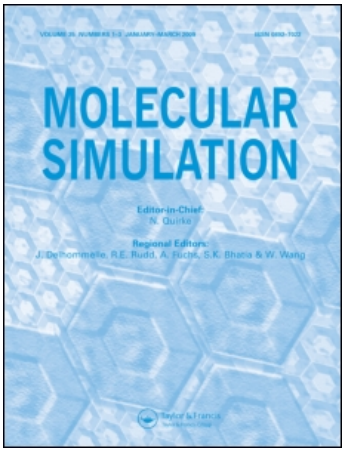

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Artificial neural network-based QSPR study on absorption maxima of organic dyes for dye-sensitised solar cells

Jie Xu[a]; Hui Zhang[a]; Lei Wang[a]; Guijie Liang[ab]; Luoxin Wang[a]; Xiaolin Shen[a]
[a] Key Laboratory of Green Processing and Functional Textiles of New Textile Materials, Wuhan Textile University, Ministry of Education, Wuhan, P.R. China [b] College of Materials Science and Engineering, Xi'an Jiaotong University, Xi'an, P.R. China

Online publication date: 28 January 2011

To cite this Article Xu, Jie , Zhang, Hui , Wang, Lei , Liang, Guijie , Wang, Luoxin and Shen, Xiaolin(2011) 'Artificial neural network-based QSPR study on absorption maxima of organic dyes for dye-sensitised solar cells', Molecular Simulation, 37: 1, 1 — 10
To link to this Article: DOI: 10.1080/08927022.2010.506513
URL: http://dx.doi.org/10.1080/08927022.2010.506513

Molecular Simulation
Vol. 37, No. 1, January 2011, 1–10

Taylor & Francis
Taylor & Francis Group

Artificial neural network-based QSPR study on absorption maxima of organic dyes for dye-sensitised solar cells

Jie Xu[a]*, Hui Zhang[a], Lei Wang[a], Guijie Liang[ab], Luoxin Wang[a] and Xiaolin Shen[a]

[a]*Key Laboratory of Green Processing and Functional Textiles of New Textile Materials, Wuhan Textile University, Ministry of Education, 430073 Wuhan, P.R. China;* [b]*College of Materials Science and Engineering, Xi'an Jiaotong University, 710049 Xi'an, P.R. China*

(*Received 13 March 2010; final version received 27 June 2010*)

A quantitative structure–property relationship study was performed between descriptors representing the molecular structures and the absorption maxima (λ_{max}) of organic dyes for dye-sensitised solar cells. The entire set of 70 dyes was divided into a training set of 53 dyes and a test set of 17 dyes according to Kennard and Stones algorithm. Seven descriptors were selected on the training set by genetic algorithm. Based on these seven descriptors, a nonlinear model with the squared correlation coefficient $R^2 = 0.991$ was developed by using artificial neural networks. The reliability of the proposed model was validated through the test set. All descriptors involved in the model were derived solely from the chemical structures of the dyes, which makes the model very useful to estimate the λ_{max} of the dyes before they are actually synthesised.

Keywords: absorption maxima; QSPR; organic dyes; dye-sensitised solar cells; artificial neural network

1. Introduction

Dye-sensitised solar cells (DSSCs) have attracted considerable attention because of their high efficiency and low cost [1–6] of conversion of sunlight into electricity. The most efficient sensitisers employed in these cells are ruthenium polypyridyl complexes, yielding solar-to-electric power conversion efficiencies of up to 11–12% under simulated sunlight [6]. However, Ru polypyridyl complexes contain a heavy metal, which is undesirable for environmental reasons [7]. Moreover, the process of synthesising the complexes is complicated and costly. As well as Ru complexes, pure organic dyes as sensitisers are also under intensive investigation, due to their high molar extinction coefficients, flexible structural modifications and low costs, and possibly good efficiency [8–15].

The primary process of DSSCs is similar to photographic sensitisation, in which the dye absorbs light and its photo-excited state injects an electron into the TiO_2 conduction band. Thus, the optimal sensitiser for DSSCs should be panchromatic, that is, absorb solar radiation strongly in the visible or near-infrared region [3]. Ideally, all photons below a threshold wavelength of about 920 nm should be harvested and converted to electric current. Moreover, the sensitiser should fulfil several demanding conditions: (1) it must be firmly grafted to the semiconductor surface and inject electrons into the conduction band with a quantum yield of unity, (2) its redox potential should be sufficiently high so that it can be regenerated rapidly via electron donation from the electrolyte or a hole conductor and (3) it should be stable enough to sustain at least 10^8 redox turnovers under illumination corresponding to about 20 years of exposure to natural sunlight [3]. Since it is very difficult to calculate all the necessary parameters and then to synthesise a dye fulfilling all of them, the method of 'trials and failures' is still an important approach to finding appropriate dye sensitisers. Methods for quantitatively predicting the absorption maxima (λ_{max}) of dyes from their molecular structures undoubtedly would be of significant utility in the molecular design for the exploration of a new sensitiser. So far, the main computational efforts to simulate the absorption spectra have been based on quantum chemistry calculations, such as density functional theory (DFT) and *ab initio* methods. However, such accurate calculation of absorption profiles is very 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 [16,17].

Alternatively, the quantitative structure–property relationship (QSPR) provides a promising approach for estimating the λ_{max} of dyes based on descriptors derived solely from the molecular structure to fit experimental data. The QSPR approach is based on the assumption that the variation of the behaviour 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' [18–21]. The advantage of this approach lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once

*Corresponding author. Email: xujie0@ustc.edu

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927022.2010.506513
http://www.informaworld.com

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 approach can expedite the process of development of new molecules and materials with desired properties. The approach has been successfully used to investigate the spectral properties of various systems, for example the prediction of the absorption maxima of second-order nonlinear optical chromophores using stepwise multilinear regression analysis [22], the estimation of the excitation and emission maxima of green fluorescent protein chromophores using artificial neural networks (ANNs) [23], the prediction of maximum absorption wavelength of flavones using heuristic methods and radial basis function neural network [24], and the modelling of relative fluorescence intensity ratio of Eu(III) complex in different solvents [25]. However, there are relatively few attempts to correlate and predict the absorption maxima of organic dyes for DSSCs.

The goal of this paper is to develop a robust QSPR model to predict the λ_{max} values of the dyes for DSSCs by using ANNs. ANN-based modelling methods could produce more accurate QSPR models compared to linear regression methods, since they have the ability to handle the possible nonlinear relationships during the training process [26].

2. Materials and method

2.1 Data-set

A total of 70 organic dyes with extensive structural diversity were selected as the data-set; their molecular structures are depicted in Figure 1. The quality and robustness of the predictive power of a QSPR model depends heavily on the diversity of the data-set. To select significant descriptors for the QSPR model that captures all the underlying interaction mechanisms, it is advisable to have as many structural features represented in the data-set as possible. The working data-set included hemicyanine, squaraine, indoline, coumarin, polyene, tetrahydroquinoline, phenyl-conjugated oligoene and so on. As the experimental λ_{max} values are usually measured in solution and depend significantly on the solvent, all λ_{max} taken from literatures [11–13,27–45] were measured in ethanol. The reported λ_{max} values span between 378 and 660 nm, as shown in Table 1.

2.2 Descriptors generation

The structures of all molecules were preoptimised using MM+ molecular mechanics force field (Polak–Ribiere algorithm) in the HYPERCHEM program [46]. 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 0.02 kcal $Å^{-1}$ mol^{-1} as a stopping criterion. A total of 1164 molecular descriptors for each molecule were calculated from the optimised molecular geometries using DRAGON software [47]. These descriptors include: (a) 0D constitutional (atom and group counts); (b) 1D functional groups and atom-centred fragments; (c) 2D topological, BCUT (Burden Chemical Abstract Service University of Texas), 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, weighted holistic invariant molecular (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 [48]) were excluded in a pre-reduction step. Thus, 777 descriptors remained to undergo subsequent descriptor selection.

2.3 Kennard and Stones algorithm

The Kennard and Stones algorithm [49] has been widely used for splitting one data-set 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 data-set 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 no validation samples fall outside the measured region. The Kennard and Stones algorithm has been considered as one of the best ways to build training and test sets [50,51]. Using the Kennard and Stones algorithm, the entire set of dyes was divided into two subsets: a training set of 53 dyes and a test set including the remaining 17 dyes.

2.4 Model development and validation

The most commonly used variable selection method in QSPR studies is the stepwise approach, which may be classified as forward or backward. However, the stepwise approach has two main disadvantages: (1) each choice heavily affects the following choices and (2) the final results are expressed by a single combination, and then no choice is given to the researchers [52].

Genetic algorithm (GA) [53] is a stochastic optimisation method inspired by genetics and Darwinian theory; it is based on the evolution of a starting random population of models, which provides an optimal or near optimal solution through mutation, cross-over and selection after a number of generations. In this work, the best subset of descriptors was selected using the genetic function approximation

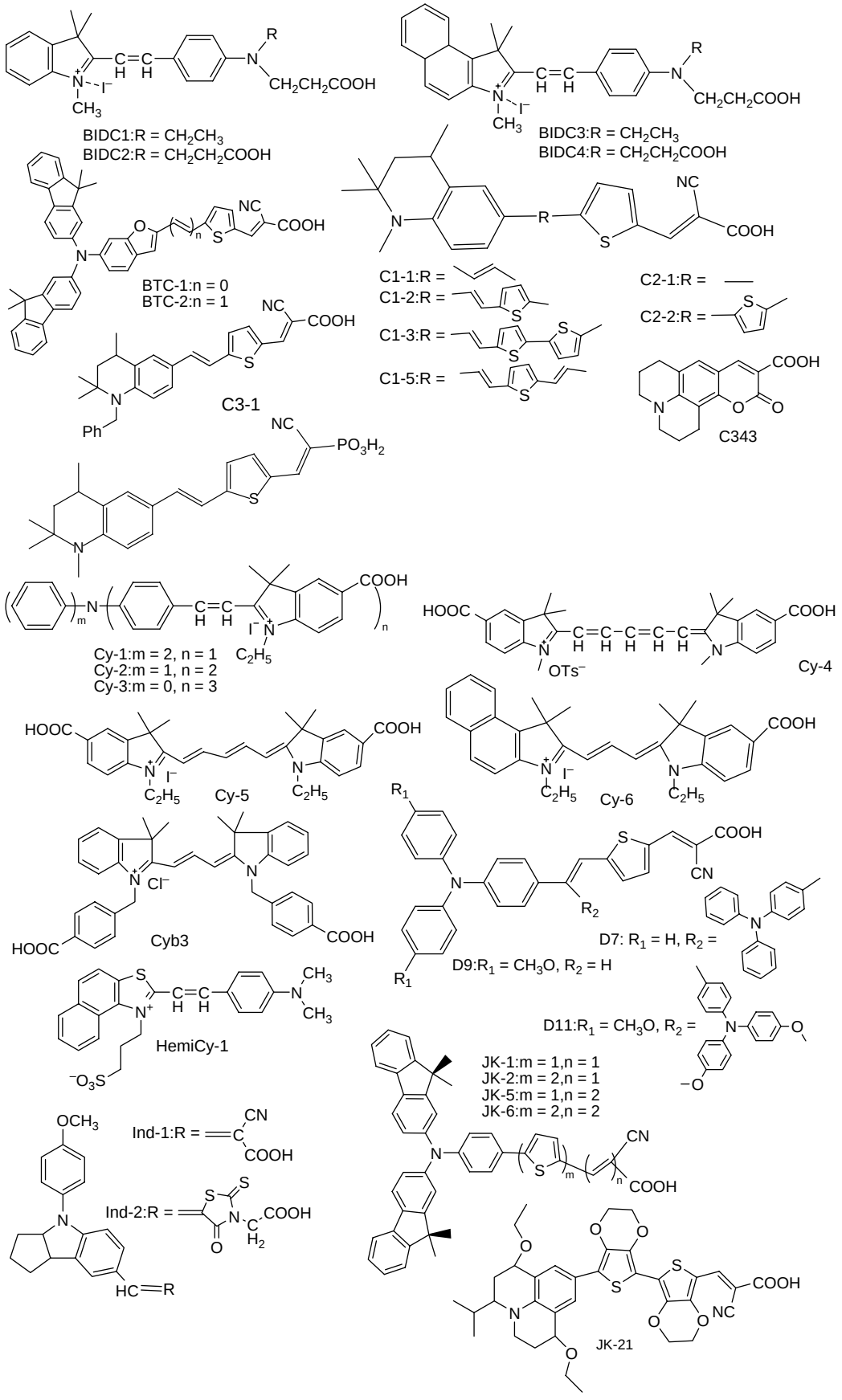

Figure 1. Structures of the dye sensitisers included in the data-set.

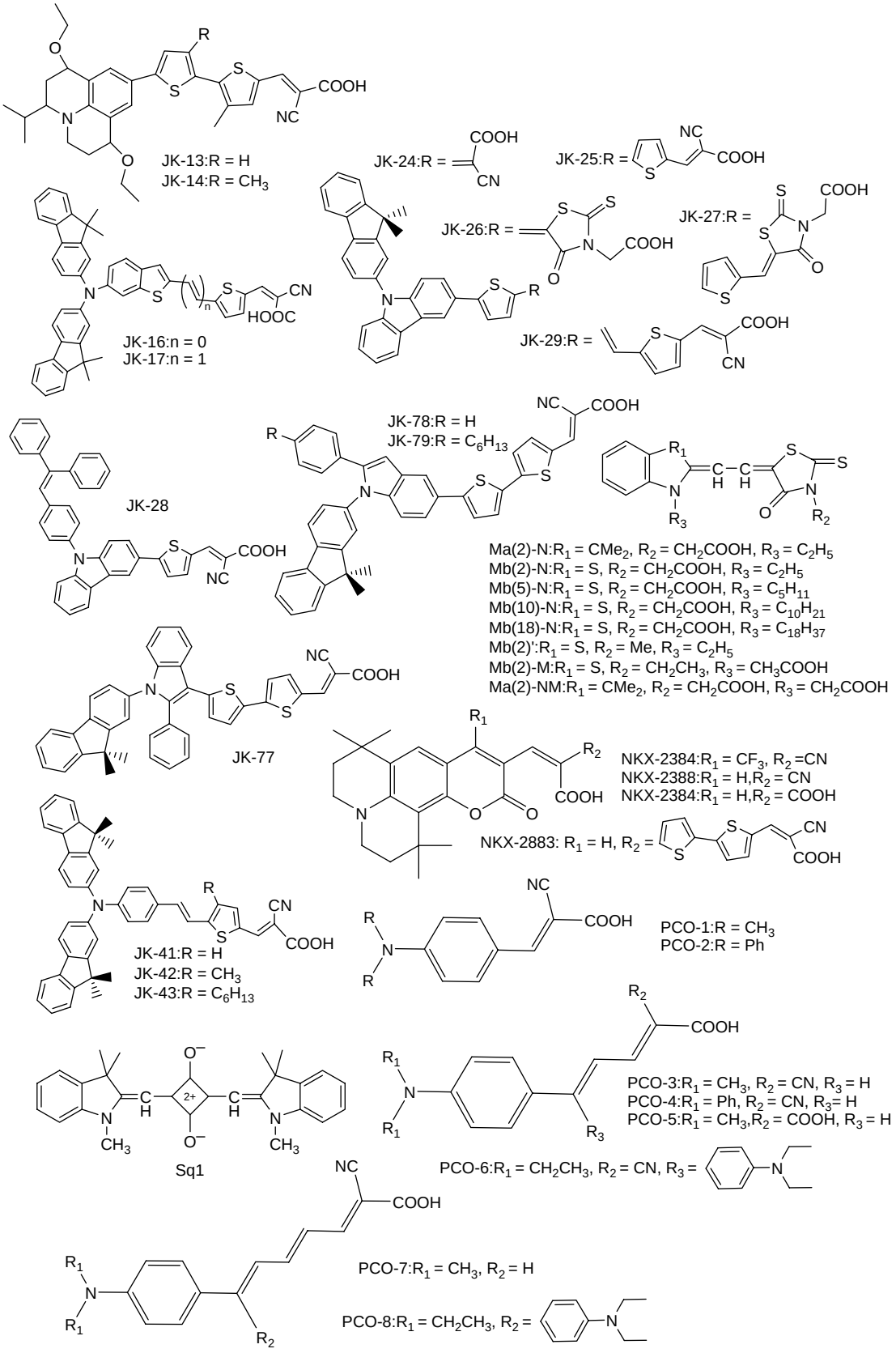

Figure 1. *Continued.*

Table 1. Dyes used in this study with experimental and calculated λ_{max} (nm).

Structure	Exp. (λ_{max})	Calc. (λ_{max})	Diff.	Structure	Exp. (λ_{max})	Calc. (λ_{max})	Diff.
BIDC1[a] [28]	561	563.1	−2.1	JK-24 [44]	411	394.8	16.2
BIDC2[a] [28]	560	558.7	1.3	JK-25 [44]	435	430.1	4.9
BIDC3[a] [28]	577	572.1	4.9	JK-26 [44]	469	469.8	−0.8
BIDC4 [28]	578	578.6	−0.6	JK-27 [44]	489	484.9	4.1
BTC-1[a] [34]	463	464.4	−1.4	JK-28 [44]	406	415.9	−9.9
BTC-2 [34]	479	477.3	1.7	JK-29 [44]	451	452.0	−1.0
C1-1 [35]	468	458.7	9.3	JK-41 [45]	461	452.2	8.8
C1-2 [35]	472	468.1	3.9	JK-42 [45]	463	457.5	5.5
C1-3 [35]	475	478.0	−3.0	JK-43[a] [45]	456	448.4	7.6
C1-5 [35]	492	495.1	−3.1	JK-5 [33]	438	449.2	−11.2
C2-1[a] [35]	441	427.4	13.6	JK-6 [33]	459	448.4	10.6
C2-2 [35]	462	471.7	−9.7	JK-77[a] [32]	433	428.0	5.0
C3-1[a] [35]	470	475.4	−5.4	JK-78 [32]	429	432.1	−3.1
C343 [29]	442	437.0	5.0	JK-79 [32]	431	447.6	−16.6
C4-1 [35]	455	456.5	−1.5	Ma(2)-N [31]	490	484.2	5.8
Cy-1 [36]	550	551.1	−1.1	Ma(2)-NM [31]	500	503.7	−3.7
Cy-2 [36]	564	568.0	−4.0	Mb(10)-N[a] [31]	520	523.9	−3.9
Cy-3[a] [36]	572	572.9	−0.9	Mb(18)-N[a] [31]	520	518.5	1.5
Cy-4 [37]	660	656.6	3.4	Mb(2)′[a] [31]	520	527.2	−7.2
Cy-5 [38]	571	570.9	0.1	Mb(2)-M[a] [31]	520	518.1	1.9
Cy-6 [38]	651	651.3	−0.3	Mb(2)-N [31]	520	522.6	−2.6
Cyb3 [39]	560	561.7	−1.7	Mb(5)-N[a] [31]	520	517.7	2.3
D11 [13]	458	457.6	0.4	NKX-2384 [29]	477	482.9	−5.9
D7 [13]	441	448.5	−7.5	NKX-2388 [29]	493	493.1	−0.1
D9[a] [13]	462	451.1	10.9	NKX-2393 [29]	486	484.5	1.5
HemiCy-1 [40]	536	533.1	2.9	NKX-2883 [30]	552	551.4	0.6
Ind-1[a] [11]	390	403.3	−13.3	PCO-1 [12]	378	378.0	0.0
Ind-2 [11]	483	482.6	0.4	PCO-2 [12]	386	390.6	−4.6
JK-1 [41]	436	443.3	−7.3	PCO-3 [12]	412	410.9	1.1
JK-13[a] [42]	462	459.2	2.8	PCO-4 [12]	417	418.2	−1.2
JK-14 [42]	422	419.3	2.7	PCO-5 [12]	416	412.6	3.4
JK-16 [43]	456	452.5	3.5	PCO-6 [12]	443	441.8	1.2
JK-17 [43]	476	478.0	−2.0	PCO-7 [12]	434	448.1	−14.1
JK-2 [41]	452	449.3	2.7	PCO-8 [12]	470	471.3	−1.3
JK-21 [42]	506	503.6	2.4	Sq1 [27]	626	625.8	0.2

[a] Data used for the test set.

approach [54] and then submitted to a three-layer fully connected feed-forward ANN. The number of input neurons was equal to that of the selected descriptors. The number of hidden neurons was optimised by trial and error procedure on calculations of the training process [55]. One output neuron was used to represent the experimental λ_{max}. The network was trained using the quasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [56]. To avoid overtraining, from the training set, 1/10 systems 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.

An external validation of the model was further performed by using the data that was not used in the prediction model development as an external test set. The $R^2_{CV,ext}$ for the test set is determined by Equation (1):

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

where $\bar{y}_{test}$ is the average value for the response variable of the test set. According to Golbraikh and Tropsha [57], a QSPR model is successful if it satisfies the following conditions:

$$\begin{aligned} &R^2_{CV,ext} > 0.5 \\ &r^2 > 0.6 \\ &\left|\frac{(r^2 - r_0^2)}{r^2}\right| < 0.1 \quad \text{or} \quad \left|\frac{\left(r^2 - r'^2_0\right)}{r^2}\right| < 0.1 \\ &0.85 \le k \le 1.15 \quad \text{or} \quad 0.85 \le k' \le 1.15. \end{aligned} \quad (2)$$

Table 2. Descriptors involved in the final model.

Descriptor	Type	Definition
HOMA	Geometrical	Harmonic oscillator model of aromaticity index
RDF095v	RDF	Radial distribution function – 9.5/weighted by atomic van der Waals volumes
Mor09u	3D MoRSE	3D-MoRSE – signal 09/unweighted
Mor26v	3D MoRSE	3D-MoRSE – signal 26/weighted by atomic van der Waals volumes
G1p	WHIM	1st component symmetry directional WHIM index/weighted by atomic polarisabilities
R2m	GETAWAY	*R* autocorrelation of lag 2/weighted by atomic masses
R6m+	GETAWAY	*R* maximal autocorrelation of lag 6/weighted by atomic masses

Here, r is the correlation coefficient between the calculated and experimental values in the test set. r_0^2 (calculated vs. observed values) and $r_0'^2$ (observed vs. calculated values) are the coefficients of determination. k and k' are slopes of regression lines through the origin of calculated vs. observed, and observed vs. calculated, respectively. Detailed mathematical definitions of these parameters can be found in Golbraikh and Tropsha [57].

3. Results and discussion

The experimental data of λ_{max} values in Table 1 were divided into the training and test sets according to the Kennard and Stones algorithm. At first, GA-MLR was applied on the training set to select the best subset of descriptors and to build linear models. The best subset of seven descriptors (shown in Table 2), harmonic oscillator model of aromaticity (HOMA) index, RDF095v, Mor09u, Mor26v, G1p, R2m and R6m+, was selected. However, no statistically significant linear models could be found (R^2 of 0.929 and 0.726 for the training and test sets, respectively). It was then decided to use an ANN approach to build a nonlinear model.

The ANN has become an important and widely used nonlinear modelling technique for QSPR studies. The mathematical adaptability of ANN commends it as a powerful tool for pattern classification and building predictive models. A particular advantage of ANN is its inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function. Among the neural network learning algorithms, the back-propagation (BP) method [58] is one of the most commonly used methods. The drawback of BP is that the training processes slowly, because the gradient-descent algorithm is usually used for 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 the training is much more rapid [59]. The seven descriptors were used as inputs to the network. The number of hidden neurons is an important parameter influencing the performances of the ANN. The usual rule of thumb is that the weights and biases should be less than the samples so that the model achieved by the network is stationary [60]. In the situation of this work, with 53 samples in the training set, the number of the hidden neurons should not, therefore, be

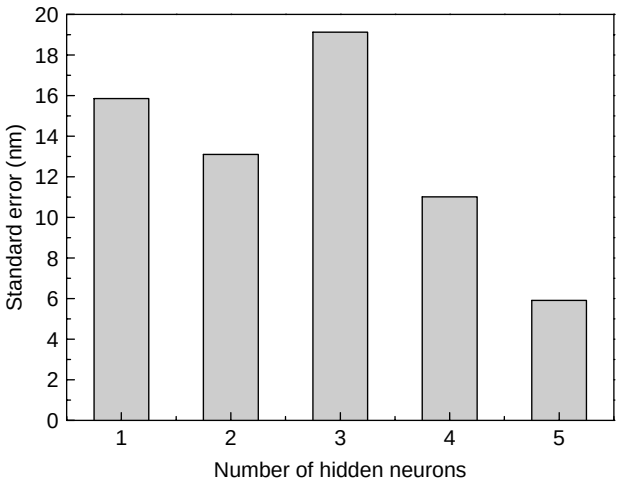

Figure 2. Standard error vs. number of neurons in the hidden layer.

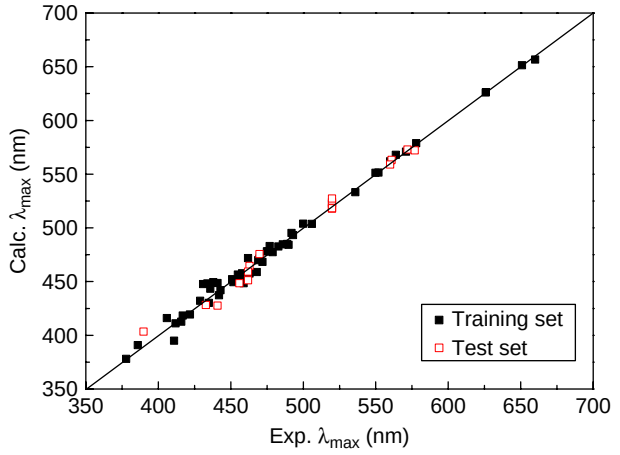

Figure 3. Plot of calculated vs. experimental λ_{max} values for the entire data-set.

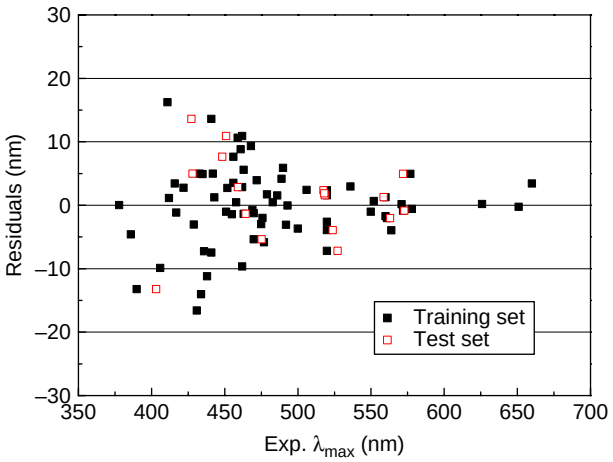

Figure 4. Plot of residuals vs. experimental λ_{max} values for the entire data-set.

greater than five. Figure 2 shows the changes in the standard error of estimation (s), while optimising the neural network architecture with respect to the number of hidden neurons. Thus, a 7-5-1 network architecture was obtained after a rigorous trial and error procedure with a learning period size of 428.

The results by the ANN model for the entire data-set are given in Table 1 and Figure 3. The obtained R^2 and s are 0.991 and 5.9 nm, respectively, which indicate a very good agreement between the correlation and the variation in the data. Leave-one-out cross-validation has also been carried out and R^2_{CV} of 0.990 was obtained, which confirms the reliability of the model by focusing on the sensitivity of the model to the elimination of any single data point. Figure 4 shows the calculated residuals against the experimental λ_{max} values for the entire data-set. As the residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the model development. The min/max values of the absolute error are 0.5/16.6 and 0.9/13.6 nm for the training and test sets, respectively. The following statistical parameters are obtained for the test set, which obviously satisfy the generally accepted conditions:

$$R^2_{CV,ext} = 0.985 > 0.5$$

$$r^2 = 0.986 > 0.6$$

$$\left|\frac{\left(r^2 - r_0^2\right)}{r^2}\right| = \left|\frac{(0.986 - 0.9997)}{0.986}\right| < 0.1 \quad \text{or}$$

$$\left|\frac{\left(r^2 - r_0'^2\right)}{r^2}\right| = \left|\frac{(0.986 - 0.9996)}{0.986}\right| < 0.1$$

$$0.85 \le k = 0.9980 \le 1.15 \quad \text{or}$$

$$0.85 \le k' = 1.002 \le 1.15.$$

The descriptors appearing in the ANN model encode 3D aspects of the molecular structure, and can be classified as follows: (1) a geometrical descriptor – HOMA index; (2) a radial distribution function (RDF) descriptor – RDF095v, RDF, 9.5/weighted by atomic van der Waals volumes; (3) two 3D MoRSE descriptors – Mor09u, signal 09/unweighted and Mor26v, signal 26/weighted by atomic van der Waals volumes; (4) a WHIM descriptor – G1p, 1st component symmetry directional WHIM index weighted by atomic polarisabilities; and (5) two GETAWAY descriptors – R2m, R autocorrelation of lag 2/weighted by atomic masses and R6m+, R maximal autocorrelation of lag 6/weighted by atomic masses.

The HOMA index is based on the degree of alternation of single/double bonds, measuring the bond length deviations from the optimal lengths attributed to the typical aromatic state [61]. The HOMA index is calculated as follows:

$$\mathrm{HOMA} = 1 - \frac{\sum_k \alpha_k \sum_{b=1}^{B_{\pi k}} \left(r_k^{opt} - r_b\right)^2}{B_\pi}, \qquad (3)$$

where the first sum runs over each aromatic bond type; $B_{\pi k}$ is the number of considered π-bond contributions of the kth aromatic bond type; r_b is the actual bond length; α_k and r_k^{opt} are numerical constants of the typical aromatic bond length referring to the kth aromatic bond type; and B_π is the total number of aromatic bonds.

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

$$\mathrm{RDF}rw = f \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j \, e^{-\beta(r-r_{ij})^2}, \qquad (4)$$

where f is a scaling factor (assumed equal to 1 in the calculations); w_i and w_j are characteristic properties of the atoms i and j (including unweighted, atomic 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 [62]. RDFrw is generally calculated at a number of discrete points with defined intervals. Besides information about interatomic distances in the entire molecule, RDF095v provides further information about atomic van der Waals volumes.

3D MoRSE descriptors are the 3D molecular representations of structure based on electron diffraction descriptors [63,64], 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 3D atomic coordinates of a molecule. The 3D MoRSE descriptor is

calculated using the following expression:

$$\mathrm{Mor}sw = \sum_{i=i}^{n\mathrm{AT}-1} \sum_{j=i+1}^{n\mathrm{AT}} w_i w_j \frac{\sin(sr_{ij})}{sr_{ij}}, \quad (5)$$

where s is the scattering angle. For the case of Mor09u, an unweighted scheme was used and $s = 8\,\text{Å}^{-1}$, while for Mor26v, a volume-weighted scheme was used and $s = 25\,\text{Å}^{-1}$.

WHIM descriptors are based on statistical indices calculated on the projections of atoms along principal axes [65,66], which are built to capture 3D information regarding size, shape, symmetry and atom distributions with respect to invariant reference frames. To calculate them, six different weighting schemes for the atoms are adopted: the same weighting schemes used in the RDF definition as well as atomic electrotopological state. For each weighting scheme, a set of statistical indices is calculated on the atoms projected onto the principal axes. Essentially, the WHIM descriptors provide a variety of principal axes with respect to a defined atomic property. G1p is the 1st component symmetry directional WHIM descriptor, which involves the atomic polarisabilities as the weighting scheme.

GETAWAY descriptors [67,68] 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. R2m and R8m+ are calculated by Equations (6) and (7), respectively, where h_{ii} and h_{jj} are the leverages of ith and jth atom, d_{ij} is the topological distance and $\delta(k, d_{ij})$ is a Dirac-delta function ($\delta = 1$ if $d_{ij} = k$, zero otherwise).

$$\mathrm{R}kw = \sum_{i=1}^{n\mathrm{AT}-1} \sum_{j>1} \frac{\sqrt{h_{ii}h_{jj}}}{r_{ij}} w_i w_j \delta(k, d_{ij}), \quad (6)$$

$$\mathrm{R}kw+ = \max_{ij}\left(\frac{\sqrt{h_{ii}h_{jj}}}{r_{ij}} w_i w_j \delta(k, d_{ij})\right). \quad (7)$$

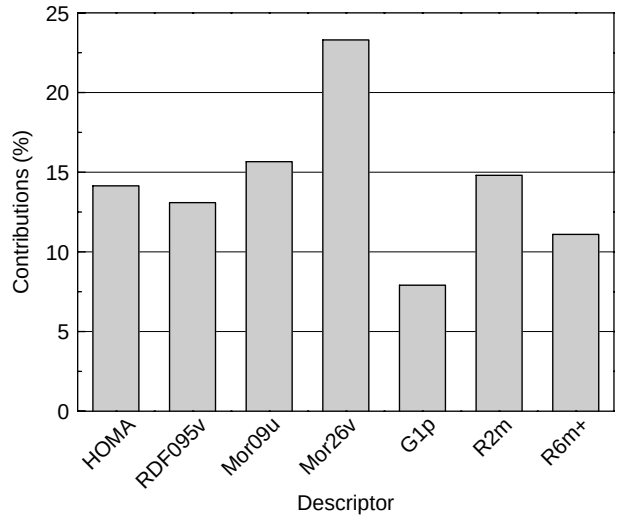

Figure 5. Relative contributions of the seven descriptors to the QSPR model.

Based on a previously described procedure [69,70], the relative contributions of the seven descriptors to the model were determined and are plotted in Figure 5. The significance of these descriptors decreases in the following order: Mor26v > Mor09u > R2m > HOMA > RDF095v > R6m+ > G1p.

Generally speaking, an increase in the extent of the π-electron system leads to a red shift in the absorption. However, it is not a trivial task to interpret the ANN model directly due to its complex modelling procedure and vague output. To make the interpretation easier, the dependency of the λ_{max} values on each descriptor was tested. The most relevant descriptor in the present ANN model is Mor26v, implying that the λ_{max} values have a significant dependence on the size of the molecules. This is further supported by the presence of RDF095v in the ANN model which is also weighted by atomic volumes, although the contribution of this descriptor is relatively low. R2m and R6m+ are the descriptors weighted by atomic masses with positive contribution. The favourable nature of these coefficients suggests that the atomic mass may play an important role in the process of light absorption. The aromaticity index HOMA reflects the planar and rigid geometry of the molecule and contributes positively to the absorption wavelength. The positive nature of this descriptor is in good agreement with the theory that an electron is easier to transfer to the molecule of an aromatic plane. G1p is related to conventional polarisability while allowing for attenuation of the influence of more remote atoms and bonds. The negative contribution of G1p indicates that the absorption wavelength would increase with increasing polarisability. Additionally, since all descriptors (including Mor09u) encode 3D information that depends on the conformation of the molecule, it is possible to argue that the λ_{max} values of the present set of dye sensitisers have a considerable dependence on conformational changes.

4. Conclusions

In this paper, an ANN-based QSPR model with $R^2 = 0.991$ and $s = 5.9$ nm was reported to predict the λ_{max} values for a set of diverse dye sensitisers for DSSCs. The descriptors appearing in the ANN model are HOMA, RDF095v, Mor09u, Mor26v, G1p, R2m and R6m+, indicating that the λ_{max} values of the dye sensitisers

depend significantly on the size, mass, polarisability and aromatic plane of the molecules. Also, the molecular conformational changes have a considerable effect on the λ_{max} values of the dye sensitisers as considering the 3D aspects of the ANN model. The model relies solely on descriptors derived from the molecular structure and, thus, it is applicable to dye sensitisers of any chemical structure. Therefore, this QSPR model should be useful in the prediction of the λ_{max} of new sensitisers for DSSCs.

Acknowledgements

This work was supported by the Foundation of Wuhan Textile University (No. 2009003), the Natural Science Foundation of Hubei Province (No. 2008CDB261) and China (No. 51003082), the Key Project of Science and Technology Research of Ministry of Education (No. 208089) and the Educational Commission of Hubei Province (Q20101606). The authors gratefully wish to express their thanks to the reviewers for critically reviewing the manuscript and making important suggestions.

References

[1] B. O'Regan and M. Grätzel, *A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films*, Nature 353 (1991), pp. 737–740.

[2] M.K. Nazeeruddin, A. Kay, I. Rodicio, R. Humpbry-Baker, E. Miiller, P. Liska, N. Vlachopoulos, and M. Grätzel, *Conversion of light to electricity by cis-X_2Bis(2,2′-bipyridyl-4,4′-dicarboxylate)ruthenium(II) charge-transfer sensitizers (X=Cl^-, Br^-, I^-, CN^-, and SCN^-) on nanocrystalline TiO_2 electrodes*, J. Am. Chem. Soc. 115 (1993), pp. 6382–6390.

[3] M.K. Nazeeruddin, P. Péchy, T. Renouard, S.M. Zakeeruddin, R. Humphry-Baker, P. Comte, P. Liska, L. Cevey, E. Costa, V. Shklover, L. Spiccia, G.B. Deacon, C.A. Bignozzi, and M. Grätzel, *Engineering of efficient panchromatic sensitizers for nanocrystalline Tio_2-based solar cells*, J. Am. Chem. Soc. 123 (2001), pp. 1613–1624.

[4] M.K. Nazeeruddin, F.D. Angelis, S. Fantacci, A. Selloni, G. Viscardi, P. Liska, S. Ito, B. Takeru, and M. Grätzel, *Combined experimental and DFT-TDDFT computational study of photoelectrochemical cell ruthenium sensitizers*, J. Am. Chem. Soc. 127 (2005), pp. 16835–16847.

[5] F. Gao, Y. Wang, D. Shi, J. Zhang, M. Wang, X. Jing, R. Humphry-Baker, P. Wang, S.M. Zakeeruddin, and M. Grätzel, *Enhance the optical absorptivity of nanocrystalline TiO_2 film with high molar extinction coefficient ruthenium sensitizers for high performance dye-sensitized solar cells*, J. Am. Chem. Soc. 130 (2008), pp. 10720–10728.

[6] C.-Y. Chen, M. Wang, J.-Y. Li, N. Pootrakulchote, L. Alibabaei, C.-H. Nagoc-le, J.-D. Decoppet, J.-H. Tsai, C. Grätzel, C.-G. Wu, S.M. Zakeeruddin, and M. Grätzel, *Highly efficient light-harvesting ruthenium sensitizer for thin-film dye-sensitized solar cells*, ACS Nano 3 (2009), pp. 3103–3109.

[7] Y. Amao and T. Komori, *Bio-photovoltaic conversion device using chlorine-e6 derived from chlorophyll from Spirulina adsorbed on a nanocrystalline TiO_2 film electrode*, Biosens. Bioelectron. 19 (2004), pp. 843–847.

[8] T. Horiuchi, H. Miura, K. Sumioka, and S. Uchida, *High efficiency of dye-sensitized solar cells based on metal-free indoline dyes*, J. Am. Chem. Soc. 126 (2004), pp. 12218–12219.

[9] K. Hara, M. Kurashige, Y. Dan-oh, C. Kasada, A. Shinpo, S. Suga, K. Sayama, and H. Arakawa, *Design of new coumarin dyes having thiophene moieties for highly efficient organic-dye-sensitized solar cells*, New J. Chem. 27 (2003), pp. 783–785.

[10] T. Horiuchi, H. Miura, and S. Uchida, *Highly-efficient metal-free organic dyes for dye-sensitized solar cells*, Chem. Commun. (2003), pp. 3036–3037.

[11] T. Horiuchi, H. Miura, and S. Uchida, *Highly efficient metal-free organic dyes for dye-sensitized solar cells*, J. Photochem. Photobiol. A Chem. 164 (2004), pp. 29–32.

[12] T. Kitamura, M. Ikeda, K. Shigaki, T. Inoue, N.A. Anderson, X. Ai, T. Lian, and S. Yanagida, *Phenyl-conjugated oligoene sensitizers for TiO_2 solar cells*, Chem. Mater. 16 (2004), pp. 1806–1812.

[13] D.P. Hagberg, J.-H. Yum, H. Lee, F.D. Angelis, T. Marinado, K.M. Karlsson, R. Humphry-Baker, L. Sun, A. Hagfeldt, M. Grätzel, and M.K. Nazeeruddin, *Molecular engineering of organic sensitizers for dye-sensitized solar cell applications*, J. Am. Chem. Soc. 130 (2008), pp. 6259–6266.

[14] H. Tian, X. Yang, R. Chen, R. Zhang, A. Hagfeldt, and L. Sun, *Effect of different dye baths and dye-structures on the performance of dye-sensitized solar cells based on triphenylamine dyes*, J. Phys. Chem. C 112 (2008), pp. 11023–11033.

[15] A. Mishra, M.K.R. Fischer, and P. Bäuerle, *Metal-free organic dyes for dye-sensitized solar cells: From structure:property relationships to design rules*, Angew. Chem. Int. Ed. 48 (2009), pp. 2474–2499.

[16] C.-R. Zhang, Z.-J. Liu, Y.-H. Chen, H.-S. Chen, Y.-Z. Wu, and L.-H. Yuan, *DFT and TDDFT study on organic dye sensitizers D5, DST and DSS for solar cells*, J. Mol. Struct. (Theochem) 899 (2009), pp. 86–93.

[17] C.-R. Zhang, Z.-J. Liu, Y.-H. Chen, H.-S. Chen, Y.-Z. Wu, W.-J. Feng, and D.-B. Wang, *DFT and TD-DFT study on structure and properties of organic dye sensitizer TA-St-CA*, Curr. Appl. Phys. 10 (2010), pp. 77–83.

[18] J. Devillers and A.T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, The Netherlands, 1999.

[19] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley-Interscience, New York, 2000.

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

[21] 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.

[22] 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.

[23] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. Tansila, T. Naenna, and V. Prachayasittikul, *Prediction of GFP spectral properties using artificial neural network*, J. Comput. Chem. 28 (2007), pp. 1275–1289.

[24] H. Liu, Y. Wen, F. Luan, Y. Gao, and X. Li, *Quantitative structure–λ_{max} relationship study on flavones by heuristic method and radial basis function neural network*, Anal. Chim. Acta 649 (2009), pp. 52–61.

[25] J. Xu, Q. Xiong, B. Chen, L. Wang, L. Liu, and W. Xu, *Modeling the relative fluorescence intensity ratio of Eu(III) complex in different solvents based on QSPR method*, J. Fluoresc. 19 (2009), pp. 203–209.

[26] A. Afantitis, G. Melagraki, K. Makridima, A. Alexandridis, H. Sarimveis, and O. Iglessi-Markopoulou, *Prediction of high weight polymers glass transition temperature using RBF neural networks*, J. Mol. Struct. (Theochem) 716 (2005), pp. 193–198.

[27] W. Zhao, Y.J. Hou, X.S. Wang, B.W. Zhang, Y. Cao, R. Yang, W.B. Wang, and X.R. Xiao, *Study on squarylium cyanine dyes for photoelectric conversion*, Sol. Energy Mater. Sol. Cells 58 (1999), pp. 173–183.

[28] Q.-H. Yao, F.-S. Meng, F.-Y. Li, H. Tian, and C.-H. Huang, *Photoelectric conversion properties of four novel carboxylated hemicyanine dyes on TiO_2 electrode*, J. Mater. Chem. 13 (2003), pp. 1048–1053.

[29] K. Hara, T. Sato, R. Katoh, A. Furube, Y. Ohga, A. Shinpo, S. Suga, K. Sayama, H. Sugihara, and H. Arakawa, *Molecular design of coumarin dyes for efficient dye-sensitized solar cells*, J. Phys. Chem. B 107 (2003), pp. 597–606.

[30] Z.-S. Wang, Y. Cui, K. Hara, Y. Dan-oh, C. Kasada, and A. Shinpo, *A high-light-harvesting-efficiency coumarin dye for stable dye-sensitized solar cells*, Adv. Mater. 19 (2007), pp. 1138–1141.

[31] K. Sayama, K. Hara, N. Mori, M. Satsuki, S. Suga, S. Tsukagoshi, Y. Abe, H. Sugihara, and H. Arakawa, *Photosensitization of a porous TiO_2 electrode with merocyanine dyes containing a carboxyl group and a long alkyl chain*, Chem. Commun. (2000), pp. 1173–1174.

[32] D. Kim, M.-S. Kang, K. Song, S.O. Kang, and J. Ko, *Molecular engineering of organic sensitizers containing indole moiety for dye-sensitized solar cells*, Tetrahedron 64 (2008), pp. 10417–10424.

[33] S. Kim, H. Choi, D. Kim, K. Song, S.O. Kang, and J. Ko, *Novel conjugated organic dyes containing bis-dimethylfluorenyl amino phenyl thiophene for efficient solar cell*, Tetrahedron 63 (2007), pp. 9206–9212.

[34] I. Jung, J.K. Lee, K.H. Song, K. Song, S.O. Kang, and J. Ko, *Synthesis and photovoltaic properties of efficient organic dyes containing the benzo[b]furan moiety for solar cells*, J. Org. Chem. 72 (2007), pp. 3652–3658.

[35] R. Chen, X. Yang, H. Tian, X. Wang, A. Hagfeldt, and L. Sun, *Effect of tetrahydroquinoline dyes structure on the performance of organic dye-sensitized solar cells*, Chem. Mater. 19 (2007), pp. 4007–4015.

[36] F.S. Meng, Q.H. Yao, J.G. Shen, F.L. Li, C.H. Huang, K.C. Chen, and H. Tian, *Novel cyanine dyes with multi-carboxyl groups and their sensitization on nanocrystalline TiO_2 electrode*, Synth. Met. 137 (2003), pp. 1543–1544.

[37] K. Sayama, S. Tsukagoshi, T. Mori, K. Hara, Y. Ohga, A. Shinpou, Y. Abe, S. Suga, and H. Arakawa, *Efficient sensitization of nanocrystalline TiO_2 films with cyanine and merocyanine organic dyes*, Sol. Energy Mater. Sol. Cells 80 (2003), pp. 47–71.

[38] M. Guo, P. Diao, Y. Ren, F. Meng, H. Tian, and S. Cai, *Photoelectrochemical studies of nanocrystalline TiO_2 co-sensitized by novel cyanine dyes*, Sol. Energy Mater. Sol. Cells 88 (2005), pp. 23–35.

[39] X. Chen, J. Guo, X. Peng, M. Guo, Y. Xu, L. Shi, C. Liang, L. Wang, Y. Gao, S. Sun, and S. Cai, *Novel cyanine dyes with different methine chains as sensitizers for nanocrystalline solar cell*, J. Photochem. Photobiol. A: Chem. 171 (2005), pp. 231–236.

[40] Q.-H. Yao, L. Shan, F.-Y. Li, D.-D. Yin, and C.-H. Huang, *An expanded conjugation photosensitizer with two different adsorbing groups for solar cells*, New J. Chem. 27 (2003), pp. 1277–1283.

[41] S. Kim, J.K. Lee, S.O. Kang, J. Ko, J.-H. Yum, S. Fantacci, F.D. Angelis, D.D. Censo, M.K. Nazeeruddin, and M. Grätzel, *Molecular engineering of organic sensitizers for solar cell applications*, J. Am. Chem. Soc 128 (2006), pp. 16701–16707.

[42] H. Choi, J.K. Lee, K.H. Song, K. Song, S.O. Kang, and J. Ko, *Synthesis of new julolidine dyes having bithiophene derivatives for solar cell*, Tetrahedron 63 (2007), pp. 1553–1559.

[43] H. Choi, J.K. Lee, K.H. Song, K. Song, S.O. Kang, and J. Ko, *Novel organic dyes containing bis-dimethylfluorenyl amino benzo[b]thiophene for highly efficient dye-sensitized solar cell*, Tetrahedron 63 (2007), pp. 3115–3121.

[44] D. Kim, J.K. Lee, S.O. Kang, and J. Ko, *Molecular engineering of organic dyes containing N-aryl carbazole moiety for solar cell*, Tetrahedron 63 (2007), pp. 1913–1922.

[45] S. Kim, H. Choi, C. Baik, K. Song, S.O. Kang, and J. Ko, *Synthesis of conjugated organic dyes containing alkyl substituted thiophene for solar cell*, Tetrahedron 63 (2007), pp. 11436–11443.

[46] *HYPERCHEM*, Hypercube, Inc., Gainesville, FL, 2000.

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

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

[49] R.W. Kennard and L.A. Stone, *Computer aided design of experiments*, Technometrics 11 (1969), pp. 137–148.

[50] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[51] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, and K.A. Prebble, *Artificial neural networks in classification of NIR spectral data: Design of the training set*, Chemom. Intell. Lab. Syst. 33 (1996), pp. 35–46.

[52] J. Devillers, *Genetic Algorithms in Molecular Modeling*, Academic Press, London, 1996.

[53] R. Leardi, *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks (Data Handling in Science and Technology)*, Vol. 23, Elsevier, Amsterdam, 2003.

[54] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.

[55] J. Xu, B. Chen, and H. Liang, *Accurate prediction of θ(lower critical solution temperature) in polymer solutions based on 3D descriptors and artificial neural networks*, Macromol. Theory Simul. 17 (2008), pp. 109–120.

[56] M.D. Wessel and P.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.

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

[58] P.A. Jansson, *Neural networks: An overview*, Anal. Chem. 63 (1991), pp. 357A–362A.

[59] L. Xu, J.W. Ball, S.L. Dixon, and P.C. Jurs, *Quantitative structure–activity relationships for toxicity of phenols using regression analysis and computational neural networks*, Environ. Sci. Chem. 13 (1994), pp. 941–951.

[60] Y.-H. Qi, Q.-Y. 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.

[61] T.M. Krygowski, M. Cyrański, A. Ciesielski, B. Świrska, and P. Leszczyński, *Separation of the energetic and geometric contributions to aromaticity. 2. Analysis of the aromatic character of benzene rings in their various topological environments in the benzenoid hydrocarbons. Crystal and molecular structure of coronene*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 1135–1141.

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

[63] 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.

[64] 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.

[65] R. Todeschini, M. Lasagni, and E. Marengo, *New molecular descriptors for 2D and 3D structures. Theory*, J. Chemom. 8 (1994), pp. 263–272.

[66] R. Todeschini, P. Gramatica, R. Provenzani, and E. Marengo, *Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons*, Chemom. Intell. Lab. Syst. 27 (1995), pp. 221–229.

[67] V. Consonni, R. Todeschini, and M. Pavan, *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 682–692.

[68] V. Consonni, R. Todeschini, M. Pavan, and P. Gramatica, *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 693–705.

[69] F. Zheng, E. Bayram, S.P. Sumithran, J.T. Ayers, C.-G. Zhan, J.D. Schmitt, L.P. Dwoskin, and P.A. Crooks, *QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release*, Bioorg. Med. Chem. 14 (2006), pp. 3017–3037.

[70] R. Guha and P.C. Jurs, *Interpreting computational neural network QSAR models: A measure of descriptor importance*, J. Chem. Inf. Model. 45 (2005), pp. 800–806.