J Math Chem (2010) 48:959–987
DOI 10.1007/s10910-010-9719-x

ORIGINAL PAPER

QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL

Alla P. Toropova · Andrey A. Toropov ·
Emilio Benfenati · Danuta Leszczynska ·
Jerzy Leszczynski

Received: 7 June 2010 / Accepted: 31 July 2010 / Published online: 15 August 2010
© Springer Science+Business Media, LLC 2010

Abstract Quantitative structure – activity relationships (QSAR) for prediction of binding affinities (pEC50, i.e., minus decimal logarithm of the 50% effective concentration) of 48 fullerene derivatives inhibitors of the HIV-1 PR (human immunodeficiency virus type 1 protease) have been developed using the software CORAL. CORAL (CORrealtions And Logic) is a freeware aimed to assist QSAR modeling by application of descriptors calculated with SMILES (simplified molecular input line entry system). Three methods of the QSAR modeling of pEC50 have been examined: 1. classic scheme, where model is constructed with a training set and checked up with a validation set; 2. the balance of correlations, where training set is separated into subtraining set and calibration set that is used as a preliminary validation set (the target function provides maximal correlation coefficients for the training and calibration sets with their minimal difference): the final estimation of predictability is based on an external validation set (structures which are not used in

Electronic supplementary material The online version of this article (doi:10.1007/s10910-010-9719-x) contains supplementary material, which is available to authorized users.

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

D. Leszczynska
Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering,
Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, USA

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

building up of the model); and 3. the balance of correlations developed by applying slopes in the plots of the experimental pEC50 versus the calculated pEC50 (separately for the subtraining and the calibration set). A validation set is also used in this case. The best prediction has been obtained for the balance of correlations with ideal slopes. These approaches have been examined for three random splits: into the subtraining set, calibration set, and the validation set. Reliability of the $R_m{}^2$ criterion, which has been suggested by P.P Roy and K. Roy for estimation of external predictability of QSAR models has been confirmed. Statistical characteristics of the best model are as follows: $n = 27$, $r^2 = 0.9030$, $q^2 = 0.8855$, $s = 0.406$, $F = 233$ (subtraining set); $n = 15$, $r^2 = 0.9720$, $R^2_{pred} = 0.9661$, $s = 0.980$, $F = 451$ (calibration set); $n = 6$, $r^2 = 0.9224$, $R^2_{pred} = 0.7956$, $s = 0.950$, $F = 48$; $R_m{}^2 = 0.7812$ (validation set).

Keywords Fullerene · QSAR · HIV-1 PR · SMILES · Optimal descriptor · Balance of correlations

1 Introduction

Various programs for calculations of molecular descriptors and the multiple linear regression analysis (MLRA) have been developed over the last few years [1–4]. Typically, QSAR models are built up through a series of steps, going from the chemical structure, to chemical descriptors/fragments, to algorithm, to validation. These steps are done using separate programs, in many cases commercial. These activities are complex, require skill, and their implementation is complex, due to the fact that integration of several components are necessary.

We have attempted to suggest convenient alternative for the QSPR/QSAR analysis [5]. It is a system based on representation of molecular structure by simplified molecular input line entry system (SMILES) [6–9]. The software (CORAL) that is a representation of the system involves a provider of correlations of optimal SMILES-based descriptors together with a data for probabilistic estimation of these correlations. SMILES gradually becomes an widely used component of databases for molecular properties which are available on the Internet [10,11]. Due to such a progress, the SMILES-based QSPR/QSAR analysis is very attractive and has been already recognized as quite promising approach [12].

Majority of the QSAR analyses are dedicated to organic substances. However, new groups of compounds are being studied using such approaches. Quite recently QSAR methods have been also succesfuly applied to nanomaterials [for a recent review see 13]. Fullerene derivatives (being de-facto organic substances) represent an example of an important group of nanoparticles. Their applications are vital in number of areas of modern life sciences and in particular, they can form quite effective anti-HIV-1 agents [14,15].

The aim of the present work is the evaluation of the CORAL as a tool for QSAR modeling of measured anti-HIV-1 activity (pEC50) of fullerene derivatives taken from Ref. [15].

2 Method

2.1 Data set

The measured binding affinity data for fullerene derivatives (minus decimal logarithm of the 50% effective concentration, pEC50) were taken from Ref. [15]. SMILES notations for these structures have been generated by ChemSketch software [16]. Three random splits of these 48 compounds into subtraining, calibration, and validation set were examined. The validation set for split 1 was taken from Ref. [15].

2.2 Descriptors

The models examined in the present work represent one-variable correlations between the endpoint under consideration (pEC50) and the optimal descriptor that is defined in CORAL as

$$\mathrm{DCW(Threshold)} = \alpha \sum \mathrm{CW}\,(\mathrm{s_k}) + \beta \sum \mathrm{CW}\,(\mathrm{ss_k}) + \gamma \sum \mathrm{CW}\,(\mathrm{sss_k}) \qquad (1)$$

where $\mathrm{s_k}$, $\mathrm{ss_k}$, and $\mathrm{sss_k}$ represent one-, two-, and three-elements SMILES attributes. The essence of the $\mathrm{s_k}$, $\mathrm{ss_k}$, and $\mathrm{sss_k}$ (if a SMILES is a sequence of 'ABCDE') can be described by the following scheme:

$$\begin{array}{lll} \mathrm{ABCDE} \rightarrow & \mathrm{A+B+C+D+E} & (\mathrm{s_k}) \\ \mathrm{ABCDE} \rightarrow & \mathrm{AB+BC+CD+DE} & (\mathrm{ss_k}) \\ \mathrm{ABCDE} \rightarrow & \mathrm{ABC+BCD+CDE} & (\mathrm{sss_k}) \end{array}$$

The SMILES elements of A,B,C,D, and E can be one or two symbols. Twelve symbols are used for the representation of these SMILES attributes. There are three zones placed in positions 1–4 (zone 1) 5–8 (zone 2) 9–12 (zone 3). The $\mathrm{sss_k}$ element involves all three zones; $\mathrm{ss_k}$ involves zone 1 and zone 2; $\mathrm{s_k}$ involves only zone 1. Unused positions are indicated by 'x' in this representation of a SMILES attribute. In the present study $\alpha = 1$, $\beta = 1$, $\gamma = 0$ values have been used.

There are SMILES-elements containing two symbols, e.g., Cl, Br, etc.. Also, there are SMILES-elements containing three symbols, e.g., %10, %11, etc.. These elements are indicators of cycles if the total number of cycles in molecular structure is more than 9 [16].

$\mathrm{CW(s_k)}$, $\mathrm{CW(ss_k)}$, and $\mathrm{CW(sss_k)}$ are correlation weights for the $\mathrm{s_k}$, $\mathrm{ss_k}$, $\mathrm{sss_k}$, respectively.

The threshold is the parameter used to define rare (noise) SMILES attributes. The rare SMILES attributes can lead to overtraining: excellent correlation for the training set accompanied by poor correlation for the validation set. Thus they can bring 'noise'. The threshold can be defined as 0, 1, 2, …N. The N is the number of compounds in training set. For instance, if threshold is defined as 5, all SMILES attributes attributes which are less frequent than 5 in the SMILES notation of the training set are classified as rare.

Since the descriptors we use are optimized through the overall procedure we describe here, we call "optimal" the descriptor.

Three methods for building up the pEC50 models have been used: classic scheme, balance of correlations, and balance of correlations with ideal slopes. Each method has individual target function for Monte Carlo optimization.

Classic scheme: maximum of R

In this case **R** represents the correlation coefficient between endpoint and optimal descriptor calculated with Eq. 1 for the training set. The software is aimed to maximize **R**.

Balance of correlations: maximum of BC

In this case the software maximizes BC

$$\mathbf{BC = R + R' - abs\left(R - R'\right){}^{*}dR - weight}$$

where R and R′ are correlation coefficient between endpoint and optimal descriptor for subtraining set and calibration set. The role of the calibration set is a preliminary validation of the model. This approach is an attempt to avoid the overtraining. In other words, in the case of balance of correlations, the training set is split into two sets: subtraining and calibration. The **dR-weight** is an empirical parameter: **dR-weight = 0.1**

The balance of correlations keeps into account the fact that different correlation coefficients may occur for the subtraining and validation set. If this occurs, it means that the model is not stable, and different results are obtained for the subtraining and validation set. The CORAL is aimed to generate minimal difference between abovementioned correlation coefficients.

Balance of correlations with ideal slopes: maximum of IS

In this case the software maximizes IS

$$\mathbf{IS = BC - abs\left(C0 + C0' + C1 - C1'\right){}^{*}dC - weight}$$

Here C0 and C0′ are intercepts for the subtraining set and calibration set; C1 and C1′ represent slopes for the subtraining set and calibration set, respectively The C0, C0′, C1, and C1′ are changing in process of the Monte Carlo optimization. The **dC-weight** is an empirical parameter: the range for **dC-weight** is **(0.01–0.005)**. Figure 1 shows the logic of these modifications for the case of the balance of correlations and for the case of the balance of correlations *with ideal slopes*.

The method with the ideal slope further improves the process to reduce the risk of overtraining. In this case, the software wants to avoid the situation where the results on the subtraining and validation sets are good, on the basis of the correlation coefficients, but the results on these two sets are unbalanced. As shown in Fig. 1a, we can imagine a situation where the correlation coefficients are good for both the subtraining and validation set, but the equations representing the results for these two sets are quite different. Indeed, this hypothetical situation is not so uncommon in our experience. An usual behavior is that the model "learns" the average values of the property to be modeled, and thus the curve of the training (or subtraining) set is in several cases closer to the correct situation, while the curve of the validation set is more "horizon-

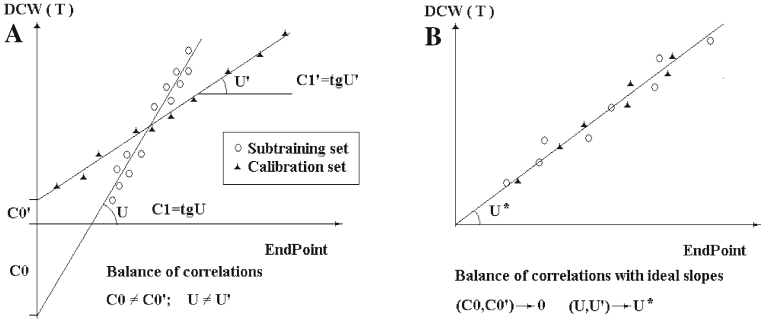


Fig. 1 Balance of correlations can be accompanied by C0 $\neq$ CO$'$ $\neq$ 0 and U $\neq$ U$'$, whereas balance of correlations *with ideal slopes* is an attempt to reach C0 = C0$'$ = 0 and U = U$'$ = U*, where C1 = tgU and C1$'$ = tgU$'$. It is to be noted that the range of U^* is (0, 90°), i.e., the balance of correlations *with ideal slopes* is not aimed to situation where U* = 45°

tal/vertical". This optimization process avoids this situation. CORAL, to optimize the model, uses Monte Carlo method (Fig. 1b).

The Monte Carlo optimization represents a number of epochs of the training with selected target function. The epoch describes the following operations. For each attribute SA (i.e., the s_k, ss_k, and sss_k), CW(SA) is determined initially by setting the start values of all CWs to 1 ± 0.01*random. The random is the generator of random value of range (0, 1). The regular order of number of attributes (i.e., 1, 2, 3, 4, 5,…) is replaced by a random sequence (e.g., 3, 1, 5, 2, 4,…). A starting value of target function (TF1) is calculated. In a generated random sequence, each attribute correlation weight CWi is modified using the following algorithm:

1. DCWi := D^*_{Start}CWi; Eps := $d^*_{Precesion}$DCWi;
2. CWi := CWi + DCWi;
3. Calculation of TF2, after modify CWi;
4. If TF2 > TF1 then TF1 := TF2; go to 2
5. CWi := CWi – DCWi;
6. DCWi := −0.5*DCWi;
7. If absolute value (DCWi) > Eps then go to 2

Then steps 1–7 are carried out for all weights of attributes which are classified as not rare. Correlation weights for rare attributes are zero.

Table 1 contains options selected for modeling by CORAL software the pEC50 for the split 1, the split 2, and the split 3.

3 Results

Table 2 contains the statistical characteristics of the models obtained for three random splits. One can see from the data presented in Table 2 that for all three splits the best statistical quality of the models for validation set is achieved in the case of the balance

Table 1 Selected options in QSAR modeling of the pEC50 for split 1, split 2, and split 3

Options	Split 1	Split 2	Split 3
N_{epoch}	77	77	33
DCW	$\alpha = 1$	$\alpha = 1$	$\alpha = 1$
	$\beta = 1$	$\beta = 1$	$\beta = 1$
	$\gamma = 0$	$\gamma = 0$	$\gamma = 0$
dR-weight	0.1	0.1	0.1
dC-weight	0.005	0.02	0.01
D_{start}	0.1	0.1	0.1
$D_{precesion}$	0.01	0.01	0.01
Start T	0	0	0
Maximum T	5	5	5
Number of probes of optimization	3	3	3

of correlations *with ideal slopes*. Figure 2 displays graphic representation of these results.

There are models which are characterized by similar correlation coefficients and different standard error, e.g., for the split 1, one can see that classic scheme with threshold 3 and balance of correlation with ideal slopes with threshold 2 have similar r, but different s. In this situation, the model that is characterized by smaller standard error and larger $R_m{}^2$ value [17] should be classified as preferable (Table 2).

Supplementary materials section contains an example of the DCW(2) calculations for split 1. This QSAR model can be described as follows:

$$\mathbf{pEC50 = 1.7795(\pm 0.0580) + 0.1044(\pm 0.0015)^{*}DCW(2)} \tag{2}$$

$$\mathrm{n = 27,\ r^2 = 0.9030,\ q^2 = 0.8855,\ s = 0.406,\ F = 233\ (subtraining\ set)}$$

$$\mathrm{n = 15,\ r^2 = 0.9720,\ R^2_{pred} = 0.9661,\ s = 0.980,\ F = 451 (calibration\ set)}$$

$$\mathrm{n = 6,\ r^2 = 0.9224,\ R^2_{pred} = 0.7956,\ s = 0.950,}$$

$$\mathrm{F = 48;\ R^2_m = 0.7812 (validation\ set)}$$

According to P.P. Roy and K. Roy the $R_m{}^2$ can be considered as a measure of predictability of the model [17]. Figure 3 shows that the $R_m{}^2$ correctly indicates the best model for all three random splits. The models obtained for pEC50 for three random splits are displayed at the Fig. 3. These models have been calculated by means of the balance of correlations with ideal slopes.

Table 3 contains experimental and calculated using Eq. 2. pEC50 values. Figure 4 shows this model graphically.

Supplementary materials section contains split 2 and split 3, details of the models for each split (that is shown in Fig. 3), and structures of the fullerene derivatives.

Table 2 Average values of statistical characteristics obtained in three runs of the Monte Carlo optimization for three random split into the subtraining, calibration, and validation sets. In case of the classic scheme training set is the combined subtraining and calibration sets

Threshold	N_{act}	Subtraining set				Calibration set				Validation set					
		n	r^2	s	F	n	r^2	s	F	n	r^2	s	F	$R_m{}^2{}_{av}$	W%
Split 1, classic scheme															
0	285	42	0.9204	0.404	463					6	0.5761	2.989	5	−0.4400	71
1	277	42	0.9206	0.403	464					6	0.5798	2.632	6	−0.3975	70
2	239	42	0.9187	0.408	452					6	0.6763	2.149	9	−0.2050	79
3	215	42	0.9105	0.428	407					6	0.8870	1.005	32	0.2207	86
4	200	42	0.9048	0.442	380					6	0.4437	1.015	3	0.1671	87
5	191	42	0.8991	0.455	357					6	0.6967	0.850	13	0.3598	89
Split 1, Balance of correlations															
0	285	27	0.9132	0.384	263	15	0.9983	1.468	7579	6	0.7108	1.483	10	0.3124	67
1	257	27	0.9135	0.383	264	15	0.9984	1.465	8137	6	0.6278	1.134	7	0.3375	71
2	214	27	0.9080	0.395	247	15	0.9921	1.261	1660	6	0.5227	0.984	5	0.2527	83
3	200	27	0.8961	0.420	216	15	0.9860	1.179	917	6	0.6680	1.711	8	0.5874	86
4	178	27	0.8901	0.432	203	15	0.9897	1.170	1248	6	0.5791	1.695	6	0.3778	90
5	169	27	0.8855	0.441	193	15	0.9889	1.138	1158	6	0.5694	1.875	5	0.4633	92
Split 1, Balance of correlations with Ideal Slopes															
0	285	27	0.9052	0.401	239	15	0.9870	1.100	999	6	0.7646	1.712	13	0.4266	67
1	257	27	0.9052	0.401	239	15	0.9877	1.090	1041	6	0.8626	1.276	29	0.7239	71
2	**214**	**27**	**0.9042**	**0.404**	**236**	**15**	**0.9735**	**0.994**	**479**	**6**	**0.8804**	**0.856**	**40**	**0.7884**	**83**
3	200	27	0.8942	0.424	211	15	0.9778	1.026	572	6	0.6499	1.874	8	0.4195	86
4	178	27	0.8880	0.436	198	15	0.9814	0.997	688	6	0.6982	1.633	9	0.5099	90
5	169	27	0.8822	0.448	187	15	0.9822	0.954	718	6	0.6740	1.712	9	0.5161	92
Split 2, classic scheme															
0	285	40	0.8996	0.429	340					8	0.5413	1.882	7	0.2230	80
1	268	40	0.8989	0.430	338					8	0.4301	2.047	5	0.0342	78
2	239	40	0.8983	0.431	336					8	0.4351	2.033	5	0.0602	85
3	213	40	0.8880	0.453	301					8	0.4317	1.487	6	0.2780	92
4	200	40	0.8890	0.451	304					8	0.4342	1.458	5	0.2078	94
5	187	40	0.8806	0.467	280					8	0.2969	1.562	3	0.1817	96
Split 2, Balance of correlations															
0	285	23	0.8966	0.354	182	17	0.9916	1.367	1783	8	0.4128	1.676	4	0.2894	71
1	250	23	0.8957	0.356	180	17	0.9913	1.367	1708	8	0.4215	1.477	4	0.3114	80
2	206	23	0.8727	0.393	144	17	0.9846	1.239	961	8	0.2588	1.590	2	0.2334	92
3	195	23	0.8720	0.394	143	17	0.9839	1.233	917	8	0.1066	1.590	1	0.0821	93
4	177	23	0.8724	0.394	144	17	0.9847	1.227	965	8	0.0801	1.636	1	0.0594	97
5	168	23	0.8690	0.399	139	17	0.9801	1.187	741	8	0.1603	1.554	1	0.1295	98

Table 2 continued

Threshold	N_{act}	Subtraining set				Calibration set				Validation set					
		n	r^2	s	F	n	r^2	s	F	n	r^2	s	F	$R_m{}^2{}_{av}$	W%
Split 2, Balance of correlations with Ideal Slopes															
0	**285**	**23**	**0.8666**	**0.403**	**136**	**17**	**0.9572**	**0.940**	**335**	**8**	**0.9510**	**0.669**	**117**	**0.6597**	**71**
1	250	23	0.8657	0.404	135	17	0.9562	0.935	328	8	0.9473	0.657	129	0.6348	80
2	206	23	0.8569	0.417	126	17	0.9538	0.977	310	8	0.2777	1.409	2	0.2381	92
3	195	23	0.8615	0.410	131	17	0.9389	0.922	231	8	0.2992	1.304	3	0.2631	93
4	177	23	0.8601	0.412	129	17	0.9392	0.923	234	8	0.3172	1.277	3	0.3012	97
5	168	23	0.8556	0.419	125	17	0.9332	0.934	210	8	0.4894	1.139	6	0.4313	98
Split 3, classic scheme															
0	285	39	0.8833	0.482	280					9	0.2368	1.478	2	0.0875	75
1	278	39	0.8839	0.481	282					9	0.1534	1.601	1	0.0550	75
2	239	39	0.8837	0.481	281					9	0.3031	1.482	3	0.1527	85
3	211	39	0.8775	0.494	265					9	0.3307	1.808	4	0.0310	93
4	196	39	0.8801	0.489	272					9	0.5025	1.559	7	0.1379	94
5	187	39	0.8765	0.496	263					9	0.8012	1.329	29	0.4713	95
Split 3, Balance of correlations															
0	285	22	0.8688	0.517	132	17	0.9943	1.106	2636	9	0.0262	1.522	0	0.0115	73
1	249	22	0.8689	0.517	133	17	0.9946	1.113	2755	9	0.0092	1.456	0	0.0051	80
2	209	22	0.8820	0.490	150	17	0.9844	1.155	945	9	0.4554	1.001	6	0.3726	91
3	183	22	0.8758	0.503	141	17	0.9840	1.151	932	9	0.5291	0.871	8	0.5028	95
4	173	22	0.8652	0.524	128	17	0.9867	1.150	1121	9	0.4912	0.903	7	0.4715	98
5	162	22	0.8665	0.521	130	17	0.9874	1.233	1188	9	0.3794	1.001	4	0.3528	99
Split 3, Balance of correlations with Ideal Slopes															
0	285	22	0.8554	0.542	118	17	0.9824	0.841	839	9	0.3669	1.033	4	0.3293	73
1	249	22	0.8542	0.545	117	17	0.9840	0.847	922	9	0.3007	1.058	3	0.2908	80
2	209	22	0.8455	0.561	110	17	0.9572	0.996	336	9	0.7530	0.643	22	0.6904	91
3	**183**	**22**	**0.8440**	**0.563**	**108**	**17**	**0.9562**	**1.002**	**328**	**9**	**0.7925**	**0.578**	**27**	**0.7640**	**95**
4	173	22	0.8351	0.579	101	17	0.9570	1.018	335	9	0.7828	0.596	25	0.7201	98
5	162	22	0.8400	0.571	105	17	0.9754	1.075	599	9	0.5786	0.823	10	0.5557	99

The n represents the number of compounds in the set, r is correlation coefficient, s is standard error of estimation, F is Fischer F-ratio. ‘s’ is indicator of subtraining set, ‘c’ is indicator of calibration set, and v is indicator of validation set. N_{act} is the number of SMILES attributes which are not blocked. W% is percent of attributes which take place in all sets. $R_m{}^2{}_{av}$ is the measure of predictability according to P.P. Roy and K. Roy [17]: the $R_m{}^2$ should be larger 0.5. Statistical characteristics of the best models are indicated by bold

4 Discussion

The comparison of the 3D approach described in Ref. [15] (CoMSIA: comparative molecular similarity indices analysis) and topological SMILES-based descriptors, calculated with Eq. 1, is able to provide methodological information that is interesting and useful from point of view of the QSAR analysis.

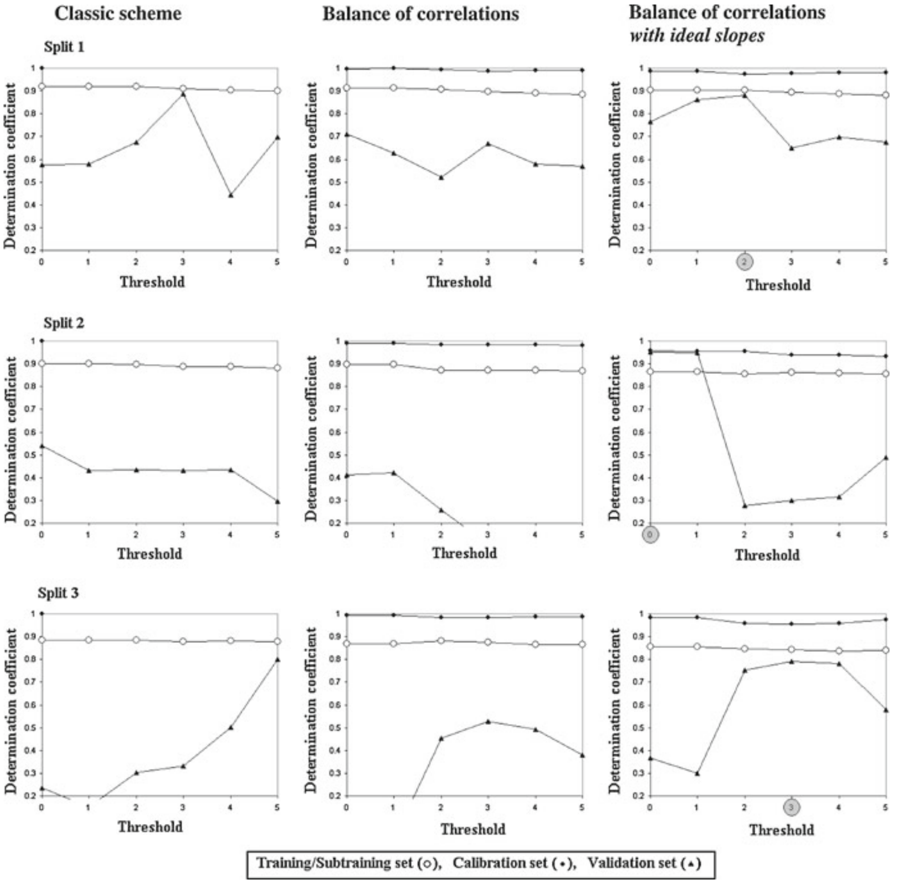


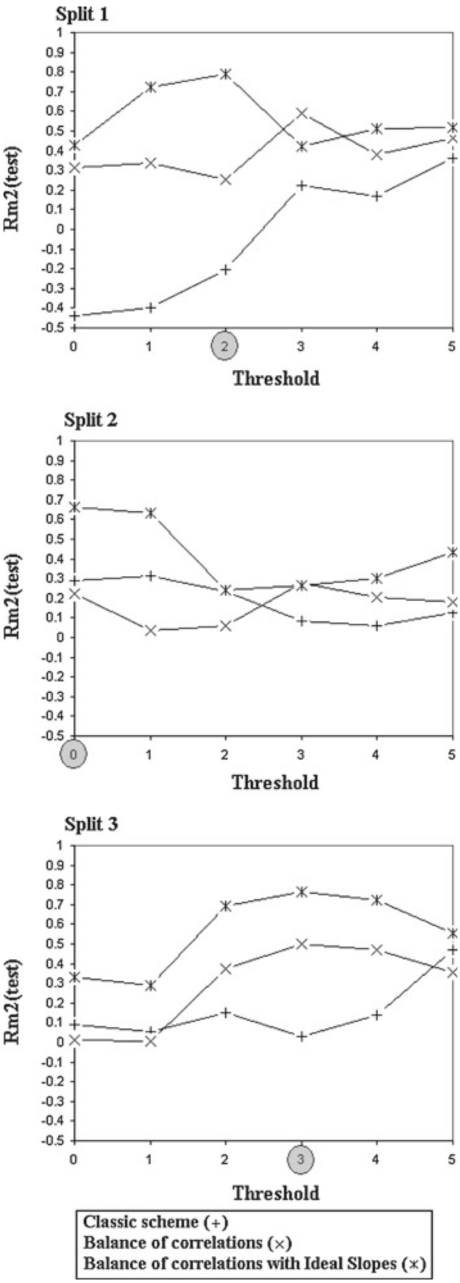
Fig. 2 Correlation coefficicients for sub training, calibration, and validation sets (three random split) obtained using threshold 0,1, …5

In fact, each SMILES attribute is a representation of molecular fragments such as carbon (‘C’ or ‘c’), chlorine atoms (‘Cl’), nitrogen (‘N’ and/or ‘n’) and others [16]. Thus, it is not surprising that the SMILES-based optimal descriptors can be predictors of an endpoint: there is a similarity (but not identity) between SMILES-based descriptors and descriptors calculated with molecular graphs.

We have defined two criteria as the most important for the estimation of a QSAR model: (1). the reproducibility of the model for a group of splitting into training and test, and (2). the statistical quality for the external validation set. The statistical characteristics of the model described in Ref. [15] are $n = 42, r^2 = 0.993, s = 0.127$ (training set). For the test set the corresponding statistical characteristics of the model are the following: $n = 6, r^2 = 0.744, s = 0.755$. Thus the statistics of the above-mentioned model for external test set are considerably worse. This model has been built by 3D/QSAR CoMSIA (comparative molecular similarity indices analysis) [15].

We have no information about statistical characteristics of the 3D/QSAR models described in Ref. [15] for cases of other split. Thus we cannot compare statistical

Fig. 3 Rm^2 values obtained for three random split using the threshold 0,1,…5



characteristics of our models (the split 2 and the split 3) with results of the 3D/QSAR analysis. However, all our models can be considered as reliable and satisfactory according to above-mentioned criterions.

Supplementary materials section contains a list of attributes and their correlation weights. There are attributes which have correlation weights larger than zero for all three probes of the Monte Carlo optimization. These attributes can be classified as promoters of the pEC50 increase. There are also attributes which have correlation weights that are less than zero. Consequently, these attributes are considered to be promoters of the pEC50 decrease. In addition, there are attributes with correlation weights both larger and smaller then zero in different probes of the Monte Carlo optimization. These (together with blocked attributes) do not have any obvious influence on the pEC50 values. The described discrimination of the SMILES attributes can be useful in the search for perspective compounds which can be effective anti-HIV-1 agents.

For instance (the split 1, Threshold = 2), analysis of list of SMILES attributes such as, Cxxx1xxx, Cxxx2xxx, …Cxxx9xxx (i.e., cycles [16]) has shown that Cxxx3xxx, Cxxx6xxx, and Cxxx8xxx are stable promoters of pEC50 increase because their correlation weights in three probes of the Monte Carlo optimization are equal to (2.38, 3.16, 2.81), (3.13, 2.68, 3.15), and (3.03, 3.86, 2.72), respectively. Vice versa, Nxxxxxxx, Nxxx = xxxxxxx, NxxxCxxxxxxx are stable promoters of pEC50 decrease, since their correlation weights in three probes of the Monte Carlo optimization are equal to (−2.07, −1.93, −1.80), (−4.88, −4.33, −4.23), and (−4.32, −4.06, −3.92), respectively. In fact, it means, that in search for a anti-HIV-1 agents (at least in the first approximation) presence of above-mentioned cycles fragments (C3, C6, and C8) is preferable, whereas presence of the above-mentioned nitrogen-containing fragments (N, N=, and NC) is objectionable.

Table 3 contains data on the number of blocked (Blk) attributes together with the total number of the attributes (All) for each SMILES. This information can be useful for adequate estimation of a split: apparently, a split is not satisfactory if percent of blocked attributes for a group of SMILES will be too large. For the given case (the split 1, Threshold = 2) maximum of blocked attributes takes place for the substance #3 (Table 3). However the Blk/All = 28/377 $<$ 8%, i.e., the percent is not too large.

5 Conclusions

The CORAL approach provides reliable models for pEC50 (minus decimal logarithm of the 50% effective concentration) of fullerene derivatives, because statistical quality of these models is satisfactory for all examined distributions (splits) into the subtraining set, calibration set, and external validation set. Balance of correlations with ideal slopes gives the best prediction for the validation set in comparison with classic scheme and the balance of correlations that is carried out without taking into account slopes on the plots of the experimental versus the calculated pEC50 values for the subtraining set and the calibration set. For the examined QSAR-models the reliability of the Rm2 criterion is confirmed. Thus, SMILES-based optimal descriptors calculated by CORAL can be considered as useful addition for 3D approaches examined in Ref. [15].

Table 3 Experimental and calculated using Eq. 2 the pEC50 values. Blk represents the number of blocked attributes in a given SMILES, All is the total number of attributes in the SMILES

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
	Subtraining set				
1	O[C@@H](c1ccccc1)C%28(c2ccccc2)C%33%29C%14C%16=C%32C%19=C%31C=9c%18c3C%34=C7c4c3c%17c%15c%23c4C=%22C=8C6=C%21c%25c5c%13C=%12C%11=C5C6=C(C7=8)C%10=C%34C=9C(=C%10%11)C=%30C=%12C(C%27c%13c%26C(C=%20C%14=C(C%15=C%16c%17c%18%19)C=%24C=%20C(=C%21C=%22C%23=%24)c%25%26)C%27%28%29)C(C=%30 %31)=C%32%33	48.6196919	7.000	6.855	8/381
2	NCCc1ccc(cc1)C%28(c2ccc(CCN)cc2)C%33%29C%14C%16=C%32C%19=C%31C=9c%18c3C%34=C7c4c3c%17c%15c%23c4C=%22C=8C6=C%21c%25c5c%13C=%12C%11=C5C6=C(C7=8)C%10=C%34C=9C(=C%10%11)C=%30C=%12C(C%27c%13c%26C(C=%20C%14=C(C%15=C%16c%17c%18%19)C=%24C=%20C(=C%21C=%22C%23=%24)c%25%26)C%27%28%29)C(C=%30%31)=C%32%33	36.3817491	5.300	5.578	0/385
5	O=C(O)CCCCCNC%31%26C%29C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%30c7c6c%28C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%27C(=C%19C=%20%21)C%22=C(C=%27C(C=%23%24)C%25%26)C%31c%28c%29%30	28.2654879	6.310	4.730	0/339

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
7	O=C(O)C(N)CCCCNC%31%26C%29C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%30c7c6c%28C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%27C(=C%19C=%20%21)C%22=C(C=%27C(C=%23%24)C%25%26)C%31c%28c%29%30	28.1645077	4.120	4.720	0/345
8	O=C(O)C(N)CCCNC(=N)NC%31%26C%29C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%30c7c6c%28C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%27C(=C%19C=%20%21)C%22=C(C=%27C(C=%23%24)C%25%26)C%31c%28c%29%30	18.9217409	3.640	3.755	0/355
10	OCC(CO)(CO)NC(=O)CCCc1ccc(cc1)C%28(c2ccccc2)C%33%29C%14C%16=C%32C%19=C%31C=9c%18c3C%34=C7c4c3c%17c%15c%23c4C=%22C=8C6=C%21c%25c5c%13C=%12C%11=C5C6=C(C7=8)C%10=C%34C=9C(=C%10%11)C=%30C=%12C(C%27c%13c%26C(C=%20C%14=C(C%15=C%16c%17c%18%19)C=%24C=%20C(=C%21C=%22C%23=%24)c%25%26)C%27%28%29)C(C=%30%31)= C%32%33	36.8916735	5.600	5.631	1/409

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
11	O=C(O)CO\N=C%30\CC %25%31C%12C=%14C=%15C%29C%17=C%28C%20=C%27C=7c%19c1C%32=C5c2c1c%18c%16c%21c2C=%22C=6C4=C%23c%13c3c%11C=%10C9=C3C4=C(C5=6)C8=C%32C=7C(=C89)C=%26C=%10C(C%25c%11c%12c%13C=%24C=%14C(C=%15C%16=C%17c%18c%19%20)=C%21C=%22C%23=%24)C(C=%26%27)=C%28C%29%31C(OC)C%30	40.9274197	6.050	6.052	7/361
13	O=C(O)CCC(=O)OCC%26C=CC(C)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C%32c%29c%30%31	38.2065472	5.660	5.768	0/365
14	O=C(O)CCC(=O)OC%29CC%25%31C%10C=%16C=%15C%30C=%28C=%14c%32c%13c2c%12C%22=C1C=5C4=C3c(c12)c%32C%27=C3C%26=C8C4=C7C=5C%23=C%11C=6c%24c9C(C=67)=C8C(C%25c9c%10c%21c%24C%17=C%11C(C=%18c%12c%19c%13C=%14C=%15C=%20C=%16C%21=C%17C=%18C%19=%20)=C%22%23)C%26=C(C%27=%28)C%30%31CC%29	32.9780679	5.200	5.222	3/357

Table 3 continued

No.	SMILES	DCW(2)	pEC50$_{\mathrm{Expr}}$	pEC50$_{\mathrm{Calc}}$	Blk/All
17	C1COCCOCCOCCOCCOc2cc(ccc2O1)C%35(c3ccccc3)C%28%34C%19=C%32C=6C=%18C=5c%17c4c%16c%15C%13=C4C=%11C=5C=%10C=6C%31=C9c%30c8c7c%29C%25=C%24C7=C%14C=%12C8=C9C=%10C=%11C=%12C%13=C%14C=%23c%15c%22c%21c%16c%20c%17C=%18C%19=C%27C%20=C%21C=%26C(=C%22C=%23%24)C%25=C(C=%26C%27%28)C%33c%29c%30C(C%31%32)C%33%34%35	20.3325419	3.860	3.902	5/391
19	CN%26CC%22%28C%19C%30c6c%27c%25c1c5C=4C%32=C1C%23=C%15C%33=C%14C%31=C3c2c%13c%12c%11c%10c2C9=C3C7=C(C=4C8C(c56)C%30%29CN(C)CC%21%29C(=C78)C9=C%20C%10=C%18c%11c%17c%16c%12C(=C%13%14)C%15=C%24C%16=C(C%17=C%22C%18=C%19C%20%21)C(C%25=C%23%24)C%27%28C%26)C%31=C%32%33	22.9615307	4.140	4.177	6/331
21	O=C(O)C%26=C(O)C(O)=C(C(=O)O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C %32c%29c%30%31	50.9482228	6.600	7.098	0/371

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
23	O=C(O)C%26=C(C(=O)O)C(C(=O)O)=C(C(=O)O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C%32c%29c%30%31	58.3802098	8.700	7.874	0/391
24	OC%26=C(O)C(O)=C(O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C%32c%29c%30%31	43.4494393	6.150	6.316	0/351
26	O=C(O)C%30N(N)C(C(=O)O)C%23%31C%16C=%10C=9C%29C=%32C=8c%27c7c2c5c%26C%25=C%20C6=C%19C3=C1C=%18c%17c%15C%11=C1C4C=%12c(c2C(C34)C56)c%13c7C=8C=9C=%14C=%10C(=C%11C=%12C%13=%14)c%15c%16c%22c%17C=%21C=%18C%19=C%20C=%24C=%21C(C%22%23)C=%28C=%24C%25=C(c%26%27)C=%32C=%28C%29%30%31	43.9249931	6.200	6.365	0/361

Table 3 continued

No.	SMILES	DCW(2)	pEC50$_{Expr}$	pEC50$_{Calc}$	Blk/All
28	NC(=O)C%30N(N)C(C(N)=O)C%23%31C%16C=%10C=9C%29C=%32C=8c%27c7c2c5c%26C%25=C%20C6=C%19C3=C1C=%18c%17c%15C%11=C1C4C=%12c(c2C(C34)C56)c%13c7C=8C=9C=%14C=%10C(=C%11C=%12C%13=%14)c%15c%16c%22c%17C=%21C=%18C%19=C%20C=%24C=%21C(C%22%23)C=%28C=%24C%25=C(c%26%27)C=%32C=%28C%29%30%31	40.4804279	6.180	6.006	0/361
30	OC%30=C(O)C(O)=C(O)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C(O)=C(O)C(O)=C(O)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25) C%30%33C%31%32	28.1973269	4.700	4.723	0/381
31	NC%30=C(N)C(N)=C(N)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C(N)=C(N)C(N)=C(N)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25) C%30%33C%31%32	26.2739426	4.770	4.522	0/381

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
33	NC(=O)CC%26=C(O)C(O)=C(CC(N)=O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C%32c %29c%30%31	45.5067706	6.550	6.530	0/375
35	NC(=O)C%32=C(N)C(N1C=CC=CC1)=C(C(N)=O)C%31%35C3=C%30C=2c%29c%10C=9C=2C=%26C3=C%34C%27=C8c7c%33c6c(c%28C=5C%31C%30=C4c%29c%11C%12=C4C=5C=%13c%28c%14c6c%15c7C%16=C8C%17=C(C=9C%18C%20c%10c%11C%19C%24=C%12C=%13C=%23C%14=C%15C%25=C%16C(=C%17%18)C%21(C(C(N)=O)=C(C(N)=C(C(N)=O)C%19%20%21)N%22C=CC=CC%22)C%25C=%23%24)C=%26%27)C%32%35C%33%34	24.9896618	4.230	4.388	3/457
37	FC%30=NC%33%32C=4C2=C1C%31c5c9C%10=C1C%11=C2C%28=C3c%12c%23C%25=C(C3=4)C%33C=%26c6c(c5c8c7c6C%27=C%29C%24=C%13C%20=C%29C7=C%19c8c9C%18=C%10C%14=C%11C%15=C%28c%12c%16c(C%13C%17%22N=C(F)C(F)=NC%21%22C(=C%14C%15C%16%17)C%18=C%19C%20%21)c%23C%24=C%25C=%26%27)C%31%32N=C%30F	15.5514702	3.360	3.403	4/345

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
38	OC%27=C(O)C(c1ccccc1)=C(O)C%33%28C%31C6C4=C%26C=%24C%18=C5C=3c%17c2c%16c%15C%13=C2C=%11C=3C=%10C(=C45)C6=C9c%32c8c7c%30C%23=C%22C7=C%14C=%12C8=C9C=%10C=%11C=%12C%13=C%14C=%21c%15c%20c%19c%16c(c%17%18)C%25=C%19C=%29C(=C%20C=%21%22)C%23=C(C=%29C(C=%24%25)C%26%27%28)C%33c%30c%31%32	39.2305970	5.730	5.875	0/365
41	O=C(O)C%27=CC(Cc1c(O)c(O)c(O)c(O)c1O)=C(C(=O)O)C%33%28C%31C6C4=C%26C=%24C%18=C5C=3c%17c2c%16c%15C%13=C2C=%11C=3C=%10C(=C45)C6=C9c%32c8c7c%30C%23=C%22C7=C%14C=%12C8=C9C=%10C=%11C=%12C%13=C%14C=%21c%15c%20c%19c%16c(c%17%18)C%25=C%19C=%29C(=C%20C=%21%22)C%23=C(C=%29C(C=%24%25)C%26%27 %28)C%33c%30c%31%32	53.5381283	7.400	7.369	0/407
42	NC(=O)C%30=C(C(N)=O)C(C(N)=O)=C(C(N)=O)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C(C(N)=O)=C(C(N)=O)C(C(N)=O)=C(C(N)=O)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25)C%30%33C%31%32	56.0018906	7.400	7.626	0/461

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
45	c%31cccc%32CC%30%35C2=C%29C=1c%28c8C=%23C=1C=%25C2=C%34C%26=C7c6c%33c5c(c%27C=4C%30C%29=C3c%28c9C%10=C3C=4C=%11c%27c%12c5c%13c6C%22=C7C=%24C=%14C(C%15c8c9C%16C%21=C%10C=%11C=%20C%12=C%13C%19=C%22C=%14C%17(Cc%18ccccc%18CC%15%16%17)C%19C=%20%21)C=%23C=%24C=%25%26)C%35(Cc%31%32)C%33%34	22.8285809	4.130	4.163	1/357
47	OC%27C(O)C(O)C(O)C%32%29C%22C%16=C%31C%15=C3C=2c%14c1c%13c%12C%10=C1C=8C=2C=7C%30=C3C(C%26C%30=C6c%25c5c4c%24C%21=C%20C4=C%11C=9C5=C6C=7C=8C=9C%10=C%11C=%19c%12c%18c%17c%13c(c%14%15)C%16=C%17C=%23C(=C%18C=%19%20)C%21=C(C%22=%23)C%28c%24c%25C%26C%27%28%29)=C%31%32	41.8325497	6.150	6.147	3/341
49	NC(=O)CC%26=C(F)C(F)=C(CC(N)=O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C%32c%29c%30%31	44.5918982	6.080	6.435	0/375

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
	Calibration set				
3	CC(C)[C@@H]%30C[C@H](O)[C@H](C(C)C)C%23%32C%10C=%14C%29=C%19C=%28c%13c%26c2c%12C%20=C1C=5C4=C3c(c12)c%27C%25=C3C%24=C8C4=C7C=5C%21=C%11C=6c%22c9C(C=67)=C8C(C%23c9c%10c%18c%22C%15=C%11C(C=%16c%12c%13C%17=C%19C=%14C%18=C%15C=%16%17)=C%20%21)C%24=C%31C%25=C(C=%28c%26%27)C%29C%30%31%32	53.5599041	6.820	7.371	28/377
4	O=C(O)CCC(=O)NCCc1ccc(cc1)C%28(c2ccc(CCNC(=O)CCC(=O)O)cc2)C%33%29C%14C%16=C%32C%19=C%31C=9c%18c3C%34=C7c4c3c%17c%15c%23c4C=%22C=8C6=C%21c%25c5c%13C=%12C%11=C5C6=C(C7=8)C%10=C%34C=9C(=C%10%11)C=%30C=%12C(C%27c%13c%26C(C=%20C%14=C(C%15=C%16c%17c%18%19)C=%24C=%20C(=C%21C=%22C%23=%24)c%25%26)C%27%28%29)C(C=%30%31)= C%32%33	39.0173849	5.140	5.853	0/437
6	O=C(O)CCC(CCCC\C=C\CCCCCCCC)NC%31%26C%29C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%30c7c6c%28C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%27C(=C%19C=%20%21)C%22=C(C=%27C(C=%23%24)C%25%26)C%31c%28c%29%30	22.0381584	2.890	4.080	4/373

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
9	O=C(O)CC(CC(=O)O)NC(=O)NC%31%26C%29C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%30c7c6c%28C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%27C(=C%19C=%20%21)C%22=C(C=%27C(C=%23%24)C%25%26)C%31c%28c%29%30	36.5651131	3.850	5.597	0/363
12	O=C(O)CO\N=C%30/CCC%26%32C%18=C5C4C%31c%29c3c2c1c%28C%24=C%23C1=C%13C=%11C2=C9C3=C4C8=C5C=%17C=7c%16c6c%15c%14C%12=C6C=%10C=7C8=C9C=%10C=%11C%12=C%13C=%22c%14c%21c%20c%15c%19c%16C=%17C%18=C%25C%19=C%20C=%27C(=C%21C=%22%23)C%24=C(C=%27C%25%26)C(c%28%29)C%31%32C%30	41.4706949	5.140	6.109	6/341
15	O=C(O)CO\N=C%30/C=CC%26%32C%18=C5C4C%31c%29c3c2c1c%28C%24=C%23C1=C%13C=%11C2=C9C3=C4C8=C5C=%17C=7c%16c6c%15c%14C%12=C6C=%10C=7C8=C9C=%10C=%11C%12=C%13C=%22c%14c%21c%20c%15c%19c%16C=%17C%18=C%25C%19=C%20C=%27C(=C%21C=%22%23)C%24=C(C=%27C%25%26)C(c%28%29)C%31%32C%30	41.8387025	5.540	6.147	6/343

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
18	OC%16CC%22%18c%30c%21c1c5C=4C%32=C1C%19=C%11C%33=C%10C%31=C3c2c9c8c7c6c2C%23=C3C%24=C(C=4C%25C%27c5c%30C%26C%17C=%15C%29C(C6=C%14c7c%13c%12c8C(=C9%10)C%11=C%20C%12=C(C%13=C(C%14=%15)C%17%18CC%16)C%22C%21=C%19%20)=C%23C(=C%24%25)C%29%28CCC(O)CC%26%27%28)C%31=C%32%33	37.0073258	4.750	5.643	6/337
22	O/C(O)=C(/O)C=%31C(O)=C(O)C%26%32C%18=C5C4C%30c%29c3c2c1c%28C%24=C%23C1=C%13C=%11C2=C9C3=C4C8=C5C=%17C=7c%16c6c%15c%14C%12=C6C=%10C=7C8=C9C=%10C=%11C%12=C%13C=%22c%14c%21c%20c%15c%19c%16C=%17C%18=C%25C%19=C%20C=%27C(=C%21C=%22%23)C%24=C(C=%27C%25%26)C(c%28%29)C%30 %32C=%31O	57.5568603	7.290	7.788	9/361
25	C%30=CC=CC%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C=CC=CC%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25)C%30%33C%31%32	27.7742839	3.330	4.679	0/337

Table 3 continued

No.	SMILES	DCW(2)	pEC50$_{\mathrm{Expr}}$	pEC50$_{\mathrm{Calc}}$	Blk/All
29	O=C(O)C%26=C(F)C(F)=C(C(=O)O)C%32%27C%30C5C3=C%25C=%23C%17=C4C=2c%16c1c%15c%14C%12=C1C=%10C=2C=9C(=C34)C5=C8c%31c7c6c%29C%22=C%21C6=C%13C=%11C7=C8C=9C=%10C=%11C%12=C%13C=%20c%14c%19c%18c%15c(c%16%17)C%24=C%18C=%28C(=C%19C=%20%21)C%22=C(C=%28C(C=%23%24)C%25%26%27)C %32c%29c%30%31	50.0333504	6.680	7.003	0/371
32	NC(=O)C%30=C(N)C(N)=C(C(N)=O)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C(C(N)=O)=C(N)C(N)=C(C(N)=O)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24 %25)C%30%33C%31%32	41.1379166	5.500	6.074	0/421
36	O=C(OC)C%32=C(N)C(N1C=CC=CC1)=C(C(=O)OC)C%31%35C3=C%30C=2c%29c%10C=9C=2C=%26C3=C%34C%27=C8c7c%33c6c(c%28C=5C%31C%30=C4c%29c%11C%12=C4C=5C=%13c%28c%14c6c%15c7C%16=C8C%17=C(C=9C%18C%20c%10c%11C%19C%24=C%12C=%13C=%23C%14=C%15C%25=C%16C(=C%17%18)C%21(C(C(=O)OC)=C(C(N)=C(C(=O)OC)C%19%20%21)N%22C=CC=CC%22)C%25C=%23%24)C=%26%27)C%32%35C%33%34	16.2198846	2.250	3.473	3/465

Table 3 continued

No.	SMILES	DCW(2)	pEC50$_{\mathrm{Expr}}$	pEC50$_{\mathrm{Calc}}$	Blk/All
40	C%30CCCC%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(CCCCC%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25)C%30%33C%31%32	23.3845009	3.060	4.221	0/329
44	OC%27=C(O)C(Cc1ccccc1)=C(O)C%33%28C%31C6C4=C%26C=%24C%18=C5C=3c%17c2c%16c%15C%13=C2C=%11C=3C=%10C(=C45)C6=C9c%32c8c7c%30C%23=C%22C7=C%14C=%12C8=C9C=%10C=%11C=%12C%13=C%14C=%21c%15c%20c%19c%16c(c%17%18)C%25=C%19C=%29C(=C%20C=%21%22)C%23=C(C=%29C(C=%24%25)C%26%27%28)C%33c%30c%31%32	42.9871367	5.610	6.267	0/367
46	O=C(O)C%29=C(C(=O)O)C(C3C1CC2CC(C1)CC3C2)=C(C(=O)O)C%35%30C%33C8C6=C%28C=%26C%20=C7C=5c%19c4c%18c%17C%15=C4C=%13C=5C=%12C(=C67)C8=C%11c%34c%10c9c%32C%25=C%24C9=C%16C=%14C%10=C%11C=%12C=%13C=%14C%15=C%16C=%23c%17c%22c%21c%18c(c%19%20)C%27=C%21C=%31C(=C%22C=%23%24)C%25=C(C=%31C(C=%26%27)C%28%29%30)C%35c%32c%33%34	55.1422281	7.000	7.536	7/415

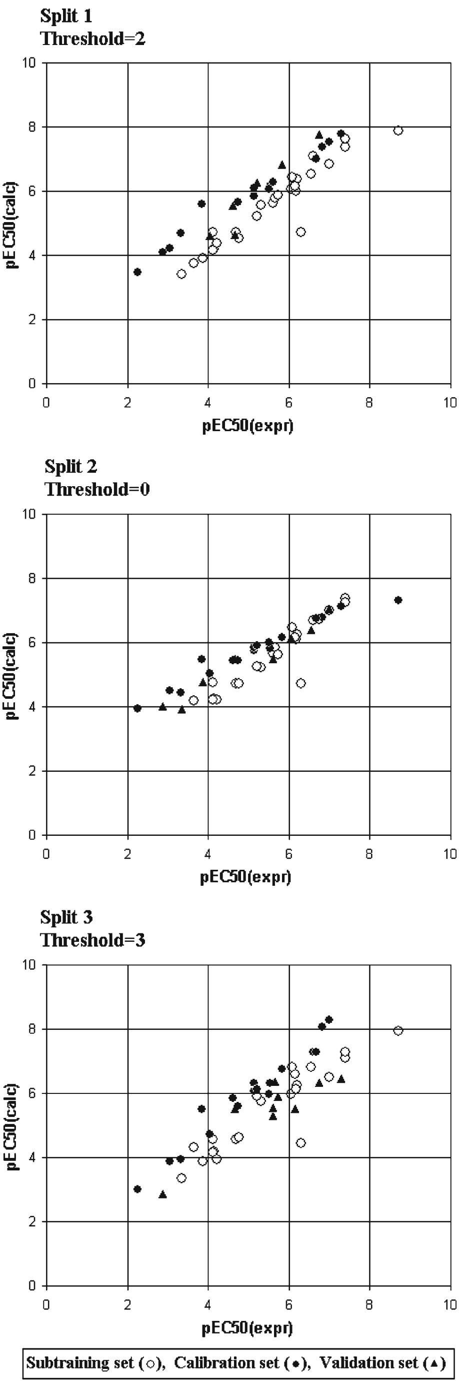
Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
	Validation set				
16	O=C(O)CO\N=C1/CCCC%29(C)C1C%20%30C%19C%28=C%27C=%33c%18c3c2c%17C=%15C%10=C2C=%31C4=C3C=%33C%26=C6C4=C%32C5=C9c8c7c5c6c%25c%21c7c%22c%13c8C%12=C%11C9=C(C%10=C%11C%16=C%14C%12=C%13C=%24C(=C%14C(C=%15%16)C%20c%17c%18%19)C%29%30C%23C%28=C(C%21=C%22C%23=%24)C %27=C%25%26)C=%31%32	27.6888565	4.660	4.670	9/355
27	F/C(F)=C(/F)C=3C(F)=C(C(=O)O)C%13%14c%15c%16c%12c%17C=%10C=%18C=8C=%19C7C%20c%31c6c%33C=5C=4C2=C1c%33c%30C%21=C1C%23=C(C2C%14(C=3C(=O)O)C=4C=%11C=9C=5C6=C7C=8C=9C=%10C=%11C%12%13)c%15c%22c%24c%16c%25c%17C=%18C=%26C=%19C%29%32C(C(=O)O)=C(C(\F)= C(/F)F)C(F)=C(C(=O)O)C%20%32C(C%28=C%21C(=C%22%23)C=%27C%24=C%25C=%26C%29C=%27%28)c%30%31	48.5912675	5.820	6.852	16/471
34	FC%30=C(F)C(F)=C(F)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C %17(C(F)=C(F)C(F)=C(F)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25)C%30%33C%31%32	27.3309894	4.040	4.633	1/381

Table 3 continued

No.	SMILES	DCW(2)	$pEC50_{Expr}$	$pEC50_{Calc}$	Blk/All
39	NC%27=C(N)C(c1ccccc1)=C(N)C%33%28C%31C6C4=C%26C=%24C%18=C5C=3c%17c2c%16c%15C%13=C2C=%11C=3C=%10C(=C45)C6=C9c%32c8c7c%30C%23=C%22C7=C%14C=%12C8=C9C=%10C=%11C=%12C%13=C%14C=%21c%15c%20c%19c%16c(c%17%18)C%25=C%19C=%29C(=C%20C=%21%22)C%23=C(C=%29C(C=%24%25)C%26%27%28)C%33c%30c%31%32	36.0952257	4.610	5.548	0/365
43	O/C(O)=C(/O)\C (\O)=C(/O)C=%31C(O)=C(O)C%26%32C%18=C5C4C%30c%29c3c2c1c%28C%24=C%23C1=C%13C=%11C2=C9C3=C4C8=C5C=%17C=7c%16c6c%15c%14C%12=C6C=%10C=7C8=C9C=%10C=%11C%12=C%13C=%22c%14c%21c%20c%15c%19c%16C=%17C%18=C%25C%19=C%20C=%27C(=C%21C=%22%23)C%24=C(C=%27C%25%26)C(c%28%29)C%30%32C=%31	57.4999434	6.740	7.782	15/383
48	O=C(O)C%30=C(O)C(O)=C(C(=O)O)C%29%33C2=C%28C=1c%27c8C=%22C=1C=%24C2=C%32C%25=C7c6c%31c5c(c%26C=4C%29C%28=C3c%27c9C%10=C3C=4C=%11c%26c%12c5c%13c6C%21=C7C=%23C=%14C(C%15c8c9C%16C%20=C%10C=%11C=%19C%12=C%13C%18=C%21C=%14C%17(C(C(=O)O)=C(O)C(O)=C(C(=O)O)C%15%16%17)C%18C=%19%20)C=%22C=%23C=%24%25)C%30%33C%31%32	43.1280974	5.220	6.282	0/421

Fig. 4 Graphical representation of best models (best statistics for validation set) for three random split



Acknowledgments The authors are grateful to the Marie Curie Fellowship for financial support (contract 39036, CHEMPREDICT). This work was in part supported by NSF through the RISE grant # HRD-0401730.

References

1. M. Randic, S.C. Basak, J. Chem. Inf. Comput. Sci. **41**, 650–656 (2001)
2. P.R. Duchowicz, E.A. Castro, Int. J. Mol. Sci. **10**, 2558–2577 (2009)
3. A.A. Toropov, B.F. Rasulev, J. Leszczynski, QSAR Comb.Sci. **26**, 686–693 (2007)
4. B.F. Rasulev, A.A. Toropov, A.T. Hamme II, J. Leszczynski, QSAR Comb. Sci. **27**, 595–6065 (2008)
5. CHEMPREDICT at: http://www.insilico.eu/coral (2010)
6. A.A. Toropov, E. Benfenati, Eur. J. Med. Chem. **42**, 606–613 (2007)
7. K. Roy, A.A. Toropov, I. Raska Jr, QSAR Comb. Sci. **26**, 460–468 (2007)
8. A.A. Toropov, D. Leszczynska, J. Leszczynski, Comput. Biol. Chem. **31**, 127–128 (2007)
9. A.A. Toropov, E. Benfenati, Comput. Biol. Chem. **31**, 57–60 (2007)
10. NIST Chemistry WebBook at: http://webbook.nist.gov/chemistry/
11. US National Laboratory of Medicine at: http://toxnet.nlm.nih.gov/
12. A.A. Toropov, E. Benfenati, Curr. Drug Disc. Tech. **4**, 77–116 (2007)
13. T. Puzyn, D. Leszczynska, J. Leszczynski, Small **5**, 2494–2509 (2009)
14. S. Durdagi, T. Mavromoustakos, M.G. Papadopoulos, Bioorg. Med. Chem. Lett. **18**, 6283–6289 (2008)
15. S. Durdagi, T. Mavromoustakos, N. Chronakis, M.G. Papadopoulos, Bioorg. Med. Chem. **16**, 9957–9974 (2008)
16. ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada (2007) http://www.acdlabs.com
17. P.P. Roy, K. Roy, Chem. Biol. Drug Des. **73**, 442–455 (2009)