Chem Biol Drug Des 2011; **77**: 343–360

Research Article

doi: 10.1111/j.1747-0285.2011.01109.x

Simplified Molecular Input-Line Entry System and International Chemical Identifier in the QSAR Analysis of Styrylquinoline Derivatives as HIV-1 Integrase Inhibitors

Alla P. Toropova[1], Andrey A. Toropov[1,*], Emilio Benfenati[1] and Giuseppina Gini[2]

[1]*Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy*
[2]*Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy*
**Corresponding author: Andrey A. Toropov, andrey.toropov@marionegri.it*

The simplified molecular input-line entry system (SMILES) and IUPAC International Chemical Identifier (InChI) were examined as representations of the molecular structure for quantitative structure–activity relationships (QSAR), which can be used to predict the inhibitory activity of styrylquinoline derivatives against the human immunodeficiency virus type 1 (HIV-1). Optimal SMILES-based descriptors give a best model with $n = 26$, $r^2 = 0.6330$, $q^2 = 0.5812$, $s = 0.502$, $F = 41$ for the training set and $n = 10$, $r^2 = 0.7493$, $r^2_{pred} = 0.6235$, $R^2_m = 0.537$, $s = 0.541$, $F = 24$ for the validation set. Optimal InChI-based descriptors give a best model with $n = 26$, $r^2 = 0.8673$, $q^2 = 0.8456$, $s = 0.302$, $F = 157$ for the training set and $n = 10$, $r^2 = 0.8562$, $r^2_{pred} = 0.7715$, $R^2_m = 0.819$, $s = 0.329$, $F = 48$ for the validation set. Thus, the InChI-based model is preferable. The described SMILES-based and InChI-based approaches have been checked with five random splits into the training and test sets.

Key words: anti-HIV-1 inhibitory activity, InChI, optimal descriptor, QSAR, SMILES

Received 23 December 2009, revised 15 February 2011 and accepted for publication 20 February 2011

Quantitative structure–property/activity relationships (QSPR/QSAR) are tools of modern research in the fields of chemistry, biochemistry, and ecology. Some models use a large number of substances (1), while in other cases, it is preferable or necessary because of the limited number of examples to use a small set of compounds (2). Establishing correlations between the molecular structure and a rare biochemical activity for a small set of compounds is just as important as for large arrays of chemicals.

The inhibitory activity of 36 styrylquinoline derivatives (Table 1) against the human immunodeficiency virus (HIV-1), studied in Ref. (3), can be used for an experiment to establish robust correlations between the molecular structure and the activity.

Representation of the molecular structure is an important component of the QSPR/QSAR analyses, and the molecular graph is the most widely used representation (4–17). Being a convenient mathematical tool, the molecular graph required operations with the adjacency matrix in which majority of elements are equal to zero (18). For this reason, the simplified molecular input-line entry system (SMILES[a]) (19–21) and IUPAC International Chemical Identifier (InChI) (22,23) are widely used in databases available on the Internet for the physicochemical and biochemical endpoints[b,c]. Thus, searching for algorithms to establish correlations between molecular structures represented by SMILES or InChI and various end-points is a logical way to develop QSPR/QSAR analyses.

Both the SMILES and the InChI are tools to describe the molecular structure by means of a sequence of symbols (19–23). The SMILES is a more convenient representation for the understanding by human. The InChI is a more complex representation able to provide a unique representation of the molecular structure (22,23). For example, the representation of 2-methylbutane by SMILES is 'CC(C)CC'[a]; the representation of this molecule by means of the InChI is 'InChI=1/C5H12/c1-4-5(2)3/h5H,4H2,1-3H3'[a]. In other words, InChI is a more detailed representation of the molecular structure (22,23).

Optimal descriptors (24–29) can be reorganized so that they can be calculated with a representation of the molecular structure by SMILES (30,31) and/or InChI (32,33). The optimal SMILES-based descriptors can provide robust prediction for toxicity (31). The optimal InChI-based descriptors can be better predictors for octanol–water partition coefficient (32) and for solubility (33) than the SMILES-based optimal descriptors.

Table 1: Molecular structure of styrylquinoline derivatives

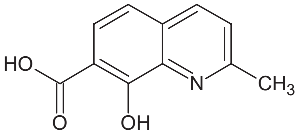
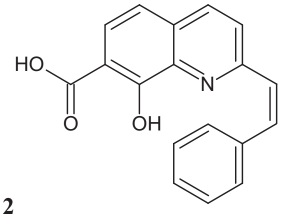
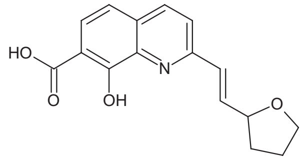
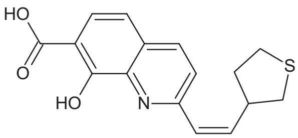
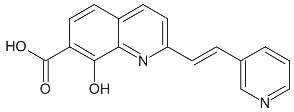
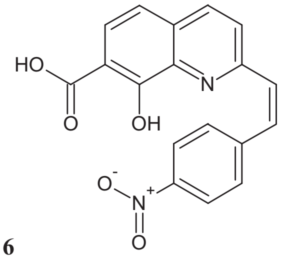
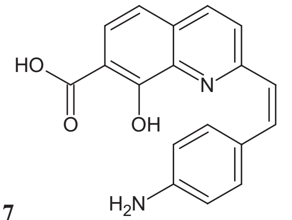
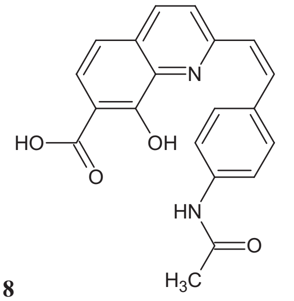
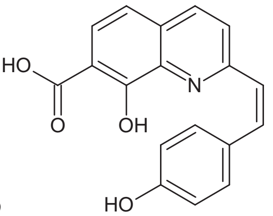
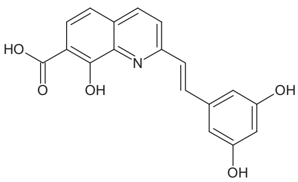
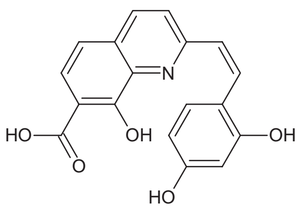
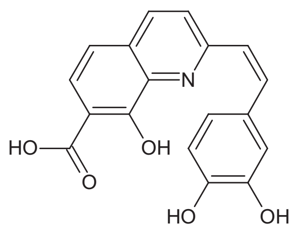
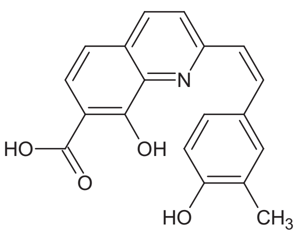
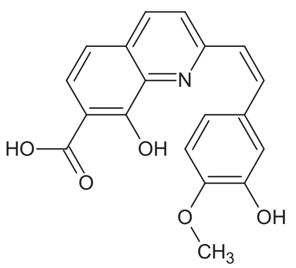
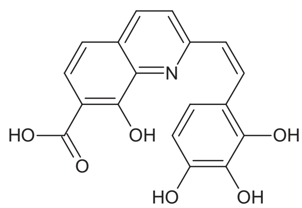
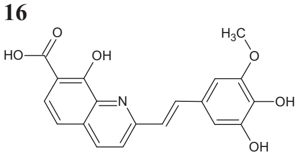
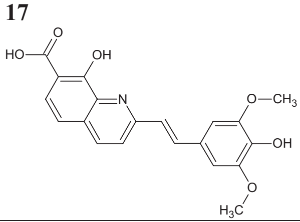
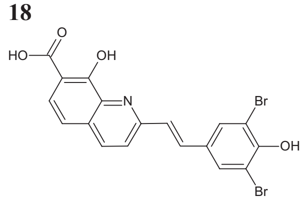
1 	 **2**	**3** 
4 	**5** 	 **6**
 **7**	 **8**	 **9**
10 	**11** 	**12** 
13 	**14** 	**15** 
16 	**17** 	**18** 

Table 1: (Continued)

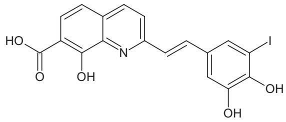
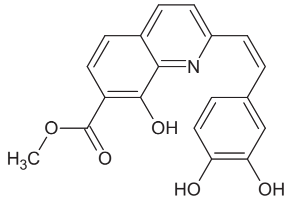
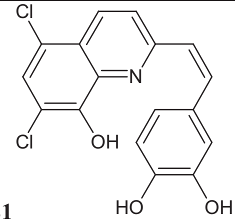
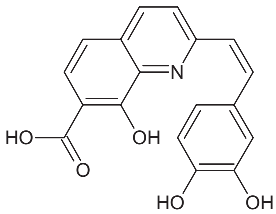
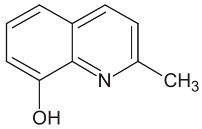
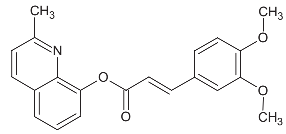
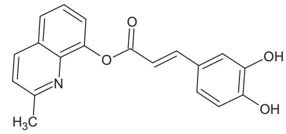
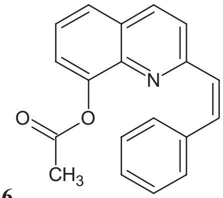
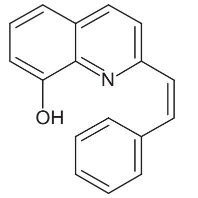
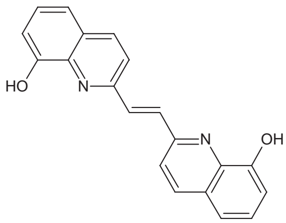
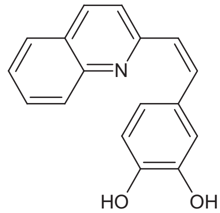
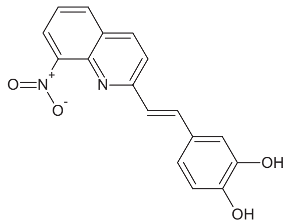
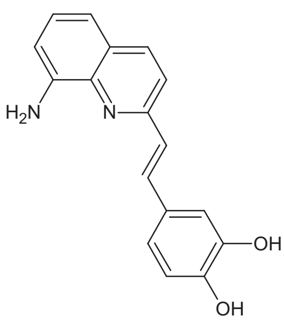
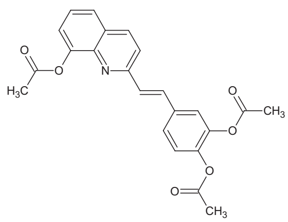
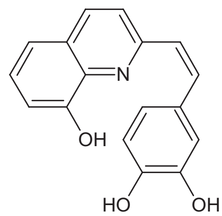
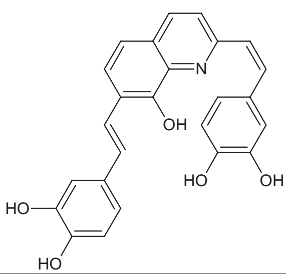
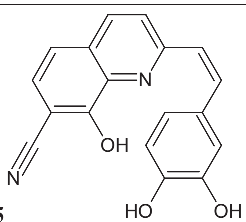
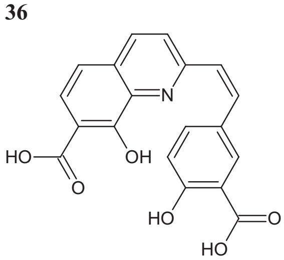
19 	**20** 	 **21**
22 	 **23**	**24** 
25 	 **26**	 **27**
6 **28** 	 **29**	**30** 
31 	**32** 	 **33**
34 	 **35**	**36** 

Table 2: Five splits into the training and test sets [split 1 has been taken from Ref. (3)]

Number of split	Training set	Test set
1	2,3,4,5,8,9,11,12,14,15,16,17,18,19,20,23, 24,25,27,28,30,31,33,34,35,36	1,6,7,10,13,21,22,26,29,32
2	3,4,5,8,9,11,12,14,15,16,17,18,19,20,23,24, 25,27,28,30,31,32,33,34,35,36	1,2, 6,7,10,13,21,22,26,29
3	2,4,5,8,9,11,12,14,15,16,17,18,19,20,23,24,25,27,28,29,30,31,33,34,35,36	1,3, 6,7,10,13,21,22,26,32
4	1,2,3,4,5, 8,10,11,12,13,14,16,17,18,20,23, 25,27,30,33,34,35,36,21,22,32	6,7,9,15,19,24,28,31,26,29
5	2,3,4,5,8,10,11,12,13,16,17,19,20,21, 22,23,24,25, 28,29,30,31,32,34,35,36	1,6,7,9,14,15,18,26,27,33

Table 3: Correlation weights of simplified molecular input-line entry system (SMILES) attributes obtained in the first probe of the Monte Carlo optimization method with threshold equal to 4. *N*(TRN) and *N*(VLD) are the numbers of SMILES that contain the given Sk, in the training and validation sets, respectively

ID	Sk	CW(Sk)	*N*(TRN)	*N*(VLD)
1	(xxxBrxx(xxx	0.0	1	0
2	(xxxClxx(xxx	0.0	0	1
3	(xxxCxxx#xxx	0.0	1	0
4	(xxxCxxx(xxx	0.3498468	22	10
5	(xxxNxxx#xxx	0.0	1	0
6	(xxxNxxx(xxx	0.0	1	0
7	(xxxOxxx(xxx	1.7962853	20	7
8	(xxxcxxx(xxx	2.4025995	14	6
9	+xxx[xxx(xxx	0.0	1	1
10	−xxx[xxx(xxx	0.0	1	0
11	1xxx2xxx(xxx	0.0	2	1
12	1xxxCxxx(xxx	0.0	0	1
13	1xxxOxxx(xxx	0.2969450	16	3
14	1xxxcxxx(xxx	1.2009262	8	3
15	1xxxcxxx/xxx	0.0	1	0
16	2xxx(xxx/xxx	0.0	1	0
17	2xxxOxxx(xxx	0.0	3	0
18	2xxxcxxx(xxx	2.3961978	19	8
19	2xxxcxxx1xxx	0.0	3	1
20	3xxxCxxx/xxx	0.0	1	0
21	3xxxcxxx2xxx	0.0	2	0
22	3xxxcxxx/xxx	0.0	1	0
23	3xxxcxxx(xxx	0.7204912	19	8
24	3xxxnxxx2xxx	1.2047657	16	8
25	4xxxcxxx(xxx	0.0	2	0
26	4xxxnxxx3xxx	0.0	1	0
27	=xxxCxxx(xxx	1.0209640	24	10
28	=xxxCxxx/xxx	1.6961277	4	0
29	=xxxOxxx(xxx	0.3048762	15	7
30	Brxxcxxx1xxx	0.0	1	0
31	Cxxx(xxx2xxx	0.9012408	13	6
32	Cxxx(xxx=xxx	0.4272772	21	9
33	Cxxx(xxx1xxx	0.0	3	3
34	Cxxx(xxxCxxx	0.0	2	2
35	Cxxx/xxxCxxx	0.0	1	0
36	Cxxx/xxx(xxx	2.1969389	4	0
37	Cxxx3xxxCxxx	0.0	2	0
38	Cxxx=xxx(xxx	0.2993007	21	9
39	Cxxx=xxxCxxx	2.0963202	4	0
40	CxxxCxxx3xxx	0.0	2	0
41	CxxxCxxxCxxx	0.0	1	0
42	CxxxCxxx(xxx	0.0	1	2
43	CxxxOxxx(xxx	0.0	3	1
44	CxxxOxxx1xxx	0.0	1	1
45	CxxxOxxx3xxx	0.0	1	0
46	CxxxSxxxCxxx	0.0	1	0
47	Cxxx\xxxCxxx	0.0	1	0
48	Ixxxcxxx1xxx	0.0	1	0
49	Nxxx#xxxCxxx	0.0	1	0
50	Nxxx[xxx(xxx	0.0	1	0
51	Nxxxcxxx1xxx	0.0	1	1
52	Oxxx(xxxNxxx	0.0	1	0
53	Oxxx(xxxCxxx	1.6497003	19	5
54	Oxxx(xxx/xxx	0.0	2	0
55	Oxxx(xxxOxxx	0.3043891	14	7
56	Oxxx=xxx(xxx	0.2985726	15	7
57	Oxxx=xxxCxxx	2.4028279	5	1
58	OxxxCxxx(xxx	0.0	3	1
59	OxxxCxxxCxxx	0.0	1	0
60	Oxxx[xxx(xxx	0.0	1	0
61	Oxxxcxxx2xxx	0.0	3	0
62	Oxxxcxxx1xxx	0.4962099	17	7
63	SxxxCxxx3xxx	0.0	1	0
64	SxxxCxxxCxxx	0.0	1	0
65	[xxx(xxx[xxx	0.0	1	0
66	[xxx(xxx=xxx	0.0	1	1
67	[xxx+xxxNxxx	0.0	1	1
68	[xxx−xxxOxxx	0.0	1	1
69	[xxxNxxx+xxx	0.0	1	1
70	[xxxOxxx−xxx	0.0	1	1
71	[xxx[xxx−xxx	0.0	0	1
72	[xxx[xxxNxxx	0.0	0	1
73	\xxxCxxx=xxx	0.0	1	0
74	\xxxCxxx3xxx	0.0	1	0
75	cxxx(xxxBrxx	0.0	1	0
76	cxxx(xxxOxxx	1.4034395	21	8
77	cxxx(xxx/xxx	0.0	1	0
78	cxxx(xxxcxxx	0.5012076	19	8
79	cxxx(xxxCxxx	0.9463535	22	10
80	cxxx(xxxNxxx	0.0	2	0
81	cxxx(xxxClxx	0.0	0	1
82	cxxx(xxx[xxx	0.0	1	0
83	cxxx/xxxCxxx	0.0	2	0
84	cxxx1xxxcxxx	2.3969701	26	10
85	cxxx1xxxCxxx	0.0	0	1
86	cxxx1xxxOxxx	0.2998894	17	5
87	cxxx1xxx2xxx	0.0	3	1
88	cxxx1xxx(xxx	2.4000354	4	3
89	cxxx2xxxOxxx	0.0	3	0
90	cxxx2xxxcxxx	2.1546596	26	10
91	cxxx2xxx3xxx	0.0	2	0
92	cxxx3xxxcxxx	0.3019973	23	9
93	cxxx3xxxOxxx	0.0	1	0
94	cxxx4xxxcxxx	0.0	2	0

Table 3: (Continued)

ID	Sk	CW(Sk)	N(TRN)	N(VLD)
95	cxxxNxxx(xxx	0.0	1	0
96	cxxxOxxx(xxx	0.0	2	2
97	cxxxOxxxCxxx	0.0	3	0
98	cxxxcxxx1xxx	0.3028519	26	10
99	cxxxcxxx4xxx	0.0	2	0
100	cxxxcxxx3xxx	0.2955818	23	9
101	cxxxcxxxcxxx	0.2951162	26	10
102	cxxxcxxx2xxx	2.1503686	26	10
103	cxxxcxxx(xxx	0.2981324	26	10
104	cxxxnxxx(xxx	1.0499671	10	2
105	cxxxnxxxcxxx	0.0	1	0
106	nxxx(xxx1xxx	0.0	1	0
107	nxxx(xxxCxxx	0.0	3	1
108	nxxx(xxxcxxx	1.4971106	6	1
109	nxxx2xxx(xxx	2.3952194	12	5
110	nxxx3xxxcxxx	1.4746850	16	8
111	nxxx4xxxcxxx	0.0	1	0
112	nxxxcxxx1xxx	2.4007048	6	1
113	nxxxcxxx2xxx	0.3019832	4	1
114	nxxxcxxx3xxx	0.0	1	0
115	nxxxcxxxcxxx	0.0	1	0

Table 4: Correlation weights of InChI attributes obtained in the first probe of the Monte Carlo optimization method with threshold equal to 2. *N*(TRN) and *N*(VLD) are the numbers of InChI, which contain the given Ik, in training and validation sets, respectively

No.	Ik	CW(Ik)	N(TRN)	N(VLD)
1	(10	1.3932368	8	4
2	(11	0.4974939	12	5
3	(12	1.9774505	13	5
4	(13	0.3129114	5	2
5	(14	0.3077129	3	1
6	(15	1.2997495	2	1
7	(16	2.1315503	3	1
8	(17	2.2335049	3	1
9	(18	0.5823361	11	6
10	(19	1.1208031	11	5
11	(20	0.3235466	12	3
12	(21	1.3010836	16	5
13	(22	0.3142406	13	6
14	(23	0.3067574	13	5
15	(24	0.3123641	9	1
16	(25	2.0198240	3	0
17	(26	0.4033249	2	0
18	(27	0.0	1	0
19	(28	0.0	1	1
20	(29	0.0	1	0
21	(30	0.0	1	0
22	(2	1.1222480	4	3
23	(3	0.0	0	1
24	(4	0.0	0	1
25	(5	0.0	0	1
26	(7	0.0	0	2
27	(8	2.3783567	4	0
28	(9	2.3835241	5	1
29	(	0.7559318	26	10
30	+	1.1153524	3	0
31	,10	0.0	1	0
32	,12	2.3837917	2	0
33	,13	0.0	0	1
34	,14	0.0	0	1
35	,15	0.0	0	1
36	,18	1.0303879	3	0
37	,19	2.3780177	4	2
38	,20	1.5712777	10	2
39	,21	1.2849163	10	4
40	,22	2.3762251	5	2
41	,23	1.5342150	7	5
42	,24	1.0657276	10	3
43	,25	0.5777748	7	0
44	,26	0.3062678	2	0
45	,27	0.0	1	0
46	,1	0.5246861	23	10
47	,2	0.3344974	7	3
48	,3	0.0	1	0
49	,7	0.0	1	0
50	,9	0.0	1	0
51	,	2.2841207	15	6
52	−10	0.4561746	24	9
53	−11	0.3149572	24	8
54	−12	0.7299485	23	9
55	−13	0.3128622	23	9
56	−14	0.3075480	21	9
57	−15	0.3107120	19	7
58	−16	1.4784451	7	4
59	−17	0.6666253	5	2
60	−18	0.3123427	3	2
61	−19	1.8412315	4	1
62	−20	0.0	1	2
63	−21	2.1556832	8	1
64	−22	0.3122587	5	3
65	−23	0.6729373	7	1
66	−24	1.4761540	2	0
67	−31	0.0	1	0
68	−1	1.2958900	4	0
69	−2	1.8402775	23	7
70	−3	0.6135869	26	9
71	−4	0.3147165	25	10
72	−5	1.4068876	26	9
73	−6	2.2009127	25	9
74	−7	2.3790340	25	7
75	−8	0.3078477	24	10
76	−9	0.8453533	25	9
77	−	0.0	1	0
78	/	1.2240344	26	10
79	0	0.3109460	11	8
80	1	0.7651237	26	10
81	2	0.3056813	22	9
82	3	0.8309131	10	5
83	4	0.3122884	15	5
84	5	1.7820982	15	8
85	6	0.3075810	21	6
86	7	0.3084866	20	9
87	8	2.3846030	14	4
88	9	0.3112684	14	5
89	Br	0.0	1	0
90	C10	0.0	1	0

Table 4: (Continued)

No.	Ik	CW(Ik)	N(TRN)	N(VLD)
91	C11	0.0	0	1
92	C16	1.2726048	2	0
93	C17	0.3952620	5	2
94	C18	2.3774174	8	4
95	C19	0.3058109	5	2
96	C20	0.3097077	3	0
97	C21	0.0	1	0
98	C23	0.0	0	1
99	C25	0.0	1	0
100	Cl	0.0	0	1
101	H11	0.0	1	1
102	H12	1.4787828	4	1
103	H13	2.3846044	8	3
104	H14	0.3050102	2	1
105	H15	0.4862685	6	2
106	H16	0.0	1	0
107	H17	0.0	1	0
108	H19	1.3432290	2	1
109	H2	0.3147637	23	9
110	H3	0.9420254	8	4
111	H9	0.0	1	1
112	H	1.6007745	26	10
113	I	0.0	1	0
114	N2	1.2219115	6	2
115	N	1.4235922	20	8
116	O2	0.3086935	2	2
117	O3	2.3847641	5	3
118	O4	0.3058276	7	1
119	O5	0.3103874	6	3
120	O6	0.5753524	4	1
121	O	2.3782819	2	0
122	S	0.0	1	0
123	b12	0.0	1	0
124	b4	0.0	1	0
125	b6	0.0	1	0
126	b7	0.0	1	0
127	c18	1.3034296	2	0
128	c20	0.0	1	0
129	c23	0.0	1	0
130	c1	1.4752837	22	10
131	h1	1.9346553	3	0
132	h2	2.2280274	14	7
133	h3	1.5709405	7	2
134	h4	0.0	1	0
135	h5	0.0	1	1

The aim of the present study was to compare the statistical characteristics of QSARs for anti-HIV-1 activity of styrylquinoline derivatives calculated with the optimal SMILES-based and InChI-based descriptors.

Method

Anti-HIV-1 integrase inhibitory activity data, minus decimal logarithm of 50% effective concentration, and pEC_{50} have been taken from a report of Leonard and Roy (3). Split into the training and the test sets from Ref. (3) and four additional random splits were examined in the present study (Table 2). It is to be noted that the absolutely random split for 36 substances that are examined in the present research is impossible, because 13 substances are characterized by the same value $pEC_{50} = 4$. Thus, five splits are organized in such a way where the mentioned 13 substances are distributed in both the training set (majority) and test set.

The optimal SMILES-based descriptors of correlation weights (DCW) are calculated as the following:

$$\mathrm{DCW(Threshold)} = \sum \mathrm{CW(Sk)} \qquad (1)$$

where Sk is SMILES attribute that includes three SMILES elements. CW(Sk) is the correlation weight of Sk. The SMILES element is one symbol of the SMILES notation or two symbols that cannot be examined separately (e.g. Br, Cl, etc.). For instance, SMILES = 'CN(C)Cl' contains the following elements: C, N, (, C,), Cl, the construction of SMILES attributes containing three elements can be represented as:

```
CxxxNxxx(xxx;

Nxxx(xxxCxxx;

(xxxCxxx)xxx;

Cxxx)xxxClxx.
```

The 'x' indicates a vacant position in the string that represents the attribute.

Table 5: Statistical characteristics of simplified molecular input-line entry system-based models for anti-HIV-1 activity, pEC_{50}. N_{act} is the number of attributes that are not blocked for the given threshold; *r*, *s*, and *F* are correlation coefficient, standard error of estimation, and Fisher *F*-ratio, respectively. The model with the best predictability is indicated in bold

			Training set, *n* = 26			Validation set, *n* = 10		
Threshold	N_{act}	Probe	r^2	*s*	*F*	r^2	*s*	*F*
Split 1								
0	115	1	0.7210	0.438	62	0.5915	0.608	12
		2	0.7232	0.436	63	0.5682	0.618	11
		3	0.7225	0.437	62	0.5893	0.608	11
		Average	0.7222	0.437	62	0.5830	0.611	11

Table 5: (Continued)

Threshold	N_{act}	Probe	Training set, n = 26 r^2	s	F	Validation set, n = 10 r^2	s	F
1	109	1	0.7217	0.437	62	0.6644	0.549	16
		2	0.7206	0.438	62	0.6756	0.541	17
		3	0.7210	0.438	62	0.6410	0.563	14
		Average	0.7211	0.438	62	0.6603	0.551	16
2	63	1	0.6741	0.473	50	0.6929	0.561	18
		2	0.6738	0.474	50	0.6971	0.564	18
		3	0.6734	0.474	49	0.6901	0.565	18
		Average	0.6738	0.474	50	0.6934	0.563	18
3	50	1	0.6739	0.474	50	0.6596	0.582	16
		2	0.6723	0.475	49	0.6546	0.587	15
		3	0.6749	0.473	50	0.6594	0.584	15
		Average	0.6737	0.474	50	0.6579	0.584	15
4	40	1	0.6330	0.502	41	0.7493	0.541	24
		2	0.6337	0.502	42	0.7476	0.548	24
		3	0.6328	0.502	41	0.7425	0.549	23
		Average	0.6332	0.502	41	**0.7464**	**0.546**	**24**
5	35	1	0.5782	0.539	33	0.6847	0.617	17
		2	0.5767	0.540	33	0.7142	0.604	20
		3	0.5741	0.541	32	0.7138	0.606	20
		Average	0.5763	0.540	33	0.7042	0.609	19
Split 2								
0	115	1	0.7270	0.448	64	0.6385	0.493	14
		2	0.7288	0.447	65	0.6339	0.496	14
		3	0.7265	0.449	64	0.6303	0.499	14
		Average	0.7275	0.448	64	0.6342	0.496	14
1	109	1	0.7254	0.450	63	0.7423	0.426	23
		2	0.7252	0.450	63	0.7438	0.425	23
		3	0.7264	0.449	64	0.7336	0.433	22
		Average	0.7257	0.450	63	0.7399	0.428	23
2	65	1	0.6606	0.500	47	0.7593	0.425	25
		2	0.6598	0.501	47	0.7529	0.428	24
		3	0.6586	0.502	46	0.7593	0.424	25
		Average	0.6596	0.501	47	0.7572	0.426	25
3	51	1	0.6379	0.517	42	0.7547	0.431	25
		2	0.6391	0.516	43	0.7582	0.429	25
		3	0.6373	0.517	42	0.7684	0.422	27
		Average	0.6381	0.516	42	0.7604	0.427	25
4	41	1	0.6146	0.533	38	0.7930	0.432	31
		2	0.6154	0.532	38	0.7971	0.430	31
		3	0.6149	0.533	38	0.7996	0.426	32
		Average	0.6150	0.533	38	**0.7966**	**0.429**	**31**
5	34	1	0.5297	0.589	27	0.7671	0.503	26
		2	0.5270	0.590	27	0.7663	0.504	26
		3	0.5269	0.590	27	0.7682	0.502	27
		Average	0.5279	0.590	27	0.7672	0.503	26
Split 3								
0	115	1	0.7382	0.435	68	0.4780	0.654	7
		2	0.7403	0.433	68	0.4899	0.649	8
		3	0.7371	0.436	67	0.4848	0.650	8
		Average	0.7385	0.435	68	0.4842	0.651	8
1	104	1	0.7392	0.434	68	0.4632	0.642	7
		2	0.7401	0.433	68	0.4440	0.649	6
		3	0.7387	0.434	68	0.4729	0.637	7
		Average	0.7393	0.434	68	0.4600	0.642	7
2	60	1	0.6789	0.482	51	0.6419	0.582	14
		2	0.6780	0.482	51	0.6446	0.575	15
		3	0.6777	0.482	50	0.6454	0.579	15
		Average	0.6782	0.482	51	**0.6440**	**0.579**	**14**
3	47	1	0.6673	0.490	48	0.5550	0.612	10

Table 5: (Continued)

			Training set, $n = 26$			Validation set, $n = 10$		
Threshold	N_{act}	Probe	r^2	s	F	r^2	s	F
2	0.6634	0.493	47	0.5338	0.621	9		
3	0.6641	0.492	47	0.5506	0.612	10		
Average	0.6649	0.492	48	0.5465	0.615	10		
4	37	1	0.6499	0.503	45	0.4414	0.672	6
		2	0.6523	0.501	45	0.4641	0.662	7
		3	0.6541	0.500	45	0.4715	0.660	7
		Average	0.6521	0.501	45	0.4590	0.665	7
5	34	1	0.6159	0.527	38	0.3886	0.693	5
		2	0.6153	0.527	38	0.3869	0.692	5
		3	0.6174	0.526	39	0.3775	0.697	5
		Average	0.6162	0.526	39	0.3843	0.694	5
Split 4								
0	115	1	0.6749	0.446	50	0.7979	0.522	32
		2	0.6770	0.444	50	0.7863	0.527	29
		3	0.6726	0.447	49	0.7800	0.529	28
		Average	0.6748	0.446	50	0.7881	0.526	30
1	107	1	0.6747	0.446	50	0.8522	0.482	46
		2	0.6740	0.447	50	0.8719	0.473	54
		3	0.6735	0.447	49	0.8388	0.496	42
		Average	0.6740	0.446	50	0.8543	0.484	47
2	56	1	0.5946	0.498	35	0.9068	0.440	78
		2	0.5954	0.497	35	0.8959	0.451	69
		3	0.5960	0.497	35	0.9027	0.439	74
		Average	0.5953	0.497	35	**0.9018**	**0.443**	**74**
3	48	1	0.5787	0.508	33	0.8767	0.433	57
		2	0.5833	0.505	34	0.8816	0.440	60
		3	0.5756	0.509	33	0.8792	0.436	58
		Average	0.5792	0.507	33	0.8792	0.437	58
4	39	1	0.4823	0.563	22	0.9140	0.408	85
		2	0.4836	0.562	22	0.9105	0.413	81
		3	0.4827	0.562	22	0.9118	0.412	83
		Average	0.4829	0.562	22	0.9121	0.411	83
5	35	1	0.4590	0.575	20	0.8625	0.468	50
		2	0.4598	0.575	20	0.8582	0.470	48
		3	0.4645	0.572	21	0.8592	0.469	49
		Average	0.4611	0.574	21	0.8599	0.469	49
Split 5								
0	115	1	0.7747	0.377	83	0.7310	0.657	22
		2	0.7763	0.376	83	0.7271	0.667	21
		3	0.7679	0.383	79	0.7559	0.657	25
		Average	0.7730	0.379	82	0.7380	0.660	23
1	110	1	0.7782	0.374	84	0.7338	0.705	22
		2	0.7662	0.384	79	0.7571	0.695	25
		3	0.7748	0.377	83	0.7518	0.694	24
		Average	0.7730	0.379	82	0.7476	0.698	24
2	64	1	0.6763	0.452	50	0.7997	0.639	32
		2	0.6789	0.450	51	0.7924	0.649	31
		3	0.6790	0.450	51	0.7902	0.646	30
		Average	0.6781	0.451	51	**0.7941**	**0.645**	**31**
3	50	1	0.6654	0.460	48	0.7818	0.643	29
		2	0.6716	0.455	49	0.7633	0.649	26
		3	0.6601	0.463	47	0.7763	0.650	28
		Average	0.6657	0.459	48	0.7738	0.647	27
4	43	1	0.6307	0.483	41	0.7602	0.679	25
		2	0.6346	0.480	42	0.7705	0.666	27
		3	0.6404	0.477	43	0.7658	0.669	26
		Average	0.6352	0.480	42	0.7655	0.671	26
5	35	1	0.5770	0.517	33	0.7122	0.691	20

Table 6: Statistical characteristics of InChI-based models for anti-HIV-1 activity, pEC_{50}. N_{act} is the number of attributes that are not blocked for the given threshold; *r*, *s*, and *F* are correlation coefficient, standard error of estimation, and Fisher *F*-ratio, respectively. The model with the best predictability is indicated in bold

			Training set, *n* = 26			Validation set, *n* = 10		
Threshold	N_{act}	Probe	r^2	*s*	*F*	r^2	*s*	*F*
Split 1								
0	135	1	0.8994	0.263	215	0.6739	0.513	17
		2	0.8957	0.268	206	0.6834	0.505	17
		3	0.8959	0.268	207	0.6737	0.513	17
		Average	0.8970	0.266	209	0.6770	0.510	17
1	125	1	0.8953	0.268	205	0.7647	0.421	26
		2	0.8948	0.269	204	0.7709	0.414	27
		3	0.9004	0.262	217	0.7590	0.427	25
		Average	0.8968	0.266	209	0.7649	0.420	26
2	95	1	0.8673	0.302	157	0.8562	0.329	48
		2	0.8663	0.303	156	0.8631	0.321	50
		3	0.8664	0.303	156	0.8646	0.318	51
		Average	0.8667	0.303	156	**0.8613**	**0.323**	**50**
3	84	1	0.8526	0.318	139	0.8356	0.375	41
		2	0.8540	0.317	140	0.8383	0.372	41
		3	0.8531	0.318	139	0.8372	0.374	41
		Average	0.8532	0.318	140	0.8370	0.374	41
4	75	1	0.8228	0.349	111	0.8558	0.359	47
		2	0.8223	0.350	111	0.8510	0.364	46
		3	0.8216	0.350	111	0.8582	0.356	48
		Average	0.8223	0.350	111	0.8550	0.360	47
5	69	1	0.7815	0.388	86	0.8373	0.404	41
		2	0.7815	0.388	86	0.8406	0.404	42
		3	0.7826	0.387	86	0.8347	0.409	40
		Average	0.7819	0.387	86	0.8376	0.406	41
Split 2								
0	135	1	0.8897	0.285	194	0.7651	0.401	26
		2	0.8891	0.286	192	0.7661	0.399	26
		3	0.8885	0.287	191	0.7674	0.399	26
		Average	0.8891	0.286	192	0.7662	0.400	26
1	128	1	0.8842	0.292	183	0.8061	0.367	33
		2	0.8853	0.291	185	0.8079	0.366	34
		3	0.8887	0.286	192	0.8029	0.372	33
		Average	0.8861	0.290	187	0.8056	0.369	33
2	98	1	0.8633	0.317	152	0.8855	0.280	62
		2	0.8637	0.317	152	0.8855	0.280	62
		3	0.8635	0.317	152	0.8834	0.282	61
		Average	0.8635	0.317	152	**0.8848**	**0.281**	**61**
3	85	1	0.8494	0.333	135	0.8782	0.299	58
		2	0.8489	0.334	135	0.8768	0.302	57
		3	0.8493	0.333	135	0.8790	0.297	58
		Average	0.8492	0.333	135	0.8780	0.299	58
4	77	1	0.8289	0.355	116	0.8553	0.321	47
		2	0.8245	0.360	113	0.8500	0.327	45
		3	0.8292	0.355	116	0.8528	0.322	46
		Average	0.8275	0.356	115	0.8527	0.323	46
5	69	1	0.7564	0.424	75	0.7923	0.382	31
		2	0.7552	0.425	74	0.7880	0.386	30
		3	0.7561	0.424	74	0.7980	0.378	32
		Average	0.7559	0.424	74	0.7928	0.382	31
Split3								
0	135	1	0.8984	0.271	212	0.6951	0.485	18
		2	0.8999	0.269	216	0.6719	0.506	16
		3	0.8996	0.269	215	0.6654	0.510	16
		Average	0.8993	0.270	214	0.6775	0.500	17
1	123	1	0.8992	0.270	214	0.7743	0.412	27

Table 6: (Continued)

			Training set, n = 26			Validation set, n = 10		
Threshold	N_{act}	Probe	r^2	s	F	r^2	s	F
2	0.8951	0.275	205	0.7694	0.415	27		
3	0.8984	0.271	212	0.7649	0.420	26		
Average	0.8976	0.272	210	0.7695	0.416	27		
2	93	1	0.8583	0.320	145	0.7985	0.386	32
		2	0.8582	0.320	145	0.8011	0.384	32
		3	0.8620	0.316	150	0.8010	0.383	32
		Average	0.8595	0.319	147	0.8002	0.384	32
3	81	1	0.8254	0.355	113	0.8161	0.378	36
		2	0.8258	0.355	114	0.8103	0.382	34
		3	0.8271	0.353	115	0.8170	0.376	36
		Average	0.8261	0.354	114	0.8145	0.379	35
4	75	1	0.8293	0.351	117	0.8571	0.331	48
		2	0.8255	0.355	114	0.8578	0.332	48
		3	0.8286	0.352	116	0.8616	0.327	50
		Average	0.8278	0.353	115	**0.8588**	**0.330**	**49**
5	68	1	0.7729	0.405	82	0.8364	0.392	41
		2	0.7745	0.404	82	0.8407	0.385	42
		3	0.7753	0.403	83	0.8418	0.385	43
		Average	0.7742	0.404	82	0.8396	0.387	42
Split4								
0	135	1	0.8030	0.347	98	0.7920	0.522	30
		2	0.8042	0.346	99	0.7968	0.515	31
		3	0.8041	0.346	99	0.7981	0.514	32
		Average	0.8038	0.346	98	0.7956	0.517	31
1	128	1	0.8029	0.347	98	0.8749	0.452	56
		2	0.8037	0.346	98	0.8726	0.449	55
		3	0.8027	0.347	98	0.8739	0.449	55
		Average	0.8031	0.347	98	0.8738	0.450	55
2	95	1	0.7893	0.359	90	0.8609	0.452	50
		2	0.7842	0.363	87	0.8698	0.446	53
		3	0.7854	0.362	88	0.8691	0.446	53
		Average	0.7863	0.362	88	0.8666	0.448	52
3	79	1	0.7276	0.408	64	0.9215	0.419	94
		2	0.7233	0.411	63	0.9217	0.418	94
		3	0.7230	0.412	63	0.9230	0.419	96
		Average	0.7246	0.410	63	0.9221	0.418	95
4	75	1	0.7008	0.428	56	0.9263	0.446	101
		2	0.7022	0.427	57	0.9264	0.444	101
		3	0.7012	0.428	56	0.9276	0.443	103
		Average	0.7014	0.427	56	**0.9268**	**0.444**	**101**
5	65	1	0.6233	0.480	40	0.8913	0.519	66
		2	0.6227	0.480	40	0.8885	0.521	64
		3	0.6229	0.480	40	0.8899	0.521	65
		Average	0.6229	0.480	40	0.8899	0.520	65
Split 5								
0	135	1	0.9229	0.221	287	0.8262	0.659	38
		2	0.9216	0.222	282	0.8274	0.658	38
		3	0.9255	0.217	298	0.8248	0.665	38
		Average	0.9233	0.220	289	0.8261	0.661	38
1	129	1	0.9223	0.221	285	0.9076	0.664	79
		2	0.9223	0.221	285	0.9033	0.661	75
		3	0.9222	0.222	285	0.9037	0.667	75
		Average	0.9223	0.222	285	0.9049	0.664	76
2	97	1	0.8754	0.281	169	0.9330	0.605	111
		2	0.8758	0.280	169	0.9316	0.604	109
		3	0.8752	0.281	168	0.9328	0.609	111
		Average	0.8755	0.280	169	0.9325	0.606	110
3	85	1	0.8488	0.309	135	0.9405	0.557	126

Table 6: (Continued)

Threshold	N_{act}	Probe	Training set, n = 26			Validation set, n = 10		
			r^2	s	F	r^2	s	F
2	0.8499	0.308	136	0.9417	0.555	129		
3	0.8519	0.306	138	0.9407	0.557	127		
Average	0.8502	0.308	136	0.9409	0.556	127		
4	75	1	0.7992	0.356	95	0.9620	0.582	202
		2	0.7993	0.356	96	0.9619	0.580	202
		3	0.7993	0.356	96	0.9596	0.584	190
		Average	0.7993	0.356	96	**0.9612**	**0.582**	**198**
5	69	1	0.7657	0.385	78	0.9501	0.542	152
		2	0.7637	0.386	78	0.9476	0.546	145
		3	0.7651	0.385	78	0.9501	0.542	152
		Average	0.7648	0.385	78	0.9493	0.543	150

Additional operations are then performed to define the list of attributes:

- Bracket ')' is changed into '(', because both brackets indicate the same molecular phenomenon (branching);
- Each system of 'AxxxBxxxCxxx' is represented by only one version (according to ASCII), in other words, only one version of a SMILES attribute is used for the modeling (not 'AxxxBxxxCxxx' together with 'CxxxBxxxAxxx').

The CW(Sk) is the correlation weight of Sk. There are numerical data for the correlation weights calculated by the Monte Carlo optimization method that indicate the maximum of correlation coefficient between DCW(Threshold) (defined in eqn 3) and the pEC_{50} for the training set. Using the numerical data on the correlation weights, one can calculate DCW(Threshold) for compounds of the training set, and then by the least squares method, one calculate the model

$$pEC_{50} = C_0 + C_1 \times \mathrm{DCW(Threshold)} \quad (2)$$

The predictability of eqn 2 must be checked with compounds of the external validation set.

Threshold is a parameter of the model intended to define rare attributes. For example, if threshold = 4, then all attributes that take place less than in four SMILES of the training set should be classified as rare, and their correlation weight should be defined as zero. Table 3 contains SMILES attributes and their correlation weights used for the QSAR analysis (the split 1).

The optimal InChI-based descriptors are calculated as follows:

$$\mathrm{DCW(Threshold)} = \sum \mathrm{CW(Ik)} \quad (3)$$

where Ik is the InChI attribute and CW(Ik) is the correlation weight of the Ik. The list of InChI attributes was prepared by means of the approach described in Refs. (32,33). Table 4 contains InChI attributes and their correlation weights used for the QSAR analysis (the split 1).

Canonical SMILES and InChI used in this study were generated with ACD/ChemSketch freeware[a]. The optimal SMILES-based descriptors were built by CORAL[d].

Results and Discussion

Table 5 shows the statistical characteristics of the models for the pEC_{50}, which have been calculated with the optimal SMILES-based descriptors. The best model (the case of the split 1) for the external validation set is obtained when the threshold is equal to 4. Table 6 shows the statistical characteristics of the models for the pEC_{50}, which have been calculated with the optimal InChI-based descriptors. The best model (the case of the split 1) for the external validation set is obtained when the threshold is equal to 2. Figure 1 shows the influence of the threshold on the correlation coefficient between DCW and pEC_{50} of the SMILES-based and of InChI-based descriptors. Table 7 gives an example of the DCW(4) calculation for the SMILES-based model. Table 8 shows an example of the DCW(2) calculation for the InChI-based model.

The SMILES-based model for the pEC_{50} with threshold equal to 4 (first probe of the Monte Carlo optimization, split 1) is as follows:

$$pEC_{50} = 2.4028(\pm 0.0682) + 0.0857(\pm 0.00225) \times \mathrm{DCW(4)} \quad (4)$$

n = 26, r^2 = 0.6330, q^2 = 0.5812, s = 0.502, F = 41 (training set); n = 10, r^2 = 0.7493, r^2_{pred} = 0.6235, R^2_m = 0.537, s = 0.541, F = 24 (validation set)

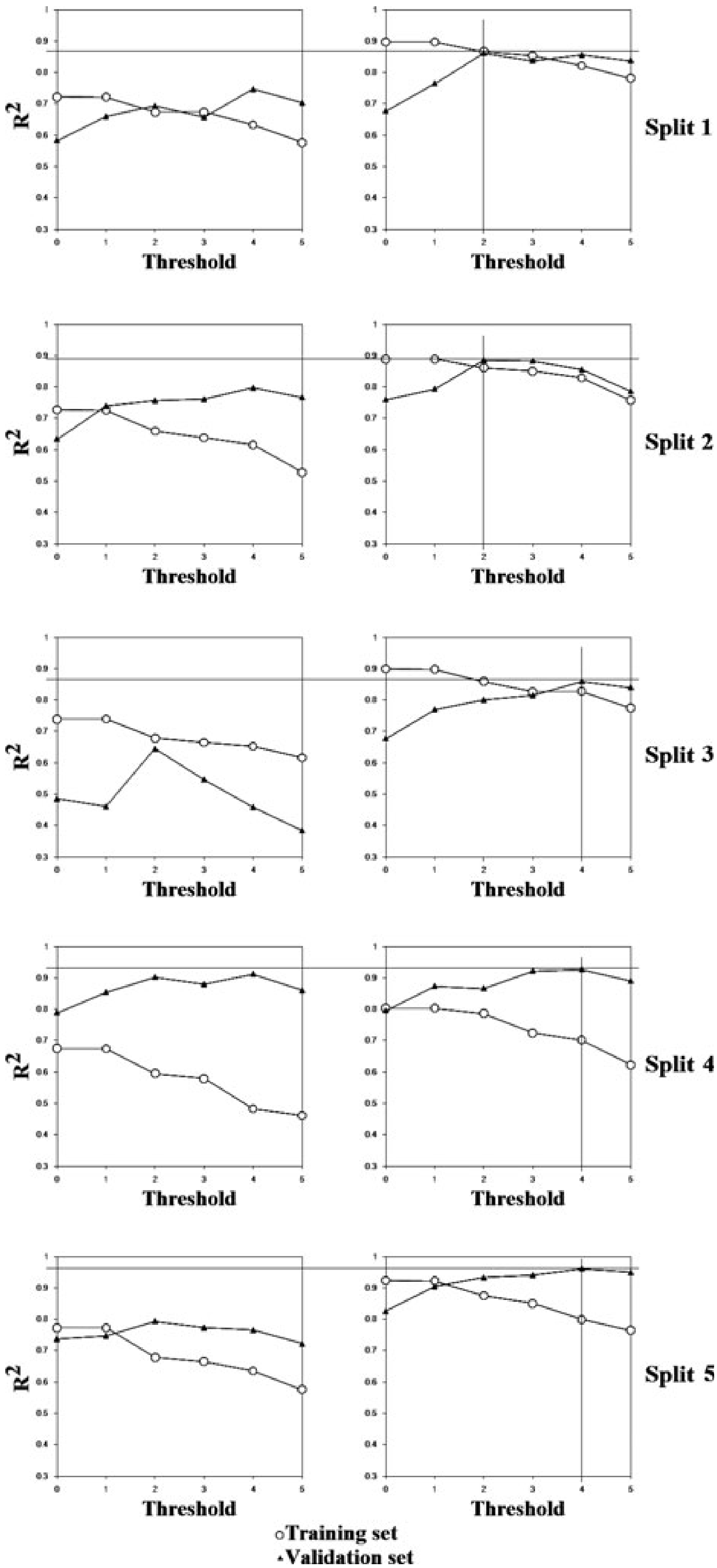


Figure 1: The statistical quality of the simplified molecular input-line entry system-based and InChI-based models, which are calculated with different thresholds.

Table 7: Example of a calculation with the correlation weights listed in Table 2: Compound **2**. Simplified molecular input-line entry system: O=C(O)c1ccc2ccc(nc2c1O)C(=C)c3ccccc3; Threshold = 4; DCW(4) = 32.3107249

Sk	CW(Sk)
Oxxx=xxxCxxx	2.4028279
=xxxCxxx(xxx	1.0209640
Oxxx(xxxCxxx	1.6497003
(xxxOxxx(xxx	1.7962853
cxxx(xxxOxxx	1.4034395
1xxxcxxx(xxx	1.2009262
cxxx1xxxcxxx	2.3969701
cxxxcxxx1xxx	0.3028519
cxxxcxxxcxxx	0.2951162
cxxxcxxx2xxx	2.1503686
cxxx2xxxcxxx	2.1546596
cxxxcxxx2xxx	2.1503686
cxxxcxxxcxxx	0.2951162
cxxxcxxx(xxx	0.2981324
nxxx(xxxcxxx	1.4971106
cxxxnxxx(xxx	1.0499671
nxxxcxxx2xxx	0.3019832
cxxx2xxxcxxx	2.1546596
2xxxcxxx1xxx	0.0
cxxx1xxxOxxx	0.2998894
1xxxOxxx(xxx	0.2969450
Oxxx(xxxCxxx	1.6497003
(xxxCxxx(xxx	0.3498468
Cxxx(xxx=xxx	0.4272772
Cxxx=xxx(xxx	0.2993007
=xxxCxxx(xxx	1.0209640
cxxx(xxxCxxx	0.9463535
3xxxcxxx(xxx	0.7204912
cxxx3xxxcxxx	0.3019973
cxxxcxxx3xxx	0.2955818
cxxxcxxxcxxx	0.2951162
cxxxcxxxcxxx	0.2951162
cxxxcxxxcxxx	0.2951162
cxxxcxxx3xxx	0.2955818

The InChI-based model for the pEC_{50} with threshold equal to 2 (first probe of the Monte Carlo optimization, split 1) is as follows:

$$pEC_{50} = -0.2515(\pm 0.0851) + 0.1029(\pm 0.00162) \times DCW(2) \quad (5)$$

$n = 26$, $r^2 = 0.8673$, $q^2 = 0.8456$, $s = 0.302$, $F = 157$ (training set); $n = 10$, $r^2 = 0.8562$, $r^2_{pred} = 0.7715$, $R^2_m = 0.819$, $s = 0.329$, $F = 48$ (validation set).

The R^2_m is the measure of predictability of a model (34). According to the report (34), model is predictable if the $R^2_m > 0.5$. Thus, the models that are calculated with eqns 4 and 5 are satisfactory according to the R^2_m.

Figure 2 shows the pEC_{50} experimental value and the pEC_{50} calculated for splits 1–5 with the optimal SMILES-based and the InChI-based descriptors. The InChI model is preferable and separates inactive compounds, with the threshold equal to two. The ratio of the number of blocked attributes (Blk) to the total number of attributes (All) is an apparent measure of uncertainty for

Table 8: Example of a calculation with the correlation weights listed in Table 3: Compound **2** ``InChI=1/C18H13NO3/c1-11(12-5-3-2-4-6-12)15-10-8-13-7-9-14(18(21)22)17(20)16(13)19-15/h2-10,20H,1H2,(H,21,22)'' Threshold = 2; DCW(2) = 55.2464323

Ik	CW(Ik)
C18	2.3774174
H13	2.3846044
N	1.4235922
O3	2.3847641
/	1.2240344
c1	1.4752837
–11	0.3149572
(12	1.9774505
–5	1.4068876
–3	0.6135869
–2	1.8402775
–4	0.3147165
–6	2.2009127
–12	0.7299485
(	0.7559318
1	0.7651237
5	1.7820982
–10	0.4561746
–8	0.3078477
–13	0.3128622
–7	2.3790340
–9	0.8453533
–14	0.3075480
(18	0.5823361
(21	1.3010836
(	0.7559318
2	0.3056813
2	0.3056813
(	0.7559318
1	0.7651237
7	0.3084866
(20	0.3235466
(	0.7559318
1	0.7651237
6	0.3075810
(13	0.3129114
(	0.7559318
1	0.7651237
9	0.3112684
–15	0.3107120
/	1.2240344
h2	2.2280274
–10	0.4561746
,20	1.5712777
H	1.6007745
,1	0.5246861
H2	0.3147637
,	2.2841207
(	0.7559318
H	1.6007745
,21	1.2849163
,22	2.3762251
(	0.7559318

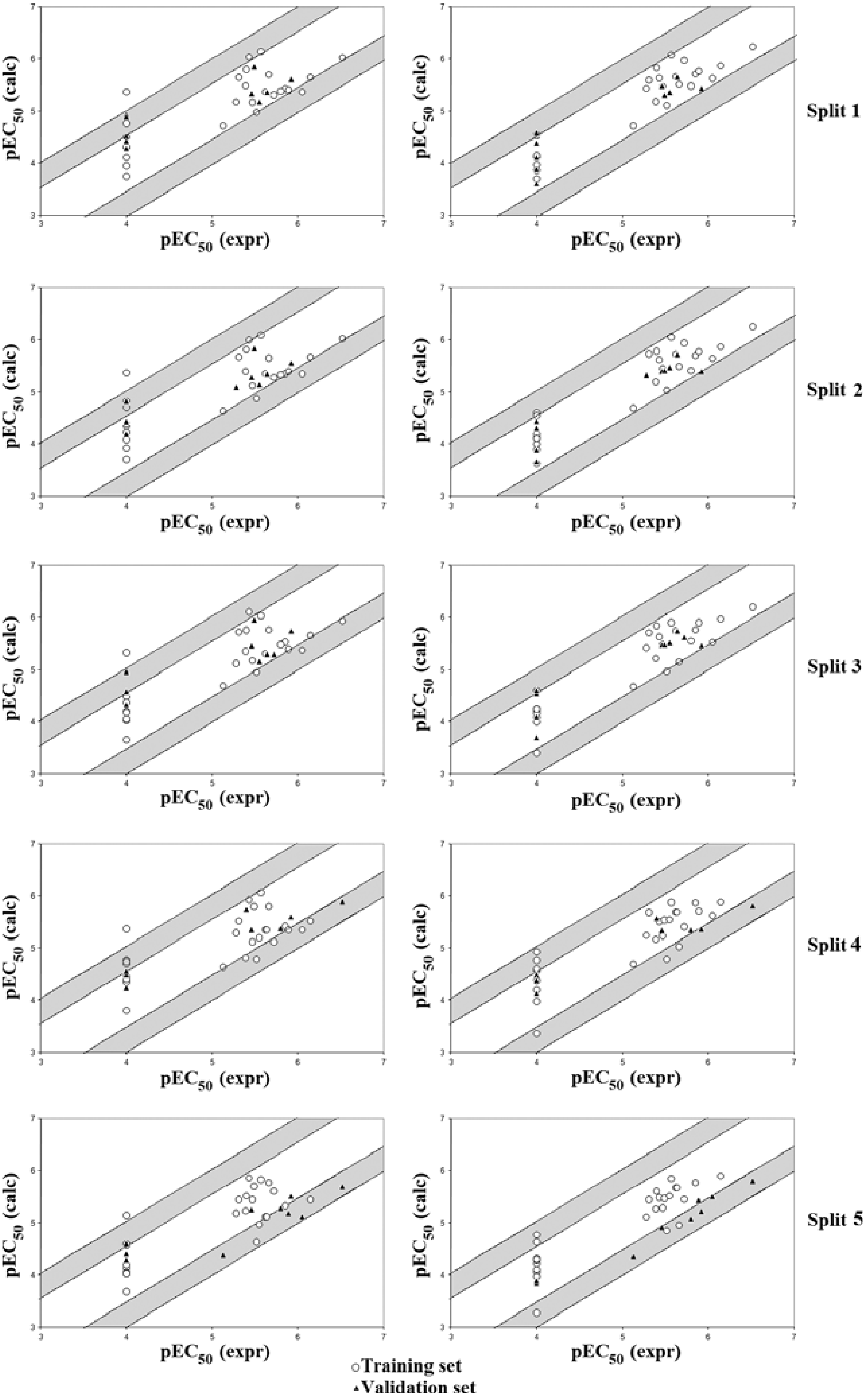


Figure 2: The pEC_{50} experimental value versus pEC_{50} calculated for splits 1–5.

Table 9: Experimental values and pEC_{50} calculated with eqn 4. Blk is the number of simplified molecular input-line entry system (SMILES) attributes that are blocked (Threshold equal to 4), and All is the total number of SMILES attributes for a given compound

ID	SMILES	DCW(4)	Exp	Calc	Exp-Calc	Blk/All
	Training set					
2	O=C(O)c1ccc2ccc(nc2c1O)C(=C)c3ccccc3	32.3107249	5.280	5.172	0.108	1/34
3	O=C(O)c2ccc1ccc(nc1c2O)/C=C/C3CCCO3	33.8957488	5.720	5.308	0.412	11/33
4	O=C(O)c2ccc1ccc(nc1c2O)/C=C\C3CCSC3	32.1996211	5.470	5.162	0.308	12/33
5	O=C(O)c3ccc2ccc(/C=C/c1cccnc1)nc2c3O	35.9228517	5.390	5.481	−0.091	8/34
8	CC(=O)Nc1ccc(cc1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	35.2408532	5.850	5.423	0.427	5/45
9	Oc1ccc(cc1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.7063371	5.800	5.377	0.423	1/39
11	Oc1ccc(c(O)c1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	42.3120428	5.430	6.029	−0.599	1/42
12	Oc1ccc(cc1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.5528364	5.620	5.364	0.256	0/40
14	COc1ccc(cc1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.5528364	6.050	5.364	0.686	1/41
15	Oc1ccc(c(O)c1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	42.1585421	6.520	6.016	0.504	0/43
16	Oc1cc(cc(OC)c1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	37.8037199	6.150	5.643	0.507	2/44
17	COc1cc(cc(OC)c1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	37.8037199	5.310	5.643	−0.333	3/45
18	Brc1cc(cc(Br)c1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.9577170	5.890	5.399	0.491	4/43
19	Ic1cc(cc(O)c1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	39.5608813	5.400	5.793	−0.393	1/43
20	Oc1ccc(cc1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)OC	34.5528364	4.000	5.364	−1.364	1/41
23	Oc1cccc2ccc(C)nc12	15.5817819	4.000	3.738	0.262	2/16
24	COc1ccc(cc1OC)C(=C)C(=O)Oc2cccc3ccc(C)nc23	18.0035331	4.000	3.946	0.054	9/40
25	Oc1ccc(cc1O)C(=C)C(=O)Oc2cccc3ccc(C)nc23	19.9501784	4.000	4.113	−0.113	5/38
27	Oc1cccc2ccc(nc12)C(=C)c3ccccc3	22.2266759	4.000	4.308	−0.308	2/28
28	Oc2cccc1ccc(nc12)/C=C/c3ccc4cccc(O)c4n3	27.5087045	4.000	4.760	−0.760	12/37
30	Oc1ccc(cc1O)C(=C)c2ccc3cccc([N+]([O−])=O)c3n2	24.3664474	4.000	4.491	−0.491	11/43
31	Oc1ccc(cc1O)C(=C)c2ccc3cccc(N)c3n2	22.3595591	4.000	4.319	−0.319	3/32
33	Oc1ccc(cc1O)C(=C)c2ccc3cccc(O)c3n2	26.9627234	5.130	4.714	0.416	0/32
34	Oc1ccc(cc1O)C(=C)c2ccc3ccc(nc3c2O)C(=C)c4ccc(O)c(O)c4	38.4091952	5.660	5.694	−0.034	8/51
35	Oc1ccc(cc1O)C(=C)c2ccc3ccc(C#N)c(O)c3n2	30.0165602	5.520	4.975	0.545	4/37
36	O=C(O)c1cc(ccc1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	43.5307697	5.570	6.133	−0.563	0/45
	Validation set					
1	O=C(O)c1ccc2ccc(C)nc2c1O	24.6204264	4.000	4.513	−0.513	2/22
6	[O−][N+](=O)c1ccc(cc1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	37.4179417	5.920	5.610	0.310	9/50
7	Nc1ccc(cc1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.2101272	5.460	5.335	0.125	2/39
10	Oc1cc(cc(O)c1)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	40.2105919	5.490	5.849	−0.359	1/42
13	Oc1ccc(cc1C)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	32.3063017	5.550	5.171	0.379	3/40
21	Oc1ccc(cc1O)C(=C)c2ccc3c(Cl)cc(Cl)c(O)c3n2	29.4981323	4.000	4.931	−0.931	6/38
22	Oc1ccc(cc1O)C(=C)c2ccc3ccc(c(O)c3n2)C(=O)O	34.5528364	5.640	5.364	0.276	0/40
26	CC(=O)Oc1cccc2ccc(nc12)C(=C)c3ccccc3	23.5617910	4.000	4.422	−0.422	4/34
29	Oc1ccc(cc1O)C(=C)c2ccc3ccccc3n2	21.9316335	4.000	4.282	−0.282	0/29
32	CC(=O)Oc1ccc(cc1OC(C)=O)C(=C)c2ccc3cccc(OC(C)=O)c3n2	28.9657538	4.000	4.885	−0.885	8/50

a given structure. For instance, in the case of SMILES, this ratio for compound **2** is 11/33, whereas in the case of InChI, it is 2/55. This situation holds for the majority of compounds (Tables 9 and 10). Thus, SMILES-based models have larger uncertainty.

The statistical characteristics of the best model for the pEC_{50} described in Ref. (3) are the following: $n = 26$, $r^2 = 0.607$, $s = 0.542$ for the training set and $n = 10$, $r^2 = 0.611$, $s = 0.550$ for the validation set. Thus, the model calculated using eqn 5 is better.

In order to use these models for the prediction of pEC_{50} value for an external substance (a styrylquinoline derivative), one should prepare SMILES or InChI for the above-mentioned substance and calculate SMILES-based DCW(4) descriptor for calculation with eqn 4 (Table 3) or InChI-based DCW(2) descriptor for calculation with eqn 5 (Table 4).

Conclusions

The optimal descriptors calculated with eqn 4 (representation of the molecular structure by SMILES) and those calculated with eqn 5 (representation of the molecular structure by InChI) give models for the anti-HIV-1 integrase inhibitory activity of styrylquinoline derivatives offering better predictability than the best model described in Ref. (3). The optimal InChI-based descriptors predict for the anti-HIV-1 integrase inhibitory activity of styrylquinoline derivatives against HIV-1 better than the optimal SMILES-based descriptors. These results are reproduced for five examined splits into the training and test sets.

Table 10: Experimental values and pEC_{50} calculated with eqn 5. Blk is the number of InChI attributes that are blocked (Threshold equal to 2), and All is the total number of InChI attributes for a given compound

ID	SMILES	DCW(2)	Expr	Calc	Expr-Calc	Blk/All
	Training set					
2	InChI=1/C18H13NO3/c1-11(12-5-3-2-4-6-12)15-10-8-13-7-9-14(18(21)22)17(20)16(13)19-15/h2-10,20H,1H2,(H,21,22)	55.2464323	5.280	5.433	−0.153	0/53
3	InChI=1/C16H15NO4/c18-15-13(16(19)20)8-4-10-3-5-11(17-14(10)15)6-7-12-2-1-9-21-12/h3-8,12,18H,1-2,9H2,(H,19,20)/b7-6+	60.4720503	5.720	5.971	−0.251	2/55
4	InChI=1/C16H15NO3S/c18-15-13(16(19)20)6-3-11-2-5-12(17-14(11)15)4-1-10-7-8-21-9-10/h1-6,10,18H,7-9H2,(H,19,20)/b4-1-	55.6352893	5.470	5.473	−0.003	5/55
5	InChI=1/C17H12N2O3/c20-16-14(17(21)22)8-5-12-4-7-13(19-15(12)16)6-3-11-2-1-9-18-10-11/h1-10,20H,(H,21,22)/b6-3+	52.7690591	5.390	5.178	0.212	2/51
8	InChI=1/C20H16N2O4/c1-11(13-3-7-15(8-4-13)21-12(2)23)17-10-6-14-5-9-16(20(25)26)19(24)18(14)22-17/h3-10,24H,1H2,2H3,(H,21,23)(H,25,26)	57.9209635	5.850	5.709	0.141	1/68
9	InChI=1/C18H13NO4/c1-10(11-2-6-13(20)7-3-11)15-9-5-12-4-8-14(18(22)23)17(21)16(12)19-15/h2-9,20-21H,1H2,(H,22,23)	55.6710957	5.800	5.477	0.323	0/56
11	InChI=1/C18H13NO5/c1-9(12-6-4-11(20)8-15(12)21)14-7-3-10-2-5-13(18(23)24)17(22)16(10)19-14/h2-8,20-22H,1H2,(H,23,24)	57.1187516	5.430	5.626	−0.196	0/59
12	InChI=1/C18H13NO5/c1-9(11-4-7-14(20)15(21)8-11)13-6-3-10-2-5-12(18(23)24)17(22)16(10)19-13/h2-8,20-22H,1H2,(H,23,24)	57.3767062	5.620	5.653	−0.033	0/59
14	InChI=1/C19H15NO5/c1-10(12-5-8-16(25-2)15(21)9-12)14-7-4-11-3-6-13(19(23)24)18(22)17(11)20-14/h3-9,21-22H,1H2,2H3,(H,23,24)	57.0625394	6.050	5.620	0.430	0/62
15	InChI=1/C18H13NO6/c1-8(10-5-7-13(20)17(23)16(10)22)12-6-3-9-2-4-11(18(24)25)15(21)14(9)19-12/h2-7,20-23H,1H2,(H,24,25)	62.9432930	6.520	6.225	0.295	0/63
16	InChI=1/C19H15NO6/c1-9(11-7-14(21)18(23)15(8-11)26-2)13-6-4-10-3-5-12(19(24)25)17(22)16(10)20-13/h3-8,21-23H,1H2,2H3,(H,24,25)	59.4590457	6.150	5.867	0.283	0/66
17	InChI=1/C20H17NO6/c1-10(12-8-15(26-2)19(23)16(9-12)27-3)14-7-5-11-4-6-13(20(24)25)18(22)17(11)21-14/h4-9,22-23H,1H2,2-3H3,(H,24,25)	56.7813359	5.310	5.591	−0.281	2/68
18	InChI=1/C18H11Br2NO4/c1-8(10-6-12(19)17(23)13(20)7-10)14-5-3-9-2-4-11(18(24)25)16(22)15(9)21-14/h2-7,22-23H,1H2,(H,24,25)	58.4291109	5.890	5.761	0.129	2/64
19	InChI=1/C18H12INO5/c1-8(10-6-12(19)17(23)14(21)7-10)13-5-3-9-2-4-11(18(24)25)16(22)15(9)20-13/h2-7,21-23H,1H2,(H,24,25)	59.0441369	5.400	5.824	−0.424	1/63
20	InChI=1/C19H15NO5/c1-10(12-5-8-15(21)16(22)9-12)14-7-4-11-3-6-13(19(24)25-2)18(23)17(11)20-14/h3-9,21-23H,1H2,2H3	46.5466162	4.000	4.538	−0.538	0/56
23	InChI=1/C10H9NO/c1-7-5-6-8-3-2-4-9(12)10(8)11-7/h2-6,12H,1H3	38.3670825	4.000	3.696	0.304	2/30
24	InChI=1/C21H19NO4/c1-13-8-9-15-6-5-7-18(20(15)22-13)26-21(23)14(2)16-10-11-17(24-3)19(12-16)25-4/h5-12H,2H2,1,3-4H3	42.5630368	4.000	4.128	−0.128	3/56
25	InChI=1/C19H15NO4/c1-11-6-7-13-4-3-5-17(18(13)20-11)24-19(23)12(2)14-8-9-15(21)16(22)10-14/h3-10,21-22H,2H2,1H3	39.9298774	4.000	3.857	0.143	0/54
27	InChI=1/C17H13NO/c1-12(13-6-3-2-4-7-13)15-11-10-14-8-5-9-16(19)17(14)18-15/h2-11,19H,1H2	42.4653651	4.000	4.118	−0.118	0/40
28	InChI=1/C20H14N2O2/c23-17-5-1-3-13-7-9-15(21-19(13)17)11-12-16-10-8-14-4-2-6-18(24)20(14)22-16/h1-12,23-24H/b12-11+	42.7119710	4.000	4.144	−0.144	2/51
30	InChI=1/C17H12N2O4/c1-10(12-6-8-15(20)16(21)9-12)13-7-5-11-3-2-4-14(19(22)23)17(11)18-13/h2-9,20-21H,1H2	46.4713181	4.000	4.530	−0.530	0/50
31	InChI=1/C17H14N2O2/c1-10(12-6-8-15(20)16(21)9-12)14-7-5-11-3-2-4-13(18)17(11)19-14/h2-9,20-21H,1,18H2	40.9936062	4.000	3.967	0.033	0/47
33	InChI=1/C17H13NO3/c1-10(12-6-8-14(19)16(21)9-12)13-7-5-11-3-2-4-15(20)17(11)18-13/h2-9,19-21H,1H2	48.2577301	5.130	4.714	0.416	0/46
34	InChI=1/C25H19NO5/c1-13(16-5-9-20(27)22(29)11-16)18-7-3-15-4-8-19(26-24(15)25(18)31)14(2)17-6-10-21(28)23(30)12-17/h3-12,27-31H,1-2H2	56.0243483	5.660	5.513	0.147	8/67
35	InChI=1/C18H12N2O3/c1-10(12-5-7-15(21)16(22)8-12)14-6-4-11-2-3-13(9-19)18(23)17(11)20-14/h2-8,21-23H,1H2	52.0398555	5.520	5.103	0.417	0/50
36	InChI=1/C19H13NO6/c1-9(11-4-7-15(21)13(8-11)19(25)26)14-6-3-10-2-5-12(18(23)24)17(22)16(10)20-14/h2-8,21-22H,1H2,(H,23,24)(H,25,26)	61.4842765	5.570	6.075	−0.505	0/69
	Validation set					
1	InChI=1/C11H9NO3/c1-6-2-3-7-4-5-8(11(14)15)10(13)9(7)12-6/h2-5,13H,1H3,(H,14,15)	40.2422546	4.000	3.889	0.111	6/42
6	InChI=1/C18H12N2O5/c1-10(11-2-6-13(7-3-11)20(24)25)15-9-5-12-4-8-14(18(22)23)17(21)16(12)19-15/h2-9,21H,1H2,(H,22,23)	55.2667781	5.920	5.435	0.485	1/60
7	InChI=1/C18H14N2O3/c1-10(11-2-6-13(19)7-3-11)15-9-5-12-4-8-14(18(22)23)17(21)16(12)20-15/h2-9,21H,1,19H2,(H,22,23)	55.7422221	5.460	5.484	−0.024	0/56
10	InChI=1/C18H13NO5/c1-9(11-6-12(20)8-13(21)7-11)15-5-3-10-2-4-14(18(23)24)17(22)16(10)19-15/h2-8,20-22H,1H2,(H,23,24)	54.0208340	5.490	5.307	0.183	0/58
13	InChI=1/C19H15NO4/c1-10-9-13(5-8-16(10)21)11(2)15-7-4-12-3-6-14(19(23)24)18(22)17(12)20-15/h3-9,21-22H,2H2,1H3,(H,23,24)	54.5613778	5.550	5.363	0.187	1/62
21	InChI=1/C17H11Cl2NO3/c1-8(9-2-5-14(21)15(22)6-9)13-4-3-10-11(18)7-12(19)17(23)16(10)20-13/h2-7,21-23H,1H2	45.1037243	4.000	4.390	−0.390	2/53
22	InChI=1/C18H13NO5/c1-9(11-4-7-14(20)15(21)8-11)13-6-3-10-2-5-12(18(23)24)17(22)16(10)19-13/h2-8,20-22H,1H2,(H,23,24)	57.3767062	5.640	5.653	−0.013	0/59
26	InChI=1/C19H15NO2/c1-13(15-7-4-3-5-8-15)17-12-11-16-9-6-10-18(19(16)20-17)22-14(2)21/h3-12H,1H2,2H3	37.4937913	4.000	3.607	0.393	0/46
29	InChI=1/C17H13NO2/c1-11(13-7-9-16(19)17(20)10-13)14-8-6-12-4-2-3-5-15(12)18-14/h2-10,19-20H,1H2	42.4362954	4.000	4.115	−0.115	1/44
32	InChI=1/C23H19NO6/c1-13(18-9-11-20(28-14(2)25)22(12-18)30-16(4)27)19-10-8-17-6-5-7-21(23(17)24-19)29-15(3)26/h5-12H,1H2,2-4H3	46.9730495	4.000	4.582	−0.582	6/63

SMILES, simplified molecular input-line entry system.

Acknowledgments

The authors thank the Marie Curie Fellowship for financial support (the contract ID 39036, CHEMPREDICT). The authors also express their gratitude to Dr J. Baggot for the English revision.

References

1. Vidal D., Thormann M., Pons M. (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model;45:386–393.
2. Toropov A.A., Toropova A.P., Raska I. Jr (2008) QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES. Eur J Med Chem;43:714–740.
3. Leonard J.T., Roy K. (2008) Exploring molecular shape analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors. Eur J Med Chem;43:81–92.
4. Liu B., Gutman I. (2007) On general Randic indices. MATCH Commun Math Comput Chem; 58: 147–154.
5. Gutman I., Durdevic J. (2008) Fluoranthene and its congeners – a graph theoretical study. MATCH Commun Math Comput Chem;60:659–670.
6. Kuz'min V.E., Muratov E.N., Artemenko A.G., Gorb L., Qasim M., Leszczynski J. (2008) The effect of nitroaromatics' composition on their toxicity *in vivo*: novel, efficient non-additive 1D QSAR analysis. Chemosphere;72:1373–1380.
7. Marrero-Ponce Y., Castillo-Garit J.A., Castro E.A., Torrens F., Rotondo R. (2008) 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: theory and QSAR applications to central chirality codification. J Math Chem;44:755–786.
8. Duchowicz P.R., Talevi A., Bruno-Blanch L.E., Castro E.A. (2008) New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg Med Chem;16:7944–7955.
9. Duchowicz P.R., Vitale M.G., Castro E.A. (2008) Partial Order Ranking for the aqueous toxicity of aromatic mixtures. J Math Chem;44:541–549.
10. Afantitis A., Melagraki G., Sarimveis H., Koutentis P.A., Markopoulos J., Igglessi-Markopoulou O. (2006) A novel QSAR model for evaluating and predicting the inhibition activity of dipeptidyl aspartyl fluoromethylketones. QSAR Comb Sci;25:928–935.
11. Afantitis A., Melagraki G., Sarimveis H., Koutentis P.A., Markopoulos J., Igglessi-Markopoulou O. (2006) Prediction of intrinsic viscosity in polymer-solvent combinations using a QSPR model. Polymer;47:3240–3248.
12. Puzyn T., Mostrag A., Suzuki N., Falandysz J. (2008) QSPR-based estimation of the atmospheric persistence for chloronaphthalene congeners. Atmos Environ;42:6627–6636.
13. Puzyn T., Suzuki N., Haranczyk M. (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine?. Environ Sci Technol;42:5189–5195.
14. Puzyn T., Suzuki N., Haranczyk M., Rak J. (2008) Calculation of quantum-mechanical descriptors for QSPR at the DFT level: is it necessary?. J Chem Inf Model;48:1174–1180.
15. Kusic H., Rasulev B., Leszczynska D., Leszczynski J., Koprivanac N. (2009) Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study. Chemosphere;75:1128–1134.
16. Gini G., Benfenati E. (2007) E-modelling: foundations and cases for applying AI to life sciences. Int J Artif Intell Tools;16: 243–268.
17. Gini G., Garg T., Stefanelli M. (2009) Ensembling regression models to improve their predictivity: a case study in qsar (quantitative structure activity relationships) with computational chemometrics. Appl Artif Intell;23:261–281.
18. Gutman I., Rucker C., Rucker G. (2001) On walks in molecular graphs. J Chem Inf Comput Sci;41:739–745.
19. Weininger D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci;28:31–36.
20. Weininger D., Weininger A., Weininger J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci;29:97–101.
21. Weininger D. (1990) Smiles. 3. Depict. Graphical depiction of chemical structures. J Chem Inf Comput Sci;30:237–243.
22. Coles S.J., Day N.E., Murray-Rust P., Rzepa H.S., Zhang Y. (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. Org Biomol Chem;3:1832–1834.
23. Murray-Rust P., Rzepa H.S., Stewart J.J.P., Zhang Y. (2005) A global resource for computational chemistry. J Mol Model;11:532–541.
24. Randic M., Basak S.C. (1999) Optimal molecular descriptors based on weighted path numbers. J Chem Inf Comput Sci;39:261–266.
25. Randic M., Pompe M. (2001) The variable connectivity index 1Xf versus the traditional molecular descriptors: a comparative study of $^{1}X^{f}$ against descriptors of CODESSA. J Chem Inf Comput Sci;41:631–638.
26. Randic M., Basak S.C. (2001) New descriptor for structure-property and structure-activity correlations. J Chem Inf Comput Sci;41:650–656.
27. Da Silva Junkes B., Arruda A.C.S., Yunes R.A., Porto L.C., Heinzen V.E.F. (2005) Semi-empirical topological index: a tool for QSPR/QSAR studies. J Mol Model;11:128–134.
28. Arruda A.C.S., Da Silva Junkes B., Souza E.S., Yunes R.A., Heinzen V.E.F. (2008) Semi-empirical topological index to predict properties of halogenated aliphatic compounds. J Chemometr;22:186–194.
29. Porto L.C., Souza E.S., da Silva Junkes B., Yunes R.A., Heinzen V.E.F. (2008) Semi-empirical topological index: development of QSPR/QSRR and optimization for alkylbenzenes. Talanta;76:407–412.
30. Toropov A.A., Benfenati E. (2007) SMILES in QSPR/QSAR modeling: results and perspectives. Curr Drug Discov Technol;4:77–116.
31. Toropov A.A., Rasulev B.F., Leszczynski J. (2008) QSAR modeling of acute toxicity by balance of correlations. Bioorg Med Chem;16:5999–6008.
32. Toropov A.A., Toropova A.P., Benfenati E. (2009) QSPR modeling of octanol water partition coefficient of platinum complexes by InChI-based optimal descriptors. J Math Chem;46:1060–1073.

33. Toropov A.A., Toropova A.P., Benfenati E., Leszczynska D., Leszczynski J. (2009) Additive InChI-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents. J Math Chem;46:1232–1251.
34. Roy P.P., Roy K. (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci;27:302–313.

Notes

[a]ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, available at: http://www.acdlabs.com, 2007.

[b]U.S. Library of Medicine (2008) available at: http://toxnet.nlm.nih.gov/.

[c]National Institute of Standard and Technology (2008) available at: http://webbook.nist.gov/chemistry/.

[d]CHEMPREDICT, CORAL freeware, available at: http://www.insilico.eu/coral/.