TITLE: Lipophilicity estimation and characterization of selected steroid derivatives of biomedical importance applying RP HPLC

AUTHORS: Lidija Jevrić, Milica Karadžić, Anamarija Mandić, Sanja Podunavac-Kuzmanović, Strahinja Kovačević, Andrea Nikolić, Aleksandar Oklješa, Marija Sakač, Katarina Penov-Gaši, Srđan Stojanović

NOTICE: This is the author's version of a work that was accepted for publication *Journal of Pharmaceutical and Biomedical Analysis.* Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Journal of Pharmaceutical and Biomedical Analysis*, Volume 134, 5 February 2017, Pages 27–35. DOI: 10.1016/j.jpba.2016.11.015

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com

Lipophilicity estimation and characterization of selected steroid derivatives of biomedical importance applying RP HPLC

Lidija R. Jevrić [a], Milica Ž. Karadžić [a,*], Anamarija I. Mandić [b], Sanja O. Podunavac Kuzmanović [a], Strahinja Z. Kovačević [a], Andrea R. Nikolić [c], Aleksandar M. Oklješa [c], Marija N. Sakač [c], Katarina M. Penov Gaši [c], Srđan Z. Stojanović [d]

[a] *Faculty of Technology Novi Sad, University of Novi Sad, Bulevar Cara Lazara 1, 21000 Novi Sad, Serbia*
[b] *Institute of Food Technology, University of Novi Sad, Bulevar Cara Lazara 1, 21000 Novi Sad, Serbia*
[c] *Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 3, 21000 Novi Sad, Serbia*
[d] *Faculty of Pharmacy, Trg Mladenaca 5, 21000 Novi Sad, Serbia*

ARTICLE INFO

Article history:
Received 21 October 2016
Received in revised form 4 November 2016
Accepted 6 November 2016
Available online xxx

Keywords:
Drug candidates
Chemometrics
Lipophilicity
Liquid chromatography
Quantitative structure retention relationships

ABSTRACT

The present paper deals with chromatographic lipophilicity determination of twenty-nine selected steroid derivatives using reversed-phase high-performance liquid chromatography (RP HPLC) combined with two mobile phase, acetonitrile-water and methanol-water. Chromatographic behavior of four groups (triazole and tetrazole, toluenesulfonylhydrazide, nitrile and dinitrile and dione) of selected steroid derivatives was studied. Investigated compounds were grouped using principal component analysis (PCA) according to their log*k* values for both mobile phases. Grouping was in the very good accordance with the polarity and lipophilicity of the investigated compounds. QSRR (quantitative structure-retention relationship) approach was used to model chromatographic lipophilicity behavior using molecular descriptors. Modeling was performed using linear regression (LR) and multiple linear regression (MLR) methods. The most influential molecular descriptors were lipophilicity descriptors that are important for molecules ability to pass through biological membranes and geometrical descriptors. All established LR-QSRR and MLR-QSRR models were statistically validated by standards, *cross*- and external validation parameters as well as with two graphical methods. According to all these assessments, MLR models were better for chromatographic lipophilicity prediction. It was shown that chromatographic systems with methanol-water were better for modeling of log*k* than systems with acetonitrile-water, as well as the systems that contained lower volume fractions of organic component in mobile phase. Modeling was performed in order to obtain lipophilicity profiles of investigated compounds as future drug candidates of biomedical importance.

1. Introduction

As one of the most important physicochemical characteristics, lipophilicity has a crucial role in pharmacological behavior and activity of drugs, with an emphasis on passive transport through biological membranes. Lipophilicity also affects the formation of complexes between a compound and blood proteins and receptors at the site of drug action in the organism [1,2]. Passive transport through biological membranes is expressed as the 1-octanol/water partition coefficient ($\log P_{o/w}$) [3]. The reference method for $\log P_{o/w}$ determination is the shake-flask method. Reversed-phase high-performance liquid chromatography (RP HPLC) represents a very good alternative method because of its good accuracy, low sample consumption, on-line detection and its ability to perform measurements even in a presence of a mixture. Additionally, chromatography has high throughput ability that is important regarding a very high number of potential drug candidates. Considering the crucial importance of the first step in selection of the drug candidates it is very valuable to obtain any information regarding physicochemical properties of selected drug candidate. The great advantage of HPLC is the possibility to use different types of stationary phases and various numbers of the mobile phases. Steroid compounds are usually present in small concentrations in various biological samples so analytical techniques of high sensitivity are needed for their detection and quantification. Chromatographic techniques are very useful in this field as well as for the lipophilicity determination. Chromatographic behavior of molecules is conditioned by functional groups presented in it and their position and orientation dictates the chromatographic conditions (mobile and stationary phases) selection. Lipophilicity determination through chromatographic analysis requires the use of strictly defined chromatographic conditions.

Retention behavior of molecule in RP chromatographic system is closely related to its lipophilicity [4]. Hence, chromatography is very often used for lipophilicity determination of various number of different molecules [5]. In RP HPLC lipophilicity is commonly derived

* Corresponding author.
Email address: mkaradza@uns.ac.rs (M.Ž. Karadžić)

http://dx.doi.org/10.1016/j.jpba.2016.11.015
0731-7085/

from the logarithm of the retention factor, log*k*:

$$\log k = \log\left(\frac{t_r - t_0}{t_0}\right) \quad (1)$$

t_r – retention time of a compound, t_0 – dead time (the first peak on the chromatogram). As chromatographic lipophilicity, logarithm of the retention factor, log*k*, could be used [6]. Because lipophilicity of selected steroid derivatives could not be determined in pure water, $\log k_w$ factor (theoretical capacity factor defined in pure water as mobile phase) would not have any physical meaning. Lipophilicity values can be also estimated using different software packages. Advantages of chromatographic lipophilicity determination compared to software lipophilicity determination is consistency of the retention parameters determined at strictly defined chromatographic conditions. Additionally, software lipophilicity determination provides several ways for log*P* calculation for every software package and therefore the result depends on the calculation procedure.

The role of lipophilicity is discussed in terms of quantitative structure-retention relationships (QSRR). Linear regression (LR) and multiple linear regression (MLR) were applied in order to establish linear relationships between the experimentally observed log*k* values and *in silico* molecular descriptors. As a classification tool, principal component analysis (PCA) has been applied to group investigated compounds regarding their log*k* values for both organic solvents.

Based on the importance of the selected steroid derivatives and their biomedical importance, chromatographic lipophilicity of twenty-nine compounds was examined under different chromatographic conditions. Chromatographic retention was used for QSRR modeling of chromatographic lipophilicity in order to conduct lipophilicity determination for further biological analysis.

2. Material and methods

2.1. Studied steroids

The synthesis of these steroid derivatives has been published earlier [7–10]. Their 2D structures and IUPAC names are presented in Table 1. The set of twenty-nine studied steroid derivatives was divided into four groups: triazole and tetrazole (molecules **I.1**–**I.7**), toluenesulfonylhydrazide (**II.8**–**II.11**), nitrile and dinitrile (**III.12**–**III.27**) and dione (**IV.28** and **IV.29**). Substituents that occur in these molecules are: OH (hydroxyl), O (oxo), Ac (acetyl) and Bn (benzyl).

2.2. Instrumentation and chemicals

For the chromatographic measurements an Agilent Technologies 1200 Series HPLC (Agilent, Santa Clara, California, USA) with diode array (DAD) and evaporative light scattering detector (ELSD) system was used. As stationary phase, ZORBAX SB-C18, 3.0 × 250 mm i.d., 5 μm particle size (Agilent, Santa Clara, California, USA) column was used. Used acetone, acetonitrile and methanol were HPLC grade, purchased from J. T. Baker (Deventer, Netherlands). Ultrapure water was obtained in the laboratory using Millipore, Elix UV system and Simplicity Water Purification System (Millipore, Molsheim, France).

2.3. Chromatographic procedure

Investigated compounds were dissolved in acetone in concentration of 1 mg/mL and filtrated throughout Captiva Econofilter (nylon membrane, 25 mm diameter, 0.2 μm pore size) (Santa Clara, California, USA). Chromatographic procedure was isocratic. A binary mixtures of acetonitrile and water and methanol and water were used as the mobile phases. For ZORBAX SB-C18 column acetonitrile volume fraction was 70–80 v/v and methanol 70–85 v/v. The flow rate and injection volume was 0.6 mL/min and 10 μL. During the analysis, the column temperature was held constant at 30 °C. DAD detection was done at 210 nm. Operating temperature of ELSD detector was 40 °C, pressure 3.5 bar and gain 5. All analysis were done in triplicate. Retention data were expressed as the log*k* values as defined by Eq. (1) and they were used as dependent variables in QSRR modeling.

2.4. In silico molecular descriptors

For molecular structure design the following software were used: for 2D structure MarvinSketch 15.3.26 and for 3D structure ChemBio3D Ultra 12.0 [11,12]. For the calculation of 380 molecular descriptors 7 programs were used: ChemBioDraw Ultra 12.0, ChemBio3D Ultra 12.0, PaDEL Descriptor, ALOGPS 2.1, PreADMET online program, Molinspiration online program, MarvinSketch 15.3.26 [11–16]. For descriptor selection, the stepwise selection (SS) method was used. In this method, after each step independent variable (descriptor) is added in the model and all other are being checked for their significance. The criterion for adding or removing the variables was root mean square error (RMSE) [17]. In this paper, minimum RMSE was set at 0.05 and each descriptor that increased the RMSE was removed from the model. Selected data were used as input data for MLR modeling [18]. The selected set of 22 molecular descriptors contains 15 lipophilicity, 4 geometrical, 2 physicochemical and 1 molecular bulkiness descriptors. The molecular descriptors were calculated on the basis of 2D structures, therefore, the structural optimization and energy minimization were not required except in the case of molecular descriptors calculated using ChemBio3D Ultra 12.0 [12]. For the formed 3D structures, energy minimization had to be done using molecular mechanics force field method (MM2). The energy minimization was performed until the root mean square (RMS) value reached a value smaller than 0.1 kcal/Å mol. The values of the selected molecular descriptors for QSRR modeling are presented in Supplementary data (Table S1).

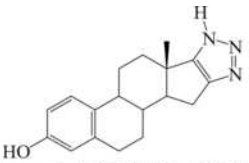
Lipophilicity descriptors (PC, ALog*P*, XLOG*P*2 and AClog*S*) are used for prediction of chromatographic behavior, biological and physicochemical characteristics of molecules [19–21]. As the lipophilicity is the main promoter regarding passive transport through biological membranes, it is a crucial characteristic when it comes to pharmacological behavior and activity of drugs. Lipophilicity descriptors are very important because they are often used as start elimination factor for design and synthesis of new pharmacologically active drugs. According to them, molecules of point of interest can be distinguished and chosen for *in vitro* and *in vivo* experiments.

Other descriptors that figure in established QSRR models are geometrical: LPMaxA (length perpendicular to the max area), MaxPR (maximal projection radius), DE (dreiding energy) and Kier3 (kappa shape index); physicochemical: FMF (the fraction of the size of the molecular framework *versus* the size of the whole molecule) and AMR (molar refractivity) and molecular bulkiness—TE (total energy [kcal/mol]).

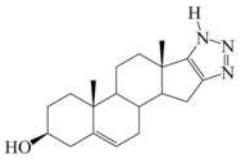
2.5. Chemometric tools

In this paper, one classification (principal component analysis) and two regression techniques (linear and multiple linear regression) were used as a chemometric tools. Principal component analysis is used in order to reduce the amount of data when there is appearance of correlation. If the variables are not correlated, this technique is not useful [22]. With this method, grouping of similar objects into clus-

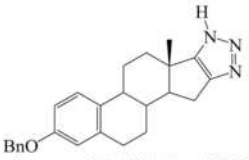
Table 1
Chemical structures and IUPAC names of investigated steroids.



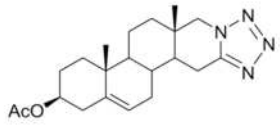
I.1. 3-hydroxyestra-1,3,5(10)-trieno[16,17-d]-1,2,3-triazole



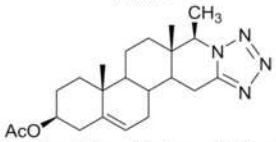
I.2. 3β-hydroxyandrost-5-eno[16,17-d]-1,2,3-triazole



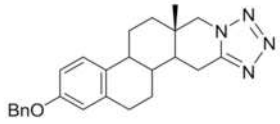
I.3. 3-benzyloxyestra-1,3,5(10)-trieno[16,17-d]-1,2,3-triazole



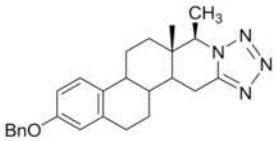
I.4. 3β-acetoxy-17-aza-17a-homoandrost-5-eno[16,17-e]tetrazole



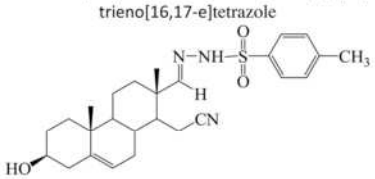
I.5. 3β-acetoxy-17-aza-17a-homo-17aβ-methyl-androst-5-eno[16,17-e]tetrazole



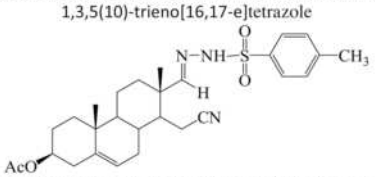
I.6. 3-benzyloxy-17-aza-17a-homoestra-1,3,5(10)-trieno[16,17-e]tetrazole



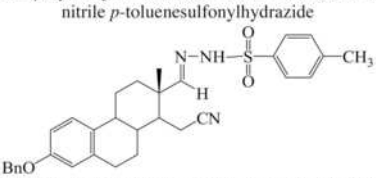
I.7. 3-benzyloxy-17-aza-17a-homo-17aβ-methylestra-1,3,5(10)-trieno[16,17-e]tetrazole



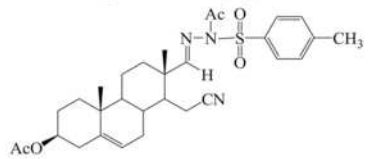
II.8. 3β-hydroxy-17-oxo-16,17-secoandrost-5-ene-16-nitrile *p*-toluenesulfonylhydrazide



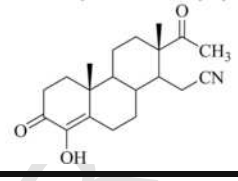
II.9. 3β-acetoxy-17-oxo-16,17-secoandrost-5-ene-16-nitrile *p*-toluenesulfonylhydrazide



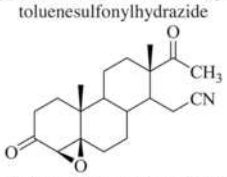
II.10. 3-benzyloxy-17-oxo-16,17-secoestra-1,3,5(10)-triene-16-nitrile *p*-toluenesulfonylhydrazide



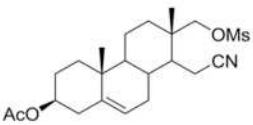
II.11. 3β-acetoxy-*N′*-acetyl-17-oxo-16,17-secoandrost-5-ene-16-nitirile *p*-toluenesulfonylhydrazide



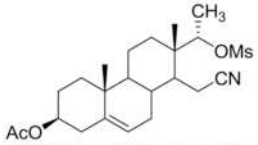
III.12. 3,17-dioxo-4-hydroxy-17-methyl-16,17-secoandrost-4-ene-16-nitrile



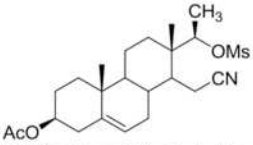
III.13. 3,17-dioxo-4β,5β-epoxy-17-methyl-16,17-secoandrostane-16-nitrile



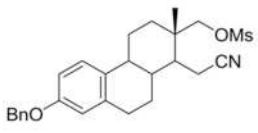
III.14. 3β-acetoxy-17-(methylsulfonyloxy)-16,17-secoandrost-5-ene-16-nitrile



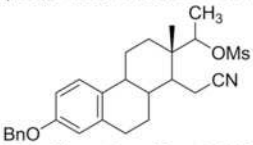
III.15. (17*S*)-3β-acetoxy-17-(methylsulfonyloxy)-17-methyl-16,17-secoandrost-5-ene-16-nitrile



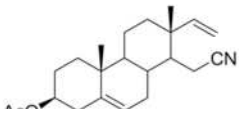
III.16. (17*R*)-3β-acetoxy-17-(methylsulfonyloxy)-17-methyl-16,17-secoandrost-5-ene-16-nitrile

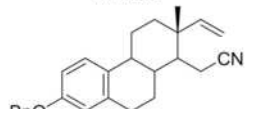


III.17. 3-benzyloxy-17-(methylsulfonyloxy)-16,17-secoestra-1,3,5(10)-triene-16-nitrile



III.18. 3-benzyloxy-17-methyl-17-(methylsulfonyloxy)-16,17-secoestra-1,3,5(10)-triene-16-nitrile





ters is possible. PCA calculates new variables (latent) combining the original variables and presents the data in a space where variables define the axes and data are projected into a few principal components (PCs) [23]. The result of PCA analysis is shown through scores and loadings plots. From scores and loadings plots, similarities among the data can be observed. Scores reflect the new coordinates of the projected objects and loadings correspond to the direction with the respect to the original variables [24]. Loadings plot present the relationships between variables and scores plot present a data overview regarding patterns or grouping. Loadings plot give information regarding variables contribution to the positioning of the objects on the scores plot.

Linear regression, as the simplest one, is used in chemometric as the first step of the regression analysis. It correlates a dependent variable and an independent variable and analyzes the relationship between them. Because it cannot correlate more than one independent variable, multiple linear regression approach has to be applied. Multiple linear regression method is used for correlation of more than one independent variable and a dependent variable. There has to be no multicollinearity between predictor variables in generated MLR models. The multicollinearity manifests through variance inflation factor (*VIF*) and values higher than 10 indicate that multicollinearity is present in established MLR model [25–27]. Variables that have high VIF values should be excluded from the model. Additionally, in MLR models the ratio of the number of data and the number of variables has to be equal or higher than 5, according to Topliss-Castello rule [28].

Established LR and MLR models were statistically validated and their predictive ability was estimated through standard, *cross*- and external validation parameters. Standard statistical parameters include: Pearson's correlation coefficient (R), determination coefficient (R^2), adjusted determination coefficient (R^2_{adj}), Fisher's test (F), root mean square error (*RMSE*) and significance level (p). *Cross*-validation statistical parameters are: leave-on-out (LOO) *cross*-validation determination coefficient (R^2_{cv}), total sum of squares (*TSS*), predicted residual error sum of squares (*PRESS*), *PRESS/TSS* ratio and standard deviation based on predicted residual sum of squares (SD_{PRESS}). External validation parameters include: Pearson's correlation coefficient of external test set (R_{test}), determination coefficient of external test set (R^2_{test}) and root mean square error of external test set ($RMSE_{test}$). All regression calculations were conducted using NCSS 2007 program [29].

3. Results and discussion

3.1. Chromatographic retention

As the most used bonded phase, octadecyl (C18) represents silica gel modified with long hydrocarbon chains with 18 carbon atoms. In this way, stationary phase gets less polar than mobile phase. As it is well known that the strong interactions occur between polar mobile phase and polar molecules. Polar molecules travel faster throughout the column and because of that, they have smaller retention than non-polar molecules that form attractions with hydrocarbon groups on the basis of van der Waals dispersion forces. Non-polar molecules are being longer retained on the column and so they have higher retention. Generally, as methanol is more polar solvent than acetonitrile, more polar molecules are going to travel faster through the column and have lower retention in methanol-water system. Since the investigated molecules are not very polar, it could be assumed that the interactions between them and acetonitrile-water mobile phase were stronger than interactions between the analyzed compounds and mobile phase with methanol. Therefore, the retention of the studied compounds in the system with acetonitrile was lower than the retention in the system with methanol. Retention data for both chromatographic systems were expressed as the logarithm of the retention factor (log$k$) values and results are shown in Supplementary data Tables S2 and S3. It can be noticed that for both systems log$k$ values are higher in chromatographic systems with lower volume fraction of organic solvent. Hence, the highest log$k$ values for acetonitrile-water system were obtained when acetonitrile 70 v/v was used, same as for methanol-water system, when methanol 70 v/v was used.

Generally, it can be noticed that as the organic solvent volume fraction increases the retention time decreases. According to the retention data, in the first group (triazole and tetrazole) it can be noticed that triazoles (compounds **I.1–I.3**) are more polar than tetrazole (**I.4–I.7**). Triazole that contain hydroxyl group (**I.1** and **I.2**) are more polar than triazole containing benzyl group (**I.3**) and they are the most polar compounds in this group. Among tetrazole, compounds with acetyl group (**I.4** and **I.5**) have higher polarity than those with benzyl group (**I.6** and **I.7**). In the second group (toluenesulfonylhydrazide) compound with hydroxyl group (**II.8**) is the most polar in its group and it is more polar than compounds with acetyl (**II.9**), benzyl (**II.10**) and double acetyl group (**II.11**). Compound containing acetyl group (**9**) has higher polarity than compounds containing benzyl (**II.10**) and double acetyl group (**II.11**). Compound with benzyl group (**II.10**) has higher polarity than compound with double acetyl group (**II.11**). The third group (nitrile and dinitrile) is the biggest group and nitrile compounds (**III.12–III.20**) are more polar than dinitrile compounds (**III.21–III.27**). Nitrile that contain hydroxyl (**III.12**) and oxo (**III.13**) group are the most polar among the nitrile. Nitrile with acetyl and mesylate group (**III.14–III.16**) are more polar than nitrile with benzyl and mesylate (**III.17** and **III.18**), acetyl (**III.19**) and benzyl (**III.20**) group. Compounds with benzyl and mesylate group (**III.17** and **III.18**) have higher polarity than compounds with acetyl (**III.19**) and benzyl group (**III.20**). Nitrile that contains acetyl group (**III.19**) has higher polarity than nitrile containing benzyl (**III.20**). Dinitrile with hydroxyl group (**III.21** and **III.22**) are more polar than dinitrile with oxo (**III.23** and **III.24**), double oxo (**III.25**), acetyl (**III.26**) and benzyl (**III.27**) group. Dinitrile that contain oxo (**III.23** and **III.24**) and double oxo (**III.25**) group are more polar than dinitrile containing acetyl (**III.26**) and benzyl (**III.27**) group. Molecule from dinitrile subgroup with acetyl (**III.26**) has higher polarity than molecule with benzyl (**III.27**). In fourth group (homoandrostan) compound with hydoxyl group (**IV.28**) has higher polarity than compound with oxo group (**IV.29**). Polarity rises along the group number and along the molecules that have the same substituents. With respect to the polarity all investigated compounds behave in accordance with their functional groups (hydroxyl > oxo > acetyl > benzyl) taking into account that hydroxyl group is the most and benzyl group the least polar functional group. All of this is in accordance with chromatographic theory.

Linear regression parameters between log$k$ values and acetonitrile volume fraction for both organic solvents are shown in Table 2.

3.2. Correlation of retention data and calculted logP

In this study 16 computationally calculated lipophilicity descriptors (Supplementary data Table S2) have been considered. Investigated steroid derivatives can be considered as lipophilic according to the obtained log$P$ values (log$P$ > 1) [6]. Values of log$k$ calculated for different chromatographic systems were correlated with computationally calculated lipophilicity descriptors. The best linear dependence was found for the relationships between log$k$ *versus* Clog$P$ and PC

Table 2
Linear regression parameters between logarithm of the retention factor (log*k*) and acetonitrile/methanol volume fraction.

Compound	$y = a \cdot x + b$							
	ZORBAX SB-C18 acetonitrile-water				ZORBAX SB-C18 methanol-water			
	a	*b*	*R*	R^2	*a*	*b*	*R*	R^2
I.1	–	–	–	–	−0.0454	3.2513	0.9999	0.9999
I.2	−0.0169	0.5839	0.9961	0.9922	−0.0442	3.3488	0.9999	0.9999
I.3	−0.0312	2.7011	0.9996	0.9993	−0.0641	5.9171	0.9997	0.9994
I.4	−0.0271	2.0990	0.9996	0.9993	−0.0512	4.2519	0.9996	0.9993
I.5	−0.0277	2.3104	0.9995	0.9990	−0.0550	4.7536	0.9994	0.9988
I.6	−0.0341	2.8585	0.9997	0.9995	−0.0598	5.2842	0.9998	0.9997
I.7	−0.0352	3.0961	0.9996	0.9993	−0.0643	5.8332	0.9997	0.9994
II.8	−0.0415	2.5228	0.9999	0.9998	−0.0607	4.4001	0.9997	0.9994
II.9	−0.0584	4.5815	0.9614	0.9243	−0.0673	5.5482	0.9994	0.9989
II.10	−0.0451	3.7767	0.9998	0.9996	−0.0742	6.4880	0.9995	0.9990
II.11	−0.0407	3.5321	0.9997	0.9995	−0.0705	6.1524	0.9995	0.9990
III.12	–	–	–	–	−0.0439	2.9020	0.9974	0.9948
III.13	−0.0307	1.8289	0.9999	0.9998	−0.0404	2.6924	0.9999	0.9999
III.14	−0.0352	2.5838	0.9999	0.9998	−0.0543	4.2709	1.0000	1.0000
III.15	−0.0352	2.6476	0.9999	0.9998	−0.0555	4.4416	0.9999	0.9998
III.16	−0.0343	2.6658	0.9998	0.9997	−0.0557	4.5713	0.9998	0.9996
III.17	−0.0412	3.3004	0.9998	0.9997	−0.0620	5.2761	0.9997	0.9994
III.18	−0.0413	3.4247	0.9999	0.9998	−0.0642	5.6133	0.9997	0.9994
III.19	−0.0341	3.1325	0.9997	0.9995	−0.0606	5.5063	0.9997	0.9994
III.20	−0.0416	3.9202	0.9998	0.9996	−0.0715	6.7420	0.9998	0.9996
III.21	–	–	–	–	–	–	–	–
III.22	−0.0560	3.2884	0.9966	0.9933	−0.0528	3.3500	0.9973	0.9947
III.23	−0.0383	2.2007	0.9988	0.9976	−0.0463	2.8926	0.9980	0.996
III.24	−0.0331	1.8315	0.9997	0.9995	−0.0499	3.2084	0.9972	0.9945
III.25	−0.0369	2.1405	0.9992	0.9985	−0.0478	3.0236	0.9983	0.9967
III.26	−0.0331	2.4725	0.9999	0.9998	−0.0518	4.0894	0.9997	0.9995
III.27	−0.0393	3.1988	0.9998	0.9997	−0.0604	5.1771	0.9998	0.9996
IV.28	–	–	–	–	–	–	–	–
IV.29	−0.0258	1.5705	0.9999	0.9999	−0.0394	2.8482	1.000	1.0000

values (Fig. 1). Good correlation of log*k* values with lipophilicity descriptors indicate that chromatographic retention (log*k*) can be used as a chromatographic lipophilicity parameter.

3.3. Principal component analysis

For the PCA, as input data, log*k* values for all five volume fractions for both organic solvents were used in order to group studied steroid derivatives regarding their similiarities. The PCA analysis was conducted using program Statistica v. 10 [30].

Obtained PCA model for log*k* for acetonitrile-water system is shown through two principal components. These two principal components describe 99.94% of total variance. PC1 contributes with 99.52% and PC2 with 0.42% of total variance. Scores and loadings plots for this PCA analysis are shown in Supplementary data (Fig. S1). From loadings plot, it can be noticed that on the position of the investigated compounds on the scores plot all five volume fractions have almost the same influence, regarding both PC1 and PC2. All five volume fractions have negative coefficients of latent variables regarding PC1 axis. Volume fractions marked as ACN 70 and ACN 72.5 have positive and volume fractions marked as ACN 75, ACN 77.5 and ACN 80 have a negative coefficient of latent variables regarding PC2 axis. From scores plot, it can be noticed that investigated compounds are grouped according to their polarity, regarding PC1 axis. On the positive end of PC1 axis more polar (with oxo and hydroxyl group) compounds are positioned and less polar (with acetyl and benzyl group) compounds are positioned on the negative and of PC1 axis.

Regarding PCA model obtained for log*k* for methanol-water system, it is expressed through two principal components. These two

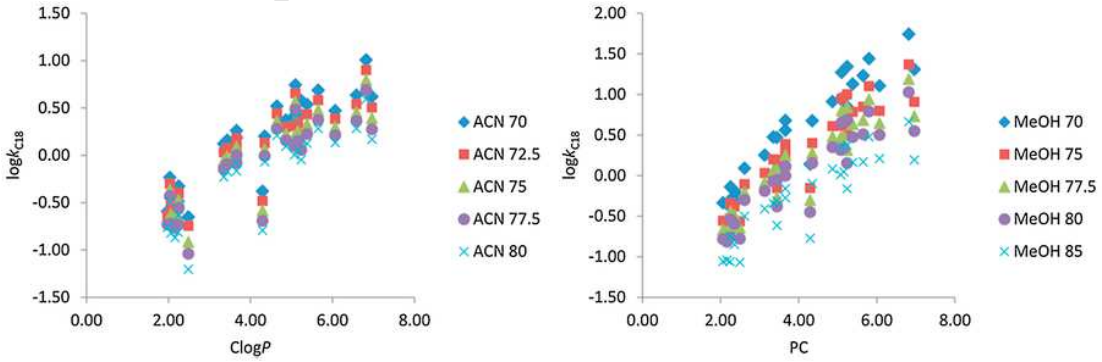


Fig. 1. Experimentally obtained log*k* *versus* calculated log*P* and PC values.

principal components describe 99.98% of total variance. PC1 contributes with 99.72% and PC2 with 0.26% of total variance. Scores and loadings plots are presented in Supplementary data (Fig. S2). From loadings plot, it can be noticed that on the position of the investigated compounds on the scores plot all five volume fractions have almost the same influence, regarding both PC1 and PC2. All five volume fractions have negative coefficients of latent variables relative to PC1 axis. Volume fractions marked as MeOH 70, MeOH 75 and MeOH 77.5 have positive and volume fractions marked as MeOH 80 and MeOH 85 have negative coefficient of latent variables relative to PC2 axis. According to PC1 axis, from scores plot it can be noticed that investigated compounds are grouped according to their polarity. On the positive end of PC1 axis more polar compounds that contain oxo and hydroxyl functional groups are positioned. Compounds that have acetyl and benzyl functional group are less polar and they are possitioned toward the negative end of PC1 axis. The results of conducted PCA analysis indicate that polarity is presponsible for discrimination between the investigated compounds.

3.4. Interpretation of selected QSRR models

QSRR-LR and QSRR-MLR models were obtained using NCSS 2007 program [29]. The best established models were selected and presented in this paper. In order to obtain statistically valid and meaningful models, internal and external validation was conducted. Investigated compounds were divided into two sets, calibration and external test set. Calibration set for acetonitrile-water system consists of 20 compounds – 80% of total number of compounds (**I.2**, **I.3**, **I.4**, **I.6**, **I.7**, **II.8**, **II.9**, **II.11**, **III.14**, **III.16**, **III.17**, **III.18**, **III.19**, **III.20**, **III.22**, **III.23**, **III.24**, **III.25**, **III.26**, **III.27**). For methanol-water system, calibration set consists also of 22 compounds − 80% of total number of compounds (**I.1**, **I.2**, **I.3**, **I.4**, **I.6**, **I.7**, **II.8**, **II.11**, **III.12**, **III.14**, **III.16**, **III.17**, **III.18**, **III.19**, **III.20**, **III.21**, **III.22**, **III.23**, **III.24**, **III.25**, **III.26**, **III.27**). External test set is the same for both mobile phases and it consists of 5 compounds – 20% of total number of compounds (**I.5**, **II.10**, **III.13**, **III.15**, **IV.29**). Molecules **I.1**, **III.12**, **III.21** and **IV.28** were excluded from calculations regarding system with acetonitrile and **III.21** and **IV.28** from system with methanol, as their retention times were not defined under given chromatographic conditions. In this paper, the best QSRR models are shown, for four linear regression and four multiple linear regression models. Obtained LR models:

$$\text{LR1: } \log k_{\text{C18ACN70}} = 0.3009\ (\pm 0.0330)\ \mathbf{CLog}\boldsymbol{P} - 1.1026\ (\pm 0.1498) \quad (2)$$

$$\text{LR2: } \log k_{\text{C18ACN80}} = 0.2871\ (\pm 0.0377)\ \mathbf{CLog}\boldsymbol{P} - 1.4187\ (\pm 0.1714) \quad (3)$$

$$\text{LR3: } \log k_{\text{C18MeOH70}} = 0.4308\ (\pm 0.0314)\ \mathbf{PC} - 1.2050\ (\pm 0.1389) \quad (4)$$

$$\text{LR4: } \log k_{\text{C18MeOH80}} = 0.3696\ (\pm 0.0312)\ \mathbf{PC} - 1.5145\ (\pm 0.1380) \quad (5)$$

Shown LR-QSRR models are obtained using calibration test. Next step was internal and external validation. Statistical parameters for linear regression models are shown in Table 3. From this table, it can be concluded that all four obtained LR models are statistically valid

Table 3
Statistical parameters of internal and external validation for QSRR-LR models.

Parameters	LR1	LR2	LR3	LR4
	C18 ACN 70	C18 ACN 80	C18 MeOH 70	C18 MeOH 80
R	0.9068	0.8735	0.9509	0.9357
R^2	0.8223	0.7630	0.9042	0.8756
R^2_{adj}	0.8124	0.7498	0.8994	0.8694
F	83.30	57.94	188.79	140.78
$RMSE$	0.2218	0.2537	0.2064	0.2051
p	0.000000	0.000000	0.000000	0.000000
R^2_{cv}	0.7902	0.7201	0.8896	0.8569
TSS	4.9820	4.8888	8.8984	6.7641
$PRESS$	1.0452	1.3682	0.9828	0.9681
$PRESS/TSS$	0.2098	0.2799	0.1104	0.1431
SD_{PRESS}	0.2286	0.2616	0.2114	0.2098
R_{test}	0.9675	0.9304	0.9760	0.9585
R^2_{test}	0.9360	0.8656	0.9526	0.9188
$RMSE_{test}$	0.1167	0.1491	0.1484	0.1476

by standard, *cross*- and external statistical parameters. As R^2 have values significantly higher than 0.64 that confirms a very strong correlation between variables. High values of R^2_{adj} (higher than 0.70) and low values of *RMSE* also indicate good statistical validity. Predictive ability of established models is confirmed by parameters of *cross*- and external validation. Relatively low *PRESS* and SD_{PRESS} values and high values of R^2_{cv} (higher than 0.60) also contribute to the quality of established models. Parameters of external validation for all four LR models additionally confirm the statistical validity and predictive ability of these models. In all four models, lipophilicity descriptors CLog*P* and PC have a dominant effect and a positive regression coefficient affecting the log*k*. This shows that retention behavior of investigated compounds in RP HPLC system depends on their lipophilicity that has the greatest influence on the molecule distribution between stationary and mobile phase.

The predictive power of established models was tested by two graphical methods. Fig. 2 shows experimental log*k versus* predicted values and experimental log*k* values *versus* residuals for four LR models. Comparison of experimental and predicted data indicates that there is good fit of data for all four LR models. Additionally, it can be noticed that the residuals are randomly distributed around y = 0 axis which indicates that prediction error is unpredictable. According to all given parameters and graphical view, as the best models, equations LR3 and LR4 can be selected. It can be noticed that for the same volume fractions of organic solvents in mobile phases, better models for log*k* prediction were derived for system methanol-water. Also, lower volume fraction of organic solvent in mobile phase gives better QSRR models. As LR models give insight only on influence of one factor on retention of the investigated compounds, it was necessary to apply multivariate regression method. Simplicity and possibility of mechanistic interpretation is the advantage of LR models but given that chromatography is very complex process, the influence of the more than one factor on the retention was investigated.

Molecular descriptor selection for QSRR-MLR models was done by stepwise selection (SS) in NCSS 2007 program [29]. In order to avoid the over-parameterization of the mathematical model and correlation between descriptors it is very important to define the number of independent variables in the model [31]. In accordance with Topliss-Costello rule, maximum number of molecular descriptors in model is four. Established MLR models were free of multicollinearity. The *VIF* values were calculated for independent variables in each model. *Cross*-validation was done using leave-one-out (LOO) method. Obtained MLR models for log*k* prediction:

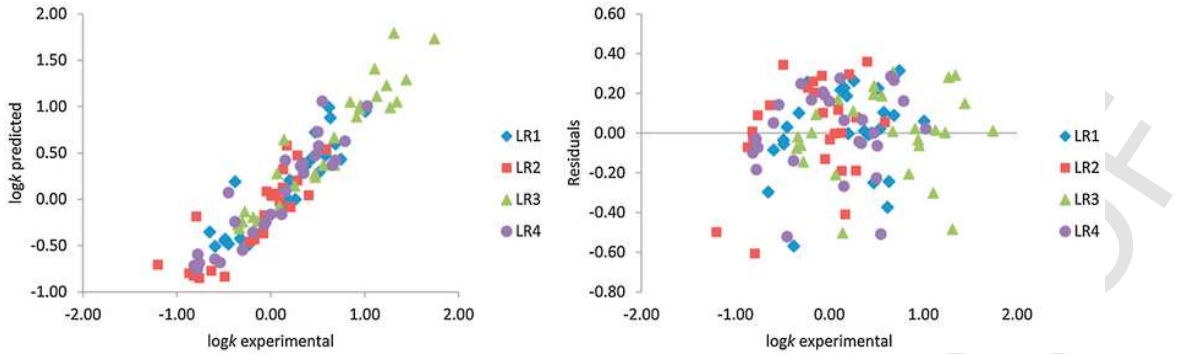


Fig. 2. Experimental log*k versus* predicted and experimental log*k versus* residuals for four LR models.

MLR1: $\log k_{C18ACN70}$ = 0.4480 (±0.0450) **PC** + 0.0798 (±0.0252) **LPMaxA** − 2.4912 (±0.6428) **FMF** − 0.0691 (±0.0386) **Kier3** − 1.2474 (±0.2907) (6)

MLR2: $\log k_{C18ACN80}$ = 0.4495 (±0.0962) **ALog*P*** + 0.0091 (±0.0032) **TE** + 0.2006 (±0.0534) **PC** + 0.1587 (±0.0490) **MaxPR** − 3.1153 (±0.3432) (7)

MLR3: $\log k_{C18MeOH70}$ = 0.0186 (±0.0035) **TE** + 0.4079 (±0.0400) **PC** − 0.0060 (±0.0013) **DE** + 0.1188 (±0.0381) **MaxPR** − 2.3605 (±0.2453) (8)

MLR4: $\log k_{C18MeOH80}$ = −0.2536 (±0.0816) **XLOG*P*2** − 0.2845 (±0.1236) **AClog*S*** + 0.0005 (±0.0019) **AMR** + 0.4558 (±0.0565) **PC** − 2.1678 (±0.4264) (9)

All established MLR models contain one or more lipophilicity descriptors (PC, ALog*P*, XLOG*P*2 and AClog*S*). The second most common group of descriptors in these MLR models, that can be found in three models, are geometrical – LPMaxA, Kier3, MaxPR and DE. In the establishment of these models also participate following molecular descriptors: physicochemical − FMF and AMR and molecular bulkiness – TE. As the most influential lipophilicity descriptors PC and ALog*P* have positive and XLOG*P*2 and AClog*S* have negative influence on log*k*. All four MLR models were validated by standard, *cross*- and external statistical parameters. Statistical parameters for multiple linear regression models are shown in Table 4. According to *VIF* values that were calculated for every independent variable in model, multicollinearity is below the given limit (*VIF* < 10). Very high values of R^2 (in range from 0.9214 to 0.9700) and R^2_{adj} (from 0.9029 to 0.99629) indicate very good predictive ability of generated MLR models. Low values of *cross*-validation statistical parameters as *PRESS* values and SD_{PRESS} and high values of R^2_{cv} (from 0.8666 to 0.9518) contribute to good statistical quality of models. The external test set gives the most confident information about the predictive power of established models. Parameters of external validation R^2_{test} higher than 0.8397 and $RMSE_{test}$ lower than 0.2074 indicate very strong correlation between variables. In Fig. 3 is shown graphical test

Table 4
Statistical parameters of internal and external validation for QSRR-MLR models.

Parameters	MLR1	MLR2	MLR3	MLR4
	C18 ACN 70	C18 ACN 80	C18 MeOH 70	C18 MeOH 80
R	0.9650	0.9645	0.9849	0.9599
R^2	0.9313	0.9303	0.9700	0.9214
R^2_{adj}	0.9130	0.9117	0.9629	0.9029
F	50.83	50.06	137.18	49.79
RMSE	0.1511	0.1507	0.1254	0.1769
p	0.000000	0.000000	0.000000	0.000000
VIF	3.5 $_{PC}$	1.2 $_{ALogP}$	2.6 $_{TE}$	6.8 $_{XLOGP2}$
	1.1 $_{LPMaxA}$	1.4 $_{TE}$	4.4 $_{PC}$	4.6 $_{AClogS}$
	2.5 $_{FMF}$	4.9 $_{PC}$	2.2 $_{DE}$	1.1 $_{AMR}$
	2.4 $_{Kier3}$	4.2 $_{MaxPR}$	4.1 $_{MaxPR}$	4.4 $_{PC}$
R^2_{cv}	0.8666	0.8716	0.9518	0.8727
TSS	4.9820	4.8888	8.8984	6.7641
PRESS	0.6646	0.6278	0.4285	0.8609
PRESS/TSS	0.1334	0.1284	0.0482	0.1273
SD_{PRESS}	0.1823	0.1772	0.1396	0.1978
R_{test}	0.9760	0.9862	0.9985	0.9164
R^2_{test}	0.9526	0.9726	0.9970	0.8397
$RMSE_{test}$	0.1004	0.0673	0.0376	0.2074

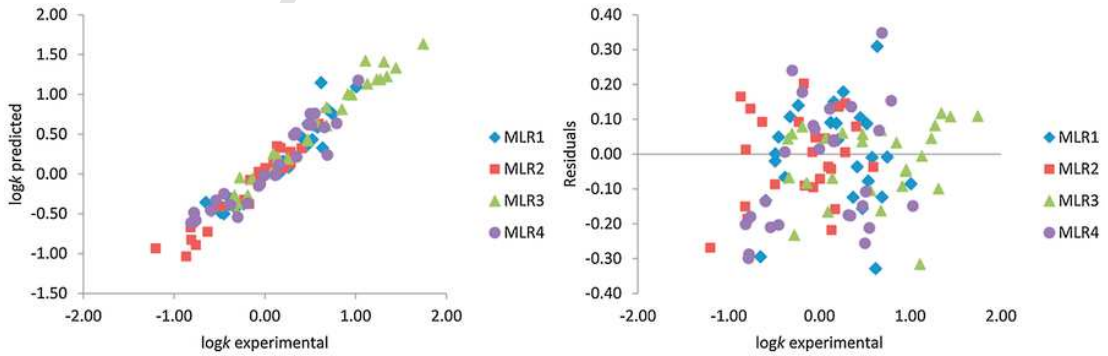


Fig. 3. Experimental log*k versus* predicted and experimental log*k versus* residuals for four MLR models.

of the predictive power of established models through experimental log*k* *versus* predicted values and experimental log*k* values *versus* residuals for four MLR models. Comparison of experimental and predicted data indicates that there is good fit of data for all models. Additionally, it can be noticed that the residuals are randomly distributed around y = 0 axis which also confirms the predictive power of established models. The same trend can be noticed at MLR as at the LR models that better models were derived for methanol-water systems and that lower volume fraction of organic solvent in mobile phase gives better MLR-QSRR models.

Established MLR models have better standard, *cross*- and external validation parameters than LR models. That gives them advantage in retention lipophilicity prediction for investigated steroid derivatives. QSRR modeling of chromatographic lipophilicity was performed in order to obtain good physicochemical profiles of investigated compounds as future drug candidates of biomedical importance.

4. Conclusion

The QSRR modeling was successfully carried out on the set of twenty-nine (triazole and tetrazole, toluenesulfonylhydrazide, nitrile and dinitrile and dione) selected steroid derivatives. According to PCA method, the best discrimination factor between the investigated compounds is their polarity. The most influential molecular descriptors in QSRR modeling were lipophilicity and geometrical descriptors. All of these descriptors affect molecules ability to pass into the cells and reach receptor site in the organism. The best LR-QSRR and MLR-QSRR models were selected and confirmed by comprehensive statistical validation. It was noticed that chromatographic systems with methanol-water and lower volume fractions of organic component in mobile phase were better for log*k* prediction. Chromatographic lipophilicity of investigated compounds as future drug candidates of biomedical importance successfully correlates with *in silico* lipophilicity descriptors. They can be presented as function of retention value log*k* and in that way reflect the lipophilicity of investigated steroid derivatives.

Acknowledgments

These results are the part of the projects Nos. 172025, 31055 and 172012, supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

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.jpba.2016.11.015.

References

[1] L. Di, E. Kerns, Profiling drug-like properties in discovery research, Curr. Opin. Chem. Biol. 7 (2003) 402–408.
[2] E.H. Kerns, L. Di, Pharmaceutical profiling in drug discovery, Drug. Discov. Today 8 (2003) 316–323.
[3] C. Hansch, T. Fujita, ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (1964) 1616–1626.
[4] T. Tuzimski, K. Sztanke, Retention data for some carbonyl derivatives of imidazo[2,1-c][1,2,4]triazine in reversed-phase systems in TLC and HPLC and their use for determination of lipophilicity. Part 1. Lipophilicity of 8-aryl-3-phenyl-6,7-dihydro-4H-imidazo[2,1-c][1,2,4]triazin-4-ones, J. Planar. Chromatogr. 18 (2005) 274–281.
[5] A. Nasal, D. Siluk, R. Kaliszan, Chromatographic retention parameters in medicinal chemistry and molecular pharmacology, Curr. Med. Chem. 10 (2003) 381–426.
[6] S.Z. Kovačević, S.O. Podunavac-Kuzmanović, L.R. Jevrić, P.T. Jovanov, E.A. Djurendić, J.J. Ajduković, Comperhensive QSRR modeling as a starting point in characterization and further development of anticancer drugs based on 17α-picolyl and 17(E)-picolinylidene androstane structures, Eur. J. Pharm. Sci. 93 (2016) 1–10.
[7] M. Sakač, A. Gaković, S. Stojanović, E. Djurendić, V. Kojić, G. Bogdanović, K. Penov-Gaši, Synthesis and biological evaluation of a series of A,B-ring modified 16,17-secoandrostane derivatives, Bioorg. Chem. 36 (2008) 128–132.
[8] M.N. Sakač, A.R. Gaković, J.J. Csanádi, E.A. Djurendić, O. Klisurić, V. Kojić, G. Bogdanović, K.M. Penov-Gaši, An intramolecular one-pot synthesis of steroidal triazoles via 1,3-dipolar cycloadditions of in situ generated diazo compounds, Tetrahedron Lett. 50 (2009) 4107–4109.
[9] K.M. Penov-Gaši, A.M. Oklješa, E.T. Petri, A.S. Ćelić, E.A. Djurendić, O.R. Klisurić, J.J. Csanádi, G. Batta, A.R. Nikolić, D.S. Jakimov, M.N. Sakač, Selective antitumor activity and ERš molecular docking studies of newly synthesized D-homo fused steroidal tetrazoles, MedChemComm 4 (2013) 317–323.
[10] A.R. Nikolić, E.T. Petri, O.R. Klisurić, A.S. Ćelić, D.S. Jakimov, E.A. Djurendić, K.M. Penov-Gaši, M.N. Sakač, Synthesis and anticancer cell potential of steroidal 16,17-seco-16,17a-dinitriles: identification of a selective inhibitor of hormone-independent breast cancer cells, Bioorgan. Med. Chem. 23 (2015) 703–711.
[11] Chem Axon, Ltd. http://www.chemaxon.com/.
[12] Cambridge Soft Corporation, Perkin Elmer Inc., 2012. ChemBioOffice Software Version 12.0. http://www.cambridgesoft.com.
[13] PaDEL Descriptors. <http://www.nus.edu.sg/>
[14] Virtual Computational Chemistry Laboratory ALOGPS 2.1 Online Program. http://www.vcclab.org/.
[15] PreADMET Software. http://www.preadmet.bmdrc.org/ .
[16] Molinspiration Cheminformatics. http://www.molinspiration.com.
[17] S.Z. Kovačević, S.O. Podunavac-Kuzmanović, L.R. Jevrić, E.A. Djurendić, J.J. Ajduković, Non-linear assessment of anticancer activity of 17-picolyl and 17-picolinylidene androstane derivatives—chemometric guidelines for further syntheses, Eur. J. Pharm. Sci. 62 (2014) 258–266.
[18] O. Deeb, Correlation ranking and stepwise regression procedures in principal components artificial neural networks modeling with application to predict toxic activity and human serum albumin binding affinity, Chemom. Intell. Lab. Syst. 104 (2010) 181–194.
[19] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, Quantitative correlation of physical and chemical properties with chemical structure: utility prediction, Chem. Rev. 110 (2010) 5714–5789.
[20] J.M. Pallicer, J. Sales, M. Rosés, C. Ráfols, E. Bosch, Lipophilicity assessment of basic drugs (log $P_{o/w}$ determination) by a chromatographic method, J. Chromatogr. A 1218 (2001) 6356–6368.
[21] N.P. Milošević, S.S. Stojanović, K. Penov-Gaši, N. Perišić-Janjić, R. Kaliszan, Reversed- and normal-phase liquid chromatography in quantitative structure retention-property relationships of newly synthesized seco-androstene derivatives, J. Pharm. Biomed. Anal. 88 (2014) 636–642.
[22] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, 6th ed., Pearson Education Limited, Harlow, UK, 2010221–247.
[23] S. Kovačević, S. Podunavac-Kuzmanović, N. Zec, S. Papović, A. Tot, S. Dožić, M. Vraneš, G. Vastag, S. Gadžurić, Computational modeling of ionic liquids density by multivariate chemometrics, J. Mol. Liq. 214 (2016) 276–282.
[24] J. Trifković, F. Andrić, R. Ristivojević, D. Andrić, Ž.Lj. Tešić, D.M. Milojković Opsenica, Structure-retention relationship study of arylpiperazines by linear multivariate modeling, J. Sep. Sci. 33 (2010) 2619–2628.
[25] D.W. Marquardt, R.D. Snee, Ridge regression in practice, Am. Stat. 29 (1975) 3–19.
[26] R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, Qual. Quant. 41 (2007) 673–690.
[27] T.M. Young, L.B. Shaffer, F.M. Guess, H. Bensmail, R.V. Léon, A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard, For. Prod. J. 58 (2008) 39–48.
[28] J.G. Topliss, R.J. Costello, Chance correlation in structure-activity studies using multiple regression analysis, J. Med. Chem. 15 (1972) 1066–1068.
[29] Hintze, J., 2007. NCSS 2007, NCSS, LLC., Kaysville, Utah, USA. http://www.ncss.com/.
[30] StatSoft Inc., 2011. STATISTICA (Data Analysis Software System), Version 10. http://www.statsoft.com/.
[31] N. Minovski, A. Jezierska-Mazzarello, M. Vračko, T. Šolmajer, Investigation of 6-fluoroquinolones activity against Mycobacterium tuberculosis using theoretical molecular descriptors: a case study, Cent. Eur. J. Chem. 9 (2011) 855–866.