QSTR STUDY OF ORGANIC PHOSPHONIUM SALTS BY MLR

SIMONA FUNAR-TIMOFEI, ADRIANA POPA

Institute of Chemistry of the Romanian Academy, 24 Mihai Viteazul Bvd.,
300223 Timisoara, Romania, e-mail: timofei@acad-icht.tm.edu.ro

ABSTRACT

The polymer-bound phosphonium salts are known as disinfectants, being important for drugs with prolonged activity and less toxicity, antifouling coatings and fiber finishing, water and air disinfection. The toxicity (expressed as the logarithm of oral lethal dose for mouse) was related to the structural features of a series of organic phosphonium salts by MLR. The structure of these compounds was modeled by molecular mechanics calculations and descriptors were then derived from the minimized structures. The structural features thus derived important for compound toxicity were derived by MLR combined with genetic algorithm for variable selection. Phosphonium salts toxicity was influenced by their electron distribution and steric factors.

KEYWORDS: phosphonium salts, toxicity, MLR, QSTR, LD_{50}

INTRODUCTION

Polyethylene glycols (PEGs) are polymers of ethylene oxide with the generalized formula $HO(CH2CH2\ O)_n$-H, "*n*" indicating the average number of oxyethylene groups. They comprise a class of compounds varying in molecular weights between 200 and over 10,000. Due to their presence in many cosmetics, an evaluation of their safety is critical, as potential exposure of consumers may be chronic and extensive [1]. PEGs and their anionic derivatives or surfactants are used as cleansing agents, emulsifiers, skin conditioners, and humectants. The biodegradable nano particles presenting poly(ethylene glycol) chains at their surface have appeared to be a particularly promising system.

In recent years, many insoluble disinfectants reported are phosphonium salts grafted on polymer [2, 3]. Polymeric disinfectants have received considerable attention in recent years with respect to important applications, such as: antifouling coatings and fiber finishing, drugs with

prolonged activity and less toxicity, water and air disinfection. The phosphonium salts grafted on polyethylene glycol were proved to have antibacterial activity against *Staphylococcus aureus* and *Escherichia coli* [4].

A new variant of synthesis of the poly(oxyethylene)s functionalized with quaternary phosphonium end groups by means of polymer-analogous quaternization reaction was reported [4, 5]. Lethal doses of the poly(oxyethylene)s functionalized with quaternary phosphonium end groups were determined by white mice and were calculated by the Probit method. According to the toxicity scale of Hodge and Steaner they can be considered as low toxic compounds [5].

In this paper 0D, 1D and 2D descriptors of organic phosphonium salts were related to their logarithm of oral mouse LD_{50} values to find out structural features which influence their toxicity.

METHODS AND MATERIALS

Molecular descriptors

Twenty eight quaternary phosphonium salts derivatives with known toxicity, the logarithm of the lethal oral dose for mouse LD_{50}, expressed in mg/Kg (calculated for the cationic structures), were employed in the quantitative structure-toxicity relationships (QSTR) study. The data were retrieved from the RTECS database (RTECS Database, MDL Information Systems, Inc. 14600 Catalina Street San Leandro, California U.S.A. 94577, http://www.ntis.gov/products/types/databases/rtecs.asp) (see table 1).

The molecular structure of the phosphonium salts (modeled as cations) was built by the ChemOffice package (ChemOffice 6.0, CambridgeSoft.Com, Cambridge, MA, U.S.A.) and energetically optimized using the molecular mechanics approach. Twenty-two types of descriptors were calculated by the Dragon software (Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy), like: constitutional, topological (PW5), walk and path count (MWC02), connectivity indices (X0A), information indices (HVcpx), 2D autocorrelations (GATS2m, GATS8v), edge adjacency indices (EPS0, ESpm03u), Burden eigenvalues, topological charge indices (GGI1), eigenvalue-based indices (VRm2,VEe1, EA2), Randic molecular profiles, geometrical, RDF descriptors (RDF030u, RDF045u,RDF090u, RDF030v), 3D-MoRSE (Mor05e, Mor28u), WHIM descriptors (P2e, P2p, E3m, Km, P1m, P2u, P1s, Kp, Vs), Getaway descroiptors (HATS3m, HATS6m, HATS3e, HATS4e, REIG, H1e, R7v+, R4m+, HATS7u, R6u, R3m+, RTm, H3v, R1e, R8v+), functional group counts, atom-centred fragments, charge, molecular properties (Neoplastic-80), 2D binary fingerprints, 2D frequency fingerprints.

Table 1. Name and the logarithm of the LD_{50} values of phosphonium salt structures

No	Phosphonium salt name	$logLD_{50}$	No	Phosphonium salt name	$logLD_{50}$
1	Phosphonium, acetonyltriphenyl-, iodide	-3.9	15	Phosphonium, (cyanomethyl)triphenyl-, chloride	-3.78
2	Phosphonium, tributyl-2-propen-1-yl-, chloride	-4.19	16	Phosphonium, (2,4-dimethylbenzyl)tributyl-, chloride	-4.3
3	Phosphonium, allyltriphenyl-, iodide	-3.89	17	Phosphonium, (2,4-dichlorobenzyl)triphenyl-, iodide	-4.48
4	Phosphonium, benzyltributyl-, chloride	-4.52	18	Phosphonium, (2,4-dichlorobenzyl)tri(p-tolyl)-, chloride	-3.95
5	Phosphonium, benzyltriphenyl-, iodide	-3.93	19	Phosphonium, (dichloromethyl)tripiperidino-, perchlorate	-4.41
6	Phosphonium, bis(p-butylamino)benzylphenyl-, iodide	-3.42	20	Phosphonium, (ethoxycarbonylmethyl)triphenyl-, bromide	-4.23
7	Phosphonium, bis (t-butylamino)methylphenyl-, iodide	-5.85	21	Phosphonium, (2-ethoxypropenyl)triphenyl-, iodide	-3.93
8	Phosphonium, (o-bromomethylbenzyl)triphenyl-, bromide	-4.22	22	Phosphonium, ethyltriphenyl-, iodide	-4.87
9	Phosphonium, (p-bromomethylbenzyl)triphenyl-, bromide	-4.47	23	Phosphonium, (o-methylbenzyl)triphenyl-, bromide	-4.15
10	Phosphonium, butyltriphenyl-, bromide	-3.85	24	Phosphonium, p-nitrobenzyltributyl-, iodide	-4.67
11	Phosphonium, butyltriphenyl-, iodide	-3.4	25	Phosphonium, (p-nitrobenzyl)triphenyl-, iodide	-4.47
12	Phosphonium, carboxymethyltriphenyl-, chloride	-3.3	26	Phosphonium, phenacyltriphenyl-, iodide	-3.96
13	Phosphonium, (p-chloromethylbenzyl)tris (dimethylamino)-, chloride	-6.13	27	Phosphonium, (3-phenoxypropyl)triphenyl-, bromide	-3.85
14	Phosphonium, chloromethyltriphenyl-, chloride	-4.04	28	Phosphonium, tetrabutyl-, iodide	-4.08

Multiple Linear Regression (MLR)

Multiple linear regression relates one experimental variable y_k to one or several structural variables x_i by the equation [6]:

$$y_k = b_o + \sum_i b_i \cdot x_{ik} + e_k \qquad (1)$$

where b represents regression coefficients and e the deviations and residuals. MLR calculations were performed by the STATISTICA (STATISTICA 7.1, Tulsa, StatSoft Inc, OK, USA) and MobyDigs [7] programs.

All the statistical tests were performed at a significance level of 5 % or less. Outliers were tested by estimating the standardized residuals of less than 3 standard deviation units.

The goodness of prediction of the MLR models was checked by the Akaike Information Criterion (AIC), the multivariate K correlation index (K_X-the multivariate correlation index of the

matrix of X descriptors and K_{XY} - the multivariate correlation index of the matrix of X descriptors and Y response variable), Y-scrambling ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$) and external validation (q^2_{ext}) parameters. All these calculations were performed by the MobyDigs software. The leave-one out cross-validation (q^2_{LOO}) procedure was, also, employed for the internal validation of models.

RESULTS AND DISCUSSION

In the MLR analysis a training set of 21 compounds and a test set of the following 7 compounds (selected randomly): 3, 4, 8, 15, 25, 26 and 28 were considered. Starting from the total set of calculated descriptors, MLR analysis has been applied to model the toxicity of the phosphonium salts. Variable selection was carried out by the genetic algorithm included in the MobyDigs program, using the RQK fitness function [8], with leave-one-out crossvalidation correlation coefficient as constrained function to be optimised, a crossover/mutation trade-off parameter T = 0.5 and a model population size P = 50. In addition, AIC - Akaike Information Criterion, the multivariate K correlation index (Kx and Kxy), Y-scrambling variables ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$), external q^2 (q^2_{ext}) values, q^2_{boot} - bootstrapping parameter and were calculated. RMSE (Root Mean Squared Errors) values were calculated for the training set (SDEC values) and test set (SDEP values). The final most stable MLR models are presented in Table 2.

Table 2. MLR results*

No	Descriptors	r^2	q^2_{LOO}	q^2_{boot}	q^2_{ext}	$r^2_{Y-scrambling}$	$q^2_{Y-scrambling}$	AIC	Kx	Kxy	SDEP	SDEC	F	s
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	P2e HATS3m HATS6m REIG	0.863	0.782	0.707	0.951	0.237	-0.498	0.138	0.26	0.40	0.321	0.254	25.26	0.291
2	PW5 RDF030u RDF045u Mor05e	0.862	0.763	0.717	0.690	0.379	-0.488	0.139	0.47	0.56	0.334	0.255	24.9	0.293
3	E3m HATS3m H1e R7v+	0.860	0.777	0.712	0.757	0.341	-0.302	0.141	0.29	0.45	0.325	0.257	24.57	0.294
4	P2p HATS3m HATS6m REIG	0.856	0.768	0.684	0.943	0.28	-0.39	0.145	0.28	0.42	0.331	0.261	23.68	0.299
5	PW5 RDF045u Mor05e HATS6m	0.855	0.749	0.690	0.699	0.338	-0.226	0.146	0.45	0.54	0.344	0.261	23.62	0.299
6	P2e HATS6m REIG R4m+	0.854	0.794	0.718	0.911	0.314	-0.32	0.146	0.27	0.37	0.312	0.262	23.46	0.3

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
7	PW5 VRm2 RDF045u Mor05e	0.848	0.739	0.654	0.751	0.312	-0.426	0.153	0.41	0.51	0.351	0.268	22.3	0.307
8	GATS2m RDF045u Mor05e HATS6m	0.848	0.761	0.671	0.755	0.331	-0.255	0.153	0.40	0.42	0.336	0.268	22.3	0.307
9	Km HATS3m HATS6m REIG	0.847	0.737	0.668	0.926	0.303	-0.242	0.154	0.29	0.43	0.352	0.268	22.17	0.308
10	X0A HVcpx RDF090u P2e HATS7u	0.841	0.748	0.724	0.735	0.426	-0.252	0.189	0.53	0.56	0.344	0.274	15.83	0.324
11	EPS0 ESpm03u VEe1 Km	0.834	0.724	0.663	0.806	0.338	-0.291	0.167	0.41	0.51	0.361	0.279	20.16	0.32
12	MWC02 ESpm03 u Km R6u	0.833	0.761	0.713	0.891	0.254	-0.474	0.168	0.28	0.43	0.336	0.281	19.94	0.322
13	X0A HVcpx RDF090 u P2e R3m+	0.829	0.724	0.679	0.827	0.301	-0.43	0.203	0.37	0.47	0.361	0.284	14.56	0.336
14	X0A HVcpx RDF090u P2e	0.826	0.745	0.728	0.755	0.411	-0.151	0.175	0.45	0.51	0.347	0.287	18.94	0.329
15	P1m H3v HATS3e REIG	0.820	0.727	0.697	0.898	0.415	-0.189	0.181	0.48	0.54	0.359	0.291	18.28	0.333
16	E3m H1e R7v+	0.818	0.721	0.715	0.759	0.162	-0.286	0.156	0.18	0.41	0.363	0.293	25.53	0.325
17	E3m R7v+ R1e	0.817	0.737	0.737	0.674	0.143	-0.418	0.157	0.14	0.39	0.352	0.294	25.27	0.327
18	X0A GATS8v RDF090 u P2e R4m+	0.812	0.716	0.638	0.853	0.311	-0.502	0.223	0.28	0.36	0.366	0.297	12.99	0.352
19	P1m H3v REIG RTm	0.806	0.718	0.687	0.899	0.308	-0.236	0.195	0.43	0.48	0.365	0.302	16.67	0.346
20	P1s H3v HATS3e REIG	0.806	0.707	0.671	0.917	0.483	0.015	0.196	0.48	0.53	0.372	0.303	16.57	0.347
21	GGI1 Km R6u	0.804	0.715	0.659	0.874	0.152	-0.469	0.168	0.20	0.42	0.366	0.304	23.29	0.338
22	E3m HATS4e	0.804	0.741	0.751	0.739	0.251	-0.118	0.144	0.26	0.57	0.349	0.304	36.82	0.329
23	P2u HATS3m H1e	0.801	0.723	0.725	0.708	0.122	-0.382	0.17	0.12	0.32	0.361	0.306	22.83	0.34
24	P1m Kp Vs REIG	0.798	0.714	0.596	0.910	0.269	-0.302	0.203	0.60	0.64	0.367	0.308	15.82	0.353

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
25	P2u R8v+ R1e	0.794	0.717	0.702	0.655	0.277	-0.2	0.166	0.29	0.44	0.359	0.307	23.09	0.339
26	ESpm03u VEA2 Km	0.793	0.710	0.688	0.872	0.163	-0.298	0.177	0.17	0.34	0.37	0.312	21.76	0.347
27	Mor28u HATS3m R6u	0.790	0.688	0.669	0.847	0.258	-0.229	0.18	0.29	0.47	0.384	0.315	21.3	0.35
28	P1m Kp REIG	0.787	0.703	0.696	0.903	0.355	-0.097	0.183	0.48	0.58	0.374	0.317	20.9	0.353
29	RDF030v Neoplastic-80	0.727	0.660	0.658	0.733	0.093	-0.337	0.19	0.18	0.49	0.394	0.353	25.24	0.38

* r^2 – correlation coefficient, SDEP – standard deviation error in prediction ($RMSE_{test}$), SDEC – standard deviation error in calculation ($RMSE_{training}$), F- Fischer test, s – standard error of estimate, AIC - Akaike Information Criterion, the multivariate K correlation index (Kx and Kxy), Y-scrambling variables ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$), q^2_{ext} - external q^2, q^2_{boot} - bootstrapping parameter, q^2_{LOO} - leave-one out cross-validation parameter

Starting from the descriptor matrix containing all variables, following descriptors were found to be significant and were included in the final MLR models: PW5 (path/walk 5 - Randic shape index), MWC02 (molecular walk count of order 02), X0A (average connectivity index chi-0), HVcpx (graph vertex complexity index), GATS2m (Geary autocorrelation - lag 2 / weighted by atomic masses), GATS8v (Geary autocorrelation - lag 8 / weighted by atomic van der Waals volumes), EPS0 (edge connectivity index of order 0), ESpm03u (Spectral moment 03 from edge adj. matrix), GGI1 (topological charge index of order 1),VRm2 (average Randic-type eigenvector-based index from mass weighted distance matrix),VEe1 (eigenvector coefficient sum from electronegativity weighted distance), VEA2 (average eigenvector coefficient sum from adjacency matrix), RDF030u (Radial Distribution Function - 3.0 / unweighted), RDF045u (Radial Distribution Function - 4.5 / unweighted), RDF090u (Radial Distribution Function - 9.0 / unweighted), RDF030v (Radial Distribution Function - 3.0 / weighted by atomic van der Waals volumes), Mor05e (3D-MoRSE - signal 05 / weighted by atomic Sanderson electronegativities), Mor28u (3D-MoRSE - signal 28 / unweighted), P2e (2nd component shape directional WHIM index / weighted by atomic Sanderson electronegativities), P2p (2nd component shape directional WHIM index / weighted by atomic polarizabilities), E3m (3rd component accessibility directional WHIM index / weighted by atomic masses), Km (K global shape index / weighted by atomic masses), P1m (1st component shape directional WHIM index / weighted by atomic masses), P2u (2nd component shape directional WHIM index / unweighted), P1s (1st component shape directional WHIM index / weighted by atomic electrotopological states), Kp (K global shape index / weighted by atomic

polarizabilities), Vs (V total size index / weighted by atomic electrotopological states), HATS3m (leverage-weighted autocorrelation of lag 3 / weighted by atomic masses), HATS6m (leverage-weighted autocorrelation of lag 6 / weighted by atomic masses), HATS3e (leverage-weighted autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities), HATS4e (leverage-weighted autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities), REIG (first eigenvalue of the R matrix), H1e (H autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities), R7v+ (R maximal autocorrelation of lag 7 / weighted by atomic van der Waals volumes), R4m+ (R maximal autocorrelation of lag 4 / weighted by atomic masses), HATS7u (everage-weighted autocorrelation of lag 7 / unweighted), R6u (R autocorrelation of lag 6 / unweighted), R3m+ (R maximal autocorrelation of lag 3 / weighted by atomic masses), RTm (R total index / weighted by atomic masses), H3v (H autocorrelation of lag 3 / weighted by atomic van der Waals volumes), R1e (R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities), R8v+ (R maximal autocorrelation of lag 8 / weighted by atomic van der Waals volumes), Neoplastic-80 (Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 80%).

Good correlations with the phosphonium salt toxicity and models with predictive power were obtained (Table 2). The best externally predictive single model, based on four descriptors, would be selected from the population of 50 models. It's very difficult to determine which one is the best for their similar and comparable performance. Considering these models, the range of q^2_{LOO} is 0.660 – 0.794, while the range of q^2_{ext} is 0.655 – 0.951. The selection criterion used in this study is that the model should have higher r^2, higher cross-validated q^2_{LOO}, higher external predictive ability, least difference between internal and external predictive ability, the fewer chemicals outside the chemical domain and the fewer chemicals with large relative errors. On the basis of the above principles, model 1 was selected as the best single model, whose regression equation was:

$$\log LD_{50} = 2.36(\pm 0.99) + 2.36(\pm 1.31)P2e - 12.37(\pm 2.88)HATS3m + 4.87(\pm 1.15)HATS6m - 11.28(\pm 1.6)REIG$$

$$N_{training} = 21\ N_{test} = 7$$

$$r^2_{training} = 0.863\ q^2_{LOO} = 0.782\ q^2_{ext} = 0.951\ r^2_{Y-scrambling} = 0.251\ q^2_{Y-scrambling} = -0.348$$

$$K_{XY} = 0.402\ K_X = 0.258\ RMSE_{training} = 0.254\ RMSE_{test} = 0.321$$

From all the statistical parameters, it can be seen that the proposed model is stable, robust and predictive. Figure 1 shows the regression plot of the best single model. Descriptors P2e,

HATS3m, HATS6m, REIG, the four most important descriptors in model 1 were found in many other individual models.

Following descriptors: HATS3m abd REIG yielded high toxicity values. A low toxicity of phosphonium compounds was derived by the P2e and HATS6m descriptors.

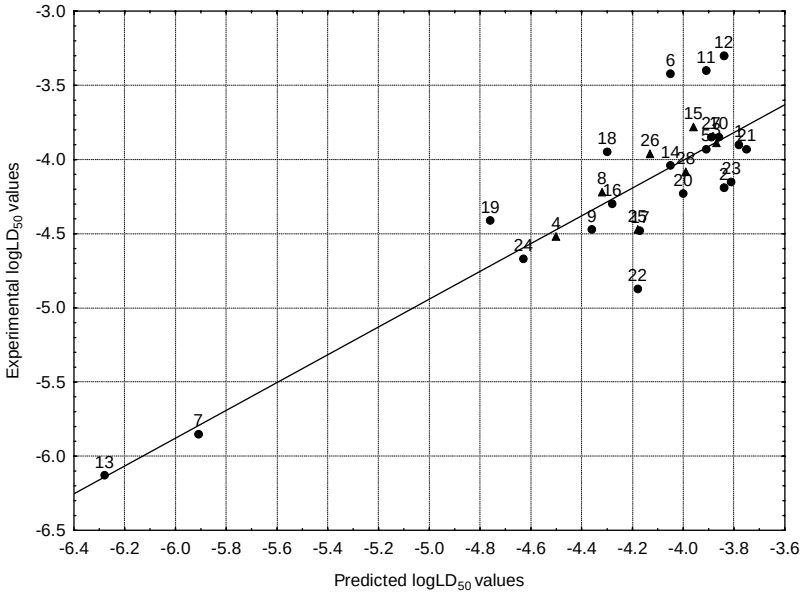

Figure 1. Experimental *versus* predicted $logLD_{50}$ values of the final MLR model 1 (Table 2). Training set is marked by circles, test set marked by blue triangles.

The Williams plot for model 1 (Table 2) is presented in figure 2. Leverage values are less than the control value (of 0.714).

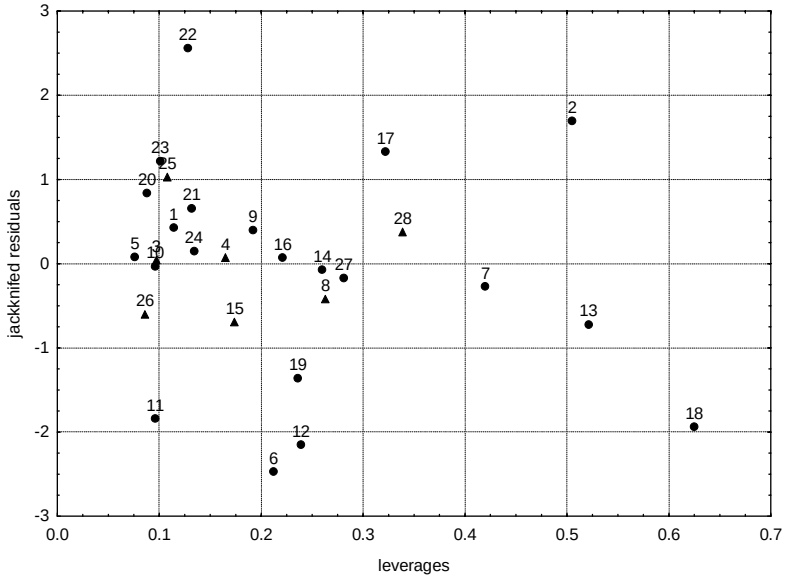

Figure 2. Williams plot: jackknifed residuals *versus* leverages of the MLR model 1 (Table 2). Training set is marked by circles, test set marked by triangles.

No outliers or influential points were found in model 1, considered to have best statistical results.

P2e (2nd component shape directional WHIM index / weighted by atomic Sanderson electronegativities) encodes information on atomic distribution and shape (weighted by electronegativity) along the main direction of the molecule. It was concluded that the electronic distribution is very important for the phosphonium salts toxicity.

GETAWAYs descriptors which are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. HATS3m (leverage-weighted autocorrelation of lag 3 / weighted by atomic masses), HATS6m (leverage-weighted autocorrelation of lag 6 / weighted by atomic masses) and REIG (first eigenvalue of the R matrix), are evaluated by considering separately all the contributions of each different path length (lag) in the molecular graph, as collected in the topological distance matrix. Therefore steric factors can be considered to influence the toxicity.

CONCLUSIONS

The quaternary phosphonium salts have several applications, being mainly used as end groups of some polymers, with disinfectant properties.

Multiple linear regressions combined with genetic algorithm for variable selection was used to correlate the logarithm of the oral lethal dose for mouse with structural features of quaternary phosphonium salts. Following descriptors: topological, walk and path count, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, topological charge indices, eigenvalue-based indices, RDF descriptors, 3D-MoRSE, WHIM descriptors, Getaway descriptors, and molecular properties were present in the final MLR models with acceptable statistical results.

The electron distribution and steric factors are found to be important for phosphonium salt toxicity.

REFERENCES

1. Fruijtier-Polloth, C. Toxicology 2005; 214: 1–38.

2. Kanazawa, A.; Ikeda, T.; Endo T. J. Polym. Sci. Pol. Chem. 1994; 32: 1997-2001.

3. Kanazawa, A.; Ikeda, T.; Endo, T. J. Appl. Polym. Sci. 1994; 53: 1237-1244.

4. Popa, A.; Davidescu, C.M.; Ilia, G.; Iliescu, S.; Dehelean, G.; Trif, R.; Păcureanu, L.; Macarie L. Rev. Roum. Chim. 2003 ; 48(1) : 41-48.

5. Popa, A. ; Trif, A. ; Curtui, V.G. ; Dehelean, G. ; Iliescu, S. ; Ilia G. Phosphorus Sulfur. 2002; 177: 2195-2196.

6. Wold, S.; Dunn III, W.J. J. Chem. Inf. Comp. Sci**.** 1983; 23: 6-13.

7. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. MobyDigs: software for regression and classification models by genetic algorithms, in: 'Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks'. (Leardi R., Ed.), Chapter 5, Elsevier, 2004, pp. 141-167.

8. Todeschini R., Consonni V., Mauri A., Pavan M. Anal. Chim. Acta 2004; 515: 199-208.