Membrane Water Treatment, Vol. 1, No. 3 (2010) 181-192
DOI: http://dx.doi.org/10.12989/mwt.2010.1.3.181

Prediction of retention of uncharged solutes in nanofiltration by means of molecular descriptors

Alicja Nowaczyk[1], Jacek Nowaczyk[2] and Stanisław Koter[2]*

[1]*Department of Organic Chemistry, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, 9 Sklodowskiej-Curie St., 85-094 Bydgoszcz, Poland*
[2]*Faculty of Chemistry, Nicolaus Copernicus University, 7 Gagarin St., 87-100 Toruñ, Poland*

(Received October 1, 2009, Accepted March 22, 2010)

Abstract. A linear quantitative structure-property relationship (QSPR) model is presented for the prediction of rejection in permeation through membrane. The model was produced by using the multiple linear regression (MLR) technique on the database consisting of retention data of 25 pesticides in 4 different membrane separation experiments. Among the 3224 different physicochemical, topological and structural descriptors that were considered as inputs to the model only 50 were selected using several criteria of elimination. The physical meaning of chosen descriptor is discussed in detail. The accuracy of the proposed MLR models is illustrated using the following evaluation techniques: leave-one-out cross validation procedure, leave-many-out cross validation procedure and Y-randomization.

Keywords: nanofiltration; retention; pesticides; QSPR analysis.

1. Introduction

Nanofiltration (NF) is a very useful membrane technique which enables the removal of solutes of diameter ca. 0.5-5 nm (*e.g.*, textile dyes, pesticides, sugars) from the solutions. The application of NF grows continuously (Scott 1995, Kołtuniewicz and Drioli 2008). Initially it has been mainly applied for the filtration of aqueous solutions (Scott 1995, Kołtuniewicz and Drioli 2008). After the development of new membranes, stable in organic solvents, NF of organic media becomes gradually more attractive. Because of variety of NF applications the problem of finding an appropriate membrane for a given purpose arises. For the characterization of retention properties of membranes usually the term "molecular weight cut off" (MWCO) is used (Koros *et al*. 1996). It works well in the case of ultrafiltration membranes which reject large solutes (diameter ca. 5-100 nm) of high molecular weight (ca. 5-100 kDa) (Trägårdh and Ölund 1986, Kim *et al.* 1994, Tsapiuk 1996). In the case of NF, where the solutes of much lower molecular weight are separated, MWCO fails. The reason is that the differences in molecular weight of solutes are small and the differences in their chemical structure are much more important. Thus, another molecular descriptor(s) of solute should be found for the characterization of NF membranes.

In this work we test a wide range of molecular descriptors to predict the retention properties of NF membranes. The test was conducted using the experimental data of Kiso *et al*. (2000, 2001). These data include the retention data of 25 pesticides differing in the molecular mass from 187.3 (molinate)

* Corresponding author, Professor, E-mail: skoter@chem.uni.torun.pl

to 350.6 (chlorpyrifos) obtained for the commercial membranes. The aim was to find the best correlated relation between rejection coefficient and descriptors, containing from 1 to 5 descriptors. The main purpose of the research is to check whether there exist a molecular descriptor, more universal than MWCO which could be used for characterization of NF membranes.

In order to formulate a model based on the experimental data we decided to use the QSPR methodology. The QSPR approach is a very useful tool in prediction of experimental physicochemical properties such as: octanol-water partition coefficient (Toropov *et al*. 2008, Li *et al*. 2008), chromatographic retention time (Hu *et al*. 2008, Jäntschi *et al*. 2007), aqueous solubility of compounds (Duchowicz *et al*. 2008, Toghiani *et al*. 2008). However, there are relatively few (Kiso *et al*. 2000, 2001, Libotean *et al*. 2008, Koter *et al*. 2009, Yangali-Quintanilla *et al*. 2009) reporting the QSPR analysis applied to prediction of rejection properties of NF membranes.

2. Material and methods

2.1 Experimental data of retention

The retention data of 25 pesticides were taken from the original papers published by Kiso *et al*. (2000, 2001). They determined the retention coefficient in the filtration of aqueous pesticide solutions of concentration 0.5-1.5 mg/L through the commercial membranes (Nitto Denko Co. Ltd., Osaka, Japan). The single solute solutions were used. They used four membranes of NaCl retention ranging from 15 to 92%. Two membranes denoted as Memb-1 and Memb-2 of high NaCl retention (92 and 60%, respectively) were made from poly(vinyl alcohol)/polyamide NTR-729HF. The membranes Memb-3 and Memb-4 of lower NaCl retention (51 and 15%, respectively) were made from sulfonated polyethersulfone. In the case of all membranes except Memb-4 the applied pressure was 1 MPa. The Memb-4 was much more permeable and for that membrane the pressure 0.5 MPa was applied.

2.2 Molecular descriptors and methods

Before the calculation of molecular descriptors the molecular 3D structure has been optimized using the Gaussian® 03 (version 6.1) package (Frisch *et al*. 2004). The three-dimensional structures of pesticides for isolated systems in vacuum have been conducted using the DFT (Density Functional Theory) method with the B3LYP functional (Becker 3-term correlation functional, Lee, Yang and Parr exchange functional) and 6-31G basis set. The force field calculations were used to ascertain whether the resulting geometries were the true energy minima structures. All the molecules were minimized until the root mean square (RMS) gradient value became smaller than 10^{-6} a.u.

The molecular descriptors have been calculated using the Talete srl, DRAGON for Windows Version 5.5-2007 package. The Dragon descriptors include 22 different logical blocks. The total number of calculated descriptors is 3224. Several criteria were used to reduce this number while optimizing the information content of the descriptors set. Firstly, descriptors for which no value was available for all the compounds were disregarded. Secondly, the descriptors showing the same value (or nearly the same) for all pesticides were excluded. For the remaining descriptors, if two descriptors showed a correlation coefficient greater than 0.9, the one showing the highest pair correlation with the other descriptors was removed. After these automatic screening procedures, a set of 541 descriptors was taken for further analysis. To reduce the vast number of descriptors to the 50 best correlated to

Table 1 Descriptors of the QSPR regression reported in this study

Descriptor blocks	PR1	PR2	PR3	PR4
	Descriptor symbol (Todeschini and Consonni 2000)			
Constitutional	MW; Sv;	MW; Sv;	Sv;	Sv;
Topological	MAXDN; SPI; PW4; PW3; PJI2; TI2;	Dz; MAXDP; PW3; D/Dr06;	D/Dr06; T(N..O); T(N..S); HNar;	Hnar; Jhete; T(N..S); T(Cl..Cl); T(N..O); D/Dr06;
Walk and path counts	MWC05;	–	–	–
Connectivity indices	XMOD; X5v	XMOD;	XMOD;	XMOD;
Information indices	IVDE; IDM; Xindex; Uindex	IDM; IVDE; Uindex	CIC1;	CIC1;
2D autocorrelations	ATS5v; ATS2m;	ATS5v; ATS2m;	GATS2e; GATS3m; GATS1m;	GATS2e; GATS1m;
Edge adjacency indices	EEig03d; ESpm03d; ESpm05r;	ESpm05r; EEig03d; ESpm03d;	EEig10d;	EEig10d; EEig03x;
Burden eigenvalues	BEHm1; BELm3; BEHp1; BEHp7; BEHm3; BEHm2;	BEHp7; BELm3; BEHm1; BEHm3;	BEHp7; BEHm3;	BEHm3;
Topological charge indices	GGI1; GGI2; JGI3; JGI1	GGI1; GGI2; SPI;	–	–
Eigenvalue based indices	LP1; VRZ1;	VRZ1;	–	–
Randic molecular profiles	DP11;	DP11;	–	–
Geometrical	RGyr; PJI3;	RGyr;SEig; H8p; QXXm;	G(N..O);	G(N..O); QXXm; G(Cl..Cl);
RDF	RDF030m; RDF055v; RDF010e; RDF045p; RDF065m; RDF040v;	RDF040v; RDF070v; RDF055v; RDF115v; RDF045p; RDF045m;	RDF075v; RDF045m; RDF070v;	RDF075v; RDF070v; RDF045m;
3D-MoRSE	Mor03m; Mor01m	Mor01m; Mor03u	Mor03u; Mor10m; Mor02v; Mor03v; Mor30m; Mor11m;	Mor10m; Mor30m; Mor03v; Mor27m; Mor03u; Mor02v;
WHIM	Dv; HATS5v; L1u;	G2m;	E1m; Vp;	E1m; Vp;
GETAWAY	R5p+; R4p; HATSp;	R5u+; R5p+; H1p;	H4m; H5p; HATSp; RARS; R5p+; H1p;	R5p+; H5p; HATSp; H1p; RARS; H4m;
Atom-centered fragments	–	–	H-050;	H-050;
Molcular properties	AMR;	GVWAI-50; ALOGP; Inflammat-80; AMR; Neoplastic-80; Hypertens-80;	AMR; MLOGP; BLTD48; Hy; TPSA(Tot);	MLOGP; BLTD48; AMR; TPSA(Tot);
2D binary fingerprints	–	B05[C-O]; B08[C-C]; B09[C-C]; B06[C-C];	B02[N-N]; B10[C-C]; B09[C-C]; B02[O-S]; B04[O-Cl]; B03[O-O]; B04[N-N]; B03[Cl-Cl];	B04[N-N]; B02[N-N]; B03[Cl-Cl]; B04[O-Cl]; B09[C-C]; B05[O-O]; B02[O-S];
2D frequency fingerprints	–		F03[Cl-Cl]; F06[C-C]; F09[C-S]; F04[N-O];	F04[N-O]; F03[Cl-Cl]; F04[O-Cl]; F03[N-P];

the experimental data the "Feature Selection and Variable Screening" methods available in Statistica® (version 8.0) software have been applied. Then the chosen descriptors were used as regressors of the model; they are collected in Table 1, their detailed description can be found in the literature (Todeschini and Consonni 2000).

2.3 Statistical analysis

The Multiple Linear Regression (MLR) analysis (Allison 1999) and correlation analysis was carried out using Statistica® (version 8.0) software. The step-wise regression analysis yielded a four models describing retention **(R_i)** as a linear combination of a number of molecular descriptors. The statistical quality of the regression equations was evaluated by parameters such as the correlation coefficient, R, the squared correlation coefficient, R^2, the adjusted squared correlation coefficient, R^2_{adj}, the RMSE (Root Mean Squared Errors) and the variance ratio F. The statistical significance (p-level) of a result was determined as $p \leq 0.01$ (Bland 2000).

The model obtained in this work was validated by calculations of the validated squared correlation coefficient (Q^2) and of the prediction error sum of squares (called S_{PRES}). Q^2 was calculated using the most general internal cross-validation procedures: "leave-one-out" (LOO), "leave-many-out" (LMO) (Baumann 2005, Golbraikh and Tropsha 2002). Abbreviations Q^2_{LOO}, Q^2_{LMO}, (and QS_{LOO}, QS_{LOO}) were used in their usual meaning for the tests listed above, respectively. Additionally the robustness of the proposed model was checked by permutation testing: parallel models were developed based on the fit to randomly reordered Y-data (Y-scrambling, Y-randomization) (Gramatica 2007).

3. Results and discussion

In this work we have limited the number of presented equations to those of the best statistics. The models are given in Table 2 together with the statistical and validation (average values) parameters. For a training set of 25 compounds the Topliss and Costello rule (Topliss and Costello 1972) allows the use of up to 5 descriptors. Therefore the presented QSPR analysis yields models with 5, 4 and 2 descriptors. The fulfilled relation $R^2_{adj} < R^2$ confirms that the models are not overparametrized. The correlation coefficient, R, of these relationships is in the range 0.95-0.99 and explains up to 91-95% of all variance data for rejection of investigated pesticides. Moreover, the F-test value together with RMSE at the p-level of 0.01 means that the equation has a good correlation with the data and is statistically significant. Every descriptor in the regression equation must be independent, consequently the correlation between each descriptor was calculated and is presented in the form of Pearson correlation matrices in Table 3 (A-D). It is seen that the descriptors selected for our models are not highly correlated. Plot of predicted rejection vs. residual values was prepared in order to determine the existence of systemic error in the model development (Fig. 1 A-D). The uniform distribution of residues indicates no systemic error (Belsley *et al*. 2005). The plots of observed rejection vs. predicted by our models are shown in Fig. 1 (E-H). These plots proves that the models developed in this work are adequate for the prediction of the property. Summing up the linear models seems to be an adequate fit to the data, all predictors have $p < 0.01$ and one can conclude that each model is independently associated with the retention coefficient.

In an attempt to determine the utility of our models three validation analyses were carried out *i.e*., LOO, LMO, Y-scrambing. The former (LOO) is one of the most general and frequently used

Table 2 Results of the QSPR analysis

PR1	5.25 (±0.77) EEig03d + 1.15 (±0.25) RDF010e − 17.99 (±2.81) ESpm05r − 5.06 (±1.06) X5v + 0.04 (±0.01) Mor01m + 144.91 (±15.86) R = 0.9860, R^2 = 0.9662, R^2_{adj} = 0.9294, F = 26.24, RMSE = 1.25, Q^2_{LOO} = 0.97, QS_{LOO} = 0.84, Q^2_{LMO} = 0.94, QS_{LMO} = 0.84
PR2	8.78 (±1.57) RDF040v + 53.92 (±9.97) RGyr − 23.34 (±5.44) ATS5v + 5.75 (±1.83) ALOGP − 54.62 (±12.68) R = 0.9688, R^2 = 0.9386, R^2_{adj} = 0.9173, F = 37.11, RMSE = 7.44 Q^2_{LOO} = 0.94, QS_{LOO} = 6.13, Q^2_{LMO} = 0.94, QS_{LMO} = 5.96
PR3	1046.23 (±169.28) R5p+ + 7.25 (±2.40) RDF075v R = 0.9638, R^2 = 0.9290, R^2_{adj} = 0.8979, F = 29.90, RMSE = 11.52 Q^2_{LOO} = 0.93, QS_{LOO} = 9.41, Q^2_{LMO} = 0.93, QS_{LMO} = 9.54
PR4	− 106.81 (±16.40) Mor27m + 35.22 (±9.74) B05[O-O] − 123.22 (±35.24) RARS − 30.96 (±7.49) B02[N-N] + 7.46 (±2.20) RDF075v + 100.91 (±30.36) R = 0.9548, R^2 = 0.9116, R^2_{adj} = 0.8730, F = 23.58, RMSE = 12.85, Q^2_{LOO} = 0.92, QS_{LOO} = 10.49, Q^2_{LMO} = 0.91, QS_{LMO} = 10.85

N = 25 is the number of compounds included in the data set
p = 0.01 the significance of the variables in the model
R the correlation coefficient
R^2 the squared correlation coefficient
R^2_{adj} the adjusted squared correlation coefficient
RMSE the Root Mean Squared Errors
F the variance ratio
Q^2_{LOO} the validated squared correlation coefficient calculated from internal cross-validation procedure "leave-one-out test" (LOO)
Q^2_{LMO} the average values the validated squared correlation coefficient calculated from internal cross-validation procedure "leave-one-out test (LMO) from ten test
QS_{LOO} the prediction error sum of squares (called S_{PRES}) values calculated from LOO test
QS_{LMO} the average values the prediction error sum of squares (called S_{PRES}) values calculated from ten of LMO test
Definitions of descriptors abbreviation are given in the text and they are to extended to be mentioned again here.

technique but it does not ultimately validate the model (Golbraikh and Tropsha 2002). In the field of statistical technique LOO and LMO are used for internal validation. Based on theory R^2 of acceptable model cannot be smaller than Q^2_{LOO} and Q^2_{LMO}. The best model is achieved when $Q^2_{LOO} \leq R^2 \geq Q^2_{LMO}$ and the cross validated squared correlation coefficients are similar: $Q^2_{LOO} \approx Q^2_{LMO}$. Commonly $Q^2_{LOO} > 0.5$ is considered as a proof of the reasonably predictive capability of the equation, the inequality $Q^2_{LOO} > 0.7$ indicates stable and predictive potential of the equation, nevertheless high Q^2_{LOO} value does not indicate high predictive power of the model. On the other hand if $R^2 < Q^2_{LOO}$ the model is overfitted. As it is seen from the statistics presented next to each model (Table 2) $R^2 > Q^2_{LOO}$, which means that our model is not overfitted. The LMO test is usually used to verify the results obtained from the LOO test. The calculations of Q^2_{LMO} were based on 10 random selections of groups of 10 objects from the 25 training observations (*e.g.*, 10fold, 80/20 cross validation). Each group of ten randomly chosen pairs of predicted and observed values was

Table 3A Pearson correlation matrix of the parameters (p-level<0.01) used in this study for model PR1

PR1	EEig03d	RDF010e	ESpm05r	X5v	Mor01m
EEig03d	1	0.16051	0.45262	0.44326	0.58459
RDF010e		1	0.36487	0.38097	0.31450
ESpm05r			1	0.49803	0.79775
X5v				1	0.66517
Mor01m					1

Table 3B Pearson correlation matrix of the parameters (p-level<0.01) used in this study for model PR2

PR2	RDF040v	RGyr	ATS5v	ALOGP
RDF040v	1	0.69117	0.73106	0.54270
RGyr		1	0.88283	0.54438
ATS5v			1	0.51138
ALOGP				1

Table 3C Pearson correlation matrix of the parameters (p-level<0.01) used in this study for model PR3

PR3	R5p+	RDF075v
R5p+	1	-0.34421
RDF075v		1

Table 3D Pearson correlation matrix of the parameters (p-level<0.01) used in this study for model PR4

PR4	Mor27m	B05[O-O]	RARS	B02[N-N]	RDF075v
Mor27m	1	0.28423	-0.09671	-0.22617	0.24774
B05[O-O]		1	0.00797	-0.29277	0.22394
RARS			1	-0.06592	-0.19692
B02[N-N]				1	-0.02061
RDF075v					1

left out and that group was predicted by the model (Afantitis *et al*. 2006). The results of the LMO test are collected in Table 4. For average over all test steps $R^2 > Q^2_{LMO}$ which is another proof that our model is not underdetermined and internally stable due to $Q^2_{LOO} \approx Q^2_{LMO}$. In order to ascertain whether the good results in models PR1-PR4 are not due to a chance correlation or structural dependency of the training set, the Y-scrambling tests were performed. The Y-randomization tests results are shown in the Table 5. The low values of R^2_Y and Q^2_Y proves that our model is valid.

The 5-parametric equation **(PR1)** defines the best model for Memb-1 - the poly(vinyl alcohol)/ polyamide nanofiltration membrane of 92% NaCl rejection. Molecular descriptors incorporated in the equation are: Radial Distribution Function - 1.0/weighted by atomic Sanderson electronegativities (RDF010e), 3D-MoRSE - signal 01/weighted by atomic masses (Mor01m) (Schuur *et al*. 1996), valence connectivity index chi-5 (X5v) (Kier and Hall 1981). Eigenvalue 03 from edge adjacency matrix weighted by dipole moments (EEig03d) and Eigenvalue 05 from edge adjacency matrix weighted by resonance integrals (ESpm05r). The last two descriptors pertain to a group called edge adjacency indices (Estrada *et al*. 1998). The remaining descriptors belong to different logical block

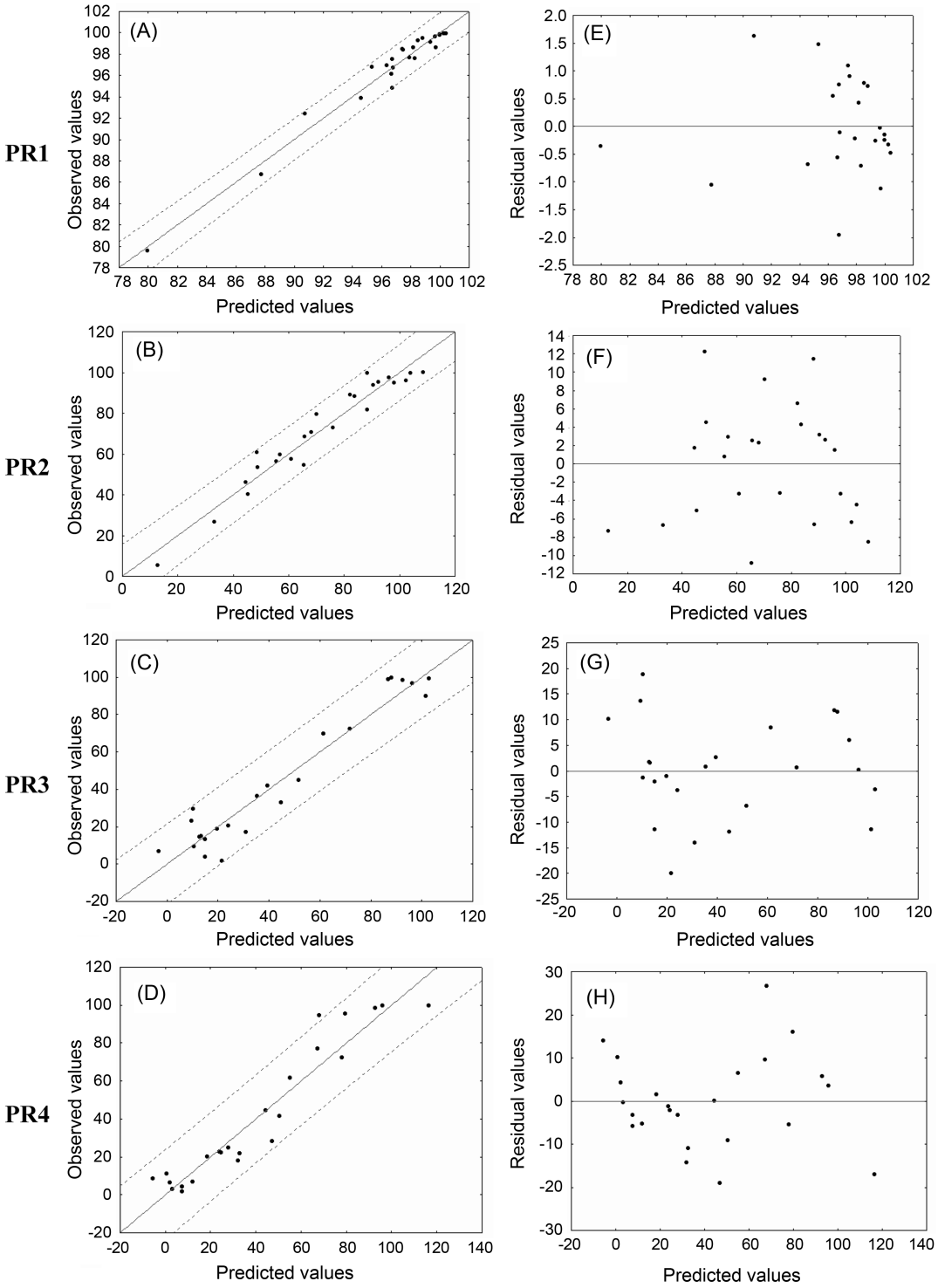

Fig. 1 Graphical representation (A-D) of observed values versus predicted given by fitted model together whit corresponding the predicted intervals; (E-H) of residuals values versus predicted

Table 4 Results of the leave-many-out (LMO) test

number of test	PR1		PR2		PR3		PR4	
	Q^2_{LMO}	QS_{LMO}	Q^2_{LMO}	QS_{LMO}	Q^2_{LMO}	QS_{LMO}	Q^2_{LMO}	QS_{LMO}
1	0.97	0.87	0.91	10.63	0.93	9.87	0.91	10.63
2	0.97	0.83	0.89	11.15	0.92	9.93	0.89	11.15
3	0.97	0.88	0.90	11.27	0.93	9.81	0.90	11.27
4	0.97	0.88	0.90	11.08	0.94	8.57	0.90	11.08
5	0.84	0.85	0.90	10.89	0.93	8.99	0.90	10.89
6	0.82	0.88	0.90	11.29	0.93	9.44	0.90	11.29
7	0.98	0.77	0.91	10.83	0.93	9.86	0.91	10.83
8	0.98	0.78	0.91	10.80	0.92	10.24	0.91	10.80
9	0.96	0.86	0.90	11.23	0.93	9.46	0.90	11.23
10	0.97	0.88	0.91	10.95	0.92	10.29	0.91	10.95
Average values	**0.94**	**0.85**	**0.90**	**11.01**	**0.93**	**9.64**	**0.90**	**11.01**

Table 5 R^2_Y and Q^2_Y values after ten Y-scrambling tests

Iteration no.	PR1		PR2		PR3		PR4	
	R^2_Y	Q^2_Y	R^2_Y	Q^2_Y	R^2_Y	Q^2_Y	R^2_Y	Q^2_Y
1	0.01	0.00	0.02	0.00	0.01	0.00	0.08	0.00
2	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00
3	0.02	0.00	0.09	0.00	0.00	0.00	0.01	0.00
4	0.03	0.00	0.05	0.00	0.00	0.00	0.00	0.00
5	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00
6	0.01	0.00	0.03	0.00	0.02	0.00	0.08	0.00
7	0.00	0.00	0.20	0.00	0.11	0.00	0.03	0.00
8	0.06	0.00	0.00	0.00	0.00	0.00	0.05	0.00
9	0.08	0.00	0.00	0.00	0.01	0.00	0.00	0.00
10	0.10	0.00	0.02	0.00	0.03	0.00	0.32	0.00

of descriptors such as: Radial Distribution Function (RDF descriptors, RDF010e) (Hemmer *et al*. 1999), 3D-Molecule Representation of Structures based on Electron diffraction *i.e*., 3D-MoRSE descriptors (Mor01m) (Schuur and Gasteiger 1996) and connectivity indices (X5v) (Kier and Hall 1986). For this permeation experiment the rejection parameter increases with a decrease of the ESpm05r, X5v and increase EEig03d, RDF010e, Mor01m. Interestingly, the signs of regression coefficient value for two edge adjacency indices parameters are opposite which suggests that the decrease of eigenvalue 05r and increase eigenvalue 03d of the molecule from edge adjacency matrix weighted by resonance integrals or by dipole moments are important to the measure property. The edge adjacency matrix is derived from the H-depleted molecular graph and encodes the connectivity between graph edges. Relatively low absolute of the regression coefficient of the Mor01m indicates that its significance in the model is not crucial. Three of the descriptors in this model (PR1) are derived from the 3D structural data of compound. The RDF010e can be interpreted as a probability distribution of finding an atom in a specific volume of radius of 1.0 and weighted by the atomic

Sanderson electronegativities. It does not depend on the number of atoms and consequently on the size of the molecule. This means that the RDF010e descriptor refers to the 3D arrangement of atoms around each other. The second descriptor, X5v, is the valence connectivity index of order 5 which accounts for heteroatoms and multiple bonds present in every 5 connected vertex. The last descriptor in this model, Mor01m, has some characteristics common with RDF. The 3D-MoRSE descriptors although are based on the theoretically calculated electron scattering curves from the known 3D structure of molecule. In our model PR1 this descriptor is scaled on atomic mass. From the above it can be concluded that the rejection for membrane 1 depends on the arrangement of both atomic mass and atomic electronegativity in molecule volume.

The 4-parametric equation **(PR2)** defines the best model for the membrane Memb-2 which is made from the same material as Memb-1 but shows lower NaCl rejection (60% vs. 92%). The molecular descriptors included in PR2 are surprisingly different from those in PR1, *i.e.*,: radius of gyration (mass weighted) (RGyr) (Katritzky *et al*. 1996, Tanford 1961), Broto-Moreau autocorrelation of a topological structure - lag 5/weighted by atomic van der Waals volumes (ATS5v) (Broto and Devillers 1990), Radial Distribution Function - 4.0/weighted by atomic van der Waals volumes (RDF040v) (Hemmer *et al*. 1999), Ghose-Crippen octanol-water partition coefficient (ALOGP) (Ghose and Crippen 1986). These parameters pertain to different logical block of descriptors such as: Geometrical descriptors (RGyr) (Tanford 1961), RDF descriptors (RDF040v) (Hemmer *et al*. 1999), 2D autocorrelations (Broto *et al*. 1984), Molecular properties (ALOGP). The RGyr is a measure of molecular compactness - the smaller is its value the more atoms are close to the center of mass. It is also a crucial parameter of this model, so it indicates that the more compact the molecule is, the lower is its retention. The second most important descriptor is ATS5v which is a sum over products of van der Waals volumes of each pair of atoms with topological distance of 5. In general this topological distance practically concerns mostly terminal atoms, consequently this descriptor is correlated with a molecular volume. Thus it is not surprising that the rejection increases with the increase of that parameter. In PR2 only one descriptor is similar to that in the PR1 model - Radial Distribution Function descriptor RDF040v. However, here it is weighted by atomic van der Waals volumes. ALOGP is a less important parameter of this model; it is connected with adsorption properties of organic compounds on membrane materials. Here it should be mentioned that Kiso *et al*. (2001) already found that the rejection does not correlate well with logP alone, and the sorption of aromatic pesticides on the membranes is not only related to logP but also is influenced by the molecular shape. The model PR2 shows that the rejection depends mostly on the molecular shape reflected by the three crucial parameters of the model: RGyr, ATS5v and RDF040v.

The 2-parametric equation **(PR3)** define the best model for Memb-3 - the sulfonated polyethersulfone membrane with the nominal value of NaCl retention of 51%. The molecular descriptors incorporated in PR3 are: R maximal autocorrelation of lag 5/weighted by atomic polarizabilities (R5p+) (Consonni and Todeschini 2002), Radial Distribution Function - 7.5/weighted by atomic van der Waals volumes (RDF075v) (Hemmer *et al*. 1999). In this model the crucial is the R5p+ descriptor that pertains to Geometry, Topology, and Atom-Weights Assembly *i.e.* GETAWAY descriptors block. It is the special eigenvalue of the influence/distance matrix weighted by atomic polarizabilities. The second parameter of this model is again one of the RDF descriptors (Hemmer *et al*. 1999).

The 5-parametric equation **(PR4)** define the best model for Memb-4, similar in the chemical composition to Memb-3 but of much higher hydrodynamic permeability and of lower NaCl rejection (15% with 15% vs. 51%). The molecular descriptors in PR4 are: R matrix average row sum (RARS), 3D-MoRSE - signal 27/weighted by atomic masses (Mor27m) (Schuur *et al*. 1996), Radial Distribution

Function - 7.5/weighted by atomic van der Waals volumes (RDF075v) (Hemmer *et al*. 1999), presence/ absence of O-O and N-N at topological distance 5 (B05[O-O]) and 2 (B02[N-N]) respectively. Two last descriptors pertain to a group called 2D binary fingerprints (Carhart *et al*. 1985). The remaining descriptors belong to different logical blocks of descriptors such as: GETAWAY descriptors (RARS) (Consonni and Todeschini 2002), 3D-Molecule Representation of Structures based on Electron diffraction (3D-MoRSE) descriptors (Mor27m) (Schuur and Gasteiger 1996) and RDF (RDF075v) (Hemmer *et al*. 1999). This model incorporates a variety of different descriptors, however the RDF descriptor (here, again, weighted by atomic van der Waals volumes) is present as in previous models.

4. Conclusions

In this work we have developed linear models for the prediction of the retention of pesticides. The resulting models fit the experimental data reasonably well explaining 91-95% of all variance. Additionally, the cross-validation coefficients reflecting the predictive power of a regression Q^2_{LOO} are between 0.92-0.97, and Q^2_{LMO} are between 0.91-0.94. The Y-scrambling test proved that good statistics obtained for equation (PR1-PR4) is not due to a chance correlation or structural dependency of the training set. The main purpose of this investigation was to determine the parameters which best describe the rejection properties. The obtained results show that the retention of these compounds is mainly determined by the Radial Distribution Function descriptors (RDF010e, RDF040v, RDF075v), 3D-MoRSE descriptors (Mor01m, Mor27m), GETAWAY descriptors (R5p+, RARS) and other. The models provide important information on the structure-activity relationships of these types of compounds at the molecular level concerning the application to remove hazard organic pollutants.

Unfortunately, the retention of the same group of pesticides is described by different sets of descriptors even for the membranes of similar chemical composition. However, one should noted that a linear model applied here is the simplest one. In the next work we will check if some basic nonlinear transformations of the retention coefficient will be more advantageous.

References

Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Markopoulos, J. and Igglessi-Markopoulou, O. (2006), "A novel QSAR model for modeling and predicting induction of apoptosis by 4-Aryl-4H-chromenes", *Bioorgan. Med. Chem.*, **14**(19), 6686-6694.

Allison, P.D. (1999), *Multiple Regression: A Primer*, Pine Forge Press.

Baumann, K. (2005), "Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging", *QSAR Comb. Sci.*, **24**(9), 1033-1046.

Belsley, D.A., Kuh, E. and Welsch, R.E. (2005), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley-Interscience.

Bland, M. (2000), *Introduction to Medical Statistics*, 3rd Edition, Oxford University Press, London.

Broto, P., Moreau, G. and Vandycke, C. (1984), "Molecular structures: perception, autocorrelation descriptor and QSAR studies. System of atomic contributions for the calculation of the n-octanol/water coefficients", *Eur. J. Med. Chem.*, **19**(1), 71-78.

Broto, P. and Devillers, J. (1990), "Autocorrelation of properties distributed on molecular graphs", *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Eds. Karcher,W. and Devillers, J., Kluwer, Dordrecht.

Carhart, R.E., Smith, D.H. and Venkataraghavan, R. (1985), "Atom pairs as molecular features in structureac

tivity studies: definition and applications", *J. Chem. Comput. Sci.*, **25**(2), 64-73.
Consonni, V., Todeschini, R. and Pavan, M. (2002), "Structure/Response Correlations And Similarity/Diversity Analysis By Getaway Descriptors. 1. Theory Of The Novel 3d Molecular Descriptors", *J. Chem. Comput. Sci.*, **42**(3), 682-692.
Consonni, V. and Todeschini, R. (2002), *Rational Approaches to Drug Design,* Vol. 42, Prous Science, Barcelona.
Duchowicz, P.R., Talevi, A., Bruno-Blanch, L.E. and Castro, E.A. (2008), "New QSPR study for the prediction of aqueous solubility of drug-like compounds", *Bioorgan. Med. Chem*., **16**(17), 7944-7955.
Estrada, E., Guevara, N. and Gutman, I. (1998), "Extension of edge connectivity index. relationships to line graph indices and QSPR applications", *J. Chem. Comput. Sci.*, **38**(3), 428-431.
Frisch, M.J., Trucks, G.W., Schlegel, H.B., *et al*. (2004), *Gaussian 03*, *Revision D.01*, Gaussian, Inc.
Ghose, A.K. and Crippen, G.M. (1986), "Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition coefficients as a Measure of hydrophobicity", *J. Comput. Chem.*, **7**(4), 565-577.
Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2!", *J. Mol. Graph. Model.*, **20**(4), 269-276.
Gramatica, P. (2007), "Principles of QSAR models validation: internal and external", *QSAR Comb. Sci.*, **26**(5), 694.
Hemmer, M.C., Steinhauer, V. and Gasteiger, J. (1999), "The Prediction of the 3D Structure of Organic Molecules from Their Infrared Spectra", *Vib. Spectrosc*., **19**(1), 151-164.
Hu, R., Yin, C., Wang, Y., Lu, C. and Ge, T. (2008), "QSPR study on GC relative retention time of organic pesticides on different chromatographic columns", *J. Sep. Sci.*, **31**(3), 2434-2443.
Jäntschi, L., Bolboacă, S.D. and Diudea, M.V. (2007), "Chromatographic retention times of polychlorinated biphenyls: from structural information to property characterization", *Int. J. Mol. Sci.*, **8**(11), 1125-1157.
Katritzky, A.R., Mu, L., Lobanov, V.S. and Karelson, M. (1996), "Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics", *J. Phys. Chem.*, **100**(24), 10400-10407.
Kier, L.B. and Hall, L.H. (1981), "Derivation and significance of valence molecular connectivity", *J. Pharm. Sci.*, **70**, 583-589.
Kier, L.B. and Hall, L.H. (1986), *Molecular Connectivity in Structure-Activity Analysis*, RSP-Wiley, Chichetser.
Kim, K.J., Fane, A.G., Ben Aim, R., Liu, M.G., Jonsson, G., Tessaro, I.C., Broek, A.P. and Bargeman, D. (1994), "A comparative study of techniques used for porous membrane characterization: pore characterization", *J. Membrane Sci.*, **87**(1-2), 35-46.
Kiso, Y., Nishimura, Y., Kitao, T. and Nishimura, K. (2000), "Rejection properties of non-phenilic pesticides with nanofiltration membranes", *J. Membrane Sci.*, **171**(2), 229-237.
Kiso, Y., Sugiura, Y., Kitao, T. and Nishimura, K. (2001), "Effects of hydrophobicity and molecular size on rejection of aromatic pesticides with nanofiltration membranes", *J. Membrane Sci.*, **192**(1-2), 1-10.
Kołtuniewicz, A.B. and Drioli, E. (2008), *Membranes in Clean Technologies: Theory and Practice*, Vol.1, 2, WILEY-VCH.
Koros, W.J., Ma, Y.H. and Shimidzu, T. (1996), "Terminology for membranes and membrane processes", *J. Membrane Sci*., **120**(2),149-159.
Koter S., Gilewicz-Łukasik, B., Nowaczyk, A., Nowaczyk, J. (2009), "Application of molecular descriptors to the prediction of retention in organic solvent nanofiltration", *Pol. J. Chem*., **83**(12), 2137-2152.
Li, L., Xie, S., Cai, H., Bai, X. and Xue, Z. (2008), "Quantitative structure-property relationships for octanol-water partition coefficients of polybrominated diphenyl ethers", *Chemosphere*, **72**(10), 1602-1606.
Libotean, D., Giralt, J., Rallo, R., Cohen, Y., Giralt, F., Ridgway, H.F., Rodriguez, G. and Phipps, D. (2008), "Organic compounds passage through RO membranes", *J. Membrane Sci.*, **313**(1-2), 23-43.
Schuur, J.H. and Gasteiger, J. (1996), *Software Development in Chemistry*, Vol. 10, Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main.
Schuur, J.H., Selzer, P. and Gasteiger, J. (1996), "The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure - spectra correlations and studies of biological activity", *J. Am. Chem. Soc.*, **36**(2), 334-344.
Scott, K. (1995), *Handbook of Industrial Membranes*, 1st Edition, Elsevier Science.
Tanford, C. (1961), *Physical Chemistry of Macromolecules*, Wiley, New York.

Todeschini, R. and Consonni, V. (2000), *Handbook of Molecular Descriptors*, Wiley-VCH New York.
Toghiani, R.K., Toghiani, H., Maloney, S.W. and Boddu, V.M. (2008), "Prediction of physicochemical properties of energetic materials", *Fluid Phase Equilibr.*, **264**(1-2), 86-92.
Topliss, J.G. and Costello, R.J. (1972), "Chance correlations in structure-activity studies using multiple regression analysis", *J. Med. Chem.*, **15**(10), 1066 -1068.
Toropov, A.A., Toropova, A.P. and Raska, I. (2008), "QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES", *Eur. J. Med. Chem.*, **43**(4), 714-740.
Trägårdh, G. and Ölund, K. (1986), "A method for characterization of ultrafiltration membranes", *Desalination*, **58**(3), 187-198.
Tropsha, A., Gramatica, P. and Gombar, V.K. (2003), "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models", *QSAR Comb. Sci.*, **22**(1), 69-77.
Tsapiuk, E.A. (1996), "Ultrafiltration separation of aqueous solutions of poly(ethylene glycol)s on the dynamic membrane formed by gelatin", *J. Membrane Sci.*, **116**(1), 17-29.
Yangali-Quintanilla, V., Verliefde, A., Kim, T.-U., Sadmani, A., Kennedy, M. and Amy, G. (2009), "Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes", *J. Membrane Sci.*, **342**(1-2), 251-262.

CC