http://dx.doi.org/10.5012/bkcs.2011.32.11.3865

QSPR Study of the Absorption Maxima of Azobenzene Dyes

Jie Xu,[†,‡,*] Lei Wang,[‡] Li Liu,[‡] Zikui Bai,[‡] and Luoxin Wang[‡]

[†]*Key Lab of Green Processing & Functional Textiles of New Textile Materials, Ministry of Education, Wuhan Textile University, 430073, Wuhan, China.* [*]*E-mail: xujie0@ustc.edu*
[‡]*College of Materials Science & Engineering, Wuhan Textile University, 430073, Wuhan, China*
Received May 3, 2011, Accepted August 31, 2011

A quantitative structure-property relationship (QSPR) study was performed for the prediction of the absorption maxima of azobenzene dyes. The entire set of 191 azobenzenes was divided into a training set of 150 azobenzenes and a test set of 41 azobenzenes according to Kennard and Stones algorithm. A seven-descriptor model, with squared correlation coefficient (R^2) of 0.8755 and standard error of estimation (s) of 14.476, was developed by applying stepwise multiple linear regression (MLR) analysis on the training set. The reliability of the proposed model was further illustrated using various evaluation techniques: leave-many-out cross-validation procedure, randomization tests, and validation through the test set.

Key Words : Azobenzene dyes, Absorption maxima, QSPR, Molecular descriptors, Multiple linear regression

Introduction

Azobenzenes are of great importance in many branches of chemistry.[1] The π-π* excitation in azobenzenes determines the absorption spectrum, and for many azobenzenes it appears in the visible region. Azobenzenes have been widely used as synthetic dyes with colors ranging from red to blue.[2,3] In addition to dyeing feature, azobenzenes possess some interesting properties such as the reversible *cis-trans* photoisomerization about the azo bond when irradiated,[4-6] and nonlinear optical (NLO) effect[7,8] related to the donor-acceptor azobenzenes. Thus, azobenzenes have attracted much attention as materials in the development of nonlinear optical and storage data devices.[9-14]

The absorption maxima (λ_{max}), one of the most important spectroscopic properties, is related to the color property and determined mainly by the structure of the dyes. Application of experimental methods to obtain the λ_{max} of dyes is the most obvious and effective method; however there are some drawbacks such as the need for laboratory facilities and the huge workload. Also, the methods cannot be easily applied for toxic, volatile, explosive or radioactive substances; and they cannot be used if the material has not been synthesized yet. For instance, several classes of dyes are considered as possible carcinogens or mutagens; the high coloring power of dyes gives rise to esthetic damage: dye concentrations lower than 1 mg/L may induce visible coloration and hence public complaint.[15] Therefore, it is necessary to develop theoretical methods to compensate the shortage of experimental methods. Methods for quantitatively predicting the λ_{max} of dyes from their molecular structures alone would be of significant utility not only in the use of dyes, but also in the molecular design of new dye exploration.

So far, the computational efforts to predict the λ_{max} were based mainly on quantum-chemistry calculations, such as density functional theory (DFT) and *ab initio* methods. However, these calculations of the absorption profiles are relatively time-consuming and complex, thus precluding the use of such methods to predict dozens of dyes in a fast and accurate manner. In addition, it has been found that the λ_{max} values of some dyes calculated by DFT gave rise to poor results.[16-20]

Alternatively, the quantitative structure-property relationship (QSPR) provides a promising method for the prediction of λ_{max} using descriptors derived solely from the molecular structure to fit experimental data. The QSPR method is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physico-chemical properties, can be correlated with numerical changes in structural features of all compounds, termed "molecular descriptors".[21-26] The advantage of this method lies in the fact that it requires only the 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 synthesized 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 successfully used to investigate the relationship between the λ_{max} and the structure of various compounds. For example, Buttingsrud *et al.*[1] developed empirical models relating bond lengths and critical points in the electron density distribution to the λ_{max} of azobenzene dyes. Liu *et al.*[27] studied the λ_{max} of flavones using heuristic method (HM) and radial basis function neural network (RBFNN). Recently, Fayet *et al.*[28] proposed satisfactory QSPR models for the prediction of the λ_{max} of a small set of 22 azobenzenes and 24 anthraquinones using quantum chemical descriptors. In our previous work, QSPR models were built to predict the λ_{max} of second-order NLO chromophores[29] and organic dyes for dye-sensitized solar cells,[30,31] respectively.

The aim of this work is to develop a robust QSPR model that could predict the λ_{max} values for a diverse set of azobenzene dyes using the general molecular descriptors and to seek the important structural features related to the λ_{max} values.

Materials and Methods

Dataset. The experimental λ_{max} values for 191 azobenzene dyes were taken from the article published by Buttingsrud *et al*.[1] The general structure of the azobenzenes is sketched in Figure 1. The detailed structures of all the studied compounds are depicted in Figure SI of Supporting Information. The experimental values span between 318 and 562 nm (Table 1).

Descriptor Generation. The chemical structure of each compound was sketched on a PC using the HYPERCHEM program[32] and preoptimized using MM+ molecular mechanics method (Polak-Ribiere algorithm). 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. Then the geometries were used as input for the generation of 1664 descriptors using the Dragon software (Version 5.4).[33] 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.

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[34]) were excluded in a pre-reduction step. Thus 840 descriptors were remained to undergo subsequent descriptor selection.

Kennard and Stones Algorithm. Kennard and Stones algorithm[35] has been widely used for splitting datasets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other. These two samples are removed from the original dataset and put into the calibration set. This procedure is repeated until the desired number of samples has been selected in the calibration set. The advantages of this algorithm are that the calibration samples always map the measured region of the input variable space completely with respect to the induced metric and that the no validation samples fall outside the measured region. Kennard and Stones algorithm has been considered as one of the best ways to build training and test sets.[36,37] Using Kennard and

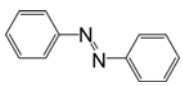

Figure 1. General structure of the studied azobenzene dyes.

Table 1. Experimental and calculated λ_{max} for the studied azobenzenes

No.	Expt. λ_{max}	Calc. λ_{max}	Residual	No.	Expt. λ_{max}	Calc. λ_{max}	Residual
1	318	319.9	-1.9	97	380	357.8	22.2
2	415	432.4	-17.4	98	388	384.1	3.9
3[a]	411	421.7	-10.7	99	386	378.0	8.0
4	407	406.5	0.5	100	387	368.2	18.8
5	407	456.5	-49.5	101	390	376.1	13.9
6	397	398.7	-1.7	102	375	372.9	2.1
7	397	398.9	-1.9	103	362	370.1	-8.1
8	399	403.8	-4.8	104	381	386.5	-5.5
9	382	393.1	-11.1	105	370	387.6	-17.6
10	405	394.8	10.2	106	409	406.7	2.3
11	402	375.9	26.1	107	460	413.5	46.5
12	390	380.0	10.0	108*	452	442.0	10.0
13	389	358.1	30.9	109	424	407.7	16.3
14	348	380.1	-32.1	110	420	428.1	-8.1
15	345	366.7	-21.7	111	457	462.8	-5.8
16	348	348.2	-0.2	112	415	408.5	6.5
17	333	348.7	-15.7	113[a]	414	417.2	-3.2
18	457	438.8	18.2	114	424	408.0	16.0
19	431	423.2	7.8	115	416	418.7	-2.7
20	435	413.3	21.7	116[a]	418	413.9	4.1
21	407	418.7	-11.7	117[a]	417	414.1	2.9
22	407	418.8	-11.8	118	424	421.6	2.4
23	407	409.3	-2.3	119	425	427.8	-2.8
24	419	418.3	0.7	120	434	441.7	-7.7
25	427	436.9	-9.9	121	434	429.0	5.0
26	451	451.8	-0.8	122	447	428.2	18.8
27	447	435.4	11.6	123	454	447.8	6.2
28	454	443.0	11.0	124[a]	450	449.4	0.6
29	462	444.3	17.7	125[a]	403	397.5	5.5
30	466	466.7	-0.7	126	407	409.6	-2.6
31	407	412.2	-5.2	127	409	398.2	10.8
32	405	409.0	-4.0	128	402	407.0	-5.0
33	405	402.6	2.4	129	402	407.5	-5.5
34	435	423.5	11.5	130	405	405.7	-0.7
35	409	419.8	-10.8	131	415	403.3	11.7
36	416	432.5	-16.5	132	417	405.9	11.1
37	433	453.8	-20.8	133	413	407.5	5.5
38	357	374.0	-17.0	134[a]	410	407.7	2.3
39	353	355.3	-2.3	135	412	418.7	-6.7
40	368	391.7	-23.7	136[a]	414	402.7	11.3
41	385	365.4	19.6	137[a]	417	423.7	-6.7
42	356	358.8	-2.8	138[a]	421	425.5	-4.5
43	354	385.4	-31.4	139	418	417.8	0.2
44	417	415.4	1.6	140[a]	419	415.8	3.2
45	326	347.3	-21.3	141	437	443.1	-6.1
46	322	339.8	-17.8	142	438	446.4	-8.4
47	412	409.8	2.2	143[a]	396	387.0	9.0
48	413	401.4	11.6	144	403	419.6	-16.6
49	407	411.8	-4.8	145[a]	395	405.4	-10.4
50	407	411.6	-4.6	146	394	389.0	5.0
51	418	413.0	5.0	147[a]	394	387.4	6.6
52[a]	418	414.7	3.3	148	395	400.6	-5.6
53	420	422.1	-2.1	149[a]	400	394.5	5.5

Table 1. Continued

No.	Expt. λ_{max}	Calc. λ_{max}	Residual	No.	Expt. λ_{max}	Calc. λ_{max}	Residual
54	424	420.5	3.5	150	400	388.6	11.4
55	421	428.4	-7.4	151	398	389.0	9.0
56	425	424.4	0.6	152[a]	406	393.4	12.6
57	423	424.6	-1.6	153	404	397.8	6.2
58[a]	429	419.9	9.1	154[a]	404	399.2	4.8
59	435	451.1	-16.1	155[a]	408	397.4	10.6
60	449	446.2	2.9	156	408	405.3	2.7
61[a]	420	421.6	-1.6	157[a]	411	413.7	-2.7
62	442	424.2	17.8	158[a]	428	434.1	-6.1
63	450	447.7	2.3	159	429	431.7	-2.7
64[a]	426	432.0	-6.0	160	418	404.3	13.7
65[a]	446	466.8	-20.8	161[a]	434	415.8	18.2
66	462	460.1	1.9	162[a]	410	412.9	-2.9
67	500	503.9	-3.9	163[a]	406	427.6	-21.6
68	514	494.1	19.9	164	410	429.1	-19.1
69	490	501.3	-11.3	165[a]	420	424.1	-4.1
70	495	492.3	2.7	166	420	425.9	-5.9
71	478	487.0	-9.0	167	428	437.3	-9.3
72	503	492.1	10.9	168	448	441.1	6.9
73	562	526.0	36.0	169[a]	448	446.3	1.7
74	405	396.2	8.8	170	450	455.9	-5.9
75	395	406.8	-11.8	171[a]	450	436.1	13.9
76[a]	395	403.6	-8.6	172[a]	425	437.6	-12.6
77	406	398.8	7.2	173	414	434.6	-20.6
78[a]	412	413.3	-1.3	174	417	439.0	-22.0
79	416	413.4	2.6	175	422	451.2	-29.2
80	434	438.0	-4.0	176	426	431.3	-5.3
81	396	422.1	-26.1	177	444	448.3	-4.3
82	349	356.9	-7.9	178[a]	448	444.6	3.4
83	351	356.9	-5.9	179[a]	461	453.7	7.3
84	349	353.5	-4.5	180	482	460.4	21.6
85	350	360.3	-10.3	181	487	472.6	14.4
86	357	342.1	14.9	182[a]	442	442.7	-0.7
87	360	358.3	1.7	183[a]	449	438.5	10.5
88	366	366.3	-0.3	184	452	439.6	12.4
89	370	361.4	8.6	185	446	454.6	-8.6
90	368	374.0	-6.0	186[a]	450	435.7	14.3
91	370	361.8	8.2	187	461	453.6	7.4
92	378	406.9	-28.9	188	464	450.2	13.8
93	372	398.0	-26.0	189	471	456.4	14.6
94	354	361.5	-7.5	190[a]	360	361.9	-1.9
95	354	367.5	-13.5	191	411	378.9	32.1
96	372	355.5	16.5				

[a]Members for the test set

Stones algorithm, the entire set was divided into two subsets: a training set of 150 compounds, and a test set including the remaining 41 compounds.

Model Development and Validation. Stepwise multiple linear regression (MLR) analysis with Leave-Many-Out (LMO) cross-validation was used to select descriptors for the QSPR models on the training set. Five samples of the original training set were removed, and the model was recalculated using the remaining n-5 samples as training set. The response was then predicted for the excluded samples. This process was repeated for all samples of the training set, obtaining a prediction for every one and thus the cross-validated R^2 (R_{CV}^2). *F*-to-enter and *F*-to-remove were 4 and 3, respectively. The models were justified by the R^2, the adjusted R^2, the cross-validated R^2, the *F* ratio values, the standard error *s* and the significance level value *p*. The adjusted R^2 is calculated using the following formula:

$$R_{adj}^2 = 1-\left[\left(\frac{n-1}{n-m-1}\right)(1-R^2)\right] \quad (1)$$

where *n* is the number of samples 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. *F* ratio is defined as the ratio between the model sum of squares and the residual sum of squares, which is a comparison between the model-explained variance and the residual variance: high values of the *F* ratio indicate reliable models. A variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as

$$\text{VIF} = \frac{1}{1-R_j^2} \quad (2)$$

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 five.[38]

Randomization tests were also carried out to prove the possible existence of chance correlation. To do this, the dependent variable was randomly scrambled and used in the experiment. Models were then investigated with all members in the descriptor pool to find the most predictive models. The resulting models obtained on the training set with the randomized λ_{max} 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.

Validation of the models was further performed by using the test set. The external $R_{CV,ext}^2$ for the test set is determined with Eq. (3):

$$R_{CV,ext}^2 = 1-\frac{\Sigma(y_i-\tilde{y}_i)^2}{\Sigma(y_i-\bar{y}_{tra})^2} \quad (3)$$

where y_i and $\tilde{y}_i$ are the observed and the calculated response values, respectively; and $\bar{y}_{tra}$ is the averaged value for the response variable of the training set; and the summation runs

over all samples in the test set. According to Golbraikh and Tropsha,[39] a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{CV,ext} > 0.5 \quad (4a)$$

$$r^2 > 0.6 \quad (4b)$$

$$(r^2 - r_0^2)/r^2 < \text{or } (r^2 - r'^2_0)/r^2 < 0.1 \quad (4c)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \quad (4d)$$

Here:
$$r = \frac{\Sigma(y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\Sigma(y_i - \tilde{y})^2 \Sigma(\tilde{y}_i - \bar{\tilde{y}}_i)^2}} \quad (5a)$$

$$r_0^2 = 1 - \frac{\Sigma(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\Sigma(\tilde{y}_i - \bar{\tilde{y}})^2} \quad (5b)$$

$$r'^2_0 = 1 - \frac{\Sigma(y_i - y_i^{r_0})^2}{\Sigma(y_i - \bar{y})^2} \quad (5c)$$

$$k = \frac{\Sigma y_i \tilde{y}_i}{\Sigma y_i^2} \quad (5d)$$

$$k' = \frac{\Sigma y_i \tilde{y}_i}{\Sigma \tilde{y}_i^2} \quad (5e)$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r_0^2 (calculated versus observed values) and r'^2_0 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively; $y_i^{r_0}$ and $\tilde{y}_i^{r_0}$ are defined as $y_i^{r_0} = k\tilde{y}_i$ and $\tilde{y}_i^{r_0} = k'y_i$, respectively; and the summations are over all samples in the test set.

Applicability Domain Analysis. The applicability domain of a QSPR model[36,40] must be defined if the model is to be used for screening new compounds. The applicability domain (AD) is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. This region is defined by the nature of the compounds in the training set, and can be characterized in various ways. In this work, the structural AD was verified by the leverage approach. The leverage h_i[41] is defined as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (6)$$

where x_i is the descriptor row-vector the i-th compound, x_i^T is the transpose of x_i, X is the descriptor matrix, X^T is the transpose of X. The warning leverage h^* is, generally, fixed at 3(m+1)/n, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation. In fact, leverage can be used as a quantitative measure of the model AD suitable for evaluating the degree of extrapolation. It represents a sort of compound distance from the model experimental space.

The Williams plot, the plot of leverage values versus standardized residuals, was used to give a graphical detection of both the response outliers (Y outliers) and the structurally influential compounds (X outliers). In this plot, the two horizontal lines indicate the limit of normal values for Y outliers (i.e. samples with standardized residuals greater than 2.5 standard deviation units, ± 2.5 *s*); the vertical straight lines indicate the limits of normal values for X outliers (i.e. samples with leverage values greater than the threshold value, $h > h^*$). For a sample in the external test set whose leverage value is greater than h^*, its prediction is considered unreliable, because the prediction is the result of a substantial extrapolation of the model. Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and experimental values is as high as that for the compounds in the training set. It is noteworthy that the response outliers can be highlighted only for compounds with known responses and the possibility of a compound to be out of the structural AD of a model can be verified for every new compound, the only knowledge needed being the molecular structure information represented by the molecular descriptors selected in the model.

Results and Discussion

Results of the MLR Model. The experimental λ_{max} values in Table 1 were divided into the training and test sets on the basis of Kennard and Stones algorithm. Stepwise MLR analysis with LMO cross-validation was applied on the training set to select the descriptors for the best model and the number of descriptors in the final QSPR model was determined on the basis of the dataset size and on the basis of the correlation coefficient R, the adjusted R, the significance test F and the standard error s. The R^2 results during the stepwise MLR analysis are shown in Figure 2. Obviously, λ_{max} is not linearly correlated with any of the molecular descriptors since univariant correlations between λ_{max} and the different descriptors have poor R^2 values. The R^2 increases gradually with the increased number of descriptors. When adding another descriptor did not significantly

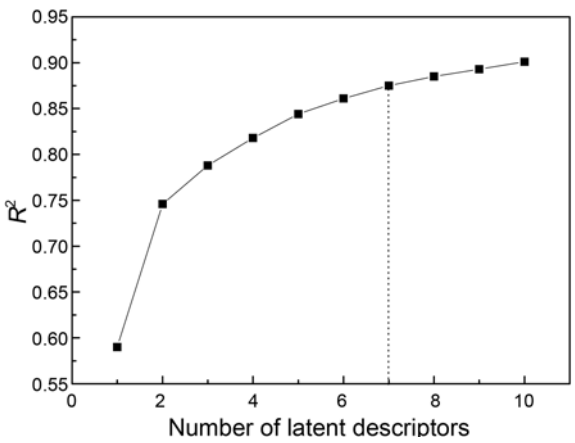

Figure 2. R^2 *vs.* number of latent descriptors in the best MLR equation.

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 best seven-parameter equation with R^2 = 0.8755, R^2_{CV} = 0.8584 and R^2_{adj} = 0.8694 was obtained, which is as the following:

$$\lambda_{max} = -72.191 - 734.508[PW5] + 64.847[piPC07] + 22.399[DISPv] + 2.621[GATS2v] + 2.833[Mor02v] + 47.546[R4u] + 17.051[nArCN] \quad (7)$$

n = 150, R^2 = 0.8755, R_{adj}^2 = 0.8694, R_{CV}^2 = 0.8584, s = 14.476, F = 142.7, $p < 0.00001$

Here, PW5 is path/walk 5 – Randic shape index; piPC07 is molecular multiple path count of order 07; GATS2v is Geary autocorrelation – lag 2/weighted by atomic van der Waals volumes; DISPv is d COMMA2 value/weighted by atomic van der Waals volumes; Mor02v is 3D-MoRSE – signal 02/ weighted by atomic van der Waals volumes; R4u is R auto-correlation of lag 4/unweighted; nArCN is number of nitriles (aromatic). More information about these descriptors can be found in Dragon software user's guide[33] and the references therein.

In general, 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 142.7 indicates that Eq. (7) does an excellent job of predicting the λ_{max} values. Eq. (7) has an adjusted R^2 value of 0.8694, which indicates very good agreement between the correlation and the variation in the data. The cross-validated correlation coefficient R_{CV}^2 = 0.8584 illustrates the reliability of the model by focusing on the sensitivity of the model to the elimination of any five data point. The model was further validated by applying the randomization tests and the obtained R^2 *vs*. the correlation coefficient between the original and permuted response data are plotted in Figure 3. The lower R^2 values indicate that the good results of the original model are not due to chance correlation or structural dependency of the training set. Some important statistical parameters (as given in Table 2) were used to valuate the involved descriptors. The *t*-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute *t*-values shown in Table 2 express that the regression coefficients of the

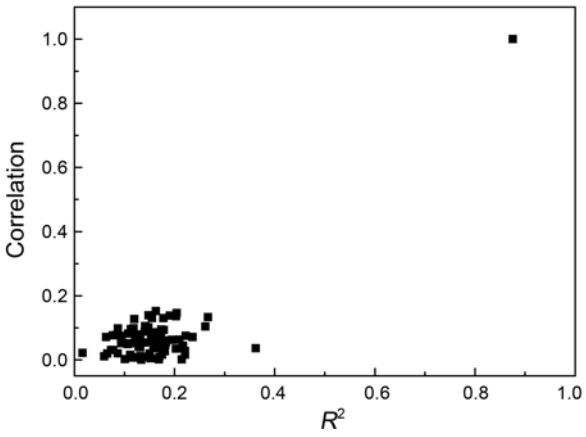

Figure 3. R^2 *vs.* the correlation coefficient between the original and permuted response data.

Table 2. Characteristics of the selected descriptors in the best MLR model

Descriptor	Descriptor type	X	DX	*t*-value	*t*-probability	VIF
Constant		-72.191	27.018	-2.635	0.010	
PW5	Topological descriptors	-734.508	223.024	-3.293	0.001	2.229
piPC07	Walk and path counts	64.847	9.743	6.656	0.000	4.895
GATS2v	2D autocorrelations	22.399	5.419	4.133	0.000	1.209
DISPv	Geometrical descriptors	2.621	0.350	7.487	0.000	1.688
Mor02v	3D-MoRSE descriptors	2.833	0.658	4.308	0.000	2.595
R4u	GETAWAY descriptors	47.546	8.540	5.567	0.000	1.490
nArCN	Functional group counts	17.051	3.223	5.291	0.000	2.052

Table 3. Correlation matrix between the selected descriptors and λ_{max}

	PW5	piPC07	GATS2v	DISPv	Mor02v	R4u	nArCN	λ_{max}
PW5	1.0000							
piPC07	0.5620	1.0000						
GATS2v	0.2628	0.1074	1.0000					
DISPv	-0.0674	0.3520	-0.0797	1.0000				
Mor02v	0.2960	0.7019	0.2379	0.2084	1.0000			
R4u	-0.0321	0.3679	-0.0830	0.0744	0.4215	1.0000		
nArCN	0.0071	0.4034	-0.0086	0.5932	0.1029	-0.0468	1.0000	
λ_{max}	0.1714	0.7788	0.1504	0.6315	0.6684	0.4527	0.5807	1.0000

descriptors involved in the MLR model are significantly larger than the standard deviation. The *t*-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i.e., descriptors' interactions). Descriptors with *t*-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance.[42] The smaller *t*-probability suggests the more significant descriptor. The *t*-probability values of the seven descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values and the correlation matrix as shown in Table 3 suggest that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

The calculated λ_{max} values from Eq. (7) for the training and test set are shown in Table 1 and Figure 4. The distributions of errors for the entire dataset are given in Figure 5. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the model development. The following statistical parameters were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the

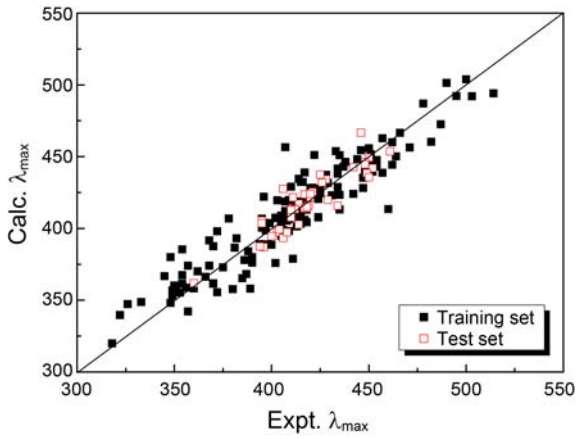

Figure 4. Plot of predicted *vs.* experimental λ_{max} for the entire dataset.

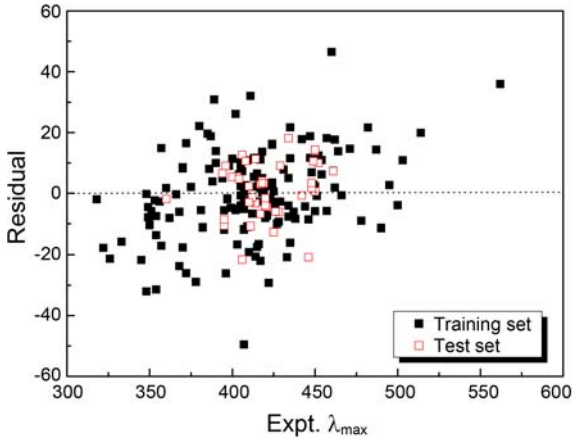

Figure 5. Plot of residual *vs.* experimental λ_{max} for the entire dataset.

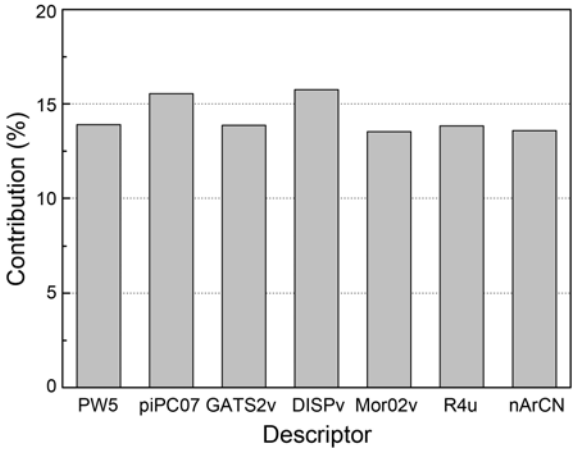

Figure 6. Relative contributions of the selected descriptors to the MLR model.

predictive power of the present model:

$R^2_{CV,ext} = 0.8397 > 0.5$

$r^2 = 0.8236 > 0.6$

$(r^2 - r_0^2)/r^2 = (0.8236 - 0.9963)/0.8236 = -0.2090 < 1$

or $(r^2 - r'^2_0)/r^2 = (0.8236 - 0.9986)/0.8236 = -0.2125 < 0.1$

$0.85 \le k = 1.002 \le 1.15$ or $0.85 \le k' = 0.997 \le 1.15$

Descriptor Contribution Analysis and Interpretation. Based on a previously described procedure,[43,44] the relative contributions of the seven descriptors to the model were determined and are plotted in Figure 6. Seven descriptors were needed in the QSPR model, showing that the analyzed dataset is quite 'noisy' within the data range (318-562 nm), although it is not against the rule of thumb for building a linear model, that is, at least five data point (samples) per descriptor must exist in the model. The significance of the descriptors involved in the model decreases in the following order: DISPv (15.8%) > piPC07 (15.5%) > PW5 (13.9%) GATSs2v (13.9%) > R4u (13.8%) > nArCN (13.6%) > Mor02v (13.5%). It should be noted that the difference in the descriptor contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive models.

The importance of atomic van der Waals volumes on the λ_{max} values is apparent, since the descriptors weighted by atomic van der Waals volumes explain 43.2% of the contributions (15.8% of DISPv, 13.9% of GATS2v and 13.5% of Mor02v). The first important descriptor is DISPv, which has a relatively high correlation coefficient with the experimental λ_{max} values ($R = 0.6315$). The positive coefficient of DISPv indicates that the azobenzenes with larger values for this descriptor would have larger λ_{max} values, since the azobenzenes with larger atomic volumes usually have longer conjugated structures. Thus, this descriptor could be an indicator for the azobenzenes that have a large λ_{max} value.

The second important descriptor is piPC07, which belongs

to the walk and path counts. This type of molecular descriptors is related to molecular branching and size or in general to molecular complexity of graph. When the molecule is bigger and its elemental composition is more complex, this descriptor increases. The coefficient of piPC07 is positive, meaning that the azobenzenes with larger values for this descriptor and accordingly more complex composition would have larger λ_{max} values.

PW5 is a topological descriptor. Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. When this descriptor increases, the λ_{max} decreases.

GATS2v belongs to the Geary autocorrelations, which has a smaller correlation coefficient with the experimental λ_{max} values (R = 0.1504). GATS2v[33] is defined by Eq. (8), where v is the atomic van der Waals volumes, $\overline{v}$ is its average value on the molecule, *nSK* is the number of non-hydrogen atoms, δ_{ij} is the Kronecker delta (δ_{ij} = 1 if d_{ij} = k, zero otherwise, d_{ij} being the topological distance between two considered atoms). Δ is the sum of the Kronecker deltas, i.e. the number of atom pairs at distance equal to k. The positive sign of GATS2v in Eq. (7) indicates that the azobenzenes containing atoms with larger atomic volumes would possess higher λ_{max}, because this descriptor increases with increased atomic volumes.

$$GATS2v = \frac{\frac{1}{2\Delta}\cdot\sum_{i=1}^{nSK}\sum_{j=1}^{nSK}\delta_{ij}\cdot(v_i - v_j)^2}{\frac{1}{(nSK-1)}\cdot\sum_{i=1}^{nSK}(v_i-\overline{v})^2} \quad (8)$$

R4u is a GETAWAY descriptor and correlates with the experimental λ_{max} values of 0.4527. The GETAWAY descriptors[45,46] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. The positive sign of R4u means that the increase in this descriptor increases the λ_{max}.

The functional group count nArCN has a positive sign in Eq. (7), pointing out that the azobenzenes with more nitrile group (aromatic) would possess larger λ_{max} values. The contribution of this descriptor to the λ_{max} values is in agreement with the contribution that one could expect for the influence of the electron-withdrawing moiety.

The last descriptor Mor02v is a 3D-MoRSE descriptor, which correlates with the experimental λ_{max} values of 0.6684. 3D-MoRSE descriptors are the 3D molecular representations of structure based on electron diffraction descriptor,[47,48] 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 0-31 Å^{-1} from the three dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using following expression:

$$\mathrm{Mor}sw = \sum_{i=i}^{nAT-1}\sum_{j=i+1}^{nAT} w_i \cdot w_j \frac{\sin(s\cdot r_{ij})}{s\cdot r_{ij}} \quad (9)$$

where s is the scattering angle, *nAT* is the number of atoms, r_{ij} is the interatomic distance between ith and jth atom, w is an atomic property, including atomic number, masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities. The coefficient for Mor02v is positive, indicating that an increase in Mor02v would result in an increase in λ_{max} values. However, the value and sign of the 3D-MoRSE descriptor depend, to a large extent, on the values of s and r_{ij}.[49] Thus, it could not be concluded that atomic volumes have a specific effect on the λ_{max} values, either negative or positive, only taking into account the coefficient sign of the descriptor. When the coefficient and the descriptor have the same sign, the contribution of the descriptor is positive, else, negative.

Applicability Domain of the MLR Model. It needs to be pointed out that no matter how robust, significant and validated a QSPR model may be, it cannot be expected to reliably predict the modeled property for the entire universe of compounds. Therefore, before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered as reliable.

The AD of the MLR model was analyzed in the Williams plot (shown in Fig. 7). There are one X outlier with leverage higher than the warning limit of 0.1589 (Compound 73) and two Y outliers with residual higher than ± 2.5 s (Compounds 5 and 107) in the training set. Removing these three outliers could improve R^2 between the experimental λ_{max} values and

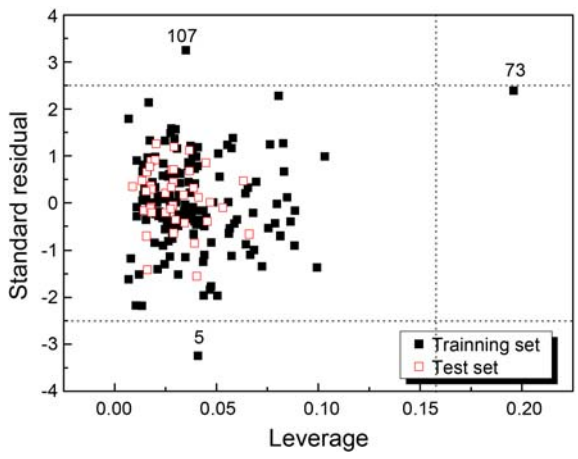

Figure 7. Williams plots of the MLR model for the entire dataset.

the selected descriptors to 0.8895 (R_{CV}^2 = 0.8760) and the standard error decreased to 13.039.

Due to its high predictive ability, the proposed model could be used to screen existing databases or virtual chemical structures to identify organic compounds with desired absorption maxima. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

Conclusions

In this paper, the QSPR method was applied to the prediction of the absorption maxima of azobenzene dyes. A seven-parameter linear model was developed by MLR, with R^2 of 0.8755 and s of 14.476 for the training set. Several validation techniques, including leave-many-out cross-validation, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors involved can be directly calculated from the molecular structure of the compound, thus the proposed model is predictive and could be used to estimate the absorption maxima of azobenzene dyes.

Supporting Information. Figure SI The structures of the studied compounds are available on request from the correspondence author. Fax: +86-27-87426559; Email: xujie0@ustc.edu

Acknowledgments. This work was supported by the Educational Commission of Hubei Province (Q20101606) and the Natural Science Foundation of China (No. 51003082).

References

1. Buttingsrud, B.; Alsberg, B. K.; Åstrand, P.-O. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2226.
2. Zollinger, H. *Color Chemistry. Synthesis, and Applications of Organic Dyes and Pigments*; Wiley-VCH: Weinheim, 2003.
3. Griffiths, J. *Colour and Constitution of Organic Molecules*; Academic Press: London, 1976.
4. Kumar, G. S.; Neckers, D. C. *Chem. Rev.* **1989**, *89*, 1915.
5. Yagai, S.; Karatsu, T.; Kitamura, A. *Chem. Eur. J.* **2005**, *11*, 4054.
6. Zheng, Z.; Xu, J.; Sun, Y.; Zhou, J.; Chen, B.; Zhang, Q.; Wang, K. *J. Polym. Sci. Part A: Polym. Chem.* **2006**, *44*, 3210.
7. Guo, B.; Su, W.; Jia, Y.; Li, Z.; Zhang, Q.; Wang, G. *Phys. Stat. Sol. (b)* **2005**, *242*, 1081.
8. Yesodha, S. K.; Pillai, C. K. S.; Tsutsumi, N. *Prog. Polym. Sci.* **2004**, *29*, 45.
9. Eich, M.; Wendorff, J. H.; Reck, B.; Ringsdorf, H. *Macromol. Chem. Rapid Commun.* **1987**, *8*, 59.
10. Berg, R. H.; Hvilsted, S.; Ramanujam, P. S. *Nature* **1996**, *383*, 505.
11. Natansohn, A.; Rochon, P.; Gosselin, J.; Xie, S. *Macromolecules* **1992**, *25*, 2268.
12. Åstrand, P.-O.; Ramanujam, P. S.; Hvilsted, S.; Bak, K. L.; Sauer, S. P. A. *J. Am. Chem. Soc.* **2000**, *122*, 3482.
13. Natansohn, A.; Rochon, P. *Chem. Rev.* **2002**, *102*, 4139.
14. Wu, S.; Duan, S.; Lei, Z.; Su, W.; Zhang, Z.; Wang, K.; Zhang, Q. *J. Mater. Chem.* **2010**, 5202.
15. .Hélène, M. P.; Faur, C.; Cloirec, P. L. *Chemosphere* **2007**, *77*, 887.
16. Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326.
17. Zhang, C.-R.; Liu, Z.-J.; Chen, Y.-H.; Chen, H.-S.; Wu, Y.-Z.; Yuan, L.-H. *J. Mol. Struct. (THEOCHEM)* **2009**, *899*, 86.
18. Zhang, C.-R.; Liu, Z.-J.; Chen, Y.-H.; Chen, H.-S.; Wu, Y.-Z.; Feng, W.-J.; Wang, D.-B. *Curr. Appl. Phys.* **2010**, *10*, 77.
19. Xu, J.; Zhang, H.; Wang, L.; Liang, G.; Wang, L.; Shen, X.; Xu, W. *Monatsh. Chem.* **2010**, *141*, 549.
20. Xu, J.; Wang, L.; Liang, G.; Bai, Z.; Wang, L.; Xu, W.; Shen, X. *Bull. Korean Chem. Soc.* **2010**, *31*, 2531.
21. Devillers, J.; Balaban, A. T., Eds.; *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999.
22. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
23. 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.
24. Xu, J.; Chen, B.; Zhang, Q.; Guo, B. *Polymer* **2004**, *45*, 8651.
25. Xu, J.; Guo, B.; Chen, B.; Zhang, Q. *J. Mol. Model.* **2005**, *12*, 65.
26. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, 2009.
27. Liu, H.; Wen, Y.; Luan, F.; Gao, Y.; Li, X. *Anal. Chim. Acta* **2009**, *649*, 52.
28. Fayet, G.; Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Rotureau, P.; Adamo, C. *J. Mol. Graphics. Modell.* **2010**, *28*, 465.
29. Xu, J.; Zheng, Z.; Chen, B.; Zhang, Q. *QSAR Comb. Sci.* **2006**, *25*, 372.
30. Xu, J.; Zhang, H.; Wang, L.; Liang, G.; Wang, L.; Shen, X.; Xu, W. *Spectrochim. Acta. A: Mol. Biomol. Spectrosc.* **2010**, *76*, 239.
31. Xu, J.; Zhang, H.; Wang, L.; Liang, G.; Wang, L.; Shen, X. *Mol. Simul.* **2011**, *37*, 1.
32. , Hypercube, Inc.: Gainesville, 2000.
33. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M., TALETE srl: Milan, 2006.
34. Liu, H.; Gramatica, P. *Bioorgan. Med. Chem.* **2007**, *15*, 5251.
35. Kemmrd, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137.
36. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69.
37. Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 35.
38. Holder, A. J.; Yourtee, D. M.; White, D. A.; Glaros, A. G.; Smith, R. *J. Comput. Aid. Mol. Des.* **2003**, *17*, 223.
39. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
40. Shen, M.; Béguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2004**, *47*, 2356.
41. Atkinson, A. *Plots, Transformations, and Regression*; Clarendon Press: Oxford, UK, 1985.
42. Ramsey, L. F.; Schafer, W. D. *The Statistical Sleuth*; Wadsworth Publishing Company: USA, 1997.
43. 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.
44. Guha, R.; Jurs, P. C. *J. Chem. Inf. Model.* **2005**, *45*, 800.
45. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
46. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693.
47. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030.
48. Schuur, J.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334.
49. Saíz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2006**, *14*, 7347.