Prediction of Thermal Decomposition Temperature of Polymers Using QSPR Methods

Davood Ajloo,* Ali Sharifian, and Hossein Behniafar

*School of Chemistry, Damghan University of Basic Science, Damghan, Iran. *E-mail: ajloo@dubs.ac.ir*

Received February 26, 2008

The relationship between thermal decomposition temperature and structure of a new data set of eighty monomers of different polymers were studied by multiple linear regression (MLR). The stepwise method was used in order to variable selection. The best descriptors were selected from over 1400 descriptors including; topological, geometrical, electronic and hybrid descriptors. The effect of number of descriptors on the correlation coefficient (R) and F-ratio were considered. Two models were suggested, one model having four descriptors ($R^2 = 0.894$, $Q^2_{\text{cv}} = 0.900$, $F = 172.1$) and other model involving 13 descriptors ($R^2 = 0.956$, $Q^2_{\text{cv}} = 0.956$, $F = 125.4$).

Key Words : Thermal decomposition temperature, Geometrical descriptors, Semi-empirical, Cross validation, Polymer

Introduction

Nowadays, before synthesizing a new material, database searching can provide information on related reactions to assist in designing viable pathways to synthesize it. In addition, computational chemistry can help to determine whether these pathways are favored from the thermodynamic point of view. Quantitative structure-property relationships (QSPR) studies can help to predict the properties of the compound of interest. The concept of QSPR is to transform searches for compounds with desired properties using chemical institution and experience into a mathematically quantified and computerized form. Once a correlation between structure and property found, any number of compounds, including those not yet synthesized, can be readily screened on computer in order to select structures with the properties desired. Thus, QSPR approach conserves resources and accelerates the process of development of new molecules for use as any purpose.

The different properties of polymer such as thermal stability, viscosity and dielectric constant, have been studied by QSPR.[1-7] Thermal stability of a material is stability against degradation upon exposure to elevated temperatures in an inert environment. Polymers are often exposed to high temperatures during processing and/or use. Thus, thermal stability is among the most important properties of polymers for a wide range of applications.[8] Thermal stability can be expressed with the temperature of ten percent of decomposition $T_{\text{d,10\%}}$, which is defined as temperature at which 10% loss of weight during pyrolysis was recorded by TGA at a heating rate of 10 °C/min.

It can be also described with the molar thermal decomposition function $Y_{\text{d,1/2}}$ that correlates $T_{\text{d,1/2}}$, with an equation: $Y_{\text{d,1/2}} = T_{\text{d,1/2}} \times M$ (M is the molecular weight per repeating unit). $T_{\text{d,1/2}}$ is defined same $T_{\text{d,10\%}}$, only for half decomposition.

Although there are numerous techniques for measuring thermal stability, it is difficulty to interpret the data and/or to compare data obtained in different laboratories or under different test condition.[8,9] Some researchers have found that $\text{Y}_{\text{d,1/2}}$ values for polymers can be estimated on the basis of quantitative structure–property relationship models.[10-13]

One of the best-known examples of the group additive approach is that of van Krevelen.[10] This method provides a rapid and computationally inexpensive approach to the estimation of $Y_{\text{d,1/2}}$ value, but is purely empirical approach and limited to systems composed only of functional groups that have been previously investigated. Furthermore, the group contributions method is only approximate since this approach fails to account for the presence of neighboring groups or conformational influences. Bicerano extended the group additive concept for the prediction of $Y_{\text{d,1/2}}$, based on 21 topological and constitutional descriptors.[8] There are too many descriptors involved in the models though the predictions are good accuracy. In addition, Sun *et al.* obtained two kinds of calculated models on polymeric $Y_{\text{d,1/2}}$ by artificial neural networks for linear chain polymers.[11] After this work, Sun *et al.* further built two models on polymeric $Y_{\text{d,1/2}}$ for the same polymeric data set by fuzzy set theory.[12] The group average method is used to calculate the descriptors for two models, and the connectivity indexes method is used for them.

On the other hand, the quantum chemical descriptors used in QSPR models encode information about the electronic structure of the molecule and thus implicitly account for the cooperative effects between functional groups, charge redistribution, and possible hydrogen bonding in the polymer.[14] The two quantum chemical descriptors, obtained from the monomers of vinyl polymers were used to predict the molar thermal decomposition function $Y_{\text{d,1/2}}$. A more physically meaningful quantitative structure-property relationship (QSPR) model obtained from the training set with multiple linear stepwise regression analysis.[7]

The goal is to find one equation that is function of a small number of structure-based molecular descriptors that accurately predicts the experimental property. We must balance between increasing in number of molecular descriptors in the selected model and increasing in accuracy of it. This paper is the first time to produce a robust QSPR model that could predict $T_{d,10\%}$ values for polymers with different structure and family by using 1497 descriptors in 18 groups, which are calculated from the monomers of polymers by PM3 semi-empirical method. The effect of number of descriptors on statistical parameters has been analysed.

Material and Methods

Data set. One of the steps necessary for this research was the assembly of a data set. A number of structurally heterogeneous addition linear polymers were selected. The reported experimental data have been taken from Refs. [15-25]. The Thermal stability evaluation of the resulting polymers was carried out by TGA/DTG in nitrogen at a heating rate of 10 °C/min. The thermograms of these polymers were almost similar with each other.

It is impossible to calculate descriptors directly for an entire molecule because all the polymers have wide distribution of molecular weight and possess high molecular weight. As we know, if the molecular weight is high enough, the terminal groups hold only a very small proportion in a polymer and its effect on the properties can be ignored. Molecular descriptors calculated directly from the structure of the repeat units can be used for the study of QSPRs for polymers, since all the properties depends on the chemical structure of the polymer molecule, and all this structure was conditioned by the repeat unit structure. Therefore, we adopt this method and focus on the following model to calculate molecular descriptors. The structures for polymers were endcapped with last group of opposing side. In the next step, the molecular structure of monomer compounds used in the polymerization of polymers, were used to determine the molecular descriptors of polymers. After providing the data set, all molecules (monomer of polymer) were drawn into Hyperchem software[26] and optimized using the PM3 semi-empirical method[4,27] in Hyperchem and thereafter, topological, geometric, and electronic features of each molecule were calculated by the descriptor generation routines Dragon.[28] Topological descriptors provide information about molecular shape and connectivity. Examples of these include the electrotopological state indices and molecular distance edge information geometric descriptors encode features such as shapes shadow projections, and solvent accessible surface areas. Electronic descriptors describe the electronic environment that exists within the molecule such as partial charges on each atom. There are also several hybrid descriptors, which combine several aspects of molecules molecular descriptors were calculated from the molecular structure of molecules by means of Dragon software.[28]

Stepwise multiple regression. In QSPR studies, the goal is to find one or more equations that are functions of a small number of structure based molecular descriptors that accurately predict the experimental property. As it is possible to generate a large number of molecular descriptors for each compound in the data set, the problem becomes how to efficiently select the set of molecular descriptors that yield an accurate relationship. If a method is sufficiently fast, it can be used to generate relationships with different numbers of molecular descriptors. The most commonly used method for this problem is the stepwise approach, which can run forward or backward.

Stepwise method was used to select the most appropriate descriptors. It is a popular technique that combines forward selection and backward elimination. It is essentially a forward selection approach, but at each step it considers the possibility of deleting a variable as in the backward elimination approach, provided that the number of model variables is greater than two. The two basic elements of the stepwise method are forward selection and backward elimination. In the forward selection the variable considered for inclusion at any step is the one yielding the largest single degree of freedom F-ratio among the variables that are eligible for inclusion. Consequently, at each step, the j^{th} variable is added to a k-size model if

$$F_j = \max_j\left(\frac{RSS_k - RSS_{k+j}}{s^2_{k+j}}\right) > F_{in} \qquad (1)$$

In the above inequality RSS is the residual sum of squares and s is the mean square error. The subscript $k + j$ refers to quantities computed when the j^{th} variable is added to the k variables that are already included in the model. In backward elimination the variable considered for elimination at any step is the one yielding the minimum single degree of freedom F-ratio among the variables that are included in the model. The variable is eliminated only if the corresponding F-ratio does not exceed a specified value F_{out}. Consequently, at each step, the j^{th} variable is eliminated from the k-size model if

$$F_j = \min_j\left(\frac{RSS_{k-j} - RSS_k}{s^2_k}\right) > F_{out} \qquad (2)$$

The subscript k-j refers to quantities computed when the j^{th} variable is eliminated from the k variables that have been included in the model so far. The variable selection was done by stepwise method which implemented in SPSS software.

Cross-validation technique. The reliability of the proposed method was explored using the cross-validation method. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects.[29-31] For each data set, an input-output model is developed, based on the utilized modeling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the leave-one-out (LOO) procedure was utilized in this study, which produces a number of models, by deleting each time one object from the training set. Obviously, the number of models produced by the LOO procedure is equal to the number of available examples n.

Prediction error sum of squares (*PRESS*) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the *PRESS* and *SSY* (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2 and $S_{\rm PRESS}$ values can be easily calculated. The formulae used to calculate all the aforementioned statistics are presented below (Eqs. (3) and (4)):

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n}(y_{\rm exp} - y_{\rm pred})^2}{\sum_{i=1}^{n}(y_{\rm exp} - \bar{y})^2} \tag{3}$$

$$S_{PRESS} = \sqrt{\frac{PRESS}{n}} \tag{4}$$

For a more exhaustive testing of the predictive power of the model, except from the classical LOO cross-validation technique, validation of the model was carried out by a leave one-out (LOO) cross validation procedure. From the training set we randomly selected one compound. Each group was left out and that group was predicted by the model developed from the remaining observations. This procedure was carried out several times.

Estimation of the predictive ability of a QSPR model. According to Tropsha[32] the predictive power of a QSPR model can be conveniently estimated by an external $R^2_{\rm cv;ext}$ (Eq. (5)).

$$R^2_{\rm cv,ext} = 1 - \frac{\sum_{i=1}^{test}(y_{\rm exp} - y_{\rm pred})^2}{\sum_{i=1}^{test}(y_{\rm exp} - \bar{y}_{tr})^2} \tag{5}$$

where $\bar{y}_{tr}$ is the averaged value for the dependent variable for the training set. Furthermore, the same group[32,33] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2_{\rm cv,ext} > 0.5 \tag{6}$$

$$R^2 > 0.5 \tag{7}$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \quad \text{or} \quad \frac{R^2 - R'^2_0}{R^2} < 0.1 \tag{8}$$

Table 1. List of monomers of polymers* used in this research and their thermal decomposition temperature $T_{\rm d,10\%}$.

No	Monomers used to prepare the polymers	$T_{\rm d,10\%}$
	Training set	
1	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and *m*-phenylenediamine	541
2	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and N-3,5-diaminophenylphthalimide	554
3	1,3-bis(N-trimellitoyl)benzene and *m*-phenylenediamine	528
4	1,3-bis(N-trimellitoyl)benzene and N-3,5-diaminophenylphthalimide	537
5	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and resorcinol	455
6	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and hydroquinone	461
7	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 4,4'-dihydroxybiphenyl	473
8	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 2,2'-dihydroxybiphenyl	477
9	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 2,2'-dimethyl-4,4'-dihydroxybiphenyl	458
10	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and bisphenol-A	451
11	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 2,7-dihydroxynaphthalene	460
12	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 1,5-dihydroxynaphthalene	468
13	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and *m*-phenylenediamine	516
14	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and 2,6-diaminopyridine	520
15	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and 4,4'-diaminophenylene ether	528
16	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and *N*-(3,5-diaminophenyl)phthalimide	528
17	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and 2,2'-bis(4-aminophenoxy)-1,1'-binaphthyl	543
18	N-[3,5-bis(3,4-dicarboxybenzamido)phenyl]phthalimide dianhydride and 4-*p*-biphenyl-2,6-bis(4-aminophenyl)pyridine	539
19	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and *m*-phenylenediamine	506
20	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and 2,6-diaminopyridine	508
21	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and 4,4'-diaminophenylene ether	513
22	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and *N*-(3,5-diaminophenyl)phthalimide	516
23	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and 2,2'-bis(4-aminophenoxy)-1,1'-binaphthyl	522
24	1,3-bis(3,4-dicarboxybenzamido)benzene dianhydride and 4-*p*-biphenyl-2,6-bis(4-aminophenyl)pyridine	518
25	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and resorcinol	566
26	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and hydroquinone	575
27	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and 4,4'-dihydroxybiphenyl	573
28	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and 2,2'-dihydroxy-1,1'-binaphthyl	577
29	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and *m*-phenylene diamine	579

Table 1. continued

No	Monomers used to prepare the polymers	$T_{d,10\%}$
30	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and *p*-phenylene diamine	594
31	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and benzidine	596
32	2,2'-bis(N-trimellitoyl)-1,1'-binaphthyl and 2,2'-diamino-1,1'-binaphthyl	604
33	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and resorcinol	398
34	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 4,4'-dihydroxybiphenyl	410
35	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and bisphenol-A	406
36	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,2'-dihydroxybiphenyl	392
37	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,2'-dihydroxy-1,1'-binaphthyl	416
38	pyromellitic dianhydride and *m*-phenylene diamine	600
39	pyromellitic dianhydride and 4,4'-diaminophenylene ether	528
40	3-trimellitimidobenzoic acid and *m*-phenylene diamine	540
41	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and *m*-phenylene diamine	495
42	3,6-dimethyl-1,4-bis(methyleneisocyanate) and benzidine	340
43	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and N-3,5-Diaminophenylphthalimide	554
44	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and *m*-phenylene diamine	541
45	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and resorcinol	455
46	2,2'-bis(4-trimellitimidophenoxy)biphenyl and *m*-phenylene diamine	507
47	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 2,6-diaminopyridine	511
48	2,2'-bis(4-trimellitimidophenoxy)biphenyl and *p*-phenylene diamine	520
49	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 4,4'-diaminophenylene ether	531
50	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 4-phenyl-2,6-bis(4-aminophenyl)pyridine	545
	Testing set	
51	2,2'-bis(4-trimellitimidophenoxy)-1,1'-binaphthyl and *m*-phenylene diamine	512
52	2,2'-bis(4-trimellitimidophenoxy)-1,1'-binaphthyl and 2,6-diaminopyridine	517
53	2,2'-bis(4-trimellitimidophenoxy)-1,1'-binaphthyl and *p*-phenylene diamine	524
54	2,2'-bis(4-trimellitimidophenoxy)-1,1'-binaphthyl and 4,4'-diaminophenylene ether	535
55	4-phenyl-2,6-bis(4-trimellitimidophenyl)pyridine and *m*-phenylene diamine	485
56	4-phenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,6-diaminopyridine	502
57	4-phenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 4-phenyl-2,6-bis(4-aminophenyl)pyridine	516
58	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and *m*-phenylene diamine	485
59	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,6-diaminopyridine	510
60	2,2'-bis(4-trimellitimidophenoxy)biphenyl and resorcinol	453
61	2,2'-bis(4-trimellitimidophenoxy)biphenyl and hydroquinone	467
62	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 2,2'-dihydroxybiphenyl	458
63	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 4,4'-dihydroxybiphenyl	470
64	2,2'-bis(4-trimellitimidophenoxy)biphenyl and bisphenol-A	449
65	2,2'-bis(4-trimellitimidophenoxy)biphenyl and 2,2'-dihydroxy-1,1'-binaphthyl	484
66	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and resorcinol	448
67	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and hydroquinone	452
68	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 4,4'-dihydroxybiphenyl	464
69	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,2'-dihydroxybiphenyl	449
70	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,2'-dimethyl-4,4'-dihydroxybiphenyl	441
71	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and bisphenol-A	433
72	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,7-dihydroxynaphthalene	459
73	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 1,5-dihydroxynaphthalene	455
74	4-*p*-biphenyl-2,6-bis(4-trimellitimidophenyl)pyridine and 2,2'-dihydroxy-1,1'-binaphthyl	471
75	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and resorcinol	391
76	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and hydroquinone	399
77	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 4,4'-dihydroxybiphenyl	404
78	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 2,2'-dihydroxybiphenyl	403
79	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and bisphenol-A	385
80	N-[3,5-bis(N-trimellitoyl)phenyl]phthalimide and 2,2'-dihydroxy-1,1'-binaphthyl	412

*Data were taken from references 15-25

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{9}$$

Mathematical definitions of R_0^2, R'^2_0, k and k' are based on regression of the observed activities against predicted activities and vice versa (regression of the predicted activities against observed activities). The definitions are presented clearly in Golbraikh *et al.*[33]

Results and Discussion

Table 1 shows data set for 80 different groups including poly(amide-imide), poly(ester-imide) and poly(urethane-imide) of polymers and their thermal decomposition temperature which are taken from Refs. [15-25].

The data of $T_{\mathrm{d,10\%}}$ in Table 1 are divided into a training set and a validation set. The former includes 50 polymeric (1-50) $T_{\mathrm{d,10\%}}$ values; while the latter with 30 polymers (51-80) data are obtained from the experimental data of $T_{\mathrm{d,10\%}}$. Training set is used to build a QSPR model, which is evaluated with the validation set. It is impossible to calculate descriptors directly for an entire molecule because all polymers have wide distribution of molecular weights and possess high molecular weights. Molecular descriptors calculated directly from the structure of the monomers can be used on the study of QSPR models for polymers, since all the properties depend on the chemical structure of the polymer molecule, and all this structure is conditioned by the monomer structure.

Stepwise method selects the descriptors step-by-step. The best 13 selected descriptors among 1497 descriptors were tabulated in Table 2. They are belongs to different class of descriptors such as autocorrelation, 3D-MoRSE, topological, shape and aromaticity.[28]

Molecular descriptors based on the autocorrelation function AC_l, defined as:

$$AC_l = \int_a^b f(x) \cdot f(x+1) \cdot dx \tag{10}$$

Where $f(x)$ is any function of variable x and l is the lag representing an interval of x; a and b defined the total studied interval of the function. The function $f(x)$ is usually time-dependent function such as a time dependent electrical signal, or a spatial-dependent function such as the population density in space. Function AC_l is a summation of the function value products calculated at x and $x + l$. To obtain spatial autocorrelation molecular descriptors, function $f(x)$ is any physico-chemical property calculated for each molecule, such as atomic mass, polarizability, etc. Therefore, the molecule atom represents the set of discrete points in space and the atomic property evaluated at those points. *AC* descriptors can also be calculated for 2D and 3D-spatial molecular geometry. In 2D case the distribution of a molecular property can be a mathematical function $f(x, y)$ for each point. The most well known autocorrelation descriptors are a) Moran coefficient ($I(d)$) and b) Geary coefficient $c(d)$.

General index of spatial Moarn autocorrelation that, if applied to a molecular graph, can be defined as:

$$I(d) = \frac{\frac{1}{\Delta}\sum_{i=1}^{A}\sum_{j=1}^{A} \delta_{ij} \cdot (w_i - \bar{w})(w_j - \bar{w})}{\frac{1}{A} \cdot \sum_{i=1}^{A} (w_i - \bar{w})^2} \tag{11}$$

In addition Geary descriptors are defined as:

$$c(d) = \frac{\frac{1}{2\Delta}\sum_{i=1}^{A}\sum_{j=1}^{A} \delta_{ij}(w_i - w_2)^2}{\frac{1}{(A-1)} \cdot \sum_{i=1}^{A} (w_i - \bar{w})^2} \tag{12}$$

Where w_{i} is any atomic property, $\bar{w}$ is its average value on the molecule. A is the atom number, d is the considered topological distance (*i.e.* the lag in autocorrelation terms), δ_{ij} is a Kronecker delta ($\delta_{\mathrm{ij}} = 1$ if $d_{\mathrm{ij}} = d$, zero otherwise), Δ is the sum of the Kronecker deltas, *i.e.* the number of vertex pairs at distance equal to d.[28] The Moran coefficient usually takes a values in the interval [−1, +1]. Positive autocorrelation corresponds to positive values of the coefficient

Table 2. Selected descriptors by stepwise method

Symbol	Definition	Class
MATS2e	Moran autocorrelation – lag 2 / weighted by atomic Sanderson electronegativities	2D autocorrelations
GATS1v	Geary autocorrelation – lag 1 / weighted by atomic van der Waals volumes	2D autocorrelations
GATS8e	Geary autocorrelation – lag 8 / weighted by atomic Sanderson electronegativities	2D autocorrelations
HATS5v	leverage-weighted autocorrelation of lag 5 / weighted by atomic van der Waals volumes	GETAWAY descriptors
MATS8p	Moran autocorrelation – lag 8 / weighted by atomic polarizabilities	2D autocorrelations
Mor28u	3D-MoRSE – signal 28 / unweighted	3D-MoRSE descriptors
BEHm2	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses	BCUT descriptors 132
DECC	Eccentric	topological descriptors
AROM	aromaticity (trial)	aromaticity indices
Mor19p	3D-MoRSE – signal 19 / weighted by atomic polarizabilities	3D-MoRSE descriptors
PJI2	2D Petitjean shape index	topological descriptors
Mor29e	3D-MoRSE – signal 29 / weighted by atomic Sanderson electronegativities	3D-MoRSE descriptors
Mor27m	3D-MoRSE – signal 27 / weighted by atomic masses	3D-MoRSE descriptors

whereas negative autocorrelation produces negative values.

The Geary coefficient, $c(d)$[28] is a distance type function varying from zero to infinity. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates into values between 0, and 1 whereas negative autocorrelation produce values larger than 1; therefore the reference "no correlation" is $c = 1$.

3D-MoRSE descriptors (3D-Molecule Representation of Structure based on Electron diffraction) are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing the practical scattering curves.[28] A generalized scattering function called the molecular transform, can be used as the functional basis for deriving, from a known molecular structure, the specific analytic relationship at both X-ray and electron diffraction.

Molecular eccentricity (ε): Among the eigenvalue-based descriptors, molecular eccentricity is a shape descriptor obtained from the eigenvalues, λ, of the inertia matrix I defined as follow[28]:

$$\varepsilon = \frac{(\lambda_1^2 - \lambda_3^2)^{1/2}}{\lambda_1} \quad 0 \le \varepsilon \le 1 \tag{13}$$

Where $\varepsilon = 0$ corresponds to spherical top molecules and $\varepsilon = 1$ to linear molecules.

It is a shape descriptor defined by analogy with the eccentricity of ellipse, which is defined:

$$\varepsilon = \frac{(l_M^2 - l_m^2)^{1/2}}{l_M} \tag{14}$$

Where l_M and l_m are the lengths of the major and minor elliptical axes, respectively.

BCUT descriptors: defined as eigenvalues of a modified connectivity matrix, which could be called Burden matrix B.[28] The B matrix is defined as the following: The diagonal elements B_ii are the atomic number Z_i of the atoms; the off-diagonal elements B_ij representing two bonded atoms i and j are equal to $\pi^*.10^{-1}$ where π^* is conventional bond order, *i.e.* 0.1, 0.2, 0.3 and 0.15 for single, double, triple and aromatic bond respectively; off-diagonal elements B_ij corresponding to terminal bonds are augmented by 0.01; all other matrix elements are set a 0.001.

One aim of this work is also studying the effect of number of descriptors on the some statistical parameters such as correlation coefficient (R) and F ratio (F). The stepwise method in SPSS automatically selects descriptors. In each step, one descriptor is added. In this method 13 descriptors were selected step by step. Each step usually named as a model. Figure 1 shows the effect of number of descriptors on the R and F variations. Increasing the number of descriptors increases the R and F up to four descriptors. After that, we see increment of R and decrement of F up to thirteenth descriptor. It must be noted that after inclusion of fourth descriptor the variation of R is not significant (the curve is smooth and has lower slope). Thus we have two regions, one before fourth descriptors another after it. We named for simplicity the model which having four descriptors as model 1 and model which having thirteen descriptors as model 2. Before model 1 we have lower R after that we have lower F. Thus model 1 which has highest F and lowest number of descriptors and regional R is the best model.

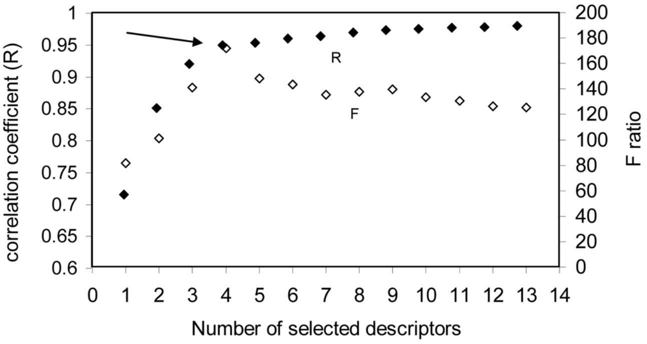


Figure 1. Effect of number of descriptors on the correlation coefficient (R) and F-ratio.

The effect of number of descriptors on the correlation coefficient (R) and F-ratio (F) shown in Figure 1. It is observed that R is increased due to increasing the number of descriptors. But, F-ratio increases up to descriptor 4 and then be decays downward to descriptor 13. The optimum model is that has higher R, F and smaller number of descriptors. Thus, in this work we selected two models, one model including four descriptors (model 1) and other including thirteen descriptors (model 2).

By carrying out the correlation between the molar thermal

Table 3. Coefficient of each descriptor and related error associated with their mean effects

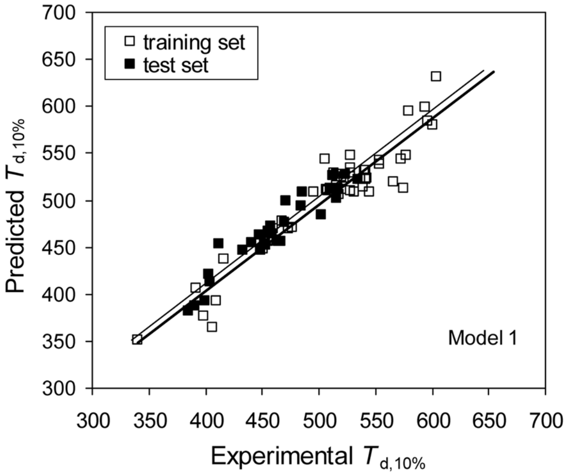
	Model 1			
	coefficient	Std. Error	Mean effect	Sig.
(Constant)	1314.393	67.220		3.6E-31
MATS2e	−460.310	45.976	−0.417	1.81E-15
GATS1v	−1061.106	83.345	−0.505	1.99E-20
GATS8e	115.397	13.055	0.323	3.02E-13
HATS5v	988.222	154.175	0.247	1.16E-08
	Model 2			
	coefficient	Std. Error	Mean effect	Sig.
(Constant)	4408.825	523.903		4.79E-12
MATS2e	−421.865	40.779	−0.382	1.9E-15
GATS1v	−1070.665	91.227	−0.510	8.28E-18
GATS8e	131.712	15.508	0.369	3.48E-12
HATS5v	495.247	148.470	0.124	0.0014
MATS8p	−245.815	66.738	−0.167	0.000466
Mor28u	30.794	5.329	0.209	2.23E-07
BEHm2	−502.815	75.351	−0.310	6.15E-09
DECC	17.355	4.058	0.242	6.26E-05
AROM	−1045.937	260.139	−0.202	0.000152
Mor19p	−18.892	6.882	−0.100	0.007784
PJI2	−136.303	49.755	−0.070	0.007907
Mor29e	−14.730	5.978	−0.101	0.016354
Mor27m	−16.625	6.917	−0.112	0.019074

decomposition function $T_{d,10\%}$ (in Table 1) with four and thirteen descriptors using stepwise regression analysis, two models obtained (Table 3). Table 3 includes coefficients of two equations for both models. Coefficient of each descriptor, related mean effect and standard error for them are listed in the table. Negative coefficient shows the reverse trend of that descriptor and thermal decomposition temperature. Mean effect represents the effect of each descriptors on $T_{d,10\%}$

Present version of Dragon software calculates 1497 descriptors in 18 different classes belong to atomic constitution, atomic correlation, molecular topology, steric properties and so on. All of them are physico-chemical properties. The selected physico-chemical properties meaning defined in the above by formula. But they may be not popular as well as molecular weight, volume, polarity and…. Correlation between $T_{d,10\%}$ and these selected descriptors were taken in Table 3. We can write it as following equation for model 1:

$$T_{d,10\%} = -460.310(MATS2e) - 1061.106(GATS1v) + 115.397(GATS8e) + 988.222(HATS5v) + 1314.393 \quad (15)$$

These four variables were briefly defined in the Table 2 and discussed with more detailed in the above. Negative and positive coefficients make decreases and increases $T_{d,10\%}$, respectively. $T_{d,10\%}$ and more customary descriptors have weak correlation, so that they have not been selected in here. Also some of reports in the literature used descriptors which are similar to our descriptors and were not discussed detailed definition of descriptors.[34]



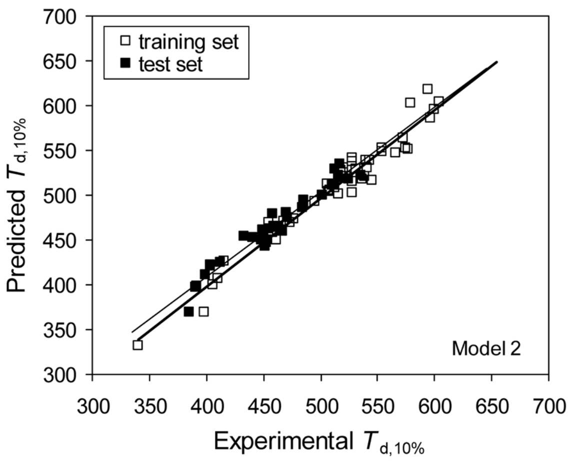


Figure 2. Predicted versus experimental thermal decomposition temperature ($T_{d,10\%}$) for two models; (a) model 1 including four descriptors and (b) model 2 including thirteen descriptors. Narrow and bold tredlines are related to testing and training set, respectively.

Variables in the model 1 and 2, are topological or geometrical. Variables in model 1 are two dimensional descriptors, namely they depend on two variable coordinates such as x and y or x and z or y and z. While variables in model 2 are two and three dimensional.

Two dimensional properties may be due to surface dependent properties, such as aromaticity. On the other hand, some part of our sample molecules have aromatic nature. Thus, significantly appearance of two dimensional properties may be due necessity for coplanarity.

The predicted versus experimental $T_{d,10\%}$ values for model 1 and 2 are shown in Figure 2. It is clearly represented that both model are satisfactory and model 2 has slightly higher correlation coefficient relative to model 1. But model 1 has lower higher F and lower number of descriptors. Thus model 1 is the better model.

The Leave-One-Out (LOO) cross-validation method is adopted to test the internally predictive ability for Eq. (3) for two models. The cross-validation $Q^2_{cv;int}$ value for medel 1 is 0.885. Then the model is evaluated externally using the test set of 30 polymers (in Table 1). The cross-validation $Q^2_{cv;ext}$ value for the external test set is 0.884. Coefficients of determination R_0^2 (predicted versus observed $T_{d,10\%}$ values) and R_0 (observed versus predicted $T_{d,10\%}$ values) are 0.885 and 0.893, respectively. The slopes of regression lines through the origin of predicted versus observed k and observed versus predicted k' values are 1.013 and 0.986, respectively. According to Refs. [35-36], a QSPR model is successful if it satisfied several criteria as follows: (1) $Q^2_{cv;int} > 0.5$, $Q^2_{cv;ext} > 0.5$, $R^2 > 0.6$; (2) R_0^2 and R'^2_0 close to R^2, (R^2

Table 4. Statistical parameters for two models

Statistics	Model 1	Model 2
No. of descriptors	4	13
R^2	0.894	0.956
F	172.1	125.4
N	80	80
Sig	5.66E-37	3.14E-41
Std. Error	18.6	12.5
Q^2	0.900	0.956
$Q^2_{cv,int}$	0.886	0.956
$Q^2_{cv,ext}$	0.884	0.931
$R_0{}^2$	0.893	0.941
$(R^2–R_0{}^2)/R^2$	0.001	0.016
$R'_0{}^2$	0.893	0.941
$(R^2-R'_0{}^2)/R^2$	0.001	0.016
k	1.013	1.011
$R^2_{cv,ext}$	0.938	0.965
k'	0.986	0.988

$- R_0^2)/R^2 < 1$, $(R^2 - R'^2_0)/R^2 < 0.1$ (3) k and k' are between 0.85 and 1.15. The cited statistical parameters for models 1 and 2 were listed in Table 4. Thus, in model 1

$$T_{d,10\%}\text{ (pred)} = 0.925\ T_{d,10\%}\text{ (exp)} + 41.04\ n = 80\ F = 172$$
$$S = 18.6 \quad R^2 = 0.894 \tag{16}$$

And in model 2

$$T_{d,10\%}\text{ (pred)} = 0.950\ T_{d,10\%}\text{ (exp)} + 28.39\ n = 80\ F = 125$$
$$S = 12.5\ R^2 = 0.956 \tag{17}$$

It is obvious that our results satisfy the generally accept condition. Therefore, the QSPR model obtained in this paper is suitable for making accurate prediction outside of the training set of polymers. According to the Sig.-test (see Table 2), the descriptors appearing in correlation equations are significant descriptors. The standard errors of estimation in Eq. (5) for model 1 and 2 are 18.6 *K* and 12.5, respectively.

Finally, an overview on the total results, shows that, increasing the number of descriptors increases the *R*, but *F* up to fourth descriptors then decreases after that descriptor. Thus, model 1 is optimum model which have lower number of descriptors and higher *F* and acceptable *R*.

Conclusion

The QSPR is a method in order to prediction of polymer properties. In this work, a number of descriptors have been selected among a large number of descriptors. The selected descriptors were belongs to different class of descriptors such as: topological, electronic, geometrical and hybrid descriptors. The effect of number of descriptors also was addressed. On the basis of this effect, two models were presented. The model which having less number of descriptors, has lower correlation coefficient, lower cross validation and higher *F*-ratio. The optimum model is one which has higher *R,* higher *F* and lower number of descriptors. Thus, It seems those model which having four descriptors is optimum and better model.

Acknowledgments. Financial support of Damghan University of Basic Science and very useful comments from Dr. Bahram Hemmtinejad are acknowledged.

References

1. Mattioni, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232.
2. Liu, A.; Wang, X.; Wang, L.; Wang, H.; Wang, H. *Eur. Polym. J.* **2007**, *43*, 989.
3. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Polymer* **2006**, *47*, 3240.
4. Gharagheizi, F. *Comp. Mater. Sci.* **2007**, *40*, 159.
5. Xua, J.; Chenb, B.; Zhang, Q.; Guob, B. *Polymer* **2004**, *45*, 8651.
6. Yu, X.; Yi, B.; Xie, Z.; Wang, X.; Liu, F. *Chemometr. Intell. Lab.* **2007**, *87*, 247.
7. Yu, X.; Xie, Z.; Yi, B.; Wang, X.; Liu, F. *Eur. Polym. J.* **2007**, *43*, 818.
8. Bicerano, J. *Prediction of Polymer Properties*. 2nd ed.; Marcelar Dekker Inc.: New York, 1996.
9. Flynn, J. H. *J. Therm. Anal.* **1988**, *34*, 367.
10. van Krevelen, D. W. *Properties of Polymers.* 3rd ed.; Elsevier: Amsterdam, 1990.
11. Sun, H.; Tang, Y. W.; Wu, G. S.; Zhang, F. S.; Chen, X. Q. *Comput. Appl. Chem.* **2003**, *20*, 210.
12. Sun, H.; Tang, Y. W.; Wu, G. S.; Zhang, F. S.; Chen, X. Q. *Comp. Appl. Chem.* **2003**, *20*, 381.
13. Yu, X.; Xie, Z.; Yi, B.; Wang, X.; Liu, F. *Eur. Polym. J.* **2007**, *43*, 818.
14. Katritzky, A. R.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *8*, 1171.
15. Behniafar, H.; Haghighat, S. *Eur. Polym. J.* **2006**, *42*, 3236.
16. Behniafar, H.; Akhlaghinia, B.; Habibian, S. *Eur. Polym. J.* **2005**, *41*, 1071.
17. Behniafar, H. *J. Appl. Polym. Sci.* **2006**, *101*, 869.
18. Behniafar, H.; Khaje-Mirzai, A. A.; Beit-Saeed, A. *Polym. Int.* **2007**, *56*, 74.
19. Banihashemi, A.; Behniafar, H. *Polym. Int.* **2003**, *52*, 1136.
20. Behniafar, H.; Habibian, S. *Polym. Int.* **2005**, *54*, 1134.
21. Behniafar, H.; Banihashemi, A. *Eur. Polym. J.* **2004**, *40*, 1409.
22. Behniafar, H.; Jafari, A. *J. Appl. Polym. Sci.* **2006**, *100*, 3203.
23. Behniafar, H. *J. Macromol. Sci. Pur. Appl. Chem.* **2006**, *43*, 813.
24. Behniafar, H.; Haghighat, S.; Farzaneh, S. *Polymer* **2005**, *46*, 4627.
25. Behniafar, H.; Banihashemi, A. *Polym. Int.* **2004**, *53*, 2020.
26. Motamedi, M.; Bathaie, S. Z.; Hemmateenejad, B.; Ajloo, D. *J. Mol. Structure (Theochem)* **2004**, *678*, 163.
27. *HyperChem Release 7.5 for Windows*, Molecular Modeling System; Hypercube, Inc.: 2002.
28. (a) Todechini, R.; Consoni, V. *Handbook of Molecular Descriptors*; Wiley VCH: 2000. (b) Talete srl, *Dragon for Windows* (Software for molecular Descriptor Calculations), Version 5.4-2006.
29. Efron, B. *J. Am. Stat. Assoc.* **1983**, *78*, 316.
30. Efroymson, M. A. *Multiple Regression Analysis* in *Mathematical Methods for Digital Computers*; Ralston, A., Wilf, H. S., Eds.; Wiley; New York, 1960.
31. Osten, D. W. *J. Chemom.* **1998**, *2*, 39.
32. Tropsha, A.; Gramatica, P.; Gombar, V. K. *Quant. Struct. Activ. Relat.* **2003**, *22*, 1.
33. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Model* **2002**, *20*, 269.
34. Brian, E. M.; Peter, C. J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232.
35. Efron, B. *J. Am. Stat. Assoc.* **1983**, *78*, 316.
36. Efroymson, M. A. *Multiple Regression Analysis* in *Mathematical Methods for Digital Computers*; Ralston, A., Wilf, H. S., Eds.; Wiley: New York, 1960.