J Math Chem (2010) 47:647–666
DOI 10.1007/s10910-009-9593-6

ORIGINAL PAPER

QSAR analysis of 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines exhibiting anticancer activity by optimal SMILES-based descriptors

A. A. Toropov · A. P. Toropova · E. Benfenati · D. Leszczynska · J. Leszczynski

Received: 6 June 2009 / Accepted: 17 August 2009 / Published online: 26 September 2009

Abstract A predictive model of the anticarcinogenic activity of a 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines series has been built, with statistical quality: $n=75$, $r^2 = 0.7688$, $s=0.48$, $F=243$ (training set); $n=25$, $r^2 = 0.8025$, $s=0{,}49$, $F=93$ (test set). The robustness of this model has been tested in three random splits into training set and test set. Correlation weights (the analogue of the contributions of substituents) of molecular attributes expressed by symbols in the simplified molecular input line entry system (SMILES) notation are able to serve as informative indicators in the search for new anticancer agents.

Keywords SMILES · QSAR · Optimal descriptor · Anticancer activity

1 Introduction

The search for anticancer agents is an important field of medicinal chemistry. Recently, quantitative structure–activity relationships (QSAR) for a series of 7- and 3-substituted 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines, which are novel antitumor quinolone agents, was carried out [1].

Typically, descriptors calculated with molecular graphs are used in QSAR analysis [2–12]. Their checking and validation [13] as well as the reduction of correlations between two-dimensional (2D) descriptors [14] are not possible when applying a

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
e-mail: aatoropov@yahoo.com

D. Leszczynska · J. Leszczynski
Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry,
Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

simple interpretation [15] of the relationship between structure and endpoint in the case of multiple linear regression (MLR).

The contributions of substituents in the free Wilson model [16] and in Fujita-Ban calculations [17] have a transparent interpretation. Semiempirical topological indexes [18] as well as optimal descriptors [12] are developed in an attempt to obtain descriptors that can indicate positive or negative influence of individual molecular fragments on endpoints. On the other hand, quantum-mechanical descriptors can lead to fundamental knowledge about the mechanisms of biochemical phenomena [19].

As an alternative, the simplified molecular input line entry system (SMILES) molecular graph [20–25] can be used for elucidation of molecular structures. Considerable knowledge about applications of SMILES in QSAR analysis has been gained in recent years [26–34].

There are a number of software systems that can generate so-called canonical SMILES; however, the canonical SMILES generated by different software packages are not the same [23,24]. One can expect that a standard for SMILES notation will be formulated in the near future. Since at the present such a standard is not available, only the SMILES which are generated by a selected software package (not a mixture of SMILES generated by several software systems) should be used for QSAR analysis. Accordingly, the SMILES used in the present study were generated with the ACD/ChemSketch program [23].

The aim of this study is to evaluate the ability of SMILES-based optimal descriptors in QSAR modeling of the potential anticancer capability of the title compounds.

2 Method

The endpoint considered herein for the studied compounds is $\log(1/\mathrm{IC}_{50})$, where IC_{50} represents the concentration of the agent necessary to reduce cell viability by 50% against Murine P388 Leukemia (in vitro cytotoxic activity). Numerical data on this endpoint were taken from Ref. [1].

Optimal SMILES-based descriptors are calculated as follows:

$$\mathrm{DCW}\,(\mathrm{limS}) = \mathrm{CW}\,(b) \times \mathrm{CW}\,(\mathrm{db}) \times \mathrm{CW}\,(\mathrm{tb}) \times \prod \mathrm{CW}\,(\mathrm{SS_k})\,, \qquad (1)$$

where b is the number of brackets in the SMILES (i.e., a measure of branching), db represents the number of '=' (the symbol indicating a double bond), tb is the number of '#' (the symbol indicating a triple bond), and SS_k is a composition of two elements of SMILES. There are elements which contain only one symbol ('c', 'C', '=', '#', '1', '2', etc.). Also, elements which contain two symbols ('Cl', 'Br', 'N+', '@@', etc.) are included. The physical meaning of these (and others) elements of SMILES notation is described in the literature [20–22] and is also available on the Internet [23–25]. $\mathrm{CW}(x)$ represents the correlation weight for x, i.e., for one of the mentioned SMILES attributes (SA). Correlation weights are calculated by the Monte Carlo optimization method. The correlation coefficient between the descriptor and anticancer activity (for a training set) is a target function of the optimization. Based on the numerical data for the correlation weights, one can calculate, using Eq. 1, the value of DCW(limS) for

Table 1 Structures and numbering for the anticancer agents

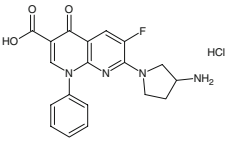
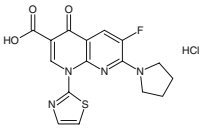
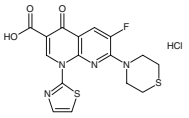
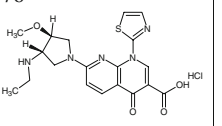
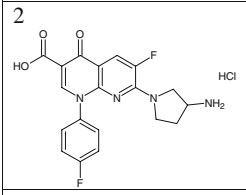
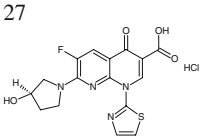
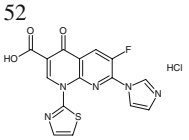
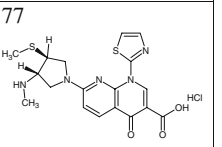
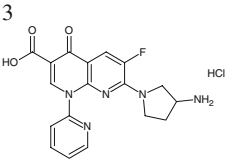
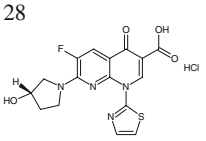
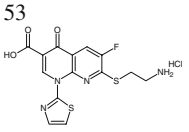
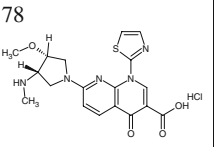
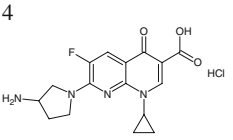
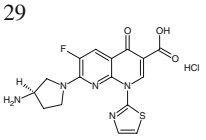
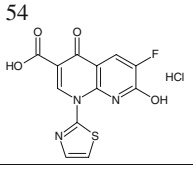
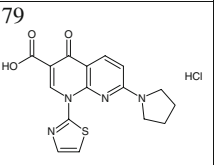
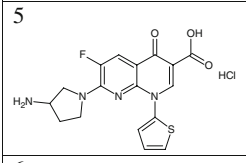
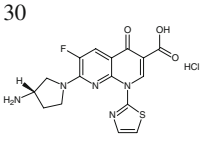
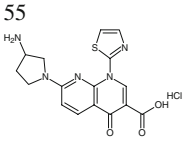
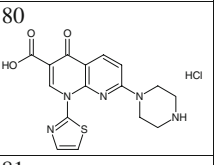
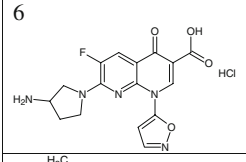
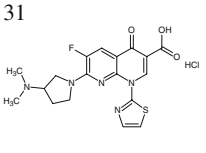
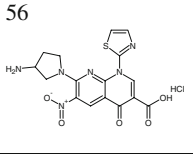
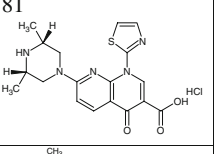
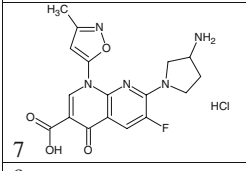
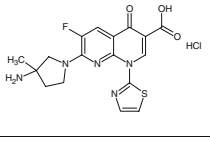
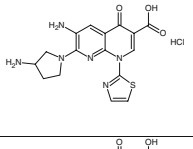
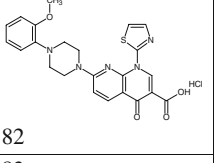
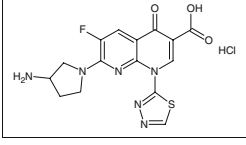
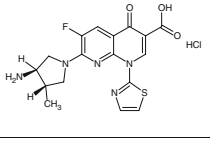
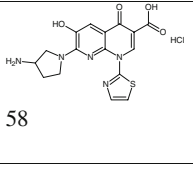
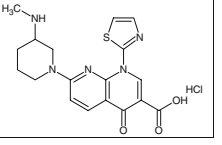
1 	26 	51 	76 
2 	27 	52 	77 
3 	28 	53 	78 
4 	29 	54 	79 
5 	30 	55 	80 
6 	31 	56 	81 
7 	32 	57 	82 
8 	33 	58 	83 

Table 1 Continued

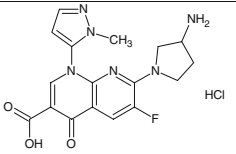
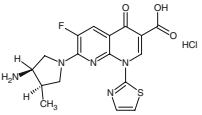
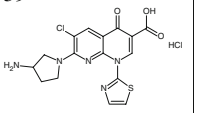
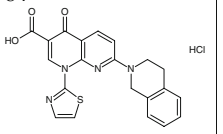
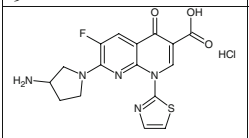
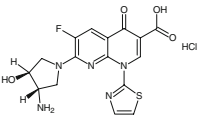
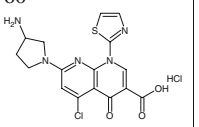
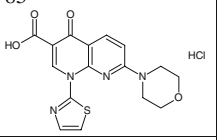
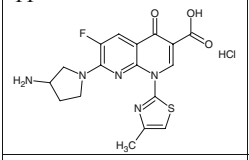
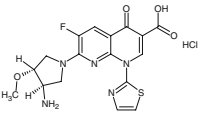
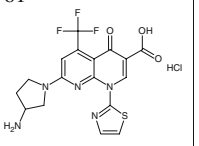
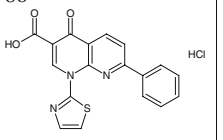
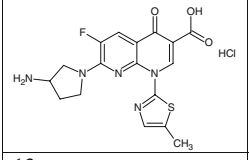
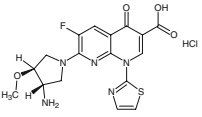
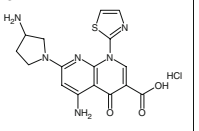
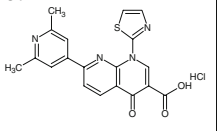
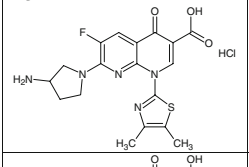
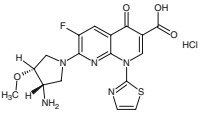
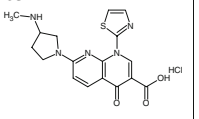
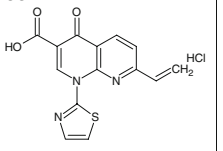
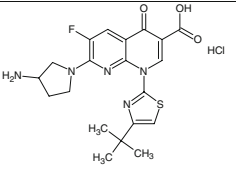
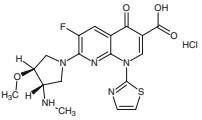
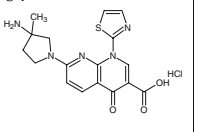
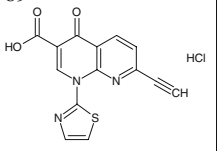
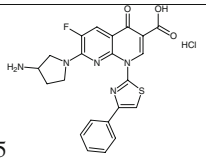
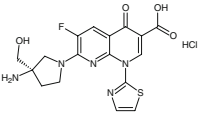
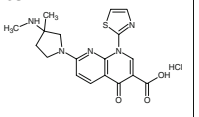
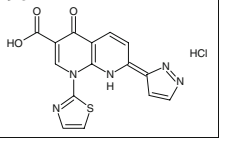
 9	34 	59 	84 
 10	35 	60 	85 
11 	36 	61 	86 
12 	37 	62 	87 
13 	38 	63 	88 
14 	39 	64 	89 
15 	40 	65 	90 

Table 1 Continued

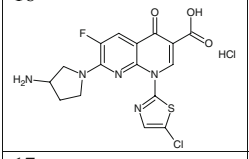
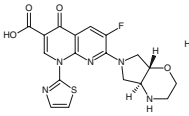
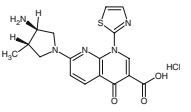
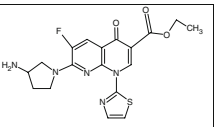
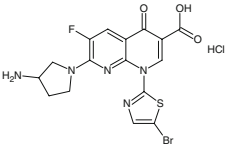
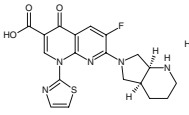
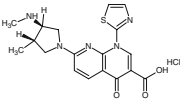
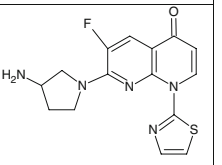
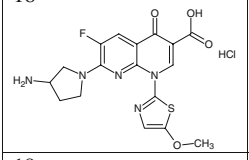
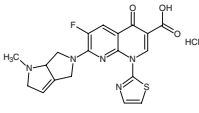
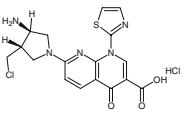
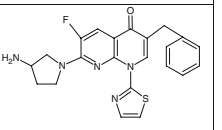
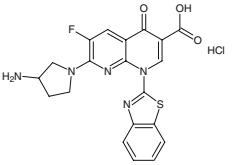
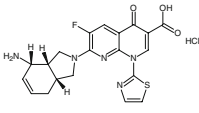
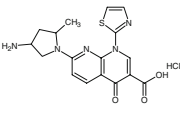
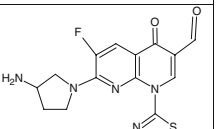
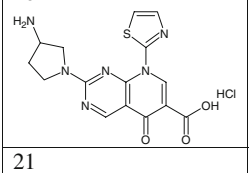
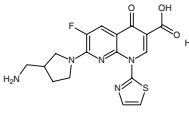
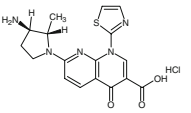
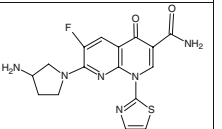
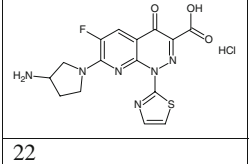
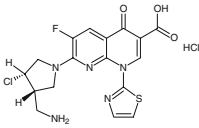
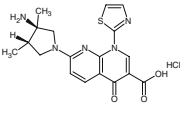
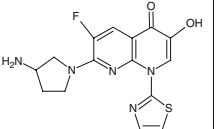
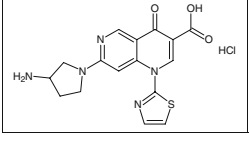
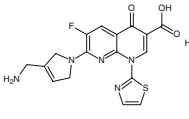
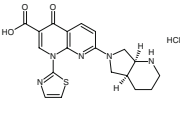
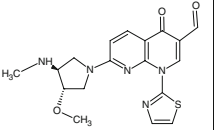
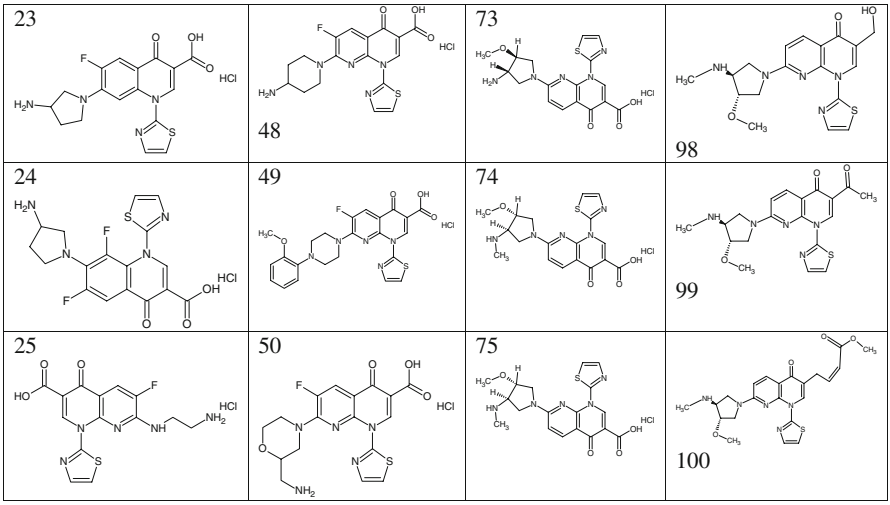
16 	41 	66 	 91
17 	42 	67 	 92
18 	43 	68 	 93
19 	44 	69 	 94
20 	45 	70 	 95
21 	46 	71 	 96
22 	47 	72 	 97

Table 1 Continued



both training and test set. Using the least-squares method, one can calculate for the training-set model

$$\mathrm{Log}\,(1/\mathrm{IC}_{50}) = \mathrm{C}_0 + \mathrm{C}_1 \times \mathrm{DCW}\,(\mathrm{limS})\,. \tag{2}$$

Using an external test set one can then estimate the predictive ability of Eq. 2.

There are SA which appear many times in the training set. However, there are also SA that are rare or absent in the training set. Rare SA can lead to overtraining (the situation in which good statistical quality of the model for the training set is accompanied by poor statistical quality for the test set). If the value of CW(x) is fixed to be 1, the influence of CW(x) is blocked. The number of SMILES in the training set that contain x can be used for classification of the SA into rare or not rare groups. The parameter limS is used for this separation into two classes. In other words, if limS = 5, all SA which appear one, two, three or four times in the SMILES notations of the training set are defined to be rare and their corresponding CW(SA) = 1.

3 Results

Table 1 shows the structures and ID of the considered anticancer agents. Three splits into training set and test set have been carried out. Table 2 shows the ID of the compounds selected for inclusion in the test set for each split.

The values in Table 3 indicate that the best limS (from the point of view of predictive statistical quality) are the following: limS = 13 (for split 1) and limS = 11 (for splits 2 and 3). Graphically these results are plotted in Figs. 1, 2, and 3.

Table 2 List of anticancer agents selected in the test set for three splits into training and test sets

Split 1	Split 2	Split 3
3	2	3
6	5	4
9	8	11
12	12	16
14	16	21
16	20	29
24	27	32
29	32	33
32	34	36
34	36	39
37	39	40
42	45	46
46	46	50
50	49	53
58	53	60
64	62	62
67	64	65
73	65	67
75	66	74
78	70	78
81	74	82
88	79	86
91	85	90
92	90	94
94	96	100

Table 4 shows that the statistical quality of the models is reproduced in three probes of the Monte Carlo optimization. Numerical data on the correlation weights for splits 1, 2, and 3 are shown in Tables 5, 6, and 7, respectively. The model obtained in probe 1 for split 1 (maximal correlation coefficient for the training set) can be described as follows (Table 8):

$$\mathrm{Log}(1/\mathrm{IC}_{50}) = -32.6796(\pm 0.2319) + 28.6643(\pm 0.2016) \times \mathrm{DCW}(9) \qquad (3)$$

$n = 75$, $r^2 = 0.7688$, $s = 0.48$, $F = 243$ (training set)
$n = 25$, $r^2 = 0.8025$, $s = 0{,}49$, $F = 93$ (test set).

4 Discussion

First of all, the data in Tables 5, 6, and 7 indicate that different splits, in fact, represent different distributions of the SA in the training set and the test set. For instance, the SMILES attribute '(004', as a rare descriptor, is blocked in the case of the split 1.

Table 3 Average values of statistical characteristics over three runs of the Monte Carlo optimization for three splits into training and test set

limS	Training set, $n=75$			Test set, $n=25$		
	R^2	s	F	R^2	S	F
Split 1						
1	0.8941	0.322	616	0.3303	1.086	12
2	0.8667	0.361	475	0.4587	0.905	20
3	0.8517	0.381	419	0.5944	0.814	34
4	0.8308	0.407	359	0.7278	0.598	62
5	0.7839	0.459	266	0.7179	0.594	60
6	0.7777	0.466	256	0.7843	0.547	85
7	0.7753	0.469	252	0.7623	0.537	75
8	0.7747	0.469	251	0.7731	0.530	79
9	**0.7658**	**0.479**	**239**	**0.7966**	**0.496**	**90**
10	0.7540	0.491	224	0.7432	0.563	67
11	0.7585	0.486	229	0.7266	0.583	61
12	0.7541	0.490	224	0.7172	0.616	59
13	0.7499	0.494	219	0.7125	0.605	59
14	0.7455	0.499	214	0.7027	0.623	55
15	0.7469	0.498	216	0.7127	0.601	57
16	0.7361	0.508	204	0.6455	0.701	42
17	0.7372	0.507	205	0.6728	0.659	48
18	0.7243	0.519	192	0.6823	0.627	49
19	0.7254	0.518	193	0.6940	0.610	52
20	0.7200	0.523	188	0.7008	0.601	54
Split 2						
1	0.9140	0.303	373	0.4175	1.079	5
2	0.8833	0.354	259	0.3878	0.900	4
3	0.8685	0.375	224	0.1813	1.080	1
4	0.8677	0.376	224	0.3088	0.934	3
5	0.8219	0.437	152	0.6341	0.552	16
6	**0.7974**	**0.466**	**128**	**0.7911**	**0.416**	**40**
7	0.7904	0.474	122	0.7533	0.460	31
8	0.7882	0.476	120	0.7619	0.451	32
9	0.7913	0.473	123	0.7775	0.430	35
10	0.7807	0.485	114	0.7129	0.495	24
11	0.7874	0.477	119	0.7148	0.492	24
12	0.7687	0.498	106	0.7326	0.482	27
13	0.7601	0.507	100	0.7523	0.458	30
14	0.7291	0.539	83	0.7151	0.504	24
15	0.7343	0.534	85	0.6985	0.529	22
16	0.7302	0.538	83	0.7140	0.499	24

Table 3 continued

limS	Training set, $n=75$			Test set, $n=25$		
	R^2	s	F	R^2	S	F
17	0.7298	0.538	83	0.7197	0.492	25
18	0.7297	0.538	83	0.7387	0.472	28
19	0.7221	0.546	80	0.7172	0.498	24
20	0.7213	0.547	79	0.7198	0.493	25
Split 3						
1	0.8896	0.319	277	0.2438	1.264	2
2	0.8594	0.360	207	0.1363	1.543	0
3	0.8228	0.405	153	0.6985	0.656	23
4	0.8105	0.419	140	0.7518	0.641	36
5	**0.7828**	**0.448**	**116**	**0.8283**	**0.539**	**52**
6	0.7606	0.471	100	0.7715	0.610	35
7	0.7599	0.471	100	0.7668	0.625	35
8	0.7507	0.480	94	0.7350	0.637	30
9	0.7571	0.474	98	0.7233	0.637	26
10	0.7333	0.497	85	0.7556	0.613	31
11	0.7406	0.490	89	0.7451	0.612	29
12	0.7363	0.494	86	0.7350	0.616	28
13	0.7324	0.497	84	0.7388	0.621	28
14	0.7246	0.505	81	0.7160	0.656	26
15	0.7111	0.517	75	0.7442	0.648	29
16	0.7110	0.517	75	0.7428	0.647	29
17	0.7062	0.521	73	0.7456	0.642	29
18	0.7103	0.518	74	0.7371	0.668	29
19	0.7047	0.523	72	0.7729	0.628	34
20	0.6977	0.529	69	0.7609	0.656	32

Best models indicated by bold

However, this attribute is active (not blocked) in splits 2 and 3. One can argue that this attribute can hardly be classified as informative from a probabilistic point of view. Indeed, in the case of split 2, the correlation weight of this attribute is less than 1 (i.e., the presence of this attribute should indicate a decrease of anticancer potential). However, in the case of split 3, the correlation weight of this attribute is more than 1 (i.e., the presence of this attribute should indicate a increase of anticancer potential) for probes 1 and 2, but less than 1 for probe 3 (Table 7).

The presence of three double bonds in a molecule (three '=' symbols in SMILES notation), indicated by the attribute '=003', is also uninformative. In spite of the prevalence of this attribute (in split 1, 67 SMILES of the training set and 23 SMILES of the test set contain this attribute; for splits 2 and 3, these numbers are 66 and 24 for the training set and test set, respectively), the correlation weight of '=003' can be less than 1 (for probes 2 and 3 of split 1 and for all probes of splits 2 and 3), but this correlation weight assumes a value of more than 1 for probe 1 of split 1 (Table 5).

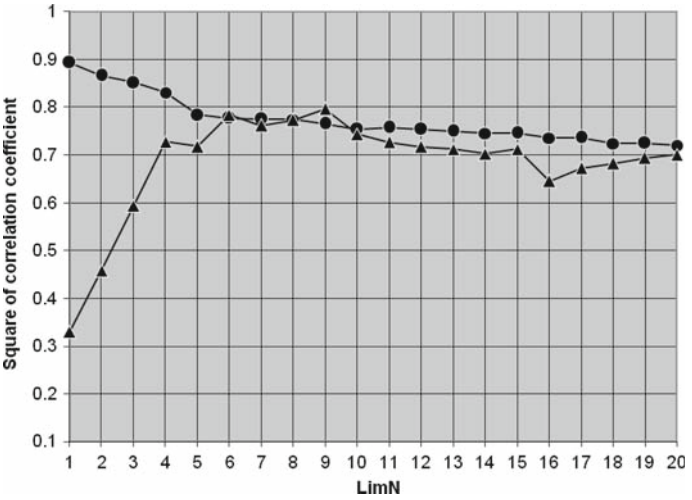


Fig. 1 Split 1: statistical quality of the models for the training (*circles*) and test (*triangles*) sets for different limS Averages of three runs of the Monte Carlo optimization values of the correlation coefficients for the training and test sets were used

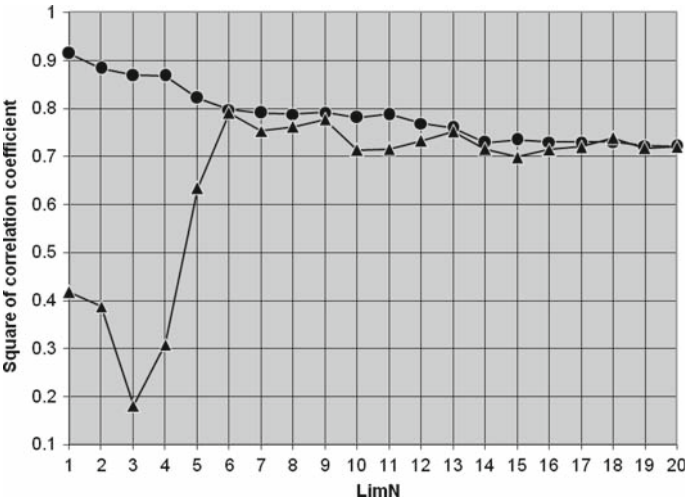


Fig. 2 Split 2: statistical quality of the models for the training (*circles*) and test (*triangles*) sets for different limS Averages of three runs of the Monte Carlo optimization values of the correlation coefficients for the training and test sets were used

Thus, small prevalence (the number of SMILES notations which contain a given attribute is slightly larger than limS) and uncertainty in the correlation weight for the attribute (e.g., both weights of more than 1 and of less then 1) are criteria that could be used to detect an uninformative SA. Vice versa, considerable prevalence and stability of the correlation weight are indicators of an informative SMILES attribute for a given split into training set and test set. Data on the correlation weights can be derived only when relevant experimental data are available. However, even without experimental

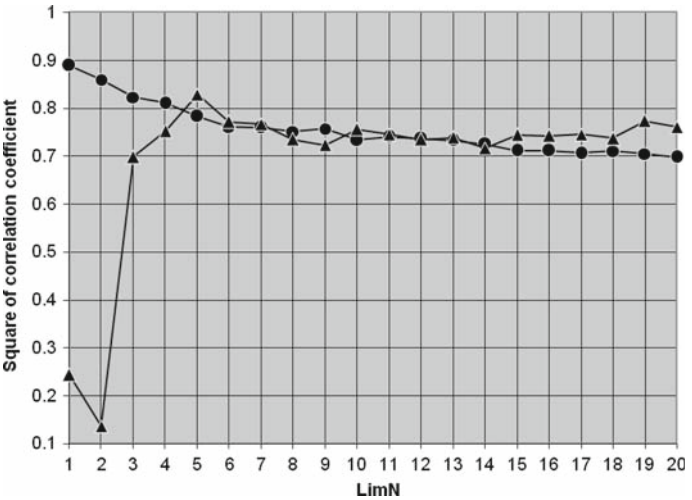


Fig. 3 Split 3: statistical quality of the models for the training (*circles*) and test (*triangles*) sets for different limS Averages of three runs of the Monte Carlo optimization values of the correlation coefficients for the training and test sets were used

Table 4 Statistical characteristics of the model for three splits into training and test set obtained with preferable limS in three runs of the optimization

limS	N_{act}	Probe	Training set, $n=75$			Training set, $n=25$		
			R^2	s	F	R^2	s	F
Split 1								
9	**59**	**1**	**0.7688**	**0.475**	**243**	**0.8025**	**0.488**	**93**
		2	0.7702	0.474	245	0.8012	0.488	93
		3	0.7584	0.486	229	0.7860	0.510	85
		Average	0.7658	0.479	239	0.7966	0.495	90
Split 2								
6	**63**	1	0.8024	0.460	132	0.7454	0.462	29
		2	0.7919	0.472	123	0.8085	0.401	43
		3	**0.7979**	**0.466**	**128**	**0.8194**	**0.383**	**47**
		Average	0.7974	0.466	128	0.7911	0.416	40
Split 3								
5	**65**	1	0.7847	0.446	117	0.7952	0.584	40
		2	**0.7847**	**0.447**	**117**	**0.8644**	**0.503**	**68**
		3	0.7790	0.452	113	0.8254	0.532	49
		Average	0.7828	0.449	116	0.8283	0.539	52

Best models indicated by bold

Table 5 Split 1: list of active SMILES attributes, their correlation weights, and their numbers in SMILES (NS) of training and test sets (limS = 9)

SMILES attribute	CW in probe1	CW in probe2	CW in probe3	NS in training set	NS in test set
tb					
#000_____	0.9538402	0.9808784	0.9614161	74	25
b					
(004_____	1.0069256	0.9917391	0.9931731	11	2
(005_____	1.0005778	0.9981406	1.0053204	40	13
(006_____	1.0102562	1.0066339	1.0177830	13	7
db					
=003_____	1.0065376	0.9964434	0.9903546	67	23
SS_k					
1___(_____	1.0001107	0.9986446	1.0011815	52	16
2___(_____	1.0103654	1.0099957	1.0163725	31	9
3___(_____	1.0462660	1.0381324	1.0458014	26	7
C=__(_____	1.0097072	1.0126836	1.0268410	56	22
C=__2_____	1.0111797	1.0040687	1.0145751	14	2
C___(_____	0.9986995	0.9992303	1.0004702	75	25
C___._____	1.0062031	1.0073227	1.0080156	18	8
C___1_____	1.0105574	1.0051111	1.0154923	53	22
C___2_____	1.0121544	1.0106909	1.0007106	21	3
C___3_____	0.9773942	0.9875076	0.9744696	18	4
C___4_____	1.0130803	1.0059339	1.0152771	32	8
C___@@_____	0.9891423	1.0029960	0.9844050	18	8
C___C=_____	1.0204134	1.0070099	1.0092012	34	16
C___@_____	0.9927229	1.0027127	0.9811036	20	7
C___C_____	1.0159151	1.0108568	1.0200656	55	16
Cl__._____	1.0149667	1.0111851	1.0211254	68	22
F___(_____	0.9925155	0.9959618	0.9911058	42	17
F___2_____	1.0254122	1.0178202	1.0339464	27	13
H___@@_____	1.0065362	1.0096221	1.0070571	17	8
H___@_____	1.0097804	1.0121132	1.0175138	20	7
O=__(_____	0.9955781	0.9933688	0.9879447	68	23
O=__._____	0.9679181	1.0040887	1.0367063	17	3
O=__2_____	1.0282465	1.0100214	1.0129371	13	2
O=__C_____	1.0051733	1.0016534	0.9926740	17	3
N___(_____	1.0052640	1.0036226	1.0074598	75	25
N___._____	0.9844198	0.9930901	0.9823370	27	11
N___2_____	1.0184709	1.0192112	1.0337232	9	1
N___3_____	1.0005210	0.9987077	0.9910126	33	16
N___4_____	1.0237013	1.0131284	1.0173496	11	1
N___C=_____	0.9947178	0.9840230	0.9810249	41	10
N___C_____	0.9986130	0.9987694	0.9976378	59	22
O___(_____	0.9965045	0.9968972	0.9952062	70	24
O___1_____	0.9865404	0.9895067	0.9810333	10	5
O___C_____	1.0028872	1.0025423	1.0048829	15	6
[___1_____	1.0121234	1.0082741	1.0185373	21	9
[___C_____	1.0028848	0.9965592	1.0044614	23	10
[___H_____	1.0067119	0.9985904	1.0102561	22	10
[___N_____	1.0158988	1.0068809	1.0211586	14	6
c___(_____	1.0125070	1.0066789	1.0127835	75	25
c___1_____	0.9850484	0.9870561	0.9871600	24	4
c___2_____	1.0047137	1.0079680	1.0022385	55	22
c___3_____	0.9859138	0.9961110	0.9941915	69	22
c___4_____	1.0037319	1.0024863	1.0059256	62	23
c___5_____	0.9863401	0.9935592	0.9890303	9	1
c___c_____	1.0055883	1.0017049	1.0037475	75	25
n___1_____	0.9684299	0.9870527	0.9804015	19	3
n___2_____	1.0220503	1.0230768	1.0205440	51	21
n___3_____	1.0084609	1.0141060	1.0228955	38	7
n___4_____	1.0068778	1.0049291	1.0087649	53	21
n___c_____	1.0098274	1.0027045	1.0086248	75	25
s___1_____	1.0637583	1.0330432	1.0469326	15	2
s___2_____	0.9887100	0.9833413	0.9787601	20	7
s___4_____	1.0264241	1.0104732	1.0345065	30	13
s___c_____	1.0016417	1.0006302	1.0007508	67	20

The blocked attributes are omitted

Table 6 Split 2: list of active SMILES attributes, their correlation weights, and their numbers in SMILES (NS) of training and test sets (limS = 6)

SMILES attribute	CW in Probe1	CW in Probe 2	CW in Probe 3	NS in training set	NS in test set
tb					
#000_____	0.9778265	0.9618957	0.9813088	74	25
b					
(003_____	0.9816578	0.9671756	0.9915241	7	3
(004_____	1.0060957	0.9990110	1.0033721	11	2
(005_____	1.0097452	1.0069479	1.0107020	40	13
(006_____	1.0216977	1.0229979	1.0179855	13	7
db					
=003_____	0.9933571	0.9773153	0.9958456	66	24
=004_____	0.9888334	0.9604921	0.9894369	6	0
SS_k					
1___(_____	1.0088002	1.0051624	1.0048705	52	16
2___(_____	1.0046415	1.0189892	1.0108256	32	8
3___(_____	1.0341271	1.0478242	1.0342879	25	8
C=__(_____	1.0171558	1.0344657	1.0149327	58	20
C=__2_____	1.0148793	1.0391856	1.0115735	11	5
C___(_____	0.9999407	0.9985423	0.9998307	75	25
C___._____	1.0067460	1.0061350	1.0052181	17	9
C___1_____	1.0132302	1.0119075	1.0071315	56	19
C___2_____	1.0004630	1.0054178	1.0050329	18	6
C___3_____	0.9828371	0.9754989	0.9831477	17	5
C___4_____	1.0142584	1.0148981	1.0094767	31	9
C___@@_____	1.0099845	0.9894355	0.9953226	20	6
C___C=_____	1.0162726	1.0262053	1.0143798	37	13
C___@_____	1.0199640	0.9860964	0.9978823	21	6
C___C_____	1.0105126	1.0139354	1.0078916	53	18
Cl__._____	1.0159930	1.0177615	1.0107319	66	24
F___(_____	0.9947550	0.9857250	0.9946057	44	15
F___2_____	1.0148933	1.0429714	1.0177395	28	12
H___@@_____	1.0114270	0.9960125	0.9994598	19	6
H___@_____	1.0003344	1.0024301	0.9957010	21	6
O=__(_____	0.9965386	1.0008260	0.9952664	67	24
O=__._____	1.0176481	1.0719661	0.9901368	15	5
O=__2_____	1.0238313	1.0338577	1.0121241	10	5
O=__C_____	1.0134058	0.9829535	1.0269374	15	5
N___(_____	0.9965036	1.0000913	1.0008129	75	25
N___._____	0.9963321	0.9915191	0.9960415	29	9
N___1_____	1.0177727	1.0088842	1.0068400	8	4
N___2_____	1.0165136	1.0329303	1.0129731	9	1
N___3_____	0.9992317	0.9950339	0.9961654	36	13
N___4_____	1.0129574	1.0332943	1.0177879	9	3
N___C=_____	0.9658609	0.9667431	0.9879131	39	12
N___C_____	0.9972448	0.9962910	0.9969706	60	21
O___(_____	0.9866033	0.9863761	0.9928568	69	25
O___1_____	0.9877221	0.9720977	0.9866998	12	3
O___C_____	1.0041509	1.0059511	1.0036623	16	5
[___1_____	1.0238752	1.0266190	1.0163632	22	8
[___C_____	0.9916985	0.9992887	0.9961809	25	8
[___H_____	0.9944438	1.0162172	1.0122050	24	8
[___N_____	1.0137984	1.0257750	1.0118613	17	3
c___(_____	1.0057637	1.0093881	1.0065657	75	25
c___1_____	0.9971611	0.9905747	0.9907554	21	7
c___2_____	0.9976407	1.0043669	1.0002008	58	19
c___3_____	0.9878292	0.9751282	0.9917536	67	24
c___4_____	1.0046855	1.0056309	1.0011469	63	22
c___5_____	0.9947451	0.9918079	0.9928074	9	1
c___c_____	0.9998288	0.9983328	1.0013955	75	25
n___1_____	0.9830864	0.9673910	0.9803039	17	5
n___2_____	1.0057613	1.0239479	1.0089735	53	19
n___3_____	1.0155376	1.0164489	1.0050820	33	12
n___4_____	1.0078733	1.0081104	1.0051191	55	19
n___5_____	0.9896612	0.9927245	0.9949830	6	1
n___c_____	1.0014870	0.9981225	1.0020674	75	25
s___1_____	1.0608505	1.0933231	1.0443981	13	4
s___2_____	0.9919999	0.9965154	0.9847545	20	7
s___4_____	1.0257410	1.0410035	1.0188340	31	12
s___c_____	1.0006827	1.0053294	1.0006227	65	22

The blocked attributes are omitted

Table 7 Split 3: list of active SMILES attributes, their correlation weights, and their numbers in SMILES (NS) of training and test sets (limS = 5)

SMILES attribute	CW in Probe1	CW in Probe 2	CW in Probe 3	NS in training set	NS in test set
tb					
#000_____	0.9621592	0.9497921	0.9816455	74	25
b					
(003_____	0.9815572	1.0044196	0.9994500	8	2
(004_____	1.0013028	1.0296071	1.0091396	11	2
(005_____	1.0197642	1.0344675	1.0195228	39	14
(006_____	1.0268523	1.0415978	1.0198709	13	7
db					
=003_____	0.9884570	0.9914248	0.9884271	66	24
=004_____	1.0054205	1.0070592	0.9860682	5	1
SS_k					
1___(_____	0.9999616	1.0086408	0.9980004	53	15
2___(_____	1.0165580	1.0347494	1.0049201	32	8
3___(_____	1.0521844	1.0697399	1.0057306	24	9
C=__(_____	1.0129000	1.0106854	1.0111276	58	20
C=__2_____	1.0113902	1.0356614	1.0202594	12	4
C___(_____	1.0009022	1.0020768	1.0015151	75	25
C___._____	1.0092467	1.0103786	1.0017177	18	8
C___1_____	1.0146842	1.0144150	0.9980776	54	21
C___2_____	1.0023651	0.9965816	0.9945243	19	5
C___3_____	0.9917420	0.9879109	0.9852441	17	5
C___4_____	1.0088102	1.0228595	1.0087346	32	8
C___@@____	0.9983925	1.0104875	1.0028854	18	8
C___C=____	1.0162671	1.0308056	1.0133050	38	12
C___@_____	1.0200239	1.0020967	0.9968486	19	8
C___C_____	1.0122098	1.0152454	1.0056880	56	15
Cl__._____	1.0165174	1.0257764	1.0114340	67	23
F___(_____	0.9997315	1.0002027	0.9997634	44	15
F___2_____	1.0225508	1.0292075	1.0060076	27	13
F___3_____	0.9565775	0.9516350	0.9791420	5	0
H___@@____	1.0153568	1.0095559	0.9953538	18	7
H___@_____	0.9951692	1.0185354	1.0006017	19	8
O=__(_____	0.9894990	0.9905463	0.9956920	67	24
O=__._____	1.0124421	0.9931883	0.9946305	16	4
O=__2_____	0.9946226	1.0029398	0.9922557	11	4
O=__C_____	0.9835682	1.0051858	1.0135789	16	4
N___(_____	0.9979574	0.9966660	0.9988118	75	25
N___._____	0.9940010	0.9880667	0.9926867	28	10
N___1_____	0.9950157	0.9970878	0.9916153	9	3
N___2_____	1.0270302	1.0356577	1.0099827	9	1
N___3_____	0.9781524	0.9858350	0.9911510	35	14
N___4_____	1.0577702	1.0593358	1.0209524	12	0
N___C=____	0.9830417	0.9673134	0.9960350	38	13
N___C_____	0.9920399	0.9903145	0.9964805	59	22
O___(_____	0.9880073	0.9827073	0.9936899	69	25
O___1_____	0.9836409	0.9765493	0.9895800	9	6
O___C_____	1.0035079	1.0034375	1.0007596	13	8
[___1_____	1.0125687	1.0245552	1.0007208	20	10
[___C_____	0.9936303	0.9966326	1.0027614	23	10
[___H_____	1.0118704	0.9944021	1.0008652	23	9
[___N_____	1.0049133	1.0204371	1.0080089	14	6
c___(_____	1.0117418	1.0168111	1.0062319	75	25
c___1_____	0.9936918	0.9955021	0.9975255	23	5
c___2_____	1.0017019	1.0014710	1.0018272	57	20
c___3_____	0.9988700	0.9942194	0.9926086	67	24
c___4_____	0.9945454	1.0009012	1.0012178	62	23
c___5_____	0.9811383	0.9781146	0.9939911	9	1
c___c_____	1.0101802	1.0102950	1.0021466	75	25
n___(_____	0.9330230	0.9064532	0.9552828	5	0
n___1_____	0.9918858	0.9764785	0.9897546	18	4
n___2_____	1.0126866	1.0260082	1.0114082	52	20
n___3_____	1.0098498	1.0265103	1.0073757	34	11
n___4_____	1.0068324	1.0126311	1.0091957	52	22
n___5_____	0.9936415	0.9967540	0.9956270	6	1
n___c_____	1.0219793	1.0186105	1.0033268	75	25
s___1_____	1.0427817	1.0706896	1.0153638	14	3
s___2_____	1.0019118	0.9876430	0.9912262	20	7
s___4_____	1.0303562	1.0534105	1.0096712	31	12
s___c_____	0.9958532	0.9911567	0.9960015	65	22

The blocked attributes are omitted

Table 8 Anticancer activity: experimental values and values calculated with Eq. 3 (split 1, probe 1, limS = 9)

ID	SMILES	DCW(9)	Expr	Calc
Training set				
1	Cl.O=C(O)C2=CN(c1nc(c(F)cc1C2=O)N3CCC(N)C3)c4ccccc4	1.1026346	−0.814	−1.073
2	Cl.O=C(O)C2=CN(c1nc(c(F)cc1C2=O)N3CCC(N)C3)c4ccc(F)cc4	1.1180538	−0.735	−0.631
4	Cl.NC1CCN(C1)c2nc3c(cc2F)C(=O)C(=CN3C4CC4)C(=O)O	1.0924168	−1.376	−1.366
5	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4cccs4	1.1243058	−0.506	−0.452
7	Cl.Cc1cc(on1)N2C=C(C(=O)O)C(=O)c3cc(F)c(nc23)N4CCC(N)C4	1.1338148	−0.191	−0.180
8	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nncs4	1.1262901	−1.165	−0.395
10	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1325841	−1.410	−0.215
11	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nc(C)cs4	1.1627775	0.883	0.651
13	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nc(C)c(C)s4	1.1604339	0.350	0.583
15	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nc(cs4)c5ccccc5	1.1582060	0.939	0.520
17	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4ncc(Br)s4	1.1559439	0.728	0.455
18	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4ncc(OC)s4	1.1537268	0.064	0.391
19	Cl.NC1CCN(C1)c2nc5c(cc2F)C(=O)C(=CN5c3nc4ccccc4s3)C(=O)O	1.1640017	0.939	0.686
20	Cl.NC1CCN(C1)c3ncc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1732082	0.882	0.950
21	Cl.NC1CCN(C1)c2nc3N(N=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.0992560	−1.424	−1.170
22	Cl.NC1CCN(C1)c2cc3N(C=C(C(=O)c3cn2)C(=O)O)c4nccs4	1.1243801	−0.447	−0.450
23	Cl.NC1CCN(C1)c2cc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1086988	−1.427	−0.900
25	Cl.O=C(O)C2=CN(c1nccs1)c3nc(NCCN)c(F)cc3C2=O	1.0990238	−1.457	−1.177
26	Cl.O=C(O)C3=CN(c1nccs1)c4nc(N2CCCC2)c(F)cc4C3=O	1.1474605	1.019	0.212
27	Cl.O[C@@H]1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1598712	0.763	0.567
28	Cl.O[C@H]1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1678218	1.275	0.795
30	Cl.N[C@H]1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1679046	1.127	0.798
31	Cl.CN(C)C1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1704181	0.850	0.870
33	Cl.C[C@H]1CN(C[C@@H]1N)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1772150	1.550	1.064
35	Cl.N[C@H]1CN(C[C@@H]1O)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1409430	−0.728	0.025

Table 8 continued

ID	SMILES	DCW(9)	Expr	Calc
36	Cl.N[C@@H]1CN(C[C@H]1OC)c2nc3N (C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1467575	0.775	0.191
38	Cl.N[C@H]1CN(C[C@H]1OC)c2nc3N (C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1546182	0.684	0.417
39	Cl.CN[C@H]1CN(C[C@@H]1OC)c2nc3N (C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1705072	1.367	0.872
40	Cl.OC[C@@]1(N)CCN(C1)c2nc3N (C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1610239	0.684	0.600
41	Cl.O=C(O)C4=CN(c1nccs1)c5nc(N2C[C@@H] 3NCCO[C@H]3C2)c(F)cc5C4=O	1.1298487	−0.379	−0.293
43	Cl.CN2CC=C1CN(CC12)c3nc4N (C=C(C(=O)c4cc3F)C(=O)O)c5nccs5	1.1862988	1.037	1.325
44	Cl.N[C@H]1C=CC[C@H]2CN(C[C@@H]12) c3nc4N(C=C(C(=O)c4cc3F)C(=O)O)c5nccs5	1.1678044	0.455	0.795
45	Cl.NCC1CCN(C1)c2nc3N(C=C(C(=O) c3cc2F)C(=O)O)c4nccs4	1.1506093	0.586	0.302
47	Cl.NCC1=CCN(C1)c2nc3N(C=C(C(=O) c3cc2F)C(=O)O)c4nccs4	1.1362022	0.664	−0.111
48	Cl.NC1CCN(CC1)c2nc3N(C=C(C(=O) c3cc2F)C(=O)O)c4nccs4	1.1506093	−0.622	0.302
49	Cl.COc1ccccc1N2CCN(CC2)c3nc4N (C=C(C(=O)c4cc3F)C(=O)O)c5nccs5	1.1215053	−0.317	−0.532
51	Cl.O=C(O)C3=CN(c2nc(N1CCSCC1) c(F)cc2C3=O)c4nccs4	1.1315070	−0.647	−0.246
52	Cl.O=C(O)C3=CN(c1nccs1) c4nc(n2ccnc2)c(F)cc4C3=O	1.1038615	−1.447	−1.038
53	Cl.O=C(O)C2=CN(c1nccs1) c3nc(SCCN)c(F)cc3C2=O	1.0947873	−1.436	−1.298
54	Cl.O=C(O)C2=CN(c1nccs1) c3nc(O)c(F)cc3C2=O	1.0659911	−1.512	−2.124
55	Cl.NC1CCN(C1)c3ccc4C(=O) C(=CN(c2nccs2)c4n3)C(=O)O	1.1421628	1.553	0.060
56	Cl.[O−][N+](=O)c3cc4C(=O)C(=CN (c1nccs1)c4nc3N2CCC(N)C2)C(=O)O	1.1193405	−0.755	−0.594
57	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O) c3cc2N)C(=O)O)c4nccs4	1.1393664	−0.433	−0.020
59	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O) c3cc2Cl)C(=O)O)c4nccs4	1.1128450	0.237	−0.781
60	Cl.NC1CCN(C1)c3cc(Cl)c4C(=O) C(=CN(c2nccs2)c4n3)C(=O)O	1.1756675	0.903	1.020
61	Cl.FC(F)(F)c1cc(nc2N(C=C(C(=O) c12)C(=O)O)c3nccs3)N4CCC(N)C4	1.1623432	0.637	0.638
62	Cl.NC1CCN(C1)c3cc(N) c4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1880775	1.661	1.376
63	Cl.CNC1CCN(C1)c3ccc4C(=O) C(=CN(c2nccs2)c4n3)C(=O)O	1.1658173	1.570	0.738
65	Cl.CC1(NC)CCN(C1)c3ccc4C(=O) C(=CN(c2nccs2)c4n3)C(=O)O	1.1882512	1.331	1.381
66	Cl.C[C@H]1CN(C[C@@H]1N) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1871711	1.588	1.350

Table 8 continued

ID	SMILES	DCW(9)	Expr	Calc
68	Cl.ClC[C@H]1CN(C[C@@H]1N) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1975109	1.228	1.646
69	Cl.CC1CC(N)CN1c3ccc4C(=O) C(=CN(c2nccs2)c4n3)C(=O)O	1.1477807	0.528	0.221
70	Cl.C[C@H]1[C@H](N)CCN1 c3ccc4C (=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1656954	0.597	0.734
71	Cl.C[C@H]1CN(C[C@]1(C)N) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1904922	1.285	1.445
72	Cl.O=C(O)C2=CN(c1nccs1)c3nc(ccc3C2=O) N4C[C@@H]5CCCN[C@@H]5C4	1.1843114	1.369	1.268
74	Cl.CN[C@@H]1CN(C[C@H]1OC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1804066	1.728	1.156
76	Cl.CCN[C@H]1CN(C[C@@H]1OC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1991929	1.220	1.694
77	Cl.CN[C@H]1CN(C[C@@H]1SC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1930665	1.320	1.519
79	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)N4CCCC4	1.1704041	0.715	0.869
80	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)N4CCNCC4	1.1488752	0.797	0.252
82	Cl.COc1ccccc1N2CCN(CC2) c4ccc5C(=O)C(=CN(c3nccs3)c5n4)C(=O)O	1.1290084	−0.334	−0.317
83	Cl.CNC1CCCN(C1)c3ccc4C (=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1843714	−0.130	1.270
84	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)N4Cc5ccccc5CC4	1.1127096	−1.393	−0.785
85	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)N4CCOCC4	1.1587309	0.475	0.535
86	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)c4ccccc4	1.1065078	−1.258	−0.962
87	Cl.Cc1cc(cc(C)n1)c3ccc4C (=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1047184	−1.287	−1.014
89	Cl.O=C(O)C2=CN(c1nccs1) c3nc(C#C)ccc3C2=O	1.1100466	−0.889	−0.861
90	Cl.O=C(O)C2=CN(c1nccs1) c3nc(ccc3C2=O)c4ccnn4	1.1023007	−1.107	−1.083
93	NC1CCN(C1)c3nc4N(C=C (Cc2ccccc2)C(=O)c4cc3F)c5nccs5	1.1491248	0.769	0.259
95	NC1CCN(C1)c2nc3N(C=C (C(=O)c3cc2F)C(N)=O)c4nccs4	1.1535595	0.144	0.386
96	NC1CCN(C1)c2nc3N(C=C(O) C(=O)c3cc2F)c4nccs4	1.1463956	0.316	0.181
97	CN[C@H]1CN(C[C@@H]1OC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C=O	1.1921655	1.096	1.493
98	CN[C@H]1CN(C[C@@H]1OC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)CO	1.1748922	0.847	0.998
99	CN[C@H]1CN(C[C@@H]1OC) c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(C)=O	1.1620083	0.829	0.629
100	CN[C@H]1CN(C[C@@H]1OC)c3ccc4C (=O)C(=CN (c2nccs2)c4n3)C\C=C/C(=O)OC	1.1751477	0.682	1.005

Table 8 continued

ID	SMILES	DCW(9)	Expr	Calc
Test set				
3	Cl.O=C(O)C2=CN(c1nc(c(F)cc1C2=O)N3CCC(N)C3)c4ccccn4	1.1107533	−1.433	−0.841
6	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4ccno4	1.0981767	−1.418	−1.201
9	Cl.Cn4nccc4N1C=C(C(=O)O)C(=O)c2cc(F)c(nc12)N3CCC(N)C3	1.1001760	−1.429	−1.144
12	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4ncc(C)s4	1.1529393	0.175	0.369
14	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nc(cs4)C(C)(C)C	1.1435082	−0.310	0.098
16	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4ncc(Cl)s4	1.1559439	0.639	0.455
24	Cl.NC1CCN(C1)c3c(F)cc4C(=O)C(=CN(c2nccs2)c4c3F)C(=O)O	1.0707181	−2.147	−1.988
29	Cl.N[C@@H]1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1599535	1.371	0.570
32	Cl.CC1(N)CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1876783	1.175	1.364
34	Cl.C[C@@H]1CN(C[C@@H]1N)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1692005	0.835	0.835
37	Cl.N[C@H]1CN(C[C@@H]1OC)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1467575	1.246	0.191
42	Cl.O=C(O)C4=CN(c1nccs1)c5nc(N2C[C@@H]3CCCN[C@@H]3C2)c(F)cc5C4=O	1.1281136	0.136	−0.343
46	Cl.NC[C@H]1CN(C[C@H]1Cl)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1519520	0.434	0.340
50	Cl.NCC1CN(CCO1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)O)c4nccs4	1.1265070	−0.630	−0.389
58	Cl.NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2O)C(=O)O)c4nccs4	1.1089551	−1.062	−0.892
64	Cl.CC1(N)CCN(C1)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1977229	1.291	1.652
67	Cl.C[C@H]1CN(C[C@@H]1NC)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1777829	1.507	1.081
73	Cl.N[C@H]1CN(C[C@@H]1OC)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1564561	1.333	0.469
75	Cl.CN[C@@H]1CN(C[C@H]1OC)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1804066	1.282	1.156
78	Cl.CN[C@H]1CN(C[C@H]1OC)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1884980	1.303	1.388
81	Cl.C[C@H]1CN(C[C@@H](C)N1)c3ccc4C(=O)C(=CN(c2nccs2)c4n3)C(=O)O	1.1813492	0.678	1.183
88	Cl.O=C(O)C2=CN(c1nccs1)c3nc(C=C)ccc3C2=O	1.0852346	−0.524	−1.572
91	NC1CCN(C1)c2nc3N(C=C(C(=O)c3cc2F)C(=O)OCC)c4nccs4	1.1574530	0.472	0.498
92	NC1CCN(C1)c2nc3N(C=CC(=O)c3cc2F)c4nccs4	1.1677941	0.277	0.794
94	NC1CCN(C1)c2nc3N(C=C(C=O)C(=O)c3cc2F)c4nccs4	1.1691781	1.229	0.834

data, the prevalence of SMILES attribute for a new compound is the parameter that governs its activity. Hence, the prevalence, which is predicted and known in advance, can be used as an indicator for selecting the robust split into training set and test set (from the probabilistic point of view) for a QSAR model of the activity of potential anticancer agents.

For instance, one can see from Tables 5, 6, and 7, that the SMILES attribute 'Cl__.' has considerable prevalence in both the training and test sets (e.g., split 1: 68–22; split 2: 66–24; and split 3: 67–23) and the correlation weight is more than 1 for all runs of the Monte Carlo optimization. Thus, this attribute is a promoter of the increase of the anticancer potential. A stable situation (from the probabilistic point of view) occurs for "c__3__", because correlation weights of this attribute are less than 1 for all runs of the optimization. The distribution of this attribute is 69–22, 67–24, and 67-24 for splits 1, 2, and 3, respectively.

It is to be noted that all the conclusions above concern the considered splits 1–3. It is quite possible that other splits may lead to different outcomes. However, there is an indicator for the rational selection of a split: the prevalence of the attributes.

Criticism of the QSAR approach appeared in recent papers [35–37]. In fact, the probabilistic logic and the reproducibility are the only arguments in the polemics related to the issue: "QSAR: Dead or alive?" [37]. Since in the present study we have applied both mentioned items for QSAR analysis of the investigated anticancer agents we strongly believe that, if properly done, QSAR approaches will be alive for years to come.

5 Conclusions

SMILES-based optimal descriptors can be used for QSAR modeling of anticarcinogenic activity for a series of 7- and 3-substituted 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines. The developed models provide satisfactory predictions for three random splits of data into training set and test set. Probabilistic analysis of the correlation weights together with the prevalence (distribution) of SA in the training set and test set can be useful in the search for mechanistic interpretations of the activity of anticancer agents.

Acknowledgments The authors thank the Marie Curie Fellowship for financial support (contract ID 39036, CHEMPREDICT). J.L. would like to thank the NSF for support through RISE grant # HRD-0734645.

References

1. M. Atanasova, S. Ilieva, B. Galabov, Eur. J. Med. Chem. **42**, 1184–1192 (2007)
2. Y. Marrero-Ponce, Y.A. Castillo-Garit, E.A. Castro, F. Torrens, R. Rotondo, J. Math. Chem. **44**, 755–786 (2008)
3. P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. **16**, 7944–7955 (2008)
4. P.R. Duchowicz, M.G. Vitale, E.A. Castro, J. Math. Chem. **44**, 541–549 (2008)
5. J.A. Castillo-Garit, O. Martinez-Santiago, Y. Marrero-Ponce, G.M. Casañola-Martín, F. Torrens, Chem. Phys. Lett. **464**, 107–112 (2008)

6. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, QSAR Comb. Sci. **25**(10), 928–935 (2006)
7. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Polymer **47**, 3240–3248 (2006)
8. T. Puzyn, A. Mostrag, N. Suzuki, J. Falandysz, Atm. Environ. **42**, 6627–6636 (2008)
9. S. Ray, C. Sengupta, K. Roy, Cent. Eur. J. Chem. **6**, 267–276 (2008)
10. K. Roy, P.P. Roy, Chem. Biol. Drug Des. **72**, 370–382 (2008)
11. K. Roy, G. Ghosh, Chem. Biol. Drug Des. **72**, 383–394 (2008)
12. A.A. Toropov, A.P. Toropova, I. Gutman, Croat. Chem. Acta **78**, 503–509 (2005)
13. D.M. Hawkins, J.J. Kraker, S.C. Basak, D. Mills, SAR QSAR Environ. Res. **19**, 525–539 (2008)
14. B. Hollas, I. Gutman, N. Trinajstić, Croat. Chem. Acta **78**, 489–492 (2005)
15. A.R. Katritzky, R. Petrukhin, D. Tatham, S.C. Basak, E. Benfenati, M. Karelson, U. Maran, J. Chem. Inf. Comput. Sci. **41**, 679–685 (2001)
16. M.-O. Fouchécourt, M. Béliveau, K. Krishnan, Sci. Total Environ. **274**, 125–135 (2001)
17. V.K. Gombar, V.K. Kapoor, Eur. J. Med. Chem. **25**, 689–695 (1990)
18. A.C.S. Arruda, B.D.S. Junkes, E.S. Souza, R.A. Yunes, V.E.F. Heinzen, J. Chemometr. **22**, 186–194 (2008)
19. T. Puzyn, N. Suzuki, M. Haranczyk, J. Rak, J. Chem. Inf. Model. **48**, 1174–1180 (2008)
20. D. Weininger, J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988)
21. D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci **29**, 97–101 (1989)
22. D. Weininger, J. Chem. Inf. Comput. Sci. **30**, 237–243 (1990)
23. ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, (2007), www.acdlabs.com
24. MDL QSAR version 2.2. MDL Information Systems Inc., San Leandro, CA (2003)
25. Daylight Chemical Information Systems, Inc. (2008), http://www.daylight.com
26. D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model **45**, 386–393 (2005)
27. D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model **46**, 836–843 (2006)
28. D. Vidal, J. Blobel, Y. Pérez, M. Thormann, M. Pons, Eur. J. Med. Chem. **42**, 1102–1108 (2007)
29. A. Toropov, K. Nesmerak, I. Raska Jr., K. Waisser, K. Palat, Comput. Biol. Chem. **30**, 434–437 (2006)
30. A.A. Toropov, A.P Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. Sec A **44**, 1545–1552 (2005)
31. A.A. Toropov, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. **441**, 119–122 (2007)
32. A.A. Toropov, A.P. Toropova, I. Raska Jr, Eur. J. Med. Chem. **43**, 714–740 (2008)
33. A.A. Toropov, E. Benfenati, Bioorg. Med. Chem. **16**, 4801–4809 (2008)
34. A.A. Toropov, A.P. Toropova, E. Benfenati, Chem. Phys. Lett. **461**, 343–347 (2008)
35. A.M. Doweyko, J. Comput. Aid. Mol. Des. **18**, 587–596 (2004)
36. S.R. Johnson, J. Chem. Inf. Model **48**, 25–26 (2008)
37. A.M. Doweyko, J. Comput. Aid. Mol. Des. **22**, 81–89 (2008)