Polish J. Chem., **83**, 2137–2152 (2009)

Application of Molecular Descriptors to the Prediction of Retention in Organic Solvent Nanofiltration

by S. Koter[1*], B. Gilewicz-Łukasik[1], A. Nowaczyk[2] and J. Nowaczyk[1]

[1]*Faculty of Chemistry, Nicolaus Copernicus University, 7 Gagarin St., 87-100 Toruń, Poland*
**e-mail: skoter@chem.uni.torun.pl, tel. +48 56 6114318*
[2]*Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, 9 Sklodowskiej-Curie St., 85-094 Bydgoszcz, Poland*

(Received May 12th, 2009; revised manuscript July 20th, 2009)

An attempt to apply the molecular descriptors for the characterization of retention of solutes in organic solvent nanofiltration has been performed. The descriptors were calculated using the program Dragon. The geometry of each solute molecule has been optimized using Gaussian®. Two linear equations relating the retention coefficient, *R*, with one or two descriptors have been tested using two sets of solutes. The first one ("soft" set) consisted of saturated and aromatic hydrocarbons (data of White, *J. Membr. Sci.*, **205**, 191 (2002)), the second one ("hard" set) contained the substituted aromatic hydrocarbons with heteroatoms (data of Geens *et al.*, *J. Membr. Sci.*, **281**, 139 (2006)). It has been found that the "soft" set of compounds is described reasonably well by both equations. The best descriptors belong to GETAWAY descriptors and Burden eigenvalues. Regarding the "hard" set of compounds only the 2-descriptors equation yields a satisfactory fitting of *R*. Here the 3D-MoRSE descriptors are the best for 7 of 14 membrane-solvent systems.

Key words: nanofiltration, organic solvent, retention, molecular descriptor

In the recent paper [1] Livingston *et al*. raised the problem of finding a standard method for the characterization of organic solvent nanofiltration membranes. Regarding the nanofiltration of aqueous solutions the methods determining the molecular weight cut off (MWCO) do exist. However, as the kind of solvent significantly changes the rejection characteristics of membranes, the problem of finding a method characterizing the membranes in many solvents is much more difficult. To determine MWCO of membranes in the organic solvent nanofiltration Livingston *et al*. [1] have proposed a homologous series of styrene oligomers which are soluble in many organic solvents. They were able to determine precisely MWCO of commercial Starmem^{TM} OSN membranes in different solvents (toluene, ethyl acetate, methanol, hexane). However, it does not mean that we will be able to predict the retention for a given membrane-solvent system when using different solutes. Such a doubt appears when we analyze the data of Geens *et al*. [2]. For example the membrane MPF44 (Koch, USA) rejects in 61% bromomethyl blue when dissolved in methanol and only 9% when in acetone. Generally, the retention of the same set of solutes in different solvents is poorly correlated – see Table 5 where the determination coefficients calculated for the Geens *et al*. retention coefficients of the same set of solutes in various membrane-solvent systems are shown. Therefore it seems to be impossible to fully characterize a membrane using one set of solutes.

As MWCO does not seem to be the best parameter for characterizing a given membrane-solvent system (the molecules of the similar MW may have different shape, functional groups, and thus different interactions with membrane and solvent), in this work the possibility of applying the physicochemical descriptors of solutes for that purpose is presented. Such descriptors are widely applied in the quantitative structure-activity and structure-property relationship studies. A number of quantitative structure-activity relationship (QSAR) studies have been reported in the medical research, which use calculated molecular descriptors in predicting anticancer [3], antiinflammatory [4] and CNS activity [5], as a set of physicochemical, pharmacological or toxicological properties of substance. The quantitative structure-property relationship (QSPR) approach is a very useful tool in prediction of experimental physicochemical properties [6] such as: octanol-water partition coefficient [7], vapor pressure of organic compounds [8], gas-phase reaction rate constants [9].

We have chosen two sources of experimental data on the organic solvent nanofiltration – published by White [10] and by Geens *et al*. [2]. In the first paper the filtration of a toluene solution containing a mixture of saturated and aromatic hydrocarbons (*n*-decane, 1-methyl-naphthalene, *n*-hexadecane, 1-phenylundecane, pristane, *n*-docosane) through a solvent resistant polyimide membrane [11] was investigated (see Fig. 1). In the second work [2] the data on the filtration of single solute solutions through the commercial membranes were published. The following membranes were used: MPF-44 and MPF-50 made from PDMS (Koch, USA), Desal-5-DK from PA (Osmonics, USA), SolSep-169 (SolSep, The Netherlands), HITK-T1 (HITK, Belgium) and FSTi-128 (VITO, Germany). The last two membranes were made from TiO_2. Among many retention data presented in [2] we have chosen those for the solutions of eusulex, 2,2-methylene-(6-tert-butyl-4-methyl-phenol), Victoria Blue, DL-α-tocopherol hydrogen succinate, bromothymol blue and erythrosine B in the solvents: methanol, ethanol, acetone, ethyl acetate, *n*-hexane (see Figs. 2 and 3).

In this paper we will test a linear relationship between the retention coefficients of solutes, *R*, and their descriptors, *d*. Two cases will be considered – *R* as a function of one and two descriptors. The applicability of a given relation will be ascertained on the base of the squared correlation coefficient (determination coefficient), r^2, and the cross-validation coefficient, Q^2_{cv}.

CALCULATION OF DESCRIPTORS

Before the calculation of descriptors, the geometry of each molecule has been optimized at the *ab initio* level of theory using Gaussian® [12] – calculation for isolated systems in vacuum has been conducted using the DFT method with the B3LYP functional and 6-31G basis set. However, one should remember that in different environments (*e.g.* solution) the molecules may have different 3D geometries and thus some of descriptors based on the detailed structure of molecule may not be fully correct. It should be marked that in the most of QSAR/QSPR analysis descriptors of the molecule are calculated from 3D geometry optimized for isolated molecule in vacuum [13].

The descriptors were calculated using the program Dragon ver. 5.5 [14] enabling the calculation of 1664 descriptors divided into 20 groups. All other calculations were performed using the program *Mathematica*® (Wolfram). The short characterization of descriptors and literature on that subject can be found on the internet site of the Virtual Computational Chemistry Laboratory [15]. One of the basic sources is the book by Todeschini and Consonni [16].

Table 1. Descriptors which can be calculated using the program Dragon ver. 5.5 [14].

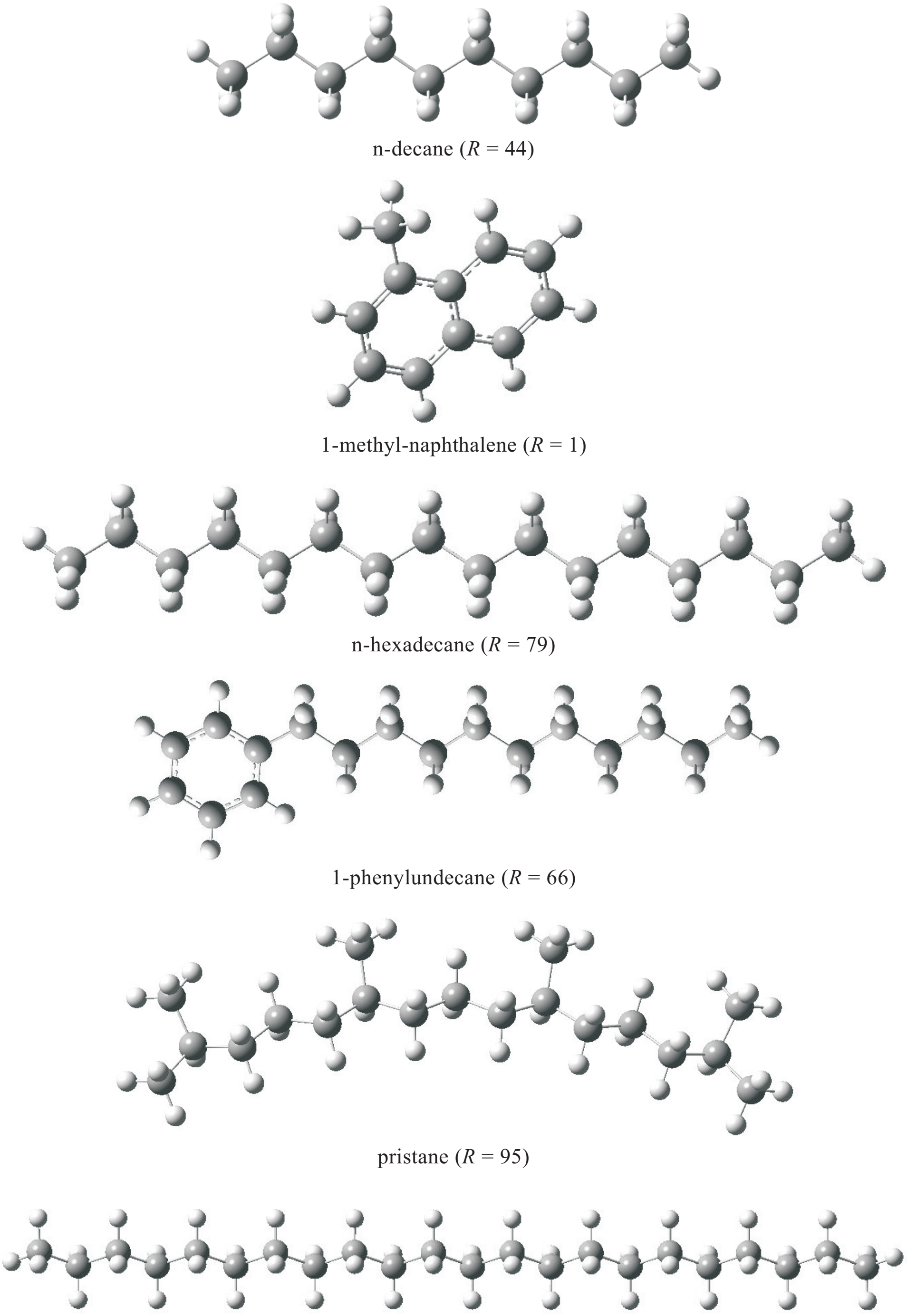
Group	No of descrptors	Our abbreviations
Constitutional descriptors	48	CD
Topological descriptors	119	TD
Walk and path counts	47	WPC
Connectivity indices	33	CI
Information indices	47	II
2D autocorrelations	96	2DA
Edge adjacency indices	107	EAI
Burden eigenvalues	64	BE
Topological charge indices	21	TCI
Eigen-value based indices	44	EI
Randic molecular profiles	41	RMP
Geometrical descriptors	74	GD
RDF descriptors	150	RDF
3D-MoRSE descriptors	160	MD
WHIM descriptors	99	WHIM
GETAWAY descriptors	197	GET
Functional group counts	154	FGC
Atom-centred fragments	120	ACF
Charge descriptors	14	ChD
Molecular properties	29	MP

Not all the descriptors were taken into account. From the calculated set of descriptors for a given set of solutes the following descriptors were removed: (*i*) of the same values for all solutes in the set of solutes, (*ii*) of value exactly zero for any of solutes in the set.

Many of the descriptors are highly correlated with each other, *e.g.* for all the solutes discussed in this paper the descriptor nBO (number of non-H bonds, the same as MWC01, MPC01 and SRW02 divided by 2) shows $r^2 > 0.9999$ with the Randic-type eigenvector-based indices from adjacency matrix (VRA1), distance matrix (VRD1), Z weighted distance matrix (Barysz matrix) (VRZ1), mass weighted distance matrix (VRm1), van der Waals weighted distance matrix (VRv1), electronegativity weighted distance matrix (VRe1) and polarizability weighted distance matrix (VRp1). However, as we also discuss the retention coefficients as a function of two descriptors from the same group, we have not removed highly correlated descriptors from different groups.

RESULTS AND DISCUSSION

Descriptors for the retention data of White. The data of White refer to the filtration of mixture of solutes in toluene [10]. The concentration of each solute was *ca.* 2 wt.%. The structures of compounds optimized at the *ab initio* level of theory are shown in Figure 1. The rejection coefficient and molality of solute are included in the parentheses.



n-decane (R = 44)

1-methyl-naphthalene (R = 1)

n-hexadecane (R = 79)

1-phenylundecane (R = 66)

pristane (R = 95)

n-docosane (R = 92)

Figure 1. Structures of solutes investigated by White [10], optimized using the Gaussian®; in parentheses the rejection coefficient R [%] is given.

One descriptor. Some of descriptors showing the highest correlation with the retention coefficient R are listed in Table 2. It is seen that for the descriptors H4p and RTu from the GET group the value of determination coefficient r^2 exceeds 0.99. It implies that the retention of species in the mixture can be approximated by these descriptors, d, using a linear function:

$$R = a_0 + a_d d \tag{1}$$

To show the significance of equation coefficients a_i, $i = 0, d$, in Table 2 the ratio of the absolute value of a_i to its standard deviation, $|a_i|/s(a_i)$, is also shown. Applying the t-Student test it is seen that $|a_d|/s(a_d)$ is much higher than the value of $t(1\text{-}\alpha/2,f) = 2.78$ for the confidence coefficient $1\text{-}\alpha = 0.95$ and the degree of freedom $f = 4$ [17]. The inequality $|a_d|/s(a_d) > t(1\text{-}\alpha/2,f)$ means that the term $a_d d$ in Eq. (1) is significant. The internal predictability of the model equation (1) is characterized by the cross-validated squared correlation coefficient, Q^2_{cv}, according to which the data in Table 2 are ordered. It is calculated according to the leave-one-out method (LOO) from the formula [18]:

$$Q^2_{cv} = 1 - \sum_{i=1}^{n} (R_i - R_{i,cv})^2 \Big/ \sum_{i=1}^{n} (R_i - \overline{R})^2 \tag{2}$$

where R_i is the experimental value of retention coefficient of solute i, $\overline{R}$ – arithmetic mean of R_i, $i = 1,..,n$. $R_{i,cv}$ is calculated from Eq. (1), where a_0 and a_d are calculated taking the data for n-1 solutes (without i-th solute). All the descriptors which yield Q^2_{cv} higher than arbitrarily assumed value 0.95 are listed in Table 2. For H4p and RTu $Q^2_{\text{cv}} \geq 0.98$. It is seen that for lower value of r^2, Q^2_{cv} can be higher than that for higer r^2 (descriptors BIC0 and RTe).

Table 2. Descriptors of $Q^2_{\text{cv}} > 0.95$ and the parameters of Eq. (1) for White's retention data (Fig. 1).

Group	Descriptor	r^2	Q^2_{cv}	a_0	$\|a_0\|/s(a_0)$	a_d	$\|a_d\|/s(a_d)$
GET	H4p	0.994	0.985	–12.7	4.1	145	27
GET	RTu	0.991	0.980	–82.9	12	5.74	21
II	BIC0	0.988	0.975	343	22	–1635	18
GET	RTe	0.990	0.974	–85.3	11	6.38	20
BE	BELm4	0.982	0.969	–136	10	127	15
BE	BELv4	0.981	0.966	–120	9	124	14
BE	BELp4	0.980	0.963	–114	9	123	14
GET	HIC	0.980	0.956	–173	10	46.1	14

Applying the F-test, it is possible to calculate that for the confidence level 0.95 and the degree of freedom 4 ($F(0.95,4,4) = 6.4$) the lower limit of r^2, which can be still regarded as similar to the highest one, is equal: $r^2_{\text{lo}} = 1 - F \cdot (1 - r^2_{\text{hi}}) = 1\text{-}6.4\cdot(1\text{-}0.994) = 0.962$. There are > 20 descriptors of r^2 higher than that limit. Most of them belong to 2 groups – GET (GETAWAY – GEometry, Topology, and Atom-Weights AssemblY) and BE descriptors (Burden eigenvalues). Because of too many descriptors and the complicated way of their calculation, it is difficult to present any analysis of the relation between the descriptor character and the retention data. Therefore only a short characteristics of the best descriptors is given below.

The descriptors H4p, H4v (not shown in Table 2), RTu and RTe of the GET group are based on the spatial autocorrelation formulas, weighting the molecule atoms by physico-chemical properties together with 3D information encoded by the elements of the molecular influence matrix **H** and influence distance matrix **R**. H4p and H4v are descriptors of H autocorrelation of lag 4 class weighted by atomic polarizabilities and by atomic van der Waals volumes, respectively. RTu and RTe are R total index (**R** is the influence/distance matrix derived from molecular influence matrix **H** defined as $\mathbf{H} = \mathbf{M}\cdot(\mathbf{M}^{T}\cdot\mathbf{M})^{-1}\cdot\mathbf{M}^{T}$, where **M** is the molecular matrix consisting of the centered Cartesian coordinates x, y, z of the molecule atoms) unweighted and weighted by atomic Sanderson electronegativities, respectively. These descriptors, as based on spatial autocorrelation, encode information on structural fragments and their influence on the molecular size and shape as well as for specific atomic properties (in this case atomic van der Waals volumes, atomic polarizabilities and atomic Sanderson electronegativities).

Two descriptors. Approximating the retention coefficient using two descriptors:

$$R_i = a_0 + a_{d1}d_{1,i} + a_{d2}d_{2,i} \qquad (3)$$

only the pairs of descriptors from the same descriptor group have been used. Comparing to Eq. (1) the degree of freedom decreases only by 1, but r^2 is significantly increased (compare Table 3 with Table 2). More than 400 pairs giving $r^2 > 0.995$ have been found. Among them the most numerous were the GET pairs (*ca*. 88%). The second group of pairs was formed by the information indices (*ca*. 5%), whereas the BE pairs were on the 3rd place (2%). In Table 3 the descriptor pairs which yield $Q^2_{cv} > 0.995$ and the pairs of low correlated descriptors ($r^2(d_1,d_2) < 0.1$) giving $Q^2_{cv} > 0.99$ are shown. As the descriptors are not orthogonal, the ratio $|a_{di}|/s(a_{di})$ cannot be used as the significance test for a_{di} [19]. It can be used only to position the descriptors in series. For example for the pair (H6v, H4p) $|a_{di}|/s(a_{di})$ for H4p is much higher than for H6v. Indeed H4p alone correlates with R very well ($r^2 = 0.994$, Table 2), therefore H6v is less significant. Regarding the constraint $r^2(d_1,d_2) < 0.1$ only the GET pairs yield a reasonably high $Q^2_{cv} > 0.99$. In the case

Table 3. The best descriptor pairs fulfilling the condition $Q^2_{cv} > 0.995$ with no constraint on $r^2(d_1,d_2)$ and the condition $Q^2_{cv} > 0.99$ and $r^2(d_1,d_2) < 0$ for the White's retention data (Fig. 1) approximated by Eq. (3).

Group	Descr. 1	Descr. 2	r^2	Q^2_{cv}	$r^2(d_1,d_2)$	$\lvert a_0\rvert/s(a_0)$	$\lvert a_{d1}\rvert/s(a_{d1})$	$\lvert a_{d2}\rvert/s(a_{d2})$
$Q^2_{cv} > 0.995$								
GET	H4v	ITH	0.9996	0.998	0.66	2.0	56	11
	H5u	HTm	0.9995	0.998	0.69	13	57	18
	H3v	R3p+	0.9987	0.997	0.26	3.5	36	9
	R4u	R7p+	0.9988	0.996	0.59	2.1	21	12
BE	BEHe7	BEHm8	0.9993	0.996	0.989	23	20	14
		BEHv8	0.9992	0.996	0.992	23	19	13
GET	H4p	H6v	0.9980	0.996	0.932	6.1	12	2.3
	HATS6p	R5u	0.9981	0.995	0.19	1.8	27	15
$Q^2_{cv} > 0.99$ and $r^2(d_1,d_2) < 0.1$								
GET	HATS4p	R4v	0.996	0.991	0.08	6.6	29	10
	R4v	R4e+	0.996	0.990	0.006	12	3.8	26

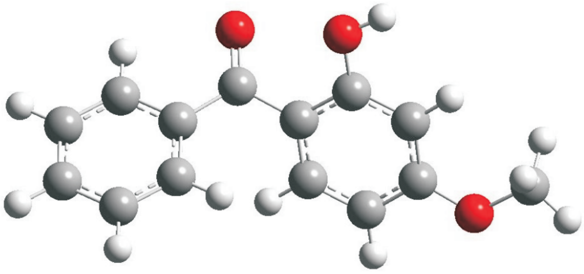
of the information indices the lowest value of $r^2(d_1,d_2)$ was 0.13, found for the pair BIC1-TIC5 which yielded $Q^2_{\text{cv}} = 0.980$. The correlation of BE descriptors in the pairs was very high ($r^2(d_1,d_2) > 0.96$), regarding the pairs fulfilling the condition $r^2 > 0.995$.

A brief description of the GET descriptors, dominant in Table 3, is as follows. They are chemical structure descriptors derived from a new representation of molecular structure. They are divided into two main groups. The first group of the GET descriptors, to which ITH (total information content on the leverage equality) belongs, have been derived by applying some traditional matrix operators and concepts of information theory both to the molecular influence matrix H and the influence/distance matrix R. Most of these descriptors are simply calculated only by the leverages (the leverage is a diagonal element of the influence matrix H which depends on the size and shape of the molecule) used as the atomic weightings. The ITH descriptor mainly encode information on molecular symmetry. The other group of GET descriptors, to which HTm (H total index / weighted by atomic masses), H5u (H autocorrelation of lag 5 / unweighted), Hαv (H autocorrelation of lag α / weighted by atomic van der Waals volumes, $\alpha = 3, 4$) and Rαp+ (R maximal autocorrelation of lag α / weighted by atomic polarizabilities, $\alpha = 3, 7$) belong, is based on the spatial autocorrelation formulas, weighting the molecule atoms by physicochemical properties together with 3D information encoded by the elements of the molecular influence matrix H and influence/distance matrix R. In other words, these descriptors are based on spatial autocorrelation and encode information on structural fragments and therefore can be suitable for describing the differences in congeneric series of molecules. Contrary to the Moreau-Broto autocorrelations, GET descriptors encodes information on the effective position of substituent and fragments in the molecular space. Additionally, as independent of molecule alignment they give some information on the molecule size and shape as well as specific atomic properties.

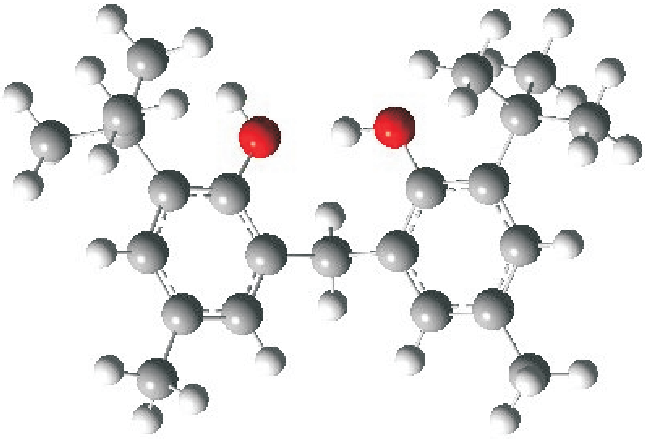
Descriptors for the retention data of Geens *et al*. In the text the names of membranes MPF-44, MPF-50, Desal-5-DK, SolSep-169, HITK-T1, FSTi-128 will be abbreviated to M44, M50, De5, S169, HIT1, F128, respectively. Ethyl acetate will be abbreviated to EA.

Contrary to the White's data (filtration of mixture of solutes) the retention coefficients were determined by Geens *et al*. [2] in the filtration of single solute solutions. According to the authors the concentration polarization does not interfere the results because of low concentration of solutes and the fact that the solvent fluxes observed during the filtration of solutions were almost equal to the pure solvent fluxes.

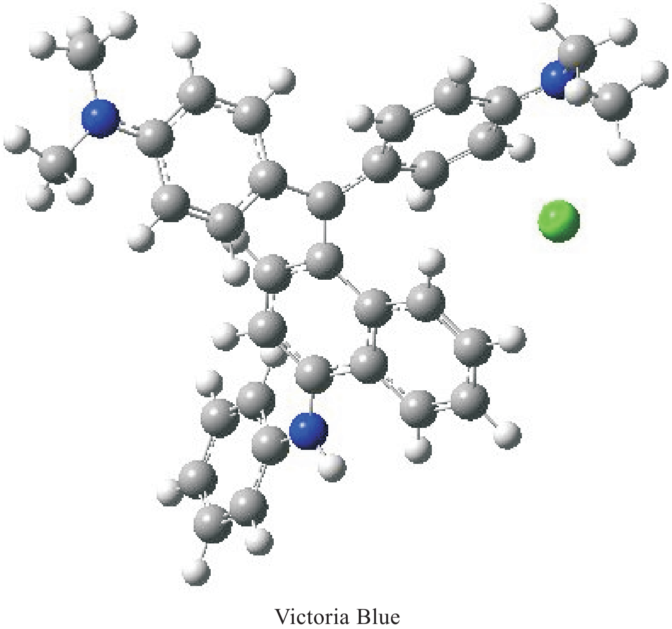
The structure of solutes, optimized by Gaussian, and the retention data for the chosen systems from [2] are shown in Figs. 2 and 3, respectively. Because of the limitations of the Dragon program which does not calculate descriptors for salts and ions, in the case of erythrosine B (EB, genuinely with Na^+) we have calculated EB with H undissociated form, instead of dissociated with the Na^+ counterion. For two systems M50-EtOH and S169-EtOH one of the solutes is not rejected by a membrane ($R = 0$). As many solutes of different properties may show $R = 0$, the results obtained for these systems should be treated cautiously.



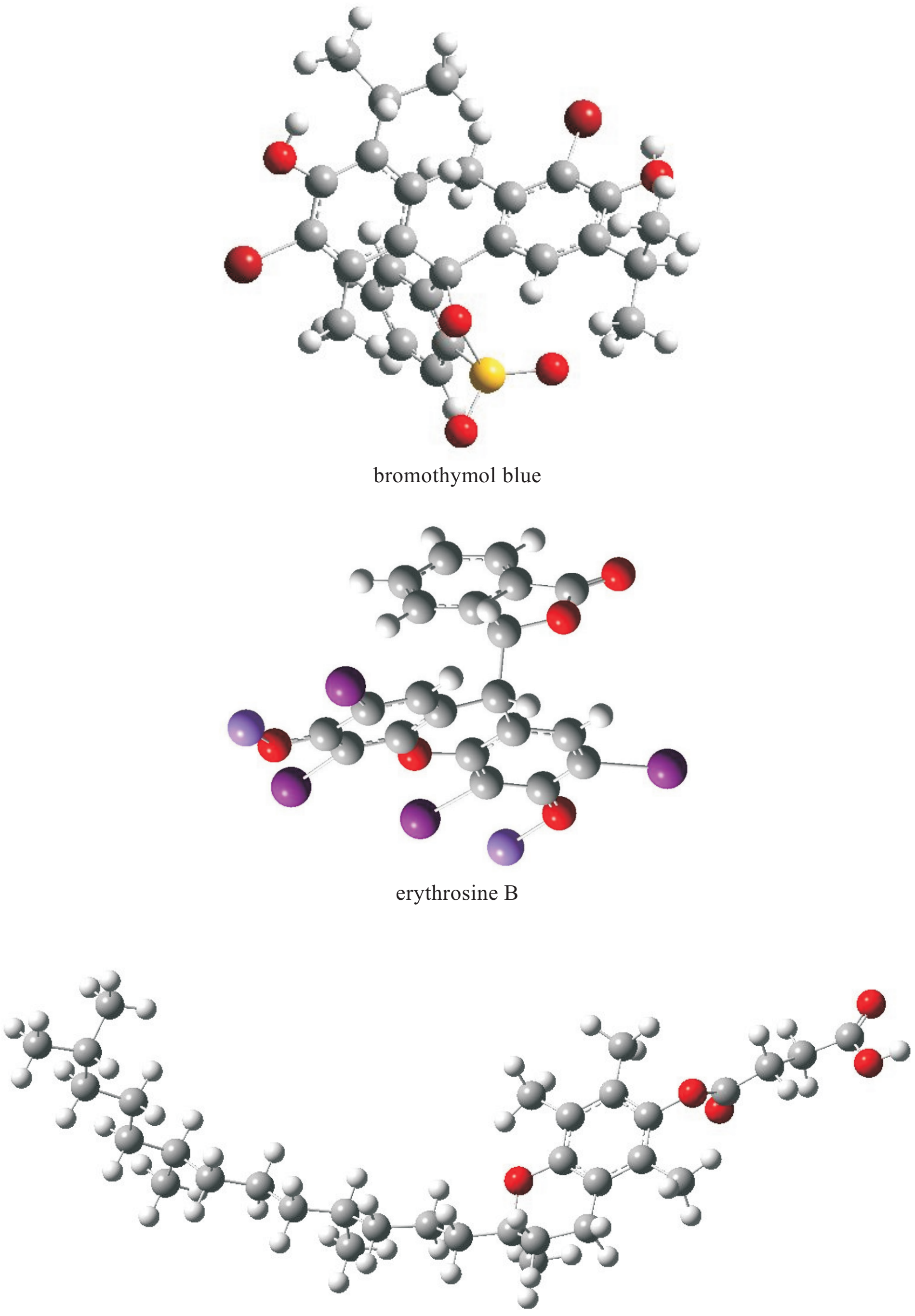
eusulex



2,2-methylene-(6-tert-butyl-4-methyl-phenol)



Victoria Blue



bromothymol blue

erythrosine B

DL- -tocopherol hydrogen succinate

Figure 2. Structures of solutes investigated by Geens *et al*. [2], optimized using the Gaussian®.

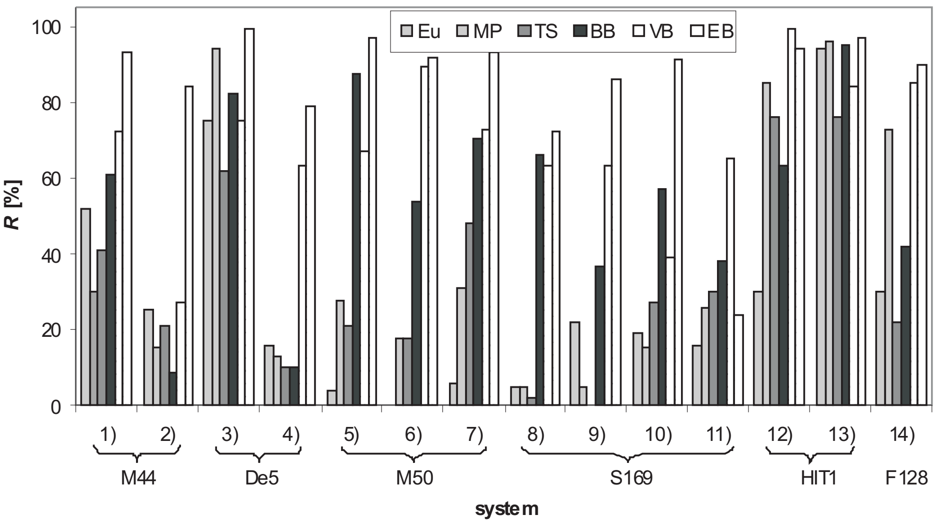


Figure 3. The data of Geens *et al*. [2]: the rejection coefficient of eusulex (Eu), 2,2-methylene-(6-tert-butyl-4-methyl-phenol) (MP), DL-α-tocopherol hydrogen succinate (TS), bromomethyl blue (BB), Victoria Blue (VB) and erythrosine B (EB) observed in the systems 1 – 14 (the numeration of systems is given in Table 4).

One descriptor. The best descriptors for the Geens' data are shown in Table 4. If the descriptors yielding the best values of r^2 and of Q^2_{cv} are different then both descriptors are listed. It is seen that for each membrane-solvent system the best descriptors are different. It is not surprising because the same descriptors should be expected only for those systems, whose retention coefficients, R, strongly correlate with each other. It is seen (Table 5) that there are only a few cases for which the correlation between R is high ($r^2(R) > 0.9$) whereas many of them do not correlate at all ($r^2(R) < 0.1$), particularly the system HIT1-acetone (13). Contrary to the White's data here there is no GET descriptors and for only one system (M50-MeOH) the BE descriptor (BEHp1) is the best one. However, it is different than those given in Table 2.

The correlation between descriptors and the Geens retention coefficients is not particularly good. Only for 3 systems – M50-MeOH, S169-EA, HIT1-acetone – $r^2 > 0.98$. Here the best descriptors are BEHp1, L2e (D/Dr06), and ICR, respectively. Surprisingly for M50-EtOH and S169-MeOH there is no better descriptor than nCIC – the number of rings. Taking into account the limit $Q^2_{\mathrm{cv}} > 0.95$ it is seen that only one system – M50-MeOH – fulfills that requirement (descriptor BEHp1). The lowest values of r^2 (< 0.9) have been obtained for the systems MPF44-acetone and F128-EA.

Table 4. The best descriptors for the retention data of Geens *et al*. (Fig. 3); in the last column "$r^2 > 0.95$" the number of descriptors fulfilling that condition is presented.

System	Group	Descriptor	r^2	Q^2_{cv}	$r^2 > 0.95$
1) M44 – MeOH	WHIM	E1u	0.945	0.86	0
2) M44 – acetone	MD	Mor21m	0.968	0.77	1
	WHIM	E2m	0.949	0.90	
3) De5 – MeOH	MP	Hy	0.76	0.42	0
4) De5 – EtOH	WPC	PCD	0.925	0.78	0
5) M50 – MeOH	BE	BEHp1	0.988	0.969	4
6) M50 – EtOH	TD	STN	0.961	0.916	2
	CD	nCIC	0.960	0.917	
7) M50 – acetone	MD	Mor05m	0.971	0.947	2
8) S169 – MeOH	CD	nCIC	0.941	0.88	0
9) S169 – EtOH	EI	VRv2=VRp2	0.970	0.935	6
10) S169 – acetone	ACF	C-026	0.973	0.947	3
11) S169 – EA	WHIM	L2e	0.984	0.916	4
	TD	D/Dr06	0.978	0.945	
12) HIT1 – MeOH	MD	Mor23p	0.948	0.83	0
13) HIT1 – acetone	TD	ICR	0.988	0.924	6
	RDF	RDF115p	0.966	0.929	
14) F128 – EA	MD	Mor16p	0.86	0.70	0

Table 5. The squared correlation coefficient, $r^2(R)$, of the Geens *et al*. retention coefficients (Fig. 3); 1) denotes M44-MeOH; grey background indicates the same membrane.

System	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)
2) M44-acetone	0.61												
3) De5-MeOH	0.13	0.28											
4) De5-EtOH	0.75	0.69	0.19										
5) M50-MeOH	0.62	0.24	0.26	0.40									
6) M50-EtOH	0.73	0.34	0.15	0.74	0.81								
7) M50-acetone	0.58	0.30	0.11	0.49	0.87	0.86							
8) S169-MeOH	0.74	0.21	0.16	0.48	0.922	0.86	0.77						
9) S169-EtOH	0.947	0.57	0.23	0.84	0.66	0.82	0.58	0.79					
10) S169-acetone	0.79	0.59	0.26	0.49	0.80	0.63	0.74	0.71	0.71				
11) S169-EA	0.07	0.04	0.06	0.12	0.16	0.37	0.24	0.27	0.12	0.00			
12) HIT1-MeOH	0.11	0.14	0.11	0.38	0.31	0.52	0.53	0.21	0.19	0.15	0.32		
13) HIT1-acetone	0.05	0.07	0.72	0.02	0.10	0.01	0.00	0.08	0.10	0.11	0.17	0.04	
14) F128-EA	0.30	0.31	0.52	0.65	0.36	0.58	0.34	0.35	0.51	0.23	0.15	0.58	0.14

As the best descriptors for different systems are not the same (Table 4), it was interesting to check whether there is any common descriptor showing adequately high r^2 for two systems having a common membrane or solvent. Regarding Eq. (1) it is obvious that such a descriptor can be found only for those pairs of systems for which the retention coefficients are well correlated. Unfortunately, according to the results shown in Table 5 the correlation of retention coefficients of two systems, even with the same solvent or membrane, is rather poor. Therefore it was not possible to find for any pair of systems a descriptor fulfilling the condition $r^2_{\mathrm{min}} > 0.95$, where r^2_{min} denotes the smallest value of r^2 observed for two systems constituting the pair.

Two descriptors. Applying Eq. (3) relating the retention coefficient with two descriptors the correlation improves significantly (compare Table 6 with Table 4). In Table 6 for each system two pairs of descriptors are shown. The first one yields the highest Q^2_{cv}, irrespectively of the correlation between the descriptors ($r^2(d_1,d_2)$), the second one – the highest Q^2_{cv} with the arbitrarily assumed condition $r^2(d_1,d_2) < 0.1$. For the systems 6, 7 and 14 both pairs are the same. It is seen that in each case the determination, r^2, and the cross-validation, Q^2_{cv}, coefficients are high and exceed 0.99. Both descriptors are important – the values of $|a_{d1}|/s(a_{d1})$ and $|a_{d2}|/s(a_{d2})$ are of comparable order. Taking into account the pairs of $r^2(d_1,d_2) < 0.1$ it can be noticed that the dominant descriptors are the MD ones (3D-MoRSE – 3D Molecule Representation of Structures based on Electron diffraction) – they occur 7 times (M44-acetone, De5-MeOH and EtOH, S169-MeOH, EtOH and ethyl acetate, HIT1-MeOH). The GETAWAY descriptors, so numerous in the case of White's data, occur only once (irrespectively of $r^2(d_1,d_2)$ – 3 times). The general formula for the MoRSE descriptor can be found in [20]. The meaning of its name "Mor*s*w" is as follows – the number *s* denotes a measure of scattering angle ($1 \le s \le 32$), the letter w – the weight of atoms (w = e, m, p, u, v; e – electronegativity, m – mass, p – polarizability, u – unweighted, v – van der Waals volume). It can be noticed that the most frequent descriptors are those weighted by mass and by volume. These properties of solute molecule are crucial regarding the rejection properties of a membrane, however alone they are not able to predict them correctly. In 3 cases (De5-MeOH, De5-EtOH and S169-EtOH) the pairs are constituted from the descriptors of the same weighting (mass or volume).

Table 6. The best pairs of descriptors for *R* approximated by Eq. (3) (the retention data of Geens *et al.*); in the last column "$r^2, Q^2_{\mathrm{cv}} > 0.99$" the numbers of pairs fulfilling these conditions are given.

System	Group	d_1	d_2	r^2	Q^2_{cv}	$r^2(d_1,d_2)$	$\frac{\lvert a_0\rvert}{s(a_0)}$	$\frac{\lvert a_{d1}\rvert}{s(a_{d1})}$	$\frac{\lvert a_{d2}\rvert}{s(a_{d2})}$	$r^2, Q^2_{\mathrm{cv}} > 0.99$
1) M44-MeOH	WPC	MWC08	MPC09	0.9999	0.9994	0.79	126	119	184	46, 5
	GD	HOMA	QXXp	0.9997	0.9987	0.03	142	97	47	
2) M44-acetone	GET	HATS4v	HATS5v	0.9991	0.990	0.84	17	45	26	172, 5
	MD	Mor21m	Mor23e	0.998	0.993	0.06	14	35	6.6	
3) De5-MeOH	BE	BELm4	BELm7	0.9976	0.992	0.908	13	29	34	6, 1
	MD	Mor16v	Mor20v	0.993	0.968	0.58	106	20	14	
4) De5-EtOH	GET	H2u	R7u+	0.9986	0.994	0.946	49	47	46	20, 1
	MD	Mor01m	Mor20m	0.991	0.974	0.04	2.6	14	14	
5) M50-MeOH	TD	w	DECC	0.9999	0.9994	0.36	103	156	145	129, 18
	BE	BEHp1	BELm1	0.9996	0.997	0.08	60	9.3	84	
6) M50-EtOH	GD	AROM	QXXe	0.9996	0.997	0.03	83.6	83	24	32, 5
7) M50-acetone	WHIM	L2p	Dm	0.9998	0.9986	0.01	45	79	85	153, 15
8) S169-MeOH	II	Xindex	Yindex	0.9996	0.9986	0.83	78	77	58	40, 7
	MD	Mor13m	Mor25v	0.9991	0.9959	0.008	28	49	30	
9) S169-EtOH	MD	Mor14m	Mor32e	0.9992	0.9963	0.13	59	46	54	49, 5
		Mor25m	Mor26m	0.9945	0.9837	0.0003	8.1	8.4	22	
10) S169-acetone	GET	HATS8e	R1p	0.9995	0.998	0.34	46	23	47	176, 9
		ITH	R7v+	0.9993	0.9975	0.003	22	27	60	

11) S169-EA	MD	Mor05v	Mor31p	0.9997	0.9995	0.21	24	103	73	107, 9
		Mor16u	Mor32p	0.998	0.992	0.08	17	30	34	
12) HIT1-MeOH	RDF	RDF085u	RDF080v	0.997	0.992	0.69	21	10	32	31, 2
	MD	Mor14m	Mor23v	0.996	0.986	0.07	2.4	17	24	
13) HIT1-acetone	EAI	ESpm04x	ESpm01r	0.9995	0.998	0.76	551	26	42	159, 14
	2DA	GATS6v	GATS7v	0.9988	0.996	0.004	19	29	38	
14) F128-EA	WHIM	P2s	E3s	0.9988	0.994	0.10	39	43	38	11, 1

In Table 7 the best descriptors pairs for the case membrane-solvent1-solvent2 are shown. The results are not so satisfactory as for the single systems (Table 6). Although there are many descriptor pairs which yield $r^2_{\mathrm{min}} > 0.95$, only for S169-MeOH-acetone a descriptor pair which gives $Q^2_{\mathrm{cv}} > 0.95$ has been found. For each membrane, except De5 ($r^2_{\mathrm{min}} < 0.9$, De5 is not listed), r^2_{min} is higher than 0.96, the highest values of r^2_{min} (> 0.98) are observed for the membranes M50, S169 and the solvent pairs EtOH-acetone, MeOH-acetone. The reason are high correlation between retention coefficients observed for these systems (Table 5). It is also a reason why for M50 the number of descriptors pair of $r^2_{\mathrm{min}} > 0.95$ is much higher than for other systems. Contrary to the single descriptor case (Eq. (1)) to obtain a high correlation the condition of high $r^2(R)$ is not necessary. The examples are HIT1-MeOH-acetone and S169-acetone-EA for which $r^2(R)$ is close to zero, whereas $r^2_{\mathrm{min}} > 0.96$.

Table 7. The best pairs of descriptors for R approximated by Eq. (3) – one membrane and 2 solvents (the retention data of Geens *et al.*); r^2_{min} and $Q^2_{\mathrm{cv,\,min}}$ are underlined.

Solvent 1	Solvent 2	$r^2(R)$	Group	d_1	d_2	r^2		Q^2_{cv}		$r^2(d_1,d_2)$	r^2_{min} > 0.95
						s.1	s.2	s.1	s.2		
					M44						
MeOH	acetone	0.61	MD	Mor28m	Mor25p	0.993	0.977	0.945	0.925	0.06	19
					HIT1						
MeOH	acetone	0.04	MD	Mor04v	Mor23p	0.965	0.980	0.57	0.76	0.24	1
					M50						
MeOH	EtOH	0.81	RDF	RDF070m	RDF045v	0.984	0.971	0.927	0.853	0.87	24
			MD	MPC07	PCD	0.965	0.984	0.920	0.944	0.37	
MeOH	acetone	0.87	WPC	MPC09	TPC	0.986	0.985	0.921	0.985	0.51	291
			TD	w	CSI	0.986	0.981	0.948	0.934	0.70	
EtOH	acetone	0.86	MD	Mor05m	Mor20p	0.994	0.990	0.971	0.941	0.0002	138
					Mor07v	0.978	0.991	0.946	0.946	0.19	
					S169						
MeOH	EtOH	0.79	EI	LP1	VRv2	0.974	0.986	0.867	0.939	0.28	37
			WHIM	E1e	P1p	0.973	0.973	0.874	0.88	0.19	
MeOH	acetone	0.71	WHIM	E2e	Dm	0.987	0.998	0.964	0.986	0.01	45
MeOH	EA	0.27	TD	PW3	D/Dr06	0.960	0.984	0.86	0.942	0.002	4
EtOH	acetone	0.71	MD	Mor01m	Mor12v	0.984	0.989	0.947	0.963	0.049	51
EtOH	EA	0.12	CD	Mv	nR06	0.999	0.983	0.993	0.86	0.32	4
acetone	EA	0.00	GET	HATS8v	R3m+	0.979	0.989	0.90	0.928	0.997	22

The descriptors are different than those for single systems. Even the groups of descriptors in many cases are also different. The dominant group is MD – it occurs 5 times on the first place, whereas the highest value of $Q^2_{\mathrm{cv,min}}$ = 0.964 has been found for the WHIM group and the combination S169-MeOH-acetone.

The results for the solvent-membrane1-membrane2 system are shown in Table 8. Similarly as for membrane-solvent1-solvent2 (Table 7) there are many descriptor pairs giving r^2_{min} > 0.95, however only for two cases – acetone-M44-HIT1 and acetone-M50-S169 – Q^2_{cv} > 0.95 have been found. Again, it may be noticed that to obtain a high value of Q^2_{cv} the condition of high $r^2(R)$ is not necessary – for the systems M44-acetone and HIT1-acetone $r^2(R) = 0.07$, whereas for M50-acetone and S169-acetone $r^2(R)$ is much higher (= 0.74). In Table 8 more analogous examples can be found. Thus, regarding Eq. (3) with two descriptors there is no correlation between $r^2(R)$ and Q^2_{cv}.

Table 8. The best pairs of descriptors for *R* approximated by Eq. (3) – one solvent and 2 membranes (the retention data of Geens *et al.*); r^2_{min} and $Q^2_{\mathrm{cv,\,min}}$ are underlined.

m.1	m.2	$r^2(R)$	Group	d_1	d_2	r^2		Q^2_{cv}		$r^2(d_1,d_2)$	r^2_{min} > 0.95
						m.1	m.2	m.1	m.2		
MeOH											
M44	M50	0.62	RDF	RDF035v	RDF060p	0.982	0.990	0.933	0.953	0.77	34
	S169	0.74	GET	R2v+	R1p+	0.970	0.984	0.82	0.945	0.991	23
				R2e+	RTv+	0.983	0.967	0.90	0.90	0.54	
De5	M50	0.26	CI	X1A	X5sol	0.955	0.977	0.85	0.87	0.62	3
M50	S169	0.922	CI	X4Av	X4sol	0.985	0.985	0.917	0.954	0.18	214
				X4A	X0A	0.997	0.984	0.987	0.935	0.0003	
	HIT1	0.31	MD	Mor08m	Mor23p	0.961	0.966	0.88	0.86	0.32	1
EtOH											
De5	M50	0.74	RDF	RDF030v	RDF040v	0.989	0.966	0.947	0.87	0.65	7
	S169	0.84	MD	Mor10p	Mor13p	0.979	0.978	0.86	0.89	0.0005	19
			RDF	RDF070u	RDF095v	0.960	0.956	0.918	0.89	0.89	
M50	S169	0.82	GET	HATS5u	R2e	0.994	0.996	0.950	0.975	0.66	310
Acetone											
M44	M50	0.30	MD	Mor05m	Mor21m	0.970	0.973	0.54	0.67	0.34	29
			RDF	RDF075e	RDF050p	0.994	0.955	0.935	0.85	0.38	
	S169	0.59	RDF	RDF065m	RDF085m	0.993	0.983	0.89	0.85	0.53	209
	HIT1	0.07	MD	Mor21m	Mor27m	0.992	0.981	0.962	0.953	0.86	13
M50	S169	0.74	GET	R3p+	H1u	0.990	0.993	0.956	0.968	0.06	460
					HATS5u	0.995	0.990	0.975	0.965	0.14	
	HIT1	0.00	TD	ICR	ZM2V	0.979	0.988	0.910	0.979	0.05	13
					w	0.978	0.988	0.923	0.914	0.40	
S169	HIT1	0.11	MD	Mor13m	Mor04v	0.961	0.968	0.83	0.90	0.08	3

CONCLUSIONS

In this paper the attempt to apply the molecular descriptors for the characterization of retention data in nanofiltration of nonaqueous media has been performed. Two sets of solutes were analyzed. The first "soft" set consisted of saturated and aromatic hydrocarbons (White's data), the second one – "hard" set – contained the substituted aromatic hydrocarbons with heteroatoms (Geens' data). The "hard" set was tested in 14 membrane-solvent systems.

It has been found that for the "soft" set of compounds it was possible to find eight descriptors, *d*, which describe satisfactorily the retention of compounds, *R*, using a linear dependence $R = a_0 + a_d d$ ($r^2 > 0.99, Q_{\mathrm{cv}}^2 > 0.95$). Regarding the "hard" set of compounds only for one membrane-solvent system (MPF50-MeOH) the cross-validation coefficient exceeded 0.95. As the retention coefficients of the set of solutes in different membrane-solvent systems in most cases show poor correlation, the best descriptors for these systems are different. Also for the same membrane and different solvents or vice versa, the correlation of retention coefficients is usually poor. Therefore it cannot be expected that the same descriptor and simple linear relation retention-descriptor can be successful in prediction of retention coefficients.

Applying the linear model with 2 descriptors – $R = a_0 + a_{d1}d_1 + a_{d2}d_2$ – the prediction of retention is significantly improved. For the "soft" set of solutes the best pairs of descriptors yield $r^2 > 0.999$ and $Q_{\mathrm{cv}}^2 > 0.995$. In the case of "hard" set of solutes for each membrane-solvent system a pair of descriptors yielding $Q_{\mathrm{cv}}^2 > 0.99$ has been found. However, such a model is still not satisfactory when applied with the same descriptors to the couples of systems even with common membrane or solvent, although for that model the condition of high correlation between retention coefficients of these systems is not needed (no correlation between $r^2(R)$ and the best values of Q_{cv}^2 has been found). Among the investigated 32 couples of systems (common membrane or solvent) only 3 descriptor pairs have given $Q_{\mathrm{cv}}^2 > 0.95$.

The best approximation of retention of the "soft" set of solutes have been obtained using two groups of descriptors – GETAWAY descriptors and Burden eigenvalues. In the case of the "hard" set of solutes the dominant group of descriptors seems to be the 3D-MoRSE group. No constitutional descriptors, like molecular weight or any other from that group, fulfilling the condition $Q_{\mathrm{cv}}^2 > 0.95$ have been found.

Taking the model $R = a_0 + a_d d$ or $R = a_0 + a_{d1}d_1 + a_{d2}d_2$ and seeking the common descriptors for a couple of systems, it automatically involves the assumption that the change of solvent or membrane will affect only the coefficients of these equations. However, the situation is more complicated and more sophisticated models should be taken into consideration.

Because of diversity of molecular descriptors it is difficult to discuss the relationship between them and the retention coefficients and the physicochemical properties of the filtration systems. Descriptors show a certain, mathematically expressed, feature of molecule but not necessarily they reflect its specific physicochemical property.

The interpretation why this or that descriptor adequately describes a given feature, in our case retention, is a slippery game. To draw reliable conclusions regarding the character of descriptors more various sets of solutes should be taken into account.

Abbreviations: De5 – De5-5-DK, EA – ethyl acetate, F128 – FSTi-128, HIT1 – HITK-T1, M44 – MPF-44, M50 – MPF-50, MWCO – molecular weight cut off, S169 – SolSep-169.
Symbols: *d* – descriptor, Q^2_{cv} – cross-validation coefficient, $Q^2_{\mathrm{cv,\,min}}$ – the smallest value of Q^2_{cv} observed for two systems membrane-solvent taken for comparison, r^2 – determination coefficient between retention coefficient and decriptor(s), r^2_{min} – the smallest value of r^2 observed for two systems membrane-solvent taken for comparison, $r^2(d_1,d_2)$ – determination coefficient between two decriptors, $r^2(R)$ – determination coefficient between retention coefficients of the same set of solutes for two membrane-solvent systems, *R* – retention coefficient.

REFERENCES

1. See Toh Y.H., Loh X.X., Li K., Bismarck A. and Livingston A.G., *J. Membr. Sci.*, **291**, 120 (2007).
2. Geens J., Boussu K., Vandecasteele C. and Van der Bruggen B., *J. Membr. Sci.*, **281**, 139 (2006).
3. Debnath B., Samanta S., Naskar S.K., Roy K. and Jha T., *Bioorg. Med. Chem. Lett.*, **13**, 2837 (2003).
4. Li X., Zhao M., Tang Y.R., Wang C., Zhang Z. and Peng S., *Eur. J. Med. Chem.*, **43**, 8 (2008).
5. Ekins S., Shimada J. and Chang C., *Adv. Drug Delivery Rev.*, **58**, 1409 (2006).
6. Liu K.P., *J. Liq. Chromatogr. Related Technol.*, **31**, 1808 (2008).
7. Persona A., Marczewska B., Fekner Z., Senczyna B., Matysiak J. and Niewiadomy A., *QSAR & Comb. Sci.*, **23**, 319 (2004).
8. Katritzky A.R., Slavov S.H., Dobchev D.A. and Karelson M., *Comput. Chem. Eng.*, **31**, 1123 (2007).
9. Öberg T., *Atmos. Environ.*, **39**, 2189 (2005).
10. White L.S., *J. Membr. Sci.*, **205**, 191 (2002).
11. White L.S., US Patent 6,180,008 (2001).
12. *Gaussian 03, Revision D.01*, Frisch M.J., Trucks G.W., Schlegel H.B., Scuseria G.E., Robb M.A., Cheeseman J.R., Montgomery J.A., Vreven J.T., Kudin K.N., Burant J.C., Millam J.M., Iyengar S.S., Tomasi J., Barone V., Mennucci B., Cossi M., Scalmani G., Rega N., Petersson G.A., Nakatsuji H., Hada M., Ehara M., Toyota K., Fukuda R., Hasegawa J., Ishida M., Nakajima T., Honda Y., Kitao O., Nakai H., Klene M., Li X., Knox J.E., Hratchian H.P., Cross J.B., Bakken V., Adamo C., Jaramillo J., Gomperts R., Stratmann R.E., Yazyev O., Austin A.J., Cammi R., Pomelli C., Ochterski J.W., Ayala P.Y., Morokuma K., Voth G.A., Salvador P., Dannenberg J.J., Zakrzewski V.G., Dapprich S., Daniels A.D., Strain M.C., Farkas O., Malick D.K., Rabuck A.D., Raghavachari K., Foresman J.B., Ortiz J.V., Cui Q., Baboul A.G., Clifford S., Cioslowski J., Stefanov B.B., Liu G., Liashenko A., Piskorz P., Komaromi I., Martin R.L., Fox D.J., Keith T., Al-Laham M.A., Peng C.Y., Nanayakkara A., Challacombe M., Gill P.M.W., Johnson B., Chen W., Wong M.W., Gonzalez C. and Pople J.A., Gaussian, Inc.: Wallingford CT, 2004.
13. Karelson M., Lobanov V.S. and Katritzky A.R., *Chem. Rev.*, **96**, 1027 (1996).
14. Talete srl, DRAGON for Windows (software for molecular decriptor calculations). Version 5.5 – 2007 – http://www.talete.mi.it/
15. http://www.vcclab.org/lab/indexhlp/
16. Todeschini R. and Consonni V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim (Germany), 2000.
17. Neter J., Wasserman W. and Kutner M.H., Applied Linear Statistical Models, IRWIN, Homewood, Illinois, 1985.
18. Mazerski J., The Basics of Chemometrics, Publisher: Gdańsk University of Technology, Gdansk, 2000 (in Polish).
19. Achnazarowa S.L. and Kafarow W.W., Optimalization of Experiment in Chemistry and Chemical Technology, WNT, Warszawa, 1982 (in Polish).
20. Gasteiger J., Schuur J., Selzer P., Steinhauer L. and Steinhauer V., *Fresenius J. Anal. Chem.*, **359**, 50 (1997).