Quantitative Structure-Property Relationships Studies on Free-Radical Polymerization Chain-Transfer Constants for Styrene

Jie Xu,[1] Lei Wang,[1] Hui Zhang,[1] Xiaolin Shen,[1] Guijie Liang[1,2]

[1]*Key Lab of Green Processing and Functional Textiles of New Textile Materials, Ministry of Education, Wuhan Textile University, 430073, Wuhan, China*
[2]*College of Materials Science and Engineering, Xi'an Jiao tong University, 710049, Xi'an, China*

Received 8 September 2010; accepted 29 January 2011
DOI 10.1002/app.34255
Published online 27 July 2011 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Quantitative structure-property relationships (QSPR) studies were performed for kinetic chain-transfer constants of 90 agents on styrene polymerization at 60°C. By using stepwise multilinear regression analysis (MLRA) and artificial neural network (ANN), linear and nonlinear models containing seven descriptors were obtained with R^2 of 0.7866 and 0.8661 for the training set, respectively. The reliability of the proposed models was validated through the test set. The descriptors involved in the models are related to the molecular conformational changes and some functional groups. As these descriptors are directly calculated from the molecular structure, the proposed models can be employed to estimate the chain-transfer constants for styrene. © 2011 Wiley Periodicals, Inc. J Appl Polym Sci 123: 356–364, 2012

Key words: QSPR; chain-transfer constants; free-radical polymerization; styrene

INTRODUCTION

The chain-transfer constant is an important parameter in polymer chemistry since an understanding of chain-transfer clarifies the microkinetic process in polymerization reactions.[1] The transfer reaction in free-radical polymerization describes a process in which the polymer radical reacts with another molecule (monomer, polymer, solvent, modifier, etc.) forming a dead polymer and new radical. This new radical can continue the kinetic chain[2]:

$$P_n^\bullet + M \xrightarrow{k_p} P_{n+1}^\bullet$$
$$P_n^\bullet + RX \xrightarrow{k_{tr,X}} P_nR + X^\bullet$$
$$X^\bullet + M \xrightarrow{k_{i,X}} P_1^\bullet$$

where P_n is a polymer molecule of chain length n, RX is a transfer agent, and M is a monomer molecule. The kinetic chain-transfer constant is defined as dimensionless quantity by the following ratio

$$C_X = k_{tr,X}/k_p$$

where k_p is the rate constant of free-radical polymerization propagation.

It is assumed that (a) all new radicals react only by formation of growing polymer radicals; (b) all polymer radicals have equal reactivity regardless of their size; (c) all rate constants are independent of solvent; (d) the consumption of monomer by initiation and transfer is negligible compared with propagation; (e) a steady-state concentration of polymer radicals is quickly established ($d[P^\bullet]/dt = 0$). Accordingly, the number-average degree of polymerization in the presence of transfer agent X (P_n) is given as follows:

$$\frac{1}{P_n} = \frac{1}{P_{n0}} + \sum C_X \frac{[X]}{[M]}$$

where P_{n0} is the number-average degree of polymerization in the absence of any transfer agent, and $[M]$ and $[X]$ are the concentrations of monomer and transfer agent, respectively.[2] It can be seen that the chain-transfer reactions modulate the molecular

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

Correspondence to: J. Xu (xujie0@ustc.edu).

Contract grant sponsor: Natural Science Foundation of Hubei Province; contract grant number: 2008CDB261.
Contract grant sponsor: Key Project of Science and Technology Research of Ministry of Education; contract grant number: 208089.
Contract grant sponsor: Natural Science Foundation of China; contract grant number: 51003082.
Contract grant sponsor: Educational Commission of Hubei Province; contract grant number: Q20101606.

weight and consequently its distribution during polymerization. Molecular weight and molecular weight distributions have considerable effects on polymer processability. Thus, control of these macromolecular features is required when polymers with high molecular weight are not suitable for a given application.[3] Knowledge of chain-transfer constants is extremely meaningful for both the laboratory and industrial scale-up of polymerization processes. Experimental C_X data have been available for a lot of systems; however, a relatively accurate prediction method is still required when experimental data is lacking.

Alternatively, the quantitative structure-property relationships (QSPR) provide a promising approach for estimating the C_X values based on descriptors derived solely from the molecular structure to fit experimental data. The QSPR approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds, termed "molecular descriptors."[4–9] The advantage of this approach lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. The statistic techniques commonly used in QSPR studies are linear and nonlinear methods. The most commonly used linear method is the stepwise multilinear regression analysis (MLRA), which can run forward or backward. The QSPR approach has been used quite extensively to predict many properties in polymer chemistry and physics, such as refractive index,[10–17] glass transition temperature,[12,18–26] lower critical solution temperature,[27–30] Flory-Huggins parameters,[31,32] intrinsic viscosity,[33,34] solubility parameters,[35] and monomer reactivity parameters.[36–38] However, there have been relatively few attempts to correlate and predict the chain-transfer constants. Ignatz-Hoover et al.[3] used the Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) program to develop linear three- and five-parameter correlations with R^2 of 0.725 and 0.818 for a set of $\log C_X$ values for 90 agents on styrene polymerization at 60°C.

Because of the complex relationships existing between the property of the molecules and the structures, nonlinear modeling methods are sometimes better to use. To deal with nonlinear behavior, different algorithms have been proposed, and among them the artificial neural network (ANN) has found much popularity in QSPR studies.[21] The goal of the present work is to produce robust QSPR models that could predict the transfer constants for styrene polymerization at 60°C using both MLRA and ANN, based on descriptors calculated by Dragon software.[39]

MATERIALS AND METHODS

Data set

The experimental chain-transfer constants of 90 transfer agents for styrene at 60°C (Table I) were taken from the article by Ignatz-Hoover et al.[3] The reported $\log C_X$ values ranged from about -1.5 to 6.5. Among them, 60 compounds were randomly chosen as the training set, and the other 30 compounds were used as the test set.

Descriptor generation

The structures of all molecules were preoptimized using MM+ molecular mechanics force field (Polak-Ribiere algorithm) in the HYPERCHEM program.[40] The final geometries of the minimum energy conformation were obtained by the semiempirical AM1 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of 0.01 $\text{kcal}\cdot\text{Å}^{-1}\cdot\text{mol}^{-1}$ as a stopping criterion. Then a total of 1664 molecular descriptors for each molecule were calculated on the resulting geometry with Dragon software.[39] These descriptors include (a) 0D-constitutional (atom and group counts); (b) 1D-functional groups and atom centered fragments; (c) 2D-topological, BCUTs, walk and path counts, autocorrelations, connectivity indices, information indices, topological charge indices, and eigenvalue-based indices; and (d) 3D-Randic molecular profiles from the geometry matrix, geometrical, WHIM, and GETAWAY descriptors.

To reduce redundant and nonuseful information, descriptors with constant or near constant values and pair intercorrelations greater than 0.99[41] were excluded in a prereduction step. The 854 remaining descriptors underwent subsequent descriptor selection.

Model development and validation

Stepwise MLRA with Leave-One-Out (LOO) cross-validation was used to select descriptors for the linear QSPR models on the training set. *F*-to-enter and *F*-to-remove were 4 and 3, respectively. The models were justified by the R^2, the cross-validated R^2, the standard error of estimation *s* and the *F* ratio values. A variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as

$$\text{VIF} = \frac{1}{1 - R_j^2} \tag{1}$$

where R_j^2 is the squared correlation coefficient between the *j*th coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with VIFs above a value of 5.[42]

The nonlinear model was then developed by submitting the selected descriptors from MLRA to a three-layer, fully connected, feed-forward ANN. The

TABLE I
Experimental and Predicted Values of logC_X

Transfer agent name	Exp. logC_X	MLRA		ANN	
		Pred. logC_X	Diff.	Pred. logC_X	Diff.
Training set					
(1,1,2,2-Tetraphenylethyl)benzene	4.30	4.45	−0.15	4.00	0.30
(3-Bromophenyl)acetonitrile	1.84	2.17	−0.33	1.53	0.31
(4-Methoxyphenyl)acetonitrile	1.71	1.31	0.40	1.70	0.01
(E)-2-butenal oxime	3.18	3.37	−0.19	3.03	0.15
(*tert*-Butyl)benzene	−1.30	−0.10	−1.20	−0.40	−0.90
1,2,3-Benzenetriol	4.02	3.84	0.18	4.58	−0.56
1,2-Benzenediol	3.13	2.42	0.71	2.61	0.52
1,2-Benzenediol	3.13	2.42	0.71	2.61	0.52
1,3,5-trinitrobenzene	5.55	4.05	1.50	4.31	1.24
1,4-Di(*sec*-butyl)benzene	1.03	1.17	−0.14	1.84	−0.81
1,4-Dibutylbenzene	0.85	0.31	0.54	0.69	0.16
1,4-Dihydroxybenzene	0.56	2.52	−1.96	1.33	−0.77
1,4-Diisopropylbenzene	0.52	0.59	−0.07	0.42	0.10
1-Chloro-2-methylpropane	0.15	0.37	−0.22	0.29	−0.14
1-Chloro-4-ethynylbenzene	2.21	1.26	0.95	1.94	0.27
1-Chlorobenzene	−0.10	0.03	−0.13	0.12	−0.22
1-Propenaloxime	4.03	0.90	3.13	1.02	3.01
2,3,5,6-Tetramethylphenol	2.76	2.61	0.15	2.71	0.05
2,4,6-Trinitrophenol	5.32	6.20	−0.88	5.25	0.07
2,5-Cyclohexadiene-1,4-dione	6.36	5.82	0.54	6.43	−0.07
2,6-Di(2-propyl)phenol	2.49	2.46	0.03	2.56	−0.07
2-Bromo-1,3,5-trinitrobenzene	5.76	5.47	0.29	5.61	0.15
2-Butanone	0.70	0.37	0.33	0.37	0.33
2-Chloroacetyl chloride	3.52	3.53	−0.01	3.72	−0.20
2-Chlorobutane	0.08	0.53	−0.45	0.47	−0.39
2-Methoxy-1,3,5-trinitrobenzene	5.31	4.74	0.57	4.75	0.56
2-Methyl-1-penten-3-one oxime	3.63	3.40	0.23	3.43	0.20
2-Methylphenol	1.63	1.95	−0.32	1.78	−0.15
2-Phenylacetic acid	0.78	1.60	−0.82	0.98	−0.20
2-Propen-1-ol	0.18	0.29	−0.11	0.03	0.15
3-Buten-2-one oxime	3.43	3.57	−0.14	3.69	−0.26
3-Chlorobenzaldehyde	1.14	1.69	−0.55	1.33	−0.19
3-Methyl-3-buten-2-one oxime	3.04	3.48	−0.44	3.18	−0.14
4-(*tert*-Butyl)-1,2-benzenediol	3.56	2.78	0.78	3.47	0.09
4-Chlorobenzaldehyde	0.94	1.73	−0.79	1.30	−0.36
4-Chlorobenzenesulfonyl chloride	3.88	4.06	−0.18	4.48	−0.60
4-Methoxybenzenesulfonyl chloride	3.49	3.76	−0.27	4.03	−0.54
4-Methyl-2-pentanone oxime	3.36	1.45	1.91	2.01	1.35
4-Methylbenzenesulfonyl chloride	3.50	3.52	−0.02	4.20	−0.70
Acetonitrile	−0.36	−0.05	−0.31	−0.12	−0.24
Acetyl bromide	3.93	3.92	0.01	3.75	0.18
Benzaldehyde	0.66	0.92	−0.26	0.49	0.17
Benzene	−1.56	−1.02	−0.54	−1.31	−0.25
Chloro(dimethyl)germane	4.52	4.52	0.00	4.20	0.32
Chloroacetic acid	1.46	1.66	−0.20	1.65	−0.19
Di(2-propenyl)propanedioate	0.72	2.02	−1.30	1.13	−0.41
Dichloro(ethyl)germane	4.76	5.46	−0.70	4.99	−0.23
Dichloroacetic acid	1.54	2.31	−0.77	1.92	−0.38
Diethyl 2,2-dibromopropanedioate	4.08	2.48	1.60	2.96	1.12
Diethyl malonate	−0.33	1.31	−1.64	0.57	−0.90
Diethyl-2,2-dichloropropanedioate	1.48	2.61	−1.13	1.07	0.41
Diethyl-2-bromopropanedioate	2.85	2.48	0.37	2.96	−0.11
Methyl 2-chloroacetate	−0.52	0.91	−1.43	0.77	−1.29
N,*N*-diethenylphenylamine	2.11	0.43	1.68	0.58	1.53
Phenylamine	0.30	−0.28	0.58	−0.64	0.94
Phenylmethanesulfonyl chloride	3.50	2.96	0.54	3.66	−0.16
Tetrabromomethane	4.33	3.92	0.41	4.57	−0.24
Tetrachloromethane	2.03	1.72	0.31	1.15	0.88
Trichloroacetic acid	1.82	2.26	−0.44	2.69	−0.87
Triphenylgermane	3.36	3.72	−0.36	4.03	−0.67

(Continued)

TABLE I. *Continued*

Transfer agent name	Exp. logC_X	MLRA		ANN	
		Pred. logC_X	Diff.	Pred. logC_X	Diff.
Test set					
(1,2-Dibromoethyl)benzene	3.29	2.33	0.96	3.33	−0.04
(4-Chlorophenyl)acetonitrile	1.82	1.58	0.24	1.15	0.67
(*sec*-Butyl)benzene	0.79	0.43	0.36	0.42	0.37
1,2,3-Benzenetriol	4.02	3.84	0.18	4.58	−0.56
1,2-Benzenediol	3.13	2.42	0.71	2.61	0.52
1-Bromo-4-ethynylbenzene	2.28	1.79	0.49	2.74	−0.46
1-Bromobenzene	0.25	0.60	−0.35	0.50	−0.25
1-Bromobutane	−1.22	0.66	−1.88	0.51	−1.73
1-Chlorobutane	−1.40	0.06	−1.46	0.37	−1.77
1-Ethylbenzene	−0.16	−0.20	0.04	−0.75	0.59
1-Ethynyl-4-nitrobenzene	3.50	2.12	1.38	3.09	0.41
2,4,6-Trinitrophenylamine	5.07	5.15	−0.08	4.38	0.69
2,5-Dimethyl-2,5-cyclohexadiene-1,4-dione	5.63	7.18	−1.55	6.26	−0.63
2-Bromoacetic acid	2.63	2.26	0.37	2.49	0.14
2-Methyl-2-propenal oxime	4.11	3.70	0.41	3.28	0.83
4-(*tert*-Butyl)-1,2-benzenediol	3.56	2.78	0.78	3.47	0.09
4-Bromobenzaldehyde	1.08	2.29	−1.21	2.40	−1.32
4-Formylbenzonitrile	1.88	1.84	0.04	1.88	0.00
4-Methylphenol	1.59	1.98	−0.39	1.75	−0.16
Acetaldehyde	0.93	−0.27	1.20	−0.24	1.17
Benzenesulfonyl chloride	3.64	3.32	0.32	3.81	−0.17
Chloro(diethyl)germane	4.50	3.87	0.63	4.79	−0.29
Dimethyl ketone	0.61	0.42	0.19	−0.01	0.62
Ethyl 2,4,6-trinitrobenzoate	5.76	6.48	−0.72	6.61	−0.85
Iodoacetic acid	3.90	2.71	1.19	2.80	1.10
Methanesulfonyl chloride	3.07	2.85	0.22	3.87	−0.80
N,*N*-dimethylacetamide	0.66	1.02	−0.36	0.23	0.43
Phenol	1.15	1.00	0.15	1.08	0.07
Trichloromethane	0.08	1.92	−1.84	0.60	−0.52
Triethylgermane	3.38	2.71	0.67	2.72	0.66

number of input neurons was equal to that of the descriptors in the linear model. The number of hidden neurons was optimized by trial and error procedure on calculations of the training process. One output neuron was used to represent the experimental logC_X. The network was trained using the quasi-Newton BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm.[43–46] To avoid overtraining, one tenth data from the training set were randomly selected as separate validation set to monitor the training process; that is, during the training of the network the performance was monitored by predicting the values for the systems in the validation set. When the results for the validation set ceased to improve, the training was stopped.

Validation of the linear and nonlinear models was also performed by using the external test set composed of data not used to develop the prediction model. The external $R^2_{CV,ext}$ for the test sets is determined with eq. (2):

$$R^2_{CV,ext} = 1 - \frac{\sum_{i=1}^{test} (y_{exp\,t} - y_{calc})^2}{\sum_{i=1}^{test} (y_{exp\,t} - \bar{y}_{test})^2} \quad (2)$$

where $\overline{y}_{test}$ is the averaged value for the response variable of the test set. According to Golbraikh and Tropsha,[47] a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{CV,ext} > 0.5 \quad (3a)$$

$$r^2 > 0.6 \quad (3b)$$

$$(r^2 - r_0^2)/r^2 < 0.1 \text{ or } (r^2 - r'^2_0)/r^2 < 0.1 \quad (3c)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \quad (3d)$$

where r is the correlation coefficient between the calculated values and experimental values in the test set. r_0^2 (calculated vs. observed values) and r'^2_0 (observed vs. calculated values) are coefficients of determination. k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively. Detailed mathematical definitions of these parameters can be found in the publication of Golbraikh and Tropsha.[47]

RESULTS AND DISCUSSION

MLRA model

The stepwise MLRA with LOO cross-validation was used to select descriptors and develop linear models

TABLE II
Characteristics of the Descriptors in the MLRA Model

Descriptor	Descriptor type	Coefficient	Error	*t*-Value	*t*-Probability	VIF
Constant		−5.418	1.254	−4.320	0.000	
SHP2	Randic molecular profiles	6.105	1.820	3.354	0.001	2.057
G2	Geometrical descriptors	0.571	0.045	12.593	0.000	1.727
Mor20u	3D-MoRSE descriptors	−1.071	0.343	−3.118	0.003	2.449
RTu+	GETAWAY descriptors	3.840	1.638	2.344	0.023	1.484
nCconj	Functional group counts	0.775	0.118	6.590	0.000	1.105
nRCOX	Functional group counts	1.829	0.729	2.510	0.015	1.212
nArOH	Functional group counts	0.768	0.187	4.097	0.000	1.047

$n = 60$, $R^2 = 0.7866$, $R^2_{CV} = 0.7355$, $s = 0.92$, $F = 27.4$.

on the training set. It is clear that univariant correlations between $\log C_X$ and different descriptors have a small value for the correlation coefficient. This indicates that $\log C_X$ is not linearly correlated with any of the molecular descriptors. The R^2 increases gradually with the increased number of descriptors. When adding another descriptor did not significantly improve the statistics of a model, it was determined that the optimum subset size had been achieved. To avoid over-parameterization of the models, such as those which contain an excess of descriptors and are difficult to interpret in terms of physical interactions, an increase of the R^2 value of less than 0.01 was chosen as the breakpoint criterion. Thus, a seven-parameter model with R^2 of 0.7866 and R^2_{CV} of 0.7355 was obtained, which is as the following:

$$\begin{aligned}\log C_X = &-5.418 + 6.105[\mathrm{SHP2}] \\ &+ 0.571[\mathrm{G2}] - 1.071[\mathrm{Mor20u}] + 3.840[\mathrm{RTu+}] \\ &+ 0.775[\mathrm{nCconj}] + 1.829[\mathrm{nRCOX}] \\ &+ 0.768[\mathrm{nArOH}] \\ &\mathrm{n} = 60, R^2 = 0.7866, R^2_{CV} = 0.7355, \\ &s = 0.92, \mathrm{F} = 27.4 \qquad (4)\end{aligned}$$

Here, SHP2 is average shape profile index of order 2; G2 is gravitational index G2 (bond-restricted); Mor20u is 3D-MoRSE signal 20/unweighted; RTu+ corresponds to R maximal index/unweighted; nCconj corresponds to the number of nonaromatic conjugated C(sp2); nCOX corresponds to the number of acyl halogenides (aliphatic); nArOH corresponds to the number of aromatic hydroxyls, respectively.[48]

The cross-validated correlation coefficient $R^2_{CV} = 0.7355$ illustrates the reliability of the model by focusing on the sensitivity of the model to the elimination of any single data point. The statistical characteristics of the seven descriptors are given in Table II, which indicate that all the descriptors are highly significant from the *t*-test values. The VIF values (less than 5) suggest that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

The calculated results of the $\log C_X$ values from eq. (4) for the training and test sets are shown in Table I and Figure 1. Only two compounds (1-propenaloxime, 1,4-dihydroxybenzene) are outliers according to 95% statistical reliability level (±1.84).

ANN model

The ANN has become an important and widely used nonlinear modeling technique for QSPR studies. The mathematical adaptability of ANN commends it as a powerful tool for pattern classification and building predictive models. A particular advantage of ANN is its inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function. Among the neural network learning algorithms, the back-propagation (BP) method[49] is one of the most commonly used methods. The drawback of BP is that the training processes slowly, because the gradient-descent algorithm is usually used for minimizing the sum-of-squares error. In this study, the quasi-Newton BFGS algorithm was used to develop nonlinear models. The

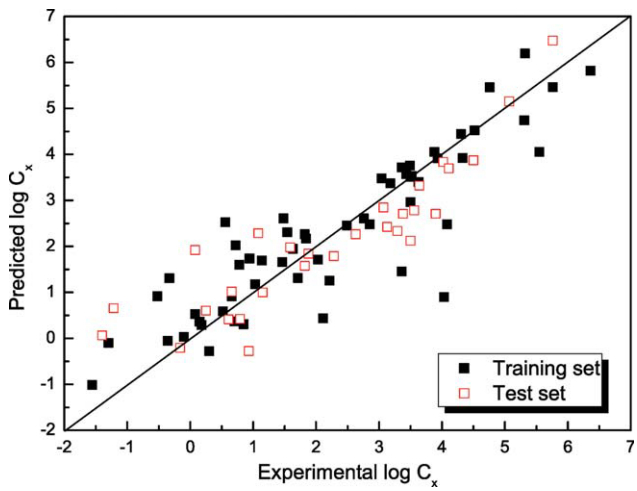

Figure 1 Plot of predicted vs. experimental values of log C_X using the MLRA model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

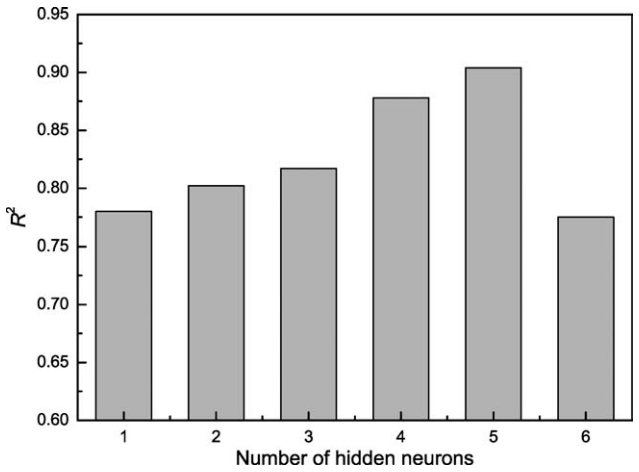

Figure 2 R^2 vs. the number of neurons in the hidden layer.

advantages of using the BFGS algorithm are that specifying rate or momentum is not necessary and the training is much more rapid.[50] The seven descriptors from the best MLRA model were used as inputs to the network. The number of hidden neurons is an important parameter influencing the performances of the ANN. The usual rule of thumb is that the weights and biases should be less than the samples so that the model achieved by the network is stationary.[51] In the situation of this work, with 60 samples in the training set, the number of the hidden neurons should not therefore be greater than six. Figure 2 shows changes in the R^2 value while optimizing neural network architecture with respect to the number of hidden neurons. Thus, a 7-5-1 network architecture (provided as supporting information) was obtained after rigorous trial and error procedure.

The obtained R^2, R^2_{CV}, and s are 0.8661, 0.8578, and 0.69, respectively, which indicate good agreement between the correlation and the variation in the data. The large F ratio of 375.2 indicates that the ANN model does an excellent job of predicting the logC_X values. The following statistical parameters are obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present ANN model:

$$R^2_{CV,ext} = 0.827 > 0.5$$
$$r^2 = 0.92^2 = 0.848 > 0.6$$
$$(r^2 - r_0^2)/r^2 = (0.848 - 0.996)/0.848 < 0.1$$
$$\text{or } (r^2 - r'^2_0)/r^2 = (0.848 - 0.998)/0.848 < 0.1$$
$$0.85 \le k = 0.963 \le 1.15 \text{ or } 0.85 \le k' = 0.974 \le 1.15$$

The calculated logC_X values by the ANN model were given in Table I and Figure 3. According to 95% statistical reliability level (±1.38), three compounds (1-propenaloxime, 1-chlorobutane, 1-bromobutane) are outliers. The distributions of the absolute errors (AEs) predicted with the present MLRA and ANN models for the entire dataset were given in Figure 4. The mean absolute errors (MAEs) from the MLRA and ANN models for the entire dataset are 0.636 and 0.513, respectively. The statistical fit of the ANN model is better than the present MLRA model and the linear five-parameter correlation by Ignatz-Hoover et al.[3] (MAE = 0.636, our calculation), which confirms practically that ANN has a better generalization performance than the traditional linear regression method in solving this kind of nonlinear problem, and can describe the relationship between the structural information and chain-transfer constants more effectively. It is noteworthy that both models give poor predictive results for 1-propenaloxime, thus this compound can be considered as an outlier perhaps due to the experimental uncertainties. The experimental C_X values used in this work

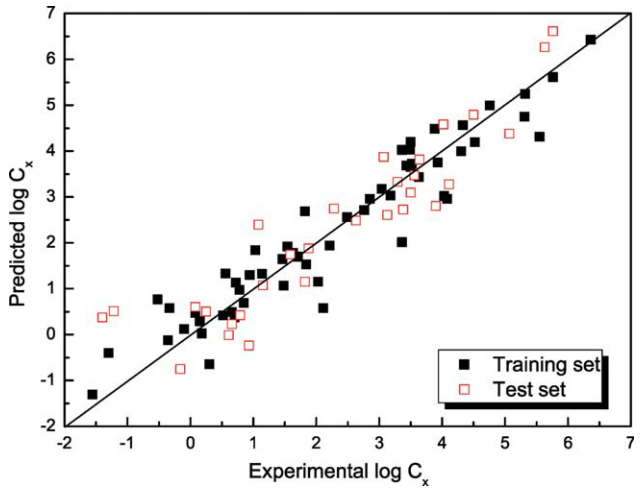

Figure 3 Plot of predicted vs. experimental values of log C_X using the ANN model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

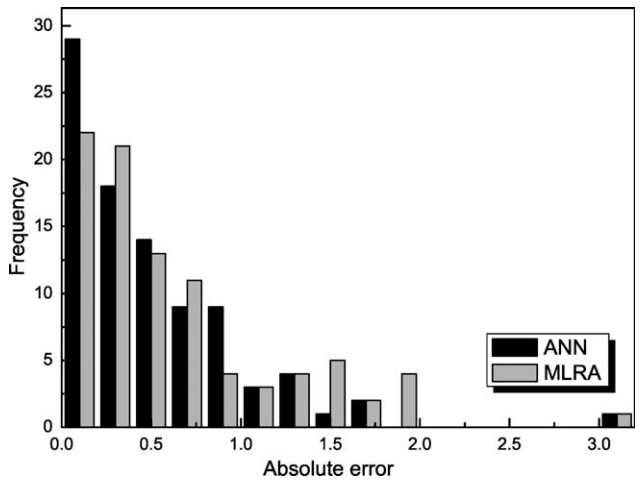

Figure 4 Distributions of relative errors predicted with the MLRA and ANN models for the whole data set.

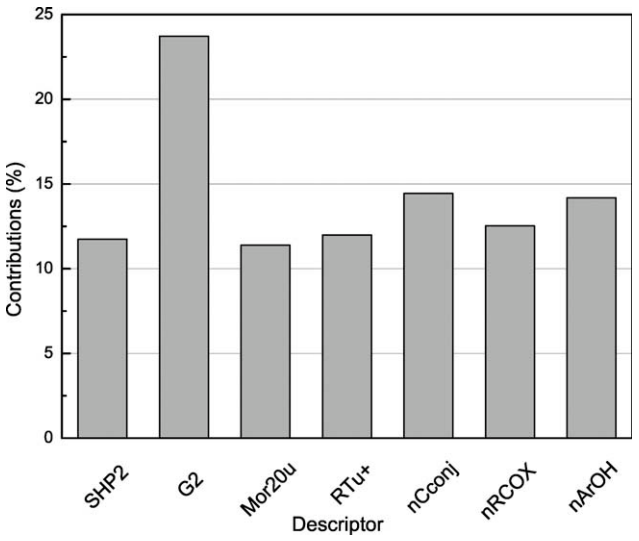

Figure 5 Relative contributions of the seven descriptors to the ANN model.

were taken from various sources and contain considerable variance since different measuring methods were employed for the molecular weight in the determination of C_X. As a result, the experimental uncertainties in C_X can be quite large.

Descriptor interpretation

The descriptors appearing in the present models can be classified as follows: (1) a Randic molecular profile, SHP2; (2) a geometrical descriptor, G2; (3) a 3D-MoRSE descriptor, Mor20u; (4) a GETAWAY descriptor, RTu+; and (4) three functional group counts, nCconj, nRCOX, and nArOH.

Based on a previously described procedure,[52,53] the relative contributions of the seven descriptors to the ANN model were determined and are plotted in Figure 5. The significance of these descriptors decreases in the following order: G2 (23.7%) > nCconj (14.4%) > nArOH (14.2%) > nRCOX (12.5%) > RTu+ (12.0%) > SHP2 (11.7%) > Mor20u (11.4%). The most significant descriptor is the gravitational index G2, which explains 23.7% contribution of the total, much higher than those of the other descriptors. However, the contribution difference between any two descriptors (except G2) used in the ANN model is not remarkable, indicating that all of the descriptors are important in generating the model.

The gravitational index G2 correlates relatively high ($R = 0.5019$) with the target experimental $\log C_X$ values. The G2 index is a molecular descriptor reflecting the mass distribution in a molecule, defined as[54]:

$$G2 = \sum_{b=1}^{n\mathrm{BT}} \left(\frac{m_i \bullet m_j}{r_{ij}^2} \right)_b \tag{5}$$

where m_i and m_j are the atomic masses of the considered atoms, r_{ij} the corresponding interatomic distances, nBT the number of chemical bonds in the molecule, respectively. The G2 index is restricted to pairs of bonded atoms, which is related to the bulk cohesiveness of the molecules. The G2 index may be regarded as responsible for the cavity-forming effects, because it accounts significantly for molecular size and the London dispersion forces.[54]

The average shape profile index of order 2 (SHP2) is the shape profile index of order 2 (SP02) divided by the number of atoms with H-depleted connectivity equal to 1 or 2, where SP02 is calculated as[55]:

$$\mathrm{SP02} = \frac{1}{2!} \bullet \frac{\sum_{i=1}^{n\mathrm{ST}} \sum_{j=1}^{n\mathrm{ST}} r_{ij}^2}{n\mathrm{ST}} \tag{6}$$

Here, r_{ij} is the geometric distance between the considered atoms, nST the number of atoms on molecular periphery (i.e., atoms with H-depleted connectivity equal to 1 or 2). The Randic index SHP2 can be interpreted as the relative molecular area of the external accessibility.[56] This area represents the total area accessible from the environment surrounding the molecule. The SHP2 in the models reflects the influence of the steric factor on the reactivity of the transfer agents, which has similar physical meaning as the first-order Kier and Hall index in the correlation by Ignatz-Hoover et al.[3]

Three-dimensional-MoRSE descriptors are the three-dimensional molecular representations of structure based on electron diffraction descriptor,[57,58] which are calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle(s) in the range of 0 to 31 Å^{-1} from the three dimensional atomic coordinates of a molecule. Mor20u is calculated using the following expression:

$$\mathrm{Mors}w = \sum_{i=1}^{n\mathrm{AT}-1} \sum_{j=i+1}^{n\mathrm{AT}} w_i \cdot w_j \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}} \tag{7}$$

where nAT is the number of atoms in the molecule, w_i and w_j the unweighted schemes of the considered atoms, and s the scattering angle. For the case of Mor20u, s was equal to 19 Å^{-1}. Mor20u increases with the increased polarizability of a molecule. The negative correlation coefficient for Mor20u indicates that the compounds with larger polarizability would have smaller C_X values.

GETAWAY descriptors have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular

influence matrix.[59,60] These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. RTu+ is calculated as:

$$\mathrm{RTu+} = \max_k\left(\max_{ij}\left(\frac{\sqrt{h_{ii}\bullet h_{jj}}}{r_{ij}}\bullet w_i \bullet w_j \bullet \delta\left(k; d_{ij}\right)\right)\right) \quad i \neq j \text{ and } k = 1, 2, ..., 8 \tag{8}$$

where h_{ii} and h_{jj} are the leverages of the ith and jth atom, d_{ij} is the topological distance, and $\delta(k, d_{ij})$ is a Dirac-delta function ($\delta = 1$ if $d_{ij} = k$, zero otherwise). The RTu+ descriptor is more correlated with the structural features like molecular size and shape.

In addition, since G2, RTu+, SHP2, and Mor20u encode also three-dimensional information that depends on the conformation of the molecule, it is possible to argue that the C_X values of the present set of chain-transfer agents have a considerable dependence on conformational changes. The importance of certain functional groups [nonaromatic conjugated C(sp2), acyl halogenides (aliphatic), and aromatic hydroxyls] on the C_X values is also apparent due to the contributions of nCconj, nArOH, and nRCOX to the models, which is in agreement with the well-known high reactivity of these groups in the chain-transfer reaction. The positive coefficients of these descriptors confirm that the transfer agents containing these functional groups more would have larger C_X values.

CONCLUSIONS

In this article, QSPR models for the prediction of free-radical polymerization chain-transfer constants (C_X) for styrene were developed by MLRA and ANN. The R^2 of the MLRA and ANN models on the training set is 0.7866 and 0.8661, respectively. The functional group counts *n*Cconj, *n*ArOH. and *n*RCOX present in the models indicate that three functional groups [nonaromatic conjugated C(sp2), acyl halogenides (aliphatic), and aromatic hydroxyls] have considerable effects on the C_X values of chain-transfer agents. Also, the C_X values depend significantly on the molecular conformational changes as considering the presence of the 3D-descriptors G2, RTu+, SHP2, and Mor20u. The accuracy and robustness of the proposed models are illustrated not only by calculating their fitness on the training set, but also by testing their predicting ability on the test set. Furthermore, it can be seen that the nonlinear ANN model can describe more accurately the relationship between the structural parameters and the property. In summary, this investigation extends the research method to predict the free-radical polymerization chain-transfer constants for styrene.

The authors gratefully wish to express their thanks to the reviewers for critically reviewing the manuscript and making important suggestions.

References

1. Wolff, E.-H. P.; Bos, A. N. R. Ind Eng Chem Res 1997, 36, 1163.
2. Brandrup, J.; Immergut, E. H.; Grulke, E. A. Polymer Handbook; John Wiley & Sons: New York, 1999.
3. Ignatz-Hoover, F.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. J Chem Inf Comput Sci 2001, 41, 295.
4. Devillers, J.;Balaban, A. T., Eds. Topological Indices and Related Descriptors in QSAR and QSPR; Gordon and Breach: The Netherlands, 1999.
5. Karelson, M. Molecular Descriptors in QSAR/QSPR; Wiley-Interscience: New York, 2000.
6. Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Chemom Intell Lab Syst 2002, 62, 217.
7. Xu, J.; Guo, B.; Chen, B.; Zhang, Q. J Mol Model 2005, 12, 65.
8. Katritzky, A. R.; Dobchev, D. A.; Karelson, M.; Z Naturforsch B: Chem Sci 2006, 61, 373.
9. Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Chem Rev 2010, 110, 5714.
10. Bicerano, J. Prediction of Polymer Properties; Marcel Dekker Inc.: New York, 1996.
11. Katritzky, A.; Sild, S.; Karelson, M. J Chem Inf Comput Sci 1998, 38, 1171.
12. García-Domenech, R.; Julián-Ortiz, J. V. J Phys Chem B 2002, 106, 1501.
13. Xu, J.; Chen, B.; Zhang, Q.; Guo, B. Polymer 2004, 45, 8651.
14. Holder, A. J.; Ye, L.; Eick, J. D.; Chappelow, C. C. QSAR Comb Sci 2006, 25, 905.
15. Yu, X.; Yi, B.; Wang X. J Comput Chem 2007, 28, 2336.
16. Gao, J.; Xu, J.; Chen, B.; Zhang, Q. J Mol Model 2007, 13, 573.
17. Xu, J.; Liang, H.; Chen, B.; Xu, W.; Shen, X.; Liu, H. Chemom Intell Lab Syst 2008, 92, 152.
18. Katritzky, A.; Sild, S.; Lobanov, V.; Karelson, M. J Chem Inf Comput Sci 1998, 38, 300.
19. Mattioni, B. E.; Jurs, P. C. J Chem Inf Comput Sci 2002, 42, 232.
20. Cao, C.; Lin, Y. J Chem Inf Comput Sci 2003, 43, 643.
21. Afantitis, A.; Melagraki, G.; Makridima, K.; Alexandridis, A.; Sarimveis, H.; Iglessi-Markopoulou, O. J Mol Struct (Theochem) 2005, 716, 193.
22. Yu, X.; Wang, X.; Wang, H.; Liu, A.; Zhang, C. J Mol Struct (Theochem) 2006, 766, 113.
23. Yu, X.; Yi, B.; Wang, X.; Xie, Z. Chem Phys 2007, 332, 115.
24. Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Polymer 2007, 48, 7121.
25. Duce, C.; Micheli, A.; Starita, A.; Tiné, M. R.; Solaro, R. Macromol Rapid Commun 2006, 27, 711.
26. Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Macromol Symp 2006, 234, 13.
27. Liu, H.; Zhong, C. Ind Eng Chem Res 2005, 44, 634.
28. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Koutentis, P. A. Markopoulos, J.; Igglessi-Markopoulou, O. J Mol Model 2007, 13, 55.
29. Xu, J.; Liu, L.; Xu, W.; Zhao, S.; Zuo, D. J Mol Graph Model 2007, 26, 352.
30. Xu, J.; Chen, B.; Liang, H. Macromol Theory Simul 2008, 17, 109.
31. Toropov, A. A.; Balakhonenko, O. I.; Nurgaliev, I. N.; Voropaeva, N. L.; Ruban, I. N.; Rashidova, S. S. Rev Roum Chim 2003, 48, 821.

32. Xu, J.; Liu, H.; Li, W.; Zou, H.; Xu, W. Macromol Theory Simul 2008, 17, 407.
33. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A. Markopoulos, J.; Igglessi-Markopoulou, O. Polymer 2006, 47, 3240.
34. Gharagheizi, F. Comput Mater Sci 2007, 40, 159.
35. Yu, X.; Wang, X.; Wang, H.; Li, X.; Gao, J. QSAR Comb Sci 2006, 25, 156.
36. Yu, X.; Yi, B.; Wang, X. Eur Polym Mater 2008, 44, 3997.
37. Yu, X.; Liu, W.; Liu, F.; Wang, X. J Mol Model 2008, 14, 1065.
38. Yu, X.; Wang, X.; Li, B. Colloid Polym Sci 2010, 288, 951.
39. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.4; TALETE SRL: Milan, 2006.
40. HYPERCHEM, Version 6.01; Hypercube, Inc.: Gainesville, 2000.
41. Liu, H.; Gramatica, P. Bioorg Med Chem 2007, 15, 5251.
42. Holder, A. J.; Yourtee, D. M.; White, D. A.; Glaros, A. G.; Smith, R. J Comput Aid Mol Des 2003, 17, 223.
43. Broyden, C. G. J Inst Maths Appl 1970, 6, 76.
44. Fletcher, R. Comput J 1970, 13, 317.
45. Goldfarb, D. Math Comput 1970, 24, 23.
46. Shannon, D. F. Math Comput 1970, 24, 647
47. Golbraikh, A.; Tropsha, A. J Mol Graph Model 2002, 20, 269.
48. Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics; Wiley-VCH: Weinheim, 2009.
49. Jansson, P. A. Anal Chem 1991, 63, 357A.
50. Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Environ Sci Chem 1994, 13, 941.
51. Qi, Y. -H.; Zhang, Q. -Y.; Xu, L. J Chem Inf Comput Sci 2002, 42, 1471.
52. Zheng, F.; Bayram, E.; Sumithran, S. P.; Ayers, J. T. Zhan, C. -G.; Schmitt, J. D.; Dwoskin, L. P.; Crooks, P. A. Bioorg Med Chem 2006, 14, 3017.
53. Guha, R.; Jurs, P. C. J Chem Inf Model 2005, 45, 800.
54. Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. J Phys. Chem 1996, 100, 10400.
55. Randic, M. J Chem Inf Comput Sci 1995, 35, 373.
56. Estrada, E. Internet Electron J Mol Des 2002, 1, 360
57. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. J Chem Inf Comput Sci 1996, 36, 1030.
58. Schuur, J.; Selzer, P.; Gasteiger, J. J Chem Inf Comput Sci 1996, 36, 334.
59. Consonni, V.; Todeschini, R.; Pavan, M. J Chem Inf Comput Sci 2002, 42, 682.
60. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. J Chem Inf Comput Sci 2002, 42, 693.