibitory activity of the NNRTI against wild-type HIV-1 RT, in terms of the best predictive power of the equation and the least number of variables involved, has the following statistics:

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 6.114(\pm 0.3) - 0.0738(\pm 0.008)\ \mathrm{Rww}\\ & + 6.437(\pm 1)\ \mathrm{LDip}\\ & + 1.408(\pm 0.2)\ \mathrm{Mor21m}\\ & + 1.000(\pm 0.2)\ \mathrm{Mor31u}\\ & + 1.242(\pm 0.3)\ \mathrm{Mor32m}\\ & - 1.210(\pm 0.3)\ \mathrm{MATS5e}\\ & - 0.768(\pm 0.2)\ \mathrm{DISPe} \qquad (1)\end{aligned}$$

$N = 154$, $R = 0.7866$, $S = 0.419$, $F = 33.861$

$R_{loo} = 0.7592$, $S_{loo} = 0.431$

$R_{l\text{-}20\%\text{-}o} = 0.6667$, $S_{l\text{-}20\%\text{-}o} = 0.496$

The details for the molecular descriptors appearing in all the proposed models are given in Table 3. On the other hand, the application of the GA-based feature selection routine produces a better seven-variable lineal model,

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 5.936(\pm 0.6) - 0.638(\pm 0.09)\ \mathrm{AECC}\\ & + 2.289(\pm 0.5)\ \mathrm{BELe2}\\ & + 0.211(\pm 0.05)\ \mathrm{TE2}\\ & + 0.121(\pm 0.03)\ \mathrm{RDF090v}\\ & + 1.651(\pm 0.3)\ \mathrm{Mor21v}\\ & - 0.658(\pm 0.1)\ \mathrm{nCq}\\ & - 0.366(\pm 0.07)\ \mathrm{nHDon} \qquad (2)\end{aligned}$$

$N = 154$, $R = 0.7953$, $S = 0.412$, $F = 35.892$

$R_{loo} = 0.7682$, $S_{loo} = 0.423$

$R_{l\text{-}20\%\text{-}o} = 0.7129$, $S_{l\text{-}20\%\text{-}o} = 0.467$

For the present data set the RM yields slightly better results than those presented for Eq. 2:

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 0.747(\pm 1) - 0.204(\pm 0.02)\ \mathrm{MDDD}\\ & + 0.364(\pm 0.06)\ \mathrm{TE2}\\ & + 1.846(\pm 0.3)\ \mathrm{Mor23e}\\ & - 3.612(\pm 0.5)\ \mathrm{Mor23v}\\ & + 0.539(\pm 0.1)\ \mathrm{Mor16e}\\ & + 0.730(\pm 0.2)\ \mathrm{Mor21m}\\ & + 0.972(\pm 0.2)\ \mathrm{MAXDP} \qquad (3)\end{aligned}$$

$N = 154$, $R = 0.7991$, $S = 0.408$, $F = 36.856$

$R_{loo} = 0.7760$, $S_{loo} = 0.417$

$R_{l\text{-}20\%\text{-}o} = 0.7233$, $S_{l\text{-}20\%\text{-}o} = 0.459$ with the fitted values shown in Table 1. As can be seen from this and all the previous equations, different types of descriptor definitions are needed to explain in the best possible manner the pIC_{90} values for wild-type. The optimum variables appearing in this equation do not make any of the training compounds to have an absolute residual (difference between the experimental and predicted value of the property) exceeding 3*S*, but compounds {**26**,**58**,**70**, **71**,**85**} exceed 2.5*S*. The *l*-20%-*o* parameters (the poorest performance of the model appeared after successive removing of 30 molecules from the set) suggest that the model has predictive power. Also, in Figure 5 the plot of the predicted property as function of the experimental data tends to be represented with a straight line trend. Figure 6 shows the residuals as function of the predicted property, thus revealing that the deviations are randomly distributed and do not follow any kind of strange pattern, and also the absence of data clustering suggests that the model is correct.

The descriptors in Eq. 3 can be classified as follows: (i) two topologicals: MDDD, the mean distance degree deviation, and MAXDP, the maximal electrotopological positive variation; (ii) four 3D-MoRSE indices: Mor23e, Mor23v, Mor16e, and Mor21m; and a charge descriptor: TE2, the topographic electronic descriptor (bond restricted). We can also rank the descriptors in model (3) according to their effect on increasing the value of *S* when removed from the model. In this case, the order found is:

$$\begin{aligned}& \mathrm{MDDD} > \mathrm{Mor23e} > \mathrm{Mor21m} > \mathrm{MAXDP} > \mathrm{TE2}\\ & \quad > \mathrm{Mor16e} > \mathrm{Mor23v} \qquad (4)\end{aligned}$$

Table 4 shows the correlation matrix among the variables of Eq. 3 and the correlation of pIC_{90} with each single variable. It reveals that the optimal descriptors are not seriously intercorrelated, thus justifying the inclusion of all the variables in the relationship.

According to Eq. 4, the most important variable is the topological quantity MDDD, defined as follows:[6]

$$\mathrm{MDDD} = 1/p\Sigma_i|s_i - 2W/p| \quad s_i = \Sigma_j d_{ij} \qquad (5)$$

Here, *W* is the Wiener index[7] for graph *G*, s_i is the ith-row sum from the topological distance matrix (having

Table 1. Experimental and predicted values of pIC$_{90}$ [mM] for the training set of 154 NNRTI of wild-type HIV-1 RT

No.	Type	Chemical structure			pIC$_{90}$	
		R1	R2	R3	Exp	Pred
1	**A**	5-F	Cp	—	5.363	5.476 (−0.113)
2	**A**	5-F	Et	—	5.379	5.448 (−0.069)
3	**A**	5-F	Pr	—	5.249	5.157 (0.092)
4	**A**	5-F	*i*Pr	—	5.178	5.367 (−0.190)
5	**A**	6-NO$_2$	Cp	—	6.081	5.646 (0.435)
6	**A**	6-NO$_2$	Et	—	5.652	5.554 (0.097)
7	**A**	6-NO$_2$	Pr	—	5.34	5.178 (0.161)
8	**A**	6-NO$_2$	*i*Pr	—	5.475	5.696 (−0.221)
9	**A**	6-NH$_2$	Cp	—	4.686	4.956 (−0.270)
10	**A**	6-NH$_2$	Et	—	4.34	4.647 (−0.308)
11	**A**	6-NH$_2$	Pr	—	4.475	4.255 (0.220)
12	**A**	6-NH$_2$	*i*Pr	—	4.561	4.591 (−0.030)
13	**A**	6-NHMe	Cp	—	5.045	4.720 (0.324)
14	**A**	6-NHMe	*i*Pr	—	4.989	4.753 (0.235)
15	**A**	6-NHAc	Cp	—	3.529	4.118 (−0.590)
16	**A**	6-NHAc	*i*Pr	—	3.301	4.174 (−0.873)
17	**A**	5-H	Cp	—	4.987	5.061 (−0.074)
18	**A**	6-F	Cp	—	5.134	5.153 (−0.019)
19	**A**	6-*i*Pr	Cp	—	4.555	4.992 (−0.437)
20	**A**	6-NMe$_2$	Cp	—	5.079	5.061 (0.017)
21	**A**	6-OCF$_3$	Cp	—	4.724	4.059 (0.665)
22	**A**	5,6-F	Cp	—	5.502	5.183 (0.318)
23	**A**	5,8-F	Cp	—	4.745	5.022 (−0.277)
24	**A**	5,6,8-F	Cp	—	4.853	4.998 (−0.146)
25	**A**	5,6,7-F	Cp	—	4.708	5.160 (−0.452)
26	**A**	6,7,8-F	Cp	—	3.379	4.516 (−1.137)
27	**A**	6-Cl,8-OMe	Cp	—	3.914	4.332 (−0.419)
28	**A**	6,8-Cl	Cp	—	4.544	5.060 (−0.517)
29	**A**	6-OMe	Cp	—	5.699	4.727 (0.972)
30	**A**	6-Cl,8-F	Cp	—	5.143	5.179 (−0.036)
31	**A**	6-Ph	Cp	—	3.604	4.030 (−0.426)
32	**A**	6-Me	Cp	—	5.148	4.990 (0.157)
33	**A**	6,7-CHCH–CHCH-	Cp	—	4.571	4.602 (−0.031)
34	**B**	Me	Cp	—	5.456	4.915 (0.540)
35	**B**	Me	Ph	—	5.102	4.844 (0.258)
36	**B**	Me	3-Py	—	5.149	4.695 (0.454)
37	**B**	H	Cp	—	5.444	5.483 (−0.040)
38	**B**	H	Ph	—	5.167	5.250 (−0.084)
39	**C**	Cp	—	—	5.444	5.370 (0.073)
40	**C**	Ph	—	—	5.086	5.165 (−0.080)
41	**C**	3-Py	—	—	5.337	5.075 (0.261)
42	**D**	6-H	Cp	Bn	3.74	4.239 (−0.500)
43	**D**	6-H	Cp	2-PyMe	2.997	3.838 (−0.841)
44	**D**	6-H	Ph	4-PyMe	3.366	3.943 (−0.577)
45	**D**	5,6-CHCH–CHCH	Cp	Bn	3.237	3.178 (0.059)
46	**D**	5-Et,6-Me	Cp	Me	3.697	3.864 (−0.167)
47	**D**	5-Et,6-Me	Cp	Et	4.022	4.173 (−0.152)
48	**D**	5-Et,6-Me	Cp	Pr	4.495	4.258 (0.236)
49	**D**	5- Et,6-Me	Cp	CpMe	4.432	3.932 (0.499)
50	**D**	5-Et,6-Me	Cp	Butyl	4.377	3.661 (0.715)
51	**D**	5-Et,6-Me	Cp	Bn	4.337	4.328 (0.009)
52	**D**	5-Et,6-Me	Cp	2,6-F-Bn	4.31	3.907 (0.402)
53	**D**	5-Et,6-Me	Cp	2-Cl,6-F-Bn	4.215	4.281 (−0.067)
54	**D**	5-Et,6-Me	Cp	2,6-Cl-Bn	3.804	3.416 (0.387)
55	**D**	5-Et,6-Me	Cp	2-F,6-OMeBn	3.796	3.595 (0.200)
56	**D**	5-Et,6-Me	Cp	2,4-Cl-Bn	3.721	3.515 (0.205)
57	**D**	5-Et,6-Me	Cp	3,5-OMe	3.237	3.497 (−0.260)
58	**D**	5-Et,6-Me	Cp	CH$_2$CN	2.854	3.923 (−1.069)
59	**E**	7-Cl	Ph	H	4.614	4.251 (0.362)
60	**E**	7-Cl	Et	H	5.041	4.353 (0.687)
61	**E**	7-Cl	*i*Pr	H	5.013	4.576 (0.437)
62	**E**	7-Cl	3-Furanyl	H	4.323	4.944 (−0.622)
63	**E**	7-Cl	Cp	H	5.041	4.738 (0.303)
64	**E**	7-Cl	Cp	*cis*-Me	5.658	5.161 (0.496)

Table 1 (*continued*)

No.	Type	Chemical structure			pIC_{90}	
		R1	R2	R3	Exp	Pred
65	**E**	7-Cl	Cp	*trans*-Me	5.155	4.959 (0.196)
66	**E**	7-Cl	Cp	*cis*-Et	5.076	4.957 (0.118)
67	**E**	7-Cl	Cp	*cis*-Pr	4.764	4.871 (−0.107)
68	**E**	7-Cl	Cp	*cis*-*i*Pr	4.684	4.690 (−0.007)
69	**E**	7-Cl	Cp	*cis*-CH_2CF_3	4.775	4.406 (0.368)
70	**E**	7-F	Cp	*trans*-Me	3.723	4.931 (−1.208)
71	**E**	7-F	Cp	*trans*-Et	4.058	5.150 (−1.092)
72	**E**	7-F	Cp	*cis*-Me	5.284	5.149 (0.135)
73	**E**	7-F	Cp	*cis*-Et	5.276	4.991 (0.285)
74	**E**	6,7-F	Cp	*cis*-Me	5.237	5.135 (0.102)
75	**F**	7-Cl	Cp	H	5.409	4.935 (0.473)
76	**F**	7-Cl	Cp	*cis*-Me	5.602	5.278 (0.323)
77	**F**	7-Cl	Cp	*cis*-Et	5.108	4.993 (0.115)
78	**F**	7-F	Cp	*cis*-Me	5.119	5.298 (−0.180)
79	**F**	7-F	Cp	*cis*-Et	5.602	5.124 (0.477)
80	**F**	6,7-F	Cp	*cis*-Me	5.284	5.478 (−0.194)
81			DPC 083		5.678	5.538 (0.140)
82			DPC 961		5.699	5.112 (0.587)
83	**A**	6-Cl	2-Py	—	5.268	4.918 (0.349)
84	**A**	6-Cl	3-Py	—	5.401	4.813 (0.587)
85	**A**	6-Cl	4-Py	—	3.991	5.036 (−1.045)
86	**A**	6-Cl	2-Furanyl	—	5.42	5.340 (0.079)
87	**A**	6-Cl	3-Furanyl	—	5.42	5.529 (−0.110)
88	**A**	6-Cl	3-Thienyl	—	5.35	5.729 (−0.380)
89	**A**	6-Cl	5-Thiazolyl	—	5.175	5.430 (−0.255)
90	**A**	6-F	2-Py	—	4.914	4.808 (0.105)
91	**A**	6-F	3-Py	—	5.592	4.766 (0.826)
92	**A**	6-F	2-Furanyl	—	5.588	5.386 (0.201)
93	**A**	6-F	3-Furanyl	—	5.599	5.573 (0.026)
94	**A**	6-F	2-Thienyl	—	5.533	5.385 (0.147)
95	**A**	6-F	3-Thienyl	—	5.593	5.567 (0.025)
96	**A**	5,6-F	2-Py	—	5.169	5.102 (0.067)
97	**A**	5,6-F	3-Py	—	5.652	5.067 (0.584)
98	**A**	5,6-F	2-Furanyl	—	5.421	5.647 (−0.226)
99	**A**	5,6-F	3-Furanyl	—	5.493	5.663 (−0.170)
100	**A**	5,6-F	2-Thienyl	—	5.652	5.787 (−0.135)
101	**A**	5,6-F	3-Thienyl	—	5.631	5.637 (−0.007)
102	**G**	6-Cl	Et	—	4.738	5.346 (−0.609)
103	**G**	6-Cl	Pr	—	5.236	5.238 (−0.003)
104	**G**	6-Cl	Butyl	—	4.992	4.820 (0.171)
105	**G**	6-Cl	Isopentyl	—	4.997	5.197 (−0.201)
106	**G**	6-Cl	CpMe	—	5.114	5.061 (0.053)
107	**G**	6-Cl	$(CH_2)_2$Cp	—	5.095	4.858 (0.236)
108	**G**	6-Cl	PhMe	—	4.747	4.718 (0.028)
109	**G**	6-Cl	$(CH_2)_2$Ph	—	3.778	4.403 (−0.626)
110	**G**	6-Cl	CH_2CHCH_2	—	5.409	5.139 (0.269)
111	**G**	6-Cl	*cis*-CH_2CHCH Me	—	5.362	5.235 (0.127)
112	**G**	6-Cl	*trans*-$CH_2CHCHMe$	—	5.252	5.010 (0.242)
113	**G**	6-Cl	CH_2CHCMe_2	—	5.572	5.358 (0.214)
114	**G**	6-Cl	CH_2CCMe	—	5.504	4.727 (0.776)
115	**G**	6-Cl	CH_2CH_2CH CH_2	—	4.864	4.979 (−0.116)
116	**G**	6-Cl	CH_2CHCCl_2	—	5.022	4.628 (0.393)
117	**G**	6-Cl	CH_2CH_2OMe	—	4.212	4.578 (−0.367)
118	**G**	6-Cl	CH_2(3-Py)	—	4.123	4.737 (−0.614)
119	**G**	6-Cl	$CH_2CF_2CF_3$	—	3.187	3.756 (−0.569)
120	**G**	6-F	CH_2CCMe_2	—	5.526	5.320 (0.205)
121	**G**	6-F	*trans*-CH_2CHCH Me	—	5.053	5.128 (−0.076)
122	**G**	6-F	CH_2CHCCl_2	—	4.923	4.711 (0.211)
123	**G**	5,6-F	CH_2CHCMe_2	—	5.812	5.613 (0.198)
124	**G**	5,6-F	CH_2CHCH_2	—	5.189	5.700 (−0.512)
125	**G**	5,6-F	CH_2CHCCl_2	—	5.197	5.099 (0.097)
126			Efavirenz		5.721	5.172 (0.548)
127	**H**	H	CpMe	—	4.77	4.787 (−0.017)
128	**H**	H	Allyl	—	4.721	4.864 (−0.144)
129	**H**	H	Bn	—	4.658	4.775 (−0.118)

(*continued on next page*)

Table 1 (*continued*)

No.	Type	Chemical structure			pIC_{90}	
		R1	R2	R3	Exp	Pred
130	**H**	H	Pr	—	4.678	4.453 (0.224)
131	**H**	H	$(CH_2)_2$Cp	—	4.276	4.356 (−0.080)
132	**H**	F	Allyl	—	4.495	4.559 (−0.064)
133	**H**	F	CpMe	—	4.678	4.467 (0.210)
134	**H**	Cl	CpMe	—	4.638	4.807 (−0.169)
135	**H**	Cl	Isobutyl	—	4.569	4.368 (0.200)
136	**H**	Cl	Allyl	—	4.284	4.925 (−0.641)
137	**H**	OMe	CpMe	—	4.071	4.334 (−0.263)
138	**H**	OMe	Allyl	—	4.721	4.602 (0.118)
139	**H**	H	COOEt	—	4.538	4.384 (0.153)
140	**H**	H	COO-*i*Pr	—	5.046	5.052 (−0.006)
141	**H**	H	COO-C(CH_2)Me	—	4.824	4.860 (−0.036)
142	**H**	H	COO-*i*Pr	—	4.959	4.628 (0.330)
143	**H**	H	COObutyl	—	3.801	4.360 (−0.560)
144	**H**	H	$COOCH_2CH\ CH_2$	—	3.979	4.194 (−0.216)
145	**H**	H	COOMe	—	4.523	4.442 (0.080)
146	**H**	Cl	COOEt	—	4.357	4.840 (−0.484)
147	**H**	Cl	COO-*i*Pr	—	4.959	5.051 (−0.092)
148	**H**	Cl	COO-C(CH_2)Me	—	4.721	5.204 (−0.483)
149	**H**	F	COOEt	—	4.824	4.452 (0.372)
150	**H**	F	COO-*i*Pr	—	5.097	5.073 (0.024)
151	**H**	F	COOC(CH_2)Me	—	5.046	4.903 (0.142)
152	**H**	H	CO-Cp-Me	—	5.097	4.582 (0.514)
153	**H**	H	CO-*i*Pr	—	4.959	4.561 (0.397)
154			GW620867X		4.959	4.809 (0.150)

Cp, cyclopropyl; Me, methyl; Et, ethyl; Ac, acethyl; Pr, propyl; *i*Pr, isoPr; Ph ,phenyl; Py, pyridyl; Bn, benzyl.

elements d_{ij}, with *i* and *j* vertices of *G*), and *p* is the number of vertices. This distance-based index takes into consideration the distribution of the topological distances in the molecular structure, and thus represents the topological shape of the compound, describing the degree of ramification, centricity, and cyclicity. The 3D-MoRSE descriptors[8] appearing in the model are important because they take into account the 3D arrangement of the atoms without ambiguities (in contrast with those coming from chemical graphs), and also because they do not depend on the molecular size, thus being applicable to a large number of molecules with great structural variance and being a characteristic common to all of them. This type of indices are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. A generalized scattering function, called the molecular transform, can be used as the functional dependence for deriving, from a known molecular structure, the specific analytic relationship of both X-ray and electron-diffraction. In order to take into account the specific contributions of the atoms to the property being studied, different atomic properties can be employed as weighting schemes. Mor23e and Mor16e correspond to signal 23 and 16, respectively, both weighted by atomic Sanderson electronegativities, while Mor21m is weighted with atomic masses and Mor23v with atomic van der Waals volumes. The index TE2[9] is obtained by connecting submolecular polarity parameters such as the atomic charges (q_i), with the molecular topography expressed by means of interatomic geometrical distances (r_{ij});

$$\mathrm{TE2} = \Sigma_{ij}(|q_i - q_j|)/r_{ij}^2 \qquad (6)$$

and the sum runs only over bonded atom pairs. The remaining variable, MAXDP, is a descriptor derived from the hydrogen-depleted molecular graph and obtained from the Kier–Hall intrinsic states of atoms[10] as:

$$\mathrm{MAXDP} = \max_i |\Delta I_i| \qquad (7)$$

where ΔI_i is the field effect on the *i*th atom due to the perturbation of all other atoms as defined by Kier and Hall:

$$\Delta I_i = \Sigma_j (I_i - I_j)/(d_{ij} + 1) \qquad (8)$$

The sum runs over all the other atoms in the molecular graph, I is the atomic intrinsic state and d_{ij} is the topological distance between the two considered atoms. The intrinsic state of an atom is calculated as the ratio between the Kier–Hall atomic electronegativity and the vertex degree, that is, the number of bonds of an atom, encoding information related to both atomic partial charges and their topological position relative to the whole molecule. Therefore, MAXDP represents the maximum positive intrinsic state difference and can be related to the electrophilicity of the molecule.

In order to further analyze the predictability of the optimal molecular descriptors found by RM on the whole set of 154 NNRTI, we performed a training set–test set partition, selecting 100 molecules for calibrating and the remainders for an external validation of the

Table 2. Experimental and predicted values of pIC_{90} [mM] for the training set of 56 NNRTI of mutant K-103N HIV-1 RT

No.	Type	Chemical Structure			pIC_{90}	
		R1	R2	R3	Exp	Pred
1	**B**	Me	Cp	—	5.125	4.900 (0.225)
2	**B**	Me	Ph	—	3.947	4.139 (−0.192)
3	**B**	Me	3-Py	—	3.833	3.953 (−0.120)
4	**B**	H	Cp	—	4.553	4.400 (0.153)
5	**B**	H	Ph	—	4.523	4.345 (0.178)
6	**C**	Cp	—	—	4.119	4.233 (−0.114)
7	**C**	Ph	—	—	4.337	4.346 (−0.009)
8	**C**	3-Py	—	—	4.119	4.137 (−0.018)
9	**D**	5-Et,6-Me	Cp	Pr	3.842	4.044 (−0.202)
10	**D**	5-Et,6-Me	Cp	Bn	3.947	3.915 (0.032)
11	**E**	7-Cl	Ph	H	2.387	2.404 (−0.017)
12	**E**	7-Cl	Et	H	2.754	2.686 (0.068)
13	**E**	7-Cl	*i*Pr	H	3.106	3.163 (−0.057)
14	**E**	7-Cl	Cp	H	2.855	2.980 (−0.125)
15	**E**	7-Cl	Cp	*cis*-Me	4.538	4.058 (0.480)
16	**E**	7-Cl	Cp	*trans*-Me	2.947	3.335 (−0.388)
17	**E**	7-Cl	Cp	*cis*-Et	3.857	3.881 (−0.024)
18	**E**	7-F	Cp	*cis*-Me	3.478	3.895 (−0.417)
19	**E**	7-F	Cp	*cis*-Et	3.932	3.724 (0.208)
20	**E**	6,7-F	Cp	*cis*-Me	4.125	3.696 (0.429)
21	**F**	7-Cl	Cp	H	3.441	3.388 (0.053)
22	**F**	7-Cl	Cp	*cis*-Me	4.456	4.452 (0.004)
23	**F**	7-Cl	Cp	*cis*-Et	4.886	4.812 (0.074)
24	**F**	7-F	Cp	*cis*-Me	3.928	4.160 (−0.232)
25	**F**	7-F	Cp	*cis*-Et	4.155	4.225 (−0.070)
26	**F**	6,7-F	Cp	*cis*-Me	4.092	4.238 (−0.146)
27			DPC 083		4.569	4.989 (−0.420)
28			DPC 961		5	4.405 (0.595)
29	**A**	6-Cl	2-Py		2.956	3.546 (−0.590)
30	**A**	6-Cl	3-Py		3.84	3.775 (0.065)
31	**A**	6-Cl	2-Furanyl		3.818	3.486 (0.332)
32	**A**	6-Cl	3-Furanyl		3.492	3.575 (−0.083)
33	**A**	6-Cl	3-Thienyl		3.974	4.057 (−0.083)
34	**A**	6-F	3-Py		3.482	3.615 (−0.133)
35	**A**	6-F	3-Furanyl		3.433	3.577 (−0.144)
36	**A**	6-F	3-Thienyl		3.714	3.901 (−0.187)
37	**A**	5,6-F	3-Py		4.227	3.902 (0.325)
38	**A**	5,6-F	3-Furanyl		3.553	3.779 (−0.226)
39	**A**	5,6-F	2-Thienyl		3.807	3.615 (0.192)
40	**A**	5,6-F	3-Thienyl		3.856	4.083 (−0.227)
41	**G**	6-Cl	Butyl		3.005	3.180 (−0.175)
42	**G**	6-Cl	Isopentyl		3.544	3.352 (0.192)
43	**G**	6-Cl	*cis*-$CH_2CHCHMe$		3.625	3.446 (0.179)
44	**G**	6-Cl	*trans*-$CH_2CHCHMe$		3.394	3.209 (0.185)
45	**G**	6-Cl	CH_2CHCMe		4.081	3.922 (0.159)
46	**G**	6-Cl	CH_2CCMe		3.509	3.251 (0.258)
47	**G**	6-Cl	CH_2CHCCl_2		3.616	3.612 (0.004)
48	**G**	6-F	CH_2CCMe_2		3.971	4.017 (−0.046)
49	**G**	6-F	*trans*-$CH_2CHCHMe$		2.94	2.903 (0.037)
50	**G**	5,6-F	CH_2CHCMe_2		4.187	4.108 (0.079)
51	**G**	5,6-F	CH_2CHCH_2		2.791	3.047 (−0.256)
52	**G**	5,6-F	CH_2CHCCl_2		3.738	3.536 (0.202)
53			Efavirenz		4.31	4.521 (−0.211)
54	**I**	H			2.752	2.714 (0.038)
55	**I**	O-4-PyMe			4.301	4.142 (0.159)
56	**I**	O-4-NH_2Bn			2.966	2.963 (0.003)

Cp, cyclopropyl; Me, methyl; Et, ethyl; Pr, propyl; Ph, phenyl; Py, pyridyl; Bn, benzyl.

model, as suggested in a recent study.[11] The members for each set (appearing in Tables 5 and 6) were chosen by analyzing 300 randomly generated cases, in such a way that the descriptors involved in Eq. 3 produce simultaneously the smallest S values for both series of compounds. Proceeding in this way, it is possible to generate two balanced sets of chemicals with the following statistics:

Table 3. Symbols for the molecular descriptors involved in the optimal models found

Molecular descriptor	Type	Description
Rww	Topological	Reciprocal hyper-detour index
LDip	Charge	Local dipole index
Mor21m	3D-MoRSE	3D-MoRSE—signal 21/weighted by atomic masses
Mor31u	3D-MoRSE	3D-MoRSE—signal 31/unweighted
Mor32m	3D-MoRSE	3D-MoRSE—signal 32/weighted by atomic masses
MATS5e	2D Autocorrelations	Moran autocorrelation—lag 5/weighted by atomic Sanderson electronegativities
DISPe	Geometrical	D COMMA2 value/weighted by atomic Sanderson electronegativities
AECC	Topological	Average eccentricity
BELe2	BCUT	Lowest eigenvalue no. 2 of Burden matrix/weighted by atomic Sanderson electronegativities
TE2	Charge	Topographic electronic descriptor (bond restricted)
RDF090v	RDF	Radial distribution function—9.0 weighted by atomic van der Waals volumes
Mor21v	3D-MoRSE	3D-MoRSE—signal 21/weighted by atomic Van der Waals volumes
nCq	Functional groups	Number of total quaternary C(sp^3)
nHDon	Functional groups	Number of donor atoms for H-bonds (with N and O)
MDDD	Topological	Mean distance degree deviation
Mor23e	3D-MoRSE	3D-MoRSE—signal 23/weighted by atomic Sanderson electronegativities
Mor23v	3D-MoRSE	3D-MoRSE—signal 23/weighted by atomic Van der Waals volumes
Mor16e	3D-MoRSE	3D-MoRSE—signal 16/weighted by atomic Sanderson electronegativities
MAXDP	Topological	Maximal electrotopological positive variation
De	WHIM	D total accessibility index/weighted by atomic Sanderson electronegativities
SRW05	Molecular walk counts	Self-returning walk count of order 05
RDF115u	RDF	Radial distribution function—11.5 unweighted
Mor02e	3D-MoRSE	3D-MoRSE—signal 23/weighted by atomic Sanderson electronegativities
H-051	Atom-centred fragments	H-attached to alfa-C
H3v	GETAWAY	H autocorrelation of lag 3/weighted by atomic van der Waals volumes
H8e	GETAWAY	H autocorrelation of lag 8/weighted by atomic Sanderson electronegativities
GATS1v	2D Autocorrelations	Geary autocorrelation—lag 1/weighted by atomic van der Waals volumes
SPAM	Geometrical	Average span R
RDF065u	RDF	Radial distribution function—6.5 unweighted
Mor21u	3D-MoRSE	3D-MoRSE—signal 21/unweighted
E2e	WHIM	2nd component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities
HATS8v	GETAWAY	Leverage-weighted autocorrelation of lag 8/weighted by atomic van der Waals volumes
N-072	Atom-centred fragments	RCO-N< / >N-X=X
Mor11u	3D-MoRSE	3D-MoRSE—signal 11/unweighted
Mor14u	3D-MoRSE	3D-MoRSE—signal 14/unweighted
RDF045p	RDF	Radial distribution function—4.5 weighted by atomic polarizabilities
Dp	WHIM	D total accessibility index/weighted by atomic polarizabilities
H8m	GETAWAY	H autocorrelation of lag 8/weighted by atomic masses
BELe4	BCUT	Lowest eigenvalue no. 4 of Burden matrix/weighted by atomic Sanderson electronegativities

RDF, radial distribution function.

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 0.598(\pm 1) - 0.199(\pm 0.03)\ \mathrm{MDDD} \\ & + 0.331(\pm 0.08)\ \mathrm{TE2} \\ & + 1.876(\pm 0.4)\ \mathrm{Mor23e} \\ & - 3.578(\pm 0.6)\ \mathrm{Mor23v} \\ & + 0.588(\pm 0.2)\ \mathrm{Mor16e} \\ & + 0.737(\pm 0.3)\ \mathrm{Mor21m} \\ & + 1.013(\pm 0.3)\ \mathrm{MAXDP} \end{aligned} \tag{9}$$

$N = 100,\ R = 0.7918,\ S = 0.421,\ F = 22.093$

$R_{loo} = 0.7700,\ S_{loo} = 0.430$

$N = 54,\ R_{val} = 0.8095,\ S_{val} = 0.421$

with *val* denoting the external test set. Eq. 9 indicates that both $S = 0.421$ and $S_{val} = 0.421$ are comparable to the value $S = 0.408$ obtained by means of the total set of compounds, suggesting that the optimal descriptors found are also able to work on a training set of 100 molecules, and these variables are capable to predict with the same degree of accuracy the molecules in the test set.

2.2. Mutant K-103N HIV-1 study

In this case, the best relationship explored with SR produces the following result for the inhibitory activity against mutant K-103N:

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 1.576(\pm 1) + 1.118(\pm 0.1)\ \text{N-072} \\ & - 1.007(\pm 0.1)\ \mathrm{Mor11u} \\ & - 1.081(\pm 0.2)\ \mathrm{Mor14u} \\ & + 0.116(\pm 0.02)\ \mathrm{RDF045p} \\ & + 7.756(\pm 2)\ \mathrm{Dp} - 4.096(\pm 0.9)\ \mathrm{H8m} \\ & - 2.040(\pm 0.7)\ \mathrm{BELe4} \end{aligned} \tag{10}$$

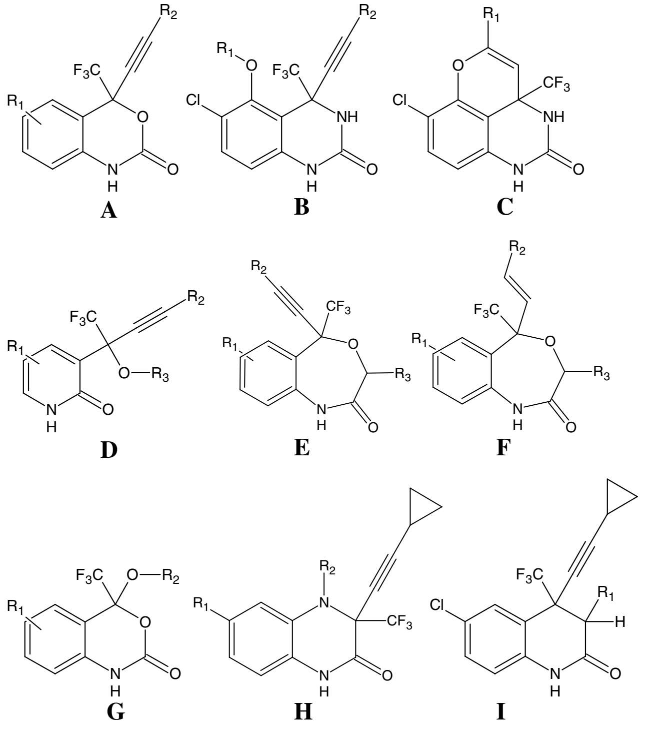

Figure 2. Structures for Efavirenz analogues under study.

$N = 56,\ R = 0.8692,\ S = 0.319,\ F = 21.197$

$R_{loo} = 0.8263,\ S_{loo} = 0.337$

$R_{l\text{-}5\%\text{-}o} = 0.7900,\ S_{l\text{-}5\%\text{-}o} = 0.370$

Again, the MLR-GA method leads to a much better model:

$$\begin{aligned} \mathrm{pIC}_{90} = {} & 15.686(\pm 1) - 1.494(\pm 0.4)\ \mathrm{GATS1v} \\ & - 31.677(\pm 2)\ \mathrm{SPAM} \\ & - 0.0531(\pm 0.01)\ \mathrm{RDF065u} \\ & + 0.697(\pm 0.1)\ \mathrm{Mor21u} \\ & + 5.073(\pm 0.8)\ \mathrm{E2e} \\ & + 12.104(\pm 2)\ \mathrm{HATS8v} \\ & - 0.858(\pm 0.07)\ \mathrm{H\text{-}051} \end{aligned} \qquad (11)$$

$N = 56,\ R = 0.9255,\ S = 0.244,\ F = 41.005$

$R_{loo} = 0.9014,\ S_{loo} = 0.259$

$R_{l\text{-}5\%\text{-}o} = 0.8778,\ S_{l\text{-}5\%\text{-}o} = 0.291$

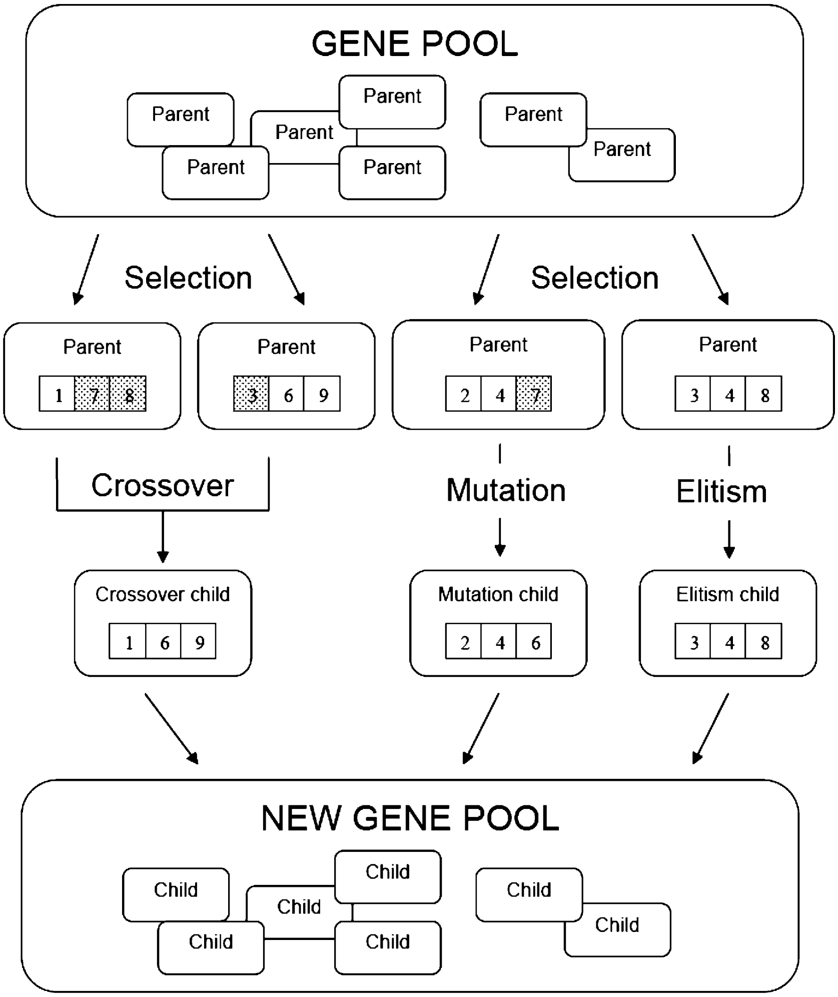

Figure 3. Flow diagram describing the strategy for the GA.

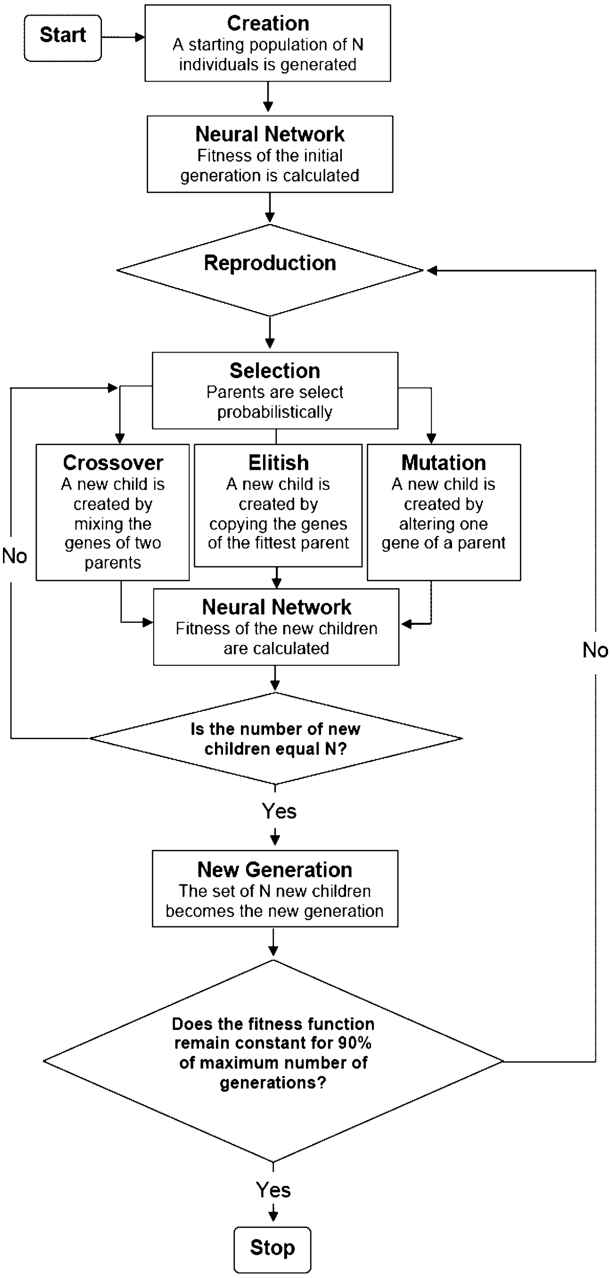
Start
Creation
A starting population of N individuals is generated
Neural Network
Fitness of the initial generation is calculated
Reproduction
Selection
Parents are select probabilistically
Crossover
A new child is created by mixing the genes of two parents
Elitish
A new child is created by copying the genes of the fittest parent
Mutation
A new child is created by altering one gene of a parent
No
Neural Network
Fitness of the new children are calculated
No
Is the number of new children equal N?
Yes
New Generation
The set of N new children becomes the new generation
Does the fitness function remain constant for 90% of maximum number of generations?
Yes
Stop

Figure 4. Schematic diagram describing the reproduction strategy in GA.

The RM was able to produce comparable results as those given by Eq. 11:

$$\begin{aligned}\mathrm{pIC}_{90} = {} & 0.880(\pm 0.6) + 6.564(\pm 1)\ \mathrm{De} \\ & + 0.0172(\pm 0.002)\ \mathrm{SRW05} \\ & - 0.255(\pm 0.08)\ \mathrm{RDF115u} \\ & - 0.0944(\pm 0.01)\ \mathrm{Mor02e} \\ & - 0.834(\pm 0.06)\ \mathrm{H\text{-}051} \\ & + 2.704(\pm 0.3)\ \mathrm{H3v} + 2.178(\pm 0.3)\ \mathrm{H8e} \quad (12)\end{aligned}$$

$N = 56$, $R = 0.9261$, $S = 0.243$, $F = 41.332$

$R_{loo} = 0.9072$, $S_{loo} = 0.252$

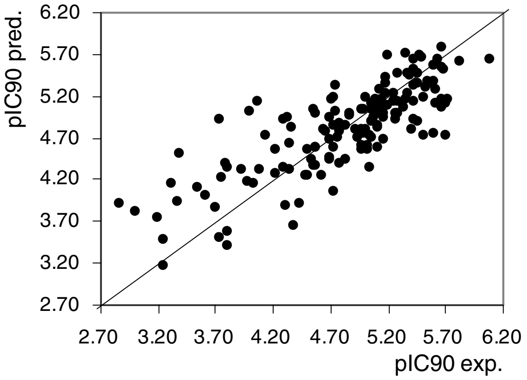

Figure 5. Predicted and experimental potencies for 154 NNRTI (wild-type HIV-1 RT).

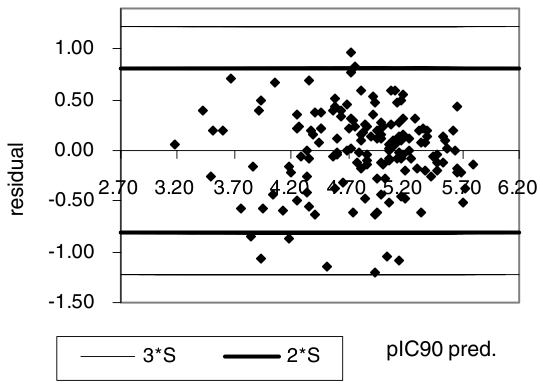

Figure 6. Dispersion plot of the residuals for 154 NNRTI of wild-type.

$R_{l\text{-}5\%\text{-}o} = 0.8802$, $S_{l\text{-}5\%\text{-}o} = 0.287$

Model (12) has no outliers exceeding 2.5S (see Table 2). As before, there is a mixture of different classes of descriptors. Figure 7 shows that a linear model can be used to correlate the molecular structure with pIC_{90} for K-103N, and Figure 8 indicates that there are no external factors that Eq. 12 cannot model, which would probably influence the predictions. The model includes the following descriptors: (i) an atom-centered fragment: H-051, the number of hydrogens attached to alfa carbon; (ii) a radial distribution function: RDF115u, of 11.5 type-unweighted; (iii) a WHIM descriptor: De, D total accessibility index/weighted by atomic Sanderson electronegativities; (iv) a 3D-MoRSE descriptor: Mor02e, signal 23/weighted by atomic Sanderson electronegativities; (v) a molecular walk counts variable: SRW05, the self-returning walk count of order 05; and (vi) two GETAWAY descriptors: H3v, H autocorrelation of lag 3/weighted by atomic van der Waals volumes, and H8e, H autocorrelation of lag 8/weighted by atomic Sanderson electronegativities. The correlation matrix (Table 7) of model (12) indicates that the optimal descriptors are not seriously intercorrelated.

The ranking of the descriptors is:

$$\begin{aligned}& \mathrm{H\text{-}051} > \mathrm{H8e} > \mathrm{SRW05} > \mathrm{H3v} > \mathrm{De} > \mathrm{Mor02e} \\ & \quad > \mathrm{RDF115u} \qquad (13)\end{aligned}$$

Table 4. Correlation matrix for Eq. 3

	pIC_{90}	MDDD	TE2	Mor23e	Mor23v	Mor16e	Mor21m	MAXDP
pIC_{90}	1.0000	0.5259	0.2041	0.2954	0.0306	0.1049	0.5172	0.2760
MDDD		1.0000	0.6215	0.3039	0.1952	0.0229	0.4313	0.7578
TE2			1.0000	0.0561	0.1237	0.0414	0.5031	0.4264
Mor23e				1.0000	0.7416	0.2537	0.1114	0.1193
Mor23v					1.0000	0.0470	0.2035	0.0146
Mor16e						1.0000	0.0832	0.0025
Mor21m							1.0000	0.2805
MAXDP								1.0000

The most significant contribution to the inhibitory potencies comes from the count of hydrogen atoms of the type H-046[12], which reveals the importance of electrostatic interactions between the NNRTI of K103-N and RT. H8e and H3v are GETAWAY descriptors derived from the molecular influence matrix (H),[13] whose elements are obtained through the atomic Cartesian coordinates values. H8e contemplates in its calculation a sum of contributions of pairs of atoms located at a topological distance of 8 and is weighted by atomic Sanderson electronegativities, while H3v deals with atoms at three bonds of distance and the sum is weighted with atomic van der Waals volumes. This type of elaborated three-dimensional descriptors are able to determine the entire shape and size of the inhibitor. The next descriptor appearing in Eq. 12, the radial distribution function[14] of an ensemble of atoms, can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius. For RDF115u, the sphere radius is of 11.5 angstroms and atomic weights are not used during its calculation; the index considers the distribution of the geometric distances in the molecular geometry. The descriptor De is a global WHIM descriptor[15] weighted by atomic Sanderson electronegativities. It is based on the projections of the atoms through principal axes or principal components, with the main purpose of capturing relevant three-dimensional information, in present case the atomic distribution with respect to the invariant reference. This numerical variable points to the influence of the molecular conformation of NNRTI during its interaction with RT. Finally, the SRW05 is a molecular walk count type descriptor,[16] counting the number of walks of length 5 that start and end at the same vertex.

3. Data set and methodology

In order to characterize the inhibitory potencies of the NNRTI we modeled the experimental activities based on whole cell antiinfectivity assays, expressed as pIC_{90} [mM] = $-\log(IC_{90}$ [mM]), and obtained on the basis of a well-known protocol.[17] The training sets comprising 154 pIC_{90} values for wild-type and 56 pIC_{90} for K-103N were collected from the literature.[18–25] All the structures under study are listed in Tables 1 and 2 together with their experimental values and are also drawn in Figure 2. These derivatives of Efavirenz involve: substitutions in the aromatic ring of Efavirenz or the replacement of the acetylenic side chain in it with a *trans*-olefin; tricyclic quinazolinones; 3-alkoxymethyl or 3-aryloxymethyl-2-pyridinones; and 3-alkylbenzoxazepinones. The data presented for Efavirenz, DPC961, DPC 083, compounds 64, 65, and 76 reflect values determined for a single enantiomer, whereas those shown for all the other compounds correspond to racemic mixtures. The experimental biological evaluation of each of the enantiomers of Efavirenz, quinozalinones, and 4,1-benzoxazepinones has determined that only the *S* enantiomer is active.

The structures of the compounds were preoptimized by means of the molecular mechanics force field (MM+) included in Hyperchem version 6.03.[26] Since various molecules contain sulfur atoms, final refined molecular structures were obtained using the semiempirical method PM3 (Parametric Method-3).[27] We chose a gradient norm limit of 0.01 kcal/Å for the geometry optimization.

Several types of molecular descriptors were derived, such as constitutional, topological, geometrical, charge, GETAWAY (GEometry, Topology, and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk counts, BCUT descriptors, 2D-Autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups, and atom centered fragments, by means of the software Dragon version 5 available in the Web for evaluation.[28] We excluded the empirical and property-based descriptors. We also decided to add three quantum-chemical descriptors to the pool (not provided by this program): homo and lumo energies, and homo–lumo gap. Descriptors with same entries for most of the training compounds were removed from the pool of variables considered, thus leading in some situations to a better validation of the models with the cross validation leave-*n*%-out (*l-n*%*-o*) and leave-one-out (*loo*) procedures. In the next subsections the number of random validation cases studied for *l-n*%*-o* was of 100,000 for any of the models proposed. We briefly describe the different techniques employed.

3.1. The forward stepwise regression

The forward stepwise regression procedure[29] consists simply in a step-by-step addition of the best descriptors to the model that leads to the smallest standard deviation (*S*), until there is no-other variable outside the equation that satisfies the selection criterion. The SR

Table 5. Experimental and predicted values of pIC$_{90}$ [mM] for the training set of 100 NNRTI of wild-type HIV-1 RT

No.	pIC$_{90}$	
	Exp	Pred
1	5.363	5.465 (−0.102)
3	5.249	5.156 (0.093)
4	5.178	5.351 (−0.173)
6	5.652	5.419 (0.233)
7	5.34	5.049 (0.291)
9	4.686	4.947 (−0.261)
11	4.475	4.248 (0.227)
12	4.561	4.567 (−0.006)
14	4.989	4.738 (0.251)
17	4.987	5.037 (−0.05)
18	5.134	5.141 (−0.007)
19	4.555	4.978 (−0.423)
20	5.079	5.057 (0.022)
21	4.724	4.053 (0.671)
22	5.502	5.184 (0.318)
23	4.745	5.021 (−0.276)
25	4.708	5.165 (−0.457)
26	3.379	4.518 (−1.139)
27	3.914	4.325 (−0.411)
28	4.544	5.065 (−0.521)
30	5.143	5.186 (−0.043)
31	3.604	4.032 (−0.428)
32	5.148	4.975 (0.173)
34	5.456	4.909 (0.547)
35	5.102	4.862 (0.24)
36	5.149	4.722 (0.427)
38	5.167	5.253 (−0.086)
39	5.444	5.321 (0.123)
40	5.086	5.125 (−0.039)
42	3.74	4.226 (−0.486)
43	2.997	3.829 (−0.832)
44	3.366	3.94 (−0.574)
45	3.237	3.172 (0.065)
46	3.697	3.831 (−0.134)
48	4.495	4.269 (0.226)
50	4.377	3.656 (0.721)
51	4.337	4.328 (0.009)
54	3.804	3.435 (0.369)
56	3.721	3.536 (0.185)
58	2.854	3.894 (−1.04)
59	4.614	4.271 (0.343)
61	5.013	4.58 (0.433)
63	5.041	4.744 (0.297)
64	5.658	5.16 (0.498)
65	5.155	4.968 (0.187)
67	4.764	4.898 (−0.134)
68	4.684	4.681 (0.003)
71	4.058	5.182 (−1.124)
73	5.276	5.02 (0.256)
74	5.237	5.142 (0.095)
75	5.409	4.947 (0.462)
76	5.602	5.285 (0.317)
81	5.678	5.536 (0.142)
82	5.699	5.114 (0.585)
83	5.268	4.929 (0.339)
84	5.401	4.832 (0.569)
85	3.991	5.049 (−1.058)
86	5.42	5.344 (0.076)
88	5.35	5.711 (−0.361)
90	4.914	4.818 (0.096)
91	5.592	4.782 (0.81)
92	5.588	5.395 (0.193)
93	5.599	5.581 (0.018)
94	5.533	5.379 (0.154)
95	5.593	5.539 (0.054)
96	5.169	5.122 (0.047)
97	5.652	5.093 (0.559)
98	5.421	5.663 (−0.242)
100	5.652	5.8 (−0.148)
101	5.631	5.623 (0.008)
102	4.738	5.317 (−0.579)
104	4.992	4.796 (0.196)
105	4.997	5.182 (−0.185)
106	5.114	5.045 (0.069)
109	3.778	4.416 (−0.638)
110	5.409	5.1 (0.309)
111	5.362	5.217 (0.145)
112	5.252	4.985 (0.267)
114	5.504	4.716 (0.788)
115	4.864	4.948 (−0.084)
117	4.212	4.556 (−0.344)
118	4.123	4.718 (−0.595)
121	5.053	5.114 (−0.061)
123	5.812	5.634 (0.178)
124	5.189	5.697 (−0.508)
127	4.77	4.79 (−0.02)
129	4.658	4.785 (−0.127)
131	4.276	4.368 (−0.092)
132	4.495	4.585 (−0.09)
133	4.678	4.455 (0.223)
138	4.721	4.593 (0.128)
139	4.538	4.36 (0.178)
140	5.046	5.052 (−0.006)
143	3.801	4.339 (−0.538)
145	4.523	4.434 (0.089)
146	4.357	4.842 (−0.485)
148	4.721	5.197 (−0.476)
151	5.046	4.908 (0.138)
152	5.097	4.562 (0.535)
153	4.959	4.547 (0.412)

technique requires much less linear regressions than a full search of optimal variables (FS).

3.2. The replacement method

Because the selection of the best small subset of descriptors from thousands of them is a time-consuming task, we proposed the replacement method (RM).[30–33] The RM yields results comparable to a full search of optimal variables with much less linear regressions. The main idea behind the RM is that one can approach the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$. In other words, we should find the global minimum of $S(\mathbf{d})$ in a subspace of $D!/[d!(D-d)!]$ points $\mathbf{d}$, where D represents the total number of available descriptors.

The RM consists of the following simple steps: we first choose a set $\mathbf{d}$ at random and do a linear regression. Then we choose one of the descriptors of this set, say X_i, and replace it with each of the D descriptors of the pool $\mathbf{D} = \{X_1, X_2, \ldots, X_D\}$, $D \gg d$ (except itself) keeping

Table 6. Experimental and predicted values of pIC_{90} [mM] for the test set of 54 NNRTI of wild-type HIV-1 RT

No.	pIC_{90}	
	Exp	Pred
2	5.379	5.434 (−0.055)
5	6.081	5.513 (0.568)
8	5.475	5.561 (−0.086)
10	4.340	4.633 (−0.293)
13	5.045	4.705 (0.340)
15	3.529	4.103 (−0.574)
16	3.301	4.151 (−0.850)
24	4.853	5.011 (−0.158)
29	5.699	4.708 (0.991)
33	4.571	4.589 (−0.018)
37	5.444	5.467 (−0.023)
41	5.337	5.032 (0.305)
47	4.022	4.148 (−0.126)
49	4.432	3.926 (0.506)
52	4.310	3.883 (0.427)
53	4.215	4.273 (−0.058)
55	3.796	3.561 (0.235)
57	3.237	3.463 (−0.226)
60	5.041	4.359 (0.682)
62	4.323	4.972 (−0.649)
66	5.076	4.983 (0.093)
69	4.775	4.413 (0.362)
70	3.723	4.936 (−1.213)
72	5.284	5.145 (0.139)
77	5.108	4.985 (0.123)
78	5.119	5.303 (−0.184)
79	5.602	5.134 (0.468)
80	5.284	5.498 (−0.214)
87	5.420	5.532 (−0.112)
89	5.175	5.435 (−0.260)
99	5.493	5.680 (−0.187)
103	5.236	5.211 (0.025)
107	5.095	4.849 (0.246)
108	4.747	4.700 (0.047)
113	5.572	5.354 (0.218)
116	5.022	4.609 (0.413)
119	3.187	3.695 (−0.508)
120	5.526	5.322 (0.204)
122	4.923	4.697 (0.226)
125	5.197	5.104 (0.093)
126	5.721	5.159 (0.562)
128	4.721	4.845 (−0.124)
130	4.678	4.478 (0.200)
134	4.638	4.813 (−0.175)
135	4.569	4.389 (0.180)
136	4.284	4.943 (−0.659)
137	4.071	4.333 (−0.262)
141	4.824	4.859 (−0.035)
142	4.959	4.604 (0.355)
144	3.979	4.206 (−0.227)
147	4.959	5.061 (−0.102)
149	4.824	4.439 (0.385)
150	5.097	5.080 (0.017)
154	4.959	4.763 (0.196)

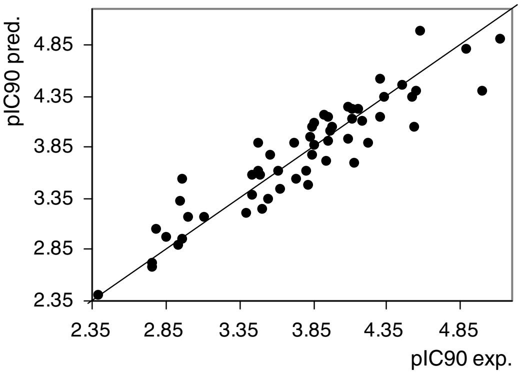

Figure 7. Predicted and experimental potencies for 56 NNRTI (K-103N HIV-1 RT).

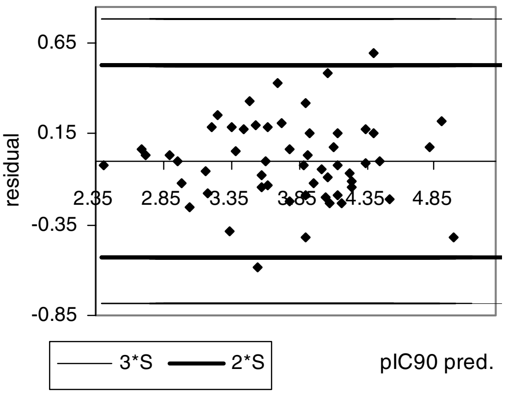

Figure 8. Dispersion plot of the residuals for 56 NNRTI of K-103N.

the best resulting set (i.e., that with smallest *S*). Since one can start replacing any of the *d* descriptors in the initial model, then there will be *d* possible paths. Choose the variable in the resulting model with greatest relative error in its coefficient (omitting the one replaced in the previous step) and replace it with all the *D* descriptors (except itself) keeping again the best set. Replace all the remaining variables in the same way by omitting those replaced in previous steps. When finishing, start again with the variable having the greatest relative error in the coefficient and repeat the whole process. Repeat this process as many times as necessary until the set of descriptors remains unchanged. At the end, we have the best model for the path *i*. Proceed in exactly the same way for all possible paths $i = 1,2,\ldots,d$, compare the resulting models, and keep the best one. Our numerical experiments show that in this way one obtains a model almost as good as the best FS one with much less than $D!/[d!(D-d)!]$ linear regressions when this combinatorial number is large. We may carry out this calculation for $d = 1,2,\ldots$ in order to obtain the overall best model.

3.3. Genetic algorithm approach

The genetic algorithms are governed by biological evolution rules.[34] They are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space.[35] The first step is to create a population of *N* individuals (Fig. 3).

Each individual encodes the same number of randomly chosen descriptors, and the fitness of each individual in this generation is determined. In the second step, a

Table 7. Correlation matrix for Eq. 12

	pIC_{90}	De	SRW05	RDF115u	Mor02e	H-051	H3v	H8e
pIC_{90}	1.0000	0.1903	0.3350	0.1390	0.1412	0.3640	0.0866	0.1308
De			0.0787	0.3149	0.0358	0.0553	0.0981	0.0807
SRW05				0.1526	0.1099	0.4125	0.2014	0.0611
RDF115u				1.0000	0.5161	0.0394	0.5786	0.4094
Mor02e					1.0000	0.2601	0.3753	0.7148
H-051						1.0000	0.3092	0.0711
H3v							1.0000	0.2819
H8e								1.0000

fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from two or one of its parents. We also included elitism which protects the fittest individual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. These selection, crossover and mutation processes are repeated until all of the *N* parents in the population are replaced by their children. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until a 90% of the generations showed the same target fitness score.[36]

The GA implemented in this paper was previously reported by two of the authors[37] and was programmed within the Matlab environment using genetic algorithm[38] and neural networks tool boxes.[39,40] The basic design of the implemented GA is summarized in the flow diagram shown in Figure 4.

4. Conclusions

The inhibitory potency pIC_{90} [mM] was modeled for 154 NNRTI of wild-type HIV-1 RT and 56 NNRTI of the K-103N mutant form, collected from the literature of the last years, using for this three different lineal regression methods: forward stepwise regression, the replacement method, and genetic algorithms. The QSAR-models obtained by exploring 1494 molecular descriptors by RM have shown the best statistical parameters. The relationships found are able to predict as potent those compounds that present favorable activity values, and as inactive compounds those molecules in agreement with experimental evidence.

Acknowledgment

P.R.D. thanks to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) for a postdoctoral research fellowship.

References and notes

1. Jonckheere, H.; Anné, J.; De Clercq, E. *Med. Res. Rev.* **2000**, *20*, 129.
2. Stazewski, S.; Morales-Ramirez, J.; Tashiima, K. T.; Rachlis, A.; Siest, D.; Stanford, J.; Stryker, R.; Johnson, P.; Labriola, D. F.; Farina, D.; Marnion, D. J.; Ruiz, N. M. *N. Engl. J. Med.* **1999**, *34*, 1865.
3. Panel on Clinical Practices for the Treatment of HIV Infection. Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents; February 5, 2001, <http://www.hivatis.org>.
4. Young, S. D.; Britcher, S. F.; Tran, L. O.; Payne, L. S.; Lumma, W. C.; Lyle, T. A.; Huff, J. R.; Anderson, P. S.; Olsen, D. B.; Carroll, S. S.; Pettibone, D. J.; O'Brien, J. A.; Ball, R. G.; Balani, S. K.; Lin, J. H.; Chen, I-W.; Scheif, W. A.; Sardana, V. V.; Long, W. J.; Brynes, V. W.; Emini, E. A. *Antimicrob. Agents Chemother.* **1995**, *39*, 2602.
5. The Body, TAGline, 5, 1998, <http://www.thebody.com/tag/oct98.html>.
6. Skorobogatov, V. A.; Dobrynin, A. A. *MATCH Commun. Math. Comput. Chem.* **1988**, *23*, 105.
7. Wiener, H. J. *J. Am. Chem. Soc.* **1947**, *69*, 17.
8. Schuur, J.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Model.* **1996**, *36*, 334.
9. Katritzky, A. R.; Gordeeva, E. V. *J. Chem. Inf. Model.* **1993**, *33*, 835.
10. Kier, L. B.; Hall, L. H.; Frazer, J. W. *J. Math. Chem.* **1991**, *7*, 229.
11. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Model.* **2003**, *43*, 579.
12. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R. *J. Chem. Inf. Model.* **1989**, *29*, 163.
13. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Model.* **2002**, *42*, 693.
14. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151.
15. Todeschini, R.; Lasagni, M.; Marengo, E. *J. Chemom.* **1994**, *8*, 263.
16. Rücker, G.; Rücker, C. *J. Chem. Inf. Model.* **1993**, *33*, 683.
17. Bacheler, L. T.; Paul, M.; Jadhav, P. K.; Otto, M.; Miller, J. *Antiviral Chem. Chemother.* **1994**, *5*, 111.
18. Patel, M.; McHugh, R. J., Jr.; Cordova, B. C.; Klabe, R. M.; Erickson-Viitanen, S. K.; Trainor, G. L.; Ko, S. S. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 3221.
19. Patel, M.; Ko, S. S.; McHugh, R. J., Jr.; Markwalder, J. A.; Srivastava, A. S.; Cordova, B. C.; Klabe, R. M.; Erickson-Viitanen, S. K.; Trainor, G. L.; Seitz, S. P. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 2805.
20. Corbett, J. W.; Pan, S.; Markwalder, J. A.; Cordova, B. C.; Klabe, R. M.; Garber, S.; Rodgers, J. D.; Erickson-Viitanen, S. K. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 211.
21. Corbett, J. W.; Kresge, K. J.; Pan, S.; Cordova, B. C.; Klabe, R. M.; Rodgers, J. D.; Erickson-Viitanen, S. K. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 309.
22. Cocuzza, A. J.; Chidester, D. R.; Cordova, B. C.; Klabe, R. M.; Jeffrey, S.; Diamond, S.; Weigelt, C. A.; Ko, S. S.; Bacheler, L. T.; Erickson-Viitanen, S. K.; Rodgers, J. D. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1389.

23. Cocuzza, A. J.; Chidester, D. R.; Cordova, B. C.; Jeffrey, S.; Parsons, R. L.; Bacheler, L. T.; Erickson-Viitanen, S. K.; Trainor, G. L.; Ko, S. S. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1177.
24. Patel, M.; McHugh, R. J., Jr.; Cordova, B. C.; Klabe, R. M.; Erickson-Viitanen, S. K.; Trainor, G. L.; Rodgers, J. D. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1729.
25. Patel, M.; McHugh, R. J., Jr.; Cordova, B. C.; Klabe, R. M.; Bacheler, R. T.; Erickson-Viitanen, S. K.; Rodgers, J. D. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1943.
26. Hyperchem, HyperCube, <http://www.hyper.com>.
27. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
28. Dragon 5.0, <http://www.disat.unimib.it/chm>.
29. Draper, N. R.; Smith, H. In *Applied Regression Analysis*; John Wiley & Sons: New York, 1981.
30. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. *Chem. Phys. Lett.* **2005**, *412*, 376.
31. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179.
32. Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. *Chemometr. Intell. Lab.* **2006**, *81*, 180.
33. Duchowicz, P. R.; Castañeta, H. M.; Castro, E. A.; Fernández, F. M.; Vicente, J. L. *Atmos. Environ.* **2006**, *40*, 2929.
34. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
35. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 5246.
36. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**, *45*, 190.
37. Caballero, J.; Fernández, M. *J. Mol. Mod.* **2005**, *12*, 168.
38. Matlab 7.0, The Math Works, Inc., 2004.
39. The MathWorks Inc. Genetic algorithm and direct search toolbox user's guide for use with MATLAB, The Mathworks Inc., Massachusetts, 2004.
40. The MathWorks Inc. Neural network toolbox user's guide for use with MATLAB, The Mathworks Inc., Massachusetts, 2004.