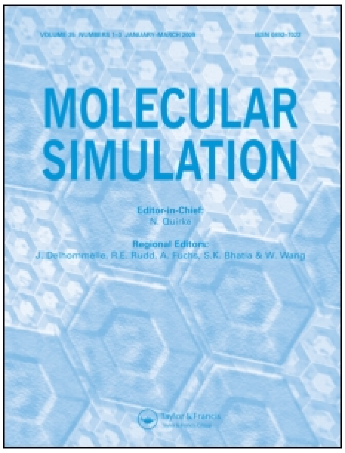

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Accurate quantitative structure-property relationship analysis for prediction of nematic transition temperatures in thermotropic liquid crystals

Jie Xu[a]; Luoxin Wang[a]; Hui Zhang[a]; Changhai Yi[a]; Weilin Xu[a]
[a] Key Lab of Green Processing and Functionalisation of New Textile Materials, Ministry of Education, Wuhan University of Science and Engineering, Wuhan, Hubei, P.R. China

First published on: 28 July 2009

To cite this Article Xu, Jie , Wang, Luoxin , Zhang, Hui , Yi, Changhai and Xu, Weilin(2010) 'Accurate quantitative structure-property relationship analysis for prediction of nematic transition temperatures in thermotropic liquid crystals', Molecular Simulation, 36: 1, 26 — 34, First published on: 28 July 2009 (iFirst)

To link to this Article: DOI: 10.1080/08927020903096064

URL: http://dx.doi.org/10.1080/08927020903096064

Molecular Simulation
Vol. 36, No. 1, January 2010, 26–34

Taylor & Francis
Taylor & Francis Group

Accurate quantitative structure–property relationship analysis for prediction of nematic transition temperatures in thermotropic liquid crystals

Jie Xu*, Luoxin Wang, Hui Zhang, Changhai Yi and Weilin Xu

Key Lab of Green Processing and Functionalisation of New Textile Materials, Ministry of Education, Wuhan University of Science and Engineering, Wuhan, Hubei 430073, P.R. China

(*Received 13 March 2009; final version received 3 June 2009*)

The quantitative structure–property relationship approach was performed to study the nematic transition temperatures (T_N) in thermotropic liquid crystals. The multi-linear regression analysis (MLRA) and artificial neural networks (ANNs) were employed to develop linear and nonlinear models, respectively. The proposed linear model contains five descriptors, with the squared correlation coefficient R^2 of 0.9837 and the standard error of estimation s of 2.31. The mean relative errors (MREs) for the training and test sets are 3.15 and 5.21%, respectively. Better predictive results were obtained from the nonlinear model: the MREs for the training and test sets are 2.03% ($R^2 = 0.9911$ and $s = 1.71$) and 1.92% ($R^2 = 0.9892$), respectively.

Keywords: QSPR; thermotropic liquid crystals; nematic transition temperature; MLRA; ANN

1. Introduction

Liquid crystals (LCs) are the aggregate states of matter whose properties are intermediate between a crystalline solid and an isotropic liquid [1]. The molecules in a crystal occupy specific sites in a lattice and point their molecular axes in specific directions. So, it may be stated that the order in a crystal is both positional and orientational. In contrast, the molecules in a liquid are not ordered and diffuse randomly throughout the container. On the other hand, the LCs flow like a liquid but due to their anisotropy they may have positional and/or orientational order as the solids. The optical, magnetic and mechanical properties of an LC depend on the direction in which these quantities are measured [2,3].

Liquid crystalline phases are known as mesophases and the molecules that are able to form these phases are called mesogens. Transitions to these phases are induced either by thermal processes yielding thermotropic LCs (TLCs) or by the effect of solvents giving lyotropic LCs. TLCs exhibit mesomorphic behaviour within a specific temperature range. They are either discotic, having planar disc-like molecules or calamitic, having cylinder-shaped rod-like molecules. The mesophases of TLCs are thermodynamically stable but only partially ordered phases. Each mesophase is described by its degree of order. If the mesophase has orientational order only, it is called nematic (*N*); if it has both orientational and positional order it is called smectic (*S*). The nematic LC phase is technologically the most important mesophase. It is used in almost all commercially available LC displays. On the other hand, the smectic LC phases have found very few commercially successful applications. Thus, the nematic LC phase must exist in an appropriate temperature range for the desired application. The nematic transition temperature (T_N), the upper temperature limit at which the nematic phase exists, is widely used as a measure of the nematic phase stability.

There have been some attempts to correlate molecular structure with T_N in TLCs. De Jeu and Van Der Veen [4] proposed the correlations of the anisotropic molecular polarisability with T_N to explain the well-known odd–even effect present in homologous series. Knaak et al. [5] developed a group contribution approach (GCA) for the prediction of T_N for structures containing two phenyl rings. This method was later refined and extended to include a wider diversity of substituents and linker groups but was still limited to two aromatic rings [6]. Kränz et al. [7] proposed a modified GCA combined with artificial neural networks (ANNs) for the prediction of T_N. The GCA can sometimes give prediction with reasonable accuracy, but a serious limitation of GCA is that this method is only applicable for the compound containing structural groups previously investigated.

An alternative approach to predicting T_N is quantitative structure–property relationship (QSPR). The QSPR is based on the assumption that the variation of the behaviour of the compounds, as expressed by any measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors [8]. The advantage of this method lies in the fact that it requires only

*Corresponding author. Email: xujie0@ustc.edu

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020903096064
http://www.informaworld.com

knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established, it can be applicable for the prediction of the property of new compounds that have not been synthesised or found. Thus the QSPR method can expedite the process of development of new molecules and materials with desired properties. The QSPR method has been used quite extensively to predict many physicochemical properties [1,9–16]. The first attempt to apply QSPR analysis for prediction of T_N is related to the work of Villanueva-García et al. [1] who obtained a nine-parameter model to predict T_N for a homologous series of 42 TLCs using multi-linear regression analysis (MLRA), with a correlation coefficient R^2 of 0.954 and a mean relative error (MRE) of 6.42% for the 29 compounds in the training set. There are too many descriptors involved in this model considering the number of samples in the training set. From a statistical viewpoint, introduction of descriptors into a QSPR should meet stringent criteria [17]. The ratio of samples to descriptors should be as high as possible, and at least 5:1 [18], whereas the ratio in this model (29:9) is less than 5:1. In addition, some of descriptors are highly inter-correlated (0.99) with one or two others and, therefore, leads to redundancy of information. The presence of the redundant features can cause many algorithms to focus attention on the idiosyncrasies of the individual samples and lose sight of the broad picture that is essential for generalisation beyond the training set [19–21]. This problem is compounded when the number of samples is also relatively small, as is often the case in molecular design. If the number of features is comparable to the number of training patterns, the parameters of the model may become unstable and unlikely to replicate if the study was to be repeated. A large number of available features also increase the risk of chance correlations [22]. More recently, Ren et al. [13] developed a five-parameter linear QSPR model for prediction of T_N for TLCs by means of heuristic method, with $R^2 = 0.9881$ and MRE $= 3.35\%$ for the training set. This equation predicts T_N for TLCs in the test set with an MRE of 9.20%. It is significant to study the nonlinear behaviour of the descriptors on T_N in TLCs as a complementary work (such as application of ANNs).

In this work, first the MLRA is applied to select the most statistically effective molecular descriptors on T_N in TLCs from a pool of many kinds of descriptors which are calculated by Dragon software [23]. Then a nonlinear model is obtained for predicting T_N in TLCs by means of ANNs, based on the results of MLRA.

2. Materials and methods

2.1 Data set

The same training and test sets of 42 TLCs as [1] were employed in this study. All the molecules contain two aromatic rings linked by an ester group, COO—, with different terminal chains (as shown in Figure 1). The molecular structures and their corresponding T_N values are listed in Table 1.

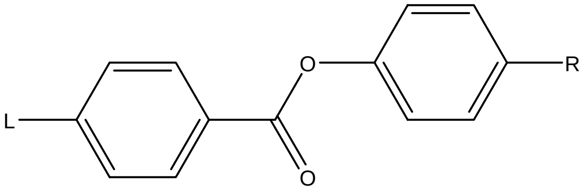

Figure 1. General structure for the studied compounds.

2.2 Descriptor calculation

The structures of all molecules were preoptimised using MM+ molecular mechanics method (Polak–Ribiere algorithm) in the HYPERCHEM program [24]. The final geometries of the minimum energy conformation were obtained by the semi-empirical AM1 method at a restricted Hartree–Fock level with no configuration interaction, applying a gradient norm limit of 0.01 kcal Å^{-1} mol^{-1} as a stopping criterion. Totally, 1664 molecular descriptors for each molecule were calculated on the resulting geometry through Dragon software [23]. These descriptors include: (a) 0D-constitutional (atom and group counts); (b) 1D-functional groups and atom centred 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.

In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.99 [25]) was excluded in a pre-reduction step. Thus, 409 molecular descriptors underwent subsequent descriptor selection.

2.3 Model development and validation

Stepwise MLRA with Leave-One-Out (LOO) cross-validation was used to select the descriptors for the linear QSPR models of the training set. F-to-enter and F-to-remove were 4 and 3, respectively. The R^2, the adjusted R^2, the cross-validated R^2, the F-ratio values, the standard error of estimation s and the significance level value p were used to measure the goodness of the models. The adjusted R^2 is calculated using the following formula:

$$R^2_{\text{adj}} = 1 - \left[\left(\frac{N-1}{N-M-1}\right)R^2\right], \qquad (1)$$

Table 1. Experimental and calculated T_N in °C for the studied TLC molecules.

				MLRA			ANN		
No.	*L*	*R*	Expt. T_N	Calc. T_N	ΔT_N[a]	RE (%)	Calc. T_N	ΔT_N[a]	RE (%)
1	C_5H_{11}—	O—CH_3	42.2	43.0	0.8	1.90	43.7	1.5	3.50
2	C_5H_{11}—	O—C_2H_5	63.4[b]	59.7	−3.7	5.83	63.6	0.2	0.34
3	C_5H_{11}—	O—C_3H_7	44.0	47.6	3.6	8.12	44.1	0.1	0.13
4	C_5H_{11}—	O—C_4H_9	57.7	52.8	−4.9	8.46	55.8	−1.9	3.24
5	C_5H_{11}—	O—C_5H_{11}	51.8	55.2	3.4	6.48	56.3	4.5	8.76
6	C_5H_{11}—	O—C_6H_{13}	59.3	59.8	0.5	0.88	59.5	0.2	0.36
7	C_5H_{11}—	O—C_7H_{15}	57.4	58.7	1.3	2.35	58.8	1.4	2.37
8	C_5H_{11}—	O—C_8H_{17}	60.6[b]	60.8	0.2	0.31	59.9	−0.7	1.22
9	C_5H_{11}—	O—C_9H_{19}	58.4[b]	61.0	2.6	4.41	58.3	−0.1	0.17
10	C_5H_{11}—	O—$C_{10}H_{21}$	60.3[b]	59.8	−0.5	0.79	60.1	−0.2	0.32
11	C_5H_{11}—	O—$C_{12}H_{25}$	60.6	58.9	−1.7	2.75	60.5	−0.1	0.19
12	C_5H_{11}—	O—$C_{14}H_{29}$	60.9	60.5	−0.4	0.61	62.1	1.2	2.01
13	C_6H_{13}—	O—CH_3	38.0[b]	34.6	−3.4	8.83	37.8	−0.2	0.47
14	C_6H_{13}—	O—C_2H_5	51.8	53.2	1.4	2.79	51.9	0.1	0.10
15	C_6H_{13}—	O—C_3H_7	36.0	42.4	6.4	17.64	38.6	2.6	7.24
16	C_6H_{13}—	O—C_4H_9	49.4[b]	45.9	−3.5	7.01	44.7	−4.7	9.57
17	C_6H_{13}—	O—C_5H_{11}	45.2[b]	50.0	5.0	11.13	46.6	1.6	3.62
18	C_6H_{13}—	O—C_6H_{13}	53.2	53.8	0.6	1.06	52.5	−0.7	1.26
19	CH_3—O—	C_3H_7—	40.0[b]	35.9	−4.1	10.22	39.9	−0.1	0.20
20	CH_3—O—	C_5H_{11}—	43.5	44.6	1.1	2.49	42.1	−1.4	3.27
21	C_6H_{13}—O—	C_5H_{11}—	63.0	59.2	−3.8	6.08	59.5	−3.5	5.49
22	C_4H_9—O—	O—C_8H_{17}	89.0	87.7	−1.3	1.46	88.2	−0.8	0.86
23	C_4H_9—O—	O—C_9H_{19}	86.0	86.9	0.9	1.04	87.9	1.9	2.22
24	C_4H_9—O—	O—$C_{10}H_{21}$	87.0	84.1	−2.9	3.37	87.4	0.4	0.42
25	C_4H_9—O—	O—$C_{12}H_{25}$	84.5	81.6	−2.9	3.47	82.0	−2.5	3.02
26	C_5H_{11}—O—	O—CH_3	72.0[b]	75.0	3.0	4.23	71.3	−0.7	0.92
27	C_5H_{11}—O—	O—C_2H_5	90.8	90.6	−0.2	0.19	92.4	1.6	1.81
28	C_5H_{11}—O—	O—C_3H_7	78.5	79.4	0.9	1.13	79.2	0.7	0.90
29	C_5H_{11}—O—	O—C_4H_9	82.0	79.9	−2.1	2.61	80.4	−1.6	1.98
30	C_5H_{11}—O—	O—C_5H_{11}	81.0	80.1	−0.9	1.17	82.2	1.2	1.52
31	C_5H_{11}—O—	O—C_6H_{13}	84.5	87.1	2.6	3.04	84.6	0.1	0.14
32	C_5H_{11}—O—	O—C_7H_{15}	82.0[b]	86.2	4.2	5.15	84.5	2.5	3.06
33	C_5H_{11}—O—	O—C_8H_{17}	85.0	81.7	−3.3	3.93	83.6	−1.4	1.69
34	C_5H_{11}—O—	O—C_9H_{19}	88.0	86.5	−1.5	1.69	85.2	−2.8	3.14
35	C_5H_{11}—O—	O—$C_{10}H_{21}$	82.0	82.1	0.1	0.15	82.2	0.2	0.29
36	C_5H_{11}—O—	O—$C_{12}H_{25}$	80.0[b]	84.1	4.1	5.08	83.1	3.1	3.88
37	C_5H_{11}—O—	O—$C_{14}H_{29}$	78.2[b]	80.7	2.5	3.25	78.4	0.2	0.24
38	C_5H_{11}—O—	O—$C_{16}H_{33}$	76.5[b]	77.6	1.1	1.46	77.2	0.7	0.94
39	C_5H_{11}—O—	O—$C_{18}H_{37}$	74.7	71.6	−3.1	4.09	74.7	0.0	0.06
40	C_6H_{13}—O—	O—CH_3	78.5	78.5	0.0	0.06	78.5	0.0	0.04
41	C_6H_{13}—O—	O—C_2H_5	95.9	95.2	−0.7	0.78	94.4	−1.5	1.61
42	$C_{10}H_{21}$—O—	O—C_6H_{13}	88.9	87.5	−1.4	1.52	87.9	−1.0	1.15

[a] ΔT_N = Calc. T_N − Expt. T_N. [b] Data used for the test set.

where *N* is the number of members of the training set and *M* is the number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2 (especially for correlations developed using small datasets), because R^2 is somewhat sensitive to changes in *N* and *M*. The adjusted R^2 corrects for the artificiality introduced when *M* approaches *N* through the use of a penalty function which scales the result. A variance inflation factor (VIF) was calculated to test if multi-collinearities existed among the descriptors, which is defined as

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

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 10 [26].

The proposed model was also checked for reliability and robustness by randomisation tests: new models were recalculated with randomly reordered T_N values. The resulting models obtained on the training set with randomised T_N values should have significantly lower R^2 values than the proposed one because the relationship between the structure and property is broken. This is a proof of the proposed model's validity as it can be reasonably excluded that the originally proposed model was obtained by chance correlation.

The nonlinear model was then developed by submitting the selected descriptors from MLRA to three-layer, fully connected, feed-forward ANNs. The number of input neurons was equal to that of the descriptors in the linear model. The number of hidden neurons was optimised by trial and error procedure on calculations of the training process. One output neuron was used to represent the experimental T_N. The network was trained using the quasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [27]. To avoid overtraining, one-tenth data from the training set were randomly selected as a 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 $R^2_{\mathrm{CV,ext}}$ for the test sets is determined using Equation (3):

$$R^2_{\mathrm{CV,ext}} = 1 - \frac{\sum_{i=1}^{\mathrm{test}} (y_{\exp t} - y_{\mathrm{calc}})^2}{\sum_{i=1}^{\mathrm{test}} (y_{\exp t} - \bar{y}_{\mathrm{test}})^2}, \qquad (3)$$

where $\bar{y}_{\mathrm{test}}$ is the averaged value for the response variable of the test set. According to [28], a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{\mathrm{CV,ext}} > 0.5,$$

$$r^2 > 0.6,$$

$$|(r^2 - r_0^2)/r^2| < 0.1 \text{ or } |(r^2 - r_0'^2)/r^2| < 0.1,$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15,$$

where r is the correlation coefficient between the calculated values and experimental values in the test set. r_0^2 (calculated versus observed values) and $r_0'^2$ (observed versus 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 [28].

3. Results and discussion

3.1 MLRA model and interpretation

A linear model was developed using stepwise MLRA with LOO cross-validation and the number of descriptors in the final model was determined on the basis of the data set size and on the basis of the R^2, adjusted R^2, cross-validated R^2, F, s and p. The R and s results during the stepwise MLRA are shown in Figure 2. Obviously, T_N is not linearly correlated with any of the molecular descriptors since univariant correlations between T_N and the different descriptors have poor R^2 and s values. The seven-descriptor equation has the best R^2 and s values. However, from a statistical viewpoint the ratio of the number of the samples (N) and the number of descriptors in the correlation (M) should not be too low. Usually, it is recommended that $N/M \ge 5$ [18]. In this study, there are 29 samples for the training set, so 5 descriptors were selected. The final correlation equation is the following:

$$\begin{aligned} T_N = {} & 262.798[\mathrm{ESpm03d}] + 0.874[\mathrm{RDF135u}] \\ & + 77.455[\mathrm{Mor26v}] - 67.699[\mathrm{Mor08p}] \\ & + 5.689[\mathrm{H-052}] - 1164.785, \qquad (4) \end{aligned}$$

$$N = 29, R^2 = 0.9837, R^2_{\mathrm{CV}} = 0.9809, R^2_{\mathrm{adj}} = 0.9831, s = 2.31, F = 321.83.$$

Here, ESpm03d is the spectral moment 03 from edge adjacency matrix weighted by dipole moments [29–33]; RDF135u is the radial distribution function (RDF) – 13.5/unweighted [34]; Mor26v is the 3D-MoRSE – signal 26/weighted by atomic van der Waals volumes [35,36]; Mor08p is the 3D-MoRSE – signal 08/weighted by atomic polarisabilities [35,36]; H-052 is the number of H atoms attached to $C^0(sp^3)$ with 1X attached to next C, where the superscript represents the formal oxidation number and X

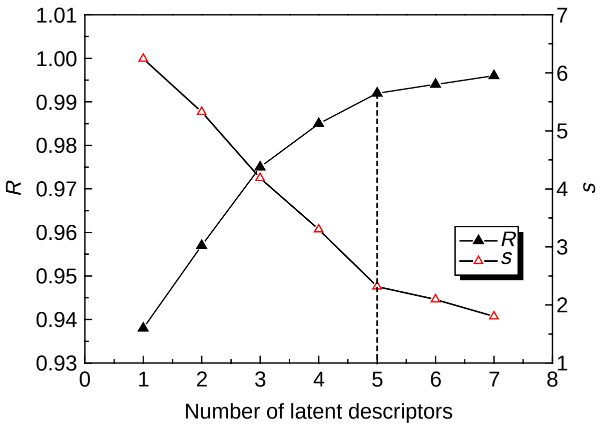

Figure 2. R and s versus number of latent descriptors in the best MLRA equation.

Table 2. Results of randomisation test.

Iteration	R^2	R^2_{CV}	Iteration	R^2	R^2_{CV}
1	0.5484	0	11	0.4362	0
2	0.5649	0	12	0.3224	0
3	0.5175	0	13	0.5138	0
4	0.3083	0	14	0.3634	0
5	0.6447	0	15	0.5845	0
6	0.6182	0.0623	16	0.3523	0
7	0.1746	0	17	0.5004	0
8	0.6407	0	18	0	0
9	0.0614	0	19	0.7151	0
10	0.3934	0	20	0	0

represents any electronegative atom (O, N, S, P, Se and halogens) [37]. The formal oxidation number of a carbon atom equals the sum of the conventional bond orders with electronegative atoms; the C—N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds; the C—X bond order in pyrrole or furan may be considered as one.

Usually, the larger the magnitude of the *F*-ratio, the better the model predicts the property values in the training set. The large *F*-ratio of 321.83 indicates that Equation (4) does an excellent job of predicting the T_N values. Equation (4) has an adjusted R^2 value of 0.9831, which indicates very good agreement between the correlation and the variation in the data. The cross-validated correlation coefficient $R^2_{CV} = 0.9809$ shows the reliability of the model by focusing on the sensitivity of the model to the elimination of any single data point [38]. The model was further validated by applying the randomisation test and several results are shown in Table 2. The low R^2 and R^2_{CV} values indicate that the good results of the original model are not due to chance correlation or structural dependency of the training set. The statistical characteristics of the five descriptors are given in Table 3, which indicate that all the descriptors are highly significant from the *t*-test values. The VIF values and the correlation matrix as shown in Table 4 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 T_N values from Equation (4) for the whole data set are shown in Table 1 and Figure 3. The distributions of relative errors (REs) are given in Figure 4. The statistical results from different QSPR models are collected in Table 5. The min/max values of the REs are 0.06%/17.64% and 0.31%/11.13% for the training and test sets, respectively. The MREs for the training and test sets are 3.15 and 5.21%, respectively, which are better than those of the HM model by Ren et al. [13] (3.35 and 9.20%) and the MLRA model by Villanueva-García et al. [1] (6.83 and 11.53%). In the training set, 82.76% of the compounds have REs less than 5%, and only compound **15** has RE larger than 10% (17.64%). In the test set, 46.15% of the compounds show REs less than 5%, and 84.62% have REs less than 9%, with compounds **9** and **11** showing large REs of 10.22 and 11.13%, respectively. For the HM model developed by Ren et al. [13], 93.10% of the compounds in the training set show REs less than 5%, but two compounds, **4** and **15**, show large REs of 14.68 and 21.93%, respectively. Within the test set, 61.53% of the compounds have REs less than 5%, and 69.23% show REs less than 9%, with compounds **8**, **13**, **17** and **19** showing large REs of 12.23, 18.76, 18.44 and 41.16%, respectively. For the MLRA model derived by Villanueva-García et al. [1], 55.17% of the compounds

Table 3. Characteristics of descriptors in the best MLRA model.

Descriptor	Descriptor type	SE	*t*-Value	*t*-Probability	VIF
Constant		126.140	−9.234	0.000 000	
ESpm03d	Edge adjacency indices	27.530	9.546	0.000 000	6.254
RDF135u	RDF descriptors	0.106	8.225	0.000 000	3.465
Mor26v	3D-MoRSE descriptors	15.158	5.110	0.000 011	3.487
Mor08p	3D-MoRSE descriptors	8.715	−7.768	0.000–000	2.734
H-052	Atom-centred fragments	0.871	6.533	0.000–000	6.809

Table 4. Correlation matrix of the selected descriptors and T_N.

	T_N	ESpm03d	RDF135u	Mor26v	Mor08p	H-052
T_N	1					
ESpm03d	0.938	1				
RDF135u	0.412	0.362	1			
Mor26v	−0.089	−0.153	−0.780	1		
Mor08p	0.434	0.620	0.559	−0.519	1	
H-052	0.882	0.823	0.327	−0.188	0.667	1

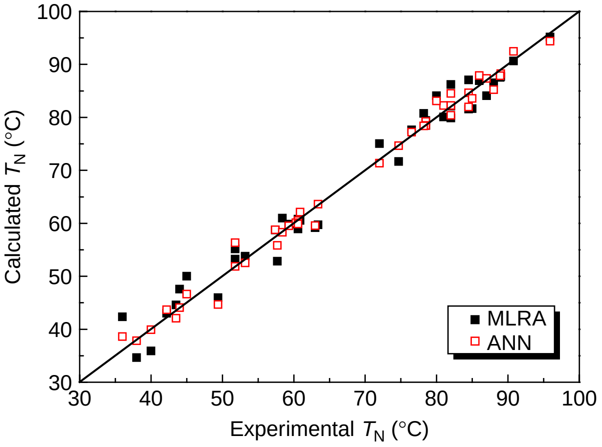

Figure 3. Calculated versus experimental T_N with MLRA and ANN for the whole data set.

in the training set show REs less than 5%, and 79.31% have REs less than 10%, with compound **15** showing a large RE of 47.78%. In the case of the test set, only 15.38% compounds show REs less than 5%, and 53.85% show REs less than 10%, with molecule 17 having a large RE of 33.78%. Therefore, the present MLRA model is suitable for making accurate prediction for compounds outside the training set, i.e. the model could be extrapolated.

It is noteworthy that all the above mentioned models give poor predictive results for compound **15**. Usually, the T_N values are difficult to be determined as experimentally measured by differential scanning calorimetry (DSC) because the transition takes place over a comparatively wide range of temperature. In addition, the obtained T_N values are strongly affected by the heating rate and other factors in the DSC measurements. Thus, this compound is considered as an outlier due to the experimental uncertainties.

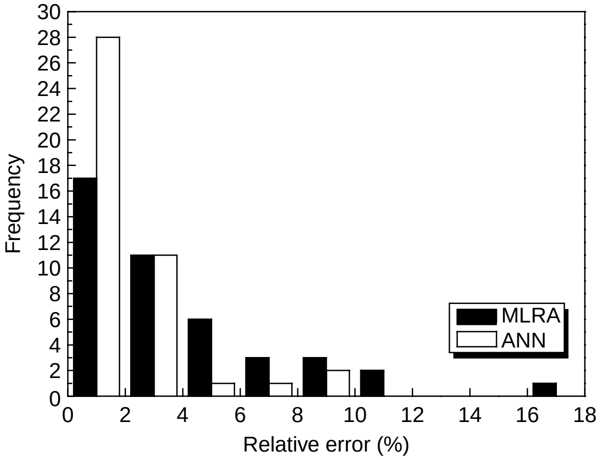

Figure 4. Distributions of REs calculated with MLRA and ANN for the whole data set.

By interpreting the descriptors involved in Equation (4), it is possible to gain some insights into the factors that may affect the T_N values in TLCs. According to the *t*-test (in Table 3), the first important descriptor in Equation (3) is the descriptor ESpm03d, which is calculated by summing the diagonal elements of the third power of the edge adjacency matrix weighted by dipole moments [29–33]. The coefficient of this descriptor is positive, meaning that compounds with greater dipole moments would have higher T_N values.

Mor26v and Mor08p are 3D representations of molecular structures based on electron diffraction descriptor (3D-MoRSE descriptor) [35,36] which is 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 0–31 Å^{-1} from the three dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using the following expression

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}}, \qquad (5)$$

where s is the scattering angle, r_{ij} is the interatomic distance between ith and jth atom, A_i and A_j are atomic properties of ith and jth atom, respectively, including atomic number, masses, van der Waals volumes, Sanderson electronegativities and polarisabilities. The presence of Mor26v and Mor08p in Equation (4) reflect the influence of atomic volumes and polarisabilities on the T_N values, respectively. The coefficient of Mor26v is positive, indicating that an increase in Mor27v would result in an increase in T_N values; while the negative sign of Mor08p indicates that an increase in this descriptor could lead to a decrease in T_N values. However, the value and sign of the 3D-MoRSE descriptor depend, to a large extent, on the values of s and r_{ij} [39]. Thus, it could not be concluded that atomic volumes and polarisabilities have a specific effect on the T_N values, either negative or positive, only taking into account the coefficient sign of the descriptor in the present MLRA model. When the coefficient and the descriptor have the same sign, the contribution of atomic volumes and polarisabilities is positive, else, negative.

RDF135u is one of the RDF descriptors which has been proposed based on a RDF [34]. The RDF descriptors can be interpreted as the probability distribution of finding an atom in a spherical volume of radius r. The general form of the RDF is represented by:

$$\mathrm{RDF}rw = f \cdot \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i \cdot w_j \cdot \mathrm{e}^{-\beta(r-r_{ij})^2}, \qquad (6)$$

Table 5. Comparison of different QSPR models for T_N.

	Training set				Test set		
Model	R^2	s	MRE (%)	Max RE (%)	R^2	MRE (%)	Max RE (%)
Present MLRA	0.9837	2.31	3.15	17.64	0.9766	5.21	11.13
Present ANN	0.9911	1.71	2.03	8.76	0.9892	1.92	9.57
García et al. (MLRA)	0.9540	6.42	6.83	47.78	0.7922	11.53	33.78
Ren et al. (HM)	0.9881	3.07	3.35	21.94	0.9216	9.20	41.25

where f is a scaling factor (assumed to be equal to one in the calculations), w_i and w_j are characteristic properties of the atoms i and j (including unweighted cases, atomic number, masses, van der Waals volumes, Sanderson electronegativities and polarisabilities), r_{ij} is the interatomic distance, and nAT is the number of atoms in the molecule. RDFrw is generally calculated at a number of discrete points with defined intervals. As an unweighted case among the RDF descriptors, RDF135u mainly provides information about interatomic distances in the entire molecule. The coefficient of RDF135u is positive, indicating that an increase in RDF135u would lead to an increase in T_N values.

The number of H atoms attached to $C^0(sp^3)$ with 1X (X=O) attached to next C (H-052) is an atom-centred descriptor calculated by knowing the molecular composition and atom connectivities. This descriptor encodes information about the hybridisation and oxidation state of the carbon atoms with an oxygen atom attached to next C. The positive sign of H-052 indicates that compounds with more H atoms attached to $C^0(sp^3)$ with 1X (X=O) attached to next C would have higher T_N values.

From the MLRA model by Villanueva-García et al. [1], the dipole moment magnitude (D) is one of the most significant descriptors, which is the same as the ESpm03d in our correlation; the dummy descriptors L and R were employed to reflect the number of oxygen atoms present in the terminal chain, which has the similar physical meaning as the relative number of oxygen atoms in a molecule in the HM model by Ren et al. [13] and the H-052 in our correlation. Thus, the present model has expressed the main factors affecting the generation of nematic phases in the TLCs.

3.2 *ANN model*

Recently, there is a growing interest in the use of ANNs for QSPR due to their inherent ability in modelling a nonlinear problem. The ANNs are especially useful when a rigid theoretical basis or mathematical relationship to describe a phenomenon to be modelled is not available. Among the neural network learning algorithms, the back-propagation (BP) method [40] is one of the most commonly used methods. The drawback of BP is that the training is processed slowly, because the gradient-descent algorithm is usually used for minimising the sum-of-squares error. In this study, the quasi-Newton BFGS algorithm was used to develop nonlinear models. The advantages of using the BFGS algorithm are that the specifying rate or momentum is not necessary and training processes are much more rapid [41].

The 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 ANNs. 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 [42]. Thus, a 5–3–1 network architecture is obtained after a rigorous trial and error procedure. The prediction results from the nonlinear ANN model are given in Table 1 and Figure 3. The distributions of REs calculated from the ANN model are also shown in Figure 4. The MREs for the training and test sets are 2.03% ($R^2 = 0.9911$ and $s = 1.71$) and 1.92% ($R^2 = 0.9892$), respectively. The min/max REs for the training and test sets are 0.04%/8.76% and 0.17%/9.57%, respectively. For the whole dataset, 90.47% of the compounds have REs less than 5%. These results show some modification (especially for the test set) in comparison to the MLRA model, which confirms the nonlinear relationship between structural information and nematic transition temperatures in thermotropic liquid.

The following statistical parameters were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the ANN model:

$$R^2_{\mathrm{CV,ext}} = 0.9866 > 0.5,$$
$$r^2 = 0.9894 > 0.6,$$

$$|(r^2 - r_0^2)/r^2| = |(0.9894 - 0.9994)/0.9894| < 0.1,$$

$$\text{or}\, |(r^2 - r'^2_0)/r^2| = |(0.9894 - 0.9995)/0.9894| < 0.1,$$

$$0.85 \le k = 0.994 \le 1.15\, \text{or}\, 0.85 \le k' = 1.005 \le 1.15.$$

4. Conclusion

In this paper, linear and nonlinear QSPR models to predict T_N in TLCs with acceptable accuracy are presented based on descriptors calculated by Dragon. A five-parameter linear model is obtained by MLRA, with an R^2 of 0.9837 and s of 2.31 for the training set. The nonlinear model appears to be more reliable than the linear model. The MREs from the linear and nonlinear models for the training set are 3.15 and 2.03%, respectively. Satisfactory prediction results for the test set (MRE of 5.21 and 1.92%, respectively) make the models very useful for the prediction of T_N of not yet synthesised TLCs. Thus, these QSPR models can expedite the process of development of new TLCs with desired nematic phase stability.

Acknowledgements

This work was supported by the Foundation of Wuhan University of Science and Engineering (No. 20073208), the Natural Science Foundation of Hubei Province (No. 2008CDB261) and the Key Project of Science and Technology Research of Ministry of Education (No. 208089). The authors gratefully wish to express their thanks to all reviewers for critically reviewing the manuscript and making important suggestions from their first submission on 29 August, 2008.

References

[1] M. Villanueva-García, R.N. Gutiérrez-Parra, A. Martínez-Richa, and J. Robles, *Quantitative structure-property relationships to estimate nematic transition temperatures in thermotropic liquid crystals*, J. Mol. Struct. (THEOCHEM) 727 (2005), pp. 63–69.

[2] S. Singh, *Phase transitions in liquid crystals*, Phys. Rep. 324 (2000), pp. 107–269.

[3] K. Binnemans, *Ionic liquid crystals*, Chem. Rev. 105 (2005), pp. 4148–4204.

[4] W.H. De Jeu and J. Van Der Veen, *Molecular structure and nematic liquid crystalline behaviour*, Mol. Cryst. Liq. Cryst. 40 (1977), pp. 1–17.

[5] L.E. Knaak, H.M. Rosenberg, and P. Servé, *Estimation of nematic-isotropic points of nematic liquid crystals*, Mol. Cryst. Liq. Cryst. 17 (1972), pp. 171–185.

[6] T. Thienmann and V. Volkmar, *Development of an incremental system for the prediction of the nematic-isotropic phase transition temperature of liquid crystals with two aromatic rings*, Liq. Cryst. 22 (1997), pp. 519–523.

[7] H. Kränz, V. Vill, and B. Meyer, *Prediction of material properties from chemical structures. The clearing temperature of nematic liquid crystals derived from their chemical structures by artificial neural networks*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 1173–1177.

[8] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, and T. Fan, *Radial basis function neural network-based QSPR for the prediction of critical temperature*, Chemom. Intell. Lab. Syst. 62 (2002), pp. 217–225.

[9] J. Xu, B. Chen, Q. Zhang, and B. Guo, *Prediction of refractive indices of linear polymers by a four-descriptor QSPR model*, Polymer 45 (2004), pp. 8651–8659.

[10] J. Xu, B. Guo, B. Chen, and Q. Zhang, *A QSPR treatment for the thermal stabilities of second-order NLO chromophore molecules*, J. Mol. Model. 12 (2005), pp. 65–75.

[11] X. Yu, X. Wang, J. Gao, X. Li, and H. Wang, *QSPR studies of polyvinyls by density functional theory*, Polymer 46 (2005), pp. 9443–9451.

[12] J. Xu, Z. Zheng, B. Chen, and Q. Zhang, *A linear QSPR model for prediction of maximum absorption wavelength of second-order NLO chromophores*, QSAR Comb. Sci. 25 (2006), pp. 372–379.

[13] Y. Ren, H. Liu, X. Yao, M. Liu, and B. Fan, *Prediction of nematic transition temperatures in thermotropic liquid crystals by a heuristic method*, Liq. Cryst. 34 (2007), pp. 1291–1297.

[14] J. Xu, L. Liu, W. Xu, S. Zhao, and D. Zuo, *A general QSPR model for the prediction of ϑ (lower critical solution temperature) in polymer solutions with topological indices*, J. Mol. Graph. Model. 26 (2007), pp. 352–359.

[15] S. Yin, Z. Shuai, and Y. Wang, *A quantitative structure–property relationship study of the glass transition temperature of OLED materials*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 970–997.

[16] M.H. Fatemi and M. Haghdadi, *Quantitative structure–property relationship prediction of permeability coefficients for some organic compounds through polyethylene membrane*, J. Mol. Struct. 886 (2008), pp. 43–50.

[17] M.T.D. Cronin and W. Schultz, *Pitfalls in QSAR*, J. Mol. Struct. (THEOCHEM) 622 (2003), pp. 39–51.

[18] J.G. Topliss and J. Costello, *Chance correlations in structure-activity studies using multiple regression analysis*, J. Med. Chem. 15 (1972), pp. 1066–1068.

[19] T.A. Andrea and H. Kalayeh, *Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors*, J. Med. Chem. 34 (1991), pp. 2824–2836.

[20] S.-S. So and G. Richards, *Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors*, J. Med. Chem. 35 (1992), pp. 3201–3207.

[21] D.T. Manallack, D.D. Ellis, and J. Livingstone, *Analysis of linear and nonlinear QSAR data using neural networks*, J. Med. Chem. 37 (1994), pp. 3758–3767.

[22] J.G. Topliss and P. Edwards, *Chance factors in studies of quantitative structure-activity relationships*, J. Med. Chem. 22 (1979), pp. 1238–1244.

[23] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, *DRAGON for Windows (Software for Molecular Descriptor Calculations)*, TALETE srl, Milan, 2006.

[24] *HYPERCHEM*, Hypercube, Inc., Gainesville, 2000.

[25] H. Liu and P. Gramatica, *QSAR study of selective ligands for the thyroid hormone receptor β*, Bioorgan. Med. Chem. 15 (2007), pp. 5251–5261.

[26] G.W. Kauffman and C. Jurs, *Prediction of inhibition of the sodium ion-proton antiporter by benzoylguanidine derivatives from molecular structure*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 753–761.

[27] M.D. Wessel and C. Jurs, *Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks*, Anal. Chem. 66 (1994), pp. 2480–2487.

[28] A. Golbraikh and A. Tropsha, *Beware of q^2!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.

[29] E. Estrada, *Edge adjacency relationships and a novel topological index related to molecular volume*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 31–33.

[30] E. Estrada, *Edge adjacency relationships in molecular graphs containing heteroatoms: a new topological index related to molar volume*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 701–707.

[31] E. Estrada and A. Ramirez, *Edge adjacency relationships and molecular topographic descriptors. Definition and QSAR applications*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 837–843.

[32] E. Estrada, *Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 844–849.

[33] E. Estrada, *Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 320–328.

[34] M.C. Hemmer, V. Steinhauer, and J. Gasteiger, *Deriving the 3D structure of organic molecules from their infrared spectra*, Vibrat. Spect. 19 (1999), pp. 151–164.

[35] J. Schuur, P. Selzer, and J. Gasteiger, *The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 334–344.

[36] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, and V. Steinhauer, *Chemical information in 3D space*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 1030–1037.

[37] V.N. Viswanadhan, A.K. Ghose, G.R. Revankar, and K. Robins, *Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics*, J. Chem. Inf. Comput. Sci. 29 (1989), pp. 163–172.

[38] R. Murugan, M.P. Grendze, J.E. Toomey, A.R. Katritzky, M. Karelson, V. Lobanov, and P. Rachwal, *Predicting physical properties from molecular structure*, Chem. Tech. 24 (1994), pp. 17–23.

[39] L. Saiz-Urra, M.P. Gonzalez, and M. Teijeira, *QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection*, Bioorg. Med. Chem. 14 (2006), pp. 7347–7358.

[40] P.A. Jansson, *Neural networks: an overview*, Anal. Chem. 63 (1991), pp. 357A–363A.

[41] L. Xu, J.W. Ball, S.L. Dixon, and C. Jurs, *Quantitative structure–activity relationships for toxicity of phenols using regression analysis and computational neural networks*, Environ. Toxicol. Chem. 13 (1994), pp. 841–851.

[42] Y. Qi, Q. Zhang, and L. Xu, *Correlation analysis of the structures and stability constants of gadolinium(III) complexes*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 1471–1475.