STUDIA UBB. CHEMIA, LV, 4, 2010

THE RELATIONSHIP BETWEEN ENERGY CALCULATIONS AND BOILING POINTS OF N-ALKANES

LORENTZ JÄNTSCHI[a], SORANA D. BOLBOACĂ[a,b]

ABSTRACT. The relationship between energy calculations and boiling points was studied on a set of fourteen n-alkanes. The correlation analysis clearly showed that the best relationship is not linear. The regression analysis showed that a dose-response logistic function provided a very good agreement between the boiling points of alkanes and their heat of formation.

Keywords: *regression, correlation, alkanes, boiling point, energy*

INTRODUCTION

Boiling point, the temperature at which the vapor pressure of the liquid equals the environmental pressure surrounding the liquid [1], of organic compounds is an important property since it can provide information about other physical properties and structural characteristics [2]. Molecules with strong intermolecular forces are known to have higher boiling points [2].

The boiling point of alkanes, chemical structures with a C_nH_{2n+2} generic formula, increases with the chain length (number of carbon atoms).

The relationship between the boiling points of alkanes and other properties or descriptors have previously been studied using simple or multiple linear regression models [3-5] or non-linear models [6]. Since the boiling point of alkanes is determined by their molecular weight, this property shows a linear relationship with the size of the molecules [7]. Kozioł obtained, on a set of fourteen n-alkanes, a non-linear model with five descriptors having a determination coefficient of 0.9993 [6]. Moreover, simple exponential models estimated the critical temperature, pressure, and volume of alkanes as function of the normal boiling point and molecular weight [8].

The present study is aimed to carry out correlation and regression analyses in order to establish the relationships between the calculated energy and the boiling points of n-alkanes (an "easy to predict" property).

[a] *Technical University of Cluj-Napoca, 103-105 Muncii Bvd., RO-400641 Cluj-Napoca, Romania, lori@academicdirect.org*

[b] *"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, RO-400023 Cluj-Napoca, Romania, sbolboaca@umfcluj.ro*

RESULTS AND DISCUSSION

The results of the correlation analyses are presented in Table 1. The dipole moment property was excluded from further analyses since the Pearson correlation coefficient was of -0.0391. The analysis of the obtained correlation coefficients revealed that Spearman and Gamma correlation coefficients had higher values compared to the Pearson correlation coefficients.

Table 1. Results of correlation analysis

X (Y= boiling point)	r (p)	ρ (p)	Γ (p)
heat-of-formation	0.9515958 ($1.67 \cdot 10^{-7}$)	1*	1*
scf-binding-energy	0.9499073 ($2.05 \cdot 10^{-7}$)	1*	1*
total-energy	0.9498675 ($2.06 \cdot 10^{-7}$)	1*	1*
scf-atom-energy	0.9498641 ($2.06 \cdot 10^{-7}$)	1*	1*
scf-electronic-energy	0.9060543 ($8.09 \cdot 10^{-6}$)	1*	1*
scf-core-energy	0.8992529 ($1.21 \cdot 10^{-5}$)	1*	1*
dipole-moment	-0.0391090 (0.8943)	0.0681 (0.8094)	0.0989 (0.9618)

Correlation coefficients: r = Person; ρ = Spearman; Γ = Gamma
* $p < 10^{-7}$;

The 0.9515958 value of the Pearson correlation coefficient revealed that the linear relationship with the heat of formation was able to explain almost 91% of boiling points variation of the studied n-alkanes, which is a good estimation. Since the Spearman correlation coefficient was equal to the Gamma correlation coefficient and both of them were higher than the Pearson correlation coefficient, the relationship between boiling points and energy calculations could be non-linear.

Non-linear regression analysis was carried out in order to identify the type of relationship between the boiling points of alkanes and energy calculations. The best performing models, in terms of determination coefficients, F-value and coefficient significance proved to be of the *dose-response logistic function* type. The top three models in terms of the above-presented criteria are shown in Table 2.

The analysis of the results in Table 2 revealed that the best performing model, able to explain the boiling points of alkanes (as estimator) used the heat of formation (as predictor, H_F) through a dose-response logistic function. As it can be observed, a four-variable equation was able to fully predict the variation of boiling points as function of the heat of formation. The smallest difference between the determination coefficient and the adjusted determination coefficient was obtained using the first equation (boiling point as function of the heat of formation). The smallest value of the standard error was of 0.33°C and provided by the first equation (boiling point as function of the heat of formation). Note that the highest t-values associated to the coefficients and the smallest values of the standard errors were obtained when the boiling points were investigated as function of the heat of formation.

Table 2. Regression analysis results

Type Y	Type X	r^2	r^2_{adj}	F (FitStErr)	C	Value [95%CI]	StErr	t
DoseRspLgstc $\hat{Y} = a_0+a_1/(1+(x/a_2)^{a_3})$								
B_P	H_F	0.999997	0.999996	1090130 (0.32797)	a_0	1142.31 [1111.59; 1173.03]	13.78	82.85
					a_1	-1435.64 [-1470.43; -1400.85]	15.61	-91.94
					a_2	-191.47 [-200.82; -182.11]	4.20	-45.59
					a_3	0.7518 [0.7386; 0.7656]	0.01	121.71
B_P	T_E	0.999864	0.999823	24478 (2.18849)	a_0	-324.89 [-367.34; -282.43]	19.06	-17.05
					a_1	1836.08 [1332.98; 2339.17]	225.80	8.13
					a_2	-179833.96 [-305299; -54369]	56313	-3.19
					a_3	-0.6190 [-0.7225; -0.5155]	0.046	-13.32
B_P	SBE	0.999857	0.999814	23351 (2.24065)	a_0	-359.58 [-416.26; -302.91]	25.44	-14.14
					a_1	1925.18 [1315.38; 2534.99]	273.70	7.03
					a_2	-14657.09 [-26730; -258]	5418.9	-2.70
					a_3	-0.5950 [-0.7137; -0.4764]	0.0532	-11.15

DoseRspLgstc = dose-response logistic function;
B_P = boiling point; H_F = heat-of-formation; T_E = total-energy; SBE = scf-binding-energy;
r^2 = determination coefficient; r^2_{adj} = adjusted determination coefficient; F = F-value;
C = coefficient; 95%CI = 95% coefficient confidence interval; StErr = standard error;
t = t-value

The graphical representation of the best performing model (B_P^= (1142.31±30.72)-(1435.6±34.79)/(1+(H_F/(-191.47±9.35))$^{(0.7518\pm0.0132)}$)) is presented in Figure 1.

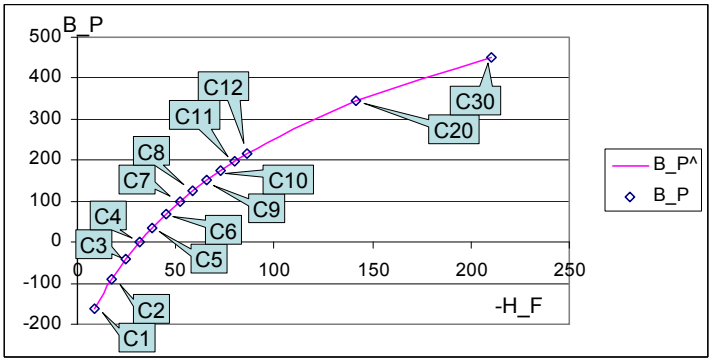

Figure 1. Boiling points of alkanes as heat of formation function

The analysis of Figure 1 revealed that the identified dose-response logistic function is the best one in estimating the relationship between the heat of formation and the boiling points of the studied n-alkanes. This statement is also supported by the value of the correlation coefficient associated to the model (see Table 2). A statistically significant linear relationship could also be identified between boiling points and the heat of formation, but this

relationship had lower performances compared to the best scoring dose-response logistic function (r^2 = 0.9062, F = 116, p = $1.6 \cdot 10^{-7}$, standard error of estimated = 52.44).

The estimated boiling points when the first equation was used (boiling point as function of the heat of formation), abbreviated as B_P^, and the measured boiling points, abbreviated as B_P, is graphically presented in Figure 2.

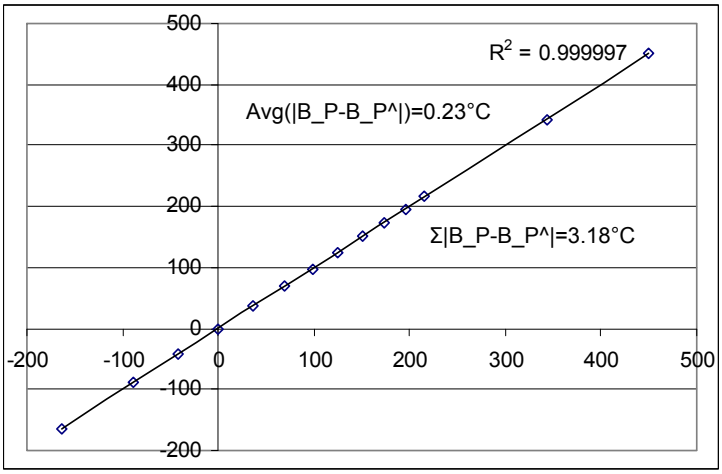

Figure 2. Estimated (horizontal) versus measured (vertical) boiling points using the dose-response logistic function

The validity and reliability of the best performing relationship obtained in the study on n-alkanes is supported by the smallest value of the absolute value of residuals (equal to 0.23°C) and by the sum of the absolute difference of residuals (equal to 3.18°C) (Figure 2). Moreover, the sum of residuals was 0.01°C while the squared sum of residuals was 1.08.

The objective of this research was met as soon as the best model able to estimate the boiling points of alkanes as functional dependence on energy calculations was identified. The value of the Person correlation coefficient, which proved to be smaller in comparison to the Spearman and Gamma correlation coefficients, determined the investigation of non-linear relationships even if the linear relationship was statistically significant. A dose response logistic function proved to better explain the boiling points as function of energy calculations for the studied n-alkanes when the molecules were prepared for analysis by applying the mm+ as molecular mechanics and the AM1 as semi-empirical method.

CONCLUSIONS

If ρ^2(Spearman), Γ^2(Gamma) >r^2(Pearson), the relationship between variables is not linear; non-linear relationships must always be checked. Thus, the best performing relationship between boiling points and the energy calculations of the investigated n-alkanes was expected not to be linear.

A functional dependence was identified between boiling points and the energy calculations of the investigated n-alkanes. This functional dependence proved to be a dose-response logistic function when mm+ molecular mechanics and AM1 semi-empirical methods were used to prepare the studied n-alkanes for analysis.

The following model was identified as the model with the highest performance:

$$B\_P\text{^} = (1142.31\pm30.72)-(1435.6\pm34.79)/(1+(H\_F/(-191.47\pm9.35))^{(0.7518\pm0.0132)}),$$

where B_P^ is the estimated boiling point and H_F is the heat of formation. The validity of the model is supported by the small value of the standard error, the high F-value and the small p-value.

EXPERIMENTAL SECTION

Fourteen normal alkanes (C_1-C_{12}, C_{20}, C_{30}), chemical compounds consisting of carbon and hydrogen elements, were analyzed (see Table 3).

Table 3. Characteristics of alkanes: boiling point, dipole-moment, total-energy, atom-energy, binding-energy, core-energy, electronic-energy, and heat-of-formation

Name	Formula	B_P	D_M	T_E	SAE	SBE	SCE	SEE	H_F
Methane	CH_4	-164	$1.12\cdot10^{-6}$	-4225	-3837	-388	4619	-8844	-9
Ethane	C_2H_6	-89	$6.87\cdot10^{-7}$	-7821	-7149	-672	13638	-21459	-18
Propane	C_3H_8	-42	$4.28\cdot10^{-3}$	-11415	-10461	-954	26313	-37727	-24
Butane	C_4H_{10}	-0.5	$1.01\cdot10^{-7}$	-15008	-13773	-1236	41607	-56615	-31
Pentane	C_5H_{12}	36	$6.28\cdot10^{-3}$	-18602	-17084	-1518	59034	-77636	-38
Hexane	C_6H_{14}	69	$3.06\cdot10^{-7}$	-22196	-20396	-1800	78191	-100387	-45
Heptane	C_7H_{16}	98	$6.57\cdot10^{-3}$	-25790	-23708	-2082	98835	-124624	-52
Octane	C_8H_{18}	125	$1.52\cdot10^{-7}$	-29383	-27020	-2364	120757	-150141	-59
Nonane	C_9H_{20}	151	$6.65\cdot10^{-3}$	-32977	-30331	-2646	143819	-176796	-66
Decane	$C_{10}H_{22}$	174	$3.95\cdot10^{-7}$	-36571	-33643	-2928	167892	-204463	-73
Undecane	$C_{11}H_{24}$	196	$8.13\cdot10^{-3}$	-40165	-36955	-3210	192888	-233052	-80
Dodecane	$C_{12}H_{26}$	216	$1.35\cdot10^{-7}$	-43758	-40267	-3492	218724	-262482	-86
Eicosane	$C_{20}H_{42}$	343	$8.61\cdot10^{-7}$	-72508	-66760	-5748	449165	-521673	-142
Triacontane	$C_{30}H_{62}$	450	$1.59\cdot10^{-6}$	-108445	-99878	-8567	779447	-887893	-210

B_P = boiling point; D_M = dipole-moment; T_E = total-energy;
SAE = scf-atom-energy; SBE= scf-binding-energy; SCE = scf-core-energy;
SEE = scf-electronic-energy; H_F = heat-of-formation.

Eight properties of the above-mentioned alkanes were investigated: boiling point [°C] [9], total-energy (T_E) [kcal/mol], dipole-moment (D_M) [Debyes], scf-atom-energy (SAE) [kcal/mol], scf-binding-energy (SBE) [kcal/mol], scf-core-energy (SCE) [kcal/mol], scf-electronic-energy (SEE) [kcal/mol], and heat-of-formation (H_F) [kcal/mol]. Except for the boiling points, all the other properties were calculated with HyperChem v. 8.0 using the following criteria: optim-converged=true, molecular mechanics method: mm+ [10], and semi-empirical method: AM1 [11].

Correlation and regression analyses were carried out in order to meet the objective of the study. Pearson ("r") [12], Spearman ("ρ") [13] and Gamma ("Γ") [14] correlation coefficients were used to find the power and the sign of the relationship between boiling points and the investigated properties.

Regression analyses were carried out with the SlideWrite Plus software. The following possibilities of regression search were used:

- *Linear:* ▪ Linear Group; ▪ Exponential Group; ▪ Power Group; ▪ Polynomial Group.
- *Nonlinear:*
  - *Standard*: ▪ User-Defined (any function defined by the user); ▪ Exponential – $Y=a_0+a_1*\exp(-x/a_2)$; ▪ Power - $Y=a_0+a_1*x\text{^}a_2$.
  - *Transitional*: ▪ 1-Site Ligant – $Y=a_0*x/(a_1+x)$;
    ▪ Cumulative – $Y=a_0+a_1*0.5*(1+\mathrm{erf}((x-a_2)/\sqrt{(2)}*a_3))$;
    ▪ DoseRspLgstc - $Y=a_0+a_1/(1+(x/a_2)\text{^}a_3)$;
    ▪ Photosynthesis - $Y=a_0*a_1*x/(a_0+a_1*x)$;
    ▪ PH Activity – $Y=(a_0+a_1*10\text{^}(x-a_2))/(1+10\text{^}(x-a_2))$;
    ▪ Sigmoidal – $Y=a_0+a_1/(1+\exp(-(x-a_2)/a_3))$.
  - *Peak*: ▪ Erfc Peak, Gaussian – $Y=a_0+a_1*\exp(-0.5*((x-a_2/a_3)^2)$;
    ▪ Logistic Peak – $Y=a_0+a_1*4*(\exp(-(x-a_2)/a_3))/(1+\exp(-(x-a_2)/a_3))^2$;
    ▪ Log-Normal – $Y=a_0+a_1*\exp(-0.5*(\ln(x/a_2)/a_3)^2)$;
    ▪ Lorentzian – $Y=a_0+a_1/(1+((x-a_2)/a_3)^2)$.
  - *Waveform*: ▪ SineWave – $Y=a_0+a_1*\sin(2*pi*x/a_3+a_2)$; SineWaveSquared – $Y=a_0+a_1*(\sin(2*pi*x/a_3+a_2))^2$
- *User-Defined:* allows to define any equation with a maximum of 7 coefficients.

ACKNOWLEDGMENTS

Financial support is gratefully acknowledged to CNCSIS-UEFISCSU Romania (project PNII-IDEI1051/202/2007).

REFERENCES

1. D.E. Goldberg, 3,000 Solved Problems in Chemistry (1st ed.). McGraw-Hill. ISBN 0-07-023684-4. Section 17.43, 1988, pp. 321.
2. M.R. Riazi, T.E. Daubert, *Hydrocarbon Processing,* **1980**, *59(3)*, 115.
3. E.S. Souza, C.A. Kuhnen, B.S. Junkes, R.A. Yunes, V.E.F. Heinzen, *Journal of Chemometrics,* **2010**, *24(3-4)*, 149.
4. D. Ciubotariu, M. Medeleanu, V. Vlaia, T. Olariu, C. Ciubotariu, D. Dragos, S. Corina, *Molecules,* **2004**, *9(12)*, 1053.
5. S.D. Bolboacă, L. Jäntschi, *International Journal of Pure and Applied Mathematics,* **2008**, *47(1)*, 23.
6. J. Kozioł, *Polish Journal of Chemistry,* **2009**, *83(12)*, 2173.
7. R.T. Morrison, R.N. Boyd. Organic Chemistry, 6th ed. (New Jersey: Prentice Hall, 1992.
8. F. Vejahati, M.B. Nikoo, S. Mokhatab, B.F. Towler, *Petroleum Science and Technology,* **2007**, 25(9), 1115.
9. C.E. Ophardt. *Virtual Chembook*, Elmhurst College, Elmhurst, IL, USA, 2003.
10. A. Hocquet, M. Langgard, *Journal of Molecular Modeling,* **1998**, *4*, 94.
11. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, *Journal of the American Chemical Society,* **1985**, *107*, 3902.
12. K. Pearson, *Philosophical Magazine,* **1900**, *50*, 157.
13. C. Spearman, *American Journal of Psychology*, **1904**, *15*, 201.
14. L.A. Goodman, W.H. Kruskal, *Journal of the American Chemical Society,* **1963**, *58*, 310.

STUDIA UBB. CHEMIA, LV, 4, 2010

DIAGNOSTIC OF A QSPR MODEL: AQUEOUS SOLUBILITY OF DRUG-LIKE COMPOUNDS

SORANA D. BOLBOACĂ[a,b], LORENTZ JÄNTSCHI[b]

ABSTRACT. A diagnostic test for a qSPR (quantitative Structure-Property Relationship) model was carried out using a series of statistical indicators for correctly classifying compounds into actives and non-actives. A previously reported qSPR model, able to characterize the aqueous solubility of drug-like compounds, was used in this study. Eleven statistical indicators like those used in medical diagnostic tests were defined and applied on training, test and overall data sets. The associated 95% confidence interval under the binomial distribution assumption was also computed for each defined indicator in order to allow a correct interpretation. Similar results were obtained in the training and test sets with some exceptions. The prior probabilities of active and non-active compounds proved not to be significantly different in the training and test sets. However, the probability of classification as active compounds proved to be significantly smaller in the training set as compared to the test set ($p = 0.0042$). The total fraction of correctly classified compounds proved to be identical in the training and test sets as well as in the overall set. Nevertheless, the overall model and the model obtained in the test set show a higher ability to correctly assign the non-active compounds to the non-active class while the model obtained in the training set has a higher ability to correctly assign the active compounds to the active class.

Keywords: *quantitative Structure-Property Relationships (qSPR), diagnostic parameters, 2×2 contingency table, solubility, drug-like compounds*

INTRODUCTION

Quantitative structure-property relationships (qSPRs) procedures able to quantitatively correlate the chemical structure with a defined property [1], are widely used in drug design [2,3], drug classification [4,5] and screening [5,6].

A series of studies were drawn in order to establish the validation methods of a qSPR model [7,8], including the principle of parsimony, selection of the simplest model, cross-validation, Y scrambling and external predictability [9]. Various procedures for variable selection have been created [10-13] and statistical

[a] *"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, RO-400023 Cluj-Napoca, Romania, sbolboaca@umfcluj.ro*

[b] *Technical University of Cluj-Napoca, 28 Memorandumului, RO-400114 Cluj-Napoca, Romania, lori@academicdirect.org*

analysis of molecular similarity matrices was developed in order to identify the best quantitative structure-activity relationships [14]. Reliability and accuracy have also been introduced for the validation of QSPR models [15,16]. The information criteria (Akaike's information criteria - AIC [17], corrected AIC [18], Schwarz (or Bayesian) Information Criterion – BIC, Amemiya Prediction Criterion – APC, and Hannan-Quinn Criterion - HQC) and Kubinyi's function [19, 20] are the parameters used to compare different qSPR/qSAR models [21-23].

The aim of this study was to carry out a diagnostic test on a qSPR (quantitative Structure-Property Relationships) model, by using a series of statistical indicators for correctly classifying compounds into actives and non-actives.

RESULTS AND DISCUSSION

Eleven statistical indicators were proposed as diagnostic parameters for qSPR models. The contingency tables used to calculate these parameters are presented in Table 1. The statistical indicators computed for the training, test and overall data sets are presented in Table 2 – 4.

Table 1. 2×2 contingency tables for the investigated qSPR model

Generic Table	Observed		
Estimated	+	-	Σ
+	TP	FP	
-	FN	TN	
Σ			n

Test Set	Observed		
Estimated	+	-	Total
+	26	10	36
-	4	29	33
Total	30	39	69

Training Set	Observed		
Estimated	+	-	Total
+	28	7	35
-	12	48	60
Total	40	55	95

Overall	Observed		
Estimated	+	-	Total
+	54	22	76
-	11	77	88
Total	65	99	164

+ = active class; - = non-active class;
Estimated = aqueous solubility estimated by Duchowitz's et al. qSPR model

The chi-squared test was applied on contingency tables in order to test the null hypotheses that the estimated class (active and non-active) is independent from the observed class (active and non-active). The value of the chi-squared statistics and associated significance level, presented at the bottom of Tables 2 - 4, supported the rejection of the null hypotheses that the estimated classification into active and non-active compounds is unrelated to the observed classification. These results sustain the ability of the qSPR model to classify compounds as actives and non-actives. The degree of association between the estimated and the observed classification of compounds proved to be a positive and moderate one, in all the investigated sets (training, test and overall set of studied compounds). The moderate association, expressed as the Φ contingency correlation coefficient, revealed that the reported qSPR [24] is not a perfect model.

Table 2. Statistical indicators for assessing the qSPR model: training set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	80.00	[71.07-87.02]
Error Rate (ER)	20.00	n.a.
Prior proportional probability of an active class	0.4211	[0.3254-0.5215]
Prior proportional probability of a non-active class	0.5789	n.a.
Sensitivity (Se)	70.00	[54.76-82.39]
False-negative rate (under-classification, FNR)	30.00	[17.61-45.24]
Specificity (Sp)	87.27	[76.39-93.96]
False-positive rate (over-classification, FPR)	12.73	[6.04-23.61]
Positive predictivity (PP)	80.00	[64.55-90.44]
Negative predictivity (NP)	80.00	[68.52-88.49]
Probability of classification		
- as active (PCA)	0.3684	[0.2766-0.4682]
- as non-active (PCIC)	0.6316	[0.5318-0.7234]
Probability of a wrong classification		
- as active compound (PWCA)	0.2000	[0.0956-0.3545]
- as non-active compound (PWCI)	0.2000	[0.1151-0.3148]
Odds Ratio (OR)	16.0000	[5.7090-45.0262]

95% CI = confidence interval at a significance level of 5%; n.a. = not available;
χ^2 = 30.2305 ($p < 0.0001$) (Chi-squared statistics); Contingency correlation coefficient Φ = 0.5641

Table 3. Statistical indicators for assessing the qSPR model: test set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	79.71	[69.04-87.79]
Error Rate (ER)	20.29	n.a.
Prior proportional probability of an active class	0.4348	[0.3225-0.5524]
Prior proportional probability of a non-active class	0.5652	n.a.
Sensitivity (Se)	86.67	[70.96-95.08]
False-negative rate (under-classification, FNR)	13.33	[4.92-29.04]
Specificity (Sp)	74.36	[59.21-85.91]
False-positive rate (over-classification, FPR)	25.64	[14.09-40.79]
Positive predictivity (PP)	72.22	[56.25-84.67]
Negative predictivity (NP)	87.88	[73.29-95.52]
Probability of classification		
- as active (PCA)	0.5217	[0.4050-0.6367]
- as non-active (PCIC)	0.4783	[0.3633-0.5950]
Probability of a wrong classification		
- as active compound (PWCA)	0.2778	[0.1533-0.4375]
- as non-active compound (PWCI)	0.1212	[0.0448-0.2671]
Odds Ratio (OR)	18.8500	[5.4919-64.5994]

95% CI = confidence interval at a significance level of 5%; n.a. = not available;
χ^2 = 22.9206 ($p < 0.0001$) (Chi-squared statistics); Contingency correlation coefficient Φ = 0.5764

The accuracy of the qSPR model proved to be almost 80% in all the investigated sets of compounds. The accuracy of the model in the training set proved not to be statistically different from the accuracy of the model in the test set (the confidence intervals overlap, see Tables 2 and 3). A similar interpretation is true when the values and associated confidence intervals of other statistical indicators are analyzed (see Tables 2 -4).

Table 4. Statistical indicators for assessing the qSPR model: overall set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	79.88	[73.22-85.43]
Error Rate (ER)	20.12	n.a.
Prior proportional probability of an active class	0.3963	[0.3238-0.4725]
Prior proportional probability of an non-active class	0.6037	n.a.
Sensitivity (Se)	83.08	[72.50-90.55]
False-negative rate (under-classification, FNR)	16.92	[9.45-27.50]
Specificity (Sp)	77.78	[68.82-85.05]
False-positive rate (over-classification, FPR)	22.22	[14.95-31.18]
Positive predictivity (PP)	71.05	[60.19-80.30]
Negative predictivity (NP)	87.50	[79.26-93.06]
Probability of classification		
- as active (PCA)	0.4634	[0.3883-0.5398]
- as non-active (PCIC)	0.5366	[0.4602-0.6117]
Probability of a wrong classification		
- as active compound (PWCA)	0.2895	[0.1970-0.3981]
- as non-active compound (PWCI)	0.1250	[0.0694-0.2074]
Odds Ratio (OR)	17.1818	[7.7989-38.1475]

95% CI = confidence interval at a significance level of 5%; n.a. = not available
$\chi^2 = 83.6385$ ($p < 0.0001$) (Chi-squared statistics); Contingency correlation coefficient $\Phi = 0.5761$

The *Z* test was applied in order to compare the statistical indicators expressed as probabilities obtained in training and test sets. The prior probabilities of active and non-active compounds proved not to be statistically different in training and test sets. The absence of statistically significant differences between prior probabilities of active and non-active compounds in training and test sets supports the correct assignment of compounds to the active/non-active sets. However, the probability of classification as active compounds proved to be statistically smaller in the training set compared to the test set ($p=0.0042$); thus, the classification model proved to perform better in terms of correct classification of active compounds when applied on test set.

The objective of this study is to propose a series of statistical indicators as diagnostic tools for the qSPR model. In achieving this, various aspects are considered:

- Analyzing the correct assignment of compounds to training and test sets: prior proportional probability of an active class & prior proportional probability of a non-active class
- Analyzing the correct classification of active and non-active compounds: all the other statistical indicators (see Table 2-4).

The proposed statistical indicators have to assess the qSPR model in training and test sets: as the indicators have similar performances in training and test sets, it could involve the model has similar classification abilities, thus being considered as a good model. The best model is the one with the highest possible accuracy and the smallest possible error rate. The best model is also the one with the highest sensitivity and specificity and the smallest false-negative and false-positive rates. In this respect, it can be observed that sensitivity is smaller than specificity in the training set while sensitivity is

higher than specificity in the test set (see Tables 2 and 3). In other words, the investigated qSPR model has a higher ability to correctly assign active compounds to the active class in the test set and a higher ability to correctly assign non-active compounds to the non-active class in the training set. An excellent classification model should also have the best possible positive and negative predictability values while the probability values of a wrong classification into active and non-active compounds should have the smallest possible values.

Similar statistical parameters are used to assess the performances of machine learning classification models: accuracy, recall (true positive rate, false positive rate, true negative rate, false negative rate, and precision) [25, 26]. These parameters are calculated based on the confusion matrix [27]. Note that the confusion matrix is the same as the generic contingency table presented in Table 1.

The present study is aimed to introduce a series of statistical indicators in order to diagnose a qSPR model. Useful information related to the assignment of compounds in the training and test sets could be obtained by using prior proportional probability of an active class & prior proportional probability of a non-active class. All the other proposed statistical indicators allow the characterization of a qSPR model in terms of total fraction of correctly classified compounds (accuracy), correct assignment to active or non-active class (sensitivity and specificity, false positive and false negative rates), etc. Statistical indicators were applied on a 2×2 confusion matrix but the same approach could also be applied on r×c confusion matrices when compounds are classified into more than two groups (e.g., non-active, active, and very active). The usefulness of this approach in diagnosing qSPR/qSAR models is currently investigated in our laboratory.

CONCLUSIONS

The total fraction of compounds correctly classified by the qSPR model proved to be identical in the training and test sets as well as in the overall set. However, the overall model and the model obtained in the test set showed a higher ability to correctly assign the non-active compounds to the negative class while the model obtained in the training set had a higher ability to correctly assign the active compounds to the active class.

EXPERIMENTAL SECTION

A previously reported qSPR model [28] able to characterize the aqueous solubility of drug-like compounds was herein used. The experimental aqueous solubility measured at 298K and expressed in mg/ml (values taken from Merck Index 13th [28]) was modeled using molecular descriptors [24].

The best model obtained in the training set (n=97) proved to be a model with 3 descriptors and the following characteristics [24]:

$R^2 = 0.871$; $S = 0.903$
$R^2_{loo} = 0.849$; $S_{loo} = 0.971$
$R^2_{val} = 0.848$; $S_{val} = 0.899$

where R^2 = determination coefficient; S = standard deviation of the model; R^2_{loo} = determination coefficient on leave one out analysis; S_{loo} = standard deviation on leave-one-out analysis; R^2_{val} = determination coefficient on validation set; S_{val} = standard deviation on validation set.

A series of statistical indicators similar with those used in medical diagnostic tests [29, 30] were defined as diagnostic parameters for the qSPR model (Table 5).

The experimental and estimated aqueous solubility of the studied compounds was transformed as dichotomial variables in order to calculate the defined statistical indicators (Table 5) using the following criteria: if experimental data ≥ 0, the compound was considered active, if experimental data < 0, the compound was considered non-active.

Table 5. Statistical indicators calculated on the 2×2 contingency table

Indicator (Abbreviation)	Formula	Definition
Accuracy / Non-error Rate (AC)	100*(TP+TN)/n	Total fraction of correctly classified compounds
Error Rate (ER)	100* (FP+FN)/n = 1-CC	Total fraction of misclassified compounds
Prior proportional probability of a class (PPP)	n_i/n	Fraction of compounds belonging to class *i*
Sensitivity (Se)	100*TP/(TP+FN)	Percentage of active compounds correctly assigned to the active class
False-negative rate (under-classification, FNR)	100*FN/(TP+FN) = 1-Se	Percentage of active compounds falsely assigned to the non-active class
Specificity (Sp)	100*TN/(TN+FP)	Percentage of non-active compounds correctly assigned to the non-active class
False-positive rate (over-classification, FPR)	100*FP/(FP+TN) = 1-Sp	Percentage of non-active compounds falsely assigned to the active class
Positive predictivity (PP)	100*TP/(TP+FP)	Percentage of compounds correctly assigned to the active class out of all compounds assigned to the active class
Negative predictivity (NP)	100*TN/(TN+FN)	Percentage of compounds correctly assigned to the non-active class out of all compounds assigned to the non-active class
Indicator (Abbreviation)	**Formula**	**Definition**
Probability of classification - as active (PCA)	(TP+FP)/n	- Probability to classify a compound as active (true positive & false

- as inactive (PCIC)	(FN+TN)/n	positive) - Probability to classify a compound as non-active (true negative & false negative)
Probability of a wrong classification - as active compound (PWCA) - as non-active compound (PWCI)	 FP/(FP+TP) FN/(FN+TN)	 Probability of a false positive classification Probability of a false negative classification
Odds Ratio (OR)	(TP*TN)/(FP*FN)	The odds of correct classification in the group of active compounds divided to the odds of an incorrect classification in the group of non-active compounds

The associated 95% confidence interval under the binomial distribution assumption [31] was also computed for the correct interpretation of the indicators [32].

ACKNOWLEDGMENTS

Financial support is gratefully acknowledged to CNCSIS-UEFISCSU Romania (project PNII-IDEI458/206/2007).

REFERENCES

1. L.P. Hammett, *Chemical reviews,* **1935**, *17*, 125.
2. I.M. Kapetanovic, *Chemico-Biological Interactions,* **2008**, *171(2)*, 165.
3. C.H. Andrade, K.F. Pasqualoto, E.I. Ferreira, A.J. Hopfinger, *Molecules,* **2010**, *15(5)*, 3281.
4. V. Potemkin, M. Grishina, *Drug Discov Today,* **2008**, *13(21-22)*, 952.
5. J. Li, P. Gramatica, *Journal of Chemical Information and Modeling,* **2010**, *50(5)*, 861.
6. M.C. Hutter, *Current Medicinal Chemistry,* **2009**, *16(2)*, 189.
7. M. Pavan, T.I. Netzeva, A.P. Worth, *SAR and QSAR in Environmental Research,* **2006**, *17(2)*, 147.
8. S. Wold, *Quantitaive Structure-Activity Relationship,* **1991**, *10*, 191.
9. R.D. Cramer III, J.D. Bunce, D.E. Patterson, I.E. Frank, *Quantitaive Structure-Activity Relationship,* **1988**, *7*, 18; Erratum **1988**, *7*, 91.
10. H. Kubinyi, *Quantitaive Structure-Activity Relationship,* **1994**, *13*, 285.
11. H. Kubinyi, *Quantitaive Structure-Activity Relationship,* **1994**, *13*, 393.
12. K. Héberger, *TrAC - Trends in Analytical Chemistry,* **2010**, *29(1)*, 101.
13. P. Ghosh, M.C. Bagchi, *Current Medicinal Chemistry,* **2009**, *16(30)*, 4032.
14. A.C. Good, S.J. Peterson, W.G. Richards, *Journal of Medicinal Chemistry,* **1993**, *36(20)*, 2929.

15. B.H. Su, M.Y. Shen, E.X. Esposito, A.J. Hopfinger, Y.J. Tseng, *Journal of Chemical Information and Modeling,* **2010**, *50(7)*, 1304.
16. L.G. Valerio Jr., *Toxicology and Applied Pharmacology,* **2009**, *241(3)*, 356.
17. H. Akaike, *Annals of the Institute of Statistical Mathematics,* **1969**, *21*, 243.
18. C.M. Hurvich, C. Tsai, *Biometrika,* **1989**, *76*, 297.
19. H. Kubinyi, *Quantitative Structure-Activity Relationships,* **1994**, *3*, 393.
20. H. Kubinyi, *Quantitative Structure-Activity Relationships,* **1994**, *13*, 285.
21. S.D. Bolboacă, L. Jäntschi, *TheScientificWorldJOURNAL*, **2009**, *9(10)*, 1148.
22. S.D. Bolboacă, M.M. Marta, C.E. Stoenoiu, L. Jäntschi, *Applied Medical Informatics*, **2009**, *25(3-4)*, 65.
23. S.D. Bolboacă, M.M. Marta, L. Jäntschi, *Folia Medica,* **2010**, *52(3)*, 37.
24. P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, *Bioorganic & Medicinal Chemistry,* **2008**, 16(17), 7944.
25. A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, E. Benfenati, *Chemistry Central Journal,* **2010**, *4 Suppl 1*, S1.
26. N. Fjodorova, M. Vracko, M. Novic, A. Roncaglioni, E. Benfenati, *Chemistry Central Journal,* **2010**, Jul 29; 4 Suppl 1:S3.
27. M. Kubat, S. Matwin, Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186, Nashville, Tennesse. Morgan Kaufmann, **1997**.
28. The Merck Index An Encyclopedia of Chemicals, Drugs, and Biologicals; Merck & Co.: NJ, **2001**.
29. S.D. Bolboacă, L. Jäntschi *Electronic Journal of Biomedicine,* **2007**, *2*, 19-28.
30. S.D. Bolboacă, L. Jäntschi, A. Achimaş Cadariu, *Applied Medical Informatics,* **2004**, *14*, 27.
31. S.D. Bolboacă, L. Jäntschi, *International Journal of Pure and Applied Mathematics*, **2008**, *47(1)*, 1.
32. L. Jäntschi, S.D. Bolboacă, *TheScientificWorldJOURNAL*, **2010**, *10*, 865.