Scientific paper

Quantitative Relationships Between Structure and Lipophilicity of Naturally Occurring Polyphenols

Vesna Rastija,[1,]* Sonja Nikolić[2] and Vijay H. Masand[3]

[1] *Faculty of Agriculture, University J. J. Strossmayer, Faculty of Agriculture, P. Svačića 1d, Osijek, 31 000, Croatia*

[2] *The Rugjer Boskovic Institute, P.O. Box 180, Zagreb, 10 002, Croatia*

[3] *Department of Chemistry, Vidya Bharati College, Camp, Amravati, Maharashtra, 444 602, India*

* *Corresponding author: E-mail: vrastija@pfos.hr*

Received: 13-05-2013

Abstract

The lipophilicity of polyphenols inherent in food, beverages, and medicinal plants was modelled by using 3D descriptors derived from optimized 3D molecular structures in combination with 2D descriptors. The training sets were generated by manual selection or by cluster formation, and statistically robust predictive models were obtained in both cases. The most relevant structural features for the lipophilicity of polyphenols are depicted by the statistically most significant variables: the number of donor atoms for the H bonds is unfavorable for lipophilicity, and the enhanced number of ring secondary C atom (sp^3) also decreases lipophilicity, while the increased atomic polarizability implies higher lipophilicity of polyphenols. The study also revealed the importance of a three-dimensional distribution of atomic electronegativity for the lipophilicity of molecules.

Keywords: Lipophilicity; polyphenols; quantitative structure-property relationships

1. Introduction

Specific groups of food and medicinal plants are rich sources of one or more subclasses of polyphenols.[1] These compounds have been reported to possess multiple biological activities including vasodilatatory, anti-inflammatory, anti-carcinogenic, anti-bacterial, and antioxidant effects.[2–6] However, the health effects of polyphenols also depend on the amount consumed and their bioavailability.[1] Bioavailability in humans differs greatly from one polyphenol to another. Gallic acid and the isoflavones are readily absorbed, followed by the catechins, the flavanones, and the quercetin glucosides, while the least well absorbed are the proanthocyanidins and the anthocyanins.[7] The oral bioavailability of pharmacologically active substances depends on their hydrophilicity-lipophilicity balance. Strong hydrophilicity of a compound implies good water solubility and good dissolution in gastrointestinal fluid. However, lipophilic compounds have the ability to diffuse passively through biological barriers owing to the lipoid nature of the cell membranes.[8] Besides, the quantitative structure-activity relationship (QSAR) studies have revealed that the lipophilicity of polyphenols is an important factor regarding their activity in biological systems.[9] Our recent QSAR studies have indicated that the antioxidant activity of flavonoids and the vasodilatory effect of phenolic acids are strongly related to lipophilicity.[3,10] The application of quantitative structure-property relationship (QSPR) approaches in developing models to predict the physicochemical properties of polyphenols using topological indices has been reported previously.[11] However, it is well know that many physical, chemical, or biological properties of compounds depend on the three-dimensional arrangement of atoms in a molecule.[12,13] Here, in view of the importance of the three-dimensional shape of molecules for passive diffusion through biological membranes, our attempt was to develop QSPR models that relate the experimentally determined lipophilicity of polyphenols to descriptors derived from optimized 3D molecular structures in combination with 2D descriptors. Due to the importance of rigid model validation, the dataset must be divided into a training set and a test set. Therefore, the goal of this study has also been to determine which of the two splitting methods (manual selection and cluster analysis) give better results.

2. Results and discussion

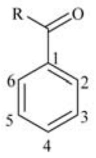
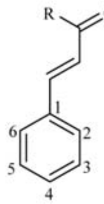
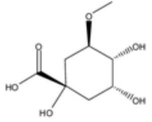
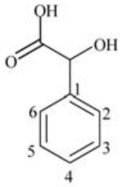
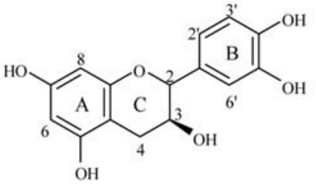
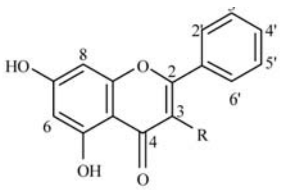
Data set I

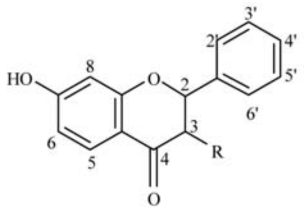
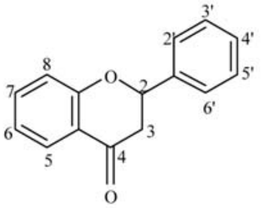
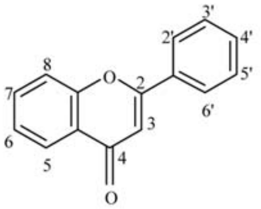
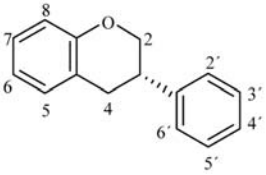
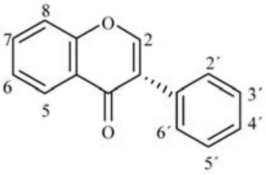
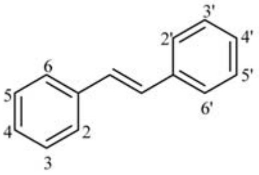
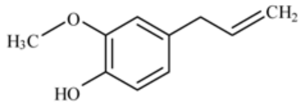
After the classification of 51 compounds into 11 structurally different classes of polyphenols, 12 compounds (Table 1) were selected for the test set manually, and the rest of the initial data set for the training set (n = 39, or 76% of the full data size). The best QSAR models with two and three descriptors (I = 2 and 3), generated by using MLR and the best-subset method, are given in Table 2.

Data set II

We have classified the initial dataset into clusters using the Tree Clustering method (Fig. 1) performed on the set of values of 580 selected descriptors. Members of the test set (n = 12) have been chosen from each cluster. The

Table 1. Chemical structure of polyphenols used in the current study

Class of polyphenol	Comp.no.	Supstituents*	Name of polyphenol	log $P_{Exp.}$
	**1**	3-OCH_3; 4-OH; R = OH	Vanillic acid	1.43
	2	3,5-OCH_3; 4-OH; R = OH	Syringic acid	1.04
	3	3,4,5-OH; R = OH	Gallic acid	0.70
	4	2-OH; R = OH	Salicylic acid	2.26
	5	4-OH; R = OH	*p*-Hydroxybenzoic acid	1.58
	6	3-OH; R = OH	*m*-Hydroxybenzoic acid	1.50
	7	3,4-OH; R = OH	Protocatehuic acid	0.86
	8	2,5-OH; R = OH	Gentisic acid	1.74
	9	2,5-OH; R = H	2,5-Dihydroxybenzaldehide	0.54
	10	3-OCH_3; R = OH	3-Methoxybenzoic acid	2.02
	11	3,4-OCH_3; R = OH	Veratric acid	1.61
	12	2,3-OH; R = OH	Pyrocatechuic acid	1.20
	13	3,4,5-OH; R = OCH_2CH_3	Ethyl gallate	1.30
	**14**	2-OH; R = OH	*o*-Coumaric acid	1.59
	15	4-OH; R = OH	*p*-Coumaric acid	1.79
	16	3-OCH_3; 4-OH; R = OH	Ferulic acid	1.51
	17	3,4-OH; R = OH	Caffeic acid	1.15
	18	3,4-OH; R = 	Chlorogenic acid	0.30
	19	*trans*-4-OH; R = OH	*trans*-*p*-Coumaric acid	1.46
	20	4-OCH_3; R = OH	4-Methoxycinnamate	2.68
	**21**		Mandelic acid	0.62
 HO 8 O 2 2' 3' OH B OH 6' A C 3 OH 6 4 OH	**22**		Catechin	0.51
 HO 8 O 2 2' 4' 5' 6' 3 R 6 4 OH O	**23**	3´,4´-OH; R = OH	Quercetin	1.82
	24	2´,4´-OH; R = OH	Morin	1.84
	25	4´-OH; R = OH	Kaempferol	3.11
	26	3´,4´-OH; R = SU1	Quercetin-3-*O*-glucoside	0.76
	27	3´,4´-OH; R = SU2	Rutin	-0.64

Class of polyphenol	Comp.no.	Supstituents*	Name of polyphenol	log $P_{Exp.}$
	**28**	5,3´-OH; 4´-OCH$_3$	Hesperetin	2.60
	29	3´,4´-OH	Fustin	0.87
	30	5,3´,4´-OH	Taxifolin	0.95
	**31**		Flavanone	3.14
	32	5,7,4´-OH	Naringenin	2.60
	33	5,7,3´,4´-OH	Eriodictyol	2.27
	34	5,4´-OH, 7-SU2	Naringin	-0.44
	35	4´-OH	4'-Hydroxyflavanone	3.20
	**36**	5,3´-OH; 4´-OCH$_3$, 7-SU2	Diosmin	0.14
	37	5,7,3´-OH; 4´-OCH$_3$,	Diosmetin	3.10
	38	5,7,4´-OH	Apigenin	2.92
	39	5,7,3´,4´-OH	Luteolin	2.53
	40	5,7-OH	Chrysin	3.52
	41		Flavone	3.56
	42	5-OH	5-Hydroxyflavone	4.30
	43	7-OH	7-Hydroxyflavone	3.62
	44	5,3´,4´-OH	5,3´,4´-Trihydroxyflavone	3.31
	**45**	7,4´-OH	Equol	3.20
	**46**	7,4´-OH	Daidzein	2.51
	47	5,7,4´-OH	Genistein	3.04
	48	7,4´-OH, 6-OCH$_3$	Glycitein	1.97
	49	5, 4´-OH, 7-SU1	Genistin	0.97
	**50**	3,5, 4´-OH	Resveratrol	3.32
	**51**		Eugenol	2.27

[*] In these compounds, the substituent groups corresponding to the SUgar moieties have been abbreviated as SU suffixed with a number as: SU1 = O-*β*-D-glucopyranosyl; SU2 = O-(6-deoxy-*á*-L-mannopyranosyl)-*β*-D-glucopyranosyl.

best QSAR models with two and three descriptors (I = 2 and 3), generated by using MLR and the best-subset method, are given in Table 2. The correlation matrix, obtained from the initial data set, given in Table 3, shows that the descriptors included in Eqs. 1–4 are independent. A scatter plot of log P_{exp} versus log P_{pred} values calculated by Eq. 4, for the studied polyphenols, is shown in Fig. 2. In order to investigate the applicability of the prediction model 4 and to detect the possible outliers, leverage of the training set was plotted against the residuals (Fig. 3). As it can be seen from the plot, there are no compounds outside the domain of applicability of the model, since their leverage values are not greater than the warning leverage (h^* = 0.308). Moreover, none of the analysed compounds were

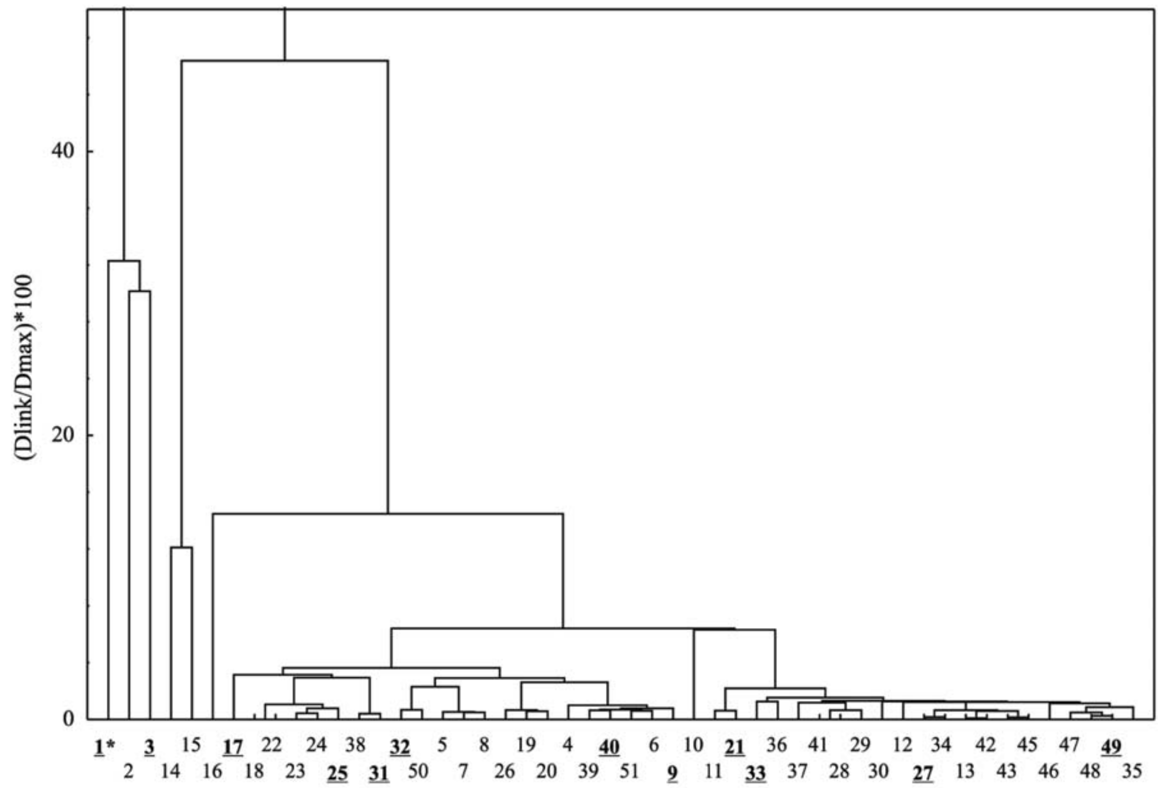


Fig. 1. Dendrogram of a cluster formation of 51 polyphenols

considered as outliers because their standardized residuals were not greater than ± 2.

The *F*-test reflects the ratio of variance explained by the model and variance that is due to an error in the model. The high values of *F*-test indicate that the model is statistically significant. The values of Fisher ratio (*F*) for QSPR models (Eqs. 1–4) ranging from 65.147 to 76.811 (Table 2) suggest that all QSPR models are statistically significant at the 95% level. However, the main disadvantage of this statistical parameter is that it is highly sensitive to the number of descriptors in an equation. An increase in the number descriptors from 2 to 3 in equations causes a reduction of *F* value. Better indicators for the statistical significance of QSPR model are the squared correlation coefficient (r^2) and the standard deviation (s). The closer the value of r^2 to the unity and the smaller the value of s, the better the QSPR models.[14,15] Although all obtained models have a high value of regression coefficients and a low standard deviation, better results are obtained when using models with three descriptors (Eqs. 2 and 4).

The stability of models is proved by the close values of R^2_{LOO} and $R^2_{\mathrm{L\text{-}10\text{-}O}}$. The robustness of the developed models was checked by using the Y-randomization technique. After five randomizations, the resulting models had a significantly lower R^2 than the original model (Table 4). This proves that neither of the developed models is a result of chance correlation. The values of experimental and predicted log *P* values obtained by using Eqs. 2 and 4 are given in Table 5.

Leonard and Roy[16] claim that the quality of the developed models depends considerably on the algorithm used for the selection of the training and test sets. They have performed a validation of QSAR models for the data sets generated by three different methods of division and the best results were obtained when the training test and the test set were selected by the *K*-means cluster. However, our best model was obtained by Eq. 2, where the training and test sets were selected by manual division, and the model obtained by Eq. 4, where the data set was divided by the tree clustering method. Both models have comparable results, such as R^2 value of 0.848 and 0.850 for models obtained by Eqs. 2 and 4, respectively. Since the difference in the parameters of quality is rather small in both models, we cannot decide about the better method for the division of data set into the training set and the test set.

The four best models include: zero-dimensional descriptors (constitutional), one-dimensional descriptors (functional groups counts), two-dimensional descriptors (information and 2D autocorrelations descriptors), and three-dimensional descriptors (RDF and GETAWAY descriptors). All descriptors have been generated from optimised three-dimensional structures of molecules.

The most relevant structural features for the lipophilicity in Eqs. 2 and 4 are depicted by the statistically most

Table 2. The best models obtained

	I	Equations	*R*	R^2	R^2_{adj}	S_{fit}	*F*	R^2_{LOO}	R^2_{L-10-O}	R_{test}	R^2_{test}	S_{test}
Data set I	2	(1) $\log P_{exp} = 2.931 + 2.537\ MATS4e - 0.121\ RDF030e$	0.888	0.788	0.776	0.515	66.947	0.757	0.763	0.875	0.765	0.586
n (training) = 39	3	(2) $\log P_{exp} = 11.048 - 0.301\ nCrs + 1.573$ $MATS4e - 5.924\ IC0$	0.921	0.848	0.835	0.436	65.147	0.817	0.810	0.904	0.817	0.517
n (test) = 12												
Data set II	2	(3) $\log P_{exp} = -4.919 + 26.689\ (\pm 2.897)\ Mp - 1.178$ $(\pm 0.217)\ HATSe$	0.900	0.810	0.800	0.465	76.811	0.782	0.793	0.306	0.093	1.289
n (training) = 39												
n (test) = 12	3	(4) $\log P_{exp} = -5.754 - 0.317\ nHDon + 1.413\ MATS4p + 13.096\ Mp$	0.922	0.850	0.837	0.414	66.251	0.821	0.828	0.894	0.799	0.605

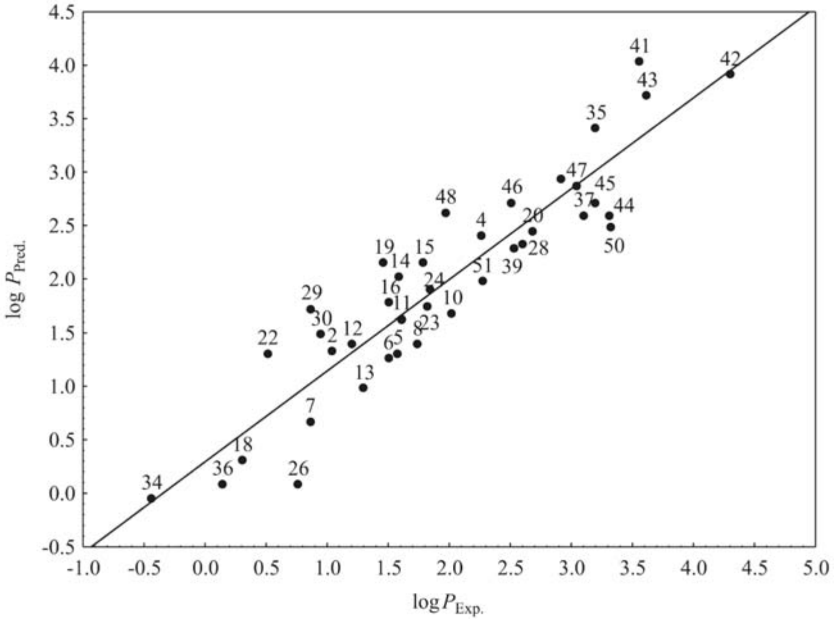


Fig. 2. Plot of predicted lipophilicity ($\log P_{Pred.}$) for the training set (39 molecules) against the observed values ($\log P_{Exp.}$) according to Eq. 4

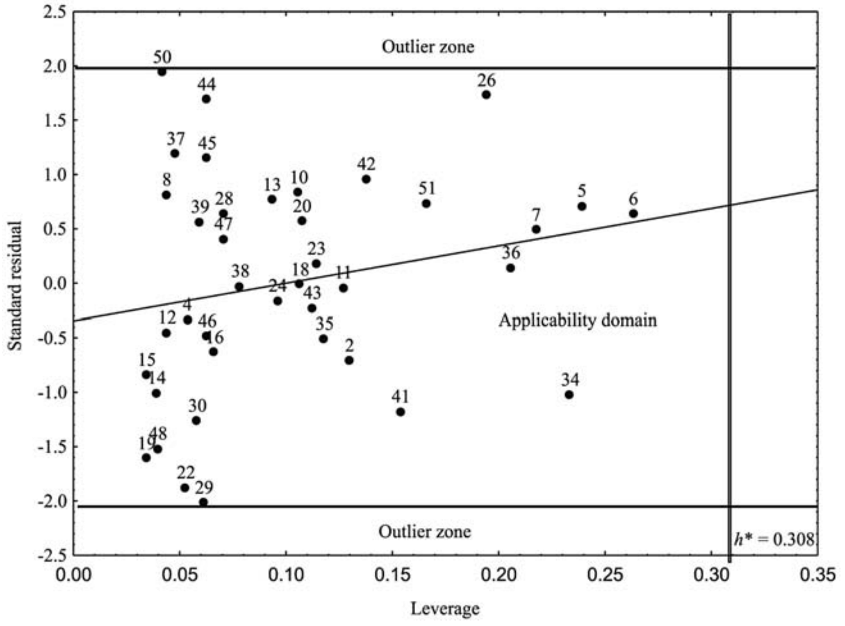


Fig. 3. Applicability domain of the QSAR model for $\log IC_{50}$ expresses by Eq. 4

Table 3. Correlation coefficient among descriptors included in Eqs. 1–4 and $\log P_{exp}$

	MATS4p	*MATS4e*	*HATSe*	*RDF030e*	*Mp*	*nHDon*	*nCrs*	*IC0*	**log** P_{exp}
MATS4p	1.00								
MATS4e	1.00	1.00							
HATSe	0.17	0.17	1.00						
RDF030e	0.37	0.37	0.34	1.00					
Mp	0.58	0.58	–0.03	–0.32	1.00				
nHDon	0.16	0.16	0.30	0.89	–0.47	1.00			
nCrs	0.22	0.22	0.24	0.90	–0.44	0.80	1.00		
IC0	–0.43	–0.43	0.20	0.26	–0.65	0.59	0.23	1.00	
log P_{exp}	0.52	0.52	–0.16	–0.48	0.82	–0.65	–0.57	–0.72	1.00

Table 4. Values of R^2 after the randomization for Eqs. 2 and 4

Y-randomization	R^2 after Y-randomization	
	Model 2	Model 4
1	0.033	0.027
2	0.015	0.057
3	0.099	0.051
4	0.015	0.048
5	0.019	0.037

significant variables (according to the absolute values of standardized regression coefficients β and t-values) – the functional group count: *nCrs* ($\beta = 0.612$; $t = 8.480$) and *nHDon* ($\beta = 0.583$; $t = 6.724$). According to Eqs. 2 and 4, an increased number of ring secondary C atom (sp^3) and an increased number of donor atoms for H bonds decrease the lipophilicity of polyphenols. The functional groups that donate atoms for H bonds, and thus increase the solubility of molecules in water, at the same time decreasing the lipophilicity of polyphenols, are the hydroxyl groups. The influence of the number of hydroxyl groups on lipophilicity, as the donor atoms for H bonds, is obvious in the case of flavanones (**31–35**). The lipophilicity of flavanones increases in the series: 5,7, 3',4'-tetrahydroxyflavanone (**33**) < 5,7,4'-trihydroxyflavanone (**32**) < 4'-hydroxyflavanone (**35**) < flavanone (**31**). The lowest lipophilicity is that of glucoside naringin (log P = –0.44).

The negative impact of an increased number of ring secondary C atoms on lipophilicity is observed in chlorogenic acid (**18**), catechin (**22**), the dihydroflavonols (**28–30**), and the flavonoid glycosides (**26**, **27**, **34**, **36**, **49**). The structures of dihydroflavonols possess a two-ring secondary C atom (C2-C3 double bond is lacking), which with the presence of hydroxyl groups in A and B rings decreases their lipophilicity. The relevance of C2–C3 double bond for increased lipophilicity is evident in the comparison of the lipophile flavonol quercetin (**23**) and the dihydroflavonol taxifolin (**30**), which posses an identical substitution pattern (5,7,3,3',4'-OH). Probably the absence of a C2–C3 double bond in taxifolin produces a significantly lower log P (0.95) than is the case with quercetin (1.82).

The flavonoid glycosides also have a maximum number of ring secondary C atoms present in sugar moieties. Therefore, the glucosides of flavonoides have the lowest log P values such as diosmin (**36**) (log P = 0.14), rutin (**27**) (log P = –0.27), and naringin (**34**) (log P = –0.44). Glycosilation increases the polarity of flavonoid molecules, which is necessary for storage in the cell vacuoles of plants.[17]

The second most significant variable in Eq. 2 is *MATS4e* ($\beta = 0.468$; $t = 6.065$), as in Eq. 4 *MATS4p* ($\beta = 0.398$; $t = 4.238$). Both descriptors belong to the 2D autocorrelation molecular descriptors that describe how a considered property is distributed along a topological molecular structure.[18] *MATS4e* and *MATS4p* correspond to the Moran autocorrelation –lag 4/weighted by atomic Sanderson electronegativities and by atomic polarizability, respectively. It is evident from the sign of regression coefficient that the descriptors *MATS4e* and *MATS4p* have contributed positively to the lipophilicity of polyphenols. It means that, for higher lipophilicity, the polyphenols should have atoms at the topological distance 4 with different electronegativities/polarizabilities as a tendency.

The least relevant in Eq. 2 is the information index *ICO,* which reflects a neighborhood symmetry of order 0. According to the Eq. 2, it is expected that the increasing values of *IC0* (possessing a negative regression coefficient) would tend to predict lower lipophilicity. Mean atomic polarizability (*Mp*) is the last relevant variable in Eq. 4. Its positive coefficient indicates that the lower values of *Mp* imply a lower lipophilicity of polyphenol compounds. Since the oxygen atom has a lower value of polarizability, molecules with more hydroxyl groups tend to have lower lipophilicity.

The three-dimensional descriptors *RDF030e* and *HATSe* are included in Eqs. 1 and 3. Descriptor *RDF030e* belongs to the RDF (Radial Distribution Function) group of descriptors, while *HATSe* belongs to the GETWAY descriptors. Both descriptors offer information about the three-dimensional distribution of electronegativity in molecules. The presence of descriptor *RDF030e* ($\beta = 0.754$; $t = 9.628$) in Eq. 1 suggests the occurrence of some linear dependence between lipophilicity and the 3D molecular distribution of electronegativity, calculated at the radius of 3.0 Å from the geometrical centers of each molecule, while descriptor *HATSe* ($\beta = 0.401$; $t = 5.422$) in Eq. 3 considers the atomic electronegativity of all atoms in molecules. The high absolute values of standardized regression coefficients β and t-values of both descriptors imply a great impact of the three-dimensional arrangement of atoms on the lipophilicity of molecules, especially of atoms with higher electronegativity, such as oxygens.

3. Materials and Methods

3. 1. Data Set

The experimentally determined lipophilicity values (expressed as log P) for 51 polyphenols were collected from the ChemIDplus Advanced database (United States National Library of Medicine, http://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsphttp://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsp) and from literature.[19,20] All experimental values were determined by the same experimental method.

3. 2. Descriptor Calculation

The 3D structures of 51 polyphenols were optimized applying the HyperChem 7.0 (HyperCube, Inc., Gainesville, FL) using the semi-empirical AM1 method.[21] The molecular structures were optimized using Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal mol^{-1}.

Table 5. Experimentally determined values (log $P_{\mathrm{Exp.}}$) and predicted values (log $P_{\mathrm{Pred.}}$) with associated residuals obtained using Eqs. 2 and 4

Comp.no. (model 2)	log $P_{\mathrm{Exp.}}$	log $P_{\mathrm{Pred.}}$	Residual	Comp.no. (model 4)	log $P_{\mathrm{Pred.}}$	Residual
1	1.43	1.10	0.33	***1**	1.16	0.24
***2**	1.04	1.41	–0.37	**2**	1.33	0.29
3	0.70	0.75	–0.05	***3**	0.47	0.23
4	2.26	2.15	0.11	**4**	2.40	–0.14
5	1.58	0.93	0.65	**5**	1.31	0.27
6	1.50	0.88	0.62	**6**	1.26	0.24
7	0.86	0.61	0.25	**7**	0.67	0.19
***8**	1.74	1.40	0.34	**8**	1.39	0.35
9	0.54	1.63	–1.09	***9**	1.93	2.47
10	2.02	1.48	0.54	**10**	1.68	0.34
11	1.61	1.45	0.16	**11**	1.63	–0.02
12	1.20	1.41	–0.21	**12**	1.39	–0.19
13	1.30	1.37	–0.07	**13**	0.98	0.32
14	1.59	2.00	–0.41	**14**	2.02	–0.43
15	1.79	2.15	–0.36	**15**	2.15	–0.36
16	1.51	1.87	–0.36	**16**	1.78	–0.27
***17**	1.15	1.67	–0.52	***17**	1.54	–0.39
18	0.30	0.16	0.14	**18**	0.31	–0.01
19	1.46	2.15	–0.69	**19**	2.15	–0.69
20	2.68	2.45	0.23	**20**	2.44	0.24
***21**	0.62	1.54	0.08	***21**	1.36	–0.74
22	0.51	1.33	–0.82	**22**	1.31	–0.80
23	1.82	2.12	–0.30	**23**	1.75	0.07
***24**	1.84	2.24	–0.40	**24**	1.91	–0.07
25	3.11	2.40	0.71	***25**	2.19	0.95
26	0.76	0.68	0.08	**26**	0.08	0.68
27	–0.64	–0.49	–0.15	***27**	–0.68	–0.39
28	2.60	1.96	0.64	**28**	2.33	0.27
29	0.87	1.41	–0.54	**29**	1.72	–0.85
30	0.95	1.49	–0.54	**30**	1.48	–0.53
31	3.14	3.08	–0.06	***31**	3.77	–0.37
32	2.60	2.31	0.29	***32**	2.67	–0.07
***33**	2.27	1.88	0.39	***33**	2.16	0.11
34	–0.44	–0.77	0.33	**34**	–0.05	–0.39
***35**	3.20	3.09	0.11	**35**	3.41	–0.21
***36**	0.14	–0.13	0.27	**36**	0.84	–0.70
37	3.10	2.55	0.55	**37**	2.59	0.51
38	2.92	2.93	–0.01	**38**	2.93	–0.01
39	2.53	2.52	0.01	**39**	2.29	0.24
40	3.52	3.42	0.10	***40**	3.59	–0.07
41	3.56	3.72	–0.16	**41**	4.03	–0.47
42	4.30	3.62	0.68	**42**	3.91	0.39
***43**	3.62	3.40	0.22	**43**	3.71	–0.09
***44**	3.31	2.53	0.78	**44**	2.59	0.72
***45**	3.20	2.98	0.22	**45**	2.71	0.49
***46**	2.51	2.98	–0.47	**46**	2.71	–0.20
47	3.04	2.86	0.18	**47**	2.87	0.17
48	1.97	2.42	–0.45	**48**	2.62	–0.65
49	0.97	1.09	–0.12	***49**	1.07	–0.10
50	3.32	3.09	0.23	**50**	2.49	0.83
51	2.27	2.83	–0.56	**51**	1.98	0.93

* compound member of the test set

After geometry optimization, several physicochemical parameters were calculated with HyperChem: the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the difference between E_{HOMO} and E_{LUMO} (GAP), the heat of formation (H_{f}), the hydration energy (E_{HYDR}), and the volume (V) of the molecule. The 2D and 3D molecular descriptors used in this study were calculated by applying the online software Parameter Client (Virtual Computational Chemistry Laboratory, http://146.107.217.178/lab/pclient/), an electronic remote version of the Dragon program.[22] 17 groups of Dragon's descriptors were used to generate the QSAR models: constitutional, topological, walk and path counts, connectivity, information, 2D autocorrelations, edge adjacency, BCUT (Burden eigenvalues), topological charge, eigenvalue-based, geometrical, RDF (Radial Distribution Function), 3D-MoRSE (3D-molecular representation of structure based on electron diffraction), WHIM (WeigHted Covariance Matrices), GETAWAY (Geometry, Topology, and Atom Weights AssemblY) descriptors, functional group counts, and molecular properties.[18]

3. 3. Training and Test Set Compounds Selection

The 51 molecules were divided into a training test (n = 39) and a test set (n = 12) in two ways:

1. (Data set I) The 51 molecules were divided into several classes of polyphenols (Table 1). Members of a certain class were selected manually for the test set.
2. (Data set II) The second training set was generated by cluster analysis using the Joining (Tree Clustering) method based on the descriptor values without log P values. As the distance measure we used the Euclidean distance with the Single linkage as a linkage rule. The cluster analysis was performed by using Statistica 7.0 (StatSoft, Inc.; Tulsa, USA).

3. 4. Regression Analysis

The selection of descriptors based on the best-subset method and the multiple regression analysis (MLR) was performed with the use of Statistica 7.0. The number of descriptors (I) in the multiple regression equation was limited to three, in consideration of the fact that the number of compounds in the training set was 39.

The initial number of 1210 calculated molecular indices and physicochemical properties was reduced to 49 descriptors using the following procedure:

1. Descriptors in which values were degenerated, and which weakly correlated to log P ($R \le 0.30$), were eliminated.
2. Further selection of the predictor variables was performed by the best-subset method for the prediction of log P.
3. In order to avoid overfitting, the terminal selection of models was based on the inter-correlation study between the variables included in the equation. Models with highly inter-correlated ($|R| \ge 0.70$) descriptors were discarded. The best models were selected based on the squared correlation coefficient (R^2), adjusted squared correlation coefficient (R^2_{adj}), standard deviation of regression (S_{fit}), and Fisher ration values (F).

If the overall model is significant for the prediction of dependent variables in multiple linear regression, the statistical significance of each independent variable in the mode can be tested separately by the t-test:

$$t = \beta_{\mathrm{j}} / S_{\beta\,\mathrm{j}} \tag{1}$$

where β_{j} is the standardized regression coefficients of the independent variable j and $S_{\beta\,\mathrm{j}}$ is the standard error of β_{j}. The higher t-test values mean that the independent variable is more significant.

In all presented equations, variables are listed according to their statistical signification according to β and t-values.[14,15]

3. 5. Validation of Models

The generated QSPR models were validated by using the classical Leave-One-Out (LOO) cross-validation technique, and also by the Leave-Many-Out (LMO), more precisely the Leave-10-Out (L-10-O) cross-validation procedure. The statistical stability of a model was revealed by the high values of correlation coefficient R^2 Leave-One-Out (R^2_{LOO}) and coefficient R^2 Leave-10-Out ($R^2_{\mathrm{L\text{-}10\text{-}O}}$). Additionally, the Y-randomization technique was applied to validate and check the robustness of MLR equation.[23] Cross-validation and Y-randomization were performed using the data mining software Weka (http://www.cs.waikato.ac.nz/ml/ weka/). Detection of outliers was carried out by investigating the applicability domain of a prediction model.[24]

3. 6. Identifying Outliers

Investigation of the applicability domain of a prediction model was performed by leverage plot (plotting residuals vs. leverage of training compounds). Detection of outliers was carried out for compounds that have values of standardized residuals greater than two standard deviation units. The leverage h^* of a compound is the measure of its influence on the model and is defined as:

$$h^* = 3 \times p' / n \tag{2}$$

where n is the number of training compounds and p' is the number of model-adjustable parameters.

4. Conclusions

Multiple linear regression is used to estimate the lipophilicity of polyphenols present in medicinal plants and food, using descriptors calculated from an optimized three-dimensional molecular structure. In order to determine the best method for dividing the data set into the training and test sets, two methods were performed: manual division and cluster analysis. No great difference was observed in the quality of models from the training and test sets obtained by using these methods. Since the best models that relate to the experimentally obtained log *P* values are generated using two-dimensional descriptors, it is possible that the three-dimensional structure of polyphenols is irrelevant for their lipophilicity. The structural lipophilicity models presented in this study revealed, besides the well known fact that the number of donor atoms for H bonds is unfavourable for lipophilicity, that the enhanced number of ring secondary C atom (sp^3) also decreases lipophilicity, while the increased atomic polarizability implies a higher lipophilicity of polyphenols. The study also confirmed the importance of a three-dimensional arrangement of atomic electronegativity for the lipophilicity of molecules. The above-mentioned descriptors could be used for further QSPR investigation of polyphenols, and the proposed models could potentially provide information about the lipophilicity of other biological active polyphenols, such as the anthocyaninis, the anthocyanidins, and the procyanidins, which are normally commercially unavailable or expensive, and their separation from the plant and food samples and their accurate identification requires techniques with especially high running costs.

5. References

1. C. Manach, A. Scalbert, C. Morand, C. Rémésy and L. Jiménez, *Am. J. Clin. Nutr.* **2004**, *79*, 727–747.
2. X. Han, T. Shen and H. Lou, *Int. J. Mol. Sci.* **2007**, *8*, 950–988.
3. I. Mundic, D. Modun. V. Rastija, I. Brizic, V. Katalinic, M. Medic-Saric and M. Boban, *Food Chem.* **2009,** *119*, 1205–1210.
4. L. H. Yao, Y. M. Jiang, J. Shi, F. A. Tomás-Barberán, N. N. Datta, R. Singanusong and S. S. Chen, *Plant. Food Hum. Nutr.* **2004**, *59*, 113–122.
5. M. Brvar, A. Perdih, V. Hodnik, M. Renko, G. Anderluh, R. Jerala and T. Solmajer, *Bioorg. Med. Chem.* **2012,** *20*, 2572–2580.
6. M. Brvar, A. Perdih, V. Hodnik, M. Renko, G. Anderluh, D. Turk and T. Solmajer, *J. Med. Chem.* **2012,** *55,* 6413–6426.
7. C. Manach, G. Williamson, C. Morand, A. Scalbert and C. Rémésy, *Am. J. Clin. Nutr.* **2005**, *81*, 230S– 242S.
8. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery. Rev*. **2001**, *46*, 3–26.
9. D. Amić, D. Davidović-Amić, D. Bešlo, V. Rastija, B. Lučić and N. Trinajstić, *Curr. Med. Chem.* **2007**, *14*, 827–845.
10. V. Rastija and M. Medić-Šarić, *Eur. J. Med. Chem.* **2009***, 44,* 400–408.
11. V. Rastija, S. Nikolić and M. Medić-Šarić, *J. Math. Chem.* **2009**, *46,* 820–833.
12. N. P. Seeram and M. G. Nair, *J. Agric. Food Chem.* **2002,** *50*, 5308–5312.
13. V. Rastija and M. Medić-Šarić, *Med. Chem. Res.* **2009**, *18*, 579–588.
14. S. Wold, *Quant. Struct.-Act. Relat*. **1991**, *10*, 191–193.
15. R. P. Verma and C. Hansch, *Eur. J. Med. Chem.* **2010**, *45*, 1470–1477.
16. J. T. Leonard and K. Roy, *QSAR Comb. Sci.* **2006**, *25*, 235–251.
17. S. A. Aherne and N. M. O´Brien, *Nutrition*, **2002**, *18*, 75–81.
18. R. Todeschini and V. Consonni, "Handbook of molecular descriptors", M. Manhold, H. Kubinyi and H. Temmerman, Eds., Wiley-VCH, Weinheim, **2000**, p. 667–673.
19. J. A. Rothwell, A. J. Day and M. R. Morgan, *J. Agric. Food Chem*. **2005**, *53*, 4355–4360.
20. Z. N. Xiang and Z. X. Ning, *LWT-Food Sci. Technol.* **2008**, *41*, 1189–1203.
21. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, 3902–3909.
22. I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. Prokopenko, *J. Comput. Aided Mol. Des.* **2005**, *19*, 453–463.
23. A. Tropsha, *Mol. Inf.* **2010**, *29*, 476–488.
24. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

Povzetek

Lipofilnost polifenolov, ki so prisotni v hrani, pijačah in medicinskih rastlinah, smo modelirali z 3D deskriptorji, ki izhajajo iz 3D molekularnih struktur v kombinaciji z 2D deskriptorji. Učni set smo ustvarili z ročno izbiro ali z gručanjem, v obeh primerih smo dobili robustne predikcijske modele. Najpomembnejše strukturne lastnosti, pomembne za lipofilnost polifenolov, so opisane z statistično najznačilnejšimi spremenljivkami: število donorskih atomov H vezi je neugodno za lipofilnost, povečano število obročev sekundarnih C atomov (sp^3) prav tako zniža lipofilnost, medtem ko povečana polarnost atomov nakazuje na povečano lipofilnost polifenolov. Študija je pokazala tudi pomembnost tridimenzionalne razporeditve elektronegativnosti atomov za lipofilnost molekul.