A Simple Three-Descriptor Model for the Prediction of the Glass-Transition Temperatures of Vinyl Polymers

Xinliang Yu,[1] Wenhao Yu,[2] Xueye Wang[3]

[1]*Department of Chemistry and Chemical Engineering, Hunan Institute of Engineering, Xiangtan, Hunan 411104, China*
[2]*School of Resource and Environmental Science, Wuhan University, Wuhan, Hubei 430079, China*
[3]*College of Chemistry, Xiangtan University, Xiangtan, Hunan 411105, China*

Received 29 March 2009; accepted 11 September 2009
DOI 10.1002/app.31423
Published online 13 November 2009 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: An artificial neural network (ANN) implementing a back-propagation algorithm was applied for the prediction of the glass-transition temperature (T_g) values of 84 polyacrylates and 21 polyvinyls. The experimental T_g data of the polymers were divided into a training set (50 polyacrylates) and a testing set (34 polyacrylates and 21 polyvinyls). Three molecule descriptors (mean atomic van der Waals volume, bond information content, and three-dimensional molecule representation of structures based on electron diffraction descriptor for signal 13/weighted by atomic masses, Mor13m) were used as input parameters of the neural network. Simulated with the optimum back-propagation ANN 3-[3-2]-1, the root mean square (rms) error for the testing set was 17.7 K, and the correlation coefficient was 0.942, which were accurate in comparison with existing models. The ANN model could be used not only to reveal the quantitative relation between T_g and the molecular structure but also to predict the T_g values of the polyacrylates and polyvinyls. © 2009 Wiley Periodicals, Inc. J Appl Polym Sci 115: 3721–3726, 2010

Key words: glass transition; modeling; structure–property relations

INTRODUCTION

The glass-transition temperature (T_g) is the temperature at which the amorphous phase of a polymer is converted between the rubbery and glassy states.[1] Below the T_g, amorphous polymers are in a glassy state, and most of their joining bonds are intact. Above T_g, polymers become soft and capable of plastic deformation without fracture. T_g is the most important and widely studied property of polymeric and composite materials.[2,3] In fact, T_g determines the temperature windows for the processing and use of these material and is a prerequisite for the prediction and understanding of the mechanical and other properties, such as hardness, modulus, heat capacity, coefficient of thermal expansion, and viscosity.

T_g is a kinetic parameter and, thus, parametrically depends on the melt cooling rate. The slower the melt cooling rate is, the lower T_g is. In addition, T_g depends on the measurement conditions, which are not universally defined. Hence, the development of theoretical methods for the prediction of T_g is needed and is interesting.

The quantitative structure–property relationship (QSPR) method has been reported quite extensively in the literature to predict T_g for polymers. van Krevelen[4] correlated T_g with the group additive properties method. The group additive properties method is a purely empirical approach and limited to systems composed only of functional groups that have been previously investigated. Chen et al.[5] introduced a comprehensive neural network model with 28 group descriptors. A network trained with 65 polymers was tested with 6 polymers and had a rms errors of 17 K ($R^2 = 0.95$) for the training set and 17 K ($R^2 = 0.85$) for the prediction set. The model was accurate, but the ratio between the samples (65) and descriptors (28) was 2.32, which was not equal to or more than the minimum ratio of 5 to 1.

Bicerano[6] related T_g with the solubility parameter and 13 structural parameters for a data set of 320 polymers and produced a regression model with a standard error of 24.65 K. However, he did not use external data set compounds to validate this model. Yu et al.[7] developed a linear model with only two descriptors. The model was tested to be accurate, with correlation coefficients of 0.953 (rms = 25.0 K) for the training set and 0.952 (rms = 20.8 K) for the test set. In addition, Katritzky et al.[8] introduced the comprehensive descriptors for structural and statistical analysis (CODESSA) method to predict T_g for 88 linear homopolymers with five parameters and

Additional Supporting Information may be found in the online version of this article.

Correspondence to: X. Yu (yxliang5602@sina.com.cn).

generated a QSPR model with a standard error of 32.9 K. Mattioni and Jurs[9] developed a 10-descriptor model and an 11-descriptor model to predict T_g values for two diverse sets of polymers. The test sets rms errors of the two models were more than 21 K.

For most models discussed previously, the molecular structures were optimized with semiempirical quantum chemical methods. Furthermore, a minuscule amount of descriptors were calculated for each molecule. In this study, the QSPR method was applied to predict the T_g of 105 polyacrylates and polyvinyls with an artificial neural network (ANN) model. The molecular descriptors used to describe the structure of the polymers were extracted from the monomers of the polymers with Dragon software.[10] The monomer structures were optimized with density functional theory (DFT) at the Becke-3-parameter-Lee-Yang-Parr (B3LYP) level of theory with a 6-31G(d,p) basis set, which means 6 primitive Gaussian type orbital (GTO) for core electrons, 3 for inner and 1 for outer valence orbitals, 1 d polarization functions added to heavy atoms and 1 p polarization functions added to H atom. A total of 1664 molecular descriptors were calculated for every molecule.

OVERVIEW OF THE ANN ARCHITECTURE

An ANN[11–13] is a highly simplified model of the biological structures found in a human brain. In recent years, the use of ANNs has become a very popular and powerful chemometric tool to solve chemical problems. ANN models are organized into a layered structure, formed by one input layer, one output layer, and at least one hidden layer. The input layer receives input data (molecular descriptors), the hidden layer performs processing and transformation of the input data, and the output layer relays the final results (T_g values; Fig. 1). Each layer has different numbers of neurons (or nodes). Each neuron has weighted inputs, a transfer function, and one output. Thus, each neuron is essentially an equation that balances inputs and outputs.

The behavior of a neural network is determined by the transfer functions of its neurons, the learning rule, and the architecture itself. A learning rule allows the network to adjust its connection weights to associate given inputs with corresponding outputs. The learning rule used in this study was the back-propagation or modified δ rule. The weights were the adjustable parameters (in that sense, a neural network is a parameterized system). The weighed sum of the inputs constituted the activation of the neuron. The activation signal was passed through the transfer function to produce a single output of the neuron. The transfer function (i.e., sigmoid function) introduces nonlinearity to the network. The internal network parameters (e.g., epoch size, momentum, learning rate, transfer function, and error function) and network geometry (i.e., the number of hidden layers and the number of nodes per hidden layer) had significant effects on the network performance.

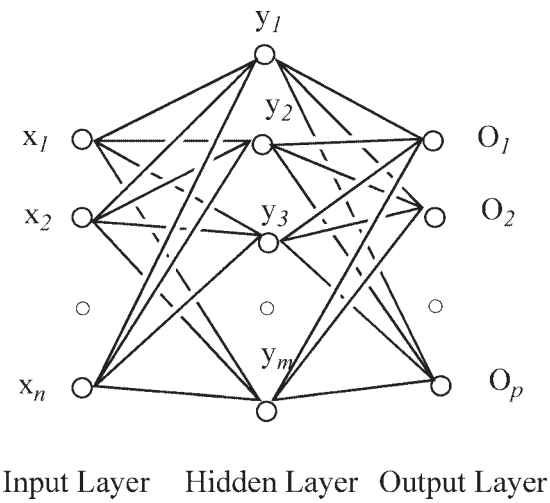

Figure 1 Structure model of a back-propagation ANN. x_n, y_m and O_p are the nth node of the input layer, the yth node of the hidden layer and the pth nodes of the output layer, respectively.

The process of optimizing the connection weights is known as *training* or *learning*. During training, the error (i.e., the sum of squared differences between the predicted and experimental values of the training set) is fed backwards through the network to adjust the weights, minimize the error, and thus prevent the same error from happening again. When the ANN produces the desired output (i.e., it is trained to a satisfactory level), the weighted links between the units are saved. These weights are then used as an analytical tool to predict results for a new set of input data. This is a recall or prediction phase when the network works only by the forward propagation of data and there is no backward propagation of error. The output of a forward propagation is the predicted model for the validation data.

Details on the principles, functioning, and applications of ANNs can be found in refs. 14 and 15.

EXPERIMENTAL

Data set

A total of 105 polyacrylates and polyvinyls and their respective experimental T_g values (see Supplemental Table) were taken from the literature.[6,16] The entire set contained a wide range of T_g values (194–420 K) and was characterized by a high degree of structural variety. The functional groups present in the side chains included halides, acetates, ethers, hydrocarbon chains, aromatic, nonaromatic rings, and so on. The experimental T_g values were divided into a

training set and a prediction set. The training set included 50 polyacrylates, whereas the test set included 34 polyacrylates and 21 polyvinyls.

Calculation of the molecular descriptors

To calculate the molecular descriptors, the polymers were represented by their corresponding monomers. For example, the structure used to calculate the descriptors for poly(acrylic acid) was the acrylic acid molecule. All polymers had wide molecular weight distributions and possessed high molecular weights. Thus, it was impossible to calculate the descriptors directly for the entire molecule. There existed correlations between the polymer properties and the monomers used in the polymerization because the properties depended on the chemical structure of the polymer molecules, and this structure was conditioned by the monomer structure. Thus, statistical methods such as multiple linear regression (MLR) and ANN could be used to develop correlations between the polymer properties and the descriptors obtained from the monomers. In fact, many researches have determined that the properties of polymers are correlated with their monomer structures.[17,18]

DFT is an extremely successful approach for the description of the ground-state properties of molecules. In addition, the computational cost of the DFT calculation, even when electron correlation is treated, is not as expensive as conventional high-level *ab initio* methods, such as configuration interaction or coupled-cluster methods. Thus, we adopted this method to optimize the models with the Gaussian 03 program[19] at the B3LYP level of theory with a 6-31G(d,p) basis set.

All of the geometries of the monomers were fully optimized without the application of symmetry or structural constraints. This was accomplished with the default Gaussian convergence criteria. All of the optimized structures were characterized as true local-energy minima on the potential energy surfaces, without imaginary frequencies. The vibrational frequencies were calculated by application of the ideal gas, rigid rotor, and harmonic oscillator approximations.

After the chemical structures were optimized, the Gaussian output files (*.out) were opened and saved as Sybyl MOL2 files (*.mol2) with GaussView 3.09. Sybyl MOL2 files (*.mol2) were then used as the input for the Dragon software.[10] A total of 1664 molecular descriptors were calculated for every molecule. More information about the types of molecular descriptors calculated with Dragon software can be found in Dragon software users' guide.[10]

Selection of the molecule descriptors

MLR is the most widely used and most well-known modeling method. A MLR model can be expressed as $Y = a_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n$, where Y is the dependent variable (e.g., T_g); X_2, X_3, …, and X_n are the dependent variables (e.g., molecular descriptors); and a_1, a_2, a_3, …, and a_n are the regression coefficients. Stepwise MLR was used to seek an optimum subset of descriptors for an MLR model with the SPSS 11.5 program.[20] This method only added one parameter to the model at a time and always in the order from most significant to least significant. Some important statistical parameters were used to evaluate the molecular descriptors. The t value (or t test) is the statistic used to test whether or not the regression coefficient is equal to the hypothesized value [$t_{\alpha/2}$(n-p-1)]. On the basis of the t value, the level value of significance (p value) was also calculated: $|t| > t_{\alpha/2}$(n-p-1) or $p < 0.05$ (the default level of significance) indicated that a descriptor was significant. Lower p values or higher $|t|$ values correspond to the relatively more significant regression coefficients. Variance inflation factors (VIFs), defined as $\mathrm{VIF} = (1 - R^2)^{-1}$, were used to identify whether excessively high multicollinearities existed among the descriptors; VIF < 10 indicated a tolerable multicollinearity among descriptors;[9] that is, the squared multicollinearity coefficient R^2 did not exceed 0.90.

Determination of the network parameters

The appropriate values of the network parameters stated previously aided network learning. In this study, three descriptors found in the MLR model were fed to ANN as input vectors; the output parameter was T_g ($p = 1$). We empirically determined the sigmoid parameter (0.9), the learning rate (0.1), the momentum parameter (0.6), the permission network error (0.00001), and the maximum number of epochs (5000).[21,22] The optimal number of hidden layers and the number of neurons in each hidden layer were determined by variation of the number of hidden layers and the number of hidden neurons and observation of rms errors. The sum of rms errors of the training set and the test set was used to evaluate the accuracy of the model. The smaller the sum of rms errors was, the higher the predictive quality was.

RESULTS AND DISCUSSION

By carrying out the correlation between the 1664 descriptors and T_g of 50 samples in the training set with stepwise MLR analysis in the SPSS 11.5 program,[20] we obtained the optimal MLR model. The MLR model included three molecule descriptors: mean atomic van der Waals volume (Mv), bond information content (BIC5; neighborhood symmetry of the fifth order), and three-dimensional (3D) molecule representation of structures

TABLE I
Characteristics of the Descriptors in the MLR Model

Descriptor	Coefficient	Standard error	p	t	VIF
Constant	73.050	38.072	0.061	1.919	—
Mv	698.016	50.816	0.000	13.736	1.010
BIC5	−278.545	31.049	0.000	−8.971	1.011
Mor13m	−54.569	9.300	0.000	−5.868	1.001

based on electron diffraction (MoRSE) descriptor for signal 13/weighted by atomic masses. The statistical parameters corresponding to the model of T_g obtained from the training set follow:

$$T_g = 73.050 + 698.016Mv - 278.545\text{BIC5} - 54.569\text{Mor13m} \quad (1)$$

$$R = 0.928, R^2 = 0.861, se = 20.9\,\text{K}, F = 95.194, N = 50$$

where N is the number of samples, R is the correlation coefficient, se is the standard error of estimation, and F is the Fischer ratio. The MLR model was used to make predictions for the test set. The characteristics of the descriptors used in the MLR model are shown in Table I. The rms errors were 20.1 K (R = 0.928) for the training set and 21.7 K (R = 0.908) for the test set.

The three descriptors were then fed to the ANN as input parameters. The final optimum ANN was obtained by trial and error. The ANN included two hidden layers. The first hidden layer comprised three nodes; the second one included two nodes. Thus, the architecture of the optimum neural network was expressed as 3-[3-2]-1. The weights matrices of the neuron links in the networks are shown in Table II. T_g values calculated with the ANN and MLR models are listed in Supplemental Table. The predicted results of the training set and the test set are depicted in Figures 2 and 3, respectively. The rms errors were 15.5 K for the training set and 17.7 K for the test set, which were superior to the results obtained from the MLR model in this study. This indicates that the correlation between T_g and the structural parameters outlined previously was nonlinear rather than linear. In comparison with previous models,[5–9] this ANN model showed better statistical quality.

Table II shows that the three descriptors were all significant descriptors from the significance test. Furthermore, the VIF value of each descriptor was less than two, which suggested that the descriptors did not contaminate each other.

According to the t test (in Table I), the most significant descriptor appearing in the MLR model was Mv (scaled on the carbon atom). Generally, the major factors affecting the T_g values of polymers are intermolecular forces and chain stiffness (or mobility). The constitutional descriptor Mv reflects the hindrance to rotation about the polymeric main chain, which affects chain stiffness (or mobility). For example, some polymers with chlorine atoms in the repeating units possess high polarity, which can enhance the interaction within a polymer chain and between different chains and result in higher T_g values, although these polymers had larger Mv values because Mv of a chlorine atom is larger than that of a hydrogen atom. Thus, the descriptor Mv bore a positive coefficient in the MLR model.

The second significant descriptor included in the model was BIC5. BIC5 is represented in eq. (2):

$$\text{BIC5} = \frac{\text{IC5}}{\log_2\left(\sum_{b=1}^{\text{nBT}} \pi_b^*\right)} \quad (2)$$

where IC5 is the neighborhood information content, nBT is the number of bonds, and π* is the conventional bond order (1 for single, 2 for double, 3 for triple, and 1.5 for aromatic bonds). Equation (2) suggests that a repeating unit with more double, triple, and aromatic bonds would have a smaller BIC5 value. However, these bonds can increase the chain stiffness and lead to a higher T_g value. Thus, the information indices descriptor BIC5 corresponds to rotatable bond fraction (i.e., single bonds) in

TABLE II
Weights Matrices of the Neuron Links in the Networks

First hidden layer			Second hidden layer		Output layer
5.573	−2.513	−3.871	2.210	−21.052	6.766
−3.493	−0.231	2.270	7.934	−15.746	−3.106
0.614	−3.689	1.598	−6.453	3.920	—

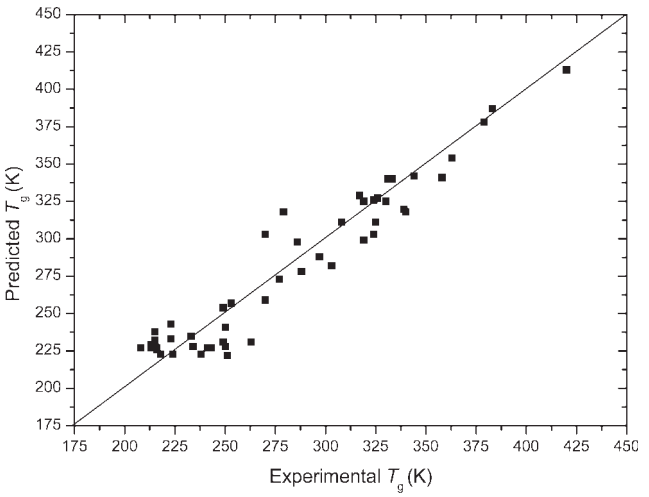

Figure 2 Plot of experimental T_g (K) versus predicted T_g (K) with the ANN model (training set).

the molecule. Moreover, a fractional increase in the rotatable bonds in a molecule is conducive for the mobility of polymer chain.

The third significant descriptor was the 3D MoRSE descriptor Mor13m.[10,23] The 3D MoRSE descriptor based on the atomic mass (m) can be expressed as:

$$\text{Morsm} = \sum_{I=1}^{\text{nAT}-1} \sum_{J=I+1}^{\text{nAT}} m_I m_J \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (3)$$

where Morsm is the scattered electron intensity, r_{ij} is the interatomic distance, and nAT is the number of atoms. The term s represents the scattering in various directions by a collection of nAT atoms. In DRAGON, it is assumed that s takes integer values in the range 0–31 (for $s = 0$, the scattering ratio is assumed to be equal to 1). Therefore, the descriptor Mor13m ($s = 13$) retains important structural features, such as the mass and the amount of branching,[10,23] although these structural features of a repeating unit have a significant effect on T_g of a polymer. Thus, the descriptor Mor13m is correlated with T_g.

Usually, it is difficult for a QSPR model to be extrapolated if the training and test sets have different structures. For example, our ANN model could not be used to predict the T_g values of polymethacrylates. However, it was successful in the prediction for 21 polyvinyls. The reason may be that polyvinyls and polyacrylates have similar repeating unit structures (—CH_2CHX—).

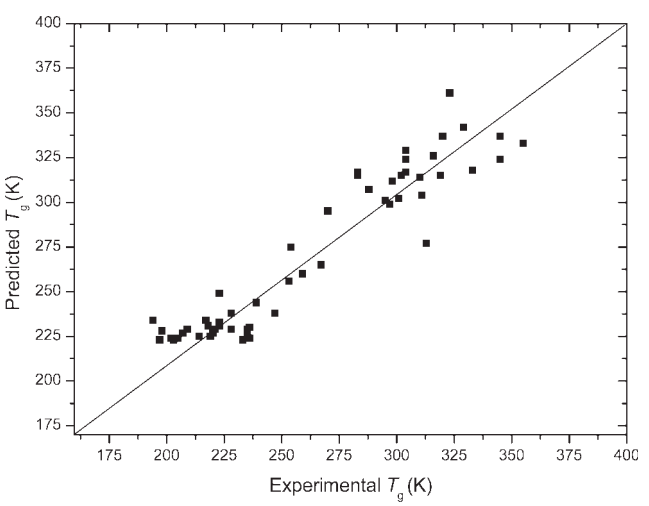

Figure 3 Plot of experimental T_g (K) versus predicted T_g (K) with the ANN model (test set).

CONCLUSIONS

An ANN model with only three descriptors was successfully developed to predict the T_g values of vinyl polymers. *Mv* described the polymer chain stiffness; whereas the 3D MoRSE descriptor Mor13m and BIC5 reflected the molecular mobility. Therefore, the three descriptors represented the essential factors governing the nature of glass transition in the polymers. The correlation between T_g and the descriptors was nonlinear rather than linear, and application of the ANN method to predict T_g values for polyvinyls and polyacrylates is feasible.

References

1. Krause, S.; Gormley, J. J.; Roman, N.; Shetter, J. A.; Wantanade, W. H. J Polym Sci Part A: Polym Chem 1965, 3, 3573.
2. Tracht, U.; Wilhelm, M.; Heuer, A.; Feng, H.; Schmidt-Rohr, K.; Spiess, H. W. Phys Rev Lett 1998, 81, 2727.
3. Schut, J.; Bolikal, D.; Khan, I. J.; Pesnell, A.; Rege, A.; Rojas, R.; Sheihet, L.; Murthy, N. S.; Kohn, J. Polymer 2007, 48, 6115.
4. van Krevelen, D. W. Properties of Polymers, Their Estimation and Correlation with Chemical Structure, 2nd ed.; Elsevier: Amsterdam, 1976.
5. Chen, X.; Sztandera, L.; Cartwright, H. M. Int J Intell Syst 2008, 23, 22.
6. Bicerano, J. Prediction of Polymers Properties, 2nd ed.; Marcel Dekker: New York, 1996.
7. Yu, X. L.; Yi, B.; Wang, X. Y.; Xie, Z. M. Chem Phys 2007, 332, 115.
8. Katritzky, A. R.; Sild, S.; Lobanov, V.; Karlson, M. J Chem Inf Comput Sci 1998, 38, 300.
9. Mattioni, B. E.; Jurs, P. C. J Chem Inf Comput Sci 2002, 42, 232.
10. Talete srl (società a responsabilità limitata). Dragon for Widows (Software for the Calculation of Molecular Descriptors), version; 5.4-2006-http://www.talete.mi.it/.
11. Ruggieri, F.; D'Archivio, A. A.; Carlucci, G.; Mazzeo, P. J Chromatogr A 2005, 1076, 163.
12. Agatonovic-Kustrin, S.; Beresford, R. J Pharm Biomed Anal 2000, 22, 717.
13. Maier, H. R.; Dandy, G. C. Environ Model Software 2000, 15, 101.
14. Zupan, J.; Gasteiger, J. Neural Networks in Chemistry and Drug Design, 2nd ed.; Wiley-VCH: Weinheim, 1999.
15. Tang, Q. Y.; Feng, M. G. Practical Statistics and DPS Data Processing System; Science: Beijing, 2002.

16. Brandrup, J.; Immergut, E. H.; Grulke, E. A. Polymer Handbook, 4th ed.; Wiley: New York, 1999.
17. Yu, X. L.; Xie, Z. M.; Yi, B.; Wang, X. Y.; Liu, F. Eur Polym J 2007, 43, 818.
18. Navarro, A.; Fernández-Liencres, M. P.; Peña-Ruiz, T.; Granadino-Roldán, J. M.; Fernández-Gómez, M.; Domínguez-Espinos, G.; Sanchís, M. J. Polymer 2009, 50, 317.
19. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 03, revision B.03; Gaussian: Pittsburgh, PA, 2003.
20. SPSS. SPSS for Windows, release 11.5.0; SPSS: Chicago, 2002.
21. Yu, X. L.; Yi, B.; Liu, F.; Wang, X. Y. React Funct Polym 2008, 68, 1557.
22. Yu, X. L.; Yi, B.; Wang, X. Y. Eur Polym J 2008, 44, 3997.
23. Schuur, J. H.; Paul Selzer, P.; Gasteiger, J. J Chem Inf Comput Sci 1996, 36, 334.