Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. *85* (1) 9–23 (2020)
JSCS–5279

UDC 635.74+66.061.14:547.596:665.52/.54+539.16:510.644
Original scientific paper

Prediction of the GC–MS retention time for terpenoids detected in sage (*Salvia officinalis* L.) essential oil using QSRR approach

BRANIMIR PAVLIĆ[1], NEMANJA TESLIĆ[2#], PREDRAG KOJIĆ[1#] and LATO PEZO[3*]

[1]University of Novi Sad, Faculty of Technology, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia, [2]Institute for Food Technology, University of Novi Sad, Boulevard cara Lazara 1, 210000 Novi Sad, Serbia and [3]Institute of General and Physical Chemistry, University of Belgrade, Studentski trg 12/V, 11000 Beograd, Serbia

(Received 22 May, revised 26 July, accepted 10 September 2019)

Abstract: This work aimed to obtain a validated model for prediction of retention time of terpenoids isolated from sage herbal dust using supercritical fluid extraction. In total 32 experimentally obtained retention time of terpenes, which were separated and detected by GC–MS were further used to build a prediction model. The quantitative structure–retention relationship was employed to predict the retention time of essential oil compounds obtained in GC–MS analysis, using six molecular descriptors selected by a genetic algorithm. The selected descriptors were used as inputs of an artificial neural network, to build a retention time predictive quantitative structure–retention relationship model. The coefficient of determination for training cycle was 0.837, indicating that this model could be used for prediction of retention time values for essential oil compounds in sage herbal dust extracts obtained by supercritical fluid extraction due to low prediction error and moderately high r^2. Results suggested that a 2D autocorrelation descriptor AATS0v was the most influential parameter with an approximately relative importance of 25.1 %.

Keywords: sage herbal dust; supercritical fluid extraction; terpenoids; QSRR; artificial neural networks.

INTRODUCTION

Sage (*Salvia officinalis* L.) represents one of the most thoroughly investigated plants of the *Lamiaceae* family,[1] which has been known for their aromatic and medicinal properties. Sage possesses many biological activities such as antimicrobial, preservative, immunomodulatory, antioxidant and anticancer properties, which are attributed to the presence of terpenoids and polyphenols, *i.e.*, two major classes of sage bioactive compounds.[2] Terpenoids are the most dominant

* Corresponding author. E-mail: latopezo@yahoo.co.uk
# Serbian Chemical Society member.
https://doi.org/10.2298/JSC190522097P

compounds in sage essential oil (EO) which could range from 0.7 to 5.2 % (m/m).[3] The major compounds in sage EO are oxygenated monoterpene ketones, such as α-thujone, β-thujone and camphor[4], however, various monoterpene hydrocarbons, sesquiterpenes and diterpenes could be also present in significant amount depending on the variety, genetic diversity, geographical origin, nutritional status of the plants, physiological age development stage, *etc*. Chemical composition of EOs is also determined by harvest timing, post-harvest drying and storage conditions.[5] Furthermore, obtaining of sage EO could be performed by different conventional (hydrodistillation and solvent extraction) and novel (supercritical fluid extraction) extraction techniques which could also have significant impact on its chemical profile and yield. Since sage EO could be used in various pharmaceutical, cosmetic and food formulations, it is necessary to determine its chemical profile in order to standardize raw material and extracts/EO obtained by each technique.

The EOs are complex mixtures of different classes of organic compounds which belong to the broad spectrum of chemical structures. They are isolated mostly from various aromatic medicinal plants and exhibit antioxidant, antibacterial, antifungal, antimicrobial and herbicidal properties due to their specific chemical profile.[5] Terpenoids are the most abundant in EOs, however, various organic volatiles (alcohols, ketones, aldehydes, esters, *etc*.) could be also present in them. EO compounds could significantly vary in structure, molecular weight, physicochemical properties (solubility, retention index, *etc.*) and bioactivity. Therefore, it is essential to provide mathematical models associated with the aforementioned feature.

The gas chromatography (GC) is a technique widely used to separate and analyze volatile compounds (*e.g.*, terpenoids from EOs) or molecules with poor volatility that can be chemically changed (*e.g.*, derivatized) to more suitable molecules for GC analysis. The GC is a powerful technique because it produces a single parameter (retention index) which can be used to identify any volatile compound under well-defined analytical conditions.[6] Elution or retention of each compound is a complex phenomenon determined by several intermolecular forces such are dipole–dipole forces, dipole-induced forces, hydrogen bonds, *etc.*[7] The retention profile of compounds can be determined by measuring various parameters, *e.g.*, retention time (*RT*), retention distance, linear-temperature retention index and Kováts retention index.[6–8]

Identification of new compounds with GC technique was significantly improved with introduction of mass spectra (MS) detectors. The possibility to separate and identify molecules according to their retention profiles and molecular mass of fragments, makes GC–MS systems more selective and sensitive, and distinguishes them from other GC configurations (*e.g.*, GC coupled with flame ionization detector). In certain cases, even the most advanced GC–MS systems are

not sufficient to accurately identify new compounds due to various reasons. In particular, certain compounds may have minor differences in their molecular structure (*e.g.*, isomers)[9], yielding similar fragments after ionization process. Also, new compounds are not always included in MS libraries, thus their identification could be rather difficult and may lead to false positive results.[7] Another issue could be related to the availability and purity of analytical standards which are used for molecules identification. This rise an idea to use models as additional tool for identification of new compounds and predictive RT. Therefore, in the light of the rapid hardware and software revolution it is easier to synthesize a compound with a definite chemical structure by computer software (*in silico*) than in the traditional way. Chemical software can correctly deduce established chemical structures and predicts the properties of the specific reaction products.[10] Quantitative structure– (chromatographic) retention relationship (QSRR) strings the chemical structure to its predicted physicochemical or biological properties. The chemical structure is presented by molecular descriptors. They are the transformation of the chemical information encoded within the symbolic representation of a molecule into a mathematical number. In order to get statistically significant relationships and avoid overfitting, it is necessary to have large sets of the property parameters. Chromatography is a unique method that yields a large number of the quantitatively comparable, reproducible and precise retention data for large sets of the analytes. In recent years the numerous publications are related to the QSRR analysis.[11–17] In the end, the connection between the molecular descriptors and the retention time can be established by numerous machine learning algorithms.[18] In this study, we used the artificial neural network (ANN) that already proved excellent predictability through the published literature.[10,19,20]

Furthemore, the aim of this work was to establish a new QSRR model for predicting the RTs of some EO compounds in sage herbal dust extracts obtained by supercritical fluid extraction by GC chromatography using the genetic algorithm (GA) variable selection method and the ANN technique.

EXPERIMENTAL

Retention time data

The analysis conditions, equipment and retention time of sage volatile compounds isolated with supercritical fluid extraction were reported in previous study.[21] In brief, sage terpenoids were identified using gas chromatography system (Agilent GC890N) equipped with mass spectrometer detector (Agilent MS 5759) and HP-5MS column (30 m×0.25 mm×0.25 μm). Mobile phase was helium with flow rate of 2 mL/min. Isolated EOs were dissolved in methylene chloride (about 1 mg/mL) and injected volume of solution was 5 μL with split ratio 30:1. The temperature of the injector was 250 °C, detector temperature 300 °C; initial temperature was 60 °C with linear increase of 4 °C/min to 150 °C. Sage terpenoids were identified using the NIST 05 and Wiley 7*n* data base.

QSRR analysis

The procedure of the QSRR model construction using molecular descriptors is showed below.

1. Collect molecular structures dataset as a .smi file (simplified molecular input line entry specification) from various free chemical compound database, such as LigandBox, Zink 12 or PubChem.
2. Split the dataset into training and test datasets in order to determine the predicted performance of the model. Data should be randomly and independently chosen.
3. Calculate specified molecular descriptors of each compound in the datasets. This can be done by several free available molecular descriptor software.[22–24] In this study we used PaDEL.[25]
4. Construct a reliable model of the training dataset to predict the retention time from PaDEL calculated descriptors using one of the regression methods such as artificial neural networks, partial least squares regression, support vector machine (SVM) and random forest.
5. Evaluate the performance of the developed model by predicting the retention time of the compounds in the test dataset that are not used for model construction. Also, check model overfitting.

The calculation was done by a four-core PC computer (i5-2500K CPU, 3.30GHz). The PaDel database was used to explore the 1875 molecular descriptors (1444 1D and 2D descriptors and 431 3D descriptors), which included: constitutional descriptors, topological descriptors, connectivity indices, information indices, 2D and 3D autocorrelations descriptors, Burden eigenvalues descriptors, eigenvalue-based indices, geometrical descriptors, WHIM descriptors, functional group counts, atom-centred fragments and molecular properties.

Since PaDel database gives an enormous amount of data for each observed compound, it was necessary to use a genetic algorithm (GA), using Heuristic Lab[26] to select the most relevant molecular descriptors for RT prediction. Genetic algorithm (GA)[27,28] is a stochastic optimization method inspired by evolution theory. In this work, it was used to select the most appropriate molecular descriptors for developing a reliable *RT* predictive model for essential oil compounds in sage herbal dust extracts obtained by supercritical fluid extraction. Heuristic Lab software was used to reduce the redundancy in the descriptor data matrix, which was gained using PaDel database. The correlation between the descriptors was examined and collinear descriptors were detected using factor analysis.

Statistical investigation of the data has been performed mainly by the Statistica 10 software.[29]

Artificial neural network (ANN)

A multi-layer perceptron model (MLP) consisted of three layers (input, hidden and output) was used, as proven and quite capable of approximating nonlinear functions.[30] Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm was used for ANN modelling. To improve the behavior of the ANN, both input and output data were normalized. All data points were randomly used to train and develop the ANN; 60 % of data points for training, 20 % of data for validations and 20 % of data for testing the process. The training data set was used for the learning cycle of ANN, and also for evaluation of the optimal number of neurons in the hidden layer, and the weight coefficient of each neuron in the network. It was assumed that the successful training was achieved when learning and cross-validation curves approached zero. ANN results, including the weight values depend on the initial assumptions of parameters necessary for ANN construction and fitting.[31,32] In the same way, the different number of hidden neurons can give different ANN outcomes. In this context, a series of different topo-

logies were used, in which the number of hidden neurons were varied from 1 to 20 and the training process of each network was run ten times with random initial values of weights and biases. The optimization process was performed on the basis of validation error minimization. ANN calculations were performed with Statistica 10.

Global sensitivity analysis

Yoon's interpretation method was used to determine the relative influence of molecular descriptors on retention time.[33] This method was applied on the basis of the weight coefficients of the developed ANN.

RESULTS AND DISCUSSION

Chemical composition and relative content of the analyzed essential oils (EO) compounds and the quantitative profile are presented in Table I.[21] Terpenoid compounds from sage EO obtained by supercritical fluid extraction (SFE) belong to the group of monoterpenes, sesquiterpenes and diterpenes which are either hydrocarbons or oxygenated. Monoterpene hydrocarbons could be acyclic (*β*-myrcene), or cyclic with cyclic monoterpene (limonene) or non-aromatic (*α*- and *γ*-terpinene) rings. Oxygenated monoterpenes could be divided into same subgroups: acyclic (*cis*-linalool oxide), aromatic (carvacrol) and cyclic non-aromatic (*α*-thujone). Sesquiterpenes detected in sage EO could be either hydrocarbons (*γ*-caryophyllene) or oxygenated (caryophyllene oxide, Table I). Diterpenes such as epirosmanol were the most complex compounds detected in sage EO due to their molecular size and number of C atoms (20). Variations in molecular structure size and physicochemical properties (Fig. S-1 of the Supplementary material to this paper) could significantly affect separation in GC and provide different retention time (*RT*).

QSRR model validation

The main step in QSRR analysis is the calculation and the identification of the structural descriptors as the numerically encoded parameters representing the chemical structures. The Authors used the PaDel database in this investigation, and a great number of molecular descriptors were examined. These descriptors could represent many aspects of the investigated compounds and have been successfully used in QSRR investigation. Prior the GA calculation, the factor analysis was performed to eliminate the descriptors with equal or almost equal values for the examined molecules. Only one of the inter-correlated descriptors remained in the GA calculation. As a result of this preliminary consideration, only 300 descriptors remained for GA calculation. GA was used to select the most appropriate molecular descriptors for RT prediction, and the selection of the most relevant descriptors was realized using the evolution simulation.[34–37] Each gene (element) of the population, defined by a "chromosome", represented a subset of the descriptors. The number of elements on each chromosome (*i.e.*, observed compounds) was equal to the number of the molecular descriptors obtained in

TABLE I. Quantitative profile of essential oil compounds in sage herbal dust extracts obtained by supercritical fluid extraction; *RT* – retention time, $RT_{pred.}$ – predicted retention time, VABC – van der Waals volume descriptor, AATSC0c – average centered Broto–Moreau autocorrelation descriptor – lag 0 / weighted by charges, AATS0v – average Broto–Moreau autocorrelation descriptor – lag 0 / weighted by van der Waals volumes, MATS6m – Moran autocorrelation descriptor – lag 6 / weighted by mass, GATS7s – Geary autocorrelation descriptor – lag 7 / weighted by I-state, ASP-1 – Chi path descriptor, average simple path, order 1

Compound	ANN cycle	RT / min	$RT_{pred.}$ / min	VABC	AATSC0c	ASP-1	AATS0v	MATS6m	GATS7s
β-Myrcene	Test	4.46	9.156	175.314	0.008	0.459	172.795	-0.030	0.364
α-Terpinene	Train	5.067	9.003	247.730	0.002	0.466	182.020	-0.186	1.354
p-Cymene	Validation	5.283	7.572	163.887	0.002	0.470	182.020	0.148	2.359
Limonene	Train	5.341	5.963	162.957	0.005	0.423	183.293	-0.524	0.000
Eucalyptol	Test	5.42	5.512	227.361	0.005	0.398	182.879	-0.157	1.219
γ-Terpinene	Train	6.106	5.104	173.607	0.003	0.518	182.020	0.219	2.182
cis-Linalool oxide	Train	6.556	5.963	162.957	0.005	0.423	183.293	-0.524	0.000
Dehydro-*p*-cymene	Train	6.997	6.819	165.594	0.008	0.415	172.795	0.601	0.000
α-Thujone	Train	7.487	11.070	206.339	0.008	0.424	180.490	-0.100	4.253
β-Thujone	Test	7.781	8.090	154.167	0.002	0.420	182.020	0.625	0.000
Camphor	Train	8.643	5.236	162.957	0.005	0.415	183.293	0.000	0.000
Borneol	Validation	9.48	5.668	159.141	0.009	0.464	195.469	-0.427	0.271
Menthol	Train	9.653	8.602	172.677	0.006	0.464	183.293	-0.360	0.294
4-Terpineol	Train	9.689	10.689	237.081	0.005	0.412	182.879	-0.008	1.788
Carvotanacetone	Train	11.826	10.239	184.104	0.013	0.448	174.249	-0.188	0.902
trans-Geraniol	Train	12.191	12.579	147.714	0.003	0.470	209.462	0.137	2.813
Bornyl acetate	Train	12.861	8.067	327.556	0.018	0.415	203.146	-0.221	1.340
Camphene	Train	13.123	10.367	165.594	0.006	0.418	172.795	0.601	0.000
Thymol	Validation	14.111	16.656	307.388	0.005	0.424	188.613	0.100	1.493
Carvacrol	Validation	14.4	22.007	238.010	0.002	0.438	182.020	-0.049	1.474
γ-Caryophyllene	Test	16.531	17.268	163.887	0.002	0.470	182.020	0.148	2.359
trans-Caryophyllene	Validation	16.942	14.779	246.801	0.005	0.436	182.879	-0.140	1.314
Aromadendrene	Train	17.53	16.119	228.290	0.001	0.414	182.020	-0.220	0.662
α-Humulene	Train	18.009	28.311	239.717	0.006	0.409	175.651	-0.139	1.407
Ledene	Train	19.24	16.906	163.887	0.002	0.470	182.020	0.230	3.675

TABLE I. Continued

Compound	ANN cycle	RT / min	RT_{pred} / min	VABC	AATSC0c	ASP-1	AATS0v	MATS6m	GATS7s
Caryophyllene oxide	Test	21.927	22.854	177.950	0.008	0.464	163.653	0.248	0.337
Viridiflorol	Train	22.386	21.829	150.350	0.002	0.470	194.598	0.057	1.889
Ledol	Train	22.586	23.297	343.334	0.005	0.497	162.834	0.009	0.704
Humulene oxide	Train	22.72	25.609	159.141	0.009	0.464	195.469	0.292	0.379
Epirosmanol	Train	34.12	22.007	238.010	0.002	0.438	182.020	-0.049	1.474
Phytol	Validation	35.324	25.067	185.034	0.009	0.516	172.795	0.182	1.151
Ferruginol	Test	40.274	28.311	239.717	0.006	0.409	175.651	-0.139	1.407

the PaDel base. The population of the first generation was selected randomly. Each gene gained value of 1 if its corresponding descriptor was included in the subset; otherwise it gained zero value. The number of the elements was kept relatively low to maintain a small subset of descriptors.[38] As a result, the probability of generating zero for a gene was set at least 60 % greater than the probability of generating the value of 1. The used operators were crossover and mutation. The probability of application of these operators was varied linearly with generation renewal (0.5 % for mutation and 90 % for crossover). A population size of 100 individuals was chosen for GA, and evolution was allowed over 50 generations. The evolution of the generations was stopped when 90 % of the generations took the same fitness. As a results, the six most significant molecular descriptors selected by GA were: van der Waals volume descriptor (VABC – which was calculated using the method proposed by Zhao, Abraham and Zissimos[39]), 2D autocorrelation descriptors (AATSC0c – average centered Broto–Moreau autocorrelation – lag 0 / weighted by charges, AATS0v – average Broto–Moreau autocorrelation – lag 0 / weighted by van der Waals volumes, MATS6m – Moran autocorrelation – lag 6 / weighted by mass, GATS7s – Geary autocorrelation – lag 7 / weighted by I-state)[38] and Chi path (ASP-1 – average simple path, order 1).[40–42]

Detailed explanations about the descriptors were found in the handbook of molecular descriptors.[38] These descriptors encode different aspects of the molecular structure and were applied to develop a QSRR model. Table II represents the correlation matrix among these descriptors.

TABLE II. The correlation coefficient matrix for the selected descriptors by GA; *p*-value – the level of marginal significance within a statistical hypothesis test

	AATSC0c	ASP-1	AATS0v	MATS6m	GATS7s
VABC	0.201	–0.197	–0.100	–0.240	0.085
	$p = 0.269$	$p = 0.279$	$p = 0.587$	$p = 0.186$	$p = 0.646$
AATSC0c		–0.114	0.029	–0.156	–0.223
		p=0.536	$p = 0.876$	$p = 0.393$	$p = 0.220$
ASP-1			–0.057	0.155	0.217
			$p = 0.757$	$p = 0.397$	$p = 0.232$
AATS0v				–0.197	0.244
				$p = 0.281$	$p = 0.178$
MATS6m					0.048
					$p = 0.794$

The calibration and predictive capability of a QSRR model should be tested through model validation. The most widely used squared correlation coefficient (r^2) can provide a reliable indication of the fitness of the model, thus, it was employed to validate the calibration capability of a QSRR model.

Artificial neural network (ANN)

In Fig. 1 the structure of three-layer ANN model is presented, showing a model of ANN (6-11-1) structure, with 6 variables (GATS7s, VABC, AATSC0c, MATS6m, AATS0v and ASP-1) in the input layer, 11 nodes in the hidden layer and 1 node (RT) in the output layer.

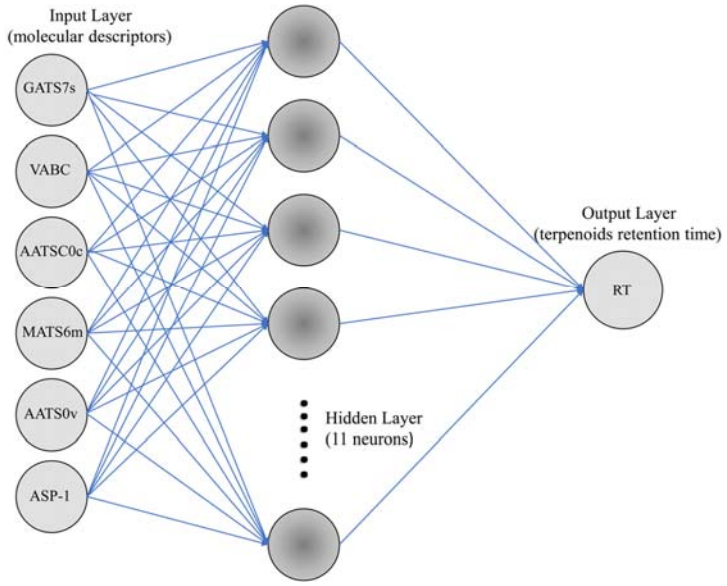

Fig. 1. The structure of a implemented three-layer ANN model with 6 nodes in input layer, 11 nodes in the hidden layer and 1 node in the output layer.

In order to explore the nonlinear relationship between RTs and the selected descriptors, ANN technique was used to build models. The ability to generalize the model was evaluated by an external test set. The statistical results of the artificial neural network MLP 6-11-1 (with 6 inputs, 11 hidden neurons and 1 output neuron) are shown in Tables III and IV and the statistical tests for the predicted *RT*s values for all the EO compounds were given in Table IV.

TABLE III. ANN model summary (performance and errors), for training, testing and validation cycles; performance term represent the coefficients of determination, while error terms indicate a lack of data for the ANN model. Net. name – network name, Train., Test., Valid. – training, testing and validation cycle of the ANN, respectively

Net. name	Performance			Error			Train.	Error	Hidden	Output
	Train.	Test.	Valid.	Train.	Test.	Valid.	algorithm	function	activation	activation
MLP 6-11-1	0.837	0.932	0.712	0.011	0.002	0.096	BFGS 31	SOS	Exponential	Logistic

TABLE IV. The "goodness of fit" tests for the developed ANN model

χ^2	*RMSE*	*MBE*	*MPE*
54.400	7.107	1.239	55.534

The quality of the model fit was tested in Table IV, with the lower χ^2, *MBE*, *RMSE* and *MPE* values showing the better fit to the experimental results.[43]

The predicted *RT*s are presented in Table I and Fig. 2 and confirm the good quality of the constructed ANN, by showing the relationship between the predicted and experimental retention values.

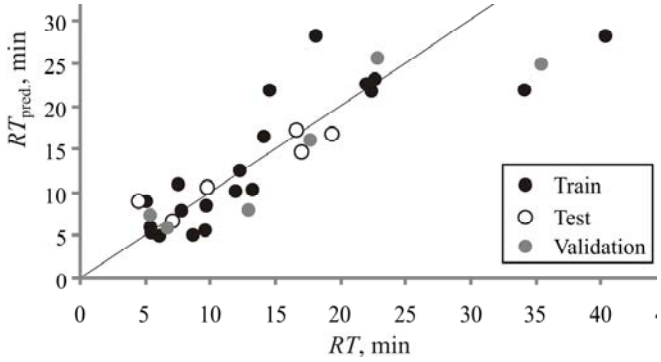

Fig. 2. Comparison of experimentally obtained *RT*s with ANN predicted values.

Obtained results reveal the reliability of the ANN models for predicting the *RT*s of EO compounds in sage herbal dust extracts obtained by SFE.

Molecular descriptors

Separation of compounds in GC and their retention indices are linked to affinity towards mobile and stationary phase. Affinity and solubility of separated molecules directly depend on their chemical structure and physicochemical properties, which could be expressed by molecular descriptors.

Six molecular descriptors we utilized for predictions of *RT* in obtained ANN model. VABC is a volume molecular descriptor which is a subgroup of geometrical descriptors and represents the volume of the area within the van der Waals molecular surface.[38] According to Zhao *et al.*[39] VABC is defined by atomic contributions and the number of atoms, bonds and rings. As it is aforementioned, heterogeneity in the molecular structure of sage terpenoids linked to the number of atoms, bonds and cyclic rings led to variations of VABC and its rather complex correlation with *RT* (Table I).

Moreau–Broto are spatial autocorrelation descriptors,[44] which could be weighted with charges (AATSC0c) and van der Waals volumes (AATS0v). These descriptors are determined by molecular structure and physico-chemical features of atoms.[45] If molecular descriptors used for QSRR analysis have high

correlation the result could be overestimated. This could be surpassed by centering molecular descriptors which leads to un-correlation and separation of different properties influence.[45] The results suggested the lack of correlation between 2D autocorrelation descriptors and VABC (Table II).

3D autocorrelation spatial molecular descriptors are defined by interatomic distances obtained within the geometry matrix which is, therefore, determined by the set of atomic characteristics.[38] Moran autocorrelation coefficients are general indices of spatial autocorrelation determined by weighted atomic property, number of atoms and topological distance between them and Kronecker delta value.[46] Moran coefficients utilized in ANN model were weighted by mass (MATS6m) (Table I). Similarly, the Geary autocorrelation coefficient is determined by the same factors influencing Morton coefficient.[47] The Geary autocorrelation index was weighted by I-state (GATS7s) and further used in obtained ANN model. According to Pearson's correlation coefficients, there was a rather poor correlation between all 3D autocorrelation descriptors (Table II). Hence, utilized molecular descriptors were appropriate to predict RT of sage terpenoids by multivariate ANN model.[48]

Chi path belongs to the group of connectivity indices which are numerical possibilities of two identical molecules encountering each other and is obtained from the bond accessibilities.[42] Chi path index used for calculation was average simple path order 1 (ASP-1).

Global sensitivity analysis – Yoon's interpretation method

In this section the influence of six most important input variables, identified using genetic algorithm on RT was studied. According to Fig. 3, 2D autocorrelation descriptor AATS0v was the most influential parameter with approximately relative importance of 25.1 %, while the influence of VABC and AATSC0c were 18.4 and 14.0 %, respectively. Moreau-Broto coefficient, weighted according to mass (MATS6m) and Geary autocorrelation (GATS7s) were influential at

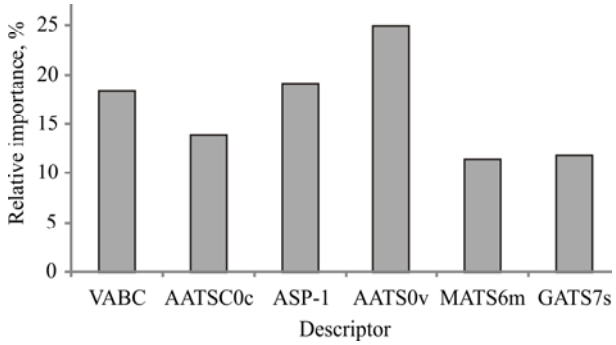

Fig. 3. The relative importance of the molecular descriptors on *RT*, determined using Yoon interpretation method.

levels 11.4 and 11.9 %, respectively, while the influence of ASP-1 reached the level of 19.2 %.

CONCLUSION

Detected terpenoids in sage herbal dust essential oil obtained by supercritical fluid extraction (SFE) belonging to the group of monoterpenes, sesquiterpenes and diterpenes which were either hydrocarbons or oxygenated were used for QSRR analysis. The following six molecular descriptors were suggested by genetic algorithm: VABC, AATSC0c, AATS0v, MATS6m, GATS7s and ASP-1 that characterize retention times of terpenoids. Selected molecular descriptors were not autocorrelated which was suggested by correlation coefficient matrix, thus descriptors were suitable for QSRR analysis. These descriptors were utilized as inputs for the artificial neural network model (ANN), for estimating the retention time (*RT*) using a set of GC–MS data from a series of 32 essential oil compounds found in sage herbal dust extracts obtained by SFE.

The results demonstrated that the ANN model was adequate in predicting the *RT*s of the terpenoid compounds in sage herbal dust extracts obtained by SFE. The coefficient of determination for training cycle was 0.837, which is a good indication that this model could be used as a fast mathematical tool for prediction of retention time values for essential oil compounds in sage herbal dust extracts obtained by supercritical fluid extraction due to low prediction error and moderately high r^2. A suitable model with high statistical quality and low prediction errors was derived, and it could be further used to estimate *RT* of newly detected compounds.

SUPPLEMENTARY MATERIAL

Variations in molecular structure size are available electronically at the pages of journal website: http://www.shd-pub.org.rs/index.php/JSCS, or from the corresponding author on request.

Acknowledgment. This research was conducted within the framework of project TR 31013 funded by the Ministry of education, science and technological development, Republic of Serbia.

ИЗВОД

ПРЕДВИЂАЊЕ РЕТЕНЦИОНОГ ВРЕМЕНА НА GC–MS ЗА ТЕРПЕНЕ ДЕТЕКТОВАНЕ У ЕТАРСКОМ УЉУ ЖАЛФИЈЕ (*Salvia officinalis* L.) КОРИШЋЕЊЕМ QSRR ПРИСТУПА

БРАНИМИР ПАВЛИЋ[1], НЕМАЊА ТЕСЛИЋ[2], ПРЕДРАГ КОЈИЋ[1] и ЛАТО ПЕЗО[3]

[1]*Технолошки факултет, Универзитет у Новом Саду, Булевар цара Лазара 1, 21000 Нови Сад,* [2]*Научни институт за прехрамбене технологије у Новом Саду, Универзитет у Новом Саду, Булевар цара Лазара 1, 21000 Нови Сад и* [3]*Институт за општу и физичку хемију, Универзитет у Београду, Студентски трг 12/V, 11000 Београд*

Циљ овог рада био је да се одреди модел за предвиђање ретенционог времена терпена изолованих из биљног праха жалфије коришћењем екстракције суперкритичним флуидом. Укупно 32 експериментално добијена ретенциона времена терпена која су одређена и детектована GC–MS техником коришћена су за израду модела предвиђања

ретенционог времена. Приступ квантитативног одређивања односа структуре и ретенционог времена коришћен је за предвиђање ретенционог времена једињења из етарских уља, која су идентификована GC–MS анализом, коришћењем шест молекуларних дескриптора, који су одређени коришћењем генетског алгоритма. Изабрани дескриптори су коришћени као улази у вештачку неуронску мрежу за формирање модела за предвиђање ретенционог времена једињења из етарског уља жалфије. Коефицијент детерминације (r^2) био је 0,837, што указује на то да би се овај модел могао користити за предвиђање вредности ретенционог времена терпена у екстрактима и етарском уљу жалфије добијеним коришћењем SFE због високе r^2 вредности и мале грешке предвиђања. Резултати су показали да је 2D аутокорелациони дескриптор AATS0v био најутицајнији дескриптор са релативном важношћу од 25,1 %.

(Примљено 22. маја, ревидирано 26. јула, прихваћено 10. септембра 2019)

REFERENCES

1. G. Miliauskas, P. R. Venskutonis, T. A. Van Beek, *Food Chem*. **85** (2004) 231 (https://doi.org/10.1016/j.foodchem.2003.05.007)
2. B. Pavlić, N. Teslić, A. Vidaković, S. Vidović, A. Velićanski, A. Versari, R. Radosavljević, Z. Zeković, *Ind. Crops Prod.* **107** (2017) 81 (https://doi.org/10.1016/j.indcrop.2017.05.031)
3. S. Glisic, J. Ivanovic, M. Ristic, D. Skala, *J. Supercrit. Fluids* **52** (2010) 62 (https://doi.org/10.1016/j.supflu.2009.11.009)
4. S. A. Aleksovski, H. Sovova, *J. Supercrit. Fluids* **40** (2007) 239 (https://doi.org/10.1016/j.supflu.2006.07.006)
5. M. Mohammadhosseini, S. D. Sarker, A. Akbarzadeh, 2017. *J. Ethnopharmacol.* **199** (2017) 257 (https://doi.org/10.1016/j.chemolab.2015.02.009)
6. J. Acevedo-Martínez, J. C. Escalona-Arranz, A. Villar-Rojas, F. Téllez-Palmero, R. Pérez-Rosés, L. González, R. Carrasco-Velar, *J. Chromatogr. A* **1102** (2006) 238 (https://doi.org/10.1016/j.chroma.2005.10.019)
7. S. J. Marrero-Ponce, Y. Barigye, M. E. Jorge-Rodríguez, T. Tran-Thi-Thu, *Chem. Pap*. **72** (2018) 57 (https://doi.org/10.1007/s11696-017-0257-x)
8. E. Kováts, *Helv. Chim. Acta.* **41** (1958) 1915 (https://doi.org/10.1002/hlca.19580410703)
9. M. Jalali-Heravi, H. Ebrahimi-Najafabadi, *J. Sep. Sci*. **34** (2011) 1538 (https://doi.org/10.1002/jssc.201100042)
10. A. Hinchliffe, *Molecular Modelling for Beginners.* John Wiley & Sons Ltd, London, 2003 (https://coulomb.umontpellier.fr/perso/lucyna.firlej/MasterPro/MMFB.pdf)
11. J. Akbar, S. Iqbal, F. Batool, A. Karim, K. W. Chan, *Int. J. Mol. Sci*. **13** (2012) 15387 (https://doi.org/10.3390/ijms131115387)
12. J. Dai, L. Jin, S. Yao, L. Wang, *Chemosphere* **42** (2001) 899 (https://doi.org/10.1016/S0045-6535(00)00178-8)
13. P. P. Dong, G. B. Ge, Y. Y. Zhang, C. Z. Ai, G. H. Li, L. L. Zhu, H. W. Luan, X. B. Liu, L. Yang, *J. Chromatogr. A* **1216** (2009) 7055 (https://doi.org/10.1016/j.chroma.2009.08.079)
14. K. Héberger, *J. Chromatogr. A* **1158** (2007) 273 (https://doi.org/10.1016/j.chroma.2007.03.108)
15. R. Kaliszan, T. Bączek, A. Buciński, B. Buszewski, M. Sztupecka, *J. Sep. Sci*. **26** (2003) 271 (https://doi.org/10.1002/jssc.200390033)
16. S. Khodadoust, *J. Chromatogr. Sep. Tech.* (2013) (https://doi.org/10.4172/2157-7064.1000149)

17. H. Noorizadeh, M. Noorizadeh, A. S. Mumtaz, *J. Saudi Chem. Soc.* **18(3)** (2014) 183 (https://doi.org/10.1016/j.jscs.2011.06.007)
18. X. J. Yao, A. Panaye, J. P. Doucet, R. S. Zhang, H. F. Chen, M. C. Liu, Z. D. Hu, B. T. Fan, *J. Chem. Inf. Comput. Sci*. **44** (2004) 1257 (https://doi.org/10.1021/ci049965i)
19. J.-L. Wolfender, G. Marti, A. Thomas, S. Bertrand, *J Chromatogr. A* **1382** (2015) 136 (https://doi.org/10.1016/j.chroma.2014.10.091)
20. R. I. J. Amos, E. Tyteca, M. Talebi, P. R. Haddad, R. Szucs, J. W. Dolan, C. A. Pohl, *J. Chem. Inf. Model.* **57** (2017) 2754 (https://doi.org/10.1021/acs.jcim.7b00346)
21. B. Pavlić, O. Bera, N. Teslić, S. Vidović, G. Parpinello, Z. Zeković, *Ind. Crops Prod.* **120** (2018) 305 (https://doi.org/10.1016/j.indcrop.2018.04.044)
22. J. Dong, D. S. Cao, H. Y. Miao, S. Liu, B. C. Deng, Y. H. Yun, N. N. Wang, A. P. Lu, W. B. Zeng, A. F. Chen, *J. Cheminform*. **7** (2015) 60 (https://doi.org/10.1186/s13321-015-0109-z)
23. J. Dong, Z. J. Yao, M. Wen, M. F. Zhu, N. N. Wang, H. Y. Miao, A. P. Lu, W. B. Zeng, D. S. Cao, *J. Cheminform*. **8** (2016) 34 (https://doi.org/10.1186/s13321-016-0146-2)
24. H. Moriwaki, Y. S. Tian, N. Kawashita, T. Takagi, *J. Cheminf.* **10** (2018) 4 (https://doi.org/10.1186/s13321-018-0258-y)
25. C. W. Yap, *J. Comput. Chem.* **32** (2011) 1466 (https://doi.org/10.1002/jcc.21707)
26. Heuristic Lab, https://dev.heuristiclab.com/trac.fcgi/ (last accessed: 10 January 2019)
27. D. E. Goldberg, *Genetic algorithms in search, optimisation and machine learning,* Addison-Wesley, Longman, Boston, MA, 1989 (ISBN:0201157675)
28. R. Leardi, R. Boggiaand, M. Terrile, *J. Chemom*. **6** (1992) 267 (https://doi.org/10.1002/cem.1180060506)
29. Statistica, v. 10. StatSoft, Inc., Tulsa, OK, 2010 (http://www.statsoft.com)
30. X. Hu, Q. Weng, *Remote Sens. Environ.* **113** (2009) 2089 (https://doi.org/10.1016/j.rse.2009.05.014)
31. D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, J. Xu, *Chemometr. Intell. Lab. Syst*. **143** (2015) 7 (https://doi.org/10.1016/j.chemolab.2015.02.009)
32. P. Kojic, R. Omorjan, *Chem. Eng. Res. Des.* **125** (2018) 398 (https://doi.org/10.1016/j.cherd.2017.07.029)
33. Y. Yoon, G. Swales, T. M. Margavio, *J. Oper. Res. Soc*. **44** (1993) 51 (https://doi.org/10.1057/jors.1993.6)
34. M. Nekoei, M. Salimi, M. Dolatabadi, M. Mohammadhosseini, *Monatsh. Chem.* **142** (2011) 943 (https://doi.org/10.1007/s00706-011-0510-x)
35. M. Nekoei, M. Mohammadhosseini, E. Pourbasheer, *Med. Chem. Res.* **24** (2015) 3037 (https://doi.org/10.1007/s00044-015-1354-4)
36. S. Ahmad, M. M. Gromiha, *J. Comput. Chem*. **24** (2013) 1313 (https://doi.org/10.1002/jcc.540030212)
37. J. Aires-de-Sousa, M. C. Hemmer, J. Gasteiger, *Anal. Chem*. **74** (2002) 80 (https://doi.org/10.1021/ac010737m)
38. R. Todeschini, V. Consonni, *Molecular descriptors for chemoinformatics*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009, p. 27 (ISBN: 978-3-527-31852-0)
39. Y. H. Zhao, M. H. Abraham, A. M. Zissimos, *J. Org. Chem*. **68** (2003) 368 (https://pubs.acs.org/doi/10.1021/jo034808o)
40. L. B. Kier, L. H. Hall, *Molecular connectivity in chemistry and drug research*, Academic Press, New York, 1976, p. 1214. (https://doi.org/10.1002/jps.2600660852)
41. L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **35** (1995) 1039 (https://pubs.acs.org/doi/pdf/10.1021/ci00028a014)

42. L. B. Kier, L. H. Hall, *J. Chem. Inf. Comput. Sci.* **40** (2000) 792 (https://doi.org/10.1021/ci990135s)
43. M. Arsenović, L. Pezo, S. Stanković, Z. Radojević, *Appl. Clay Sci.* **115** (2015) 108 (https://doi.org/10.1016/j.clay.2015.07.030)
44. G. Moreau, P. Broto, *Nouv. J. Chim.* **4** (1980) 359
45. B. Hollas, *MATCH* **45** (2002) 27 (http://match.pmf.kg.ac.rs/electronic_versions/Match45/match45_27-33.pdf)
46. P. A. Moran, *Biometrika* **37** (1950) 17 (https://doi.org/10.1093/biomet/37.1-2.17)
47. R. C. Geary, *Inc. Stat.* **5** (1954) 115 (https://www.jstor.org/stable/pdf/2986645.pdf)
48. P. A. Azar, M. Nekoei, S. Riahi, M. R. Ganjali, K. Zare, *J. Serb. Chem. Soc.* **76** (2011) 891 (https://doi.org/10.2298/JSC100219076A).