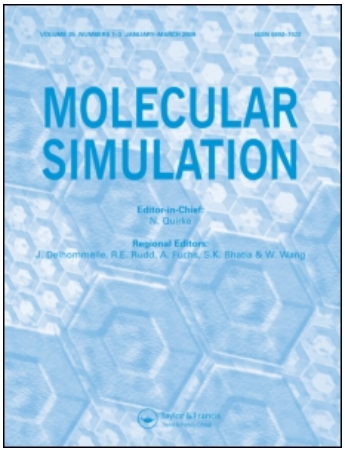


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

QSPR model to predict the thermal stabilities of second-order nonlinear optical (NLO) chromophore molecules

F. Luan[a]; H. T. Liu[a]; Y. Gao[a]; X. Y. Zhang[b]
[a] Department of Applied Chemistry, Yantai University, Yantai, P.R. China [b] Department of Chemistry, Lanzhou University, Lanzhou, P.R. China

To cite this Article Luan, F. , Liu, H. T. , Gao, Y. and Zhang, X. Y.(2009) 'QSPR model to predict the thermal stabilities of second-order nonlinear optical (NLO) chromophore molecules', Molecular Simulation, 35: 3, 248 — 257
To link to this Article: DOI: 10.1080/08927020802378928
URL: http://dx.doi.org/10.1080/08927020802378928

Molecular Simulation
Vol. 35, No. 3, March 2009, 248–257

Taylor & Francis
Taylor & Francis Group

QSPR model to predict the thermal stabilities of second-order nonlinear optical (NLO) chromophore molecules

F. Luan[a]*, H.T. Liu[a], Y. Gao[a] and X.Y. Zhang[b]

[a]*Department of Applied Chemistry, Yantai University, Yantai, P.R. China;* [b]*Department of Chemistry, Lanzhou University, Lanzhou, P.R. China*

(*Received 21 February 2008; final version received 21 July 2008*)

We aimed to establish the quantitative structure–property relationship (QSPR) with thermal stabilities of 90 structurally diverse second-order nonlinear optical (NLO) chromophore molecules using easily understood and obtained physicochemical molecular descriptors. These descriptors include number of C atoms (NC), number of F atoms (NF), relative number of N atoms (RNN) and relative number of triple bonds (RNTB). Multiple linear regression (MLR) and support vector machine (SVM) were used to develop the linear and nonlinear models, respectively. The obtained linear model had a square of correlation coefficient $R^2 = 0.928$ with a root mean square error (RMS) error of 19.28 for the training set, while $R^2 = 0.828$, RMS = 29.79 for the test set. The RMS error in prediction for overall data set is 21.65. The nonlinear model gave better results: for the training set $R^2 = 0.949$, RMS = 16.16, and for the test set $R^2 = 0.867$, RMS = 25.40. The RMS error in prediction for overall data set is 18.38. The proposed model comprises only four descriptors, which can be obtained by simple calculation and makes it more convenient to predict the thermal stabilities of NLO materials.

Keywords: QSPR; thermal stabilities; nonlinear optical (NLO) chromophores

1. Introduction

Organic second-order nonlinear optical (NLO) materials have gained much attention due to their larger susceptibilities, faster response time, ease of processing and versatility of molecular structural modifications [1–3]. Attempts are being made to incorporate NLO materials into various devices with diverse purposes, such as optical communication, computing and data storage, and image processing [4]. The design and synthesis of optimised molecules is a key goal in this area of research [5–8]. These materials are typically made from small organic molecules, namely chromophores, incorporated into polymer matrices and poled with an electric or optical field to realise a non-centrosymmetric dipole alignment. The physical and thermodynamic properties of organic compounds are needed in the above processes, especially for the unknown or un-synthesised ones. These fundamental properties which are required include high molecular nonlinearity, good solubility, good thermal stability, low cut-off wavelength and so on [9–12].

Experimental techniques are usually developed to measure the properties. For example, the T_d values for chromophores are measured by thermo gravimetric analysis (TGA) or differential scanning calorimetry (DSC) [13]. As we know, the experimental determination of the property is time-consuming and expensive. Besides, there may be other reasons for us to obtain the desired property in advance, for example the property of interest has for some reason not been measured experimentally, or that we want to estimate it before the same had been synthesised. Hence, it is convenient to have simple models that provide estimates for experimental values. An interesting solution to this problem has been found in (quantitative structure–property relationship) QSPR models based on experimental values for structurally similar compounds. The advantages of this approach lie in the fact that it requires only the knowledge of chemical structure and is not dependent on any experiment properties once the model was built.

There have been a number of attempts to model the relationship between fundamentally required properties and structure of organic second-order NLO materials. Oberg et al. predicted the NLO quantities, second and third harmonics (β and γ), using a QSPR approach. Molecular orbital *ab initio* calculations were applied to generate easily accessible variables to be used in the partial least-squares analysis. A successful QSPR model was developed for the prediction of the nonlinearities of 22 chromophores with five quantum-chemical descriptors involved [14]. Zeng et al. [15] obtained a three-parameter correlation to predict the nonlinearities for 32 para-disubstituted benzenes. Bicerano et al. [16] performed a QSPR relationship for the prediction of T_d for a set of 140 polymers, with 21 descriptors involved, using the molar

*Corresponding author. Email: fluan@sina.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020802378928
http://www.informaworld.com

thermal decomposition function Y_d as the dependent variable. Li et al. investigated the relationship between the experimental λ_{max} of 31 azo dyes and their structural parameters generated by the PM3 computation [17]. Recently, a linear QSPR model was reported by Xu et al. to predict the maximum absorption wavelength of 72 second-order NLO chromophores [18]. Later, the same authors conducted a systematic study between descriptors representing the molecular structures and thermal decomposition temperatures (T_d) for a diverse set of 90 second-order NLO chromophores [19].

All of the previous studies attempt to produce an easy, accurate and predictive model using different descriptors or methods. However, some models have been developed for relatively small data set of compounds. Some models, however, include a large number of descriptors, leading to the difficulty of the explanation of the physical meaning of the descriptors. Additionally, previous studies scarcely use the nonlinear statistical technique to build the model.

In the present work, constitutional descriptors, the simplest molecular descriptors, were used for the prediction of thermal stabilities of 90 structurally diverse second-order NLO chromophore molecules. Multiple linear regression (MLR) was used for the pre-selection of appropriate molecular descriptors and to build the linear model. Nonlinear support vector machine (SVM) model was also used to explore the possibility of establishing an accurate QSPR model. The purpose of this work is to improve the general model development by simple descriptors, to establish a simple, yet acceptable model and to seek for the structural features related to the thermal stabilities of second-order NLO compounds.

2. Methods

2.1 *Data set and* Y_d *values*

This study was performed on a data set taken from [19]. The molecular structures of the 90 NLO chromophores were of extensive structural diversity, and they were shown in Figure 1. The experimental molar thermal decomposition functions Y_d (defined as T_d multiplied by the molecular weight M) as the dependent variable were presented in Table 1. The entire set of compounds was divided into two subsets randomly for both linear and nonlinear models: a training set of 72 compounds, whose information was used to build the actual models, and the test set including the remaining 18 compounds, which was used to validate the models once they were built. Each of the compounds in test set was also marked with an asterisk in Table 1. For the data sets, the number of samples vs. the experimental Y_d is shown in Figure 2, which illuminates the diversity of the molecules in the training and prediction sets. As can be seen from the figure, the compounds are diverse in both sets. The training set with a broad representation of the chemistry space was adequate to ensure models' stability and the diversity of prediction set can prove the predictive capability of the model.

2.2 *Structure entry and descriptor generation*

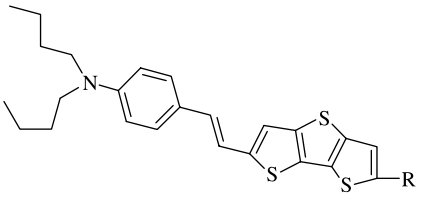
All structures of the NLO chromophore molecules were drawn with the HyperChem program [20] and exported in a file format suitable for MOPAC [21]. All calculations were carried out at restricted Hartree–Fock level with no configuration interaction. The molecular structure of each compound was optimised using the Polak–Ribiere algorithm until the RMS gradient was 0.1. A more precise optimisation is done with the semi-empirical PM3 method in MOPAC6.0. The resulting geometry was transferred into software CODESSA, developed by the Katritzky group [22,23], which can calculate constitutional, topological, geometrical, electrostatic and quantum chemical descriptors.

Among these descriptors, constitutional descriptors are the simplest ones. They depend only on the constitution of the molecule, providing information about which atoms or bonds the molecule consists of. One such descriptor may, for instance, be the number of carbon or oxygen atoms in molecule. The advantages of these descriptors are that, they can be easily calculated, that is, they can be obtained through counting only the number of the atoms or bonds in the structure. In the present work, 35 constitutional descriptors were calculated.

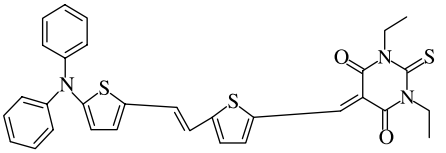
3. Methodology

3.1 *Multiple linear regression*

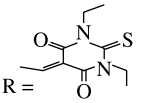
After calculating of the constitutional descriptors, the heuristic method in CODESSA was used to select descriptors [23,24]. The advantages of this method are: the high speed usually produces correlations 2–5 times faster than other methods, with comparable quality [25]; and no software restrictions on the size of the data set. It can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly inter-correlated. The heuristic method of the descriptor selection proceeds with a pre-selection of descriptors by eliminating: those descriptors that are not available for each structure; descriptors having a small variation in magnitude for all structures; descriptors that give a F-test's value below 1.0 in the one-parameter correlation; and descriptors whose t-values are less than the user-specified value, etc. This procedure orders the

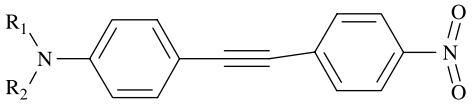


DBDTTC: R = CHO

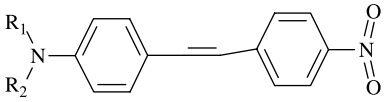


DTDPD

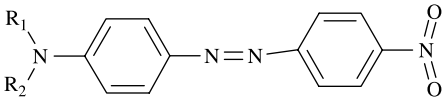
DBDTTDC: R =  DBDTTDET: R = 



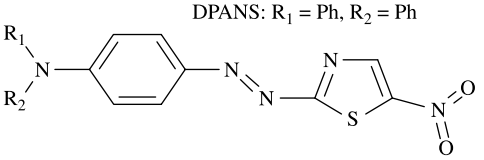
DMNPA: R_1 = Me, R_2 = Me
DPNPA: R_1 = Ph, R_2 = Ph



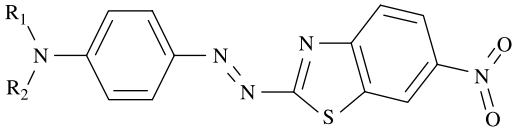
DANS: R_1 = Me, R_2 = Me
DPANS: R_1 = Ph, R_2 = Ph



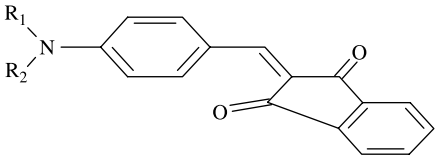
DMNPAPA: R_1 = Me, R_2 = Me
DENPAPA: R_1 = Et, R_2 = Et
DPNPAPA: R_1 = Ph, R_2 = Ph



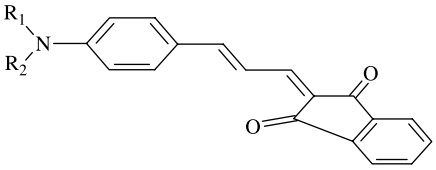
ENTPAH: R_1 = Et, R_2 = HO $(CH_2)_6$
NTPDA: R_1 = Ph, R_2 = Ph



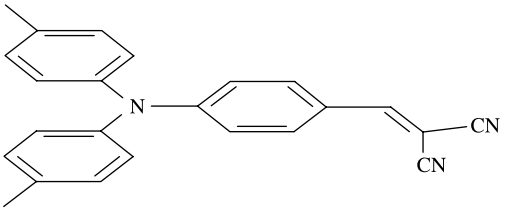
ENBTPAH: R_1 = Et, R_2 = $HO(CH_2)_6$
NBTPDA: R_1 = Ph, R_2 = Ph



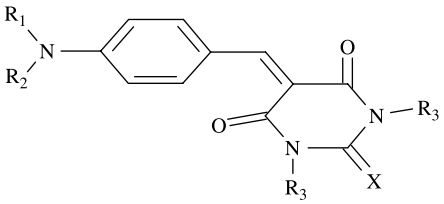
DEBID: R_1 = Et, R_2 = Et
p-DTBID: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4



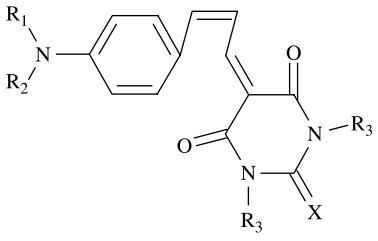
DEPAID: R_1 = Et, R_2 = Et
p-DTPAID: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4



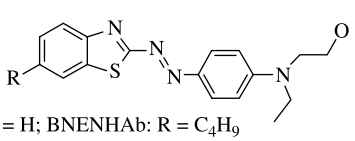
p-DTABM



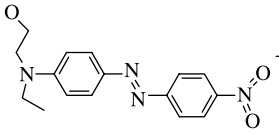
DEBDMPT: R_1 = Et, R_2 = Et, R_3 = Me, X = O
p-DTBDMPT: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4, R_3 = Me, X = O
DMBDETPD: R_1 = Me, R_2 = Me, R_3 = Et, X = S
DEBDETPD: R_1 = Et, R_2 = Et, R_3 = Et, X = S
p-DTBDETPD: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4, R_3 = Et, X = S
DEBDPTPD: R_1 = Et, R_2 = Et, R_3 = Ph, X = S
p-DTBDPTPD: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4, R_3 = Ph, X = S



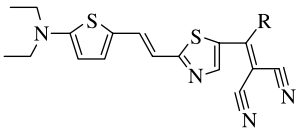
DEPADPPT: R_1 = Et, R_2 = Et, R_3 = Ph, X = O
p-DTPADPPT: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4, R_3 = Ph, X = O
DEPADPTPD: R_1 = Et, R_2 = Et, R_3 = Ph, X = S
p-DTPADPTPD: R_1 = 4-MeC_6H_4, R_2 = 4-MeC_6H_4, R_3 = Ph, X = S



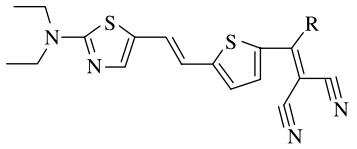
BNENHAa: R = H; BNENHAb: R = C_4H_9
BNENHAc: R = CF_3; BNENHAd: R = C_4F_9
BNENHAe: R = C_8F_{17}



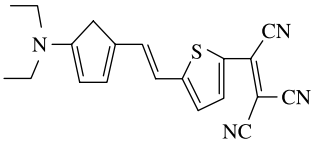
DR_1



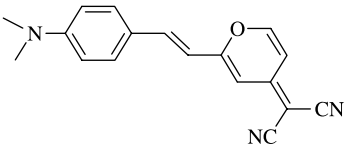
DETZVTPAa: R = H; DETZVTPAb: R = CN



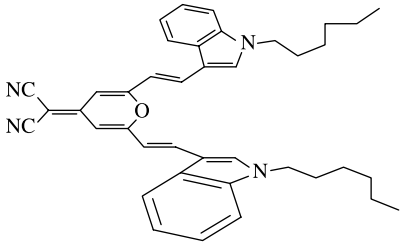
DETPVTZAa: R = H; DETPVTZAb: R = CN



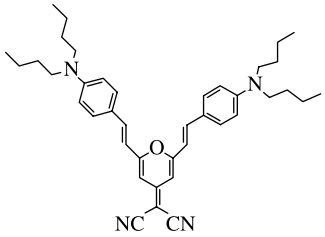
DECVTB



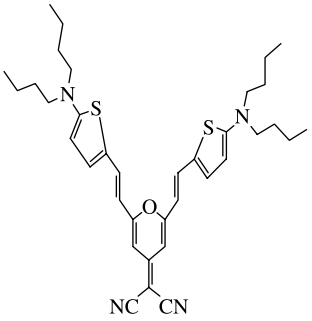
DCM



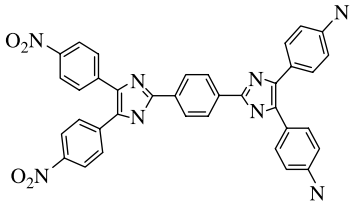
DADIH



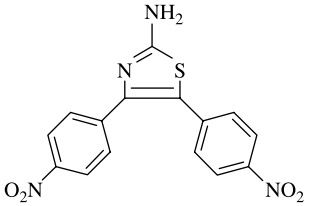
DADB



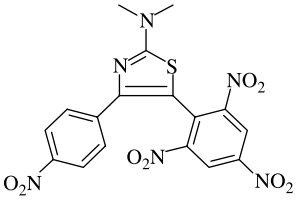
DADTB



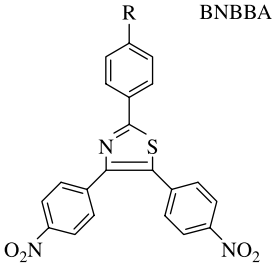
BNBBA



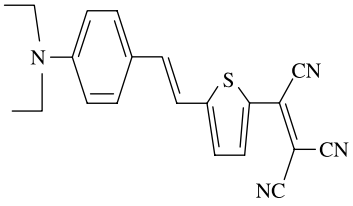
BNPTA



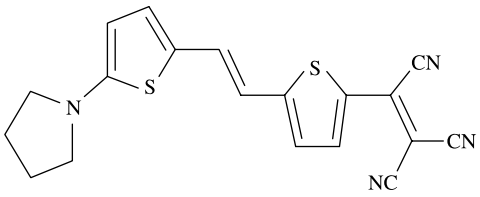
DMNPTPTA



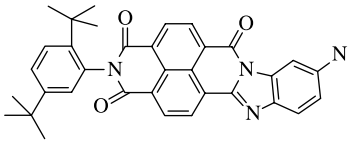
BNPIPA: R = NH_2; BNPIP : R = OH



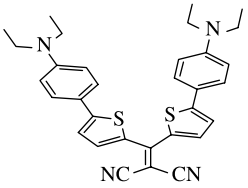
CDVTBE



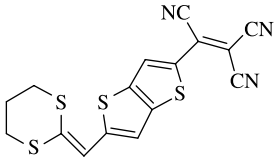
CPiVTBE



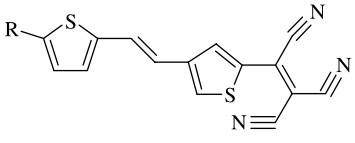
ABBIPT



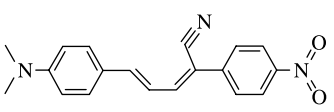
BDTMM



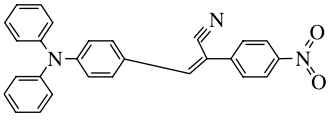
CDTTBE



CPyVTBE: R = piperidyl



DMPNPPD



DPPNPA

CDPVTBE: R = diphenylamino

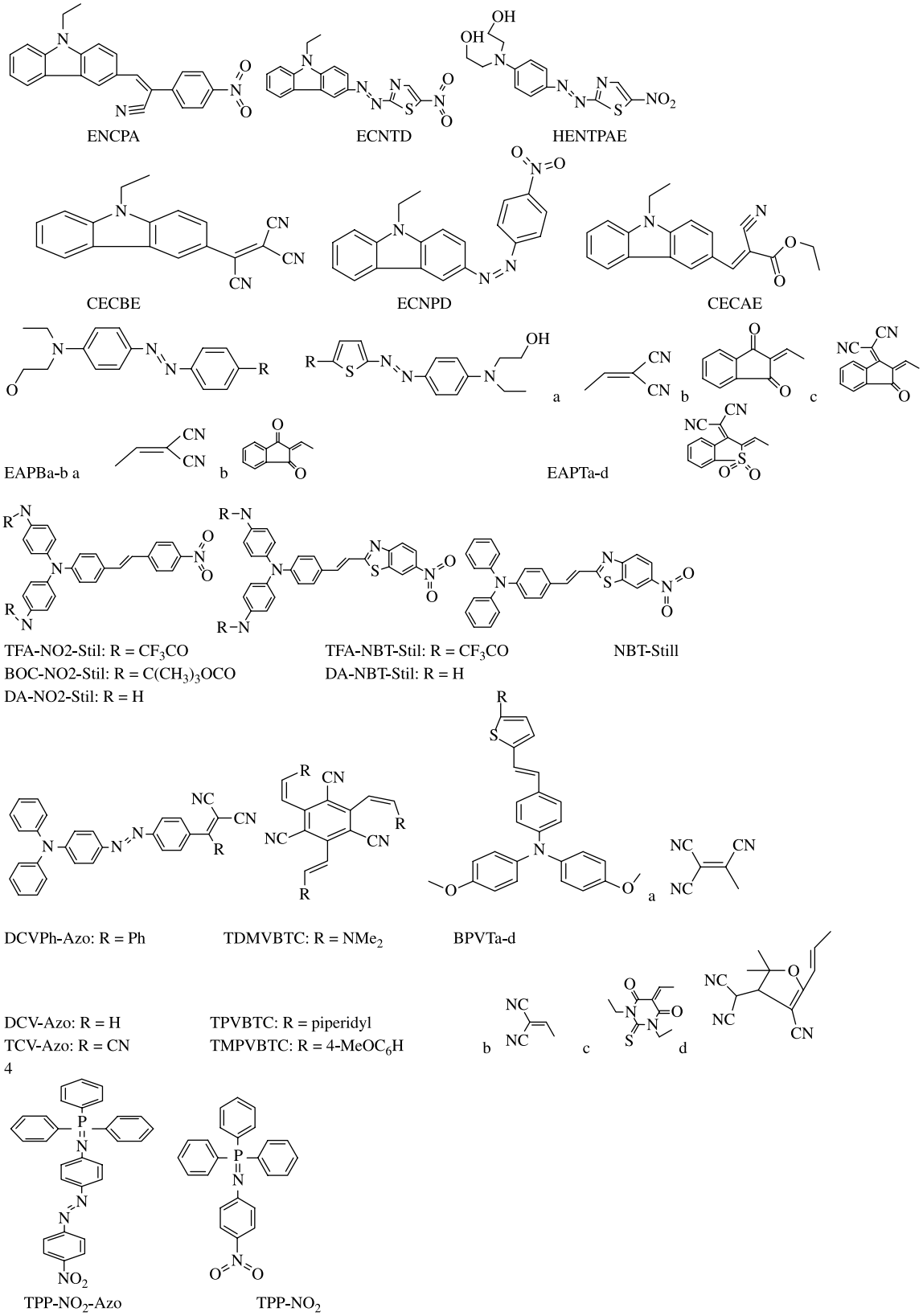


Figure 1. The chromophores' structures included in the data set.

Table 1. Compounds, descriptors value, experimental and calculated Y_d.

No.	Chromophore	Descriptors				Y_d (exp)	Y_d (cal)	
		NC	NF	RNN	RNTB		HM	SVM
1*	DBDTTC	25	1	0	0	281.80	219.762	212.971
2	DBDTTDC	28	3	0	0.031	309.64	260.620	281.900
3	DBDTTDET	33	3	0	0	333.96	317.142	323.188
4	DTDPD	32	3	0	0	289.52	310.054	315.333
5	DMNPA	16	2	0	0.029	149.96	140.898	152.093
6*	DPNPA	26	2	0	0.02	237.83	238.776	252.191
7	DANS	16	2	0	0	151.10	148.700	143.486
8	DPANS	26	2	0	0	247.69	244.493	245.568
9	DMNPAPA	14	4	0	0	156.81	166.759	158.927
10	DENPAPA	16	4	0	0	177.56	176.745	173.635
11*	DPNPAPA	24	4	0	0	262.75	250.716	247.911
12	ENTPAH	17	5	0	0	183.50	188.594	202.206
13	NTPDA	21	5	0	0	228.08	238.155	234.327
14	ENBTPAH	21	5	0	0	227.50	223.814	234.327
15	NBTPDA	25	5	0	0	284.06	271.794	270.288
16*	DEBID	20	1	0	0	170.75	170.911	161.630
17	p-DTBID	30	1	0	0	287.41	272.764	260.099
18	DEPAID	22	1	0	0	180.01	190.726	182.149
19	p-DTPAID	32	1	0	0	273.85	293.135	276.001
20	p-DTABM	24	3	0	0.042	228.23	224.900	233.942
21*	DEBDMPT	17	3	0	0	157.42	167.228	167.402
22	p-DTBDMPT	27	3	0	0	282.67	262.453	269.020
23	DMBDETPD	17	3	0	0	161.79	167.228	167.402
24	DEBDETPD	19	3	0	0	176.56	183.187	186.371
25	p-DTBDETPD	29	3	0	0	302.34	280.515	288.704
26*	DEBDPTPD	27	3	0	0	244.04	264.870	269.020
27	p-DTBDPTPD	37	3	0	0	349.65	361.716	348.047
28	DEPADPPT	29	3	0	0	250.86	283.575	288.704
29	p-DTPADPPT	39	3	0	0	337.43	381.454	355.895
30	DEPADPTPD	29	3	0	0	257.15	283.575	288.704
31*	p-DTPADPTPD	39	3	0	0	365.98	381.454	355.895
32	BNENHAa	17	4	0	0	179.58	185.767	181.760
33	BNENHAb	21	4	0	0	212.36	214.072	218.167
34	BNENHAc	18	4	3	0	202.00	216.358	204.147
35	BNENHAe	25	4	17	0	382.02	384.812	384.198
36*	BNENHAd	21	4	9	0	283.19	286.754	302.141
37	DR1	16	4	0	0	179.85	175.206	173.635
38	DETZVTPAa	17	4	0	0.05	178.94	171.663	172.379
39	DETZVTPAb	18	5	0	0.073	173.69	188.382	175.832
40	DETPVTZAa	17	4	0	0.05	188.28	171.663	172.379
41*	DETPVTZAb	18	5	0	0.073	184.25	188.382	175.832
42	DECVTB	20	4	0	0.068	179.51	191.048	190.166
43	DCM	18	3	0	0.053	183.58	167.759	181.410
44	DADIH	40	4	0	0.022	395.62	388.531	393.471
45	DADB	40	4	0	0.021	374.46	386.557	392.598
46*	DADTB	36	4	0	0.022	352.44	345.635	365.315
47	BNBBA	36	8	0	0	412.61	393.081	410.444
48	BNPTA	15	4	0	0	200.66	177.320	166.008
49	DMNPTPTA	17	6	0	0	236.24	210.253	234.074
50	BNPIPA	21	4	0	0	222.82	223.814	218.167
51*	BNPIP	21	3	0	0	193.60	210.473	206.538
52	CDVTBE	21	4	0	0.067	187.53	200.798	195.686
53	CPiVTBE	19	4	0	0.073	189.63	184.785	187.093
54	ABBIPT	34	4	0	0	343.57	339.291	338.325
55	BDTMM	32	4	0	0.027	336.62	309.218	326.505
56*	CDTTBE	16	3	0	0.088	216.65	142.340	208.812
57	CPyVTBE	20	4	0	0.068	178.14	192.446	190.166
58	CDPVTBE	27	4	0	0.058	247.86	261.408	249.987
59	DMPNPPD	19	3	0	0.024	183.68	183.278	185.856
60	DPPNPA	27	3	0	0.019	274.75	260.539	272.564

Table 1 – *Continued*

No.	Chromophore	Descriptors				Y_d (exp)	Y_d (cal)	
		NC	NF	RNN	RNTB		HM	SVM
61*	ECNPA	23	3	0	0.021	227.84	222.445	227.543
62	ECNTD	17	5	0	0	208.42	207.233	202.206
63	HENTPAE	13	5	0	0	175.14	164.990	177.322
64	CECBE	19	4	0	0.081	197.70	189.451	190.990
65	ECNPD	20	4	0	0	220.45	215.984	208.594
66*	CECAE	20	2	0	0.023	195.21	178.089	187.015
67	EAPBa	20	5	0	0.044	180.87	210.990	193.718
68	EAPBb	26	3	0	0	212.80	253.672	258.793
69	EAPTa	18	5	0	0.047	177.52	193.829	179.634
70	EAPTb	24	3	0	0	211.07	234.536	237.911
71*	EAPTc	27	5	0	0.034	239.85	274.452	265.867
72	EAPTd	23	3	0	0	225.43	223.288	227.392
73	TFA-NO2-Stil	30	4	6	0	364.48	349.390	362.366
74	BOC-NO2-Stil	36	4	0	0	310.20	354.911	351.040
75	DA-NO2-Stil	26	4	0	0	248.48	265.995	267.942
76*	TFA-NBT-Stil	31	5	6	0	426.55	368.317	378.040
77	DA-NBT-Stil	27	5	0	0	287.80	286.155	288.510
78	NBT-Stil	27	3	0	0	287.76	266.219	269.020
79	DCVPh-Azo	34	5	0	0.03	343.66	344.168	341.531
80	DCV-Azo	28	5	0	0.036	279.18	288.496	273.744
81*	TCV-Azo	29	6	0	0.054	287.03	303.878	259.392
82	TDMVBTC	21	6	0	0.059	221.02	220.378	193.435
83	TPVBTC	30	6	0	0.04	284.62	300.272	289.877
84	TMPVBTC	36	3	0	0.042	361.73	337.913	363.896
85	BPVTa	31	4	0	0.048	292.37	297.686	302.147
86*	BPVTb	30	3	0	0.032	274.73	282.446	304.186
87	BPVTc	35	3	0	0	393.70	338.886	337.030
88	BPVTd	38	4	0	0.038	389.30	366.250	380.632
89	TPP-NO2	24	2	0	0	239.09	224.423	224.669
90	TPP-NO2-Azo	30	4	0	0	305.60	303.565	306.105

Note: *Test set.

descriptors by decreasing correlation coefficient when used in one-parameter correlations.

Following the pre-selection of descriptors, MLR was developed in a stepwise procedure. Thus, descriptors and correlations are ranked according to the values of the *F*-test and the correlation coefficient. Starting with the top descriptor from the list, two-parameter correlations are calculated. In the following steps new descriptors are added one-by-one until the pre-selected number of descriptors in the model is achieved. The final result is a list of the 10 best models according to the values of the *F*-test and correlation coefficient. The goodness of the correlation is tested by the coefficient regression (R^2), the *F*-test (*F*) and the standard deviation (s^2).

3.2 Support vector machine

The support vector machines (SVM) are gaining popularity due to many attractive features and promising empirical performance [26]. It can solve high-dimension problems and therefore avoid the 'curse of dimensionality'. A detailed description of the theory of SVM can be referred in several excellent books and tutorials [27,28]. The basic idea and its performance were simply introduced here.

SVM classifiers are generated by a two-step procedure: first, the sample data vectors are mapped to a very high-dimensional space. The dimension of this space is significantly larger than that of the original data space. Then, the SVM algorithm finds a hyperplane in this space with the largest margin separating classes of data. The decision function is

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{l} y_i \alpha_i k(x, x_i) + b\right), \qquad (1)$$

where sign is simply a sign function which returns $+1$ for positive argument and -1 for a negative argument; y_i are input class labels that take a value of -1 or 1, x_i is a set of descriptors and $k(x, x_i)$ is a kernel function, whose value is equal to the inner product of two vectors $\mathbf{x}$ and $\mathbf{x}_i$ in the feature space $\phi(x)$ and $\phi(x_i)$ That is,

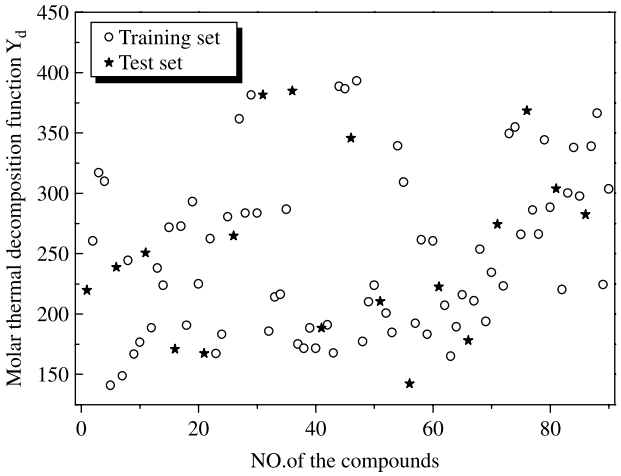


Figure 2. Scatter plot of the compounds in this research.

$k(x, x_i) = \phi(x)\cdot\phi(x_i)\cdot\alpha_i$ is Lagrangian multipliers. SVM can also be applied to regression by the introduction of an alternative loss function. The decision function of regression is as follow:

$$f(x) = \left(\sum_{i=1}^{l} y_i\alpha_i k(x, x_i) + b\right). \tag{2}$$

The constraints are the same as those of Equation (1). For regression tasks, the radial basis function kernel is often used because of its effectiveness and speed in training process. The form of the Gaussian function in R is: $\exp\{-\gamma(u-v)^2\}$. The generalisation performance of SVR depends on a good setting of parameters: C, ε and the kernel type and their corresponding kernel parameters. The overall performances of SVM were evaluated in terms of root mean square error (RMS), which was defined as

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}}, \tag{3}$$

where d_i are the desired outputs in the validation set, o_i are the actual outputs and n is the number of samples in the validation set. All calculation programs implementing SVM were written in R-file based on R script for SVM (http://www.r-project.org/). All scripts were compiled using R1.7.1 compiler running Windows XP operating system on a Pentium IV with 256M RAM.

4. Results and discussion

4.1 Results of the MLR

MLR was applied to the training set ($n = 72$) using the molar thermal decomposition function (Y_d) values as response variables and the calculated descriptors as descriptive variables. The MLR model was built using the training set and validation using an external prediction set. The obtained linear model consists of four descriptors, number of C atoms (NC), number of F atoms (NF), relative number of N atoms (RNN) and relative number of triple bonds (RNTB). The model is shown in Table 2. The correlation matrix of the four selected descriptors is shown in Table 3. From Table 3, it can be seen that the linear correlation coefficient value of each of the two descriptors is <0.80, which means the descriptors are independent in the analysis [23].

With the model built, the prediction results of test set ($n = 18$) were also obtained. The linear model produced the RMS error of 19.28, 29.78 and 21.65 for the training set, the test set and the whole data set and the corresponding square of correlations coefficients (R^2) were $R^2 = 0.928$, 0.828 and 0.910, respectively. The value of the four descriptors, calculated and experimental values of Y_d are given in Table 1, the scatter plot is shown in Figure 3.

4.2 Result of SVM

To see whether we can get a more accurate prediction model, SVM is used to develop a non-linear model based on the same subset of descriptors using the same training set as MLR. To obtain better results, the parameters that influence the performance of SVM were optimised. The selection of the value for SVM was performed by systemically changing its value in the training step. The value which gives the best leave-one-out (LOO) cross-validation result was used in the model. The parameters here include capacity parameter C, ε of ε-insensitive loss function and γ which controls the amplitude of the Gaussian function. Since three parameters exhibits strong

Table 2. Descriptors, coefficients, standard error and t-test values for the linear model.

No.	Descriptor	Coefficient	Standard error	t-test
0	Intercept	−55.333	16.862	−3.281
1	Number of C atoms	10.561	0.423	24.941
2	Number of F atoms	8.076	1.079	7.482
3	Relative number of N atoms	631.03	107.77	5.855
4	Relative number of triple bonds	−345.23	98.519	−3.504

$N = 72$, $R^2 = 0.928$, $F = 226.97$, $s^2 = 372.50$.

Table 3. Correlation matrix of the four descriptors used in this work.

	RNTB	RNN	NC	NF
RNTB	1.000			
RNN	0.171	1.000		
NC	0.149	0.013	1.000	
NF	−0.247	−0.117	−0.005	1.000

interactions, grid search (GS), which has been used either formally or informally for SVM parameter selection, was performed in this study. In the grid search, we considered the parameter γ from 0.0001 to 0.01 with 0.0002 as the increment. The parameter C was chosen from values between 100 and 1000 with 100 as the increment and 1000 as the increment within 1000 to 10,000. The parameter ε was searched with 0.01 increments within 0.01 to 0.1. LOO cross-validation was performed for parameters selection; the overall performances of SVM were evaluated in terms of RMS. The γ, ε and C for this data set were finally fixed to 0.02, 0.03 and 500, respectively.

Using the optimised parameters selected above, the predicted results with SVM were obtained. They are shown in Table 1 and Figure 4. The RMS errors in prediction for the training, test and overall data sets are 16.16, 25.40 and 18.38 and the square of correlation coefficient are 0.949, 0.867 and 0.933, respectively.

By comparison of the correlation models obtained by MLR and SVM, it can be seen that the performance of SVM is a little better than that of MLR. The proposed linear model indicated that constitutional descriptors could represent the structure character of NLO chromophore. Additionally, nonlinear SVM model based on the same sets of descriptors showed better predictive ability. To further test the suitability of the QSPR approach constructed in our study, the obtained decomposition function Y_d was compared with those calculated in [19].

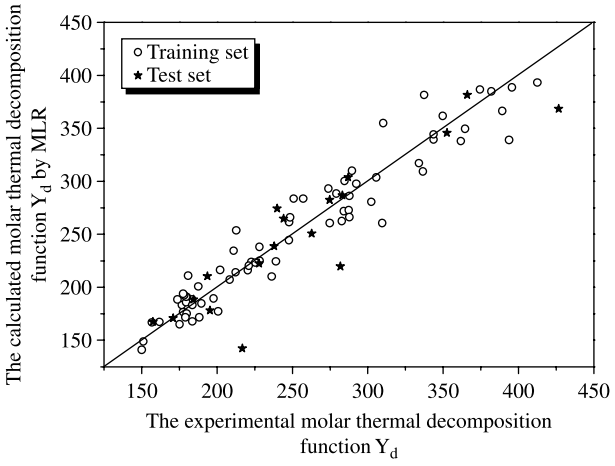


Figure 3. Predicted vs. experimental Y_d by MLR.

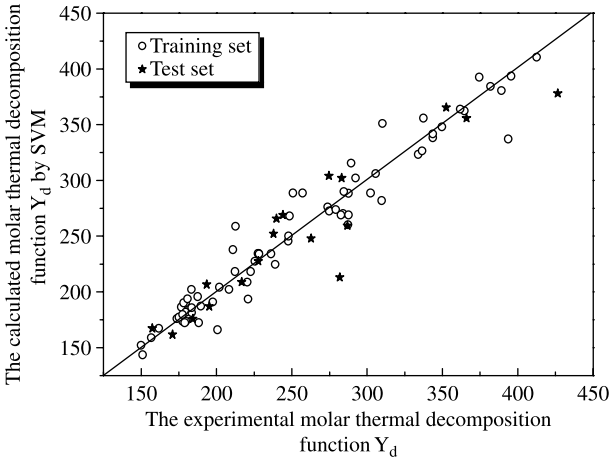


Figure 4. Predicted vs. experimental Y_d by SVM.

The statistic parameters were: $R^2 = 0.964$; SEE = 14.01 for the model in [19]. It can be seen that the performance of the model described in [19] ($R^2 = 0.964$; SEE = 14.01) is a little better than the models built by us. There are four descriptors in our model, while there are seven descriptors in the model built in [19]. The descriptors selected are easily understood and obtained. In addition, from the viewpoint of model building and the statistic parameter evaluating of the model, this approach is a suitable and alternative QSPR study.

4.3 Discussion of the input parameters

Thermal stability of a material is the stability against degradation upon exposure to elevated temperatures in an inert environment. Organic second-order NLO materials are often exposed to high temperatures during processing or use. Thus thermal stability is among the most important properties of them for a wide range of applications. Thermal stability can be described with the molar thermal decomposition function Y_d, that correlates T_d, with an equation: $Y_d = T_d \cdot M$ (M is the molecular weight). By interpreting the descriptors in the QSPR model, it is possible to gain some insight into factors that are likely to govern the thermal stability. This model contains four descriptors and these descriptors encoded different aspects of the molecular structure.

NC partially accounts for the size and shape of compounds. The larger the descriptors value is, the larger is the molecular weight. Thus, an increase in the descriptor value leads to an increase in Y_d. NF and RNN represent the properties of N and F atoms, as we know, they both have a high electronegativity, that helps in increasing the bonding energy of the molecule. Since their positive coefficients are in the linear model, increasing this descriptor also increases the molar thermal decomposition function values. Additionally, the larger of the two descriptors can also lead to

the increase in molecular weight, which has a positive influence on the thermal decomposition function values. The last one is RNTB, which is calculated as the number of triple bonds divided by the number of bonds. The RNTB of course is one important factor influencing the bonding energy of the molecule. However, opposite to NF and RNN, the negative coefficient in the model implies that increasing the value of this descriptor can lead to a decrease in the independent value.

5. Conclusion

Simple, yet acceptable, QSPR models have been developed to predict the thermal stabilities of 90 structurally diverse second-order NLO chromophore molecules using the MLR and SVM, respectively. A comparison of the results obtained by the models proved that nonlinear SVM model gave better results. Furthermore, the present work clearly demonstrates that QSPR equation based on easily understood and obtained physicochemical molecular descriptors such as constitutional descriptors can also be developed for the prediction of Y_d values for the diverse set of chromophores. Therefore, this QSPR model should be useful and more effective in the development of new NLO chromophores.

References

[1] J. Zyss, *Molecular Nonlinear Optics: Materials, Physics and Devices*, Academic Press, Boston, 1994.

[2] P.N. Prasad and D.J. Williams, *Introduction to Nonlinear Optical Effects in Molecules and Polymers*, John Wiley, New York, NY, 1991.

[3] D.S. Chemla and J. Zyss, *Nonlinear Optical Properties of Organic Molecules and Crystals*, Academic Press, New York, NY, 1987.

[4] S. Shi, *Molecular-structures macroscopic nonlinear-optical effects, and photonic devices*, Contemp. Phys. 35 (1994), pp. 21–36.

[5] B. Beck and U.W. Grummt, *Semiempirical calculations of first-order hyperpolarizabilities: Testing the performance of different methods in comparison to experiment*, J. Phys. Chem. B 102 (1998), pp. 664–670.

[6] A. Willetts, J.E. Rice, D.M. Burland, and D.P. Shelton, *Problems in the comparison of theoretical and experimental hyperpolarizabilities*, J. Chem. Phys. 97 (1992), pp. 7590–7599.

[7] D.R. Kanis, M.A. Ratner, and T.J. Marks, *Design and construction of molecular assemblies with large 2nd-order optical nonlinearities – quantum-chemical aspects*, Chem. Rev. 94 (1994), pp. 195–242.

[8] S.R. Marder, D.N. Beratan, and L.T. Cheng, *Approaches for optimizing the 1st electronic hyperpolarizability of conjugated organic-molecules*, Science 252 (1991), pp. 103–106.

[9] D.M. Burland, R.D. Miller, O. Reiser, R.J. Twieg, and C.A. Walsh, *The design, synthesis, and evaluation of chromophores for second-harmonic generation in a polymer waveguide*, J. Appl. Phys. 71 (1992), pp. 410–417.

[10] D.M. Burland, R.D. Miller, and C.A. Walsh, *Second-order nonlinearity in poled-polymer systems*, Chem. Rev. 94 (1994), pp. 31–75.

[11] L.R. Dalton, A.W. Harper, R. Ghosn, W.H. Steier, M. Ziari, H. Fetterman, Y. Shi, R.V. Mustacich, A.Y. Jen, and K.J. Shea, *Synthesis and processing of improved organic second-order nonlinear optical materials for applications in photonics*, Chem. Mater. 7 (1995), pp. 1060–1081.

[12] Q. Pan, C. Fang, F. Li, Z. Zhang, Z. Qin, X. Wu, and Q.Y. Gu, *Thermally stable chromophores for nonlinear optical applications*, Mater. Res. Bull. 37 (2002), pp. 523–531.

[13] K.V. Katti, K. Raghuraman, N. Pillarsetty, S.R. Karra, R.J. Gulotty, M.A. Chartier, and C.A. Langho, *First examples of azaphosphanes as efficient electron donors in the chemical architecture of thermally stable new nonlinear optical materials*, Chem. Mater. 14 (2002), pp. 2436–2438.

[14] K. Oberg, A. Berglund, U. Edlund, and B. Eliasson, *Prediction of nonlinear optical responses of organic compounds*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 811–814.

[15] X.D. Zeng, X. Xu, B.F. Wang, and B.C. Wang, *Calculations of hyperpolarizabilities for* para*-disubstituted benzenes with the QSPR*, Chin. Chem. Lett. 6 (2004), pp. 753–756.

[16] J. Bicerano, *Prediction of Polymer Properties*, Marcel Dekker, New York, NY, 1996.

[17] G.Z. Li, J. Yang, H.F. Song, S.S. Yang, W.C. Lu, and N.Y. Chen, *Semiempirical quantum chemical method and artificial neural networks applied for λ_{max} computation of some azo dyes*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 2047–2050.

[18] J. Xu, Z. Zheng, B. Chen, and Q.J. Zhang, *A linear QSPR model for prediction of maximum absorption wavelength of second-order NLO chromophores*, QSAR Comb. Sci. 25 (2006), pp. 72–379.

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

[20] HyperChem 6.01, Hypercube, Inc., 2000.

[21] MOPAC, v.6.0 Quantum Chemistry Program Exchange, Program 455, Indiana University, Bloomington, IN, 1999.

[22] A.R. Katritzky, V.S. Lobanov, and M. Karelson, *CODESSA: Training Manual*, University of Florida, Gainesville, FL, 1995.

[23] A.R. Katritzky, V.S. Lobanov, and M. Karelson, *CODESSA: Reference Manual*, University of Florida, Gainesville, FL, 1994.

[24] M. Oblak, M. Randic, and T. Solmajer, *Quantitative structure–activity relationship of flavonoid analogues. 3. Inhibition of $p56^{lck}$ protein tyrosine kinase*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 994–1001.

[25] A.R. Katritzky, R. Petrukhin, R. Jain, and M. Karelson, *QSPR analysis of flash points*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 1521–1530.

[26] V.N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, NY, 1998.

[27] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[28] B. Schōlkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.