Chemical Physics Letters 542 (2012) 134–137

Contents lists available at SciVerse ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

The average numbers of outliers over groups of various splits into training and test sets: A criterion of the reliability of a QSPR? A case of water solubility

Alla P. Toropova [a], Andrey A. Toropov [a,*], Emilio Benfenati [a], Giuseppina Gini [b], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
[b] Department of Electronics and Information, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
[c] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St., Jackson, MS 39217-0510, USA
[d] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J.R. Lynch St., P.O. Box 17910, Jackson, MS 39217-0510, USA

ARTICLE INFO

Article history:
Received 5 April 2012
In final form 30 May 2012
Available online 4 June 2012

ABSTRACT

The validation of quantitative structure–property/activity relationships (QSPR/QSAR) is an important challenge of modern theoretical chemistry. Analysis of QSPRs which are obtained with various distribution into sub-systems of training and of testing can be a useful approach to estimate reliability of QSPR predictions. The balance of correlation is an approach for the building up of QSPR with using three components of available data: (a) sub-training set (developer), (b) calibration set (critic), and (c) test set (estimator). Computational experiments have shown that the probabilistic interdependence between the distribution of available data into sub-training set, calibration set, and test set and the average numbers of outliers in the test set exists.

1. Introduction

Water solubility is an important physicochemical property that plays a significant role in various physical and biological processes [1]. It is not surprisingly, that a large number of quantitative structure–property relationships (QSPR) were dedicated to water solubility. The systematization of these works may be carried out with different emphasis, e.g., according to classes of organic compounds, such as esters [2,3], alcohols [4], polychlorinated biphenyl [5], and compounds with a large structural diversity [6]. The systematization of researches on QSPR for water solubility can be based on different approaches such as multiple regression analysis (MRA) [7], partial least squares method (PLS) and artificial neural networks (ANN) [8]. Other important aspect of the QSPR models for solubility is functionality of substances, e.g., drug-like [9–11], or environmentally important ones [12].

The distribution of substances into training and test sets has influence upon the predictive potential of a QSPR/QSAR and consequently the algorithm of the splitting data into the training and test set is an important component of a QSPR/QSAR analyses [12]. There are various algorithms to split available data into the training and test sets [13–15], however random splits are preferable approach to detect probabilistic interdependence between statistical characteristics of the training and test sub-systems.

The aim of this Letter is the analysis of influence of various distributions (splits) of the set of available substances into training and test sub-systems upon statistical quality of the solubility models.

2. Method

2.1. Data

The numerical values for water solubility were taken from the US National Library of Medicine [16]. The negative decimal logarithm of the solubility expressed in mg/L has been examined as the endpoint. The total number of substances is 488. Table 1 shows random distributions of these substances into the sub-training, calibration, and test sets which we analyzed. Supplementary materials section contains random splits of 488 organic compounds characterized by various distributions (Table 1) of substances in the sub-training set, calibration set, and test set.

2.2. Optimal descriptor

We have used the following version of the optimal descriptor calculated with the CORAL software [17]:

$$DCW(\text{Threshold},N_{\text{epoch}}) = \sum W(S_k) + \sum W(VD_k) + W(NOSP) + W(HALO) \quad (1)$$

where Threshold is integer value to classify molecular features extracted from simplified molecular entry system (SMILES) [18,19]

* Corresponding author.
E-mail addresses: aatoropov@yahoo.com, andrey.toropov@marionegri.it (A.A. Toropov).

0009-2614/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cplett.2012.05.073

Table 1
The denoting (ID) of distributions into sub-training, calibration, and test sets which were examined. The distributions were prepared with random number generator.

ID for distribution	Sub-training set (%)	Calibration set (%)	Test set (%)
101 080	10	10	80
105 040	10	50	40
501 040	50	10	40
303 040	30	30	40
206 020	20	60	20
602 020	60	20	20
404 020	40	40	20

or from hydrogen suppressed molecular graph as rare or not rare. Correlation weight of each rare molecular feature is fixed equal to zero, i.e., the rare feature is not taken into account for the building up of model; N_{epoch} is the number of cycles of the Monte Carlo optimization aimed to build up the model; S_k is SMILES element, i.e., one or two symbols which cannot be examined separately (e.g., 'Cl', 'Br', etc.); VD_k is vertex degree in hydrogen suppressed graph; NOSP and HALO are global SMILES attributes [20] which are mathematical functions of presence of nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, and bromine; W(x) is the correlation weight for a molecular feature 'x' [20–28].

The correlation coefficient between water solubility (or arbitrary other endpoint) and descriptor that is calculated with Eq. (1) is a mathematical function of correlation weights. One can calculate (by the Monte Carlo method) numerical values of the W(x) which give maximum of the above-mentioned correlation coefficient for the training set hoping that the descriptors calculated with those weights also will be correlated with the endpoint for external test set [29–31]. The Monte Carlo optimization can be carried out with improved target functions with taking into account the possibility of overtraining [20,27,28]. The models were built up with the CORAL software [17] with the following parameterization: threshold = 1, 3, and 5; D_{start} = 0.1; $d_{precision}$ = 0.01; and N_{epoch} = 50. Figure 1 shows the flowchart for building the CORAL models.

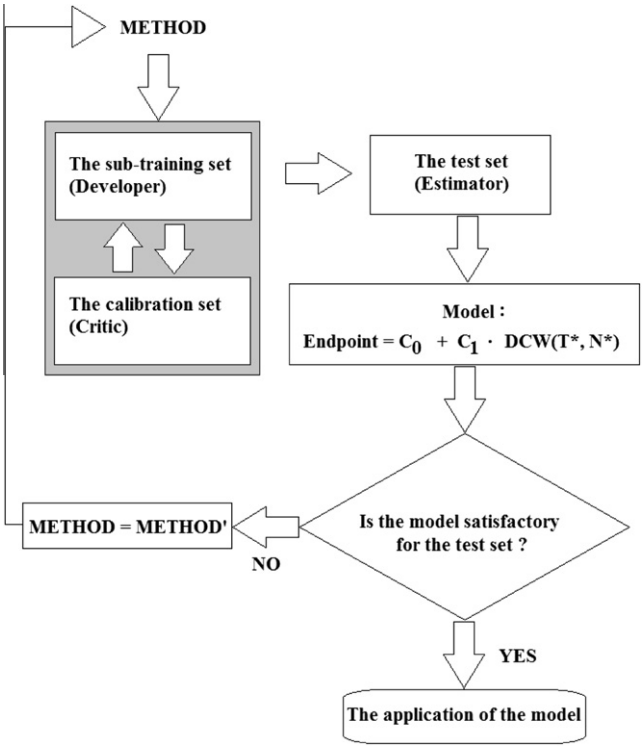

Figure 1. Flowchart of the building up and application of a CORAL model.

3. Results and discussion

Table 2 contains the statistical quality of the prediction of solubility for models calculated with various distribution of substances into the sub-training, calibration, and test sets (Table 1). For each distribution average values obtained in three runs of the Monte Carlo optimization are represented.

Having a completed quantitative structure–property relationship (QSPR), one can remove one substance from the test set. It can lead to a change of the mean square error for the test set. We classify as 'leader of outliers', the substance that gives the maximal decrease of the mean square error for the test set. The reduced test set can contain a next 'leader of outliers'. If the QSPR model is satisfactory, the removing of several 'leaders of outliers' can lead to situation where mean square error of the training set becomes larger than mean square error of the test set. We denoted this situation as 'ideal split'. The average number of 'leaders of outliers' (for several runs of the Monte Carlo optimization) which must be removed to obtain 'ideal split' $\overline{N_L}$ can be considered a measure of reliability for a QSPR model. One can see (Table 2) that distribution 101 080 gives the maximal number of 'leaders of outliers' in comparison with others. Vice versa, distribution 404 020 gives the minimal value of 'leaders of outliers'. It is to be noted, Table 2 contains average numerical values of the correlation coefficient and mean square error for various splits and for three starts of the Monte Carlo optimization with the threshold [17,20] that is equal to 1, 3, and 5. Thus, the numerical data can be a basis for significant statistical conclusions. Functions of sets which are involved in the building up of the model may be formulated as the following: sub-training set is the developer of a model; the calibration set is the critic for tuning the model; the test set is an estimator of a model (Figure 1).

In fact, QSPR analysis of one split into training and test sets is an analogy of a measurement that is carried out one time only: consequently, no information about the dispersion of the statistical criteria becomes available after this action. However, the experience shows that the dispersion exists and the numerical value of this dispersion is important for the estimation of the true reliability of various approaches. One cannot argue that the cross-validation technique gives an adequate estimation of the dispersion of statistical criteria, because the list of descriptors (in the case of multiple regression analysis) or the list of molecular fragments (in the case of optimal descriptors) which are involved in the building up of a model can be various for various splits into the training and test sets.

The domain of applicability is a component of the modern QSPR/QSAR. We have suggested the alternative to the applicability domain. The definition of the domain of applicability is list of substances for which a model gives determinist prediction with given accuracy. We have suggested the empirical technique in order to estimate the probability of outliers. This task is simpler than the definition of the domain of applicability, however, the solution of this task can be the useful for comparisons of various models. In addition, we have noticed apparent influence of a split into the sub-training, calibration, and test upon the statistical quality of the QSPR model for the water solubility: (i) the majority of substances (≈80%) should be placed in the sub-training and calibration sets; and (ii) the numbers of substances in the sub-training set and in the calibration set should be approximately equivalent (Table 2, Figure 2). According to this rationale one should expect the most reliable model in the case of 404 020 distribution. Average values of the numbers of 'leaders of outliers' and of standard error of estimation related to examined distributions into

Table 2
Statistical quality of predictions for various distributions substances into sub-training, calibration and test sets. $\overline{N_L}$ is the average number of 'leaders of outliers' which should be removed in order to obtain the average root mean square error of test set ($\overline{S_{test}}$) less than $(S_{sub\text{-}training} + S_{calibration})/2$; N_{test} is the number of compounds in the test set; the correct prediction is the percentage of substances, in the test set, which are not outliers.

Split	Threshold	$\overline{N_L}$	N_{test}	$\overline{r^2_{test}}$	$\overline{S_{test}}$	Correct prediction (%)
101080-1	1	82.3	388	0.9197	0.5118	78.9
101080-2	1	66.7	393	0.8642	0.6177	83.0
101080-3	1	40.0	373	0.8592	0.6465	89.3
101080-1	3	65.0	388	0.8961	0.5678	83.2
101080-2	3	40.0	393	0.8544	0.6448	89.8
101080-3	3	21.7	373	0.8331	0.6949	94.2
101080-1	5	62.0	388	0.8943	0.5841	84.0
101080-2	5	35.3	393	0.8450	0.6722	91.0
101080-3	5	22.7	373	0.8186	0.7256	93.9
105040-1	1	14.0	195	0.8805	0.6606	92.8
105040-2	1	18.0	198	0.8672	0.6276	90.9
105040-3	1	38.0	216	0.8940	0.5926	82.4
105040-1	3	9.0	195	0.8657	0.6886	95.4
105040-2	3	13.7	198	0.8481	0.6665	93.1
105040-3	3	18.7	216	0.8616	0.6738	91.4
105040-1	5	7.7	195	0.8529	0.7096	96.1
105040-2	5	10.0	198	0.8096	0.7452	94.9
105040-3	5	16.3	216	0.8477	0.6990	92.4
501040-1	1	39.3	181	0.8969	0.5615	78.3
501040-2	1	20.3	214	0.8733	0.5960	90.5
501040-3	1	29.3	200	0.8906	0.5768	85.3
501040-1	3	41.7	181	0.8864	0.5635	77.0
501040-2	3	11.7	214	0.8658	0.6213	94.5
501040-3	3	24.3	200	0.8880	0.5853	87.8
501040-1	5	37.7	181	0.8822	0.5792	79.2
501040-2	5	9.3	214	0.8655	0.6298	95.6
501040-3	5	20.0	200	0.8840	0.6002	90.0
303040-1	1	5.0	181	0.8547	0.6602	97.2
303040-2	1	13.0	189	0.8599	0.6411	93.1
303040-3	1	9.0	184	0.8872	0.6476	95.1
303040-1	3	2.3	181	0.8416	0.6867	98.7
303040-2	3	11.0	189	0.8552	0.6600	94.2
303040-3	3	6.7	184	0.8787	0.6685	96.4
303040-1	5	1.0	181	0.8357	0.6972	99.4
303040-2	5	12.0	189	0.8442	0.6873	93.6
303040-3	5	7.0	184	0.8723	0.6841	96.2
206020-1	1	6.0	103	0.8697	0.6508	94.2
206020-2	1	7.0	109	0.8248	0.6523	93.6
206020-3	1	11.3	99	0.8746	0.6358	88.5
206020-1	3	6.0	103	0.8630	0.6740	94.1
206020-2	3	2.3	109	0.8087	0.6878	97.8
206020-3	3	9.0	99	0.8634	0.6636	90.9
206020-1	5	5.0	103	0.8459	0.7059	95.1
206020-2	5	1.0	109	0.8116	0.6905	99.0
206020-3	5	4.0	99	0.8370	0.7160	95.9
602020-1	1	7.0	100	0.8778	0.6532	93.0
602020-2	1	2.7	98	0.8543	0.6601	97.2
602020-3	1	8.0	94	0.8917	0.6266	91.4
602020-1	3	6.7	100	0.8768	0.6560	93.3
602020-2	3	1.3	98	0.8463	0.6772	98.6
602020-3	3	7.0	94	0.8883	0.6408	92.5
602020-1	5	6.0	100	0.8693	0.6694	94.0
602020-2	5	1.0	98	0.8509	0.6816	98.9
602020-3	5	7.0	94	0.8844	0.6517	92.5
404020-1	1	1.3	99	0.8685	0.6598	98.6
404020-2	1	5.3	101	0.8851	0.6477	94.7
404020-3	1	3.0	117	0.8752	0.6517	97.4
404020-1	3	1.0	99	0.8676	0.6618	98.9
404020-2	3	5.0	101	0.8814	0.6562	95.0
404020-3	3	2.0	117	0.8697	0.6663	98.2
404020-1	5	1.0	99	0.8582	0.6836	98.9
404020-2	5	4.0	101	0.8792	0.6702	96.0
404020-3	5	2.0	117	0.8700	0.6659	98.2

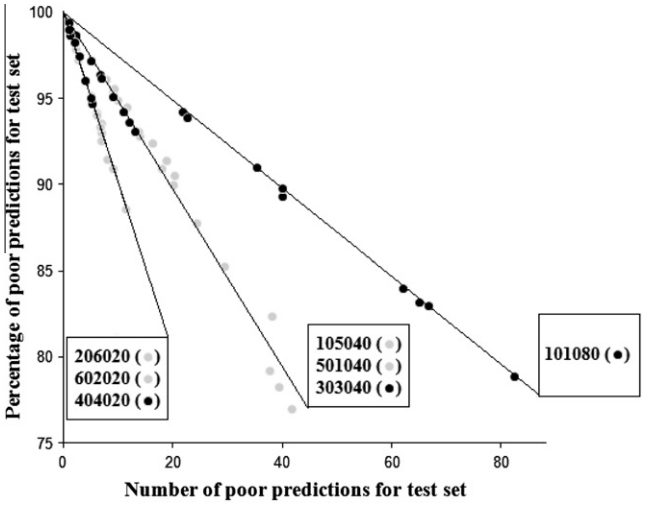

Figure 2. The plot of the number of outliers in the test set vs. percentage of poor predictions (in the test set) for different distributions (Table 1) into the sub-training set, the calibration set, and test set.

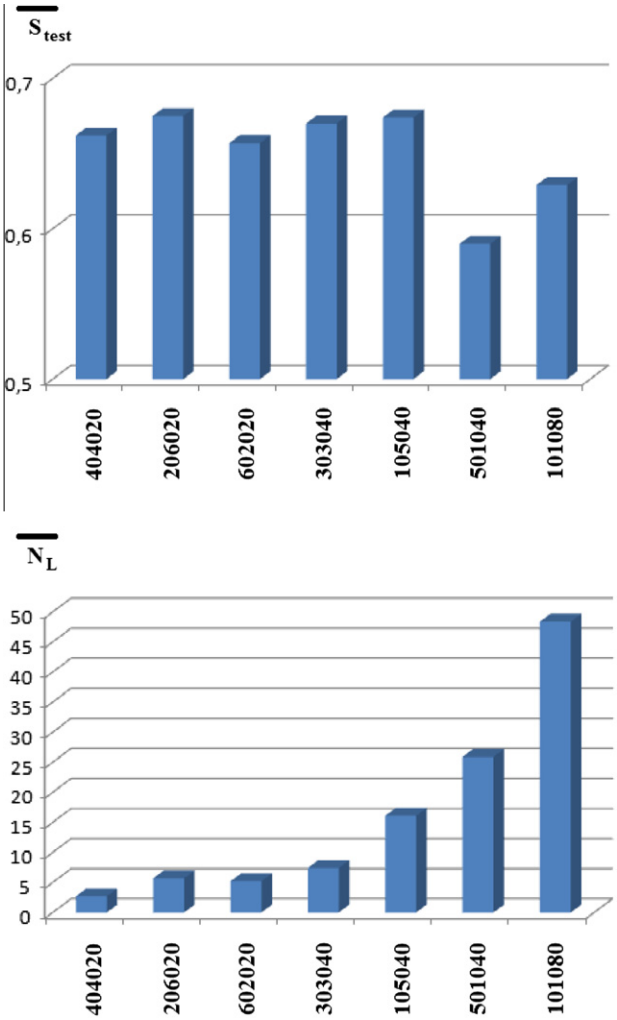

Figure 3. The average number of outliers $\overline{N_L}$ and average standard error of estimation $\overline{S_{test}}$ for various distributions (Table 1) of available data into sub-training set, calibration set, and test set.

the sub-training set, calibration set, and test set (Figure 3) confirm this hypothesis. The situation can be interpreted as the following: (i) the increase of the number of substances in the test set leads to increase of the number of outliers; (ii) in the case of equivalent numbers of substances in the sub-training set and in the calibration set (e.g., 404020 and 303040) the number of outliers is

minimal in comparison with splits where percentage of substances in the test set is equivalent, but the percentage of substances in the sub-training set and in the calibration set, are different (e.g., 602 020, 206 020, 501 040, and 105 040); and (iii) it should be noted that there are no correlations between the number of substances in the test set and the standard error of estimation calculated in the above-mentioned 'ideal splits'.

Statistical characteristics of QSPR model for solubility of 193 acyclic compounds [7] ($n = 193$, $r^2 = 0.946$) are better than statistical quality of CORAL models for 404 020-splits ($n = 105.7$, $r^2 = 0.87$, Table 2). However, one should take into account, (i) the above-mentioned model [7] has been calculated for all 193 compounds, without external checking up; (ii) only acyclic compounds were studied in this Letter [7]. QSPR that was built up with Artificial Neural Networks [8] is characterized by $n = 879$, $r^2 = 0.95$ (training set), $n = 412$ $r^2 = 0.92$ (validation set), and $n = 21$, $r^2 = 0.90$ (test set). Unfortunately, there are no data on statistical quality of such model for other splits. Statistical quality of model of water solubility of drug-like compounds [11] is the following: $n = 100$, $r^2 = 0.774$ (training set) and $n = 48$, $r^2 = 0.598$ (test set). Thus, comparison of the CORAL model with above-mentioned [7,8,11] indicates that statistical quality of suggested model (Table 2) is reasonably good.

The reliability of a model often conflicts with the precision of the model. Apparently, the reliable model should be classified as more useful than model with high precision, but not reliable. Thus, the search for criteria to estimate the reliability of a model is an important task. The solutions of this task can be found by means of involving the Gaussian principle [32] and with using of the PRECLAV software [33].

The suggested analysis of models obtained with groups of various splits and groups of various distributions (i.e., splits with various percentage of data in the training set and in the test set) is an approach that can improve the theoretical tools for the building up of models of different endpoints in general and for water solubility in particular, because, it gives possibility to obtain a measure of the reliability of QSPR/QSAR: the average value of outliers in the external test set.

4. Conclusions

The average number of 'leader of outliers' $\overline{N_L}$ can be used as a measure of reliability of a model, instead of the definition of the domain of applicability. The $\overline{N_L}$ decreases with increase of the number of substances distributed in the sub-training and calibration sets. Distributions into sub-training set (developer of the model) and into calibration set (critic of the model) should be equivalent: the preference for the 'developer' as well as the preference for 'critic' decrease the predictive potential of the model, because the number of outliers will increase in both these cases (Table 2, Figure 3). The distribution (40% in the sub-training set; 40% in the calibration set; and 20% in the test set) that gives most reliable model of water solubility has been defined by means of the computational experiment.

Acknowledgements

We thank ANTARES (Project No. LIFE08-ENV/IT/00435), and the National Science Foundation (NSF/CREST HRD-0833178, and EPS-CoR Award #:362492-190200-01\NSFEPS-090378) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cplett.2012.05.073.

References

[1] C. Xue-Qing, S.J. Cho, Y. Li, S. Venkatesh, J. Pharm. Sci. 91 (2002) 1838.
[2] T.F. Parkerton, W.J. Konkel, Ecotox. Environ. Safety 45 (2000) 61.
[3] I. Cousins, D. Mackay, Chemosphere 41 (2000) 1389.
[4] B. Ren, J. Chem. Inf. Comput. Sci. 42 (2002) 858.
[5] N.N. Nirmalakhandan, R.E. Speece, Environ. Sci. Technol. 23 (1989) 708.
[6] A.R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson, J. Chem. Inf. Comput. Sci. 38 (1998) 720.
[7] K. Roy, A. Saha, Internet Electron. J. Mol. Des. 2 (2003) 475.
[8] I.V. Tetko, V.Yu. Tanchuk, T.N. Kasheva, A.E.P. Villa, J. Chem. Inf. Comput. Sci. 41 (2001) 1488.
[9] J. Ghasemi, S. Saaidpour, Chem. Pharm. Bull. 55 (2007) 669.
[10] L. Du-Cuny, J. Huwyler, M. Wiese, M. Kansy, Eur. J. Med. Chem. 43 (2008) 501.
[11] P.R. Duchowicz, A. Talevi, C. Bellera, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. 15 (2007) 3711.
[12] P.P. Roy, J.T. Leonard, K. Roy, Chemometr. Intell. Lab. 90 (2008) 31.
[13] T. Puzyn, A. Mostrag-Szlichtyng, A. Gajewicz, M. Skrzyński, A.P. Worth, Struct. Chem. 22 (2011) (2011) 795.
[14] P. Gramatica, E. Giani, E. Papa, J. Mol. Graph. Modell. 25 (2007) (2007) 755.
[15] K. Roy, I. Mitra, Mini-Rev. Med. Chem. 12 (2012) 491.
[16] US National Library of Medicine, http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp (accessed 28.02.2012).
[17] Correlation And Logic (CORAL), http://www.insilico.eu/coral/ (accessed 9.01.2012).
[18] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31.
[19] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. 29 (1989) 97.
[20] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, J. Comput. Chem. 32 (2011) 2727.
[21] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, Chemometr. Intell. Lab. 110 (2012) 81.
[22] J.C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, Int. J. Mol. Sci. 12 (2011) (2011) 9354.
[23] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, J. Mol. Graph. Model. 31 (2011) 10.
[24] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, Chemometr. Intell. Lab. 107 (2011) 269.
[25] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, Cent. Eur. J. Chem. 9 (2011) 75.
[26] A.P. Toropova, A.A. Toropov, E. Benfenati, D. Leszczynska, J. Leszczynski, J. Math. Chem. 48 (2010) 959.
[27] A.A. Toropov, B.F. Rasulev, J. Leszczynski, Bioorg. Med. Chem. 16 (2008) 5999.
[28] A.A. Toropov, A.P. Toropova, E. Benfenati, Eur. J. Med. Chem. 45 (2010) 3581.
[29] D.J.G. Marino, P.J. Peruzzo, E.A. Castro, A.A. Toropov, Internet Electron. J. Mol. Des. 1 (2002) 115.
[30] P.J. Peruzzo, D.J.G. Marino, E.A. Castro, A.A. Toropov, Internet Electron. J. Mol. Des. 2 (2003) 334.
[31] A.A. Toropov, A.P. Toropova, Internet Electron. J. Mol. Des. 1 (2002) 108.
[32] M.V. Putz, Chem. Centr. J. 5 (2011) 29.
[33] L. Tarko, M.V. Putz, J. Theor. Comput. Chem. 11 (2012) 265.