

Scientific paper

QSRR Analysis in Characterization of Some Benzimidazole Derivatives

Milica Karadžić,* Sanja Podunavac-Kuzmanović, Lidija Jevrić and Strahinja Kovačević

Faculty of Technology Novi Sad, University of Novi Sad, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia

* Corresponding author: E-mail: mkaradzic@hotmail.com
phone: +381 64 9415248; fax: +381 21 450413

Received: 16-10-2014

Abstract

In this paper, quantitative structure-retention relationship study has been applied in order to correlate obtained retention parameter R_M^0 and two groups of molecular descriptors, for eleven investigated benzimidazole derivatives. Principal component analysis (PCA), followed by hierarchical cluster analysis (HCA), linear regression (LR) and multiple linear regression (MLR), was applied in order to identify the most important molecular descriptors. Mathematical models were established and the best models were further validated by leave-one-out (LOO) technique as well as by the calculation of the statistical parameters. Statistically significant models were established.

Keywords: Benzimidazole, linear regression, multiple linear regression, principal component analysis, hierarchical cluster analysis

1. Introduction

Benzimidazoles, as biologically active compounds, are frequently studied group of molecules. It has been confirmed that benzimidazole molecules have antibacterial, antifungal and herbicidal activity.¹⁻³ In last decade it is noticeable that they have anticancer, *in vitro* anti-HIV and antiparasitic activity.⁴⁻⁶ Because of different range of their activities, the chromatographic behavior and physicochemical characteristics of a number of benzimidazole derivatives were studied, applying thin-layer chromatography (TLC).^{7,8}

In chromatographic processes, strict control of experimental conditions can be obtained and that qualifies reversed-phase thin-layer chromatography as suitable technique for estimating physicochemical parameters and biological activity of molecules.⁹⁻¹² For understanding the chromatographic processes, it is very convenient to establish mathematical models. Quantitative structure-retention relationship (QSRR) is a useful technique for determining relationships between chromatographic properties of investigated molecules and molecular descriptors. Established QSRR models can be widely applied for identification of the most useful structural descriptors, prediction of the retention of new synthesized molecules and

identification of unknown analytes.¹³ In QSRR analysis, correlation between retention data (R_M^0 values) and structural parameters (molecular descriptors), can be examined by linear regression (LR) and multiple linear regression (MLR), principal component regression (PCR), partial least squares regression (PLS) and automated neural networks (ANN).

In this study, R_M^0 values were correlated with two groups of descriptors, molecular and *in silico* ADME (absorption, distribution, metabolism and excretion) descriptors. LR and MLR were used for establishing the equations and principal component analysis (PCA) and hierarchical component analysis (HCA) were carried out for data overview.

The goals of this study were to evaluate the retention data by multivariate statistical methods and to find the possible relationship between retention characteristics and molecular and *in silico* ADME descriptors of the investigated benzimidazole derivatives.

2. Material and Methods

The steps in QSRR analysis were: molecular structure optimization using the computer software, molecular

descriptors computation, selection of molecular descriptors, generation of structure-retention models using LR and MLR method and statistical validation.

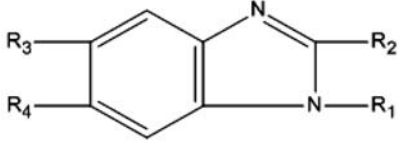
2. 1. Studied Compounds

The chemical structures of investigated benzimidazole derivatives are presented in Table 1. Compounds are divided in three groups: molecules 1–4, 5–8 and 9–11, according to their chemical structure. The compounds were synthesized by a procedure described elsewhere.¹⁴ Experimental procedure of RP TLC separation with C₁₈ silica gel plates and obtained retention data (R_M^0) of analyzed compounds were reported previously.⁷

2. 2. Molecular Modeling and Molecular Descriptors

Two groups of descriptors, molecular and *in silico* ADME descriptors were derived from the chemical struc-

Table 1. Chemical structures of eleven studied benzimidazole derivatives.



Compound	R ₁	R ₂	R ₃	R ₄
1	H	NH ₂	CH ₃	CH ₃
2	C ₂ H ₅	NH ₂	CH ₃	CH ₃
3	<i>n</i> -C ₄ H ₉	NH ₂	CH ₃	CH ₃
4	<i>n</i> -C ₆ H ₁₃	NH ₂	CH ₃	CH ₃
5	H	NH ₂	H	H
6	C ₂ H ₅	NH ₂	H	H
7	<i>n</i> -C ₄ H ₉	NH ₂	H	H
8	<i>n</i> -C ₆ H ₁₃	NH ₂	H	H
9	H	H	CH ₃	CH ₃
10	C ₂ H ₅	H	CH ₃	CH ₃
11	<i>n</i> -C ₄ H ₉	H	CH ₃	CH ₃

Table 2. The values of the molecular and *in silico* ADME descriptors.

Comp.	Molecular descriptors						<i>In silico</i> ADME descriptors						
	MR	P	MV	HE	SAG	TE	DM	GPCR	ICM	KI	NRL	PI	EI
1	49.75	18.93	526.64	-8.14	343.06	17.03	3.796	-0.44	-0.01	-0.06	-1.55	-0.85	-0.05
2	59.39	22.6	625.31	-3.65	391.86	96.7	3.774	-0.47	-0.34	-0.19	-1.40	-0.88	-0.20
3	68.42	26.27	718.97	-2.06	442.6	173.92	3.768	-0.16	-0.15	-0.08	-1.04	-0.54	-0.01
4	77.62	29.94	825.44	-1.23	501.48	254.22	3.76	-0.01	-0.10	0.05	-0.78	-0.33	0.08
5	41.19	15.26	427.16	-10.45	289.29	25.42	3.774	-0.55	0.03	-0.14	-2.00	-1.00	-0.06
6	50.62	18.93	521.51	-5.58	335.97	177.24	3.752	-0.61	-0.38	-0.34	-1.80	-1.07	-0.26
7	59.85	22.6	622.05	-4.52	389.06	257.21	3.75	-0.26	-0.15	-0.21	-1.36	-0.67	-0.04
8	69.05	26.27	729.84	-3.92	452.76	339.01	3.748	-0.08	-0.06	-0.04	-1.02	-0.44	0.06
9	42.25	17.58	491.37	-4.3	319.78	166.02	3.79	-0.66	-0.12	-0.45	-1.58	-1.12	-0.30
10	51.69	21.25	588	-0.25	367.67	242.357	3.768	-0.06	-0.33	-0.41	-0.91	-1.02	-0.22
11	60.92	24.92	692.28	0.66	425.21	232.041	3.765	-0.26	-0.16	-0.26	-0.60	-0.64	-0.01

ture. Modeling of studied compounds was performed by ChemBioDraw Ultra 12.0 for 2D structures and ChemBio3D for 3D molecular structures.¹⁵ Derived 3D molecular structures were subjected to the energy minimization using molecular mechanics force field method (MM2). The minimization was performed until the root mean square gradient (RMS) reached a value smaller than 0.1 kcal/Åmol.

Three types of molecular descriptors were derived (Table 2): variables that describe the physicochemical properties of the whole molecules such as molar refractivity (MR), molar volume (MV), hydration energy (HE) and surface area grid (SAG); total energy (TE) that is a quantum chemical property; polarizability (P) and dipole momentum (DM) as electronic features of the molecules. *In silico* ADME descriptors were calculated on the basis of 2D structures, using the Molinspiration online program.¹⁶ Calculated *in silico* ADME descriptors are (Table 2): G protein-coupled receptors ligand (GPCR), ion channel modulator (ICM), kinase inhibitor (KI), nuclear receptor ligand (NRL), protease inhibitor (PI) and enzyme inhibitor (EI).

2. 3. Chemometric Methods

In QSRR analysis, correlation between retention data and various empirical, semi-empirical and non-empirical structural parameters, are usually examined by the MLR.¹³ The main aim in QSRR analysis is to reduce the number of variables and to detect structure in the relationships between variables, by various statistical methods of explorative analysis, classification methods and regression methods.^{17,18}

PCA is a useful statistical technique for reducing the amount of data when there is correlation present, retaining as much as information as possible. This statistical technique calculates new, latent variables by a combination of the original variables. The data are projected into a few principal components (PCs) that are linear combinations of the original variables. Each PC is characterized by sco-

res that are the new coordinates of the projected objects and loadings that reflect the direction with respect to the original variables.¹⁹

HCA is a method for dividing a group of objects into clusters so that similar objects are in the same cluster. This type of analysis searches for objects which are close together in the variable space. Cluster hierarchy is displayed as a tree diagram named dendrogram, where the horizontal axis represents the distance or dissimilarity between the clusters.

LR is used for establishing the relationship between dependent variable and just one independent variable. It attempts to model the relationship between two variables by fitting a linear equation to observed data. General LR model can be written using following equation:

$$y = a \cdot x + b \quad (1)$$

where y is dependent variable (quantitative property to predict), a the slope, x an independent variable (descriptor) and b the intercept.

MLR is used for quantification of the relationship between more than one independent variables and a dependent variable. A great problem in MLR modeling is how to avoid multicollinearity. As the diagnostic tool, variance inflation factor (VIF) is used to check the impact of multicollinearity in the MLR models. In the literature it is considered that VIF factor greater than 10 indicates multicollinearity.^{13,20} Very important aspect of QSRR study is model validation. Standard statistical parameters for model validation were used: Pearson's correlation coefficient (r), F-test (Fisher's value) and standard error of estimation (s), and *cross-validation* parameters (*cross-validation* coefficient of determination (r^2_{cv}), adjusted coefficient of determination (r^2_{adj}), predicted residual sum of squares (PRESS), total sum of squares (TSS) and stan-

dard deviation based on predicted residual sum of squares (S_{PRESS}).²¹ High values of these statistical characteristics (r^2_{cv} , $r^2_{adj} > 0.5$) indicate high predictive power of the equations.²²

3. Results and Discussion

3.1. PCA

PCA was performed on both sets of molecular descriptors in order to reveal some similarities among studied molecules. The analysis was carried out by Statistica v. 10 program.²³ For molecular descriptors PCA resulted in a model that explains 89.78% of total variance with two significant PCs. The first principal component accounted for 77.05% of data variance and the second one for 12.73% (Figure 1a). As it can be observed from the loading graph (Figure 2a), all descriptors have a significant negative influence on PC1 while only DM has a high positive influence. Along the PC2 axis, TE descriptor has the most positive influence while DM has the highest negative influence. From score plots, any type of grouping of the molecules cannot be observed along the PC1 or PC2 axis.

For *in silico* ADME descriptors, the model explains 89.25% of total variance, also with two significant PCs. The first principal component accounted for 64.26% of data variance and the second one for 24.99% (Figure 1b). As it can be observed from the loading graph (Figure 2b), all descriptors have a significant negative influence on PC1. Along the PC2 axis, NRL descriptor has the most positive influence while ICM has the highest negative influence. From score plot for molecular descriptors, any type of grouping of the molecules cannot be observed along the PC1 or PC2 axis. On score plot for *in silico* ADME descriptors, two outliers can be observed, molecules 2 and 5.

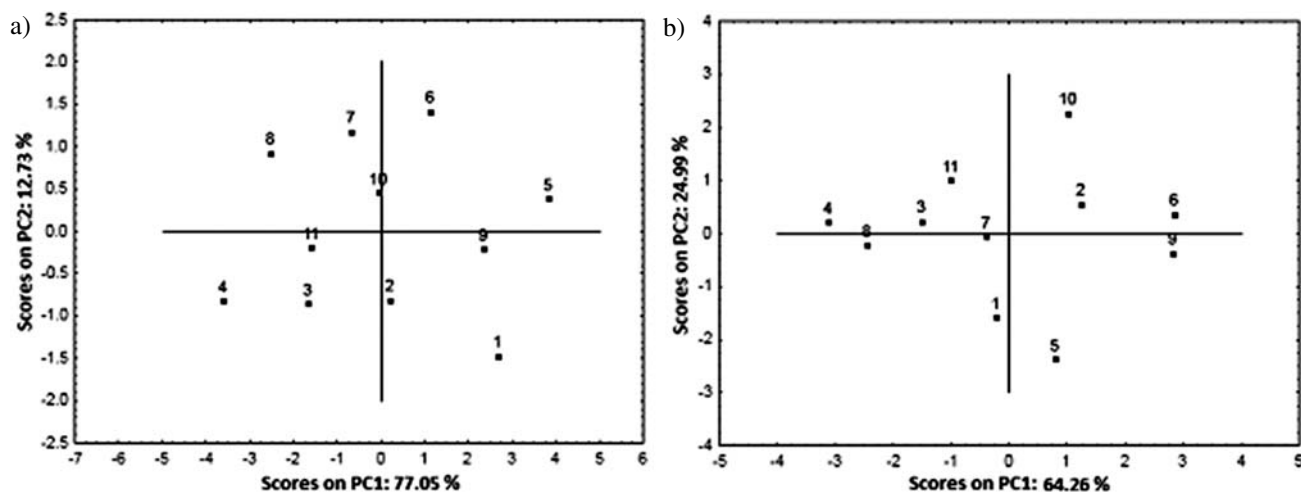


Figure 1. Score plots of molecular (a) and *in silico* ADME (b) descriptors.

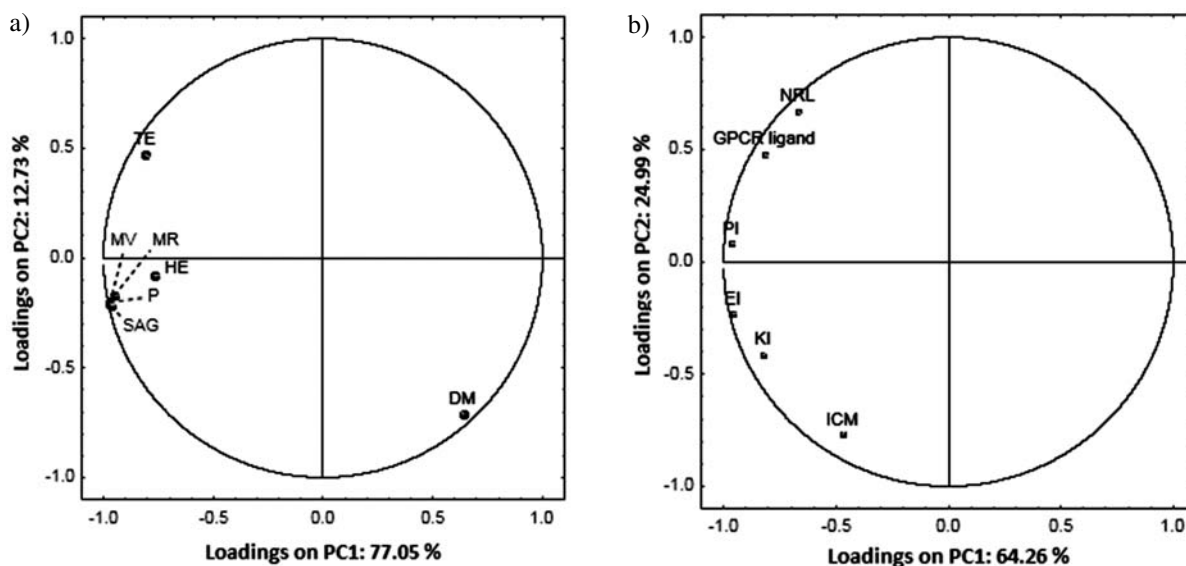


Figure 2. Loading plots of molecular (a) and *in silico* ADME (b) descriptors.

3. 2. HCA

Clustering is based on Ward's linkage method and Euclidean distance. HCA was conducted by using NCSS 2007 and GESS 2006 software.²⁴ Dendrogram based on molecular descriptors (Figure 3a) shows two well-separated clusters. One cluster consists of basic molecules in every group (5, 9, 1) that have hydrogen in position R_1 . Their molar refractivity is significantly different from the other molecules, as it is confirmed by calculated values. Second cluster contains compounds that have alkyl groups (ethyl, butyl and hexyl group) in position R_1 . It can be concluded that obtained dendrogram is the same as on the PC1-PC2 score plot (Figure 1a). Dendrogram based on *in silico* ADME descriptors resulted in two main clusters

(Figure 3b). The first cluster consists of molecules 10, 9, 6 and 2, that have the highest enzyme inhibition ability and in second cluster compounds with lower values of enzyme inhibition ability are positioned. Compounds in HCA are grouped same as in PCA (Figure 1b).

3. 3. LR and MLR

LR and MLR were conducted using NCSS 2007 and GESS 2006 software.²¹ For MLR models, two molecular descriptors that have the low value of intercorrelation coefficient were used. Each constructed LR and MLR model had to be statistically valid. In the present study, models that contain two independent variables were chosen,

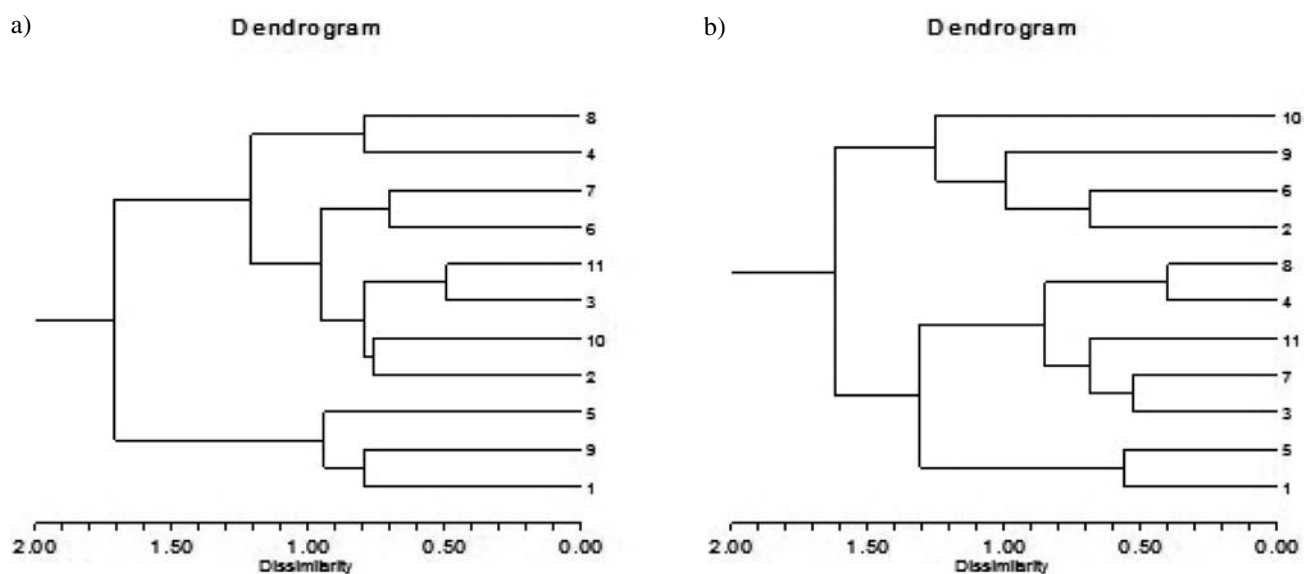


Figure 3. Clustering of examined compounds in the space of molecular (a) and *in silico* ADME descriptors (b).

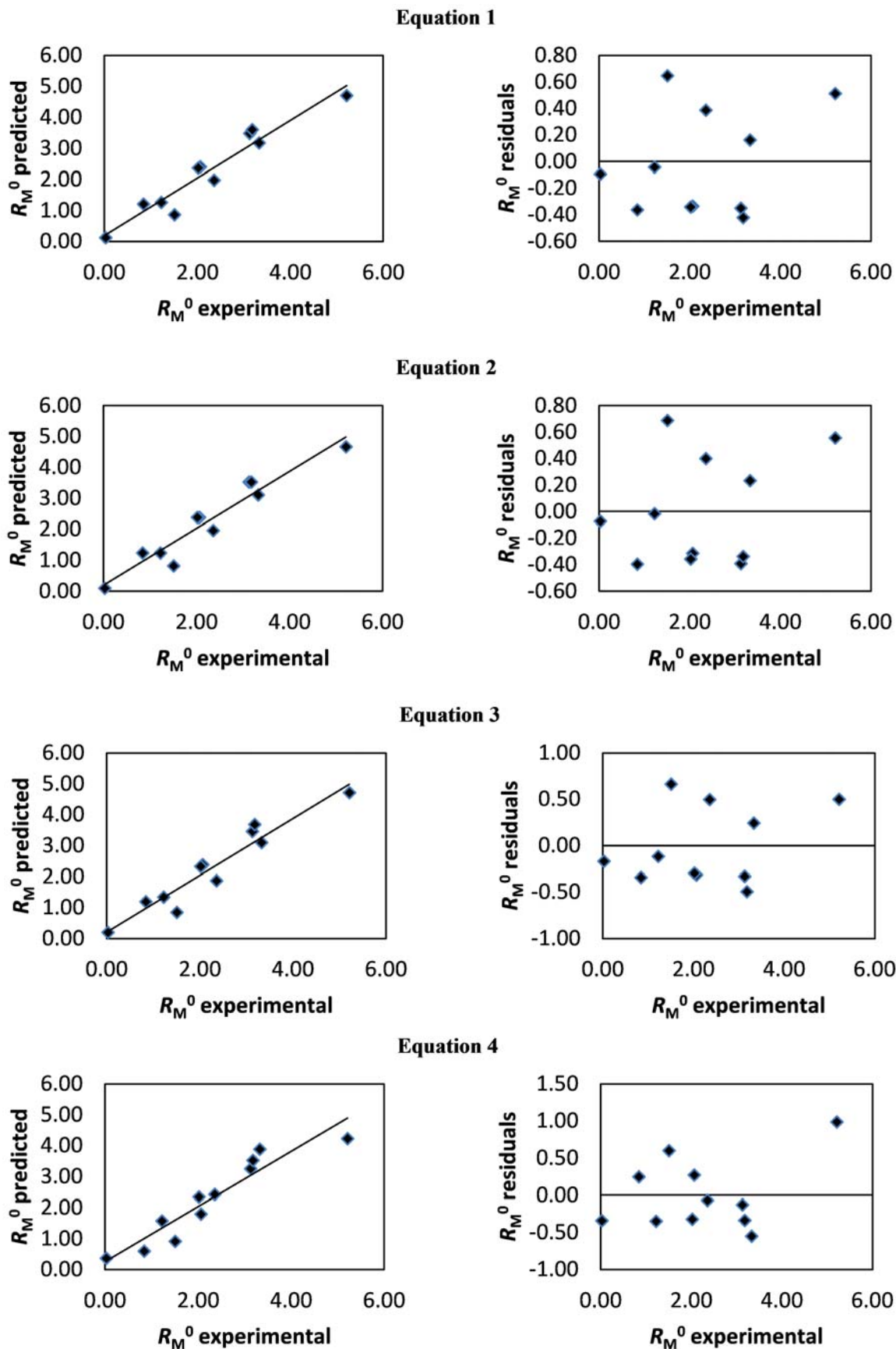


Figure 4. Plots of predicted versus experimentally observed R_M^0 values and plots of residual values versus the experimentally observed R_M^0 values.

Table 3. Statistical parameters for linear dependence between R_M^0 and calculated descriptors.

Variables		Linear Regression: $y = a + b \cdot x$					Eq.
y	x	a	b	r	F	s	
R_M^0	MV	-4.7905	0.0115	0.9247	110.59	0.4112	1
R_M^0	P	-4.6407	0.3106	0.9185	101.45	0.4278	2

Table 4. Statistical parameters for multilinear dependence between R_M^0 and calculated descriptors.

Variables			Multiple Linear Regression: $y = a + b \cdot x_1 + c \cdot x_2$						Eq.	
y	x_1	x_2	a	b	c	r	F	s	VIF	
R_M^0	NRL	PI	6.4774	1.9323	2.2457	0.8891	32.08	0.5293	1.9	4
R_M^0	KI	NRL	6.1691	2.4616	2.6865	0.8651	25.65	0.5839	6.2	5

Table 5. Cross-validation parameters for models 1–4.

Eq.	r_{cv}^2	r_{adj}^2	PRESS	TSS	PRESS/TSS	S_{PRESS}
1	0.8732	0.9164	2.5629	20.2172	0.1268	0.4827
2	0.8635	0.9095	2.7600	20.2172	0.1365	0.5009
3	0.7571	0.8614	4.9116	20.2172	0.2429	0.6682
4	0.6682	0.8314	6.7071	20.2172	0.3318	0.7809

according to the number of studied compounds. Established LR and MLR equations, with both sets of descriptors, free of multicollinearity ($VIF < 10$) and statistically significant are presented (Table 3 and 4). The statistical quality of the generated models was determined by r , s and F for statistical significance.

Equations 1–4 were *cross*-validated by leave-one-out method (Table 5). High values of r_{cv}^2 and r_{adj}^2 ($r_{cv}^2, r_{adj}^2 > 0.5$) and PRESS values significantly less than TSS for all four models indicates that these models have very good predictive power.²⁵ In equations 1–4, all descriptors have a positive influence on the retention.

Usefulness of the established models can be confirmed by the plots of predicted *versus* experimentally observed R_M^0 values and the plots of residual values *versus* the experimentally observed R_M^0 values (Figure 4). The plots of residual values *versus* the experimentally observed R_M^0 values shows that the residuals are randomly distributed around the $y = 0$ axis. On the result of given *cross*-validation parameters and plots, it can be concluded that better models are obtained with molecular than with *in silico* ADME descriptors. The best models are obtained with equations 1 and 2 and based on the same criteria, models 3 and 4 are satisfactory.

4. Conclusion

The aim of this study was to evaluate the retention data, obtained by RP TLC, by multivariate statistical methods and to find the best established models. PCA and HCA were carried out and mathematical models were de-

veloped. PCA did not show grouping among the studied molecules with both sets of molecular descriptors. The results of HCA showed two well-separated clusters in both cases. The usefulness of the established models was confirmed by standard and *cross*-validation statistical parameters. Comparison of the experimental and predicted, and experimental and residual values confirmed that established MLR models can be successfully used in the prediction of R_M^0 values. In addition, on the basis of presented results it can be concluded that the molecular and *in silico* ADME descriptors could be successfully used for predicting of the retention parameters obtained by RP TLC. Predictive ability of presented models allows us to estimate the retention behavior for structurally similar compounds and reduces the analysis time of investigated compounds.

5. Acknowledgment

These results are the part of the projects No. 31055, No. 172012 and No. 172014, supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia and the project No. 114-451-1156/2014-02 and No. 114-451-3503/2011-2014, financially supported by the Provincial Secretariat for Science and Technological Development of Vojvodina.

6. References

1. S. O. Podunavac-Kuzmanović, D. D. Cvetković, D. J. Barna, *J. Serb. Chem. Soc.* **2008**, *73*, 967–978.

2. G. Ayhan-Kilcigil, N. Altanlar, *Turk. J. Chem.* **2006**, *30*, 223–228.
3. S. O. Podunavac-Kuzmanović, D. D. Cvetković, *Centr. Eur. J. Occupat. Environ. Med.* **2006**, *12*, 55–60.
4. M. Boiani, M. Gonzales, *Med. Chem.* **2005**; *5*: 409–424.
5. E. Iwao, K. Yamamoto, F. Hirayama, K. Haga, *J. Infect. Chemother.* **2004**; *10*: 90–96.
<http://dx.doi.org/10.1007/s10156-004-0299-1>
6. S. M. Rida, S. A. El-Hawash, H. T. Fahmy, A. A. Hazzaa, M. M. El-Meligy, *Arch. Pharm. Res.* **2006**; *29*: 826–833.
<http://dx.doi.org/10.1007/BF02973901>
7. N. U. Perišić-Janjić, S. O. Podunavac-Kuzmanović, *J. Planar Chromat.* **2008**; *21*: 135–141.
<http://dx.doi.org/10.1556/JPC.21.2008.2.11>
8. N. U. Perišić-Janjić, S. O. Podunavac-Kuzmanović, J. S. Balaž, Đ. Vlaović, *J. Planar Chromat.* **2000**, *13*, 123–129.
9. R. Kaliszán, *J. Chromatogr. A.* **1993**, *656*, 417–435.
[http://dx.doi.org/10.1016/0021-9673\(93\)80812-M](http://dx.doi.org/10.1016/0021-9673(93)80812-M)
10. R. Kaliszán, M. A. van Straten, M. Markuszewski, C. A. Cramers, H. A. Claessens, *J. Chromatogr. A.* **1999**, *855*, 455–486. [http://dx.doi.org/10.1016/S0021-9673\(99\)00742-6](http://dx.doi.org/10.1016/S0021-9673(99)00742-6)
11. M. Markuszewski, R. Kaliszán, *J. Chromatogr. B.* **2002**, *768*, 55–66.
[http://dx.doi.org/10.1016/S0378-4347\(01\)00485-6](http://dx.doi.org/10.1016/S0378-4347(01)00485-6)
12. K. Héberger, *J. Chromatogr. A.* **2007**, *1158*, 273–305.
<http://dx.doi.org/10.1016/j.chroma.2007.03.108>
13. S. Z. Kovačević, L. R. Jevrić, S. O. Podunavac-Kuzmanović, E. S. Lončar, *Centr. Eur. J. Chem.* **2013**, *11*, 2031–2039.
<http://dx.doi.org/10.2478/s11532-013-0328-y>
14. Đ. Vlaović, J. Čanadanović-Brunet, J. Balaž, I. Juranić I, D. Đoković, K. Mackenzie K, *Biosci. Biotech. Biochem.* **1992**, *56*, 199–206. <http://dx.doi.org/10.1271/bbb.56.199>
15. ChemBioOffice 2010, PerkinElmer Informatics, 2175 Mission College Boulevard, Santa Clara, California, USA, <http://www.cambridgesoft.com>
16. Molinspiration Cheminformatics, Calculation of molecular properties and bioactivity score, <http://www.molinspiration.com/>
17. T. Đaković-Sekulić, Z. Lozanov-Crvenković, N. Perišić-Janjić, *Novi Sad J. Math.* **2008**, *38*, 39–46.
18. A. Z. Dudek, T. Arodz, K. Gálvez, *J. Comb. Chem. High Throughput Screen.* **2006**, *9*, 213–228.
<http://dx.doi.org/10.2174/138620706776055539>
19. S. Z. Kovačević, S. O. Podunavac-Kuzmanović, L. R. Jevrić, *Acta Chim. Slov.* **2013**, *60*, 756–762.
20. R. M. O'Brien, *Qual. Quant.* **2007**, *41*, 673–690.
<http://dx.doi.org/10.1007/s11135-006-9018-6>
21. S. Z. Kovačević, L. R. Jevrić, S. O. Podunavac-Kuzmanović, N. D. Kalajdžija, E. S. Lončar, *Acta Chim. Slov.* **2013**, *60*, 420–428.
22. D. A. Belsley, *Am. Stat.* **1984**, *38*, 73–77.
23. StatSoft Inc., 2300 East 14th Street, Tulsa, Oklahoma, USA, <http://www.statsoft.com/>
24. NCSST Statistical Software, <http://www.ncss.com/>
25. S. O. Podunavac-Kuzmanović, L. R. Jevrić, S. Z. Kovačević, N. D. Kalajdžija, *APTEFF* **2012**, *43*, 273–282.
<http://dx.doi.org/10.2298/APT1243273P>

Povzetek

V tem delu smo proučevali povezavo med strukturo in retencijo z namenom, da bi lahko korelirali dobljeni retencijski parameter R_M^0 in dve skupini molekulskih deskriptorjev za enajst benzimidazolovih derivatov. Za identifikacijo najpomembnejših molekularnih deskriptorjev smo uporabili analizo glavnih komponent (Principal component analysis, PCA), ki je ji sledila hierarhična klastrska analiza (hierarchical cluster analysis, HCA), linearna regresija (LR) in večkratna linearna regresija (multiple linear regression, MLR). Postavili smo matematični modele ter najboljšega validirali z »leave-on-out« (LOO) metodo kot tudi z izračunom statističnih parametrov.