

QSAR modeling of anxiolytic activity taking into account the presence of keto- and enol-tautomers by balance of correlations with ideal slopes

Research Article

Alla. P. Toropova¹, Andrey A. Toropov^{1*}, Emilio Benfenati¹,
Giuseppina Gini², Danuta Leszczynska³, Jerzy Leszczynski⁴

¹Institute of Pharmacological Researches "Mario Negri", 20156 Milan, Italy

²Department of Electronics and Information,
Polytechnical Institute of Milan, 20133 Milan, Italy

³Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental
Engineering, Jackson State University, Jackson, MS 39217-0510, USA

⁴Interdisciplinary Nanotoxicity Center, Department of Chemistry and
Biochemistry, Jackson State University, Jackson, MS 39217, USA

Received 2 March 2011; Accepted 11 May 2011

Abstract: Optimal descriptors calculated with simplified molecular input line entry system (SMILES) have been examined as a tool for prediction of anxiolytic activity. Descriptors calculated with SMILES (a) of keto-isomers; (b) of enol-isomers; and (c) of both keto-isomers together with enol-isomers have been studied. Three approaches have been compared: 1. classic 'training-test' system 2. balance of correlations and 3. balance of correlations with ideal slopes. The best statistical characteristics for the external validation set took place for optimal descriptors calculated with SMILES of both keto-form and enol-form (*i.e.*, molecular structure was represented in the format: 'SMILES of keto-form . SMILES of enol-form') by means of balance of correlations with ideal slopes. The predictive potential of this model was checked with three random splits.

Keywords: QSAR • SMILES • Tautomerism • Anxiolytic activity • Balance of correlation

© Versita Sp. z o.o.

1. Introduction

Tautomerism is an important phenomenon in chemistry and biochemistry. By taking this phenomenon into account one can improve the statistical characteristics of the quantitative structure – activity relationships (QSAR), which are used for prediction of the biochemical behaviour of substances.

Anxiolytic agents are widely used in medicine. The search for new anxiolytic agents is an important problem. QSAR prediction of the anxiolytic activity is possible [1]. These calculations can be useful in both practice and theory.

QSAR analysis has both many aims and approaches [2-8]. The validation of a QSAR model becomes a very important aspect of the QSAR analysis [9-11]. In the

present study we have used the probabilistic approach to validate a model calculated with the simplified molecular input line entry system (SMILES) [12,13]. In other words, models were examined with three random splits into sub-training, calibration, and validation sets.

The aim of the present study is the estimation of SMILES-based optimal descriptors as a tool to predict anxiolytic activity.

2. Experimental procedure

2.1. Method

A group of 67 pyrido[1,2-a]benzimidazole derivatives and their anxiolytic activities (pIC₅₀ values measured in the absence of γ -aminobutyric acid) were taken from [1]. The

* E-mail: aatoropov@yahoo.com

Supplementary materials section contains the molecular structures of these compounds.

Three versions of the SMILES-based optimal descriptors [13-15] were examined:

$$\text{DCW(Threshold)} = F(A) \quad (1)$$

$$\text{DCW(Threshold)} = F(B) \quad (2)$$

$$\text{DCW(Threshold)} = F(A,B) \quad (3)$$

where F is a mathematical function; A is SMILES for the keto-form of a given substance; B is SMILES for the enol-form of a given substance, and Threshold is a parameter that is used to classify SMILES attributes into two categories, *i.e.*, rare or active [12,13]. Rare attributes do not contain sound information and bring noise to the model. In order to avoid this influence of the rare (noise) SMILES attributes, one can fix zero value of the correlation weight of each rare attribute (Eq. 4).

Thus, Eq. 1 is the model for anxiolytic activities that is based on the keto-form of compounds, Eq. 2 is the model that is based on the enol-form, and finally Eq. 3 is the model that is based on both the keto- and enol-forms.

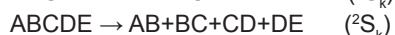
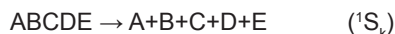
Three approaches to the calculation of optimal descriptors were examined. These are the classic training set – validation set scheme [14,15], the balance of correlations [12,13], and the balance of correlations with ideal slopes [16,17].

Classic scheme (CS). We have used optimal SMILES-based descriptors, which are calculated with the correlation weights (descriptor of correlation weights = DCW) as follows

$$\text{DCW(Threshold)} = \sum_{k=1}^E W(^1S_k) + \sum_{k=1}^{E-1} W(^2S_k) + \sum_{k=1}^{E-2} W(^3S_k) \quad (4)$$

where 1S_k , 2S_k , 3S_k are one-, two-, and three-element SMILES attributes. The majority of SMILES elements contain one character (*e.g.* 'C', 'c', 'N', etc.). There are SMILES elements which contain two characters (*e.g.* 'Cl', 'Br', '@@', etc.). In other words, the SMILES element encodes some part of the string which cannot be divided. However, the CORAL software used in this study (<http://www.insilico.eu/CORAL/>) reserves a standard twelve characters for a SMILES attribute and four positions in the standard string for each element, because, generally, a SMILES element can involve three ('Na+'), four ('[O-]'), or even larger numbers of characters ('[Cu+2]') [18-20]. Fortunately, the majority of attributes can be expressed by combining four (or less) characters. $W(^xS_k)$ is the correlation weight for a SMILES attribute ($x=1,2,3$).

The process of calculating 1S_k , 2S_k , 3S_k can be represented by the scheme:



For instance, SMILES = 'CCCN' is represented by nine strings. Table 1 shows strings encoded with 1S_k , 2S_k , and 3S_k for the above SMILES. Thus, each SMILES is converted in a group of SMILES attributes (Table 1). When the preparation of all attributes which occur in all substances is completed, the system of building up the model is provided with the list of SMILES attributes for which the correlation weights $W(^xS_k)$ should be calculated. It is to be noted that each SMILES attribute is a representation of some molecular fragment.

Using the Monte Carlo method, one can calculate the $W(^xS_k)$ values that produce the maximum correlation coefficient between DCW(Threshold) and the pIC_{50} for the training set. Having numerical data for optimal $W(^xS_k)$, one can calculate DCW(Threshold) for all compounds (*i.e.*, both for the training set and validation set). By the least squares method one can calculate a model of pIC_{50} :

$$\text{pIC}_{50} = C_0 + C_1 * \text{DCW(Threshold)} \quad (5)$$

The predictive potential of the model calculated with Eq. 5 should be checked with the external validation set.

Balance of correlations (BC). The classic scheme can lead to overtraining (overfitting), *i.e.*, a situation when high correlation for the training set is accompanied by poor correlation for the validation set. The correlation balance is aimed to avoid the overtraining. The essence of the method is the following: (a) the training set should be split into a sub-training set and calibration set; (b) instead of the Monte Carlo optimization of the correlation coefficient between DCW(Threshold) and pIC_{50} , one can use the Monte Carlo optimization with the target function calculated as

$$\text{BC} = R + R' - \text{abs}(R-R') * 0.1 \quad (6)$$

where R and R' are the correlation coefficients for the sub-training set and calibration set, respectively. In fact, the calibration set is a preliminary validation set. A low value of the correlation coefficient for the calibration set leads to a decrease of BC. In fact, the search for the maximum of BC as calculated with Eq. 6 is an attempt to obtain the same correlation coefficients for the sub-training set and the calibration set.

Table 1. Example of SMILES attributes (1S_k , 2S_k , and 3S_k) for SMILES represented by "CCCN"

1S_k	2S_k	3S_k
Cxxxxxxxx*		
Cxxxxxxxx	CxxxCxxxxxx	
Cxxxxxxxx	CxxxCxxxxxx	CxxxCxxxCxxx
Nxxxxxxxx	CxxxNxxxxxx	CxxxCxxxNxxx

^{*)} The 'x' is used to indicate the vacant place in the string of symbols used for representation of a SMILES attribute.

Balance of correlations with ideal slopes (IS). Good correlations between DCW(Threshold) and pIC_{50} can take place for considerably different slopes in plots of pIC_{50} (experiment) versus pIC_{50} (calculated) for the sub-training set and calibration set. Fig. 1 shows this situation. In order to avoid this situation, one can use the following target function for the Monte Carlo optimization:

$$IS = BC - [\text{abs}(C0) + \text{abs}(C0') + \text{abs}(C1 - C1')] * 0.005 \quad (7)$$

where C0 and C0' are intercepts for the sub-training set and calibration set, respectively, and C1 and C1' are slopes for the sub-training and calibration set (the C0, C0', C1, C1' are calculated by the least squares method). In fact, the optimization with the target function that is calculated with Eq. 7 is an attempt to obtain intercepts for the sub-training set and the calibration set which are equal to zero, as well as identical slopes for the sub-training set and the calibration set. Unfortunately, this is an ideal situation which can be obtained only approximately [16,17].

The coefficients of 0.1 (Eq. 6) and 0.005 (Eq. 7) were defined empirically. This shows that the correlation coefficients provide a larger contribution to the quality of the model. However, the influence of the intercepts (C0, and C0') and slopes (C1, and C1') is also relevant, because the models which have been calculated with Eq. 7 are more accurate (Table 2) than models based on the balance of correlations (Eq. 6). Using 0.1 or even 0.01 instead of 0.005 leads to ineffective optimization based on the target function calculated with Eq. 7.

The algorithm of the Monte Carlo optimization that is used for all three approaches mentioned above (i.e., CS, BC, and IS) was described in [21].

3. Results and Discussion

Table 2 shows the statistical quality of the models. One can see that models calculated with a separate keto-form (i.e., using Eq. 1) or with a separate enol-form (i.e., using Eq. 2) have similar statistical quality for the external test set. However, statistical quality of the

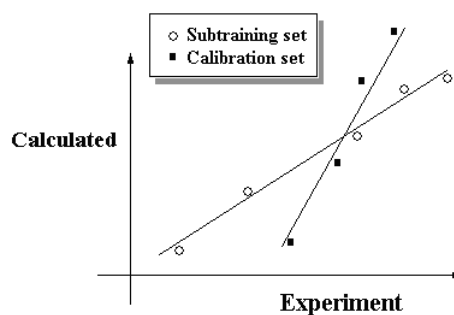


Figure 1. Good correlations which are accompanied by different slopes for the sub-training set and the calibration set in plots of experiment versus calculated values of an endpoint.

model calculated by taking into account the molecular architecture of both forms (i.e., using Eq. 3) is superior. In addition, the models calculated with the CS scheme have modest statistical quality (the range of r^2_{test} is 0.6747-0.8063); the statistical quality of the models calculated with BC is better (the range of r^2_{test} is 0.7069-0.8567); and the statistical quality of models calculated with IS is the best (the range of r^2_{test} is 0.7974-0.8763).

Fig. 2 shows diagrams of the observed correlation coefficients for sub-training, calibration, and test sets for the range of the threshold 1-20 (in the case of the CS, the diagram contains data for the training and test sets, since the calibration set is not used). One can see the best predictions (maximum r^2_{test}) are obtained with IS for all three splits, but the optimal threshold values are different. These are 5, 6, and 8 for split 1, split 2, and split 3, respectively.

The best model arises for split 3. This model is calculated as follows:

$$pIC_{50} = 2.2888(\pm 0.1307) + 0.05654(\pm 0.00142)^* \quad (8)$$

*DCW(8)

n=32, $r^2=0.6291$, $q^2=0.5805$, s=0.630, F=51 (sub-training set)

n=23, $r^2=0.7242$, s=0.677, F=55 (calibration set)

n=12, $r^2=0.8750$, s=0.490, F=70 (validation set)

Table 3 contains the values of pIC_{50} found experimentally and calculated with Eq. 8. Table 4 shows an example of the calculation of DCW(8). Fig. 3 depicts the model graphically. The *Supplementary materials* section contains data on correlation weights for the calculation of the DCW(8).

Table 2. Average correlation coefficients for the pIC_{50} models (external validation sets) obtained with three probes of the Monte Carlo optimization [F(A) is the model based on keto-form; F(B) is the model based on enol-form; and F(A,B) is the model calculated by taking into account both the keto-form and the enol-form]. The best models are indicated in bold text.

Split	Classic scheme			Balance of correlations			Balance of correlations with Ideal Slopes		
	F(A)	F(B)	F(A,B)	F(A)	F(B)	F(A,B)	F(A)	F(B)	F(A,B)
1	0.6246	0.6362	0.6747	0.6510	0.6603	0.7069	0.7277	0.7291	0.7974
2	0.6132	0.5771	0.7650	0.6346	0.5535	0.7687	0.6240	0.6204	0.8029
3	0.7889	0.7841	0.8063	0.7997	0.7993	0.8567	0.8566	0.7763	0.8763

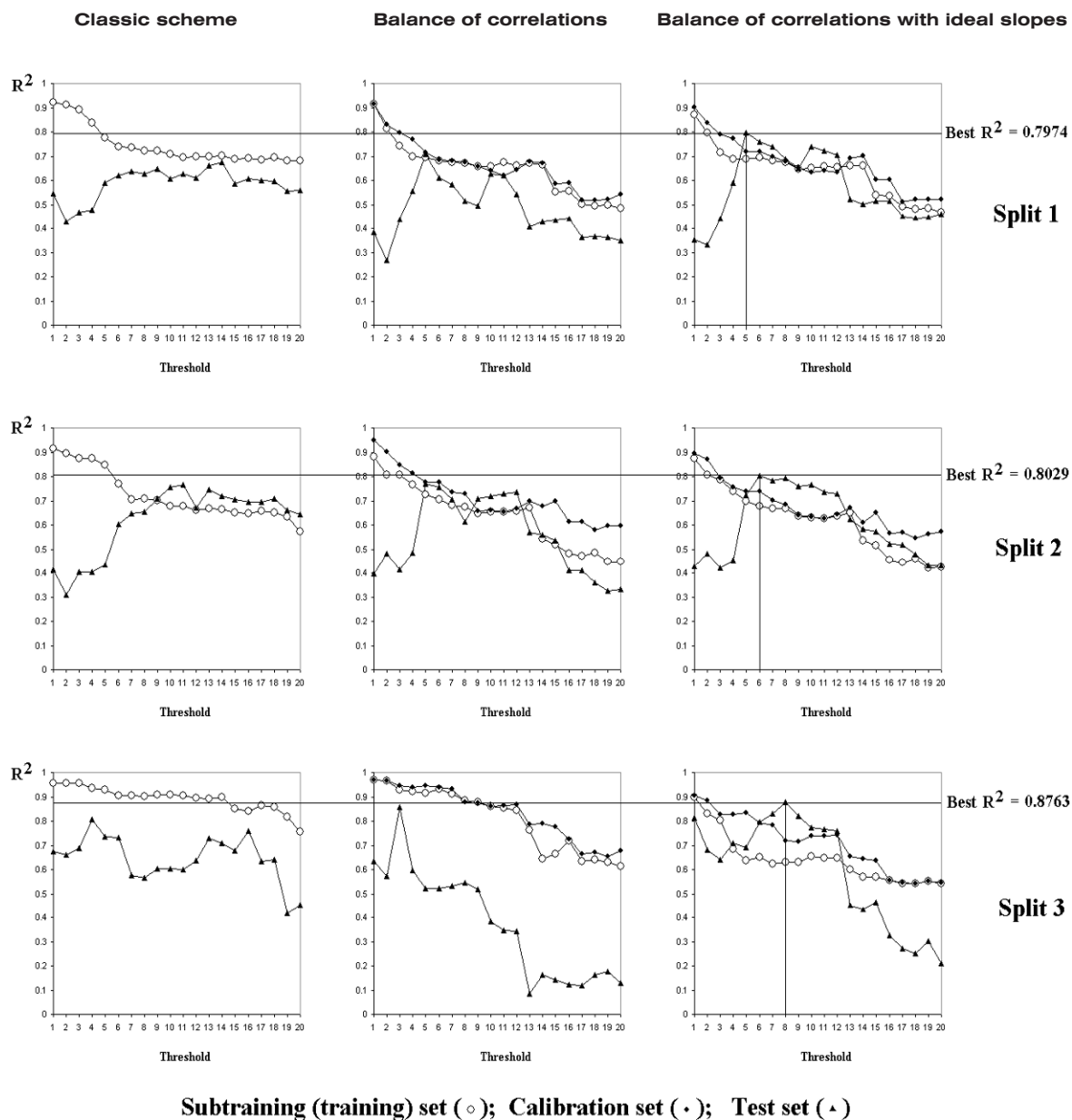


Figure 2. Diagrams of correlation coefficients versus the Threshold values for three random splits. One can see that the balance of correlations with ideal slopes gives the maximum correlation coefficient between DCW and pIC_{50} for the validation sets. The descriptors were calculated with Eq. 3

Table 3. Values of anxiolytics activity (pIC_{50}) from experiments and calculated with Eq. 8. The molecular structure is represented in the format 'SMILES of keto-form . SMILES of enol-form'.

ID	SMILES	DCW(8)	Expr	Calc
Sub-training set				
P1	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	88.1308382	8.040	7.268
P4	<chem>C1c4ccccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.C1c4ccccc4NC(=O)C=2c3nc1ccccc1n3CCC=2</chem>	81.8014747	7.920	6.911
P8	<chem>COc4ccccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.COc4ccccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	81.8824544	6.000	6.915
P9	<chem>C1c4ccccc(C)c4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.C1c4ccccc(C)c4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	81.7637028	6.490	6.908
P14	<chem>NC(=O)C2=C3Nc1ccccc1N3CCC2=O.NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	67.7360730	5.620	6.116
P16	<chem>S=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.S=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	86.6191127	6.440	7.183
P19	<chem>O=C(Sc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Sc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	89.6077880	7.540	7.352
P21	<chem>Fc4ccccc4NC(=O)C2=C3Nc1ccccc1N3CC=CC2=O.Fc4ccccc4NC(=O)c2c3nc1ccccc1n3ccc2O</chem>	134.4648060	9.640	9.886
P23	<chem>Oc1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.Oc1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	88.0426091	7.340	7.263
P25	<chem>Nc4ccccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.Nc4ccccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	72.7510016	7.300	6.399
P26	<chem>CN(C)c1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.CN(C)c1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	75.2744925	6.570	6.542
P28	<chem>CN(C)c4ccccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.CN(C)c4ccccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	76.4201682	5.000	6.607
P29	<chem>CN(C)c4ccc(NC(=O)C2=C3Nc1ccccc1N3CCC2=O)c(F)c4.CN(C)c4ccc(NC(=O)C=2c3nc1ccccc1n3CCC=2O)c(F)c4</chem>	64.4565559	6.920	5.931
P30	<chem>CN(C)c4ccc(NC(=O)C2=C3Nc1ccccc1N3CCC2=O)c(C)c4.CN(C)c4ccc(NC(=O)C=2c3nc1ccccc1n3CCC=2O)c(C)c4</chem>	59.7550663	5.400	5.665
P31	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	77.6569524	6.800	6.676
P33	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	85.8202534	7.230	7.138
P36	<chem>Fc4cncccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.Fc4cncccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	94.1884845	7.280	7.610
P37	<chem>C1c4ncccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.C1c4ncccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	86.2527330	7.170	7.162
P38	<chem>O=C(NCc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(NCc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	83.2318150	7.190	6.991
P43	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4C=CC3=O.O=C(Nc1ccccc1)c3c4nc2ccccc2n4ccc3O</chem>	81.0926322	6.590	6.871
P45	<chem>Oc1cc2NC3=C(C(=O)CCN3c2cc1)C(=O)Nc4ccccc4.Oc1cc2nc3C(=O)CCN3c2cc1)C(=O)Nc4ccccc4</chem>	88.8405337	7.390	7.308
P47	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(O)Cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(O)Cccc1n3CCC=2O</chem>	99.0822515	8.110	7.887
P48	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(O)Cccc1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(O)Cccc1n3CCC=2O</chem>	98.9914751	6.810	7.882
P49	<chem>COc1cc2NC3=C(C(=O)CCN3c2cc1)C(=O)Nc4ccccc4.COc1cc2nc3C(=O)CCN3c2cc1)C(=O)Nc4ccccc4</chem>	86.1286565	7.400	7.155
P52	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(cccc1C)N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(cccc1C)n3CCC=2O</chem>	97.8295180	7.020	7.816
P53	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(ccc1N3CCC2=O)C(F)F.Fc4ccccc4NC(=O)C=2c3nc1c(ccc1n3CCC=2O)C(F)F</chem>	75.3170896	6.740	6.544
P55	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(Cl)Cccc1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(Cl)Cccc1n3CCC=2O</chem>	104.5503841	8.960	8.196
P58	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(Cl)Cccc1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(Cl)Cccc1n3CCC=2O</chem>	103.6759025	8.200	8.146
P60	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(F)Cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(F)Cccc1n3CCC=2O</chem>	101.8920635	8.720	8.046
P61	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(F)Cccc1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(F)Cccc1n3CCC=2O</chem>	101.8012871	8.200	8.041
P63	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(ccc(F)c1F)N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(ccc(F)c1F)n3CCC=2O</chem>	101.3254463	8.960	8.014
P65	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1c(cc(F)c1F)N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1c(cc(F)c1F)n3CCC=2O</chem>	100.4509647	7.520	7.964
Calibration set				
P5	<chem>Fc4ccccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	102.6413963	8.770	8.088
P6	<chem>COc1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.COc1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	85.3307319	7.390	7.110
P7	<chem>COc1ccccc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.COc1ccccc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	86.2052135	7.590	7.159
P10	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	102.5506199	8.550	8.083
P11	<chem>O=C(NC1CCCC1)C3=C4Nc2ccccc2N4CCC3=O.O=C(NC1CCCC1)C=3c4nc2ccccc2n4CCC=3O</chem>	85.3196446	6.850	7.109
P12	<chem>O=C(NC1CCC1)C3=C4Nc2ccccc2N4CCC3=O.O=C(NC1CCC1)C=3c4nc2ccccc2n4CCC=3O</chem>	92.8600262	7.520	7.535
P13	<chem>O=C(NC1CC1)C3=C4Nc2ccccc2N4CCC3=O.O=C(NC1CC1)C=3c4nc2ccccc2n4CCC=3O</chem>	96.6302170	7.800	7.748
P15	<chem>O=C(OCC)C2=C3Nc1ccccc1N3CCC2=O.O=C(OCC)C=2c3nc1ccccc1n3CCC=2O</chem>	71.1054957	6.170	6.306
P17	<chem>CN(c1ccccc1)C(=O)C3=C4Nc2ccccc2N4CCC3=O.CN(c1ccccc1)C(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	68.3790025	5.000	6.152
P20	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4C=CC3=O.O=C(Nc1ccccc1)c3c4nc2ccccc2n4ccc3O</chem>	91.5665180	7.620	7.462
P22	<chem>O=C(O)c1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.O=C(O)c1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	65.5834716	5.000	5.994
P24	<chem>Nc1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.Nc1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	71.5913527	4.890	6.334
P27	<chem>CN(C)c1ccccc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.CN(C)c1ccccc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	76.1489741	7.430	6.591
P32	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	86.7475832	7.290	7.190
P34	<chem>C1c4cncccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.C1c4cncccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	71.5269520	6.660	6.330
P39	<chem>O=C(NCc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(NCc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	92.3224458	6.820	7.505
P42	<chem>O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	77.8758430	6.320	6.689
P50	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(O)Cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(O)Cccc1n3CCC=2O</chem>	98.2077699	8.210	7.838
P51	<chem>Cc4ccccc1c4NC2=C(C(=O)CCN12)C(=O)Nc3ccccc3.Cc4ccccc1c4nc2C(=O)CCN12)C(=O)Nc3ccccc3</chem>	98.6620685	8.000	7.863
P59	<chem>C1c1ccc2NC3=C(C(=O)CCN3c2cc1)C(=O)Nc4ccccc4.C1c1ccc2nc3C(=O)CCN3c2cc1)C(=O)Nc4ccccc4</chem>	86.0089241	6.760	7.148
P62	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(ccc(F)c1F)N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(ccc(F)c1F)n3CCC=2O</chem>	101.4162227	8.850	8.019
P64	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(cc(F)c1F)N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(cc(F)c1F)n3CCC=2O</chem>	100.5417411	7.700	7.969
P66	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(F)c(F)c1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(F)c(F)c1n3CCC=2O</chem>	92.6339461	6.070	7.523

Continued Table 3. Values of anxiolytics activity (pIC_{50}) from experiments and calculated with Eq. 8. The molecular structure is represented in the format 'SMILES of keto-form . SMILES of enol-form'.

ID	SMILES	DCW(8)	Expr	Calc
Validation set				
P2	<chem>Clc1ccc(cc1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.Clc1ccc(cc1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	83.4281413	6.210	7.002
P3	<chem>Clc1cccc(c1)NC(=O)C3=C4Nc2ccccc2N4CCC3=O.Clc1cccc(c1)NC(=O)C=3c4nc2ccccc2n4CCC=3O</chem>	84.3026229	6.920	7.052
P18	<chem>O=C(Oc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Oc1ccccc1)C=3c4nc2ccccc2n4CCC=3O</chem>	83.4242062	6.700	7.002
P35	<chem>Cc4ncccc4NC(=O)C2=C3Nc1ccccc1N3CCC2=O.Cc4ncccc4NC(=O)C=2c3nc1ccccc1n3CCC=2O</chem>	70.8855742	5.150	6.294
P40	<chem>O=C(Nc1cnccn1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1cnccn1)C=3c4nc2ccccc2n4CCC=3O</chem>	84.4369984	6.440	7.059
P41	<chem>O=C(Nc1ncccn1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ncccn1)C=3c4nc2ccccc2n4CCC=3O</chem>	84.5618052	7.270	7.067
P44	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(O)cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(O)cccc1n3CCC=2O</chem>	98.1968569	7.980	7.837
P46	<chem>Fc4ccccc4NC(=O)C2=C3Nc1cccc(O)c1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1cccc(O)c1n3CCC=2O</chem>	98.1968569	7.850	7.837
P54	<chem>Fc4ccccc4NC(=O)C2=C3Nc1c(Cl)cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1c(Cl)cccc1n3CCC=2O</chem>	104.6411605	8.150	8.201
P56	<chem>Clc1cc2NC3=C(C(=O)CCN3c2cc1)C(=O)Nc4ccccc4.Clc1cc2nc3C(=C(O)CCn3c2cc1)C(=O)Nc4ccccc4</chem>	84.2260659	7.180	7.048
P57	<chem>Fc4ccccc4NC(=O)C2=C3Nc1cc(Cl)cccc1N3CCC2=O.Fc4ccccc4NC(=O)C=2c3nc1cc(Cl)cccc1n3CCC=2O</chem>	103.7666789	8.290	8.152
P67	<chem>Fc4ccccc(F)c4NC(=O)C2=C3Nc1ccc(F)c(F)c1N3CCC2=O.Fc4ccccc(F)c4NC(=O)C=2c3nc1ccc(F)c(F)c1n3CCC=2O</chem>	92.5431697	7.180	7.517

Table 4. Example of DCW(8) calculation for a substance (P1) which is represented by the SMILES: O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O DCW(8) = 88.1308382

SMILES attribute with one element, 1S_k	$W(^1S_k)$	SMILES attribute with two elements, 2S_k	$W(^2S_k)$	SMILES attribute with three elements, 3S_k	$W(^3S_k)$
Oxxxxxxxxx	1.2719842				
=xxxxxxxxx	-2.2014019	Oxxx=xxxxxx	-2.1994232		
Cxxxxxxxxx	-0.3206889	Cxxx=xxxxxx	0.3667137	Oxxx=xxxCxxx	0.0000000
(xxxxxxxxx	-0.8829287	Cxxx(xxxxxx	-1.4836818	=xxxCxxx(xxx	2.1374485
Nxxxxxxxxx	-2.0876386	Nxxx(xxxxxx	3.7006953	Nxxx(xxxCxxx	1.4792241
cxxxxxxxxx	-0.7405851	cxixNxxxxxx	-1.5267925	cxixNxxx(xxx	0.0000000
1xxxxxxxxx	5.8997748	cxix1xxxxxx	1.4457884	Nxxxxxx1xxx	-2.3039632
cxixxxxxxxxx	-0.7405851	cxix1xxxxxx	1.4457884	cxix1xxcxxx	1.8464055
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxx1xx	-2.1506885
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxxcxix	1.2647377
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxxcxix	1.2647377
1xxxxxxxxx	5.8997748	cxix1xxxxxx	1.4457884	cxixcxxx1xx	-2.1506885
(xxxxxxxxx	-0.8829287	1xx(xxxxxx	0.2671641	cxix1xxx(xxx	1.3956548
Cxxxxxxxxx	-0.3206889	Cxxx(xxxxxx	-1.4836818	Cxxx(xxx1xxx	1.0026046
3xxxxxxxxx	5.9294417	Cxxx3xxxxxx	-1.0581548	3xxxCxxx(xxx	0.8880894
=xxxxxxxxx	-2.2014019	=xxx3xxxxxx	-0.2418664	Cxxx3xxx=xxx	-0.7277957
Cxxxxxxxxx	-0.3206889	Cxxx=xxxxxx	0.3667137	Cxxx=xxx3xxx	0.7038408
4xxxxxxxxx	0.6960386	Cxxx4xxxxxx	-0.9386474	=xxxCxxx4xxx	-1.2726292
Nxxxxxxxxx	-2.0876386	Nxxx4xxxxxx	0.4028821	Nxxx4xxxCxxx	-1.7787158
cxixxxxxxxxx	-0.7405851	cxixNxxxxxx	-1.5267925	cxixNxxx4xxx	-1.7385920
2xxxxxxxxx	5.6377804	cxix2xxxxxx	2.6149517	Nxxxxxx2xxx	-1.7110720
cxixxxxxxxxx	-0.7405851	cxix2xxxxxx	2.6149517	cxix2xxcxxx	2.1285531
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxx2xxx	2.5239971
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxxcxix	1.2647377
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxxcxix	1.2647377
cxixxxxxxxxx	-0.7405851	cxixcxxxxxx	-1.0453451	cxixcxxxcxix	1.2647377
2xxxxxxxxx	5.6377804	cxix2xxxxxx	2.6149517	cxixcxxx2xxx	2.5239971

Continued Table 4. Example of DCW(8) calculation for a substance (P1) which is represented by the SMILES: O=C(Nc1ccccc1)C3=C4Nc2ccccc2N4CCC3=O.O=C(Nc1ccccc1)C=3c4nc2ccccc2n4CCC=3O DCW(8) = 88.1308382



SMILES attribute with one element, 1S_k	$W(^1S_k)$	SMILES attribute with two elements, 2S_k	$W(^2S_k)$	SMILES attribute with three elements, 3S_k	$W(^3S_k)$
Nxxxxxxxx	-2.0876386	Nxxx2xxxxx	1.0975444	cxxx2xxxNxxx	0.7788656
4xxxxxxxx	0.6960386	Nxxx4xxxxx	0.4028821	4xxxNxxx2xxx	-1.8247358
Cxxxxxxxx	-0.3206889	Cxxx4xxxxx	-0.9386474	Nxxx4xxxCxxx	-1.7787158
Cxxxxxxxx	-0.3206889	CxxxCxxxxx	0.5510065	CxxxCxxx4xxx	2.2541704
Cxxxxxxxx	-0.3206889	CxxxCxxxxx	0.5510065	CxxxCxxxCxxx	-2.1154130
3xxxxxxxx	5.9294417	Cxxx3xxxxx	-1.0581548	CxxxCxxx3xxx	-0.4492216
=xxxxxxxx	-2.2014019	=xxx3xxxxx	-0.2418664	Cxxx3xxx=xxx	-0.7277957
Oxxxxxxxx	1.2719842	Oxxx=xxxxx	-2.1994232	Oxxx=xxx3xxx	-1.8742464
.xxxxxxxx	6.3036180	Oxxx.xxxxx	2.0704994	=xxxOxxx.xxx	1.2961041
Oxxxxxxxx	1.2719842	Oxxx.xxxxx	2.0704994	Oxxx.xxxOxxx	0.0000000
=xxxxxxxx	-2.2014019	Oxxx=xxxxx	-2.1994232	=xxxOxxx.xxx	1.2961041
Cxxxxxxxx	-0.3206889	Cxxx=xxxxx	0.3667137	Oxxx=xxxCxxx	0.0000000
(xxxxxxxx	-0.8829287	Cxxx(xxxxx	-1.4836818	=xxxCxxx(xxx	2.1374485
Nxxxxxxxx	-2.0876386	Nxxx(xxxxx	3.7006953	Nxxx(xxxCxxx	1.4792241
cxxxxxxxx	-0.7405851	cxxxNxxxxx	-1.5267925	cxxxNxxx(xxx	0.0000000
1xxxxxxxx	5.8997748	cxxx1xxxxx	1.4457884	Nxxxccc1xxx	-2.3039632
cxxxxxxxx	-0.7405851	cxxx1xxxxx	1.4457884	cxxx1xxxccx	1.8464055
cxxxxxxxx	-0.7405851	cxxxccccxxx	-1.0453451	ccccccc1xxx	-2.1506885
cxxxxxxxx	-0.7405851	cxxxccccxxx	-1.0453451	cccccccxxx	1.2647377
cxxxxxxxx	-0.7405851	cxxxccccxxx	-1.0453451	cccccccxxx	1.2647377
cxxxxxxxx	-0.7405851	cxxxccccxxx	-1.0453451	cccccccxxx	1.2647377
1xxxxxxxx	5.8997748	cxxx1xxxxx	1.4457884	ccccccc1xxx	-2.1506885
(xxxxxxxx	-0.8829287	1xxx(xxxxx	0.2671641	cxxx1xxx(xxx	1.3956548
Cxxxxxxxx	-0.3206889	Cxxx(xxxxx	-1.4836818	Cxxx(xxx1xxx	1.0026046
=xxxxxxxx	-2.2014019	Cxxx=xxxxx	0.3667137	=xxxCxxx(xxx	2.1374485
3xxxxxxxx	5.9294417	=xxx3xxxxx	-0.2418664	Cxxx=xxx3xxx	0.7038408
cxxxxxxxx	-0.7405851	cxxx3xxxxx	6.1923868	ccc3xxx=xxx	1.9870684
4xxxxxxxx	0.6960386	cxxx4xxxxx	1.7015689	4xxxccc3xxx	-1.4351969
nxxxxxxxx	-1.1419529	nxxx4xxxxx	2.9733890	nxxx4xxxccc	2.3494223
cxxxxxxxx	-0.7405851	nxxxccccxxx	-1.5660261	ccccnxxx4xxx	4.0961712
2xxxxxxxx	5.6377804	cxxx2xxxxx	2.6149517	nxxxccc2xxx	4.4522769
cxxxxxxxx	-0.7405851	cxxx2xxxxx	2.6149517	ccc2xxxccc	2.1285531
cxxxxxxxx	-0.7405851	ccccccccxxx	-1.0453451	ccccccc2xxx	2.5239971
cxxxxxxxx	-0.7405851	ccccccccxxx	-1.0453451	cccccccxxx	1.2647377
cxxxxxxxx	-0.7405851	ccccccccxxx	-1.0453451	cccccccxxx	1.2647377
cxxxxxxxx	-0.7405851	ccccccccxxx	-1.0453451	cccccccxxx	1.2647377
2xxxxxxxx	5.6377804	ccc2xxxxx	2.6149517	ccccccc2xxx	2.5239971
nxxxxxxxx	-1.1419529	nxxx2xxxxx	-0.0017811	nxxx2xxxccc	0.8517795
4xxxxxxxx	0.6960386	nxxx4xxxxx	2.9733890	4xxxnxxx2xxx	-1.9961987
Cxxxxxxxx	-0.3206889	Cxxx4xxxxx	-0.9386474	nxxx4xxxCxxx	2.9259553
Cxxxxxxxx	-0.3206889	CxxxCxxxxx	0.5510065	CxxxCxxx4xxx	2.2541704
Cxxxxxxxx	-0.3206889	CxxxCxxxxx	0.5510065	CxxxCxxxCxxx	-2.1154130
=xxxxxxxx	-2.2014019	Cxxx=xxxxx	0.3667137	CxxxCxxx=xxx	-2.2971478
3xxxxxxxx	5.9294417	=xxx3xxxxx	-0.2418664	Cxxx=xxx3xxx	0.7038408
Oxxxxxxxx	1.2719842	Oxxx3xxxxx	-1.4111115	Oxxx3xxx=xxx	2.1204695

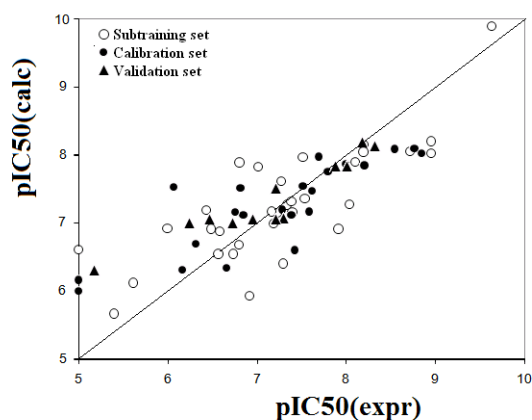


Figure 3. Graphical representation of the model calculated with Eq. 8.

The anxiolytic activity of different substances is important [1,22-28]; however, the QSAR analysis of this endpoint is carried out in few studies [1,22-24]. The statistical characteristics of the model of pIC_{50} described in [1] are the following: $n=67$, $r^2=0.951$, $s=0.246$, $F=140$. In other words, external validation is absent in this work [1]. The model has been built by the multiple linear regression analysis (MLRA) method with eight descriptors. However, the external checking of an MLRA model can help avoid overtraining [9]. The statistical characteristics of the best model for anxiolytic activity which has been obtained using parameters from quantum chemistry and neural networks [22] are the following: $n=33$, $r^2 = 0.8305$, $s=0.5700$ (training set), and $n=15$, $r^2=0.8154$, $s=0.7242$ (test set). The model of anxiolytic activity based on the PLS method [23] is characterised by $n=47$, $r^2=0.866$ (training set) and $n=7$, $r^2=0.681$ (test set). The QSAR model for anxiolytic agents described in Ref. 24 is characterized by $q^2=0.58$, *i.e.*, the statistical quality of this model is similar to the statistical quality of Eq. 8. Thus, one can consider the

statistical quality of the model calculated with Eq. 8 and the statistical quality of the above-mentioned models [1,22-24] to be similar, in spite of differences in the approaches used.

4. Conclusions

1. The SMILES-based optimal descriptors calculated with the balance of correlations (which is a system consisting of the sub-training set, the calibration set, and the external test set) yield better predictions of anxiolytic activity than the descriptors calculated with the “classic scheme” (which is a system consisting of the training set and test set, without the calibration set);

2. The optimal SMILES-based descriptors calculated by taking into account intercepts and slopes in the sub-training set and in the calibration set improves the accuracy of the prediction of the anxiolytics activity obtained by the balance of correlations. This is carried out without taking into account the intercepts and the slopes;

3. The SMILES-based descriptors calculated by taking into account both the keto-form and the enol-form of substances yield better prediction of the anxiolytic activity than the descriptors which are calculated using only one of the two aforementioned forms.

Acknowledgements

The authors express their gratitude to OSIRIS and the NSF CREST Interdisciplinary Nanotoxicity Center NSF-CREST - Grant # HRD-0833178 for financial support. Also, the authors express their gratitude to Dr. L. Cappellini and Dr. G. Bianchi for technical assistance and to Dr. J. Baggot for the English revisions.

References

- [1] J. W. Zou, C.C. Luo, H-X. Zhang, J. Mol. Graph. Model. 26, 494 (2007)
- [2] P.P. Roy, K. Roy, QSAR Comb Sci. 27, 302 (2008)
- [3] A.A. Toropov, A.P. Toropova, E. Benfenati, Eur. J. Med. Chem. 45, 3581(2010)
- [4] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorg Med Chem. 16, 7944 (2008)
- [5] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, QSAR Comb Sci. 25, 928 (2006)
- [6] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Polymer 47, 3240 (2006)
- [7] T. Puzyn, N. Suzuki, M. Haranczyk, J. Rak, J. Chem. Inform. Model. 48, 1174 (2008)
- [8] A.A. Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. 444, 209 (2007)
- [9] A.A. Toropov, B.F. Rasulev, J. Leszczynski, Bioorg. Med. Chem. 16, 5999 (2008)
- [10] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, E. Benfenati, Chemosphere 73, 1701 (2008)
- [11] G. Gini, E. Benfenati, Int. J. Artif. Intell. T. 16, 243 (2007)
- [12] A.A. Toropov, A.P. Toropova, E. Benfenati, Eur. J. Med. Chem. 44, 2544 (2009)

- [13] A.A. Toropov, A.P. Toropova, E. Benfenati, *Int. J. Mol. Sci.* 10, 3106 (2009)
- [14] A.A. Toropov, A.P. Toropova, *J. Mol. Struct. (Theochem)*. 578, 129 (2002)
- [15] A.A. Toropov, E. Benfenati, *Bioorg. Med. Chem.* 14, 3923 (2006)
- [16] A.A. Toropov, A.P. Toropova, E. Benfenati, *Eur. J. Med. Chem.* 45, 3581 (2010)
- [17] A.P. Toropova, A.A. Toropov, E. Benfenati, D. Leszczynska, J. Leszczynski, *J. Math. Chem.* 48, 959 (2010)
- [18] D. Weininger, *J. Chem. Inf. Comput. Sci.* 28, 31 (1988)
- [19] D. Weininger, A. Weininger, J.L. Weininger, *J. Chem. Inf. Comput. Sci.* 29, 97 (1989)
- [20] D. Weininger, *J. Chem. Inf. Comput. Sci.* 30, 237 (1990)
- [21] A.A. Toropov, A.P. Toropova, E. Benfenati, *Mol. Divers.* 14, 183 (2010)
- [22] B. Xia, W. Ma, B. Zheng, X. Zhang, B. Fan, *Eur. J. Med. Chem.* 43, 1489 (2008)
- [23] M.P. Freitas, *Chemometr. Intell. Lab.* 91, 173 (2008)
- [24] M.P. Freitas, S.D. Brownb, J.A. Martinsa, *J. Mol. Struct.* 738, 149 (2005)
- [25] B. Costa, E. Da Pozzo, B. Chelli, N. Simola, M. Morelli, M. Luisi, M. Maccheroni, S. Taliani, F. Simorini, F. Da Settimo, C. Martini, *Psychoneuroendocrino* 36, 463 (2011)
- [26] O. Grundmann, J.-I. Nakajima, K. Kamata, S. Seo, V. Butterweck, *Phytomedicine* 16, 295 (2009)
- [27] E. Ognibene, P. Bovicelli, W. Adriani, L. Saso, G. Laviola, *Prog. Neuro-Psychoph.* 32, 128 (2008)
- [28] A. Zamilpa, M. Herrera-Ruiz, E. Del Olmo, J.L. Lopez-Pe´rez, J. Tortoriello, A. San Feliciano, *Bioorg. Med. Chem. Lett.* 17, 4016 (2007)