Bull. Mater. Sci., Vol. 36, No. 7, December 2013, pp. 1307–1313. © Indian Academy of Sciences.

Quantitative structure-property relationships of electroluminescent materials: Artificial neural networks and support vector machines to predict electroluminescence of organic molecules

ALANA FERNANDES GOLIN and RICARDO STEFANI*

Laboratório de Estudos de Materiais (LEMAT), Instituto de Ciências Exatas e da Terra, Av. Governador Jaime Campos 6390, Campus Universitário do Araguaia, Universidade Federal de Mato Grosso, 78600-00 Barra do Garças – MT. Brazil

MS received 13 February 2012; revised 13 December 2012

Abstract. **Electroluminescent compounds are extensively used as materials for application in OLED. In order to understand the chemical features related to electroluminescence of such compounds, QSPR study based on neural network model and support vector machine was developed on a series of organic compounds commonly used in OLED development. Radial-basis function-SVM model was able to predict the electroluminescence with good accuracy ($R = 0{\cdot}90$). Moreover, RMSE of support vector machine model is approximately half of RMSE observed for artificial neural networks model, which is significant from the point of view of model precision, as the dataset is very small. Thus, support vector machine is a good method to build QSPR models to predict the electroluminescence of materials when applied to small datasets. It was observed that descriptors related to chemical bonding and electronic structure are highly correlated with electroluminescence properties. The obtained results can help in understating the structural features related to the electroluminescence, and supporting the development of new electroluminescent materials.**

Keywords. **QSPR; neural networks; SVM; electroluminescence; OLED; organic materials.**

1. Introduction

Electroluminescent materials (EL) are among the most promising modern materials with a wide range of technology applications (Xue and Luo 2003; So *et al* 2009). One of the most promising EL applications, is the design and fabrication of organic light-emitting diodes (OLEDs) (Akcelrud 2003). OLEDs have demonstrated manufacturing and market potential in small and medium device applications. Thus, OLED can become one of the mainstream display technologies, competing directly with LCD (liquid crystal display) technology (Wen *et al* 2005). For high-quality OLED displays, highly efficient and low-cost electroluminescent materials are of great importance, since, to gain market share over LCD displays, OLED devices need to be efficient and to have low prices to the final costumer. Many pyran-containing, polyaromatic hydrocarbons (PAH) and porphyrin type compounds are used in OLED fabrication and these compounds may be polymeric itself or used as a dopant to allow thinfilms to become electroluminescent (Mi *et al* 2002). Understanding of the physical and chemical features related to the electroluminescence of such materials, can help in the design and development of new chemical compounds with improved electroluminescence features. In order to develop new organic compounds that can be used in OLED applications, computational methods, such as quantitative–structure properties relationships (QSPR) have emerged as a fast and reliable method to predict and study physical–chemical properties of materials.

Quantitative–structure properties relationships (QSPR) models can be used to predict with good accuracy, key physical and chemical features from chemical compounds. QSPR methods are based on the existing correlation between groups of mathematical values (descriptors), representing certain features of a chemical structure and a target chemical property. The advantage of QSPR model is that it is based solely on the knowledge of chemical structure and it requires no additional experimental data and once the correlation is established, it can be used for the prediction of properties of new compounds that have not been prepared (Yu *et al* 2008). Thus, QSPR models can be used to assist material design, since one can predict the properties of a certain material before its synthesis. As the development of new materials involves extensive experimental work, the ability to predict the properties of materials is of great value, because, it provides a guide to the development process and speeds up the development cycle, allowing time and reagent saving (Yu *et al* 2008). Thus, many research groups have been developing QSPR models in order to assist material discovery and design (Morril and Byrd 2008; Taherpour 2009; Fourches *et al* 2010; Yu 2010). The advantage of using QSPR models over traditional computational methods is that description calculation is quite easy and requires little computation time.

*Author for correspondence (rstefani@ufmt.br)

A quantitative structure-properties relationships (QSPR) study has been developed with the use of artificial neural networks (ANN) and support vector machines (SVM) in order to understand the chemical features related to the electroluminescence of a series of commonly used red dopants and non-dopants used in OLED development (Chen 2004). The performance of support vector machine models was compared to the performance of artificial neural networks.

2. Methodology

2.1 *Dataset*

The electroluminescence data for 18 compounds (figure 1) were selected from the literature (Chen 2004) and its electroluminescence data (λ_{maxEL}) is summarized in table 1. The free software ChemSketch 12.0 was used to draw 2-D representation of the molecules. As many of the computed descriptors depend explicitly on 3-D structures of the molecules, 3-D structures of the selected compounds were calculated using E-Corina (Tetko *et al* 2005).

2.2 *Descriptor calculation and selection*

In QSPR development, descriptor calculation and selection is a critical step towards a good predictive model. After 3-D structure optimization, a total of 1668 physical–chemical descriptors were generated with E-DRAGON (Tetko *et al* 2005). In order to select the most relevant descriptors to electroluminescence, constant and near-constant descriptors were eliminated, as such descriptors are not significant and do not affect any final model. A variance–covariance matrix was calculated for all descriptors and one of two descriptors with pair wise correlation greater than 0·9 ($R > 0{\cdot}9$) was eliminated. Thus, 1201 descriptors remained in the dataset. The final selection was made with a genetic algorithm feature selection performed with WEKA 3.4 Software (Hall *et al* 2009). The following procedure was applied to feature selection: –log of the maximum wavelength of electroluminescence value (λ_{maxEL}) (table 1) of each compound was calculated and set as the dependent variable, while the remaining descriptors were set as independent variables. Then, the genetic algorithm parameters were adjusted to allow a cross-over probability of 0·8 and the maximum generations and population to 100. After run, 25 statistically significant descriptors (F-value $> 5{\cdot}000$ and $p < 0{\cdot}05$) remained in the final selection (table 2). These descriptors were further used to develop artificial neural networks and support vector machines QSPR models.

2.3 *Artificial neural network model development*

To predict electroluminescence, a multi-layer perceptron (MLP) feed-forward neural network with back propagation of errors was developed. In a multi-layer perceptron (MLP) network, the individual processing units are known as perceptrons and they are usually arranged into three layers: input, hidden and output (Han *et al* 2005). The number of neurons in the input and output layers is determined by the number of independent, in this case the chemical features, and by the number of dependent variables, in this study, the electroluminescence. The software WEKA (Hall *et al* 2009) was used to develop artificial neural networks models that could predict electroluminescence with good accuracy. It generated five artificial neural networks models, using feed-forward with back propagation of errors and radial-basis function kernels. The performance of each generated model was evaluated with standard statistical procedures and cross-validation. All artificial neural networks models were validated with 10-fold leave-one-out cross-validation (LOO–cv). Leave-one-out cross-validation is a validation method that uses a single observation from an original sample as a validation data and the remaining observations as a training data (Jun *et al* 2011). In a 10-fold leave-one-out cross-validation, this is repeated until 10 observations in the samples are used as validation data. It is difficult to divide small datasets into train and test sets, therefore, cross-validation is an accurate way to validate any QSPR model developed with small datasets. All cross-validation data were used to calculate the correlation coefficient, the mean absolute error, error mean and the standard deviation of errors for each generated model.

2.4 *Support vector machine model development*

Support vector machine (Smola and Schölkopf 2004) is a new method of machine learning that has been successfully applied in chemistry and materials science. The main advantage of support vector machine is that it has a few tunable parameters, such as the capacity constant, the control to the complexity of functions and the type of kernel function used for transforming the original input space into a high dimensional feature space (Xu *et al* 2011).

As stated before, one of the aims of the current study is to compare artificial neural networks performance with the performance of support vector machines. For this reason, QSPR models based on support vector machines were developed using the same software (WEKA) that was used for the development of QSPR artificial neural networks models. In this study, several support vector machine models with radial-basis function and sigmoid kernel were developed and the kernel parameters C, ν and γ were adjusted in each experiment, while the parameter ε was fixed to 0·01. The parameter ε is important because it can tune the generalization capacity of each model. All models were cross-validated with the same method used to cross-validate artificial neural networks models.

3. Results and discussion

In table 2, all significant descriptors to electroluminescence are listed. It is important to note that all selected features are highly related with chemical bonding, molecular volume

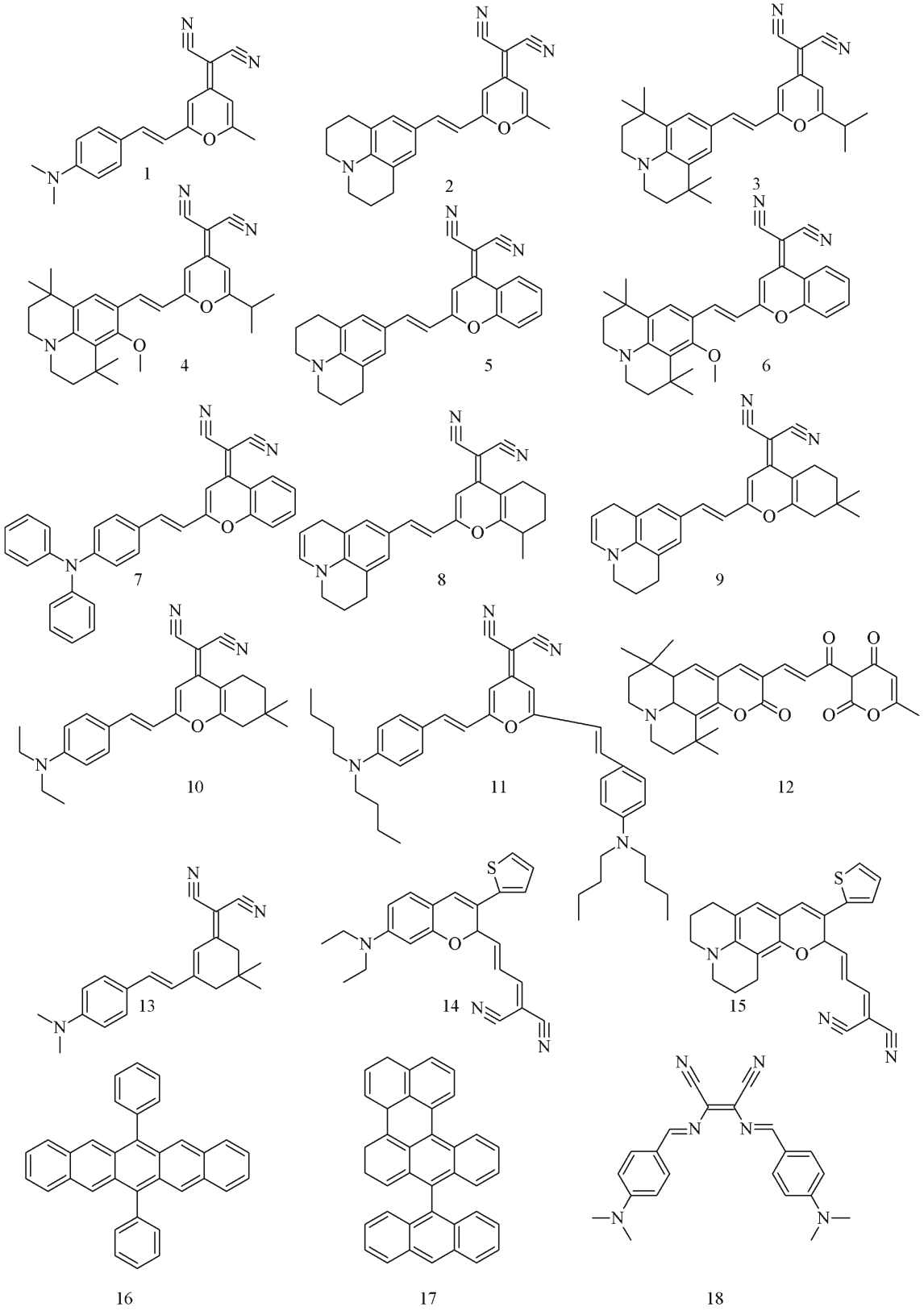

Figure 1. Chemical structures of 18 electroluminescent compounds selected to ANN and SVM developments.

and electronic structure of the studied compounds. The first three significant descriptors are *R*7*p*+, *R*7*m*+ and *R*8*e*+ (table 2), which are a type of GETAWAY (geometry, topology and atom weights assembly) descriptor weighted, respectively by atomic polarizabilities, molar mass and electronegativity (Todeschini and Consonni 2009). These

descriptors are calculated from the leverage matrix obtained by the centred atomic coordinates and are closely related to molecule geometry and the position of substituents in the molecule. Therefore, they are suitable to describe differences in molecular properties caused by the position and alignment of substituent groups and its electronic features. The second three significant descriptors are BEHm1, BELm2 and BEHm3 (table 2). They are in the group of Burden eigenvalues index (Todeschini and Consonni 2009) descriptors, which are topological indices that are weighted by relative atomic masses. Burden indices describe molecules in a topological way. Thus, the molecule topology has influence on electroluminescence. *P*1*p* and *P*1*m* (table 2) are weighted holistic invariant molecular descriptors (WHIM), weighted respectively by atomic polarizability and relative atomic mass. WHIM descriptors are geometrical descriptors derived from the geometrical distance between the atoms of a molecule. These descriptors are based on statistical indices calculated on the projections of atoms along the principal axes, allowing to capture the 3-D information regarding size, shape, symmetry and atom distributions with respect to invariant reference frames. Thus, it can provide information about the correlation between the three-dimensional structure of a molecule and a certain molecule property. *Ks* (table 2) is also a WHIM descriptor, but it is weighted by atomic electrotopological state (E-STATE). The weight function can provide information on the electronic and topological state of each atom in the molecule (Todeschini and Consonni 2009). The state of each atom is a ratio of π and lone-pair electrons over the count of σ bonds of the considered atom in

Table 1. Experimental EL performance values expressed in nm (λ_{maxEL}) and normalization (−log) to QSPR model development.

Sl. No.	Compound λ_{maxEL} (nm)	−log (λ_{maxEL})
1	645	−2·81291
2	774	−2·80884
3	630	−2·79939
4	624	−2·79518
5	645	−2·81291
6	660	−2·81954
7	670	−2·82607
8	660	−2·81954
9	680	−2·83250
10	670	−2·82607
11	653	−2·81491
12	629	−2·79934
13	650	−2·81291
14	622	−2·79379
15	624	−2·79518
16	625	−2·79588
17	616	−2·78958
18	616	−2·78958

Table 2. Statically significant selected descriptors

Descriptor symbol	Descriptor name	*F*-value	*t*-test	*p*-value
*R*7*p*+	*R* maximal autocorrelation of lag 7/weighted by atomic polarizabilities	6·070884	836·207	0·005001
*R*7*m*+	*R* maximal autocorrelation of lag 7/weighted by atomic masses	5·517095	824·289	0·008043
*R*8*e*+	*R* maximal autocorrelation of lag 8/weighted by atomic Sanderson electronegativities	6·353988	797·765	0·004851
BEHm1	Highest eigenvalue no. 1 of burden matrix/weighted by atomic masses	5·757409	254·122	0·006832
BELm2	Lowest eigenvalue no. 2 of burden matrix/weighted by atomic masses	7·765038	229·706	0·002941
BEHm3	Highest eigenvalue no. 3 of burden matrix/weighted by atomic masses	5·391306	224·527	0·007982
*P*1*p*	1st component shape directional WHIM index/weighted by atomic polarizabilities	5·891804	147·394	0·005626
*P*1*m*	1st component shape directional WHIM index/weighted by atomic masses	6·044614	140·618	0·005087
HOMA	Harmonic oscillator model of aromaticity index	5·861169	105·839	0·005742
Ks	*K* global shape index/weighted by atomic electrotopological states	8·535021	104·418	0·001318
CIC0	Complementary information content index (neighbourhood symmetry of 0-order)	6·776535	88·190	0·003811
GATS4v	Geary autocorrelation of lag 4 weighted by Van der Waals volume	8·318028	46·724	0·001479
GATS7e	Geary autocorrelation of lag 7 weighted by Sanderson electronegativity	6·571426	30·832	0·003639
MAXDP	Maximal electrotopological positive variation	5·770581	21·439	0·008751
HTu	H total index/unweighted	9·599787	20·738	0·000774
RDF095v	Radial distribution function-095/weighted by Van der Waals volume	5·861374	11·577	0·005832
RDF060m	Radial distribution function-060/weighted by atomic masses	5·214013	9·780	0·009907
RDF125v	Radial distribution function-125/weighted by Van der Waals volume	8·109778	7·743	0·001654
RDF125p	Radial distribution function-125/weighted by polarizability	7·235529	7·690	0·002708
RDF135v	Radial distribution function-135/weighted by Van der Waals volume	5·742541	5·580	0·006901
RDF135p	Radial distribution function-125/weighted by polarizability	9·871466	5·148	0·000935
RDF135u	Radial distribution function-35/unweighted	5·613742	3·797	0·009681
RDF145e	Radial distribution function-145/weighted by Sanderson electronegativity	6·450780	3·429	0·005735
RDF145u	Radial distribution function-145/unweighted	6·450780	3·402	0·005735
RDF155u	Radial distribution function-155/unweighted	6·759098	2·814	0·004773

the molecular graph. Therefore, the number of π electrons in a molecule is related to its electroluminescence. This is an expected feature, since intermolecular π-stacking can lead to aggregation in solid state, which can cause changes in electroluminescence (Chen 2004). Thus, the knowledge of the electrotopological structure can help in the design and development of new molecules which are less prone to aggregation in solid state. MAXDP (table 2) is a descriptor that represents the maximal electrotopological positive variation, confirming the great influence of electronic topology in electroluminescence. HOMA (table 2) is the harmonic oscillator model of aromatic. This is a descriptor strictly related to the degree of delocalization of π electrons in an aromatic system and an extension of delocalization. Hence, the degree of π electron delocalization has influence on electroluminescence. CIC0 (table 2) is an information index related to the 0-order neighbourhood of atom and it can be considered a descriptor related to the position and pattern of substitution. τ (table 2) denotes H total index, i.e., the total index of hydrogen atoms in the molecule. Another type of selected descriptors is 2D-autocorrelation descriptors represented by GATS4v and GATS7e (table 2) descriptors. These descriptors are weighted by molar volume and electronegativity, respectively. Thus, molar volume and electronegativity have influence on electroluminescence. The last group of representative descriptors is the radial distribution function (RDF) descriptor type (Gasteiger *et al* 1996). These descriptors are 3-D descriptors that encode the molecular 3D-structure and can be unweighted or weighted by some function. In this work, most RDF descriptors (table 2) are weighted by polarizability, electronegativity and molar volume. This corroborates, along other descriptors, the influence of the electronic structure and state in electroluminescence.

Analysing the performance of each machine learning experiment, the best support vector machine and artificial neural network models have been selected to performance comparison and discussion. In table 3, the performance of support vector machine models is given and the best support vector machine model (3) was that with radial–basis function kernel and the parameters $C = 10{\cdot}000$, $\nu = 0{\cdot}500$ and $\gamma = 0{\cdot}001$. The comparison between the experimental and calculated electroluminescence value is depicted in figure 2, with a correlation coefficient of 0·90167 and mean absolute error (MAE) of 0·00117. In table 4, the performance of artificial neural networks models is given and the best artificial neural networks model (2) was a network with radial-basis function kernel multilayer perceptron architecture with 25 neurons in the input layer, 5 neurons in the hidden layer and one neuron in the output layer. The artificial neural network was optimized with a train error of 0·043521, a selection error of 0·000147 and a test error of 0·263794. The comparison between the experimental and calculated electroluminescence value is depicted in figure 3 with a correlation coefficient of 0·80671.

Based on the obtained results, the two methods: artificial neural networks and support vector machine were compared. The comparison between MAE, RMSE, q^2 and the regression coefficient for the two best artificial neural networks and support vector machine models are listed in table 5. The

Table 3. Comparison between performances of trained SVM models.

Model numbers	C	ν	γ	Kernel	Coefficient	Correlation
1	10·000	0·300	0·001	Sigmoid	0·1	0·80584
2	10·000	0·500	0·001	Sigmoid	0·01	0·22163
3	10·000	0·500	0·001	RBF	–	0·90167
4	5·000	0·500	0·001	RBF	–	0·27272

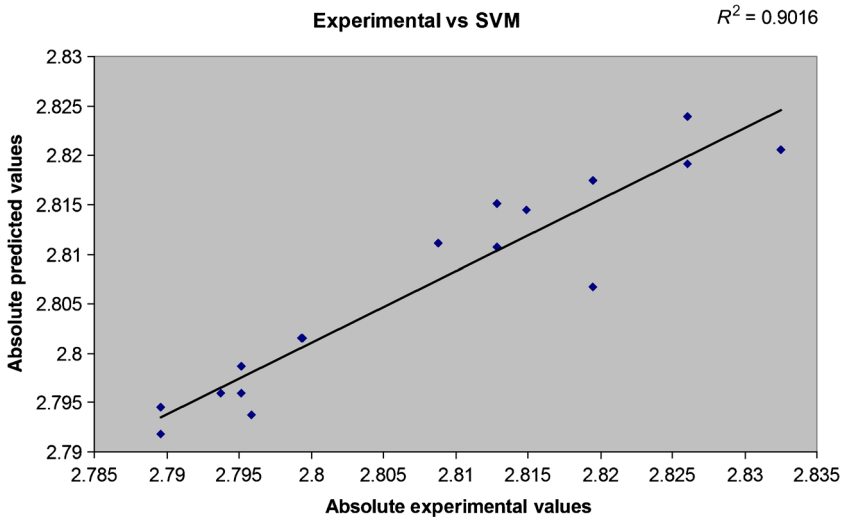

Figure 2. Plot of experimental vs calculated SVM (λ_{maxEL}) values.

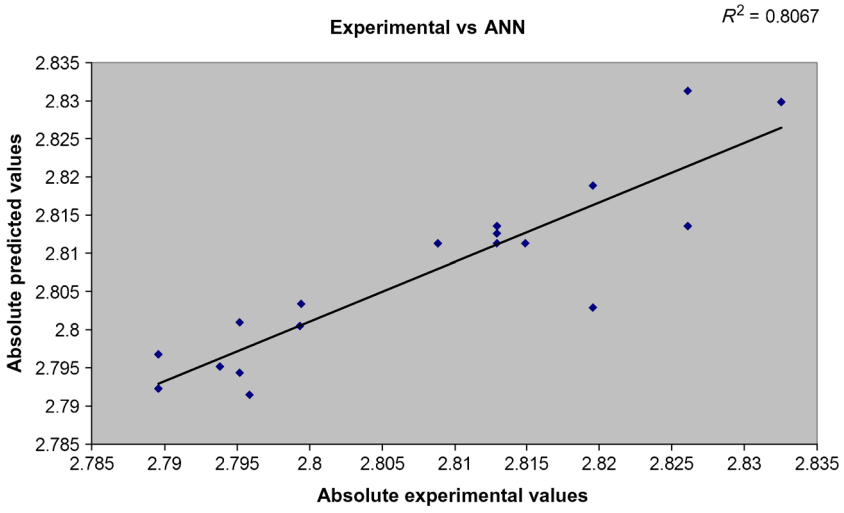

Figure 3. Plot of experimental vs calculated ANN (λ_{maxEL}) values.

Table 4. Comparison between performances of trained ANN models.

Model numbers	Architecture	Train error	Select error	Test error	Kernel	Correlation
1	MLP 25:5:1	0·073037	0·000061	0·375799	Sigmoid	0·75757
2	RBF 25:5:1	0·043521	0·000147	0·263794	RBF	0·80671
3	RBF 25:3:1	0·410573	0·062348	0·776414	RBF	0·70496
4	MLP 25:8:1	0·162391	0·056762	0·468712	Sigmoid	0·72583

Table 5. Comparison of experimental and calculated value of electroluminescence data (λ_{maxEL}) using the best ANN and SVM models and its performance.

Compounds	−log (λ_{maxEL}) (Exp)	SVM RBF kernel (calcd)	ANN RBF kernel (calcd)
1	−2·81291	−2·81076	−2·81361
2	−2·80884	−2·81104	−2·81122
3	−2·79939	−2·80151	−2·80338
4	−2·79518	−2·79872	−2·80091
5	−2·81291	−2·81508	−2·81137
6	−2·81954	−2·80676	−2·80291
7	−2·82607	−2·82391	−2·81355
8	−2·81954	−2·81738	−2·81881
9	−2·83250	−2·82053	−2·82991
10	−2·82607	−2·81915	−2·83122
11	−2·81491	−2·81441	−2·81126
12	−2·79934	−2·80151	−2·80043
13	−2·81291	−2·81074	−2·81256
14	−2·79379	−2·79595	−2·79524
15	−2·79518	−2·79592	−2·79433
16	−2·79588	−2·79372	−2·79152
17	−2·78958	−2·79175	−2·79677
18	−2·78958	−2·79455	−2·79219
Mean absolute error (MAE)	–	0·00117	0·00395
Standard deviation (SD) of error	–	0·00347	0·00577
Root mean square error (RMSE)	–	0·03420	0·06285
q^2		0·85885	0·80311
Correlation	–	0·90169	0·80672

best support vector machine model ($R^2 = 0{\cdot}90167$, MAE = 0·00117, RMSE = 0·03420 and $q^2 = 0{\cdot}85885$) exhibit better performance than the best artificial neural networks model ($R^2 = 0{\cdot}80395$, MAE = 0·00395, RMSE = 0·06285 and $q^2 = 0{\cdot}80311$). As the higher values of R^2 and q^2 indicate higher predictive power of the model (Shahbazikhah *et al* 2011), it is clear that the support vector machine model has higher predictive power than artificial neural networks model. Support vector machine regression is suitable to be applied in QSPR studies with small datasets, since it can interpret the nonlinear relationships between a molecular structure and its properties (Juna *et al* 2010), while artificial neural networks generally need larger datasets to exhibit good performance. Thus, artificial neural networks performance can be improved and can even be similar to the support vector machine performance if large datasets are available.

4. Conclusions

In this study, QSPR models to predict the electroluminescence of OLED materials were developed by using support vector machine and artificial neural networks. Since, QSPR studies is based solely on the chemical structure, the obtained results can help in the understating of the structural features related to electroluminescence, supporting the development of new electroluminescent materials. The statistical results of the radial–basis function-support vector machine model indicate a good predictive model. The difference between the observed RMSE for artificial neural networks and support vector machine models is significant from the point of view of model confidence, as the dataset is small. Thus, when applied to small datasets, Support vector machine is a good method to build QSPR models to predict the electroluminescence of chemical compounds.

Acknowledgements

The authors acknowledge the financial support from CAPES (grant for A F G) and FAPEMAT (Proc. No. 835372/2009).

References

Akcelrud L 2003 *Prog. Polym. Sci.* **28** 875
Chen C T 2004 *Chem. Mater.* **16** 4389
Fourches D, Pu D, Tassa C, Weissleder R, Shaw S Y, Mumper R J, Tropsha A 2010 *ACS Nano* **4** 5703
Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L and Steinhauer V 1996 *J. Chem. Inf. Comput. Sci.* **36** 1030
Hall M, Frank E, Holmes G, Pfahringer G, Reutmann P and Witten I H 2009 *The WEKA data mining Software: An Update. SIGKDD Explorations* **11**
Han I-S, Han C and Chung C-B 2005 *J. Appl. Polym. Sci.* **95** 967
Jun Qi, Wei J, Sun C and Pan T 2011 *Front. Earth. Sci.* **5** 245
Juna Q, Chang-Honga S and Jiac W 2010 *Proc. Env. Sci.* **2** 1429
Mi B X *et al* 2002 *J. Mater. Chem.* **12** 1307
Morril J A and Byrd E F C 2008 *J. Mol. Graph. Model* **27** 349
Shahbazikhah P, Asadollahi-Baboli M, Khaksar R, Alamdaria R F, Zare-Shahabadic V 2011 *J. Braz. Chem. Soc.* **22** 1446
Smola A J and Schölkopf B 2004 *Stat. Comput.* **14** 199
So K H, Park H-T, Shin S C, Lee S G, Lee D H, Oh H Y, Kwon S K, Kim Y-H 2009 *Bull. Korean Chem. Soc.* **30** 1611
Taherpour A 2009 *Chem. Phys. Lett.* **483** 233
Tetko I V *et al* 2005 *J. Comp. Aid. Mol. Des.* **19** 453
Todeschini R and Consonni V 2009 *Molecular descriptors for chemoinformatics* (Wiley-VCH: Weinheim)
Wen S-W, Lee M-T and Chen C H 2005 *J. Disp. Technol.* **1** 90
Xu J, Wang L, Wang L U, Shen X and Xu W 2011 *J. Comp. Chem.* **12** 1
Xue C and Luo F T 2003 *Tetrahedron* **59** 5193
Yu X 2010 *Fiber Polym.* **11** 757
Yu X, Yi B, Yu W and Wang X 2008 *Chem. Papers* **62** 623