

# Development of QSAR models for predicting anti-HIV-1 activity using the Monte Carlo method

Research Article

Andrey A. Toropov<sup>1\*</sup>, Alla P. Toropova<sup>1</sup>,  
Ivan Raska Jr.<sup>2</sup>, Emilio Benfenati<sup>1</sup>, Giuseppina Gini<sup>3</sup>

<sup>1</sup>Institute of Pharmacological Researches "Mario Negri",  
20156 Milan, Italy

<sup>2</sup>3rd Department of Medicine - Department of Endocrinology and Metabolism,  
First Faculty of Medicine, Charles University in Prague and General University  
Hospital in Prague, 12808 Prague 2, Czech Republic

<sup>3</sup>Department of Electronics and Information,  
Polytechnical University of Milan, 20133 Milan, Italy

Received 6 July 2012; Accepted 22 October 2012

**Abstract:** The CORAL software (<http://www.insilico.eu/coral/>) has been examined as a tool for modeling anti-HIV-1 activity by quantitative structure – activity relationships (QSAR) for three different sets: (i) TIBO derivatives (n=82) (ii) anti-HIV-1 activity of 2-amino-6-arylsulfonylbenzotriazoles and their congeners (n=64), and (iii) the measured binding affinity for fullerene-based HIV-1 PR inhibitors (n=48). A new global invariant ATOMPAIR of the molecular structure which can be calculated with the simplified molecular input line entry system (SMILES) was studied. The ATOMPAIR is an indicator of the joint presence of pairs of chemical elements (F, Cl, Br, N, O, S, and P) and three types of bonds (double covalent bond, triple covalent bond, and stereo chemical bond). Six random splits into sub-training, calibration, and test set were examined for each set. For the three aforementioned sets, the use of ATOMPAIR in the modeling process improves the predictive potential of the models for six random splits.

**Keywords:** QSAR; anti-HIV-1 activity • SMILES • Balance of correlations • CORAL software

© Versita Sp. z o.o.

## 1. Introduction

The treatment of Human Immunodeficiency Virus (HIV) infection is a well-known important problem [1-17]. The tetrahydroimidazo[4,5,1-jk][1,4]benzodiazepinone (TIBO) derivatives, as non-nucleoside reverse transcriptase inhibitors, have a significant role in the treatment of HIV infection [5]. A series of 2-amino-6-arylsulfonylbenzotriazoles are also known as effective anti-HIV-1 agents [6]. Finally, the molecular structures together with the anti-HIV-1 activity of a series of effective fullerene based inhibitors have been described [17].

Quantitative structure-property/activity relationships (QSPRs/QSARs) approaches have become very useful and largely widespread for the prediction of various

endpoints, in general [18-21] and for prediction of anti-HIV-1 activity, in particular [7-17].

Recently, the CORAL software has been suggested as a tool for QSPR/QSAR analysis [22]. A few models calculated with CORAL have been reported in the literature [23-26]. The representation of the molecular structure by the simplified molecular input line entry system (SMILES) is used to build the CORAL models [27-29].

The validation of QSPR/QSAR has been discussed many times [30-43]. Many authors agree that external validation is necessary, but typically only one split into the training and test sets is used for demonstration and validation of an approach. The necessity of examination of multiple splits into the training and test sets is discussed by Tropsha [44].

\* E-mail: [andrey.toropov@marionegri.it](mailto:andrey.toropov@marionegri.it)

Our studies show that using a series of splits can give important information about a QSPR/QSAR model, because there are 'successful' and 'unsuccessful' splits into the training and test sets [45-47]. Thus, the statistical characteristics of a QSPR/QSAR are a mathematical function of the split into the training and test sets.

The above-mentioned CORAL software is a realization of the QSPR/QSAR approach based on the correlation weights of molecular fragments. The numerical data for the correlation weights are calculated by the Monte Carlo method. We deem that a robust model for many endpoints can be calculated by CORAL. However, for this aim, it is necessary to obey two principles. Principle 1: one should use as a criterion of robustness of a model the statistical quality for the external test set; Principle 2: for robust estimation of an approach, one should use a large number of splits into the training and test sets (as large as possible).

The aim of the present study is the estimation of the reliability of models for (i) anti-HIV activity of TIBO derivatives; (ii) anti-HIV-1 activity of 2-amino-6-arylsulfonylbenzonnitriles, and (iii) fullerene-based HIV-1 inhibitors, which are calculated using the CORAL software. The stability of the statistical quality of the CORAL models for a series of splits into the training and test sets has been used as a criterion of the reliability of the approach. The comparison of the classic scheme (training - test) and the balance of correlations (sub-training - calibration - test) as two different approaches to building a model for anti-HIV-1 activity is an additional aim of the study. Finally, the ability to take into account the physicochemical situations in molecules (presence/absence of different chemical elements) to improve the predictive potential of this approach has been studied.

## 2. Experimental procedure

### 2.1. Data

Three sets of data were studied: (i) SET 1. The anti-HIV activity of the TIBO compounds has been expressed as the compound's ability to protect MT-4 cells against the cytopathic effect of the virus. The concentration of the compound leading to 50% effect has been measured and expressed as  $pIC_{50}$ . The decimal logarithm of the inverse of this parameter has been used as the biological endpoint (*i.e.*,  $-\log IC_{50} = pIC_{50}$ ) in the QSAR studies [5]; (ii) SET 2. Data on negative decimal logarithm of  $IC_{50}$  (50% inhibitory concentration, in  $\text{mol L}^{-1}$ ) of the 2-amino-6-arylsulfonylbenzonnitriles ( $-\log IC_{50} = pIC_{50}$ ) [6]; and (iii) SET 3. Median effective concentration ( $EC_{50}$ ), expressed as the negative

logarithm ( $pEC_{50}$ ) for fullerene  $C_{60}$  derivatives [17]. For each set (*i.e.*, for the above-mentioned SET 1, SET 2, and SET 3), six random splits into the sub-training, calibration, and test sets were studied. The canonical version of the SMILES notation (ACD/ChemSketch Freeware, v. 11.00, Inc., Toronto, Canada, www.acdlabs.com, 2007) has been used in this study.

### 2.2. Descriptors

The CORAL model is a one-variable model of an endpoint  $Y$ , calculated as

$$Y = C_0 + C_1 \text{DCW}(\text{Threshold, Nepoch}) \quad (1)$$

where  $\text{DCW}(\text{Threshold})$  is the optimal SMILES-based descriptor;  $C_0$  and  $C_1$  are regression coefficients.

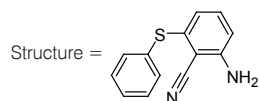
The  $\text{DCW}(\text{Threshold, Nepoch})$  is calculated as follows:

$$\begin{aligned} \text{DCW}(\text{Threshold, Nepoch}) = & w * \text{CW}(\text{ATOMPAIR}) + \\ & + x * \text{CW}(\text{BOND}) + y * \text{CW}(\text{NOSP}) + z * \text{CW}(\text{HALO}) + \quad (2) \\ & + \alpha * \sum \text{CW}(S_k) + \beta * \sum \text{CW}(SS_k) + \gamma * \sum \text{CW}(SSS_k) \end{aligned}$$

where ATOMPAIR, BOND, NOSP, and HALO are SMILES attributes which are defined according to [45].

For SET 1, the descriptor was calculated with  $w=1$ ;  $x=0$ ;  $y=0$ ;  $z=0$ ;  $\alpha=1$ ;  $\beta=1$ ; and  $\gamma=1$ . For SET 2 and SET 3, the descriptor was calculated with  $w=1$ ;  $x=1$ ;  $y=1$ ;  $z=1$ ;  $\alpha=1$ ;  $\beta=1$ ; and  $\gamma=1$ . In other words, the BOND, NOSP, and HALO do not improve the predicting power of the CORAL model for SET 1.

SMILES is a sequence of symbols which are a representation of a molecular structure. There are symbols which themselves are representations of a molecular feature, *e.g.* 'c', 'C', 'N', etc. There are undivided pairs of symbols which represent a molecular feature, *e.g.* 'Cl', 'Br', '@@', etc. We have denoted both of these types of information as SMILES atoms ( $S_k$ ).  $SS_k$  and  $SSS_k$  are combinations of two and three SMILES atoms (*e.g.* if the SMILES representation is ABCDE, the  $SS_k$  are AB, BC, CD, and DE; similarly  $SSS_k$  are ABC, BCD, and CDE. In order to avoid situations where the same molecular fragment is represented twice (*i.e.*, AB and BA), the  $SS_k$  and  $SSS_k$  are ordered according to ASCII codes of symbols).  $\text{CW}(x)$  is the correlation weight for a SMILES attribute  $x$  ( $x = \text{ATOMPAIR, BOND, NOSP, HALO, } S_k, SS_k, \text{ and } SSS_k$ ). Each SMILES attribute is represented by a sequence of twelve symbols. The first four symbols are the first zone; the second four symbols are the second zone; finally, the third four symbols are the third zone. All three zones are necessary for the SMILES attributes which involve three SMILES atoms (*i.e.*,  $SSS_k$ ). The  $SS_k$  are represented in the first and

**Table 1.** An example of the DCW(Threshold,  $N_{\text{epoch}}$ ) calculation: SET 1, split A.SMILES = Nc2cccc(Sc1ccccc1)c2C#NDCW(Threshold,  $N_{\text{epoch}}$ ) = 18.6437

$SA_k$	CW( $SA_k$ )	$N_{\text{TRN}}$	$N_{\text{CLB}}$	$N_{\text{TST}}$	$SA_k$	CW( $SA_k$ )	$N_{\text{TRN}}$	$N_{\text{CLB}}$	$N_{\text{TST}}$
N.....	0.0542	37	12	15	C...C.....	-0.6491	37	12	15
C.....	-0.6194	37	12	15	C...C.....	-0.6491	37	12	15
2.....	-0.4981	37	12	15	C...C.....	-0.6491	37	12	15
C.....	-0.6194	37	12	15	C...1.....	-0.7105	37	12	15
C.....	-0.6194	37	12	15	1...(.....	0.8246	27	9	8
C.....	-0.6194	37	12	15	C...(.....	0.4540	37	12	15
C.....	-0.6194	37	12	15	C...2.....	-0.6625	37	12	15
(.....	-0.5510	37	12	15	C...2.....	1.6020	34	10	11
S.....	-0.6540	37	12	15	C...#.....	0.1605	37	12	15
C.....	-0.6194	37	12	15	N...#.....	0.0437	37	12	15
1.....	-0.7230	37	12	15	N...c...2...	0.0	1	1	0
C.....	-0.6194	37	12	15	C...2...C...	-0.7041	37	12	15
C.....	-0.6194	37	12	15	C...C...2...	1.2980	37	12	15
C.....	-0.6194	37	12	15	C...C...C...	2.1520	37	12	15
C.....	-0.6194	37	12	15	C...C...C...	2.1520	37	12	15
C.....	-0.6194	37	12	15	C...C...(...	0.3510	37	12	15
1.....	-0.7230	37	12	15	C...(...S...	2.1980	14	2	5
(.....	-0.5510	37	12	15	C...S...(...	3.7750	11	3	2
C.....	-0.6194	37	12	15	S...c...1...	2.1990	4	2	3
2.....	-0.4981	37	12	15	C...1...C...	-0.6115	37	12	15
C.....	1.3530	37	12	15	C...C...1...	-0.7105	37	12	15
#.....	0.2165	37	12	15	C...C...C...	2.1520	37	12	15
N.....	0.0542	37	12	15	C...C...C...	2.1520	37	12	15
C...N.....	0.0	1	2	0	C...C...C...	2.1520	37	12	15
C...2.....	-0.6625	37	12	15	C...C...1...	-0.7105	37	12	15
C...2.....	-0.6625	37	12	15	C...1...(...	1.2000	27	9	8
C...C.....	-0.6491	37	12	15	C...(...1...	1.7980	18	5	6
C...C.....	-0.6491	37	12	15	2...C...(...	0.0282	35	11	12
C...C.....	-0.6491	37	12	15	C...2...C...	1.1437	34	10	11
C...(.....	0.4540	37	12	15	2...C...#...	1.1532	34	10	11
S...(.....	1.9950	37	11	13	N...#...C...	0.3010	37	12	15
C...S.....	2.0459	11	4	4	<b>NOSP10000000*</b>	1.8655	8	3	4
C...1.....	-0.7105	37	12	15	<b>HALO00000000</b>	2.4720	20	8	7
C...1.....	-0.7105	37	12	15	<b>BOND01000000</b>	2.4845	11	4	4
C...C.....	-0.6491	37	12	15	<b>+++N---B3==</b>	-0.6103	37	12	15

\*NOSP10000000 indicates the presence of nitrogen and absence of oxygen, sulphur, and phosphorus;

HALO00000000 indicates the absence of fluorine, chlorine, and bromine;

BOND01000000 indicates the presence of triple covalent bonds and absence of double bonds and stereo chemical bonds;

+++N---B3== indicates the presence of nitrogen and triple covalent bonds;

 $N_{\text{TRN}}$ ,  $N_{\text{CLB}}$  and  $N_{\text{TST}}$  are the numbers of a given  $SA_k$  in the sub-training, calibration, and test sets, respectively.

second zones. The  $S_k$  are located in the first zone. Vacant positions in this twelve-symbols representation are indicated by dots (Table 1).

The values of  $CW(x)$  are calculated with the Monte Carlo method. The classic scheme is to build a model that is satisfactory for the training set and evaluate whether the model is also appropriate for the external test set. However, the balance of correlations seems a more realistic approach. This approach, based on the split of the training set into sub-training and calibration sets, aims to avoid overtraining by means of the control of the statistical quality of the model for the calibration set. Thus, the calibration set plays the role of a 'preliminary test set'.

The correlation weights of rare molecular fragments and physicochemical situations lead to improvement of the statistical quality for compounds which are involved in the sub-training or calibration sets. Thus, a reliable model must be based on molecular fragments which are not rare. For this reason, we introduced a threshold to select SMILES attributes which are 'not rare'. If the threshold is set to five, then all SMILES attributes (including ATOMPAIR, BOND, NOSP, HALO,  $S_k$ ,  $SS_k$ , and  $SSS_k$ ) which take place only in four (or less) SMILES representations within the training set will be classified as rare. Correlation weights for these attributes will be defined as zero.

In addition, the number of epochs of the Monte Carlo optimization (Nepoch) is an important parameter: if Nepoch is large, overtraining can result (*i.e.*, very good statistical characteristics for training set and poor statistics for the test set); conversely, if Nepoch is small, one can obtain poor statistics for both training and test sets. Thus, an average value of the Nepoch parameter should be defined for the CORAL models.

### 3. Results and discussion

The statistical quality of QSAR models for the three above-mentioned endpoints calculated with the CORAL software is a mathematical function of two parameters of the Monte Carlo optimization: *i.e.*, the threshold that is used to define rare and not rare attributes, and the number of epochs of the optimization [45-47]. Fig. 1 shows representations of  $R^2_{test} = F(\text{Threshold}, N_{epoch})$  for SET 1, SET 2, and SET 3. The preferred values of the threshold ( $T^*$ ) and the number of epochs of the Monte Carlo optimization ( $N^*$ ) vary for each split into the sub-training, calibration and test set. We deem that the statistical quality of a CORAL model can be robust if the majority of molecular features (extracted from SMILES) take place in the sub-training and calibration sets.

Table 2 shows best models which were obtained by the classic scheme, the balance of correlations without use of the global attributes (ATOMPAIR for SET 1, ATOMPAIR, BOND, NOSP, and HALO for SET 2 and 3), and the balance of correlations with use of the global attributes. One can see (Table 2) that the balance of correlations method produced a superior model in comparison with the classic scheme, and the use of the aforementioned global attributes improves the predictive potential of the CORAL models for SET 1, SET 2 and SET 3.

The statistical characteristics of the models for SET 1, SET 2, and SET 3 were verified with the predictive ability criteria suggested by Golbraikh and Tropsha [48] and by Roy and Roy [49]. The verification is presented in the *Supplementary materials* section. The majority of the suggested models show predictive ability according to above-mentioned criteria.

The statistical quality of QSAR models for SET 1 [5] is  $n=39$ ,  $r^2=0.96$ ,  $s=0.286$  (training set) and  $n=15$ ,  $r^2=0.89$ ,  $s=0.489$  (test set). Thus, the statistical quality of the CORAL model can be better estimated (Table 2). In [6], the best  $R^2_{pred}$  ( $r^2_{LOO}$ ) for 13 compounds of the test set is 0.520. Hence, the CORAL models are better for the endpoint of SET 2 (Table 2). The predictive model for pEC50 of fullerene derivatives described by Durdagi *et al.* [17] is statistically characterized by  $n=48$ ,  $r^2=0.993$ ,  $s=0.127$  (training set) and  $r^2=0.744$ ,  $s=0.755$  (test set). The statistical quality of the CORAL models is similar [17].

It is to be noted that for SET 1, all models show predictive ability according to the criteria of Golbraikh and Tropsha [38] and Roy and Roy [39], but for SET 2 and SET 3, there are splits which have models which are not predictive according to these criteria (SET 2: splits A, B, D, and E; SET 3: splits A, B, C, and D). According to the criteria of Golbraikh and Tropsha, the following splits do not have predictive ability: B, D, E for SET 2 and A, C, D for SET 3. According to the criterion suggested by Roy and Roy, split B and split D in SET 3 do not have predictive ability. Thus the criteria of Golbraikh and Tropsha are in agreement with the criterion suggested by Roy and Roy only for split D in SET 3. In any case, the above mentioned criteria are very useful for comparison of a group of splits. In particular, one can see (Table 2) that SET 1 has stable statistical quality for all splits, whereas SET 2 and SET 3 have a wide range of statistical quality that considerably varies by different splits of compounds into the training and test sets.

We have selected the following models (Fig. 2):

For SET 1, split A

$$pIC_{50} = 0.0008 (\pm 0.02529) + 0.1212 (\pm 0.0006) * DCW(4,10) \quad (3)$$

**Table 2.** Correlation coefficients between experimental and predicted endpoints value for models constructed with the classic scheme (i.e., 'training-test system') and models obtained by means of balance of correlations (i.e., 'sub-training-calibrations-test system'), with and without correlation weights of the ATOMPAIR attribute.

Classic system	Balance of correlations without ATOMPAIR			Balance of correlations with ATOMPAIR				
Correlation coefficients between experimental and predicted values of logIC50 for TIBO derivatives (SET 1)								
split	training	Test	Sub-training	Calibration	Test	Sub-training	Calibration	Test
A	0.9746	0.9088	0.9258	0.9685	0.9181	0.9494	0.9672	0.9598
B	0.9779	0.8955	0.9003	0.9306	0.8971	0.9393	0.9837	0.9247
C	0.9767	0.9081	0.9310	0.9544	0.9325	0.9295	0.9561	0.9427
D	0.9875	0.8581	0.9331	0.9859	0.9167	0.9554	0.9900	0.9395
E	0.9612	0.9045	0.9217	0.9788	0.9218	0.9319	0.9850	0.9418
F	0.9752	0.8932	0.9423	0.9468	0.9573	0.9454	0.9486	0.9619
Correlation coefficients between experimental and predicted values of logIC50 for 2-amino-6-arylsulfonylbenzonitriles (SET 2)								
split	training	Test	Sub-training	Calibration	Test	Sub-training	Calibration	Test
A	0.7092	0.7346	0.6249	0.9682	0.8575	0.6025	0.9783	0.8658
B	0.7203	0.8187	0.4746	0.8565	0.9374	0.4653	0.8598	0.9414
C	0.7674	0.9084	0.7214	0.9363	0.9159	0.7291	0.9336	0.9206
D	0.5163	0.9555	0.4285	0.9054	0.9684	0.4064	0.8988	0.9720
E	0.6638	0.7576	0.5373	0.9765	0.9730	0.5263	0.9821	0.9770
F	0.6801	0.9010	0.5055	0.9575	0.8988	0.4785	0.9687	0.9065
Correlation coefficients between experimental and predicted values of the median effective concentration (pEC50) for fullerene C60 derivatives (SET 3)								
split	training	Test	Sub-training	Calibration	Test	Sub-training	Calibration	Test
A	0.7798	0.8902	0.8540	0.9789	0.9579	0.8523	0.9780	0.9611
B	0.7303	0.3892	0.7248	0.8640	0.9635	0.7272	0.8598	0.9824
C	0.8026	0.2182	0.8522	0.9676	0.8219	0.8564	0.9670	0.8259
D	0.7699	0.5168	0.9453	0.8261	0.9648	0.9545	0.8206	0.9838
E	0.6765	0.8616	0.8672	0.6533	0.9601	0.8633	0.6571	0.9689
F	0.2672	0.8748	0.5925	0.3340	0.9854	0.6045	0.3646	0.9869

n=47,  $r^2=0.9583$ ,  $q^2=0.9543$ ,  $s=0.293$ ,  $F=1035$   
(sub-training set);  
n=19,  $r^2=0.9722$ ,  $s=0.311$  (calibration set);  
n=16,  $r^2=0.9465$ ,  $s=0.351$  (test set);

n=27,  $r^2=0.8504$ ,  $q^2=0.8210$ ,  $s=0.504$ ,  $F=142$   
(sub-training set);  
n=15,  $r^2=0.9757$ ,  $s=0.728$  (calibration set);  
n=6,  $r^2=0.9541$ ,  $s=1.48$  (test set).

For SET 2, split A

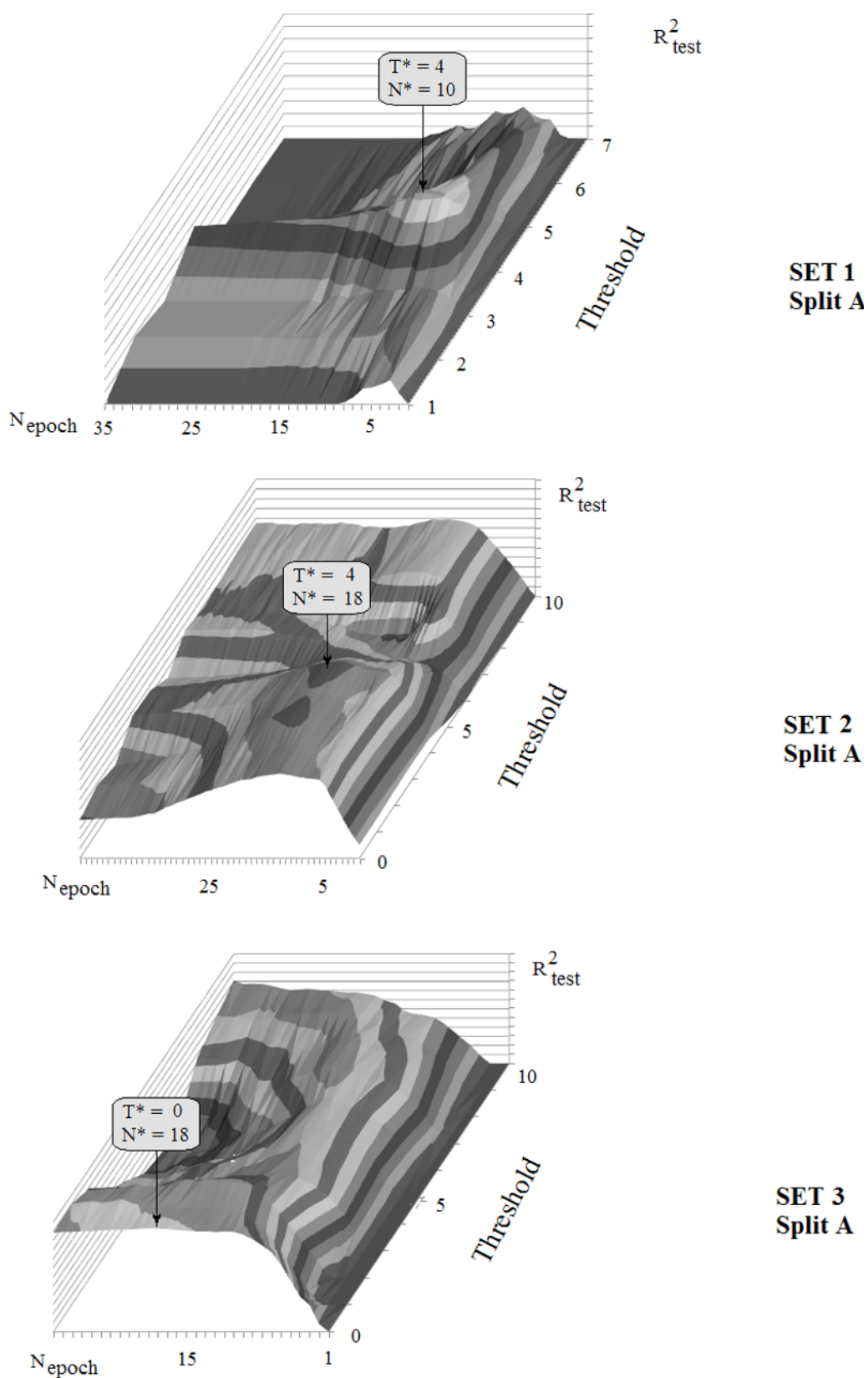
$$pIC_{50} = 0.6353 (\pm 0.0482) + 0.0599 (\pm 0.0012) * DCW(4,18) \quad (4)$$

n=37,  $r^2=0.6007$ ,  $q^2=0.5595$ ,  $s=0.629$ ,  $F=53$   
(sub-training set);  
n=12,  $r^2=0.9786$ ,  $s=0.576$  (calibration set);  
n=15,  $r^2=0.8619$ ,  $s=0.484$  (test set);

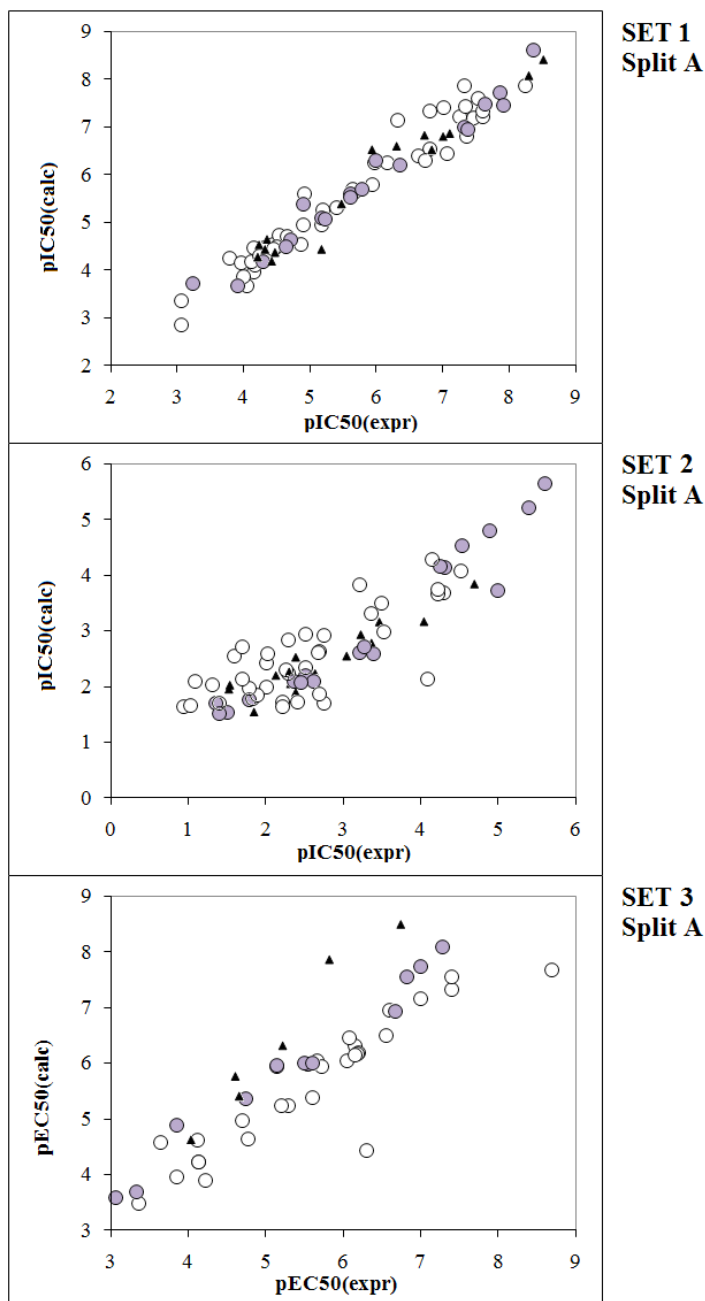
For SET 3, split A

$$pEC_{50} = 0.0672 (\pm 0.1181) + 0.0231 (\pm 0.0005) * DCW(0,18) \quad (5)$$

Having data on the correlation weights obtained in three (or more) runs of the Monte Carlo optimization method, one can extract SMILES attributes which have positive correlation weights in all runs of the optimization. Such SMILES attributes can be qualified as stable promoters of an endpoint increase. The stable promoters of increase for an endpoint can be used to construct a hypothetical compound with large value of an endpoint. However, an attribute should be a promoter of increase (i) for all runs of the Monte Carlo optimization (ii) for all splits into the sub-training, calibration, and



**Figure 1.** The graphical representation of the selection of the preferred threshold ( $T^*$ ) and the number of epochs ( $N^*$ ) for SET 1, SET 2, and SET 3.



**Subtraining set (○) Calibration set (◐) Test set (▲)**

**Figure 2.** QSAR models for anti-HIV-1 activity ( $pIC_{50}$ ) for SET 1 (Split A) and SET 2 (Split A). QSAR models for the median effective concentration ( $pEC_{50}$ ) for fullerene C60 derivatives for SET 3 (Split A).

test set, and (iii) the attribute should have considerable prevalence in the sub-training set and in the calibration set. In addition, hypothetical promising compounds should be highly similar to compounds characterized by high activity. Table 3 contains examples of SMILES attributes which are promoters of increase for endpoints related to SET 1, SET 2, and SET 3. Table 4 shows

examples of compounds which can be effective anti-HIV agents.

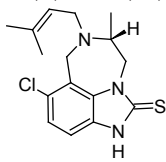
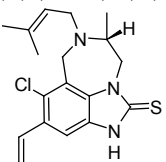
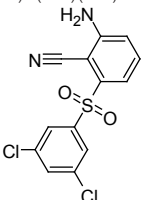
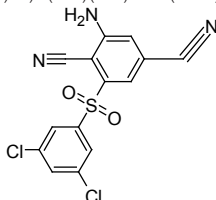
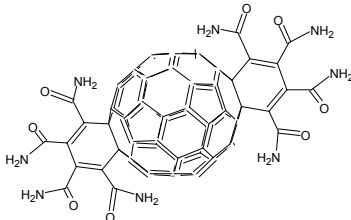
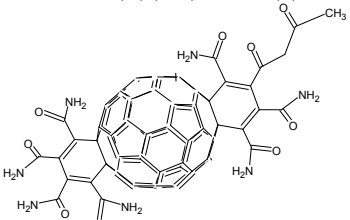
The *Supplementary materials* section contains the statistical characteristics of all described models, graphical representations of all described models, and lists of six splits for the three aforementioned datasets.



**Table 3.** The SMILES attributes which have stable positive values of the correlation weights for all starts of the Monte Carlo method optimization and for all splits.

SET	Promoter of increase for endpoints	Comments
1	C...=.....	Presence of double bonds involved carbon atom
	c...3.....	Presence of an aromatic fragment together with three rings
	N...(C...)	Presence of branching in the molecular skeleton that starts from carbon to nitrogen or vice versa
	S...=.....	Presence of sulphur atom connected with double covalent bond
2	/.....	Presence of cis- or trans- isomerism
	N...#...C...	Presence of -C≡N
	2...C...#...	Presence of two rings and triple bonds
	c...c...c...	Presence of an aromatic fragment
3	c...c...2...	Presence of an aromatic fragment connected to a ring
	=...C...(C...	Presence of fragment with the branching together double bonds
	+++ +O---B2==	Presence of (i) oxygen; and (ii) presence of double bond, which can be unconnected with the oxygen

**Table 4.** The search for the promising compound according to the calculated value of the endpoints.

SMILES, structure, ID and experimental endpoint value	SMILES, structure and calculated endpoint value	Comment
<chem>C/C(C)=C/CN1Cc3c(Cl)ccc2NC(=S)N(C[C@H]1C)c23</chem>  ID=3, SET 1 pIC50=8.37	<chem>C/C(C)=C/CN1Cc3c(Cl)c(cc2NC(=S)N(C[C@H]1C)c23)C=C</chem>  pIC50=9.048	The fragment C=C- is added
<chem>Clc1cc(cc(Cl)c1)S(=O)(=O)c2cccc(N)c2C#N</chem>  ID=59, SET 2 pIC50=4.155	<chem>Clc1cc(cc(Cl)c1)S(=O)(=O)c2cc(C#N)cc(N)c2C#N</chem>  PIC50=4.513	The fragment N≡C- is added
<chem>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</chem>  ID=42, SET 3 pEC50=7.40	<chem>CC(=O)CC(=O)C%29=C(C(N)=O)C%31%30C%17=C3C4C%28c%27c5c%33C=6C=7C2=C1c%33c%26C%24=C1C%23=C%22C2C8%32C(C(N)=O)=C(C(N)=O)C(C(N)=O)=C(C(N)=O)C%13%32c%21c%15c%12c%14C=9C(C3=C%11C4=C5C=6C=%10C(C=78)=C(C=9C=%10%11)C%12%13)=C%18c%14c%16c%15c%20C%19=C%16C(=C%17%18)C%31C%25=C%19C(=C%23c%20c%21%22)C%24=C%25C(c%26%27)C%28%30C(C(N)=O)=C%29C(N)=O</chem>  pEC50=7.758	The fragment -NH <sub>2</sub> is changed to -CH <sub>2</sub> -C(=O)-CH <sub>3</sub>



## 4. Conclusions

The CORAL software produced robust models for three data sets (SET 1: anti-HIV-1 activity of TIBO derivatives; SET 2: anti-HIV-1 activity of 2-amino-6-arylsulfonylbenzotriazoles; and SET 3: median effective concentration of fullerene derivatives). The balance of correlations method yielded better prediction than the classic scheme for all six random splits for each set of data. The global SMILES attributes (ATOMPAIR, BOND, NOSP, and HALO) can improve the predictive potential of the CORAL models for the aforementioned sets. The split of available data into the training set (used to build the model) and external test set can have a

considerable influence upon the statistical quality of the model. The CORAL models allow a mechanistic interpretation (based on the list of molecular features extracted from SMILES representations with stable positive correlation weights) and can be used to search for hypothetical effective anti-HIV agents (Table 4).

## Acknowledgements

We thank the EC project NanoBRIDGES. We also express our gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations on the computer science aspects of the work, and to J. Baggott for English editing.

## References

- [1] H. Zhang, L. Vrang, K. Backbro, P. Lind, C. Sahlberg, T. Unge, B. Oberg, *Antivir. Res.* 28, 331 (1995)
- [2] E.A. Castro, F. Torrens, A.A. Toropov, I.V. Nesterov, O.M. Nabiev, *Mol. Simul.* 30, 691 (2004)
- [3] K. Roy, J.T. Leonard, *Bioorg Med Chem.* 12, 745 (2004)
- [4] J.T. Leonard, K. Roy, *QSAR Comb. Sci.* 23, 23 (2004)
- [5] R. Darnag, E.L. Mostapha Mazouz, A. Schmitzer, D. Villemin, A. Jarid, D. Cherqaoui, *Eur. J. Med. Chem.* 45, 1590 (2010)
- [6] R. Hu, F. Barbault, M. Delamar, R. Zhang, *Bioorg. Med. Chem.* 17, 2400 (2009)
- [7] V. Ravichandran, V.K. Mourya, P.K. Agrawal, *Internet Electron. J. Mol. Des.* 11, 363 (2012)
- [8] N.S. Sapre, N. Pancholi, S. Gupta, *Internet Electron. J. Mol. Des.* 11, 55 (2012)
- [9] S. Vadivelan, T.N. Deeksha, S. Arun, P.K. MacHiraju, R. Gundla, B.N. Sinha, S.A.R.P. Jagarlapudi, *Eur. J. Med. Chem.* 46, 851 (2011)
- [10] S. Pirhadi, J.B. Ghasemi, *Eur. J. Med. Chem.* 45, 4897 (2010)
- [11] P. Lu, X. Wei, R. Zhang, *Eur. J. Med. Chem.* 45, 3413 (2010)
- [12] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, *Eur. J. Med. Chem.* 45, 1387 (2010)
- [13] N.S. Sapre, S. Gupta, N. Pancholi, N. Sapre, *J. Comput. Chem.* 30, 922 (2009)
- [14] N.S. Sapre, N. Pancholi, S. Gupta, N. Sapre, *J. Comput. Chem.* 29, 1699 (2008)
- [15] J.T. Leonard, K. Roy, *Eur. J. Med. Chem.* 43, 81 (2008)
- [16] D.J.G. Marino, E.A. Castro, A.A. Toropov, *Cent. Eur. J. Chem.* 4, 135 (2006)
- [17] S. Durdagi, T. Mavromoustakos, N. Chronakis, M.G. Papadopoulos, *Bioorg. Med. Chem.* 16, 9957 (2008)
- [18] P.R. Duchowicz, E.A. Castro, *Int. J. Mol. Sci.* 10, 2558 (2009)
- [19] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, P.A. Koutentis, G. Kollias, *Chem. Biol. Drug Des.* 76, 397 (2010)
- [20] J.A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. Rotondo, *J. Mol. Graph. Model.* 26, 32 (2007)
- [21] I. Gutman, B. Furtula, M. Petrović, *J. Math. Chem.* 46, 522 (2009)
- [22] CORAL, 2012 at <http://www.insilico.eu/coral/> Accessed on March 17, 2012
- [23] A.P. Toropova, A.A. Toropov, E. Benfenati, D. Leszczynska, J. Leszczynski, *J. Math. Chem.* 48, 959 (2010)
- [24] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Mol. Divers.* 15, 249 (2011)
- [25] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, *Eur. J. Med. Chem.* 45, 4399 (2010)
- [26] A.A. Toropov, A.P. Toropova, E. Benfenati, *Eur. J. Med. Chem.* 45, 3581 (2010)
- [27] D. Weininger, *J. Chem. Inf. Comput. Sci.* 28, 31 (1988)
- [28] D. Weininger, A. Weininger, J.L. Weininger, *J. Chem. Inf. Comput. Sci.* 29, 97 (1989)
- [29] D. Weininger, *J. Chem. Inf. Comput. Sci.* 30, 237 (1990)
- [30] A.A. Toropov, A.P. Toropova, I. Raska, E. Benfenati, *Eur. J. Med. Chem.* 45, 1639 (2010)
- [31] A. A. Toropov, E. Benfenati, *J. Mol. Struct. (Theochem)* 676, 165 (2004)
- [32] A.A. Toropov, A.P. Toropova, E. Benfenati, *Mol. Divers.* 14, 183 (2010)
- [33] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, *J. Comput. Chem.*

- 31, 381 (2010)
- [34] A.A. Toropov, A.P. Toropova, E. Benfenati, *Cent. Eur. J. Chem.* 7, 846 (2009)
- [35] A.A. Toropov, A.P. Toropova, E. Benfenati, *J. Math. Chem.* 46, 1060 (2009)
- [36] A.A. Toropov, A.P. Toropova, E. Benfenati, A. Manganaro, *Mol. Divers.* 14, 821 (2010)
- [37] A.A. Toropov, A.P. Toropova, E. Benfenati, A. Manganaro, *Mol. Divers.* 13, 367 (2009)
- [38] A.A. Toropov, A.P. Toropova, E. Benfenati, *Int. J. Mol. Sci.* 10, 3106 (2009)
- [39] I. Mitra, P.P. Roy, S. Kar, P.K. Ojha, K. Roy, *J. Chemometr.* 24, 22 (2010)
- [40] P.P. Roy, S. Paul, I. Mitra, K. Roy, *Molecules*, 14, 1660 (2009)
- [41] C. Porcelli, E. Boriani, A. Roncaglioni, A. Chana, E. Benfenati, *Environ. Sci. Technol.* 42, 491 (2008)
- [42] T. Oberg, *Chem. Res. Toxicol.* 17, 1630 (2004)
- [43] J.S. Duca, A.J. Hopfinger, *J. Chem. Inf. Comput. Sci.* 41, 1367 (2001)
- [44] A. Tropsha, *Mol. Inf.* 29, 476 (2010)
- [45] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski *J. Comput. Chem.* 32, 2727 (2011)
- [46] A.P. Toropova, A.A. Toropov, R. Gonella Diaza, E. Benfenati, G. Gini, *Cent. Euro. J. Chem.* 9, 165 (2011)
- [47] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, *Chemometr. Intell. Lab.* 105, 215 (2011)
- [48] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* 20, 269 (2002)
- [49] P.P. Roy, K. Roy, *QSAR Comb. Sci.* 27, 302 (2008)