The Effect of Leverage and Influential on Structure-Activity Relationships

Sorana D. BOLBOACĂ[1] and Lorentz JÄNTSCHI[2,3,*]

[1]"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Cluj, Romania.
[2]Technical University of Cluj-Napoca, Department of Chemistry, 103-105 Muncii Bdv., 400641 Cluj-Napoca, Romania.
[3]University of Agricultural Science and Veterinary Medicine Cluj-Napoca, 3-5 Calea Mănăştur, 400372 Cluj-Napoca, Romania.
E-mails: sbolboaca@umfcluj.ro; lorentz.jantschi@gmail.com

*Author to whom correspondence should be addressed; E-Mail: lorentz.jantschi@gmail.com;
Tel.: +4-0264-401-775; Fax: +4-0264-401-768.

Running title: Leverage and Influential on QSARs

The Effect of Leverage and Influential on Structure-Activity Relationships

Abstract

Quantitative Structure-Activity Relationship approaches have established as the main computational molecular modeling method. In spirit of reporting valid and reliable models the aim of our research was to assess how the analysis of leverage with Hat matrix (h_i) and of the influential using Cook's distance (D_i) of QSAR models reflects in the model reliability and its characteristics. The datasets included in this research was collected from previously published manuscripts. Seven datasets accomplished the imposed inclusion criteria and were analyzed. Three models were obtained for each dataset (full-model, h_i-model and D_i-model) and several validation criteria (statistical criteria) were defined to assess and to compare the model. The analysis of the obtained results revealed that in 5 out of 7 sets the correlation coefficient increase when both compounds with h_i and respectively D_i higher than thresholds were removed. The number of withdrawn compounds varied from 2 to 4 for h_i-model and from 1 to 13 for D_i-model. The analysis of validation statistics showed that D_i-models obtained systematically better results compared to both full-models and h_i-models. Identification of influential compound in data set could significantly improve the model and should be conducted any time when a regression analysis is desired. Cook's distance approach is recommended to be used to identify influential compounds in dataset whenever the linear regression analysis for QSAR models is applied.

Keywords: model sensitivity; quantitative structure-activity relationship (QSAR); leverage (h_i); Cook's distance (D_i); model validation.

Introduction

Translation of structural features of chemical compounds in the activity by incorporation of physico-chemical mechanisms into statistical models led to development of QSAR/QSPR (Quantitative Structure-Activity/Property Relationship) computational molecular modeling methodologies. In view of the fact that the capabilities of collecting and storing (such as PubChem) from one hand and analyzing of data from other hand due to rapid development of information and communication technologies have significant increased, QSAR modeling could be seen as an approach of statistical analyses as well as application of data-mining.

Guidance regarding the correct procedures in QSAR development has been published in scientific literature [1-3].The detailed description of QSAR modeling techniques, methodologies and trends is beyond the aim of the present manuscript. It is well known that the main characteristic of a QSAR model is its predictivity, translated in how well the model is able to predict the activity on compounds not used to develop the model. Guidelines for validation of QSAR models have been developed by experts [4-6]. Beside good practice principles, other QSARs problems were addressed by researchers. Mekenyan and Veith [7] pointed out two general problems of QSAR: various environments used to study the property/activity and proliferation of molecular descriptors. Dearden and co-authors identified 21 types of errors in QSAR modeling, errors classified according to OECD principles [8]. From statistical point of view, the identified errors were as follow [8]:

- Collinearity of molecular descriptors which is mainly reflected in the instability of the regression coefficients [1,2].
- Outlier detection and removal. Removal of a significant outlier led to a more significant model [9].
- Lack of/inadequate statistics. In most of published QSAR models, neither considerations of linear regression assumptions nor considerations of distribution of residuals are addressed [10,11]. Recommended statistics are as follow: n (sample size), r^2 or R^2 (determination coefficient), q^2 or Q^2 (determination coefficient in leave-one-out analysis); R^2_{adj} (adjusted determination coefficient), s (standard error of estimate - measure of error) and F-statistics (including p-value) [8]. Moreover, other methods of error are recommended: standard deviation, root mean square error and mean absolute error (ignore the sign of an error – provide information about random error [12]), mean error (consider the sign of an error – very low value indicates the absence of systematic error [12]; similar mean error and absolute mean error indicate the presence of systematic errors).
- Misuse/Misinterpretation of statistics. The application of linear regression technique without investigation of its assumption is one of most frequent misuse of statistics [13]. The inclusion in the model of additional independent variable(s) is another example [14].

Staying in the field of statistics for QSAR/QSPR models the following was the hypothesis of the present research: Model sensitivity analysis translated through influential point(s) could identify a stable and reliable QSAR/QSPR. Our aim was to assess how the analysis of leverage and influential using Cook's distance of QSAR models reflects in the model reliability and its characteristics.

Materials and Methods

Data sets

Several datasets previously published in International Journal of Molecular Science (MDPI Publishing, Basel, Switzerland) were included in our analysis. The search was conducted on April 2012 using the following search strategy:

Where? (Field)	What?
Title/Keyword	QSAR OR Quantitative Structure-Activity Relationship
Journal	IJMS
Article Type	Article OR Review
Time period	2000 to date

There were included in the study the dataset available in the previously published manuscripts that respected the following inclusion criteria: ① quantitative continues dependent variable AND ② values of descriptors provided in manuscript or supplementary material(s) AND ③ sample size > 20 AND ④ simple/multiple linear regression model with determination coefficient higher than or equal to 0.6.

Analysis of Influential

Model sensitivity in linear regression analysis refers to how estimates are affected by subgroups of the data. Three main issues could be used to assess the model sensitivity: residuals (large value identify the outliers), leverage (large value identify the point significantly far from the center point of the predictor space) and influential (large effect on an estimate) but just two of them are addressed in the present research.

The following steps were applied to accomplish the aim of the research:

- ***Step 1:*** Test the normality of observed/measured activity using Kolmogorov-Smirnov [15] and/or Chi-Square goodness-of-fit [16] → If data normal construct the SLM (Simple Linear Regression) / MLR (Multiple linear regression)
- ***Step 2:*** Identify the bets SML / MLR model → If $R^2 < 0.5$ STOP analysis. The dataset is removed from further analysis.

- ***Step 3:*** Identify the influential using:
    a. Hat matrix - leverage (h_i). Leverage are "a measure of the geometric distance of the i^{th} predictor point (X_{i1}, X_{i2}, ..., X_{ik}) from the center point of the predictor space" [17]. The formula applied to identify the leverage was: $h_i = 1/n+(x_i-x_m)^2/sum[(x_i-x_m)^2]$, where h_i = leverage of the i^{th} compound, n = sample size, x_i = the value of predictor variable for the i^{th} compound, x_m = the average mean for predictor x. The leverage indicates those compounds that may have potential influence in the model being used also as applicability domain of the QSAR models [18,19]. The leverage threshold (h_t) was set to 2*(k+1)/n for regression models with intercept and 2*k/n for models without intercept (where k= number of descriptors in the model; n=sample size) [17]. → If $h_i > h_t$ withdrawn the influential till no leverage exceed the threshold value or no improvement in the determination coefficient is observed.
    b. Cook's distance (D_i). Cook's distance combines residual and leverage in one indicator to identify influential in regression models [20,21]. Any compound was considered as influential if $D_i > 4/n$ (where n = sample size) [22]. → If $D_i > 4/n$ withdrawn the influential till no exceed of the threshold value is observed or no increase in the determination coefficient is observed.
- ***Step 4:*** Construct and evaluate the final SLM / MLR. The criteria used for assessment and validation of QSAR models are presented in Table 1. The correlated correlation analysis was apply to test if correlation coefficients obtained by full-model, h_i-model and D_i-model are statistically significant different at a significance level of 5% [23].
- ***Step 5:*** Take two sets of compounds and split the dataset in training (~2/3 compounds) and test set using a simple random approach [24] (leave-many-out analysis) in order to assess the behavior of the full-model and respectively model with higher correlation coefficient and smaller standard error.

To test the overall performances of leverage and influential withdrawn on QSAR models compared to full-model the Fisher's Chi-Squared (abbreviated as F-C-S) was applied [32]. The F-C-S- text was applied to test the following null hypothesis "The correlation coefficient on a specific model (such as h_i-model or D_i-model) is statistically higher compared to another model (full-model or h_i-model when D_i-model was compared to h_i-model)".

Table 1. Criteria for validation of regression models.

Criterion	Interpretation/Remark
Goodness-of-fit	
R^2 = determination coefficient	A descriptive measure. It does not measure the quality of the regression model. The higher the better
R^2_{adj} = adjusted determination coefficient	Its value decrease if an added predictor does not reduce the unexplained variance Used as a measure of usefulness of introducing a new variable in the model Closeness to the R^2 the better
R^2_{loo} = determination coefficient in leave-one-out analysis [25]	Internal validity of the model Underestimates the true predictive error when small samples are used to develop the model [26] Closeness to the R^2 the better
s = standard error of estimate	Measure of the dispersion around the regression line of observed values Smaller the better
s_{loo} = standard error of predicted	
F-value (p-value)	Ration between explained and unexplained variance of a given number of df – degrees of freedom p-value associated to F-value as significance of the level of correlation [27] The higher the better
F_{loo} (p-value)	
t-value (p-value)	Significance of the coefficients in the regression model t-value - the higher the better vs. p-value - the lower the better
Validation statistics	
RMS = residual mean square	Error variance The lower the better
APV = average prediction variance [28]	The lower the better
TSE = total squared error [29]	The lower the better
APMSE = Average Prediction Mean Squared Error [30]	The lower the better
%PredErr = percentage prediction error [31]	Defined as prediction error (module of the difference between observed and estimated) divided by the highest activity
Predictive Power – Fisher's approach (t_{PP} - p_{PP})	Evaluate if the mean of residuals is statistically different by the expected mean (where expected mean = 0); p_{PP}: the lower the better
RMSE = root-mean-square error	Measures the average magnitude of the error The lower the better
MAE = mean absolute error	Measures the average magnitude of the errors Could be also used to compare two models - The lower the better
MAPE = mean absolute percentage error	The lower the better
SEP = standard error of prediction	The lower the better
REP% = relative error of prediction	The lower the better

Results

Sixty-four manuscripts were identified using the applied search strategy. Fifteen manuscripts provided the experimental/observed values as well as values of molecular descriptors. Seven manuscripts accomplished all inclusion criteria and their sets of compounds were included in the analysis. The main characteristics of the previously published models (not necessary linear models) are presented in Table 2.

The identified sets of compounds were investigated in order to assess how the influential affect the model validity and characteristics. The best performing regression models for each set on the whole data set, on the sample after removal of compounds with leverage higher than threshold and

on the sample after removal of compounds with Cook's distance higher than threshold are presented in Table 3.

Table 2. Datasets included in analysis and basic summary of previously reported models.

Set [Ref]	Model characteristics
Set1 [33]	R^2=0.9992; s=0.929; F=3534; n=60; k=5
Set2 [34]	R^2=0.7779; F=133; R^2_{loo}=0.774; n=79; k=2
Set3 [35]	R^2=0.820; R^2_{loo}=0.716, s=0.440, F=22.805; n=31,k=5 (outliers: 5 & 15)
Set4 [36]	R^2=0.9571; R^2_{cv}=0.8521; s=0.2825; F=28.8207; n=29; k=5
Set5 [37]	R^2=0.840; R^2_{cv}=0.777; F=31.54; s=0.034; n=36; k=5
Set6 [38]	n.a.
Set7 [39]	R_{tr}=0.870; s=0.206; R_{test}=0.835, s_{test}=0.232; R_{loo}=0.925; s_{loo}=0.198; n=46; k=5

R=correlation coefficient; R^2=determination coefficient; loo=leave-one-out analysis; s=standard error of estimate; F=Fisher's statistics; n=sample size; k=number of independent variables used by the reported model; tr=training set; test=test set; n.a. = not available

Table 3. Regression characteristics: full-model (whole dataset), h_i-model (withdrawn of compounds with $h_i > h_t$) and D_i-model (withdrawn of compounds with $D_i > 4/n$, where n = sample size).

Set1: $\hat{Y}_{HF} = a + b_1 \times {}^2\chi + b_2 \times H^* + b_3 \times J^*$ where $\hat{Y}$ = estimated heat of formation; ${}^2\chi$ = generalized connectivity index; H^* = Harary index; J^* = Balaban index; HF = heats of formation; a = intercept; b_i = regression coefficients		
n=60	whole dataset	R^2=0.985; R^2_{adj}=0.985;s=3.46; F=1256 ($p=2.63\cdot10^{-55}$); R^2_{loo}=0.983; s_{loo}=3.76, F_{loo}=1061 ($p=2.91\cdot10^{-51}$); RMS=11.774; APV= 0.003; TSE= 4; APMSE= 0.210; %PredErr= 6.190;$t_{pp}=5.23\cdot10^{-14}$ (p_{PP}=1); RMSE=3.462; MAE=2.881; MAPE=0.396; SEP=3.191; REP(%)=26.581
n=56	h_i>2*(k+1)/n withdrawn (1, 38, 39, 40)	R^2=0.987; R^2_{adj}=0.986;s=3.35; F=1318 ($p=5.08\cdot10^{-49}$); R^2_{loo}=0.986; s_{loo}=3.54, F_{loo}=882 ($p=3.67\cdot10^{-49}$); RMS= 10.980; APV= 0.003; TSE= 4; APMSE= 0.211; %PredErr= 5.529; $t_{pp}=2.71\cdot10^{-13}$ (p_{PP}=1); RMSE=3.345; MAE=2.758; MAPE=0.354; SEP=3.253; REP(%)=27.928
n=54	D_i>4/n withdrawn (1, 2, 3, 16, 20, 23)	R^2=0.989; R^2_{adj}=0.988; s=3.04; F=1441 ($p=1.57\cdot10^{-48}$); R^2_{loo}=0.987; s_{loo}=3.24; F_{loo}=1268 ($p=2.67\cdot10^{-49}$); RMS= 9.059; APV= 0.003; TSE= 4; APMSE= 0.181; %PredErr= 4.866; $t_{pp}=-1.25\cdot10^{-13}$ (p_{PP}=1); RMSE=3.040; MAE=2.517; MAPE=0.290; SEP=2.749; REP(%)=29.9702
Set2: $\hat{Y}(\log(1/EC_{50})) = a + b_1 \times \log P + b_2 \times MTD^*$ where $\hat{Y}(\log(1/EC_{50}))$ = estimated $\log(1/EC_{50})$ - EC_{50} = level that produces a 50% protection of MT-4 cells against HIV-1 cytopathic effect; logP = hydrophobicities; MTD^* = minimal topological difference descriptor [34]; a = intercept; b_i = regression coefficients		
n=79	whole dataset	R^2=0.754; R^2_{adj}=0.747; s=0.68; F=116 ($p=7.59\cdot10^{-24}$); R^2_{loo}=0.733; s_{loo}=0.70; F_{loo}=104 ($p=9.75\cdot10^{-23}$); RMS=0.4516; APV=0.4630; TSE=3; APMSE=0.0059; %PredErr= 4.636; $t_{pp}=1.74\cdot10^{-14}$ (p_{PP}=1); RMSE=0.676; MAE=0.503; MAPE=0.083; SEP=0.668; REP(%)=10.546
n=77	h_i>2*(k+1)/n withdrawn (57, 61)	R^2=0.761; R^2_{adj}=0.754; s=0.66; F=118 ($p=1.01\cdot10^{-23}$); R^2_{loo}=0.714; s_{loo}=0.68, F_{loo}=106 ($p=1.06\cdot10^{-22}$); RMS=0.4275; APV=0.4386; TSE=3; APMSE=0.0058; %PredErr=4.636; $t_{pp}=-1.40\cdot10^{-14}$ (p_{PP}=1); RMSE=0.658; MAE=0.482; MAPE=0.080; SEP=0.650; REP(%)=10.336
n=66	D_i>4/n withdrawn (14,34,50,51,57-62,64,71,75)	R^2=0.899; R^2_{adj}=0.895; s=0.41; F=279 ($p=4.83\cdot10^{-32}$) R^2_{loo}=0.891; s_{loo}=0.43, F_{loo}=256 ($p=1.45\cdot10^{-31}$) RMS=0.1964; APV=0.2024; TSE=3; APMSE=0.0031; %PredErr=2.719; $t_{pp}=-2.66\cdot10^{-1}$ (p_{PP}=0.7910); RMSE=0.412; MAE=0.353; MAPE=0.060; SEP=0.440; REP(%)=7.059

Set3: $\hat{Y}(\log K_i) = a + b_1 \times L + b_2 \times B_1 + b_3 \times B_3 + b_4 \times FPSA_3 + b_5 \times \rho$ where $\hat{Y}$ = estimated activity; K_i = binding affinity; L = sterimol parameter; B_1, B_3 = sterimol width parameters; $FPSA_3$ = fractional charged partial surface area; ρ = density; a = intercept; b_i = regression coefficients		
n=33	whole dataset	R^2=0.524; R^2_{adj}=0.436; s=0.69; F=6 (p = 0.001); R^2_{loo}=0.287; s_{loo}=0.88; F_{loo}=1.52 (p =0.2155); RMS=0.4588; APV=0.5283; TSE=6; APMSE=0.0170; %PredErr=37.0833; t_{PP}=-1.39·10^{-14} (p_{PP}=1); RMSE=0.690; MAE=0.494; MAPE=0.679; SEP=0.634; REP(%)=39.601
n=31	h_i>2*(k+1)/n withdrawn (1, 8)	R^2=0.555; R^2_{adj}=0.466; s=0.65; F=6 (p=0.001); R^2_{loo}=0.254; s_{loo}=0.96; F_{loo}=0.81 (p=0.5518); RMS=0.4993; APV=0.3267; TSE=5; APMSE=0.0192; %PredErr=34.5228; t_{PP}=-1.35·10^{-14} (p_{PP}=1); RMSE=0.698; MAE=0.490; MAPE=0.680; SEP=0.660; REP(%)=41.023
n=26	D_i>4/n withdrawn (1, 2, 5, 13, 15, 21, 30)	R^2=0.858; R^2_{adj}=0.821; s=0.41; F=23 (p=1.87·10^{-7}); R^2_{loo}=0.767; s_{loo}=0.52, F=13 (p=1.05·10^{-5}); RMS=0.3267; APV=0.3831; TSE=3; APMSE=0.0192; %PredErr=17.0907; t_{PP}=-2.48·10^{-14} (p_{PP}=1); RMSE=0.427; MAE=0.289; MAPE=0.294; SEP=0.529; REP(%)=32.175
Set4: $\hat{Y}(MPmg) = a + b_1 \times RPCG + b_2 \times Q10 + b3 \times F_{H2O}$ where $\hat{Y}$(MPmg) = estimated mutagenic potencies for *M. gilvum*; RPCG = (charge of the most positively charged atom) / (sum of total positive charge); Q10 = charges on position 10; F_{H2O} = desolvation free energy for waterA; a = intercept; b_i = regression coefficients		
n=29	whole dataset	R^2=0.652; R^2_{adj}=0.610; s=0.41; F=16 (p=6.38·10^{-6}); R^2_{loo}=0.477; s_{loo}=0.51; F=7 (p=0.0013) RMS=0.1610; APV=0.1776; TSE=4; APMSE=0.0064; %PredErr=3.834; t_{PP}=8.80·10^{-16} (p_{PP}=1); RMSE=0.4091; MAE=0.1443; MAPE=3.2906; SEP=0.3866; REP(%)=116.6703
n=27	h_i>2*(k+1)/n withdrawn (10, 26)	R^2=0.643; R^2_{adj}=0.596; s=0.64; F=14 (p=2.34·10^{-5}); R^2_{loo}=0.495; s_{loo}=0.46; F=7 (p=0.0016); RMS=0.1401; APV=0.1557; TSE=5; APMSE=0.0061; %PredErr=3.2149; t_{PP}=3.25·10^{-14} (p_{PP}=1); RMSE=0.399; MAE=0.125; MAPE=1.228; SEP=0.360; REP(%)=89.281
n=23	D_i>4/n withdrawn (10,13,16,26)	R^2=0.568; R^2_{adj}=0.506; s=0.34; F=9 (p=4.38·10^{-4}); R^2_{loo}=0.407; s_{loo}=0.41; F=4 (p=0.0145); RMS=0.1104; APV=0.1236; TSE=4; APMSE=0.0053; %PredErr=2.6091; t_{PP}=-2.62·10^{-15} (p_{PP}=1); RMSE=0.340; MAE=0.097; MAPE=2.390; SEP=0.318; REP(%)=102.864
Set5: $\hat{Y}(pKI) = b_1 \times {}^2AIC + b_2 \times NBR + b_3 \times NCA$ where $\hat{Y}(pK_i)$ = estimated inhibitory activity against CA II isozyme; ^{2}AIC = average information content (order 2); NBR = number of benzene rings; NCA = number of C atoms; b_i = regression coefficients		
n=38	whole dataset	R^2=0.586; R^2_{adj}=0.533; s=0.29; F=16 (p=8.79·10^{-7}); R^2_{loo}=0.532; s_{loo}=0.31; F=13 (p=8.99·10^{-6}); RMS=0.0816; APV=0.0880; TSE=5; APMSE=0.0024; %PredErr=3.4285; t_{PP}=-0.0453 (p_{PP}=0.9641); RMSE=0.2856; MAE=0.0751; MAPE=0.1334; SEP=0.2778; REP(%)=14.7242
n=34	h_i>2*k/n withdrawn (C23, C24, C25, C32) b_2 - p=0.1093	R^2=0.448; R^2_{adj}=0.380; s=0.29; F=8 (p=3.42·10^{-4}); R^2_{loo}=0.360; s_{loo}=0.32; F=5 (p=4.12·10^{-3}); RMS=0.0863; APV=0.0939; TSE=5; APMSE=0.0029; %PredErr=3.0648; t_{PP}=0 (p_{PP}=1); RMSE=0.2938; MAE=0.0787; MAPE=0.1307; SEP=0.2847; REP(%)=14.5028
n=37	D_i>4/n withdrawn (C8)	R^2=0.597; R^2_{adj}=0.544; s=0.28; F=17 (p=8.60·10^{-7}); R^2_{loo}=0.541; s_{loo}=0.31; F=13 (p=9.48·10^{-6}); RMS=0.0810; APV=0.0875; TSE=5; APMSE=0.0025; %PredErr=3.3326; t_{PP}=0 (p_{PP}=1); RMSE=0.2845; MAE=0.0744; MAPE=0.1321; SEP=0.2765; REP(%)=14.5992
Set6: $\hat{Y}(HE\text{-}Mlog(1/MRC_{50})) = b_1 \times logP + b_2 \times E_{tot}$ where $\hat{Y}(HE\text{-}Mlog(1/MRC_{50}))$ = estimated toxicity on *Hydractinia echinata*; logP = hydrophobicity; E_{tot} = total optimized energy; b_i = regression coefficients		
n=28	whole dataset	R^2=0.631; R^2_{adj}=0.579; s=1.25; F=22 (p=2.81·10^{-6}); R^2_{loo}=0.550; s_{loo}=1.42; F_{loo}=15 (p=5.32·10^{-5}); RMS=1.5644; APV=1.6761; TSE=4; APMSE=0.0626; %PredErr=3.4705; t_{PP}=-0.0574 (p_{PP}=0.9546); RMSE=1.2507; MAE=1.4526; MAPE=2.4174; SEP=1.2274; REP(%)=35.2585
n=26	h_i>2*k/n withdrawn	R^2=0.692; R^2_{adj}=0.638; s=1.19; F=27 (p=9.26·10^{-7});

	(C8, C25) b_1 - p > 0.05	R^2_{loo}=0.649; s_{loo}=1.30; F=21 (p=6.28·10^{-6}); RMS=1.4097; APV=1.5182; TSE=4; APMSE=0.0613; %PredErr=3.0244; t_{PP}=0.7156 (p_{PP}=0.4804); RMSE=1.1873; MAE=1.3013; MAPE=2.3819; SEP=1.1633; REP(%)=32.9849
n=23	D_i>4/n withdrawn (C5, C8, C21, C25, C27) b_1 - p > 0.05	R^2=0.674; R^2_{adj}=0.611; s=1.13; F=22 (p=9.72·10^{-6}); R^2_{loo}=0.627; s_{loo}=1.24; F_{loo}=17 (p=5.51·10^{-5}); RMS=1.2801; APV=1.3914; TSE=4; APMSE=0.0640; %PredErr=2.4588; t_{PP}=1.5758 (p_{PP}=0.1267); RMSE=1.1314; MAE=1.1688; MAPE=3.1657; SEP=1.1054; REP(%)=31.6591
Set7: $\hat{Y}(\log ED_{50}) = b \times DCW^3$ where $\hat{Y}$ = estimated antiepileptic activities (dose at which 50% of individuals reach the desired effect); DCW^3 = descriptor calculated with Monte Carlo simulation [39]; b = regression coefficient		
n=51	whole dataset	R^2=0.737; R^2_{adj}=0.717; s=0.21; F=140 (p=5.65·10^{-16}); R^2_{loo}=0.729; s_{loo}=0.21; F_{loo}=131 (p=1.89·10^{-15}); RMS=0.0427; APV=0.0435; TSE=3; APMSE=0.0009; %PredErr=3.2254; t_{PP}=-0.1155 (p_{PP}=0.9085); RMSE=0.2066; MAE=0.0418; MAPE=0.1116; SEP=0.2066; REP(%)=12.9209
n=	h_i>2*k/n withdrawn (none)	no h_i value higher than threshold was identified
n=48	D_i>4/n withdrawn (C2, C19, C26, C36, C46, C51)	R^2=0.838; R^2_{adj}=0.816; s=0.15; F=228 (p=8.22·10^{-19}); R^2_{loo}=0.835; s_{loo}=0.16; F=213 (p=1.75·10^{-18}); RMS=0.0230; APV=0.0235; TSE=3; APMSE=0.0005; %PredErr=2.2033; t_{PP}=-0.1733 (p_{PP}=0.8632); RMSE=0.1516; MAE=0.0225; MAPE=0.0892; SEP=0.1516; REP(%)=9.6954

R^2= determination coefficient; R^2_{adj}= adjusted correlation coefficient; s= standard error of estimate; F=F-value (p= p-valye); R^2_{loo}= determination coefficient in leave-one-out analysis; s_{loo}= standard error of predicted; F_{loo}= Fisher's value and associated significance in leave-one-out analysis; RMS= residual mean square; APV= average prediction variance; TSE= total squared error; APMSE= average prediction mean squared error; %PredErr= prediction error; t_{PP}, p_{PP}= t-statistics for intercept and regression coefficients; RMSE= root-mean-square error; ME= mean error; MAE= mean absolute error; MAPE= mean absolute percentage error; SEP= standard error of prediction; REP(%)=relative error of prediction

Classification of QSAR models (full-model, model obtained after withdrawn of compound(s) with h_i - h_i-model and respectively with D_i higher than thresholds - D_i-model) according to applied validation statistics is presented in Figure 1.

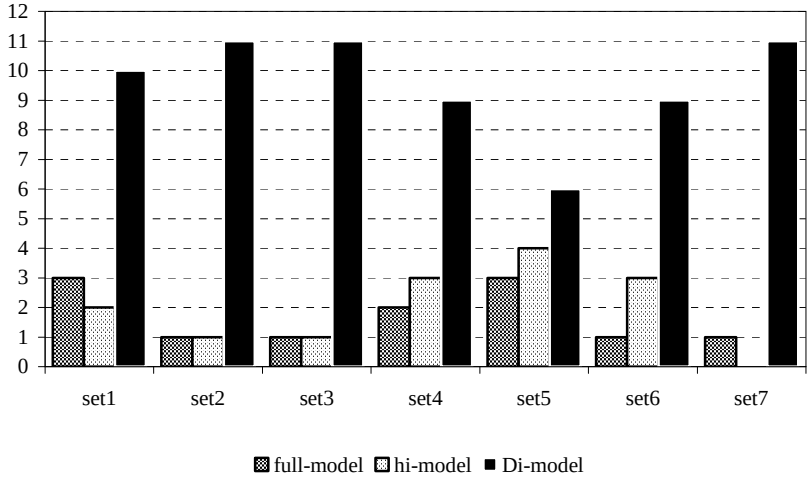

Figure 1. Full model & h_i-model & D_i-model: classification according to validation criteria.

The highest correlation coefficient was obtained in 5 cases out of 7 by the model after removal the compounds with the Cook's distance higher than threshold. The full model obtained the higher

correlation coefficient in the fourth set, while the model obtained after removal of the compounds with leverage higher than threshold obtained the higher correlation coefficient in the sixth set. The evolution of correlation coefficients is presented in Figure 2.

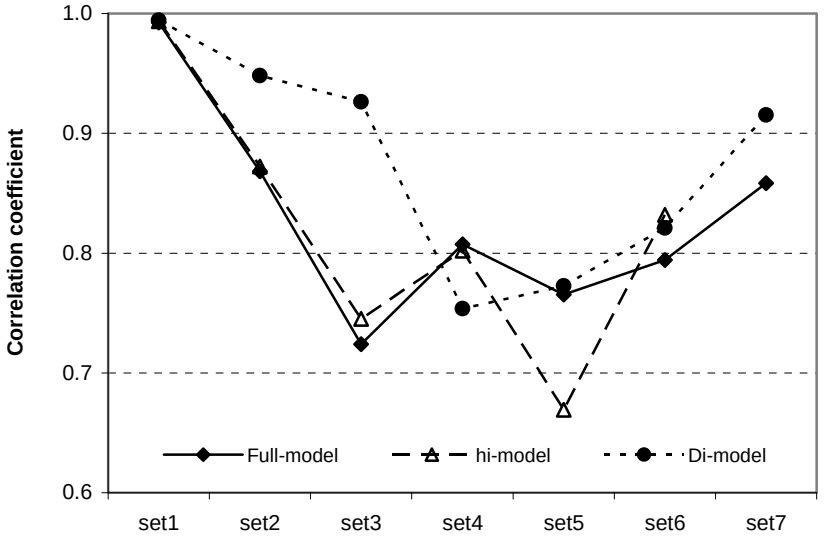

Figure 2. Full model - h_i model - D_i model: evolution of correlation coefficient

Statistical significant increases in correlation coefficient have been identified in the second and third sets when both the full-model and the h_i-model were compared to D_i-model (Table 4). The Fisher's Chi-Square statistic (F-C-S) was applied to test if overall one model is better than other and the results are presented in Table 4.

Table 4. Steiger's Z test for correlation coefficients comparisons and overall significance: results

Set	Full-model vs. h_i-model Z (p-value)	Full-model vs. D_i-model Z (p-value)	h_i-model vs. D_i-model Z (p-value)
set1	0.3760 (0.3535)	0.855 (0.1963)	0.4820 (0.3149)
set2	0.1040 (0.4586)	2.861 (0.0021)	2.7450 (0.0030)
set3	0.1740 (0.4309)	2.583 (0.0049)	2.3810 (0.0086)
set4	0.0560 (0.4777)	0.465 (0.3210)	0.4040 (0.3431)
set5	0.8100 (0.2090)	0.073 (0.4709)	0.8760 (0.1905)
set6	0.3840 (0.3505)	0.255 (0.3994)	0.1130 (0.4550)
set7	n.a.	1.312 (0.0948)	n.a.
F-C-S (p-value)	6.0139 (0.4216)	18.2757 (0.0108)	15.2359 (0.0185)

F-C-S = Fisher's Chi-Square statistic; p-value = probability

The leave-many-out analyses were conducted on set1 and set2 to assess the usefulness of influential identification and withdrawn on the QSARs abilities. Characteristics of the obtained models are presented in Table 5.

Table 5. Leave-many-out analysis: results.

Set	Split	n	R^2	F	p_F	n	R^2	F	p_F
		full-model				D_i-model			
set1	Training	40	0.9875	950	$2.59 \cdot 10^{-34}$	38	0.9890	1020	$2.32 \cdot 10^{-33}$
	Test	20	0.9802	223	$2.89 \cdot 10^{-13}$	16	0.9869	300	$1.50 \cdot 10^{-11}$
set2	Training	53	0.7539	77	$5.55 \cdot 10^{-16}$	45	0.9097	211	$1.18 \cdot 10^{-22}$
	Test	26	0.7609	33	$1.58 \cdot 10^{-7}$	21	0.8810	67	$4.77 \cdot 10^{-9}$

n = sample size; R^2 = determination coefficient;
F = Fisher's statistics; p_F = significance of F statistics;

The plot of full-model versus D_i-model for set1 and set2 are presented in Figures 3 and 4.

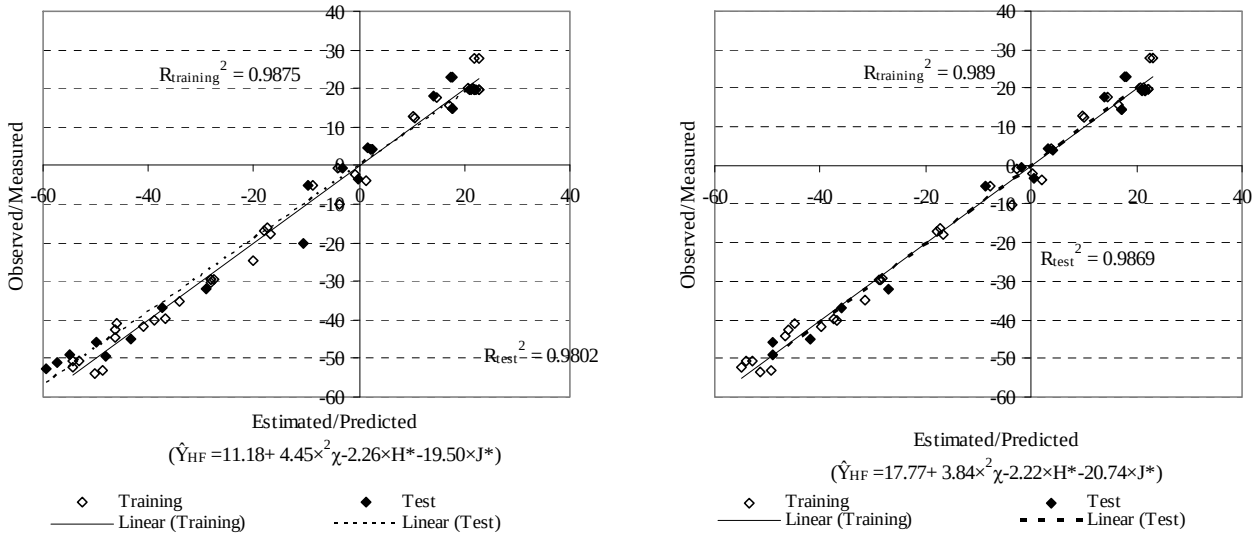

Figure 3. Set1 full-model (left-hand) vs. D_i-model (right-hand): observed/measured vs. estimated/predicted

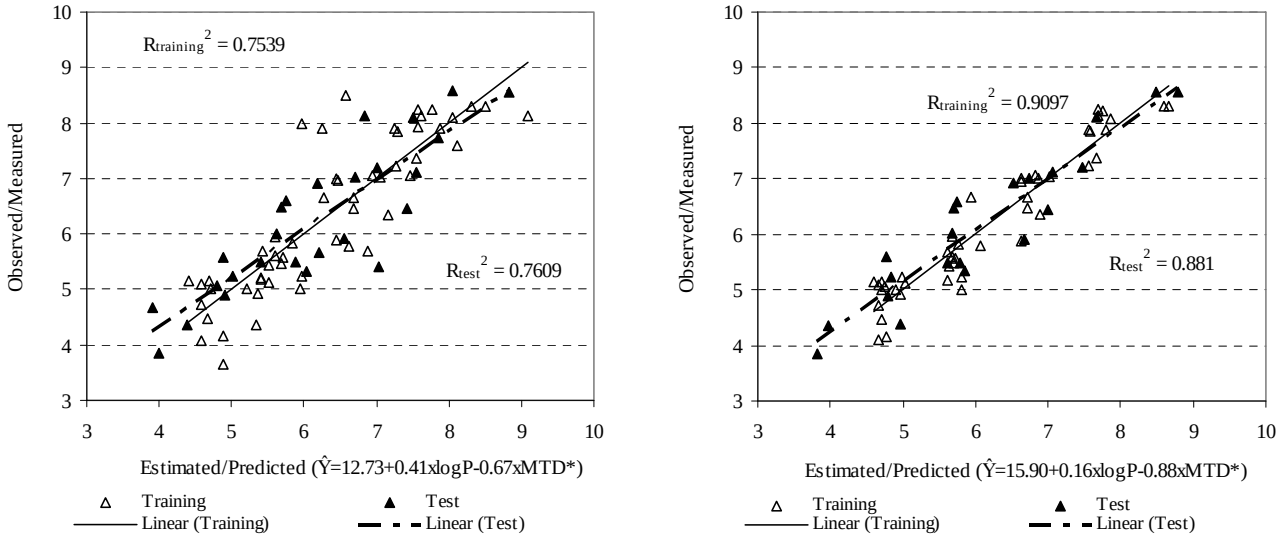

Figure 4. Set2 full-model (left-hand) vs. D_i-model (right-hand): observed/measured vs. estimated/predicted

The leave-one-out cross-validation determination coefficient for training set1 was of 0.9846 while for training set set2 was of 0.7208 when full-models were investigated. The leave-one-out cross-validation determination coefficient for training set1 was of 0.9870 while for training set2 was of 0.8975 when the D_i-models were investigated. A statistically significant increase of correlation coefficient has been identified for the training set of the set2 in D_i-model compared to full-model (Z = 2.609, p-value = 0.0045).

Discussion

The assessment of influential withdrawn using leverage and Cook's distance has successfully accomplished. Seven data sets with sample sizes range from 28 (set6) to 79 (set2) were analyzed. Three linear regression models were investigated for each set included in analyzes whenever appropriate (full-model, h_i-model and D_i-model). The present study tried to answer to the following research question: "Hat-matrix approach is more appropriate than Cook's distance approach to identify influential in regression analysis?".

The analysis of the obtained results revealed that in 5 out of 7 sets the correlation coefficient increase when both compounds with h_i and respectively D_i higher than thresholds were removed (Table 3). The number of withdrawn compounds varied from 2 to 4 for h_i-model and from 1 to 13 for D_i-model (Table 3). In just few cases the same compound was identified as influential by both leverage and Cook's methods: 1 compound (in set1, set2, and set3) and 2 compounds (in set set4 and set6).

Some independent variable proved not to have a statistically contribution to the model (see Table 3): h_i-model set5 (translated also to a lower determination coefficient compared to full-model) and set6 and D_i-model set6 (the determination coefficient had a higher value for h_i-model compared to D_i-model for set6). In these cases, it is correct to construct the models without those descriptors identified with no statistically contribution to the model. With one exception represented by set4, determination coefficients for D_i-models were higher than determination coefficients obtained in full-models (Table 3 and Figure 2). The highest increase of determination coefficient was observed in D_i-model of set3. The difference between determination coefficient and adjusted determination coefficient varied from 0 to 0.088 (for full-model – set3), 0.089 (for h_i-model – set3) and 0.063 (for D_i-model – set6). The difference between determination coefficient and its corresponding value in leave-one-out analysis varied from 0.002 to 0.237 (full-model), 0.001 to 0.148 (h_i-model), and from 0.002 to 0.161 (D_i-model).

The analysis of validation statistics showed that D_i-models obtained systematically better results compared to full-models (Table 3 and Figure 1). Furthermore, even if goodness-of-fit is not a good

statistics for model predictivity [40,41], no statistically significant differences between correlation coefficients obtained in full-model compared to those obtained in h_i-models were identified (Table 4). However, the correlation coefficients obtained by D_i-models proved statistically significant higher compared to those obtained in both full-model and hi-model for set2 and set3 (Table 4). Furthermore, the F-C-S statistic showed that overall, the D_i-model was better than both full-model and h_i-model ($p < 0.05$, Table 4). The above-presented facts let to the conclusion that analysis of influential should be conducted by applying the Cook's distance approach.

The external validation of the Cook's distance approach was furthermore assessed in leave-many-out analysis on two datasets (set1 and set2), one with statistically increase of correlation coefficient (set2) and one without statistically increase of correlation coefficient (set1). Similar results are obtained when training and test sets are compared (Table 5). The significant increase of determination coefficient in both training and test sets is transmitted also in leave-many-out analyzes for the second dataset (set2), the increase being of 0.156 for training set and 0.120 for test set. The spread of point in the plots of full-model and D_i-model is similar for set1 (Figure 3) but the difference are obvious when set2 is investigated (Figure 4). A reliable and valid regression model must look as set2 D_i-model not as set2 full-model (Figure 4).

Scientifically literature recommend not to trust a QSAR model when correlation coefficient is lower than 0.6, which known to be is an insufficient condition for assessment of predictive power of a model [42]. This analysis show that a determination coefficient < 0.6 could be significantly improved with analyses and withdrawn of influential in order to obtain a model with good performance in prediction (see Table 3, set3). In our opinion, the predictivity power of a model stands in correct application of statistical methods to identify the QSAR models. Identification of influential compound in data set could significantly improve the model and should be conducted any time when a regression analysis is desired. Fit the model with and without the influential compound(s) and look to the effect on regression characteristics (R^2, R^2_{adj}, F-value (p-value), s, regression coefficients and their significance, validation criteria presented in Table 1) as well as on the plot of the models. It is the task of a statistician to examine the influential compounds and to identify important cases before presentation of results but this task could be done by any researcher with experience in statistics. Based on the presented results, it is showed that Cook's distance approach is more suitable to proper identification of influential in dataset and we recommend its application in linear regression analysis for QSAR models. The leverage approach could be used on the D_i-model to analyze the membership of compounds in the model to the structural model domain [43].

Based on the results obtained in this study we recommend that either to accept (if leave-one-out, leave-many-out analyses and external validation sustain the model) or to reject the QSAR model

obtained after removal of influential(s) and never accept a model that contains influential compounds (their presence lead to instability of the QSAR model).

Conclusion

The use of leverage methodology led to improvement of QSAR models characteristic and performances. Better QSAR models were obtained when Cook's distance approach was used compared to both full-model and h_i-model. Cook's distance approach is recommended to be used to identify influential compounds in dataset whenever the linear regression analysis for QSAR models is applied.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for reliability and uncertainty assessment for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Persp.*, **2003**, *111*, 1361-1375.

[2] Walker, J.D.; Jaworska, J.; Comber, M.H.I.; Schultz, T.W.; Dearden, J.C. Guidelines for developing and using quantitative structure-activity relationships. *Environ. Toxicol. Chem.*, **2003,** *22*, 1653-1665.

[3] Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.*, **2007**, *13*, 3494-3504.

[4] Jaworska, J.S.; Comber, M.; Auer, C.; Van Leeuwen, C.J. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ. Health Persp.*, **2003**, *111*, 1358-1360.

[5] Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models. [online] [Accessed 31 May 2012]. ENV/JM/MONO (OECD Environment Health and Safety Publications) 2007;2. Available from: URL: http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en

[6] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.*, **2010**; *29(6-7)*, 476-488.

[7] Mekenyan, O.G.; Veith, G.D. Synergism between QSAR Modeling and Physico-Chemical Principles. *SAR QSAR Environ. Res.*, **1995**, *4*, 155-165.

[8] Dearden, J.C.; Cronin, M.T.D.; Kaiser, K.L.E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.*, **2009**, *20(3-4)*, 241-266.

[9] Cronin, M.T.D.; Schultz, T.W. Pitfalls in QSAR. *J. Theoret. Chem. (Theochem)*, **2003**, *622*, 39-51.

[10] Bolboacă, S.D.; Jäntschi, L. Distribution Fitting 3. Analysis under Normality Assumptions. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, **2009**, *66(2)*, 698-705.

[11] Bolboacă, S.D.; Jäntschi, L. Modelling the property of compounds from structure: statistical methods for models validation. *Environ. Chem. Lett.*, **2008**, *6*, 175-181.

[12] Abraham, M.H.; Sanchez-Moreno, R.; Cometto-Muniz, J.E. A quantitative structure-activity analysis on the relative sensitivity of the olfactory and the nasal trigeminal chemosensory systems. *Chem. Senses*, **2007**, *32*, 711-719.

[13] Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1177-1207.

[14] Topliss, J.G.; Costello, R.J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.*, **1972**, *15*, 1066-1068.

[15] Kolmogorov, A. Confidence Limits for an Unknown Distribution Function. *Ann. Math. Stat.*, **1941**, *12(4)*, 461-463.

[16] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **1900**, *50*, 157-175.

[17] Kleinboum, D.G.; Kupper, L.L.; Muler, K.E.; Nizam, A. *Applied Regression Analysis and other Multivariate Methods*. Duxbury Press: Pacific Grove, **1998**.

[18] Tropsha, A.; Gramatica, P.; Gombar, V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.*, **2003**, *22(1)*, 69-77.

[19] Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.*, **2007**, *26(5)*, 694-701.

[20] Cook, R.D. Detection of Influential Observations in Linear Regression. *Technometrics (American Statistical Association)*, **1977**, *19(1)*, 15-18.

[21] Cook, R.D. Influential Observations in Linear Regression. *J. Amer. Statistical Assoc.*, **1979**, *74(365)*, 169-174.

[22] Bollen, K.A.; Jackman, R. *Regression diagnostics: An expository treatment of outliers and influential cases*. In: *Modern Methods of Data Analysis*, Fox, J.; Scott, J. Long, Eds. Sage: Newbury Park, **1990**, pp. 257-91.

[23] Steiger, J.H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, **1980**, *87*, 245-251.

[24] Bolboacă, S.D. Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique. *Appl. Med. Inform.*, **2010**, *28(2)*, 9-14.

[25] Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence **1995**, *2(12)*, pp. 1137-1143.

[26] Martens, H.A.; Dardenne, P. Validation and verification of regression in small data sets. *Chemometr. Intell. Lab.*, **1998**, *44*, 99-121.

[27] Besalú, E.; de Julián-Ortiz, J.V. Equivalence of the Pecka-Ponec Correlation Probability and the Statistical F Significance for MLR Models. *J. Math. Chem.*, **2004**, *36(4)*, 361-363.

[28] Mallow, C.L. Some comments on Cp. *Technometrics*, **1973**, *15*, 661-675.

[29] Gorman, J.W.; Toman, R.J. Selection of variables for fitting equations to data. *Technometrics*, **1966**, *8*, 27-51.

[30] Tukey, J.W. Discussion. *J. R. Statisti. Soc.*, **1967**, *29*, 47-48.

[31] Milac, A.-L.; Avram, S.; Petrescu, A.-J. Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors: Application to HIV-1 protease inhibitors. *J. Mol. Graph. Model.*, **2006**, *25(1)*, 37-45.

[32] Fisher, R.A. Combining independent tests of significance. *Amer. Statistician*, **1948**, *2*, 30.

[33] Mercader, A.; Castro, E.A.; Toropov, A.A. Maximum Topological Distances Based Indices as Molecular Descriptors for QSPR. 4. Modeling the Enthalpy of Formation of Hydrocarbons from Elements. *Int. J. Mol. Sci.*, **2001**, *2*, 121-132.

[34] Duda-Seiman, C.; Duda-Seiman, D.; Dragos, D.; Medeleanu, M.; Careja, V.; Putz, M.V.; Lacrama, A.-M.; Chiriac, A.; Nutiu, R.; Ciubotariu, D. Design of Anti-HIV Ligands by Means of Minimal Topological Difference (MTD) Method. *Int. J. Mol. Sci.*, **2006**, *7*, 537-555.

[35] Kim, D.; Hong, S.-I.; Lee, D.-S. Triazoloquinazolines as Human A3 Adenosine Receptor Antagonists: A QSAR Study. *Int. J. Mol. Sci.*, **2006**, *7*, 485-496.

[36] Kim, D.; Hong, S.-I.; Lee, D.-S. The Quantitative Structure-Mutagenicity Relationship of Polycylic Aromatic Hydrocarbon Metabolites. *Int. J. Mol. Sci.*, **2006**, *7*, 556-570.

[37] Eroglu, E. Some QSAR Studies for a Group of Sulfonamide Schiff Base as Carbonic Anhydrase CA II Inhibitors. *Int J Mol Sci.*, **2008**, *9(2)*, 181-197.

[38] Chicu, S.A.; Putz, M.V. Köln-Timişoara Molecular Activity Combined Models toward Interspecies Toxicity Assessment. *Int. J. Mol. Sci.*, **2009**, *10*, 4474-4497.

[39] Garro Martinez, J.C.; Duchowicz, P.R.; Estrada, M.R.; Zamarbide, G.N.; Castro, E.A. QSAR Study and Molecular Design of Open-Chain Enaminones as Anticonvulsant Agents. *Int. J. Mol. Sci.*, **2011**, *12*, 9354-9368.

[40] Novellino, E.; Fattorusso, C.; Greco, G. Use of comparative molecular field analysis and cluster analysis in series design Original. *Pharm. Acta Helv.*, **1995**, *70(2)*, 149-154.

[41] Norinder, U. Single and domain mode variable selection in 3D QSAR applications. *J Chemomet.*, **1996**, *10(2)*, 95-105.

[42] Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Mod.*, **2002**, *20(4)*, 269-276.

[43] Atkinson, A.C. *Plots, Transformations and Regression*. Oxford: Clarendon Press, **1985**.