TITLE: Comprehensive QSRR modeling as a starting point in characterization and further development of anticancer drugs based on 17α-picolyl and 17(E)-picolinylidene androstane structures

AUTHORS: Strahinja Z. Kovačević, Sanja O. Podunavac Kuzmanović, Lidija J. Jevrić, Pavle T. Jovanov, Evgenija A. Đurendić, Jovana J. Ajduković

Elsevier Editorial System(tm) for European Journal of Pharmaceutical Sciences
Manuscript Draft

Manuscript Number: EJPS-D-16-00507R2

Title: Comprehensive QSRR modeling as a starting point in characterization and further development of anticancer drugs based on 17α-picolyl and 17(E)-picolinylidene androstane structures

Article Type: Research Paper

Keywords: Androstane derivatives; Anticancer compounds; Chemometrics; Lipophilicity; Molecular Modeling; Reversed-Phase HPLC.

Corresponding Author: Dr. Strahinja Z Kovačević, PhD

Corresponding Author's Institution: University of Novi Sad, Faculty of Technology Novi Sad

First Author: Strahinja Z Kovačević, PhD

Order of Authors: Strahinja Z Kovačević, PhD; Sanja O Podunavac-Kuzmanović, PhD; Lidija R Jevrić, PhD; Pavle T Jovanov, PhD; Evgenija A Djurendić, PhD; Jovana J Ajduković, PhD

Manuscript Region of Origin: SERBIA

Abstract: The selection of the most promising anticancer compounds from the pool of the huge number of synthesized molecules is a quite complex task. There are many compounds characterization approaches which can suggest the best structural features of a molecule with the highest antiproliferative effect on the certain type of cancer cell lines. One of these approaches is the lipophilicity determination of compounds and the analysis of its correlation with the anticancer activity. Since the importance of the lipophilicity is underlined in many earlier studies, this study is focused on determination of lipophilicity of previously synthesized 17α-picolyl and 17(E)-picolinylidene androstane derivatives by using reversed-phase high performance liquid chromatography (RP-HPLC) as a very fast, effective and relatively cheap method. Determination of the chromatographic lipophilicity of the studied androstanes can be considered as the part of their physicochemical characterization, which is a very important step in their further selection as drug candidates. The present study does not neglect the in silico approach. The determined chromatographic lipophilicity was analyzed by quantitative structure-retention relationship (QSRR) approach in order to reveal which molecular characteristics contribute mostly to the typical behavior of the androstanes in the applied chromatographic system, and thus to their lipophilicity. Classical statistical approach and Sum of Ranking Differences method were used for selection of the best QSRR models which should be used in prediction of chromatographic lipophilicity of studied androstane derivatives.

*Abstract

ABSTRACT:

The selection of the most promising anticancer compounds from the pool of the huge number of synthesized molecules is a quite complex task. There are many compounds characterization approaches which can suggest the best structural features of a molecule with the highest antiproliferative effect on the certain type of cancer cell lines. One of these approaches is the lipophilicity determination of compounds and the analysis of its correlation with the anticancer activity. Since the importance of the lipophilicity is underlined in many earlier studies, this study is focused on determination of lipophilicity of previously synthesized 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives by using reversed-phase high performance liquid chromatography (RP-HPLC) as a very fast, effective and relatively cheap method. Determination of the chromatographic lipophilicity of the studied androstanes can be considered as the part of their physicochemical characterization, which is a very important step in their further selection as drug candidates. The present study does not neglect the *in silico* approach. The determined chromatographic lipophilicity was analyzed by quantitative structure-retention relationship (QSRR) approach in order to reveal which molecular characteristics contribute mostly to the typical behavior of the androstanes in the applied chromatographic system, and thus to their lipophilicity. Classical statistical approach and Sum of Ranking Differences method were used for selection of the best QSRR models which should be used in prediction of chromatographic lipophilicity of studied androstane derivatives.

KEYWORDS:

Androstane derivatives, Anticancer compounds, Chemometrics, Lipophilicity, Molecular Modeling, Reversed-Phase HPLC

Title:

Comprehensive QSRR modeling as a starting point in characterization and further development of anticancer drugs based on 17α-picolyl and 17(*E*)-picolinylidene androstane structures

Authors:

1. Strahinja Z. Kovačević [a,*]
2. Sanja O. Podunavac-Kuzmanović [a]
3. Lidija R. Jevrić [a]
4. Pavle T. Jovanov [b]
5. Evgenija A. Djurendić [c]
6. Jovana J. Ajduković [c]

*Corresponding author: e-mail: strahko@uns.ac.rs; phone: +381 64 2839686; fax: +381 21 450413

Affiliations:

a University of Novi Sad, Faculty of Technology, Department of Applied and Engineering Chemistry, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia

b University of Novi Sad, Institute of Food Technology, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia

c University of Novi Sad, Faculty of Sciences, Department of Chemistry, Biochemistry and Environmental Protection, Trg Dositeja Obradovića 3, 21000 Novi Sad, Serbia

Abbreviations:

ADME – Absorption, Distribution, Metabolism, Excretion
ANN – Artificial Neural Networks
AR- – Androgen Receptor negative
BP – Boiling Point
CP – Critical Pressure
CRRN – Comparison of Ranks by Random Numbers
CT – Critical Temperature
DE – Dreiding Energy
EP – Electrostatic potential surface
GSA – Global Sensitivity Analysis
HCA – Hierarchical Cluster Analysis
HILI – Hydrophilic-lipophilic surface
HOMO – Highest Occupied Molecular Orbital
HPLC – High Performance Liquid Chromatography
LOO – Leave-One-Out
LR – Linear Regression
LUMO – Lowest Unoccupied Molecular Orbital
LV – Latent Variables
MLR – Multiple Linear Regression
MP – Melting Point
PCA – Principal Component Analysis
PCR – Principal Component Regression
PLS – Partial Least Squares
PR – Polynomial Regression
PRESS – Predicted Residual Sum of Squares
PSA – Polar Surface Area
QSAR – Quantitative Structure Activity Relationship
QSRR – Quantitative Structure Retention Relationship
RMSE – Root Mean Square Error
SASA – Solvent Accessible Surface Area
SRD – Sum of Ranking Differences
SS – Stepwise Selection
TSS – Total Sum of Squares
vdWSA – van der Waals Surface Area
VIF – Variance Inflation Factor
VIP – Variance Importance in Projection
WWR – Wald-Wolfowitz runs test

1. INTRODUCTION

Cancer, as one of still growing health problems in the World, is for decades in the focus of many medical, chemical and interdisciplinary studies [1-5]. A number of compounds, which were assumed to have anticancer activity, have been synthesized by organic chemists, mostly based on trial-and-error approach. However, lots of them have been rejected not only on the basis of their negligible antiproliferative activity, but also due to mutagenic and/or teratogenic effect. In this context, it is necessary to highlight that this approach can be time-consuming and in many cases unavailing [6].

The first step in the drug selection candidates is certainly crucial for further drug development. In the last two decades the computational modeling and chemometrics (particularly quantitative structure-activity relationship – QSAR approach) have become one of the essential components in the first stage of drug selection candidates [7]. They can give significant guidelines for the synthesis of new compounds with desired biological activity or predict the biological activity of already synthesized compounds.

In often too long drug discovery process every information about physicochemical properties of potential drug candidates is precious. Besides the computational characterization, the experimental analysis of certain physicochemical parameters is still needed.

Lipophilicity (often expressed as log*P* parameter) is one of the most analyzed physicochemical parameters of small biologically active compounds [8]. It is generally related to the ability of a compound to achieve its site of action in a biological system by passing through the lipophilic cell membranes. The determination of the lipophilicity of small molecules can be achieved experimentally and by computational techniques. Nowadays, the computational approach is often used than the classical experimental methods, such as shake-flask method, potentiometric titrations, filter-probe method, etc [9,10]. Chromatographic techniques have become a very popular experimental method for the lipophilicity determination (so-called chromatographic lipophilicity) [11-13]. Chromatographic lipophilicity is often in a very good relation with the computational lipophilicity [14,15]. Reversed-phase high performance liquid chromatography (RP-HPLC) with C18 stationary phase and strictly controlled chromatographic conditions is one of the most reliable experimental techniques for lipophilicity determination of steroidal compounds [16]. RP-HPLC, as a very fast, effective and relatively cheap method, has been applied for determination of the lipophilicity of many compounds [17,18]. The determined chromatographic lipophilicity was further used in quantitative structure-retention relationship (QSRR) studies aimed to correlate the molecular features with the retention behavior in the applied chromatographic system. Particularly, the basics and importance of QSRR analysis was pointed out in review papers by Héberger [19] and Kaliszan [20].

The present study is focused on the RP-HPLC determination of lipophilicity of earlier synthesized 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives, computational modeling of their structures and a comprehensive QSRR analysis. These derivatives express significant antiproliferative activity toward androgen-receptor negative (AR-) prostate cancer cells, PC-3 [21-23], and this study can be considered as the contribution to their physicochemical characterization and to selection of androstanes with the most prominent anticancer activity, which should be forwarded to the more detailed *in vivo* biological examinations as potential anticancer medicines.

2. MATERIAL AND METHODS

2.1. Studied compounds and computational modeling

The analyzed series of eleven 17α-picolyl and thirteen 17(*E*)-picolinylidene androstane derivatives have been synthesized according to the procedures described earlier [21-23]. The determination of their structures can be found in literature [21-23]. The 2D molecular structures of the androstane derivatives are given in Table 1. The compounds contain different substituents in positions 3, 4, 5 and 6, such as hydroxyl, acetoxy, epoxy, oxo, nitro and methoxy groups. These groups, besides the 17α-picolyl and 17(*E*)-picolinylidene groups which are connected to D ring of steroidal core, strongly affect the total lipophilicity. Since the 17α-picolyl and thirteen 17(*E*)-picolinylidene androstane derivatives have shown significant *in vitro* antiproliferative activity toward prostate cancer, breast cancer and colon cancer, they have become an interesting basis of further development of androstane-based anticancer drugs [21-23].

Computational modeling of the structures of studied androstanes was carried out by using the suitable software for molecular design. 2D and 3D structures were drawn by using ChemBioDraw v. 14 and ChemBio3D v. 14 programs [24]. 3D structures were energetically minimized applying molecular mechanics force field method (MM2), and the cutoff for structure optimization was set at a gradient of 0.0001 kcal/Åmol. Modeling of hydrophilic-lipophilic surfaces and surface of electrostatic potential was done by Bioluminate® program [25]. Highest occupied molecular orbitals (HOMO) and lowest unoccupied molecular orbitals (LUMO) and some physicochemical and topological descriptors were modeled by ChemBio3D v. 14 program. The other programs used for the calculation of molecular descriptors are the following: PreAMDET [26], Simulation Plus ADMET Predictor™ [27], ALOGPS 2.1. [28], Avogadro 1.0 [29], ADRIANA.Code [30], Marvin Sketch 6.1 [31] and Parameter Client Program [32]. The number of molecular descriptors (physicochemical, topological, lipophilicity and absorption, distribution, metabolism and excretion – ADME) which were used in the analysis is 143. The list of molecular descriptors is given in Supplementary data (Table S1).

2.2. Chromatographic analysis

Prior to chromatographic analysis, the samples of the analyzed androstanes were diluted in methanol (BAKER HPLC Analyzed® HPLC gradient grade) in the concentration of 2 mg/ml. The prepared solutions were afterwards filtered by Captiva Econofilter (nylon membrane, 25 mm diameter, 0.45 μm pore size, 1000/pk). Isocratic chromatographic analysis was carried out applying RP-HPLC system (Agilent Technologies 1200 Series HPLC) with ZORBAX Eclipse XDB-C18 column (4.6 x 50 mm, 1.8 micron) and Diode Array Detector (DAD) and Evaporative Light Scattering Detector (ELSD). Two mobile phases were used, consisting of methanol and water mixtures in the ratio of 70 : 30 and 90 : 10 with the flow rate of 0.600 ml/min. The pH of mobile phases was maintained on 7 by 0.01 M phosphate buffer (Na_2HPO_4, KH_2PO_4, Lach-Ner, *p.a.*). The column temperature was 25 °C. Injection volume was set at 10 μL. The detection of the compounds was done by DAD detector on 210 and 230 nm. The temperature and pressure of ELSD detector were 40 °C and 4.5 bar, respectively. The retention of the compounds was measured in triplicate. The signal from ELSD detector was the control detection signal. The capacity factor (k) was calculated by the following equation:

$$k = (t_a - t_m) / t_m \quad (1)$$

t_a – retention time of a compound (detection on 210 nm, DAD), t_m – dead time (the first disturbance on the chromatogram)

In the present study, the chromatographic lipophilicity of the analyzed androstane derivatives was defined as log$k$. The extrapolation of the retention factor to pure water or buffer was not done, since this definition of lipophilicity can be considered as a kind of manipulation. The retention of 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives could not be measured in pure water (buffer), therefore the parameter logk_W (capacity factor defined in pure water as a mobile phase, theoretically) would not have the physical meaning. Hence, this paper favors log$k$ value as the measure of lipophilicity of 17α-picolyl and thirteen 17(*E*)-picolinylidene androstane derivatives instead of logk_W value because of practical reasons.

2.3. Chemometric tools

QSRR analysis took into account several chemometric methods. In the first step, classification or pattern recognition methods was used: hierarchical cluster analysis (HCA) and principal component analysis (PCA) in order to reveal similarities or dissimilarities among the molecules. In the next step, the selection of molecular descriptors, which are the most suitable for regression analysis, was achieved by stepwise selection (SS) procedure. Afterwards, the regression analysis was carried out by using different regression methods: linear regression (LR), polynomial regression (PR), multiple linear regression (MLR), principal component regression (PCR), partial least squares regression (PLS) and

artificial neural networks regression (ANN). The obtained QSRR models were evaluated by internal and external validation, including cross-validation procedure. Ultimately, the QSRR models were ranked by Sum of Ranking Differences (SRD) approach introduced by Heberger and Kollar-Hunek [33].

HCA is a very useful method for detection of clusters of similar objects. It put the objects which are close together in the variable space in the same cluster. The distance among the objects is usually defined by Euclidian distances [34]. PCA has certain advantage over the HCA, since it can show which trait the object share in the variable space. PCA actually reduces the amount of data when there is correlation present among the variables. Principal components (PCs) are linear combinations of original variables [35]. QSRR procedure usually starts with the simple LR and/or PR analysis. If there are no high-quality or at least acceptable LR and PR models, it is necessary to apply multivariate regression approaches. MLR is suitable when there is no correlation present among the predictor variables. However, the multicollinearity is desirable in PCR and PLS modeling. Despite the limitations of MLR, it can be very suitable for explanation of certain chromatographic phenomena and it is usually used in QSRR analysis [36]. PCR is used when there can be detected a considerable degree of correlation between the independent variables [34,37]. The PCs in PCR are selected so they can describe as much of the variation in the independent variables as possible. However, in PLS, which has similar principle as PCR, the extra weight is given to the variables that are highly correlated with the response variable [34]. Both, PCR and PLS can be used only if a significant number of independent variables exists. The non-linear approach, which implies the ANN modeling, is very useful when the complex relationships between the variables exist. ANN method has become a very useful tool in modern QSRR analysis. Sometimes, ANNs are the best and the only solution for the precise prediction of the retention times (t_r) or chromatographic lipophilicity (log$k$) of analytes in certain chromatographic system [36,38,39].

SRD analysis, as one of relatively new models evaluation approaches, can give an insight into consistency of QSRR and QSAR models. It is entirely general and simple procedure and can be used together with classical comparison of statistical parameters in evaluation of QSRR and QSAR models [40-42]. The SRD method measures the distance of a model or an object from the defined reference value (“golden standard”). The smaller SRD values is, the better the model. SRD procedure can be applied in order to reveal the grouping of similar objects or models as well.

The software used for chemometric calculations in this study is the following: NCSS 2007 [43] (for SS and MLR), MATLAB R2013a with PLS_Toolbox [44] (for PCA, PCR and PLS modeling), Statistica v. 10 [45] (for ANN modeling) and Microsoft Excel 2013 [46].

Since detailed explanation of basics of the chemometric methods used in this study would require too much space and would probably draw reader’s attention from the main topic, it is omitted from the

manuscript. Therefore, the reading of the additional literature is strongly suggested [34-39], particularly for those who are not familiar with chemometric methods and QSRR methodology.

3. RESULTS AND DISCUSSION

3.1. Chromatographic lipophilicity of androstane derivatives

Chromatographic lipophilicity analysis by RP chromatography is based on the assumption that non-polar compounds express high affinity toward non-polar stationary phase and therefore have high retention in the system than polar compounds. Modifications in the structural core that increase the polarity of a molecule lead to more intensive interactions between a molecule and components of mobile phase. In this study, the analyzed compounds have low or moderate polarity, while more polar component in the mobile phase is water and methanol was used as a modifier. The results of chromatographic analysis are given in Supplementary data (Table S2). The experimentally obtained capacity factors (k) of studied androstanes are given in Figure 1, which indicates the higher retention in the system with the mobile phase which contains higher amount of water. Also, it can be seen that the compounds of 17(*E*)-picolinylidene group have higher retention than compounds of 17α-picolyl group. This leads to the conclusion that the picolinylidene function makes a molecule more non-polar than the picolyl group.

In the 17α-picolyl group the compound **10** has the highest retention in the applied chromatographic system. In 17(*E*)-picolinylidene group the highest retention expresses the compound **24**. The mutual characteristics of these compounds is the absence of substituents in A and B rings and double bonds in position 3 and 5. The compound **10** however exhibits weaker affinity toward C-18 stationary phase than the compound **24** due to the presence of hydroxyl group in 17β position. In Figure 1 it also can be observed that the compounds **2**, **7**, **12**, **18** and **21** are significantly retained in the system having low polarity. These compounds have acetoxy group in position 3. Despite the fact that compound **17** also has acetoxy group in the position 3, it has smaller retention than the aforementioned compounds. The reason for this is very polar nitro group in position 4 in compound **17**. The compounds **5**, **11**, **9**, **8**, **23**, **22** and **16** have relatively small retention since they have a number of polar functional groups in their structure.

The influence of particular substituent or functional group on chromatographic lipophilicity can be determined in the case of pairs of compounds whose structures differ only in the presence of that substituent or functional group. This influence can be described by Δk factor which is the difference between the capacity factor of a compound with specific substituent or functional group and a compound without it (Table 2). If the hydroxy group in position 3 of androstane core is substituted by acetoxy group, as in the case of pair of compounds **1** and **2**, **6** and **7**, **12** and **13**, the lipophilicity would significantly increase. The introduction of N-oxide function decreases the lipophilicity (pairs of compounds **14** and **16**, **12** and **18**), while more effective decrease in the lipophilicity is achieved by introduction of nitro group in position 4 (pair of compounds **21** and **17**). Methoxy group in position 4 leads to the slight increase of lipophilicity (pair of compounds **14** and **19**). The shift of the double

bonds position from position 5 into position 4 induce the decrease of lipophilicity (pairs of compounds **1** and **6**, **12** and **21**, **13** and **20**). The exception is the pair of compounds **2** and **7**. It must be emphasized that inductive and resonance effects can affect total lipophilicity of a compound. This can explain the effect of the position of double bonds in androstane core on total lipophilicity. Generally, the 17(*E*)-picolinylidene group have higher lipophilicity than the compounds which belong to the 17α-picolyl group. This is obvious if we compare the compounds that only differ in these type of substituents (pair of compounds **10** and **24**, **1** and **13**, **2** and **12**, **8** and **22**, **4** and **15**, **3** and **14**), and since they have a polar hydroxyl group at 17β-position.

The correlation between the k obtained by using mobile phase with volume fraction of methanol φ = 0.90 and the k obtained with mobile phase with φ = 0.70 is described by correlation coefficinet (R) of 0.9725. It indicates a good concurrence between retention factors (lipophilicity) determined by using two different mobile phases.

In the next step of analysis, the chromatographic data were processed by Wald-Wolfowitz runs (WWR) test in order to reveal if the 17α-picolyl group and 17(*E*)-picolinylidene group differ significantly in chromatographic lipophilicity. According to the results of WWR test (Supplementary data, Table S3), these two groups differ only in mobile phase with φ = 0.70. Therefore, $k_{0.70}$ factor can be used as discrimination factor of these two groups.

3.2. Computational lipophilicity of androstane derivatives

The 3D structures of studied androstanes were analyzed experimentally by X-ray crystallography and nuclear magnetic resonance [47]. The computational modeling of 3D structures was carried out so the modeled compounds are in agreement with experimentally predicted ones. The lipophilicity descriptors (log*P*, log*D* and log*S*) are obtained on the basis of 2D and 3D structures. Log*P* describes the distribution of a neutral compound between water and *n*-octanol layer, while log*D* takes into account ionic forms. Log*S* is solubility of a compound in water. According to the log*P* values, the analyzed androstane derivatives can be considered as lipophilic (log*P* > 1).

In order to visualize molecular characteristics which can affect total lipophilicity of a compound, Poisson-Boltzmann electrostatic potential (EP) surface, hydrophilic-lipophilic (HILI) surfaces and HOMO-LUMO orbitals were projected. In Figure 2 these characteristics were shown for compound **1**, and for other compounds are given in Supplementary data (Table S4). The Poisson-Boltzmann maps of electrostatic potential can give an insight into possible association of molecules and their polarity. The electronic effects can be quantified by the analysis of HOMO and LUMO orbitals. HOMO energy (E_{HOMO}) is an indicator of distribution of π-electrons in a molecule and π–π interactions, while LUMO energy (E_{LUMO}) is a measure of interactions based on electron transfer and H-bond formation effects.

The difference between E_{LUMO} and E_{HOMO} energy is so-called energy gap (E_{GAP}). E_{GAP} is an indicator of molecular stability (reactivity).

Observing the electrostatic potential maps of androstane derivatives, it can be seen that polar centers are located on N atom of the pyridine ring and around the introduced hydroxy, methoxy, acetoxy, nitro, oxo and epoxy groups, including the location of double bonds (π-electrons). According to the E_{GAP} measure, the compound **5** is chemically most stable (E_{GAP} = 12.9150 eV), while the compounds **8** and **22** have the lowest stability with the energy gap of 0.8770 eV and 0.4650 eV, respectively. The modeling of hydrophilic-lipophilic surfaces showed that majority of the analyzed androstanes has strongly non-polar or lipophilic character (lipophilic surface is significantly larger than hydrophilic in the majority of analyzed androstanes), which is in agreement with the calculated log*P* values. Hydrophilic surface marks the parts of a molecule which form dipole–dipole interactions with the components of mobile phase (methanol and water). HILI surfaces indicate the parts of the molecules of the analyzed androstanes which should be changed in order to increase or decrease their lipophilicity. Van der Waals interactions exhibit between the polar parts of a compound and C-18 stationary phase. It must be emphasized that residual non-modified silanol groups can affect the determination of chromatographic lipophilicity of compounds, despite the fact that hydrophobic surface of C-18 chains is much larger than the hydrophilic surface of residual silanol groups (Figure 3). Therefore, it is useful to examine the correlation between the experimentally obtained lipophilicity (*k* or log*k*) and computational lipophilicity (log*P*). This step is usually considered as the first step of QSRR analysis. It is aimed to confirm the concurrence between the experimental and computational lipophilicity.

3.3. Molecular features affecting the chromatographic lipophilicity of androstane derivatives – QSRR approach

The relationship between the experimental lipophilicity ($\log k_{0.90}$ and $\log k_{0.70}$) and computational lipophilicity (Average log*P*) was examined in the first step of QSRR analysis. Average log*P* was calculated on the basis of all calculated log*P* parameters (ALOGPs, AClogP, ALOGP, KOWWIN, etc.). The obtained models can be seen in Figure 4. These quite good correlations confirm the assumption that the chromatographic factors ($\log k_{0.90}$ and $\log k_{0.70}$) can be considered as lipophilicity parameters of the analyzed 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives.

The QSRR analysis started from the simples models, such as LR and PR, but followed by more complex approaches as MLR, PCR, PLS and ANN regressions. All these models were validated by internal and external validation procedures. Prior to QSRR analysis, the set of compounds was divided into training set and external test set (compounds **2**, **6**, **12**, **19** and **23**), except in ANN modeling where the additional test set is required. The statistical parameters used as indicators of the quality of QSRR

models in this study are the following: Pearson's correlation and determination coefficients (R, R^2), adjusted determination coefficient (R^2_{adj}), leave-one-out (LOO) cross-validation determination coefficient (R^2_{cv}), Fisher's test (F), p-value, root mean square error ($RMSE$), total sum of squares (TSS), predicted residuals sum of squares ($PRESS$), $PRESS/TSS$ ratio, standard deviation of cross-validation (SD_{PRESS}), $RMSE$ and correlation coefficient of the test set ($RMSE_{test}$, R_{test} respectively). PCR and PLS models were characterized by cumulative fraction of sum of squares of all the Ys explained by the component (R^2Y_{cumul}) and cumulative fraction of the total variation of the Ys that can be predicted by the component (Q^2Y_{cumul}). In ANN modeling, the additional test set must be used to determine generalization error, while validation set is used to find the best ANN configuration and training parameters by comparing validation set error and training set error during training. The optimal values of the aforementioned parameters are given in Supplementary data (Table S5).

3.3.1. Linear QSRR modeling

The molecular features which affect the chromatographic lipophilicity of androstane derivatives were revealed on the basis of the highest R value in LR and PR analysis, while the SS procedure was used in multivariate calibration. It must be highlighted that prior to QSRR modeling, PCA was carried out in order to see if there is a significant difference between the compounds of 17α-picolyl group and the compounds of 17(*E*)-picolinylidene group. If the difference exists, the QSRR models should be formed for separate groups. In this case, PCA based on molecular descriptors pointed out that there is no strict separation between these two groups in the variable space (results given in Supplementary data, Figures S1-S3), therefore the QSRR modeling is carried out on the both groups together. The established LR, PR and MLR models are the following:

LR1: $\log k_{0.90}$ = 0.35797 (±0.04005) **ALOGPs –** 1.33462 (±0.17507) (2)
R^2 = 0.8246

LR2: $\log k_{0.90}$ = 0.44518 (±0.04913) **Average log*P*** – 1.78180 (±0.22085) (3)
R^2 = 0.8285

PR: $\log k_{0.90}$ = 0.000018 (±0.000015) **MP^2** – 0.03202 (±0.02075) **MP** + 13.62574 (±7.36103) (4)
R^2 = 0.7496

MLR1: $\log k_{0.90}$ = 0.39009 (±0.03960) **Average log*P*** – 0.00644 (±0.00128) **CT** + 0.00205 (±0.00128) **DE** + 4.28539 (±1.33568) (5)
R^2 = 0.9538

MLR2: $\log k_{0.90}$ = 0.31379 (±0.03399) **XLOGP3** – 0.00685 (±0.00135) **CT** – 1.02150 (±0.36183) **Jhetv** + 6.68200 (±1.46962)
(6)
R^2 = 0.9485

MLR3: log$k_{0.90}$ = 40.04356 (±3.93576) **SCAA3** – 0.11597 (±0.02074) **CP** + 0.00212 (±0.00077) **DE** + 2.89430 (±0.31819) (7)

R^2 = 0.9210

MLR4: log$k_{0.90}$ = 0.36138 (±0.05960) **Average log*P*** + 0.00277 (±0.00076) **DE** – 0.00276 (±0.00083) **MP** + 0.19114 (±0.81162) (8)

R^2 = 0.9282

MLR5: log$k_{0.90}$ = 0.45822 (±0.04500) **Average log*P*** – 0.02847 (±0.01322) $\boldsymbol{E}_{\mathbf{sr}}$ – 1.96110 (±0.21706) (9)

R^2 = 0.8670

MLR6: log$k_{0.90}$ = 0.36361 (±0.03008) **ALOGPs** – 0.01969 (±0.00554) $\boldsymbol{E}_{\mathbf{max}}$ + 0.02356 (±0.01097) $\boldsymbol{E}_{\mathbf{GAP}}$ – 1.21035 (±0.17564)

(10)

R^2 = 0.9155

MLR7: log$k_{0.90}$ = 0.39182 (±0.03467) **ALOGPs** – 0.01714 (±0.00556) $\boldsymbol{E}_{\mathbf{max}}$ – 0.06415 (±0.02959) $\boldsymbol{E}_{\mathbf{HOMO}}$ – 1.78929 (±0.38008) (11)

R^2 = 0.9159

MLR8: log$k_{0.90}$ = 0.32753 (±0.03418) **ALOGPs** – 17.49395 (±5.68512) **FPSA3** – 0.60873 (±0.27587) (12)

R^2 = 0.8898

MLR9: log$k_{0.90}$ = 0.35765 (±0.02826) **ALOGPs** – 0.01475 (±0.00363) $\boldsymbol{\Delta E}$ + 0.03507 (±0.01086) $\boldsymbol{E}_{\mathbf{GAP}}$ – 0.98360 (±0.19327) (13)

R^2 = 0.9259

MLR10: log$k_{0.90}$ = 0.20208 (±0.04283) **ALOGPs** – 0.01641 (±0.00345) **PSA** + 0.00436 (±0.00120) **vdWSA** – 2.50896 (±0.56880) (14)

R^2 = 0.9301

MLR11: log$k_{0.90}$ = 0.33171 (±0.04212) **ALOGPs** + 0.00424 (±0.00127) **TE** – 0.00436 (±0.00141) **BP** + 1.79538 (±1.22230) (15)

R^2 = 0.9360

Since the best QSRR models were obtained for log$k_{0.90}$, it can be considered further as the best chromatographic lipophilicity descriptor of 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives. The quality of the obtained models is evaluated by statistical parameters given in Table 3. It can be said that MLR models generally have better prediction performance that LR and PR, but LR and PR models, which are statistically significant, emphasized the influence of melting point (MP) and

lipophilicity (ALOGPs, Average log*P*) on the log$k_{0.90}$. However, MLR models reveal which particular electrostatic, topological and physicochemical properties can affect retention behavior of androstanes in the applied chromatographic system. Those properties are critical temperature (CT), Balaban-type index from van der Waals weighted distance matrix (Jhetv), surface weighted charged area on acceptor atoms 3rd type (SCAA3), critical pressure (CP), Dreiding energy (DE), mean value of electrostatic potential (E_{sr}), maximum value of electrostatic potential (E_{max}), E_{GAP}, E_{HOMO}, fractional charged partial positive surface area 3rd type (FPSA3), the difference between maximum and minimum electrostatic potential (ΔE), polar surface area (PSA), van der Waals surface area (vdWSA) and boiling point (BP). The highest regression coefficient of lipophilicity parameters in MLR equations confirms the assumption that the retention behavior is mostly affected by distribution of a compound between non-polar stationary phase and more polar mobile phase. This is another confirmation of log$k_{0.90}$ as lipophilicity measure of studied androstane derivatives.

Besides the numerical data, the predictive performance of LR, PR and MLR models has been evaluated by graphical comparison of experimental and predicted log$k_{0.90}$ values, as well as by the residuals analysis (Supplementary data, Figures S4-S6). These results show that MLR models make better fitting of the data than LR and PR models, which was confirmed by statistical measures given in Table 3. Variance Inflation Factor (*VIF*) suggest that there is no multicollinearity present in MLR models (*VIF* < 10). The randomness of the residuals indicate the unpredictable error and it was confirmed for all LR, PR and MLR models.

Unlike MLR analysis, PCR and PLS modeling were carried out on the basis of intercorrelated descriptors. Selection of the independent variables was done by the correlation matrix. The best PCR and PLS models was obtained by using thirteen lipophilicity descriptors (ALOGPs, AClogP, ALOGP, MLOGP, KOWWIN, XLOGP2, XLOGP3, Average log*P*, miLog*P*, logP_{vg}, logP_{klop}, logP_{phys}, logP_{wgt}) and five physicochemical descriptors (vdWV, PSA, vdWSA, SASA1.4, MR). The selection of the number of PCs or latent variables (LVs) was done on the basis of the lowest *RMSE* of LOO cross-validation ($RMSE_{CV}$) (Supplementary data, Figure S7). In the case of PCR the lowest $RMSE_{CV}$ corresponds to the number of 8 PCs (99.66% variability), while the PLS model includes 6 LVs (98.20% of variability). Regression coefficients of PCR and PLS models and comparisons of experimental and predicted log$k_{0.90}$ values are given in Supplementary data, Table S6 and Figure S8, while statistical characteristics are presented in Table 4. The obtained results imply the significance of lipophilicity parameters and PSA. In PLS model this was confirmed by Variance Importance in Projection (VIP) (Supplementary data, Figure S9). Generally, it can be concluded that PCR and PLS models make better fitting of the data than LR, PR and MLR models, having low random residuals.

3.3.2. Non-linear QSRR modeling

The next step of the QSRR modeling included non-linear approach based on artificial neural networks. This approach can find complex relationships between the variables [37]. The input variables for ANN

modeling were the same as the independent variables in MLR models. The input variables were normalized by *min-max* normalization method [48]. The number of hidden neurons varied in the range of 2-200. The following MLP activation functions were combined for hidden and output neurons: logistic (*Lgt*), identity (*Idt*), exponential (*Exp*), tangent (*Tanh*) and sinusoidal (*Sine*). During ST-ANN regression modelling with Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm 28 000 networks have been trained. The test set contains compounds **7**, **10** and **16**, while the compounds **12**, **14** and **24** belong to the validation set.

The best five ANN models were selected on the basis of statistical parameters given in Table 5. The statistics undoubtedly indicate the ANN models as the best solution for prediction of $\log k_{0.90}$ parameter of studied androstanes in the used chromatographic system. The concurrence between the experimental and predicted data, as well as the random distribution of residuals, for representative ANN model are presented in Figure 5 and for the other ANN models are shown in Figure S10 in Supplementary data. The input variables of ANN models are characterized by global sensitivity analysis (GSA) coefficients, which present the ratio between the network error when the observed variable is omitted and the network error when the observed variable is present in the model [49]. The variable should be omitted from the ANN model if the GSA coefficient is equal to or less than 1 [49]. In this case, all the input variables are significant (Supplementary data, Figure S11). The highest GSA coefficients distinguish the lipophilicity descriptors (XLOGP3 and Average log*P*) as the most significant input variables in ANN models. This is another confirmation of the hypothesis that $\log k_{0.90}$ certainly can be considered as lipophilicity descriptor of 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives.

3.4. Ranking of the QSRR models: a novel point of view toward models quality

After the definition of relationships between selected molecular features and chromatographic lipophilicity, the selection of the most reliable QSRR models was carried out. In order to rank the models by SRD method, the experimental and average $\log k_{0.90}$ values were applied as the reference ranking. The average values as the reference ranking contain less bias than ranking by any of the individual vectors [33]. The aim of ranking analysis based on the experimental data was to describe fits, not experimental errors. The ranking was completely validated by CRRN (comparison of ranks by random numbers) method and seven-fold cross-validation procedures. The results of SRD analysis are presented in Table 6.

The ranking based on the experimental values indicate that the ANN models are the closest to the experimental data and make the best data fitting. This is completely in agreement with the statistical parameters, which distinguished the ANN models as the best solution for precise prediction of $\log k_{0.90}$ values. The worst one is the PR model and should be avoided in prediction of the chromatographic

lipophilicity of studied androstane derivatives. The second approach, which included the average values as the reference, showed that MLR10 model is the closest to the reference ranking, while the PR model is the farthest. Low prediction ability of the PR model is definitely confirmed, therefore it cannot be recommended for further application. Although the majority of the MLR models and PCR and PLS models are closer to the reference ranking than the EXP values, it is not a definitive indication of overfitting, since the phenomenon can be simply the consequence of random noise [42]. Very low probabilities, p(%) given in Table 6, describe the established QSRRs as non-random models.

Generally, taking into account calculated statistical parameters and ranking analysis, the ANN models are the best tool for prediction of the chromatographic lipophilicity ($\log k_{0.90}$) of 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives.

4. CONCLUSION

Since the androstane derivatives, studied in this paper, have a great anticancer potential, the presented results of experimental lipophilicity determination and its prediction are the first step in their further biological analysis *in vivo*. The obtained results describe 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives as lipophilic compounds and present the best mathematical (QSRR) models which can be used for precise prediction of lipophilicity of these compounds. The QSRR analysis pointed out the molecular features which influence the total molecular lipophilicity. The obtained QSRR models can be extremely useful in assessing chromatographic lipophilicity of new 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives as potential anticancer compounds.

Acknowledgement: This study is financially supported by the research projects of the Ministry of Education, Science and Technological Development of the Republic of Serbia (No. 172012 and No. 172021) and the research project of the Provincial Secretariat for Science and Technological Development of Vojvodina (No. 114-451-347/2015-02).

References

[1] M.K. Barton, Nicotinamide found to reduce the rate of nonmelanoma skin cancers in high-risk patients, Ca-Cancer J. Clin. 66 (2016) 91-92.

[2] G.K. Alderton, Therapeutic resistance: Multiple mechanisms to keep going, Nat. Rev. Cancer 11 (2015) 635-635.

[3] M. Ammad-ud-din, E. Georgii, M. Gönen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, S. Kaski, Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization, J Chem Inf Model. 54 (2014), 2347-2359.

[4] 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.

[5] S. Pesonen, L. Kangasniemi, A. Hemminki, Oncolytic Adenoviruses for the Treatment of Human Cancer: Focus on Translational and Clinical Data, Mol. Pharmaceutics, 8 (2011) 12-28.

[6] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini et al. QSAR modeling: Where have you been? Where are you going to? J. Med. Chem. 57 (2014) 4977-5010.

[7] R. Zanni, M. Gálvez-Llompart, J. Gálvez, R. García-Domenech, QSAR multi-target in drug discovery: a review, Curr Comput Aided Drug Des. 10 (2014) 129-136.

[8] G. Caron, G. Ermondi, Lipophilicity: Chemical nature and biological relevance, in: R. Mannhold (Ed.), Molecular drug properties, Wiley-VCH Verlag GmbH & Co. Weinheim, 2008, pp 315-328.

[9] A. Nasal, D. Siluk, R. Kaliszan, Chromatographic retention parameters in medicinal chemistry and molecular pharmacology, Curr. Med. Chem. 10 (2003) 381-426.

[10] R. Kaliszan, P. Wiczling, M.J. Markuszewski, M.A. Al-Haj, Thermodynamic vs. extrathermodynamic modeling of chromatographic retention, J Chromatogr A 1218 (2011) 5120-5130.

[11] C. Sârbu, R.D. Naşcu-Briciu, D. Casoni, A. Kot-Wasik, A. Wasik, J. Namieśnik, Chromatographic lipophilicity determination using large volume injections of the solvents non-miscible with the mobile phase, J Chromatogr A 1266 (2012) 53-60.

[12] S.Z. Kovačević, S.O. Podunavac-Kuzmanović, L.R. Jevrić, E.S. Lončar, Assessment of Chromatographic Lipophilicity of Some Anhydro-D-Aldose Derivatives on Different Stationary Phases by QSRR Approach, J. Liq. Chromatogr. Relat. Technol. 38 (2015) 492-500.

[13] D. Benhaim, E. Grushka, Effect of n-octanol in the mobile phase on lipophilicity determination by reversed-phase high-performance liquid chromatography on a modified silica column, J Chromatogr A 1209 (2008) 111-119.

[14] X. Liu, H. Tanaka, A. Yamauchi, B. Testa, H. Chuman, Determination of lipophilicity by reversed-phase high-performance liquid chromatography. Influence of 1-octanol in the mobile phase, J Chromatogr A, 1091 (2005) 51-59.

[15] L. Ayouni, G. Cazorla, D. Chaillou, B. Herbreteau, S. Rudaz, P. Lantéri, P.A. Carrupt, Fast Determination of Lipophilicity by HPLC, Chromatographia 62 (2005) 251-255.

[16] S. Kowalska, K. Krupczyńska, B. Buszewski, Some remarks on characterization and application of stationary phases for RP-HPLC determination of biologically important compounds, Biomed. Chromatogr. 20 (2006) 4-22.

[17] A. Guillot, Y. Henchoz, C. Moccand, D. Guillarme, J.L. Veuthey, P.A. Carrupt, S. Martel, Lipophilicity determination of highly lipophilic compounds by liquid chromatography, Chem Biodivers. 6 (2009) 1828-1836.

[18] V. Dohnal, K. Musílek, K. Kuča, Retention Behavior of Pyridinium Oximes on PFP Stationary Phase in High-Performance Liquid Chromatography, J Chromatogr Sci 52 (2014) 246-251.

[19] K. Héberger, Quantitative structure-(chromatographic) retention relationships, J Chromatogr A 1158 (2007) 273-305.

[20] R. Kaliszan, QSRR: Quantitative structure-(chromatographic) retention relationships, Chem. Rev. 107 (2007) 3212-3246.

[21] E. Djurendić, J. Daljev, M. Sakač, J. Čanadi, S. Jovanović Šanta, S. Andrić, O. Klisurić, V. Kojić, G. Bogdanović, M. Djurendić-Brenesel, S. Novaković, K.P. Gaši, Synthesis of some epoxy and/or N-oxy 17-picolyl and 17-picolinylidene-androst-5-ene derivatives and evaluation of their biological activity, Steroids 73 (2008) 129-138.

[22] E.A. Djurendić, J.J Ajduković, M.N. Sakač, J.J. Čanadi, V.V. Kojić, G.M. Bogdanović, K.M. Penov Gaši, 17-Picolinylidene-substituted steroid derivatives and their antiaromatase and cytotoxic activity, ARKIVOC 13 (2009) 311-323.

[23] E.A. Djurendić, J.J. Ajduković, M.N. Sakač, J.J. Csanádi, V.V. Kojić, G.M. Bogdanović, K.M. Penov Gaši, Synthesis and cytotoxic activity of some 17-picolyl and 17-picolinylidene androstane derivatives, Eur. J. Med. Chem. 54 (2012) 784-792.

[24] CambridgeSoft Corporation, PerkinElmer Inc., ChemBio3D software version 14.0, 2014, http://www.cambridgesoft.com/

[25] Biologics Suite 2015-1, BioLuminate version 1.7, Schrödinger, LLC, New York, NY, 2015, http://www.schrodinger.com/

[26] PreADMET software, http://preadmet.bmdrc.org/

[27] Simulations Plus, Inc. http://www.simulations-plus.com/

[28] Virtual Computational Chemistry Laboratory ALOGPS 2.1 online program, http://146.107.217.178/lab/alogps/start.html

[29] Avogadro software 1.1.1, http://avogadro.cc/wiki/

[30] ADRIANA.Code software - Calculation of Molecular Descriptors, https://www.molecular-networks.com/

[31] MarvinSketch 6.1, 2013, ChemAxon, http://www.chemaxon.com

[32] Virtual Computational Chemistry Laboratory, Parameter Client online program, http://www.vcclab.org/

[33] K. Héberger, K. Kollár-Hunek, Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers, J. Chemom. 25 (2011) 151-158.

[34] J.N. Miller, J.C. Miller, Statistics and chemometrics for analytical chemistry, sixth ed., Pearson Education Limited, Harlow, UK, 2010, pp 221-247.

[35] K.H. Esbensen, Multivariate data analysis – in practice, fifth ed., CAMO Software AS, Oslo, Norway, 2009.

[36] R. Put, Y. Vander Heyden, Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure-retention relationships, Anal. Chim. Acta 602 (2007) 164-172.

[37] S. Kovačević, S.O. Podunavac-Kuzmanović, L.R. Jevrić, Multivariate regression modelling of antifungal activity of some benzoxazole and oxazolo[4,5-b]pyridine derivatives, Acta Chim. Slov. 60 (2013) 756-762.

[38] Z. Garkani-Nejad, Use of self-training artificial neural networks in a QSRR study of a diverse set of organic compounds, Chromatographia 70 (2009) 869-874.

[39] A.G. Fragkaki, E. Farmaki, N. Thomaidis, A. Tsantili-Kakoulidou, Y.S. Angelis, M. Koupparis, C. Georgakopoulos, Comparison of multiple linear regression, partial least squares and artificial neural networks for prediction of gas chromatographic relative retention times of trimethylsilylated anabolic androgenic steroids, J Chromatogr A 1256 (2012) 232-239.

[40] K. Kollár-Hunek, K. Héberger, Method and model comparison by sum of ranking differences in cases of repeated observations (ties), Chemom. Intell. Lab. Syst. 127 (2013) 139-146.

[41] M. Vračko, N. Minovski, K. Héberger, Ranking of QSAR Models to Predict Minimal Inhibitory Concentrations Toward Mycobacterium tuberculosis for a Set of Fluoroquinolones, Acta Chim. Slov. 57 (2010) 586-590.

[42] S.Z. Kovačević, S.O. Podunavac-Kuzmanović, L.R. Jevrić, E.A. Djurendić, J.J. Ajduković, S.B. Gadžurić, M.B. Vraneš, How to rank and discriminate artificial neural networks? Case study: Prediction of anticancer activity of 17-picolyl and 17-picolinylidene androstane derivatives, J Iran Chem Soc 13 (2016) 499-507.

[43] Hintze J, NCSS 2007, NCSS, LLC. Kaysville, Utah, USA, 2007, www.ncss.com

[44] MATLAB R2013a, The MathWorks Inc, Natick, Massachusetts, United States, 2013, www.mathworks.com

[45] StatSoft Inc, STATISTICA (data analysis software system), version 10, 2011, www.statsoft.com

[46] Microsoft. Microsoft Excel. Computer Software. Redmond, Washington, 2013.

[47] J. Ajduković, Synthesis and biological activity of 17-substituted androstane derivatives, Doctoral Dissertation, Faculty of Sciences, University of Novi Sad, Serbia, 2013.

[48] T. Jayalakshmi, A. Santhakumaran, Statistical normalization and back propagation for classification, IJCTE 3 (2011) 89-93.

[49] M.H. Shojaeefard, M. Akbari, M. Tahani, F. Farhani, Sensitivity analysis of the artificial neural network outputs in friction stir lap joining of aluminium to brass, Adv. Mater. Sci. Eng. 2013 (2013) 1-7.

Table 1. Structures of the studied 17α-picolyl and 17(*E*)-picolinylidene androstane derivatives

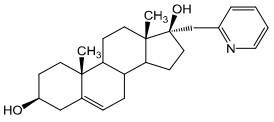
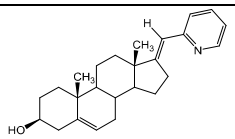
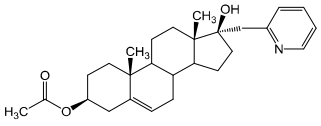
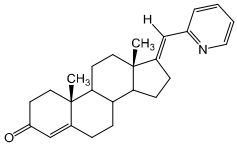
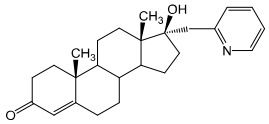
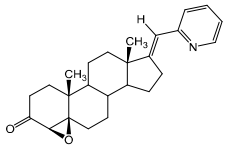
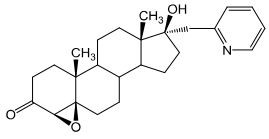
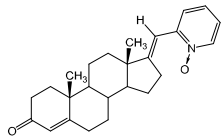
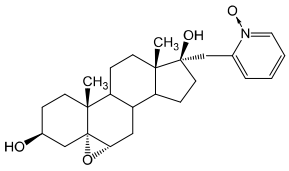
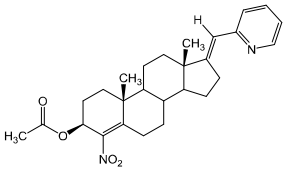
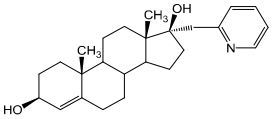
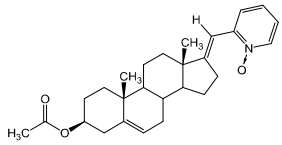
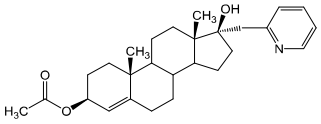
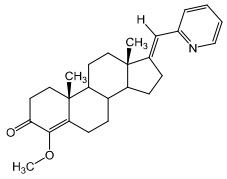
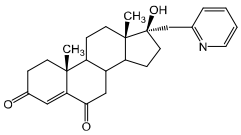
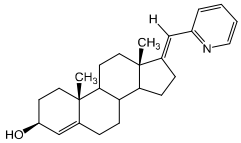
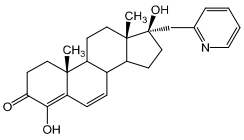
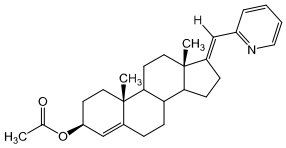
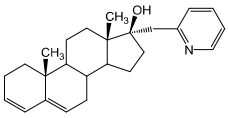
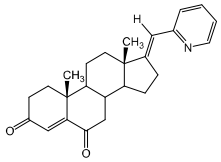
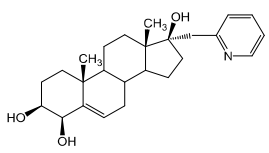
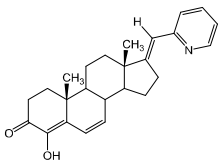
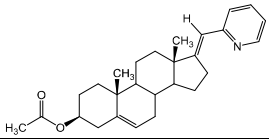
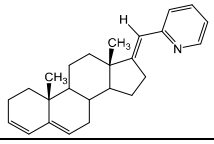
Compound No.	2D structure	Compound No.	2D structure
1		**13**	
2		**14**	
3		**15**	
4		**16**	
5		**17**	
6		**18**	
7		**19**	
8		**20**	
9		**21**	
10		**22**	
11		**23**	
12		**24**	

Table 2. Δ*k* values of the substitution of introduction of different functional groups (substituents) or the change in the position of double bonds (*φ* – volume fraction of the organic modifier)

The compared compounds		Substitution or introduction of substituents / the difference between the position of double bonds	Δ*k* (*φ* = 0,90)	Δ*k* (*φ* = 0,70)
The comparison of compounds which belong to the same group	**1** and **2**	3-OH → 3-AcO	1.880	3.074
	6 and **7**	3-OH → 3-AcO	2.003	3.301
	1 and **6**	5-en → 4-en	-0.372	-0.146
	2 and **7**	5-en → 4-en	-0.249	0.081
	13 and **12**	3-OH → 3-AcO	3.681	9.127
	14 and **16**	N → N→O	-1.001	-0.728
	12 and **18**	N → N→O	-4.146	-7.296
	21 and **17**	4-NO_2	-4.635	-8.425
	12 and **21**	5-en → 4-en	-0.612	-2.272
	13 and **20**	5-en → 4-en	-0.225	-0.213
	14 and **19**	4-OCH_3	0.350	0.127
The comparison of compounds which belong to different groups	**10** and **24**	17α-picolyl-17β-hydroxy ↓ 17(*E*)-picolinylidene	10.012	10.524
	1 and **13**		1.701	2.129
	2 and **12**		3.502	8.182
	8 and **22**		0.435	0.756
	4 and **15**		0.607	0.572
	3 and **14**		0.778	0.986

Table 3. Statistical characteristics of LR, PR and MLR QSRR models

Parameter	LR1	LR2	PR	MLR1	MLR2	MLR3	MLR4	MLR5	MLR6	MLR7	MLR8	MLR9	MLR10	MLR11
R	0.9081	0.9102	0.8658	0.9766	0.9739	0.9597	0.9634	0.9311	0.9568	0.9570	0.9433	0.9622	0.9644	0.9675
R^2	0.8246	0.8285	0.7496	0.9538	0.9485	0.9210	0.9282	0.8670	0.9155	0.9159	0.8898	0.9259	0.9301	0.9360
R^2_a	0.1754	0.1715	0.2504	0.0462	0.0515	0.0790	0.0718	0.1330	0.0845	0.0841	0.1102	0.0741	0.0699	0.0640
R^2_{adj}	0.8142	0.8184	0.7183	0.9445	0.9382	0.9052	0.9138	0.8504	0.8986	0.8990	0.8760	0.9110	0.9161	0.9232
F-test	79.90	82.10	23.94	103.20	92.10	58.26	64.65	52.15	54.16	54.43	64.59	62.43	66.53	73.16
RMSE	0.1909	0.1887	0.2351	0.1043	0.1101	0.1364	0.1300	0.1713	0.1410	0.1407	0.1559	0.1321	0.1283	0.1227
p-value	0.000000	0.000000	0.000015	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
VIF	-	-	-	2.13 (Av. log*P*) 1.67 (CT) 1.38 (DE)	1.71 (XLOGP3) 1.67 (CT) 1.05 (Jhetv)	1.28 (SCAA3) 1.16 (CP) 1.32 (DE)	3.10 (Av. log*P*) 1.41 (DE) 3.05 (MP)	1.02 (Av. log*P*) 1.02 (*E*sr)	1.03 (ALOGPs) 1.01 (*E*max) 1.04 (*E*gap)	1.38 (ALOGPs) 1.02 (*E*max) 1.39 (*E*homo)	1.09 (ALOGPs) 1.09 (FPSA3)	1.04 (ALOGPs) 1.14 (Δ*E*) 1.16 (*E*gap)	2.53 (ALOGPs) 4.65 (PSA) 2.55 (vdWSA)	2.68 (ALOGPs)
R^2_{CV} (LOO)	0.7781	0.7746	0.6476	0.9191	0.9067	0.8700	0.8439	0.8055	0.8707	0.8752	0.8501	0.8923	0.8883	1.44 (TE)
$\lvert R^2 - R^2_{CV}\rvert$	0.0465	0.0539	0.1020	0.0347	0.0418	0.0510	0.0843	0.0615	0.0448	0.0407	0.0397	0.0336	0.0418	2.06 (BP)
TSS	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	3.5303	0.8957
PRESS	0.7832	0.7958	1.2440	0.2855	0.3292	0.4590	0.5511	0.6868	0.4565	0.4407	0.5291	0.3803	0.3945	0.0403
PRESS/TSS	0.2219	0.2254	0.3524	0.0809	0.0932	0.1300	0.1561	0.1945	0.1293	0.1248	0.1499	0.1077	0.1117	3.5303
SD_{PRESS}	0.2030	0.2047	0.2559	0.1226	0.1316	0.1554	0.1703	0.1901	0.1550	0.1523	0.1669	0.1415	0.1441	0.3683
R_{test}	0.9489	0.9199	0.8147	0.9160	0.9494	0.9096	0.9239	0.9714	0.9829	0.9398	0.9172	0.9548	0.9720	0.1043
$RMSE_{test}$	0.0963	0.1067	0.1601	0.1258	0.1011	0.1684	0.1261	0.0688	0.0612	0.1213	0.1444	0.0971	0.0989	0.1392

Table 4. Statistical characteristics of PCR and PLS models

Parameters	**PCR**	**PLS**
R^2Y_{cumul}	0.9754	0.9789
Q^2Y_{cumul}	0.9189	0.9047
R	0.9879	0.9896
R^2	0.9760	0.9793
R_{test}	0.9777	0.9759
R^2_{test}	0.9558	0.9524
R^2_{CV}	0.9210	0.9075
$RMSE$	0.0667	0.0619
$RMSE_{CV}$	0.1213	0.1317
$RMSE_{test}$	0.1418	0.1357
$PRESS/TSS$	0.0790	0.0925
F-vrednost	512.00	577.24
p-vrednost	0.000000	0.000000

Table 5. Statistical characteristics of ANN models

Parameters	ANN1	ANN2	ANN3	ANN4	ANN5
Architecture (input - hidden layer - output)	3-48-1	3-123-1	3-115-1	3-3-1	3-28-1
Input variables	Average log*P* CT DE	XLOGP3 CT Jhetv	XLOGP3 CT Jhetv	XLOGP3 CT Jhetv	XLOGP3 CT Jhetv
R_{calib}	0.9766	0.9906	0.9877	0.9815	0.9810
R_{test}	0.9875	0.9881	0.9956	0.9999	0.9998
R_{valid}	0.9990	1.0000	0.9998	1.0000	0.9999
$RMSE_{calib}$	0.003	0.001	0.002	0.002	0.002
$RMSE_{test}$	0.006	0.003	0.001	0.000	0.001
$RMSE_{valid}$	0.001	0.003	0.002	0.001	0.001
F-test	630.43	1288.03	1353.00	1061.87	1015.55
Algorithm	BFGS (36)*	BFGS (56)	BFGS (35)	BFGS (29)	BFGS (39)
p-value	0.000000	0.000000	0.000000	0.000000	0.000000
Hidden activation function	*Exp*	*Tanh*	*Tanh*	*Lgt*	*Lgt*
Output activation function	*Sine*	*Idt*	*Idt*	*Sine*	*Sine*

**the number in brackets is the number of the training cycles after which the best neural architecture was achieved*

Table 6. The ranking of the analyzed QSRR models obtained by SRD analysis (XX1 – first icosaile, 5%, Q1 – first quartile, Q3 – last quartile, XX19 – last icosaile, 95%)

Rank	QSRR model	Absolute SRD value	Probability (p%) $x_1 < SRD \leq x_2$		QSRR model	Absolute SRD value	Probability (p%) $x_1 < SRD \leq x_2$	
	Reference: Experimental log$k_{0.90}$ values				**Reference: Average log$k_{0.90}$ values**			
1	ANN5	24	$1.91 \cdot 10^{-9}$	$3.26 \cdot 10^{-9}$	MLR10	18	$3.82 \cdot 10^{-10}$	$6.55 \cdot 10^{-10}$
2	ANN3	26	$3.26 \cdot 10^{-9}$	$5.20 \cdot 10^{-9}$	MLR9	20	$6.55 \cdot 10^{-10}$	$1.07 \cdot 10^{-9}$
3	ANN4	28	$5.20 \cdot 10^{-9}$	$9.05 \cdot 10^{-9}$	MLR2	22	$1.07 \cdot 10^{-9}$	$1.91 \cdot 10^{-9}$
4	ANN2	32	$1.53 \cdot 10^{-8}$	$2.45 \cdot 10^{-8}$	MLR7	22	$1.07 \cdot 10^{-9}$	$1.91 \cdot 10^{-9}$
5	ANN1	34	$2.45 \cdot 10^{-8}$	$4.04 \cdot 10^{-8}$	PLS	24	$1.91 \cdot 10^{-9}$	$3.26 \cdot 10^{-9}$
6	MLR10	38	$6.78 \cdot 10^{-8}$	$1.08 \cdot 10^{-7}$	MLR8	28	$5.20 \cdot 10^{-9}$	$9.05 \cdot 10^{-9}$
7	PLS	38	$6.78 \cdot 10^{-8}$	$1.08 \cdot 10^{-7}$	MLR11	28	$5.20 \cdot 10^{-9}$	$9.05 \cdot 10^{-9}$
8	PCR	40	$1.08 \cdot 10^{-7}$	$1.71 \cdot 10^{-7}$	MLR1	30	$9.05 \cdot 10^{-9}$	$1.53 \cdot 10^{-8}$
9	MLR1	42	$1.71 \cdot 10^{-7}$	$2.84 \cdot 10^{-7}$	MLR6	30	$9.05 \cdot 10^{-9}$	$1.53 \cdot 10^{-8}$
10	MLR2	42	$1.71 \cdot 10^{-7}$	$2.84 \cdot 10^{-7}$	PCR	30	$9.05 \cdot 10^{-9}$	$1.53 \cdot 10^{-8}$
11	MLR7	42	$1.71 \cdot 10^{-7}$	$2.84 \cdot 10^{-7}$	**EXP***	32	$1.53 \cdot 10^{-8}$	$2.45 \cdot 10^{-8}$
12	MLR8	44	$2.84 \cdot 10^{-7}$	$4.51 \cdot 10^{-7}$	MLR3	32	1.53 · 10-8	$2.45 \cdot 10^{-8}$
13	MLR9	44	$2.84 \cdot 10^{-7}$	$4.51 \cdot 10^{-7}$	ANN5	32	$1.53 \cdot 10^{-8}$	$2.45 \cdot 10^{-8}$
14	MLR11	44	$2.84 \cdot 10^{-7}$	$4.51 \cdot 10^{-7}$	LR1	34	$2.45 \cdot 10^{-8}$	$4.04 \cdot 10^{-8}$
15	MLR6	46	$4.51 \cdot 10^{-7}$	$6.86 \cdot 10^{-7}$	ANN4	34	$2.45 \cdot 10^{-8}$	$4.04 \cdot 10^{-8}$
16	MLR5	50	$1.12 \cdot 10^{-6}$	$1.77 \cdot 10^{-6}$	ANN1	36	$4.04 \cdot 10^{-8}$	$6.78 \cdot 10^{-8}$
17	MLR4	54	$2.66 \cdot 10^{-6}$	$4.21 \cdot 10^{-6}$	ANN3	36	$4.04 \cdot 10^{-8}$	$6.78 \cdot 10^{-8}$
18	LR2	56	$4.21 \cdot 10^{-6}$	$6.57 \cdot 10^{-6}$	MLR5	38	$6.78 \cdot 10^{-8}$	$1.08 \cdot 10^{-7}$
19	MLR3	56	$4.21 \cdot 10^{-6}$	$6.57 \cdot 10^{-6}$	ANN2	40	$1.08 \cdot 10^{-7}$	$1.71 \cdot 10^{-7}$
20	LR1	60	$9.84 \cdot 10^{-6}$	$1.49 \cdot 10^{-5}$	LR2	42	$1.71 \cdot 10^{-7}$	$2.84 \cdot 10^{-7}$
21	PR	84	$1.03 \cdot 10^{-3}$	$1.44 \cdot 10^{-3}$	MLR4	48	$6.86 \cdot 10^{-7}$	$1.12 \cdot 10^{-6}$
22	XX1	148	4.72	5.50	PR	72	$1.13 \cdot 10^{-4}$	$1.60 \cdot 10^{-4}$
	Q1	174	23.11	25.65	XX1	148	4.72	5.50
	Median	190	48.91	52.04	Q1	174	23.11	25.65
	Q3	208	72.68	75.12	Median	190	48.91	52.04
	XX19	232	94.72	95.57	Q3	208	72.68	75.12

**Experimental values (EXP) were used as a detector of overfitted models*

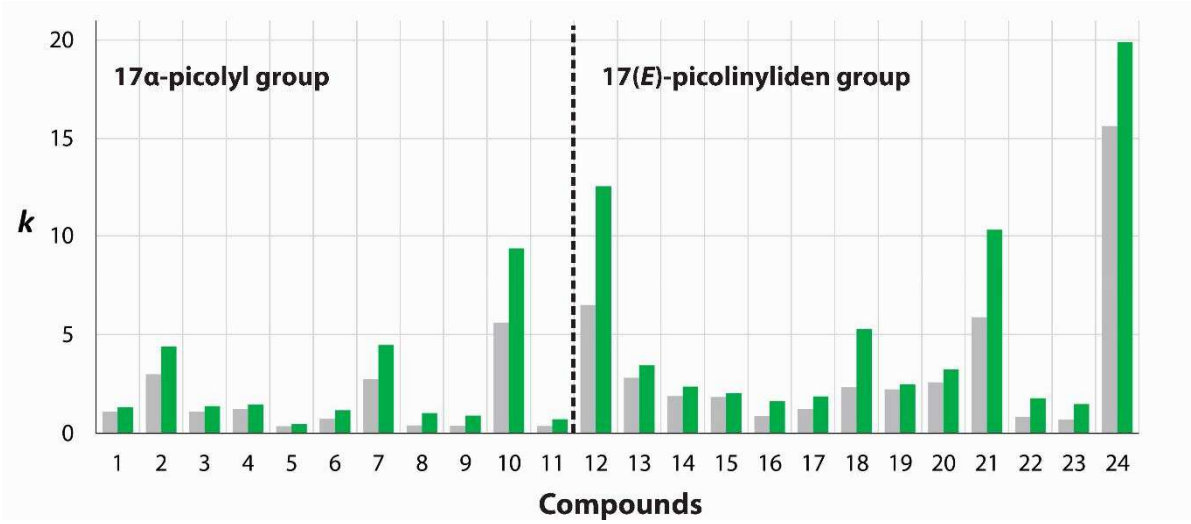


Figure 1. Capacity factors (*k*) of the analyzed androstane derivatives obtained by using two mobile phases with different methanol/water ratio (- methanol/water = 90/10, - methanol/water = 70/30)

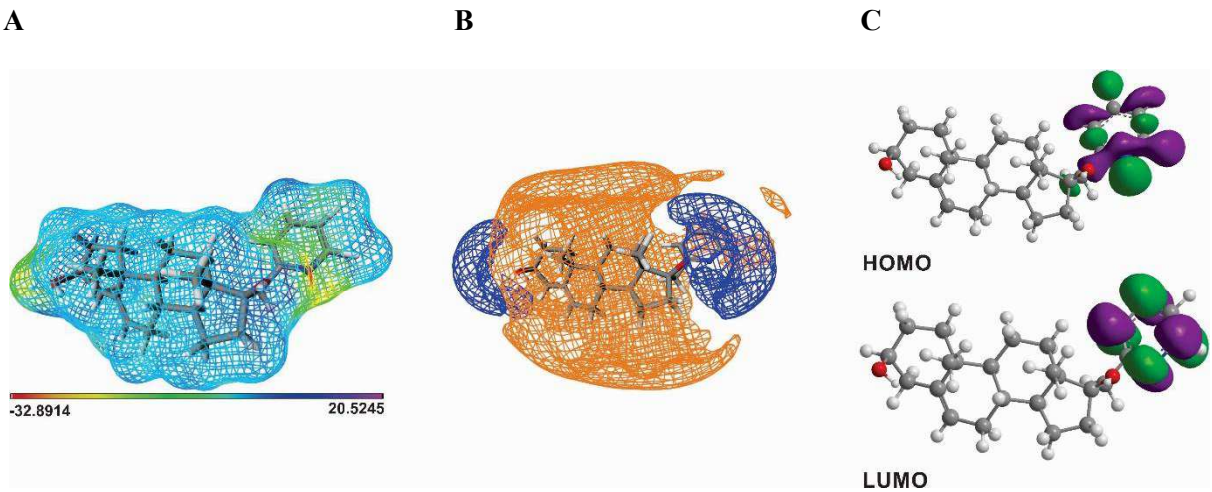


Figure 2. Computational modeling of Poisson-Boltzmann electrostatic potential surface (A), hydrophilic ( ) and lipophilic ( ) surfaces (B) by Bioluminate® program and HOMO-LUMO orbitals of 17α-picolyl-androst-5-en-3β,17β-diol (compound **1**) projected by ChemBio3D v. 14 program (C)

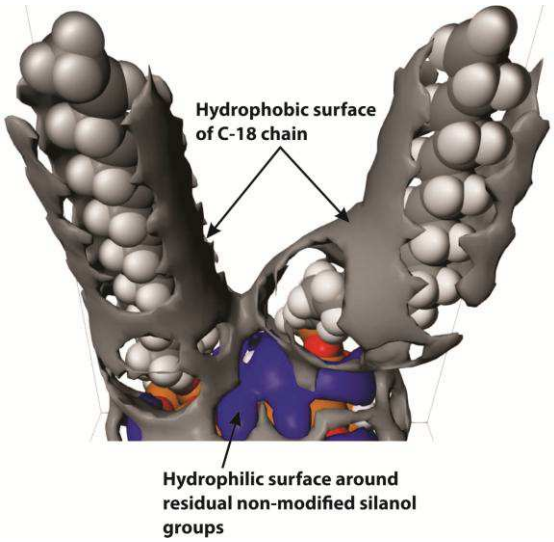


Figure 3. 3D model of hydrophilic and lipophilic surfaces of a segment of C-18 stationary phase with redundant silanol groups, modeled by Bioluminate® progam

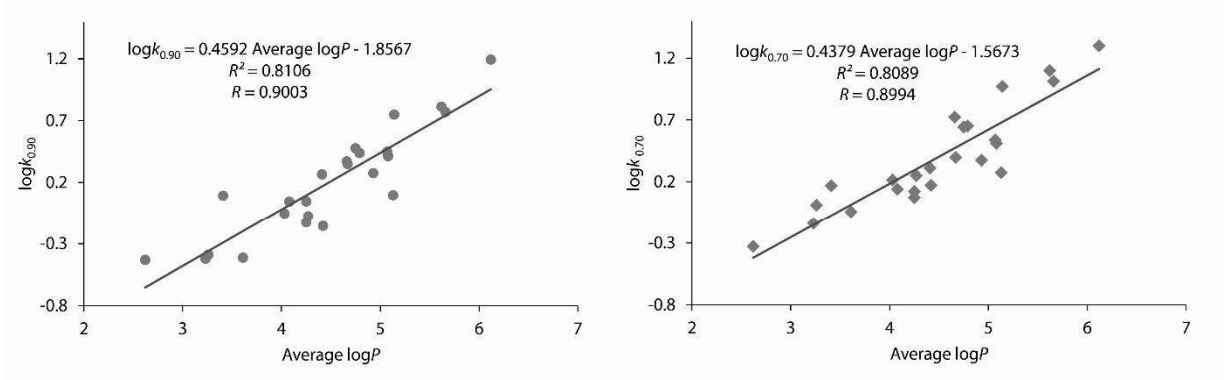


Figure 4. Linear correlation between *in silico* lipophilicity (Average log*P*) and experimental chromatographic lipophilicity (log$k_{0.90}$ and log$k_{0.70}$) defined by using two mobile phases (φ = 0.70 and φ = 0.90)

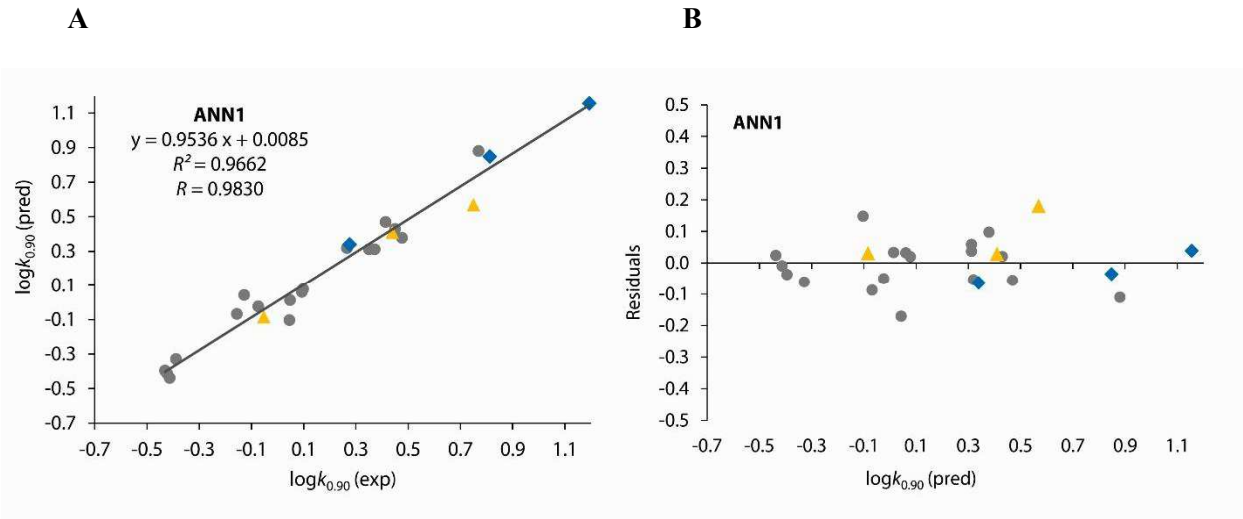


Figure 5. The comparison of experimental log$k_{0.90}$ and log$k_{0.90}$ predicted by ANN1 model (A) and distribution of the residuals (B) (● - calibration set, ▲ - test set, ◆ - validation set)

Figure 1

Click here to download high resolution image

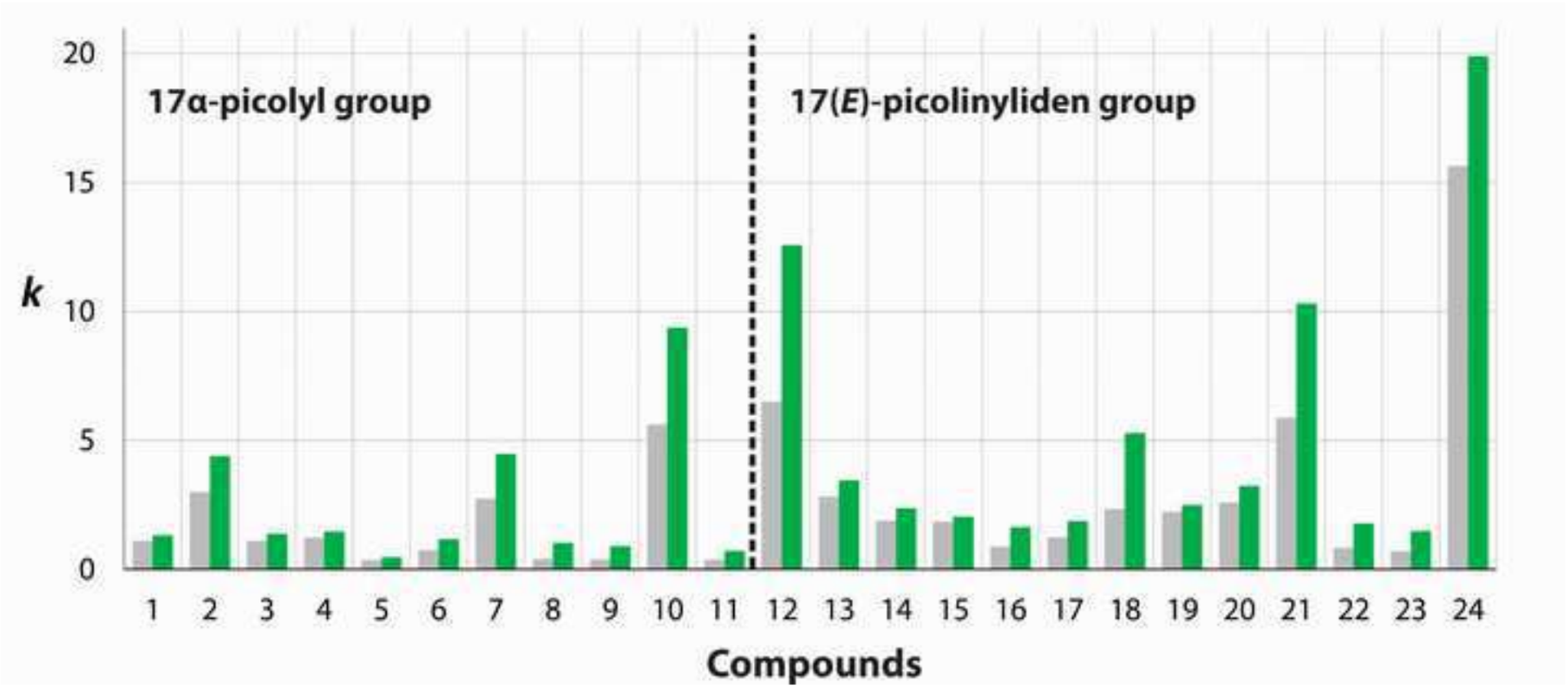


Figure 2
Click here to download high resolution image

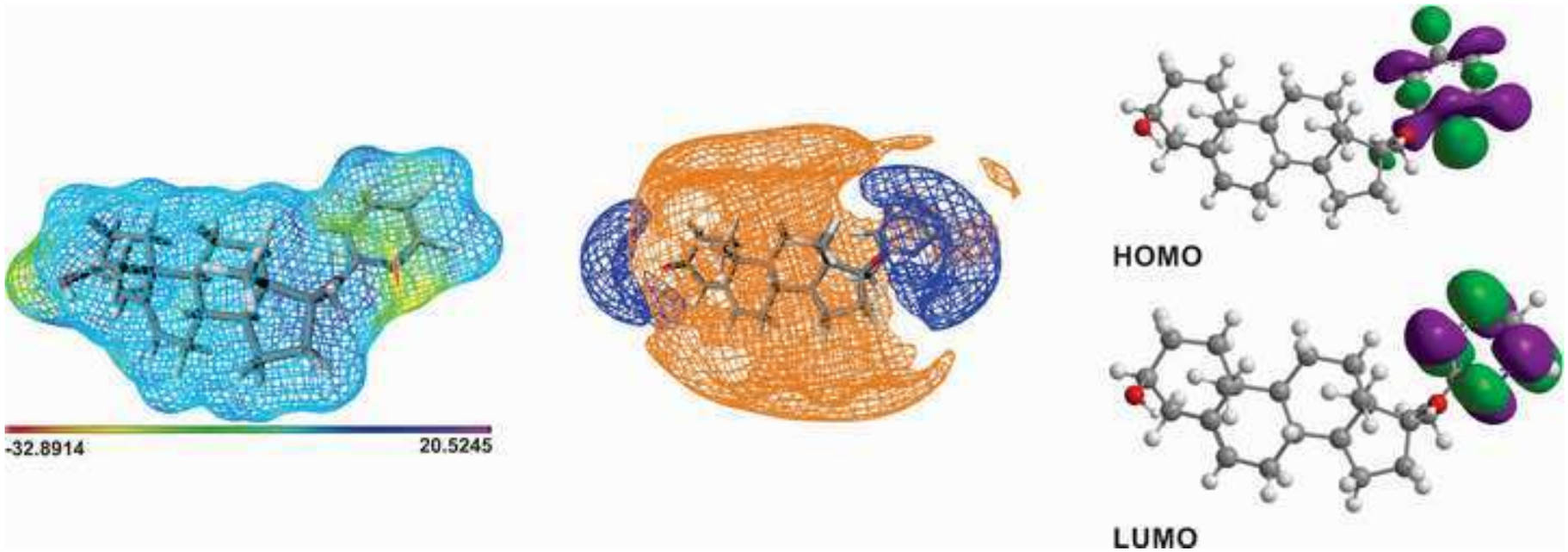


Figure 3

Click here to download high resolution image

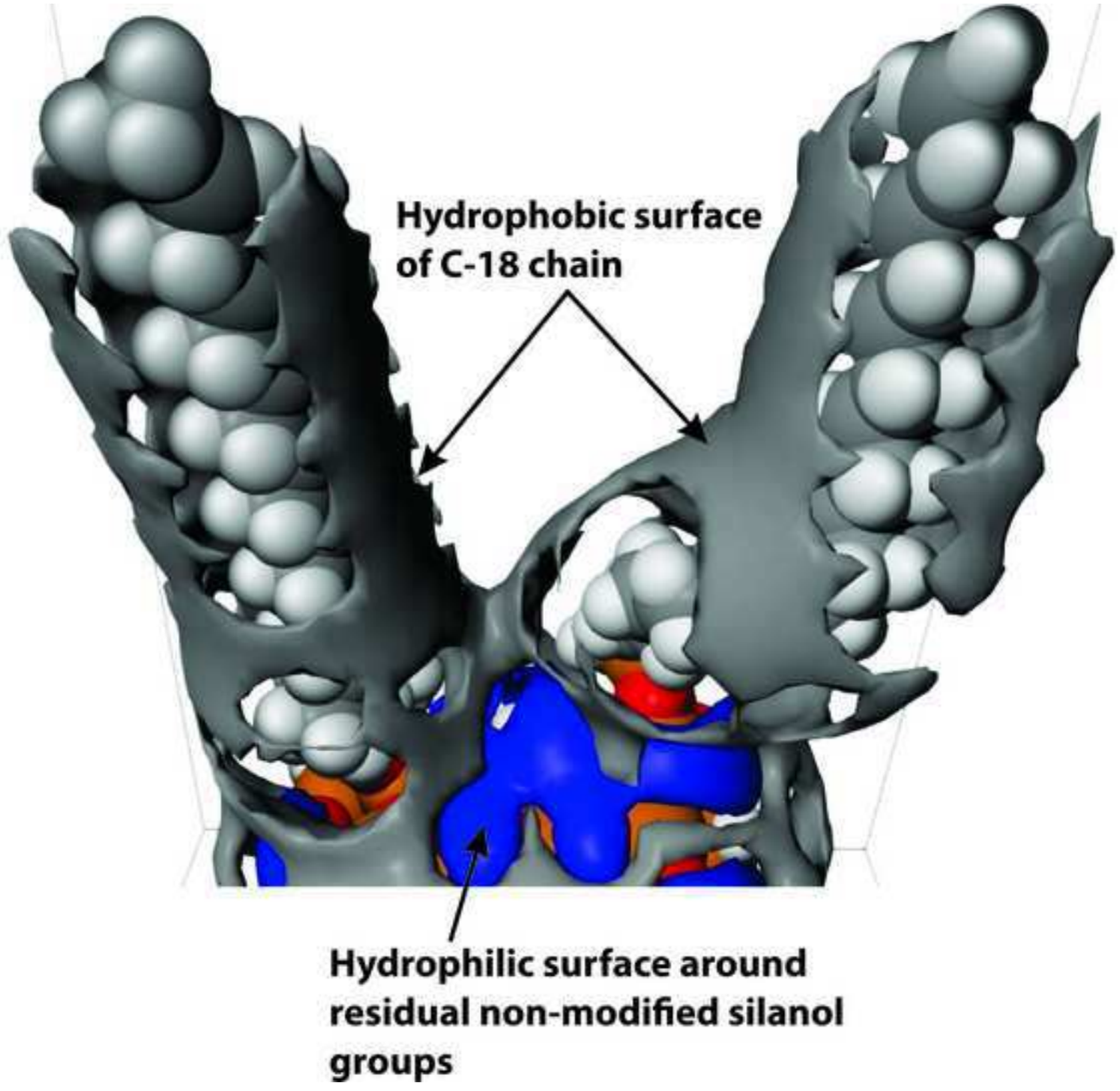


Figure 4

Click here to download high resolution image

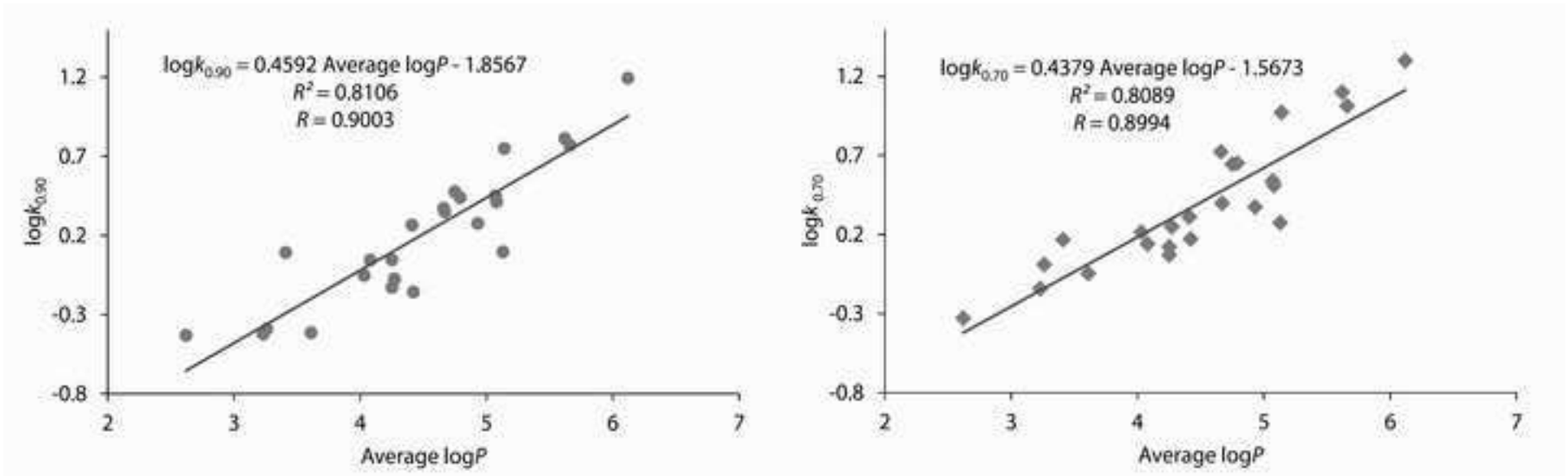
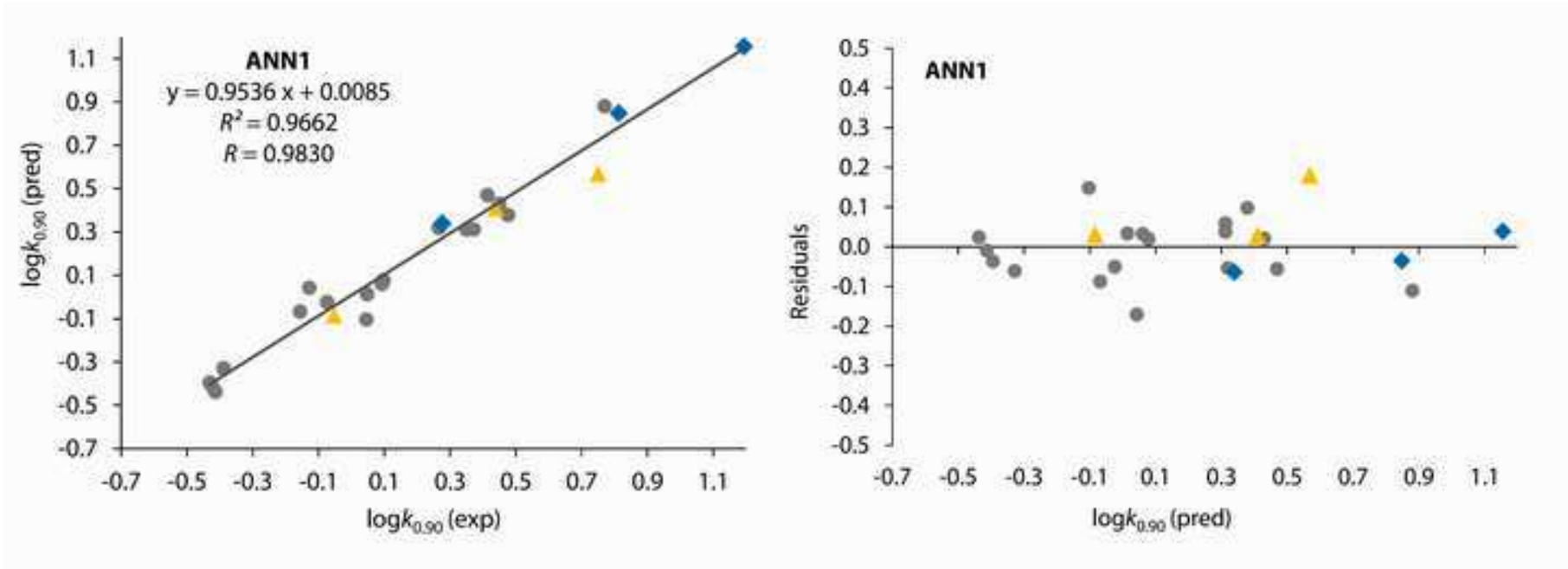


Figure 5

Click here to download high resolution image



Supplementary Material

Click here to download Supplementary Material: Kovacevic et al - Supporting Information.pdf