ACADEMIA ROMÂNĂ
Revue Roumaine de Chimie
http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2015, *60*(2-3), 189-195

Dedicated to Professor Zeno Simon
on the occasion of his 80th anniversary

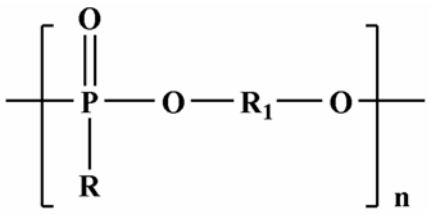
THE INVESTIGATION OF STRUCTURE-FLAMMABILITY RELATIONSHIPS OF POLYPHOSPHONATES BY PLS ANALYSIS

Luminita CRISAN,* Smaranda ILIESCU and Simona FUNAR-TIMOFEI

Timişoara Institute of Chemistry of Romanian Academy, 24 Mihai Viteazul Avenue, 300223 Timişoara, Roumania

Received August 25, 2014

To date many studies have been focused on the development of fire retardant mechanism and estimation of polymer flammability. In this analysis a series of 28 polyphosphonates has been investigated by Partial Least Square (PLS) methodology using previously calculated descriptors as independent variables, and the limiting oxygen index (LOI) as dependent variable. R and S polyphosphoester stereoisomers were previously derived using molecular mechanics calculations. Reliable statistics $R^2_{Y(CUM)}$ = 0.879 (for R series); $R^2_{Y(CUM)}$ =0.809 (for S series), and $Q^2_{(CUM)}$ = 0.840 (for R series); $Q^2_{(CUM)}$ = 0.776 (for S series) are obtained. The HATS6v (GETWAY) descriptor is the most relevant for both models, having the best VIP (Variables Importance in the Projection) value: VIP_{HATS6v} =1.320 (for R series) and VIP_{HATS6v} =1.353 (for S series). The PLS model obtained for the R stereoisomers was more predictive compared to the S stereoisomer one.



INTRODUCTION

There is great interest on polyphosphoesters due to their attractive properties that make them useful for a wide range of applications. One of the present interests lies in the intensification of the efforts concerning fire proofing of this macromolecular compounds.[1] The presence of phosphorus in the backbone polymer gives the self-extinguishing behavior to the polymer, but the flame retardancy, also, depends on the structure of the used bisphenol. Limiting oxygen index (LOI) measurements indicated their potential application as fire retardant materials being a precision method for determining the relative flammability of various materials by measuring the minimum concentration of oxygen required to support combustion. LOI values were determined on the powdered sample in according to modified ASTM D2863-70.[2] Few quantitative structure-property relationship studies for flammability properties of chemicals were reported, being usually related to flash point, autoignition temperature and lower and upper flammability limits.

In a previous study[3] the flammability of 28 polyphosphoesters has been investigated and the structure-flammability study oh these polymers was used to correlate the limiting oxygen index to structural descriptors by multiple linear regression, artificial neuronal networks and support vector machines. Monomer R chiral structures gave more stable and predictive models compared to the S isomers in all approaches. Stable and predictive ones were obtained by MLR, Multiple Linear

* Corresponding author: lumi_crisan@acad-icht.tm.edu.ro

Regression (best model having following statistical results: $R^2 = 0.882$, leave-one-out $Q^2 = 0.802$) and somewhat inferior fits were found in the nonlinear modelling by ANNs (Artificial Neural Networks) and SVMs (Support Vector Machines). It was concluded that monomer geometry influences the flame retardancy.

In this study a quantitative structure-flammability relationship for the same series reported in references[3] was modeled by the Partial Least Squares (PLS) approach by relating the limiting oxygen index to structural parameters obtained previously. New information on the flammability mechanism is reported.

MATERIALS AND METHODS

In the present paper we will concentrate on finding relations between one data block Y (the limiting oxygen index (LOI)) and data blocks X (Dragon and InstantJchem molecular descriptors), having a large number of variables.

Data Set. In this study a series of 28 polyphosphonates (R and S stereoisomers) was selected (see Table 1). Data for experimental limiting oxygen index (LOI), used as dependent variable in this study, was earlier reported.[3]

Molecular Descriptors Calculation. Four hundred and sixty-nine molecular descriptors were previously[3] calculated by the Dragon[4] (four hundred and seventeen descriptors) and InstantJchem[5] (fifty two descriptors) software for the R and S sets of twenty eight polyphosphonates derivatives. The molecular descriptors were verified to eliminate constant or near-constant variables.

Training and test set selection. The clustering technique is frequently used in QSAR/QSPR techniques to split the data sets into training and test sets.[6] The polyphosphonates derivatives were divided into training and test sets by means of the partition against medoids (PAM) algorithm[7] ("cluster" package available in R[8]) based on Euclidian distance. Twenty one out of twenty eight polyphosphonates derivatives were chosen as training set and supplied to PLS modeling. The limiting oxygen index (LOI) of the training set ranges (in %) from 0.18 to 0.55, and for the test set ranges from 0.22 to 0.47. This indicates that the data (structurally and LOI range) comprised in the test set and the training set are commensurate.

Table1

The experimental limiting oxygen index (LOI) and monomer smiles notation

No	Smiles code	LOI (%)
1	CC(C)(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1	38
2	O=P(OC1=CC=CC=C1)(OC1=CC=C(C=C1)C1=CC=CC=C1)C1=CC=CC=C1	35
3[a,b]	O=P(OC1=CC=CC=C1)(OC1=CC=C(CC2=CC=CC=C2)C=C1)C1=CC=CC=C1	30
4	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1\N=N/C1=CC=CC=C1)C1=CC=CC=C1	48
5	CC(C)(C1=CC(Br)=CC(Br)=C1)C1=CC(Br)=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C(Br)=C1	55
6	O=P(OC1=CC=CC=C1)(OC1=CC=C(SC2=CC=CC=C2)C=C1)C1=CC=CC=C1	28
7	O=P(OC1=CC=CC=C1)(OC1=CC=C(C=C1)S(=O)(=O)C1=CC=CC=C1)C1=CC=CC=C1	29
8	O=P(OC1=CC=CC=C1)(OC1=CC=C(C=C1)C1(CCCCC1)C1=CC=CC=C1)C1=CC=CC=C1	42
9	FC(F)(F)C(C1=CC=CC=C1)(C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1)C(F)(F)F	44
10	ClC(Cl)=C(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1	50
11[a]	O=P(OC1=CC=CC=C1)(OC1=CC=C(C=C1)C(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1)C1=CC=CC=C1	47
12	CC(C)(C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1)C1=CC(=CC=C1)C(C)(C)C1=CC=CC=C1	32
13	CC(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1	40
14[a,b]	CC(C)(C1=CC=CC=C1)C1=CC=C(C=C1)C(C)(C)C1=CC=C(OP(=O)(OC2=CC=CC=C2)C2=CC=CC=C2)C=C1	33
15	CC(C)(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1	28
16	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(C=C1)C1=CC=CC=C1	25
17[a,b]	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(CC2=CC=CC=C2)C=C1	22
18[a]	O=P(Oc1ccc(cc1)/N=N/c1ccccc1)(Oc1ccccc1)Oc1ccccc1	36
19	CC(C)(C1=CC(Br)=CC(Br)=C1)C1=CC(Br)=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C(Br)=C1	40
20	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(SC2=CC=CC=C2)C=C1	18
21[b]	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(C=C1)S(=O)(=O)C1=CC=CC=C1	20
22[a,b]	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(C=C1)C1(CCCCC1)C1=CC=CC=C1	31
23[a,b]	FC(F)(F)C(C1=CC=CC=C1)(C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1)C(F)(F)F	33
24	ClC(Cl)=C(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1	50
25[b]	O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=C(C=C1)C(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1	48
26	CC(C)(C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1)C1=CC(=CC=C1)C(C)(C)C1=CC=CC=C1	23
27	CC(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1	37
28	CC(C)(C1=CC=CC=C1)C1=CC=C(C=C1)C(C)(C)C1=CC=C(OP(=O)(OC2=CC=CC=C2)OC2=CC=CC=C2)C=C1	24

[a] Componds used as test set in the R stereoisomer PLS model; [b] Componds used as test set in the S stereoisomer PLS model

PLS Method. PLS regression is a commonly used technique that combines features from principal component analysis and multiple regression.[9] In this analysis the aim is to find components from the "X" variables matrix that are also relevant for the response "Y" variables matrix and to reduce the dimensionality of the "X" variables.[10] The major advantage for this method is to work with very large data sets and to overcome the problem of overfitting, multicollinearity and outliers. The most important statistical parameters for PLS are: (1) the squared correlation regression coefficient R^2 (the significant value > 0.8); (2) the squared cross-validated correlation coefficient, Q^2 (the significant value > 0.6), which measure of the quality and validity for the final PLS model; (3) the Variables Importance in the Projection (VIP) values and the (3) sign of the coefficients of the variables are more relevant in explaining the activity/property mechanism. In this study the QSPR matrix (Y matrix (the LOI value) and the X matrix (N = 28 rows/compounds and K = 469 columns/descriptors) was submitted to the SIMCA P+12 package[11] to perform initially PCA (Principal Component Analysis), and afterwards a PLS analysis.[12]

Optimization of the models: detection of outliers and significant variables. In order to optimize the PLS models the procedure for the detection of outliers, and the variables with significant coefficient different by zero was applied. The detection of outlier follows data points that exceeded ±3 standard deviations but no compound was removed as outlier. The variable selection was performed, preserving in the new model only the descriptors significantly different from zero (for elimination of noise).

Model Validation. For internal validation results several measures of robustness were employed: determination coefficient ($R^2Y(cum)$), leave-7-out crossvalidation (Q2(cum); values higher than 0.7 were considered as acceptable),[13] Y-scrambling and leave-more-out (q^2_{LMO}) cross-validation coefficient (carried out for 30% of data out of training, each run). A QSAR model can be considered robust when its performance remains satisfactory and stable when heavy perturbations (for instance by leave-many-out) in the training composition is made. Y-scrambling[14] testing is a technique for checking the robustness of a QSAR model and the statistical significance of the estimated predicted power. In this test, the dependent variable vector, Y-vector, is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The process was repeated 999 times. The risk of chance correlation was verified, also, by the Y-scrambling procedure.

The data over fitting and model applicability was controlled by comparing the root-mean-square errors (RMSE) and the mean absolute error (MAE)[15] of training and validation sets.

To test the robustness and predictive power of the model the concordance correlation coefficient (CCC)[16] was employed.

Among other statistical measures to check the model predictivity, other parameters[17] were used for the external test set, which assume that: 1) squared correlation coefficient (r^2) between the predicted and observed activities as well as squared correlation coefficient by cross-validation (Q^2); 2) coefficient of determination for linear regressions with intercepts set to zero, i.e. r_0^2 (predicted versus observed activities), and $r_0'^2$ (observed versus predicted activities); 3) slopes k and k' of the above mentioned two regression lines. All these measures were applied over the test set compounds. The following conditions should be satisfied for a model with acceptable predictive ability: (i) $Q^2 > 0.5$; (ii) $r^2 > 0.6$; (iii) $\frac{(r^2 - r_0^2)}{r^2} < 0.1$ and $0.85 \le k \le 1.15$; (iv) $\frac{(r^2 - r_0'^2)}{r^2} < 0.1$ and $0.85 \le k' \le 1.15$; (v) $\left|r_0^2 - r_0'^2\right| < 0.3$. For the external validation several types of variances explained in external prediction: Q^2_{F1} [18], Q^2_{F2} [19], Q^2_{F3} [20], r^2_m parameter[21] and R^2_{pred} [22] were calculated too. Models with values higher than 0.7 for Q^2_{F1}, Q^2_{F2} and Q^2_{F3}, higher than 0.85 for CCC and higher than 0.65 for r^2_m were considered as acceptable, these comparable thresholds values being rigorously determined by a simulation study.[23] For R^2_{pred} a threshold of 0.5 was considered.[22]

RESULTS AND DISCUSSION

In order to correlate the limiting oxygen index (LOI) values with calculated descriptors, the PLS calculations with the SIMCA P+12 package were performed. A PCA model was built for the whole X matrix (including N = 28 compounds and X = 469 descriptors for R stereoisomers and X = 471 descriptors for S stereoisomers). From thirteen significant principal components resulting from this analysis, the first three components already explained 56.6 % (R stereoisomers), 51,6% (S stereoisomers) of the information content of the

descriptor matrix. PLS calculations were also performed by the same program using 21 R stereoisomers as a training set and 7 R stereoisomers as a test set. The difference between $R^2_{Y(CUM)}$ and $Q^2_{(CUM)}$ values, of 0.311 (R stereoisomers) and 0.290 (S stereoisomers), demonstrated the overfit of the model, and suggest the need for enhancement of the model quality. The noise variables (the variables with coefficient values insignificantly different from 0) have been removed and the final resulted PLS_R and PLS_S models are robust and include only one principal component and 30 respectively 22 significant variables (see Table 2).

Table 2

Statistical parameters used for internal validation*

Model	R^2X (cum)	R^2Y(cum)	Q^2(cum)	A	K	N	$RMSE_{CALC}$	MAE_{CALC}	CCC_{CALC}
PLS_R	0.344	0.879	0.840	1	30	21	0.11	0.08	0.936
PLS_S	0.372	0.809	0.776	1	22	21	0.04	0.03	0.895

* R^2_X (CUM) and R^2_Y (CUM) are the cumulative sum of squares of all the *X* and *Y* values, respectively, explained by all extracted principal components; Q^2(CUM) is the fraction of the total variation of the *Y* values that can be predicted for all the *A* extracted principal components in the cross-validation procedure (7 rounds) used to establish the number of significant principal components, CALC – used for calibration set, N – number of compounds; root mean squared errors (RMSE), mean absolute error (MAE).

Table 3

The most important molecular descriptors according to VIP value*

PLS_R				PLS_S			
Var ID	VIP[1]	CoeffCS[1] (LOI(%))	Descriptor type	Var ID (Primary)	VIP[1]	CoeffCS[1] (LOI(%))	Descriptor type
HATS6v	1.32	0.071	GETAWAY descriptor	HATS6v	1.353	0.092	GETAWAY descriptor
Mor13u	1.257	-0.067	3D-MoRSE descriptor	HATS6m	1.259	0.085	GETAWAY descriptor
PW5	1.19	0.064	Topological index	HATS5v	1.185	0.08	GETAWAY descriptor
PW4	1.185	0.064	Topological index	Mor02m	1.153	-0.078	3D-MoRSE descriptor
HATS6m	1.162	0.062	GETAWAY descriptor	Jhetv	1.135	0.077	Topological indices
GATS6m	1.131	0.061	2D autocorrelation	MSD	1.106	-0.075	Topological index
HATS5v	1.127	0.06	GETAWAY descriptor	PW5	1.104	0.075	Topological index
MSD	1.121	-0.06	Topological index	PW4	1.103	0.075	Topological index
MAXDP	1.115	0.06	Topological index	Mor02e	1.027	-0.07	3D-MoRSE descriptor
L3v	1.107	0.059	WHIM descriptor				
Mor13v	1.103	-0.059	3D-MoRSE descriptor				
Mor13m	1.066	-0.057	3D-MoRSE descriptor				
GATS3m	1.03	-0.055	2D autocorrelation				
Jhetv	1.026	0.055	Topological index				
B06[C-P]	1.012	0.054	2D binary fingerprint				
G(S..P)	1.012	-0.054	Geometrical descriptor				

*HATS6v (leverage-weighted autocorrelation of lag 6 / weighted by van der Waals volume); Mor13u (3D- MoRSE signal 13 / unweighted); PW5 (path/walk 5 – Randic shape index); PW4 (path/walk 4 – Randic shape index); HATS6m (leverage-weighted autocorrelation of lag 6 / weighted by mass); GATS6m (Geary autocorrelation of lag 6 weighted by mass); HATS5v (leverage-weighted autocorrelation of lag 5 / weighted by van der Waals volume); MSD (mean square distance index (Balaban)); MAXDP (maximal electrotopological positive variation); L3v (3rd component size directional WHIM index / weighted by van der Waals volume); Mor13v (signal 13 / weighted by van der Waals volume); Mor13m (3D-MoRSE signal 13 / weighted by mass); GATS3m (Geary autocorrelation of lag 3 weighted by mass); Jhetv (Balaban-type index from van der Waals weighted distance matrix); B06[C-P] (Presence/absence of C – P at topological distance 6); G(S..P) (sum of geometrical distances between S..P); Mor02m (3D-MoRSE signal 2 / weighted by mass); Mor02e (3D-MoRSE signal 2 / weighted by Sanderson electronegativity).

In the analysis of the final PLS_R and PLS_S models we used as criteria the Variables Importance in the Projection (VIP) values and the sign of the coefficients (CoeffCS) (see Table 3). The VIP values computed by a module of the SIMCA P+12 program reflect the importance of descriptors in the PLS models. Variables with higher VIP scores (VIP>1) are most relevant for each model.

Y-randomization tests and leave-7-out crossvalidation runs were performed to check the robustness and internal predictive ability of the PLS models. The risk of chance correlation was verified by the Y-scrambling procedure, which was repeated 999 times. No chance correlation for the chosen PLS_R and PLS_S models, was confirmed by the low R^2 (0.190/0.157) and Q^2 (-0.217/-0.219) values of the randomized models.

The observed versus predicted LOI values obtained by the PLS_R and PLS_S models, respectively are presented in Fig. 1a and 1b respectively.

The predictive ability of final models was tested in accordance to several tests (Tables 4 and 5). The RMSE values provide a more reliable indication of the fitness of the model, independently of the applied splitting.[24] The calculated RMSE values for the training and validation sets of the PLS_R and PLS_S models demonstrate a predictive ability (low values) and sufficient generalizability (similar values). Higher R^2 values must be accompanied by Q^2 values as close to the R^2 ones as possible[24] to avoid overfitting, which was, also, checked by the RMSE and MAE values. The PLS_R model is completely satisfactory in the fitting and has high predictive power.

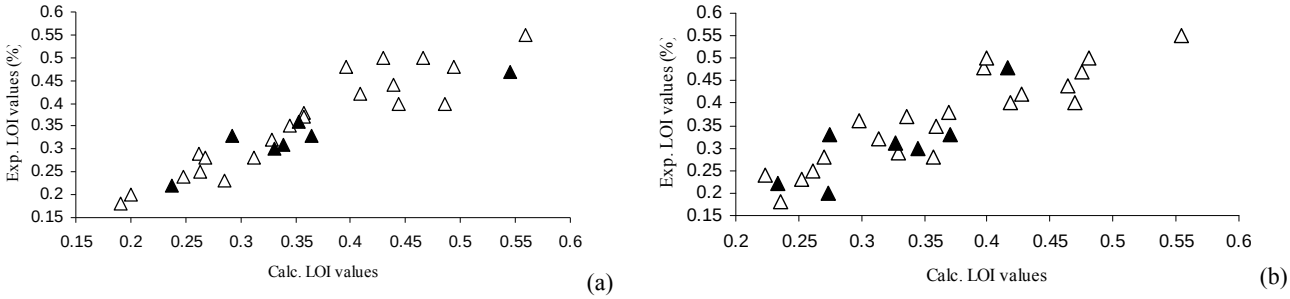


Fig. 1 – Experimental versus calculated LOI values by PLS_R (a) and PLS_S (b) models.
Training compounds are marked by white triangles and test compounds by black triangles.

Table 4

Statistical parameters used for external validation*

Model	$RMSE_{EXT}$	MAE_{EXT}	CCC_{EXT}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	r^2_{mEXT}	R^2_{pred}
PLS_R	0.11	0.07	0.886	0.734	0.692	0.869	0.665	0.769
PLS_S	0.05	0.04	*0.777*	0.769	*0.666*	0.761	*0.587*	0.794

* EXT – used for external set; concordance correlation coefficient (CCC); variances explained in external prediction: Q^2_{F1}, Q^2_{F2}, Q^2_{F3}; r^2m_{EXT} and R^2_{pred} – other predictivity parameters; underlined value surpasses the predictivity threshold; root mean squared errors (RMSE), mean absolute error (MAE).

Table 5

Golbraikh and Tropsha parameters calculated for external set*

Model	r^2	$\frac{r^2 - r_0^2}{r^2}$	$\frac{r^2 - r_0'^2}{r^2}$	k	k'	$\lvert r_0^2 - r_0'^2 \rvert$
PLS_R	0.887	0.070	0.013	0.930	1.067	0.051
PLS_S	0.700	0.037	*0.563*	0.977	1.000	*0.368*

* r^2 – correlation coefficient between the predicted and observed activities for the external set; r_0^2 – determination coefficient forcing the origin for (predicted versus observed activities); $r_0'^2$ – determination coefficient forcing the origin (observed versus predicted activities); slopes *k* and *k*' of regression lines (predicted versus observed activities, and observed versus predicted activities) through the origin; underlined values surpass the predictivity thresholds.

The leave-7-out crossvalidation results highlights that the model is stable, not obtained by chance, in fact the difference between $R^2Y(cum)$ and $Q^2(cum)$ is small: 4 % in case of PLS_R model and 3% for the PLS_S model.

The PLS_R model has predictive power, in accordance to the Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC_{EXT} values and all calculated parameters of Golbraikh and Tropsha. PLS_S is less predictive compared to the PLS_R model (see Tables 4 and 5).

Similar fitting and predictivity results were obtained by the PLS_R model, compared to the previously MLR results obtained for the monomers R chiral structures.[3] In addition to the previous MLR results the PLS analysis gave more information on the structural descriptors which influence the flammability. 3D-MoRSE, WHIM, GETWAY, 2D-Autocorrelations and Topological descriptors dominate in the PLS equations. *3D-Morse* (3D-Molecule Representation of Structure based on Electron diffraction) descriptors are based on the idea of obtaining information from 3D atomic coordinates by the function employed in electron diffraction studies to construct scattering curves.[25] *WHIM* (Weighted Holistic Invariant Molecular Descriptors) descriptors are built in order to capture the relevant molecular 3D information regarding the shape, molecular size, symmetry, and atom distribution with respect to some invariant reference frame.[26] *GETAWAY* (GEometry, Topology, and Atom-Weights AssemblY) variables[27] describe the 3D molecular geometry and encode information related to molecular fragments and substituents. The most important among the *2D-autocorrelation* descriptors involved in our model are the Geary coefficients, which are expressed by a distance type function varying from zero to infinity.[28] In the final PLS model, the selected molecular descriptors captures 3D information (3D-Morse, GETAWAY, WHIM), supplying information about interatomic distances, topological distances, types of atoms (2D-Fingerprints, 2D-Autocorrelation).

CONCLUSIONS

Flammability (expressed by the limiting oxygen index) was related to polyphosphonate descriptors (obtained previously by molecular mechanics calculations) using the partial least squares (PLS) approach. Several statistical parameters were employed to test the goodness of fit and predictivity of the resulted models. PLS model obtained for R stereoisomers gave better fitting and predictivity results compared to the S stereoisomers. The significant molecular descriptors (3D-MoRSE, GETAWAY, WHIM, 2D-autocorrelation, etc) that correlate with the limiting oxygen index, comprise different 2D and 3D aspects of the molecular geometry. Compared to the previously published results[3], obtained for the R chiral structures, the PLS model gave similar fitting and predictivity results compared to the MLR model, but included more molecular structural information. The final PLS model brought new insights in the fire retardant mechanism of polyphosphonates and can be used to predict new structures with better flammability.

Acknowledgments: The authors would like to thank to Chemaxon Ltd., for providing a free academic license to their software. This project was financially supported by the Project No. 1.1/2014 of the Institute of Chemistry of Roumanian Academy, Timişoara.

REFERENCES

1. S. Iliescu, A. Pascariu, N. Plesu, A. Popa, L. Macarie and G. Ilia, *Polym. Bull.*, **2009**, *63*, 485-495.
2. M. Makoska, *Pure Appl. Chem.*, **2000**, *72*, 1399-1403.
3. S. Funar-Timofei, S. Iliescu and T. Suzuki, *Struct. Chem.*, **2014**, *25*, 1847-1863.
4. Dragon Professional 5.5, **2007**, Talete S.R.L., Milano, Italy.
5. Instant JChem 5.10.0, **2012**, ChemAxon, http://www.chemaxon.com
6. J. T. Leonard and K. Roy, *SAR Comb. Sci.* **2006**, *25*, 235-251.
7. L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley, New York, 1990.
8. R Development Core Team. A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, **2012**.
9. H. Abdi, "Partial least square regression (PLS regression)", in: N. Salkind, (Ed.), "Encyclopedia of Measurement and Statistics", Sage, Thousand Oaks, CA, 2007, p. 792–795.
10. S. Wold, "PLS for multivariate linear modeling", in: H. van de Waterbeemd, (Ed.), Chemometric Methods in Molecular Design", VCH, Weinheim, 1995, p. 195–218.
11. SIMCA P+ 12.0.0.0, May20 **2008**, Umetrics, Sweeden, http://www.umetrics.com/
12. M. Daszykowski, K. Kaczmarek, V. Heyden and B. Walczak, *Chemom. Intel .Lab. Syst.*, **2007**, *85*, 2007, 203-219.
13. F. Lindgren, B. Hansen, W. Karcher, M. Sjöström and L. Eriksson, *J. Chemometr.*, **1996**, *10*, 521-532.
14. R. Todeschini, V. Consonni and A. Maiocchi, *Chemometr. Intell. Lab.*, **1999**, *46*, 13-29.
15. P. Gramatica, "On the Development and Validation of QSAR Models", in: B. Reisfeld and A. N. Mayeno (Eds.) "Computational Toxicology: Volume II (Methods in Molecular Biology, Vol. 930)", Springer, 2013, p. 499-526.
16. N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, **2011**, *51*, 2320-2335.

17. A. Tropsha and A. Golbraikh, "Predictive Quantitative Structure–Activity Relationships Modeling: Development and Validation of QSAR Models", in: "Handbook of Chemoinformatics Algorithms", J.-L. Faulon and A. Bender (Eds.), Chapman & Hall / CRC, London, 2010, p. 213-233.
18. L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland and D. M. Sheehan, *J. Chem. Inf. Model*., **2001**, *41*, 186-195.
19. G. Schüürmann, R. U. Ebert, J. Chen, B. Wang and R. Kühne, *J. Chem. Inf. Model*., **2008**, *48*, 2140-2145.
20. V. Consonni, D. Ballabio and R. Todeschini, *J. Chem. Inf. Model.*, **2009**, *49*, 1669-1678.
21. P. K. Ojha, I. Mitra, R. N. Das and K. Roy, *Chemom. Intell. Lab. Syst.*, **2011**, *107*, 194-205.
22. P. P. Roy, S. Paul, I. Mitra and K. Roy, *Molecules*, **2009**, *14*, 1660-1701.
23. N. Chirico and P. Gramatica, *J. Chem. Inf .Model*., **2011**, *51*, 2320-2335.
24. P. Gramatica, *QSAR Comb. Sci*., **2007**, *26*, 694-701.
25. L. Saiz-Urra, M. P.Gonzalez and M. Teijeira, *Bioorg. Med. Chem*., **2006**, *14*, 7347-7358.
26. M. P. González, P. L. Suárez, Y. Fall and G. Gómez, *Bioorg. Med. Chem. Lett.*, **2005**, *15*, 5165-5169.
27. V. Consonni, R. Todeschini and R. Pava, *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 682-692.
28. R. Todeschini and V. Consoni, in: R. Mannhold, H. Kubinyi and H. Timmerman (Eds.) "Handbook of molecular descriptors", Wiley- VCH, Weinheim, 2000, p. 17-19, 366, 482, 513.