DOI: http://dx.doi.org/10.12989/aer.2012.1.4.257

Quantitative structure activity relationship (QSAR) between chlorinated alkene E_{LUMO} and their chlorine

Walter Z. Tang* and Fang Wang

Department of Civil and Environmental Engineering, Florida International University,
10555 West Flagler Street, Miami, FL 33174, USA

(Received June 23, 2012, Revised November 30, 2012, Accepted December 1, 2012)

Abstract. QSAR models for chlorinated alkenes were developed between E_{LUMO} and their chlorine and carbon content. The aim is to provide valid QSAR model which is statistically validated for E_{LUMO} prediction. Different molecular descriptors, N_{Cl}, N_C and E_{HOMO} have been used to take into account relevant information provided by molecular features and physicochemical properties. The best model were selected using Partial Least Square (PLS) and Multiple Linear Regression (MLR) led to models with satisfactory predictive ability for a data set of 15 chlorinated alkene compounds.

Keywords: chlorinated alkene; E_{LUMO}; QSAR; uncertainty

1. Introduction

The chlorine substitution in aliphatic compounds results, by its electron attracting effect, in destabilization in alkanes and stabilization in alkenes. Thus, in alkenes, the stability of the molecule increases with the number of chlorine substitutions. Quantitative Structure Activity Relationship (QSAR) has been used intensively to screen and predict fate and toxicity of chemicals related to the environment. The essential assumption for QSAR studies is that biological, chemical and physical properties of compounds heavily dependent on their structural. Among various properties, energy of lowest unoccupied molecular orbital's (E_{LUMO}) is of critical importance for describing the ability to gain electron from other sources.

The objective of this study is to describe and discuss new externally predictive Multiple Linear Regression (MLR) QSAR models for the prediction of E_{LUMO} of chlorinated alkenes. Models have been developed including N_{Cl} as a molecular descriptor together with other theoretical descriptors, such as N_C and E_{HOMO}, while the other is to test the robustness of obtained model through some statistical methods. Furthermore, decision support has to take into account the uncertainty of the models. Since some of relevant study on QSAR often neglected the uncertainty analysis. A QSAR model without uncertainty test would be confined in practical prediction should not be used to interpret the toxicity behavior.

According to the density functional theory, chemical potential (μ), chemical hardness (η) are defined as

*Corresponding author, Professor, E-mail: tangz@fiu.edu

$$\chi = -\mu = -\left(\frac{\partial E}{\partial N}\right)_{\upsilon(\vec{r})} \tag{1}$$

$$\eta = \frac{1}{2}\left(\partial_2 E / \partial N_2\right)_v = \frac{1}{2}\left(\partial\mu / \partial N\right)_v \tag{2}$$

where E is the total energy of the system, N is the number of electrons in the system and $\upsilon(r)$ is the external potential. μ is the negative of the electronegativity (χ) as defined by Iczkowski and Margrave (1961). E versus N plots is not straight lines but is generally convex upward. Their curvatures define another property of substantial importance, the hardness (η).

The vertical ionization potential (I) of a system is the change of energy when an electron is removed from the system and variation of the energy when an electron is added to the system, known as electron affinity (A). In both electron withdrawing and electron addition, the external potential has to be kept fixed. Hence, using finite difference approximation, Eqs. (1) and (2) become

$$\mu = -\frac{(I + A)}{2} \tag{3}$$

$$\eta = \frac{I - A}{2} \tag{4}$$

However, to obtain the *IP* and *EA* and hence μ and η, one needs three energy values (E_N, E_{N+1} and $E_{N\text{-}1}$). To save computational time, chemical potential and chemical hardness have been calculated by using Koopmans' theorem as

$$\mu = \frac{E_{LUMO} + E_{HOMO}}{2} \tag{5}$$

$$\eta = \frac{E_{LUMO} - E_{HOMO}}{2} \tag{6}$$

where E_{LUMO} is the lowest unoccupied molecular orbital's energy and E_{HOMO} is the highest occupied molecular orbital's energy.

2. Dataset and methods

2.1 Data set

Data were collected for fifteen derivations of chlorinated alkene. Fig. 1 shows the structure of alkene compounds. QSAR model can be modeled due to their structural similarity which expected fall into similar mode of toxic action. From this set, two subsets were constructed by taking at random 12 compounds as the training sample and the remaining 3 compounds as the prediction sample. This proportion amounts for 80% of the compounds in the training set (Table 1). E_{HOMO} and E_{LUMO} values were calculated using SPARTAN Spartan Pro (Wavefunction Inc.). *Ab initio* calculations based on the orbital approximation were used and referred as Hartree-Fock method. Hartree-Fock molecular orbital models are generally satisfactory for a wide variety of thermo-

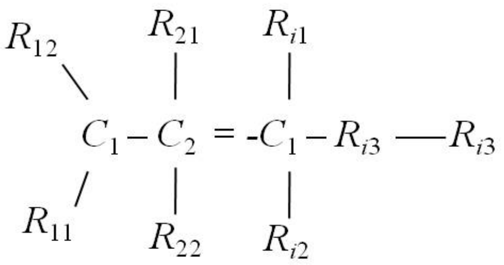

Fig. 1 Overview of structures of chlorinated alkenes used in this study

chemical and kinetic comparisons. This method provides better results for the larger number of atomic orbitals employed. Hartree-Fock models are particularly attractive for the structure determinations of medium-size organic molecules such as chlorinated alkenes studied in this paper. These chlorinated alkene congeners differ in the number of chlorine and carbon on the carbon chain. Regression models are developed for the alkene congeners with calculated E_{LUMO}, the chemical potential (μ) and the chemical hardness (η), number of chlorine (N_{Cl}), number of carbon (N_C) and E_{HOMO} were taken as independent variables.

Those descriptor pairs with high collinear relationships ($r \geq 0.9$) were eliminated. Multiple

Table 1 Molecular property of 15 alkene congeners

No.	Compounds	N_{Cl}	N_C	E_{HOMO}	E_{LUMO}	μ	η
1	Tetrachloroethylene	4	2	-0.3634	0.0975	-0.133	0.230451
2	Cis-1,2-dichloroethylene	2	2	-0.3633	0.1358	-0.11374	0.249547
3	1,1,2,3,3-Pentachloropropene	5	3	-0.3711	0.0771	-0.14702	0.22412
4	1,2-dichloroethylene	2	2	-0.3625	0.1312	-0.11564	0.246871
5	2-chloropropene	1	3	-0.3568	0.1616	-0.09761	0.259201
6	1,1-dichloropropene	2	3	-0.3523	0.1429	-0.10469	0.247636
7	Hexachlorocyclohexene	6	6	-0.3800	0.0768	-0.15165	0.228404
8	Chloroethylene	1	2	-0.3687	0.1562	-0.10624	0.262442
9	1,1-dichloroethylene	2	2	-0.3705	0.1337	-0.11841	0.2521
10	3,4-dichloro-1-butene	2	4	-0.3864	0.1462	-0.12009	0.266343
11	Hexachlorocyclopentadiene	6	**5**	-0.3388	0.0317	-0.15356	0.185223
12	Trichloroethylene	3	2	-0.3636	0.1142	-0.12469	0.238862
13	Trans-1,2-dichloroethylene	2	3	-0.3625	0.1313	-0.11562	0.246875
14	1,3-dichloropropene	2	3	-0.3691	0.1318	-0.11862	0.250462
15	Tetrachlorocyclopropene	4	3	-0.3780	0.1072	-0.13541	0.242608

linear regression analysis and variable selection were performed by the SAS software using the Partial Least Square regression (PLS) and Principle Component Regression (PCR). The acceptable linear models were subjected to a Cross-Validation (CV) analysis by a Leave-One-Out (LOO) and *K*-fold procedures to ensure that the models were not overfitted or underfitted. Many authors considered that a high r^2_{cv} (for example, $r^2_{cv} > 0.5$) an indicator that the model is highly predictive (Golbraikh and Tropsha 2002). To define the relative importance of the QSAR models, we have also performed several other tests beyond the r^2_{cv}. These include the residual standard deviation and leverage, these two distance measures were then used to decide if an object was within the domain of application (Öberg 2005).

2.2 Cross-validation technique

Cross-validation is a popular technique used to test the predictive ability of statistical models. Assuming that a training data set consisting of *n* available compounds, for each data set, an input-output model is developed based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data. In particular, the leave-One-Out (LOO) and the *K*-fold procedures were utilized in this study. Leave-One-Out Cross-Validation (LOO-CV) removes a single observation from the training data at a time and derives the prediction rule based on the remaining *n*-1 objects. This is done in turn till each object has been removed once, the maximum number of models produced by the LOO procedure is equal to the number of available examples *n*. While the *K*-fold version of cross-validation randomly allocates the training set into *k* (≤ *n*) groups, predicting in turn the observations in each group from the *k*-1 remaining groups. If $k = n$, the *K*-fold version and LOO-CV are identical. R^2_{cv} (cross-validation correlation coefficient) is the standard index to measure the accuracy of a modeling method using cross-validation technique, an additional measure to define the accuracy of the proposed QSARs is the and RMS_{CV} (Root Mean Square of Errors of cross-validation) that summarizes the overall error of the model. The formulae used to calculate R^2_{cv} and RMS_{CV} are presented below

$$R^2_{cv} = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i-1}^{n} \left(y_i - \tilde{y}_{i,cv}\right)}{\sum_{i=1}^{n} \left(y_i - \bar{y}_{tr}\right)^2} \tag{7}$$

$$\text{RMS}_{cv} = \left\{\sum \left[y_i - \tilde{y}_{i,cv}\right] / n_{cv}\right\}^{1/2} \tag{8}$$

where *N* is the number of training patterns, and y_i, $y_{i,cv}$ and $\bar{y}_{tr}$ are the measured, predicted, and tested values of the *i*-th compound, respectively. The mean of resulting R^2_{cv} and RMS_{CV} were used to assess the generalization ability of the underlying models.

2.3 Uncertainty estimation

Monte Carlo simulation has many advantages over conventional methods in the estimation of uncertainty, especially that of complex measurement system' outputs. However, this method is regarded as a black box due to none detailed exposition of the method and choosing the inputs distributions is the hardest part. This paper examines and analyses the conventional uncertainty estimation method for linear systems. The conventional uncertainty estimation demands the estimation of the separate effect of each input quantity on the final result through a sensitivity

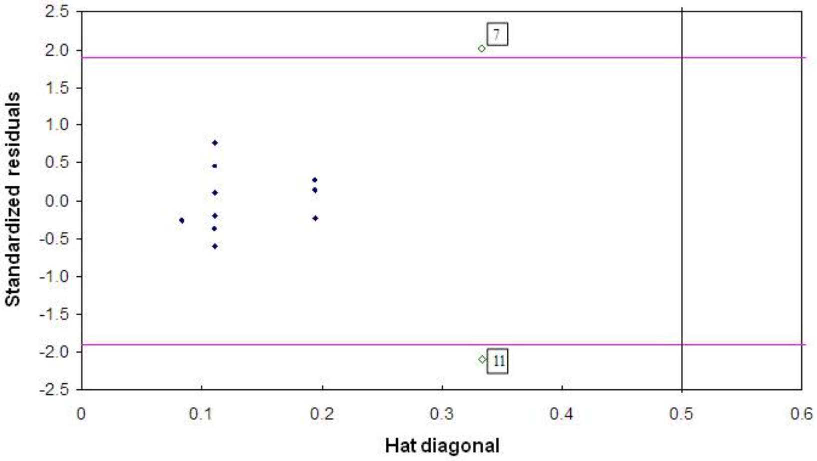

Fig. 2 Outlier detection of Model 1 for alkene

analysis. When input variables in a complex measurement system are closely correlated, sensitivity and uncertainty analysis become extremely difficult and even more unreliable.

The functional expression between the measure and Y and the input quantities X_i in a general linear model is given by: $y = f(x1, x2, x3\ldots, x_N)$. The function f includes not only corrections for systematic effects but also accounts for sources of variability. The standard uncertainty in y can be obtained by Taylor method

$$u(y) = \sigma(y) = \sqrt{\sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma(x_i)^2 + 2\sum \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\sigma(x_i x_j)} \tag{9}$$

This equation gives the uncertainty as a standard deviation irrespective of whether or not the measurement of x_i is independent and of the probability distribution. By the Guide to the Expression of Uncertainty in Measurement (GUM), the standard deviation $\sigma(x_i)$ is referred to as the standard uncertainties associated with the input estimates x_i.

The partial derivatives $((\partial f/\partial x_i)^2)$ are called sensitivity coefficients, which give the effects of each input quantity on the output. For a level of confidence of approximately 95% for normal distributed measurement, y is between $y \pm 2\sigma(y)$.

Table 2 ANOVA for a linear model describing the contribution of variable to E_{LUMO}

Source	*DF*	*SS*	*MS*	*F*	*Pr > F*
Model	1	0.00807	0.00807	169.19684	< 0.0001
Error	8	0.00038	0.00005		
Corrected total	9	0.00845			
Factors					
Intercept	1	0.00199	0.00199	90.6411	< 0.0001
N_{Cl}	1	0.00608	0.00608	259.1608	< 0.0001

Table 3 Goodness of fit statistics

Observations	10	MSE	0.00002
Sum of weights	10	RMSE	0.00485
R^2	0.97006	MAPE	2.69905
Adjusted R^2	0.96631	Press RMSE	0.00557

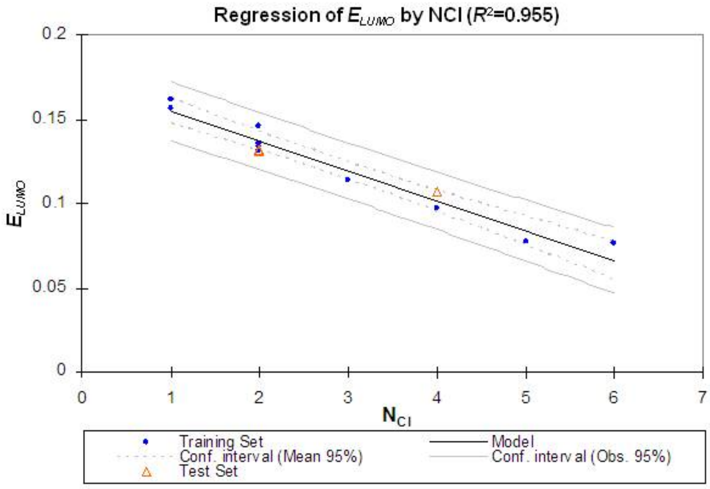

Fig. 3(a) N_{Cl} as descriptors of E_{LUMO} for the alkene

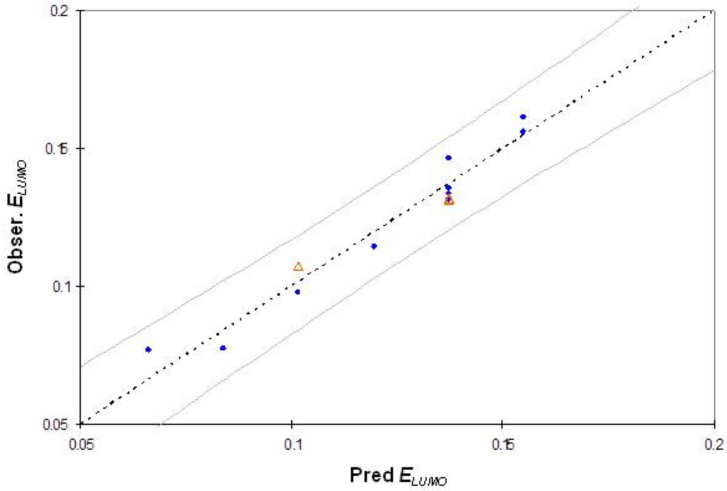

Fig. 3(b) Relationship between observed and predicted alkenes data. Points falling off the dominant straight line structure are outliers. The error bars represent 95% confidence interval. The linear regression model is as follows

3. Results and discussion

The applicability of various statistical and machine learning approaches for creating QSAR models was explored including MLR, PLS and PCR. The data were fit by either multiple linear regression (MLR) or partial least squares regression (PLS). For a theoretical background on MLR and PLS, see References (Martens and Naes 1989).

3.1 Evaluation of molecular descriptors

We will first examine a 1+1 predictor fit: intercept and 1 predictor. In Model 1, the single response variable is E_{LUMO}, and potential predictor variable is N_{Cl}. $E_{LUMO}=a_1* N_{Cl}+k_1$.

3.1.1 Outlier detection

The Williams plot, or leverage plot, is a plot of standardized residuals versus leverages (hat diagonals). In this plot the horizontal and vertical straight lines indicate the normal limits of outlier and influential chemicals values, respectively. Prediction for compounds with high leverage values must be considered unreliable. According to extrapolation method, we consider as reliable only the predictions of the compounds, whose leverages lie within the domain of applicability. Fig. 2 shows that the majority of compounds of the training set are inside of this square area. However, two compounds (7 and 11) have leverages greater than $h^*(h^*=0.5)$, and these two compounds response outliers can be identified in the training set. By removing these two outliers, R^2 value is improved to 0.9701 from 0.9281. Those two points will be discarded in the following analysis.

3.1.2 Linear regression

The final models were then developed on a training set of 12 compounds and validated on 3 external compounds. With linear regression method, a regression equation consisting of coefficients is produced. The linear regression results from an ANOVA for the alkene are summarized in Table 2.

For chlorinated alkene compounds, Fig. 3(a) demonstrated that the descriptor number of chlorine is negative interrelated to the E_{LUMO}, E_{LUMO} decrease as the number of chlorines increase. As E_{LUMO} decreases, the ability of a compound to undergo reduction increases; therefore, an increase in chlorines increases the reactivity of the molecule. E_{LUMO} represents 97.01% of the variance in the linear regression equation, the slope of the QSAR model is -0.02056, and the intercept is 0.1789. The correlation indicates that as the number of chlorine increase from 1 to 6 in an alkene compound, that E_{LUMO} will decrease 0.1299. The probability of the alkene getting a correlation of -0.9849 for a sample size of 12 is less than 0.01%. Fig. 3(b) shows a plot of observed E_{LUMO} against fitted values for Model 1, the correlation of the scatter is -0.9849. The observed over fitted plot should be routinely reported for any (QSAR) regression. The great advantage is that it can always be made, no matter how many predictor variables are included.

$$\begin{aligned} &E_{LUMO}=-0.02056N_{cl}+0.1790 \\ &N=10, R^2=0.9701, F=169.197, RMSE=0.00485, P=0.0000 \end{aligned} \tag{9}$$

We will examine three two-predictor models. The first 1+2 predictor model that combines N_{Cl}, and N_C with E_{LUMO} is: $E_{LUMO}= a_2*N_{Cl}+b_2*N_C+k_2$. For Model 2, leverage value can be calculated for both training compounds and new compounds. Fig. 4(a) shows that the majority of compounds of the training set are inside this area. It is important to note that chemical: hexacholorocyclopentadiene (compound 11) can be identified as the outlier with the standardized residual value greater than the cutoff value in the training set.

After removing this outlier and refitting the model, the regression coefficients are plotted in Fig. 4(b) shows that number of chlorine contributes negatively to E_{LUMO}, and number of carbon

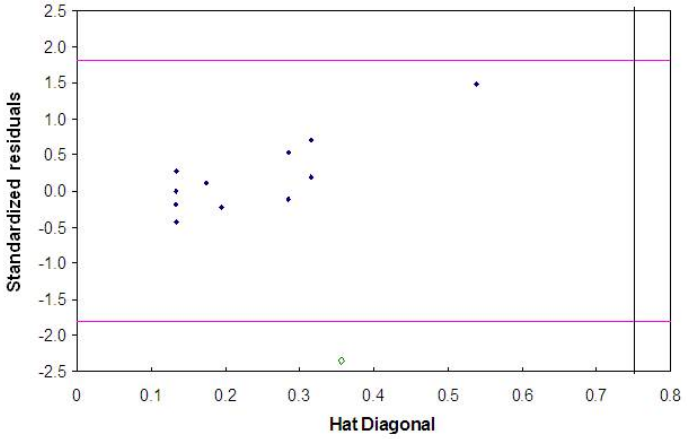

Fig. 4(a) Outlier detection of Model 2 for alkenes

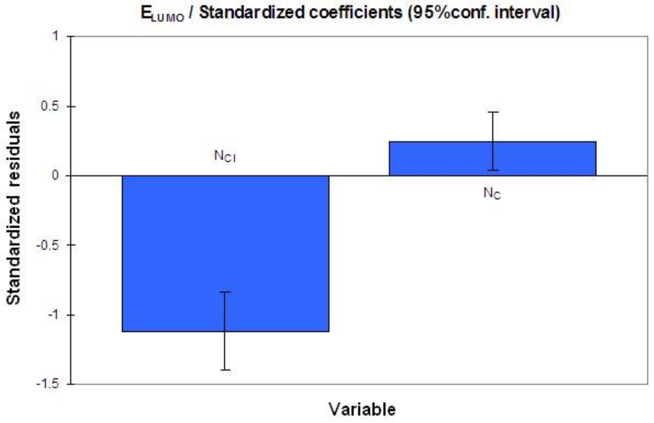

Fig. 4(b) Regression coefficients of scaled and centered variables for Model 2

contributes positively to E_{LUMO}. Eq. (5) indicates that introducing N_C into QSAR model, there is better correlation coefficient that using the one-predictor. The relationship between E_{LUMO} and two descriptors (N_{Cl} and N_C) is shown as follows

$$E_{LUMO} = -0.0205\mathrm{N}_{\mathrm{Cl}} + 0.00592\mathrm{N}_{\mathrm{C}} + 0.1641$$
$$N = 11, R^2 = 0.9956, F = 102.869, RMSE = 0.00188, P = 0.0000 \quad (10)$$

The next two-predictor models take μ and η into as response vector: $\mu = a_3*N_{Cl}+ b_3*N_C+k_3$ and $\eta = a_4*N_{Cl}+ b_4*N_C+k_4$.

The Williams plot (Fig. 5(a) and 6(a)) verified the presence of outlier (with standardized residuals greater than 2 standard deviation units), and this responding outlier can be identified in the training set of Model 3 and Model 4: 1,1-dichloropropene (compound 6) and hexachlorocyclo-pentadiene (compound 11), respectively.

After removing the outlier and refitting of the model, the regression coefficients are plotted in Figs. 5(b) and 6(b) show that number of chlorine has the negative effect to E_{LUMO}, and number of

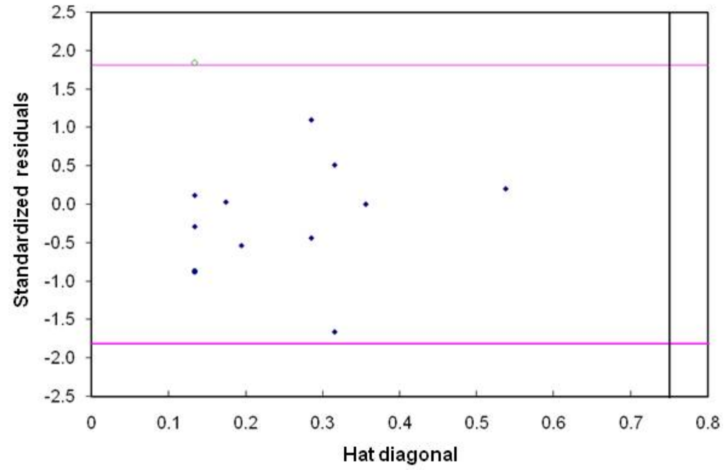

Fig. 5(a) Outlier detection of Model 3 for alkenes

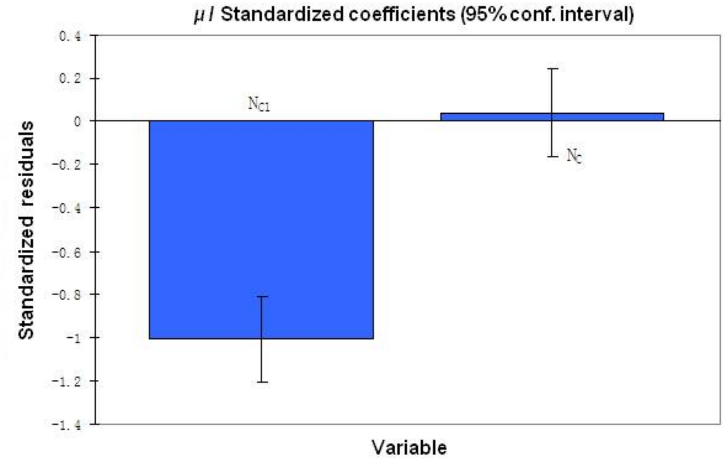

Fig. 5(b) Regression coefficients of scaled and centered variables for Model 3

carbon has the positive effect to E_{LUMO}. PLS model for the relationship between μ, η and two descriptors (N_{Cl} and N_C) is shown as follows

$$\mu = -0.0101 * N_{Cl} + 0.00053 * N_C - 0.0961$$
$$N = 11, R^2 = 0.9268, F = 50.661, RMSE = 0.0043, P = 0.0000 \tag{11}$$

$$\eta = -0.00995 * N_{Cl} + 0.00494 * N_C + 0.2592$$
$$N = 11, R^2 = 0.9268, F = 50.661, RMSE = 0.0043, P = 0.0000 \tag{12}$$

3.2 Development of QSAR model

Activities of molecules in the biological systems are highly influenced by their inherent electronic properties. Hence, highest occupied molecular orbital (E_{HUMO}) energy along with N_{Cl}

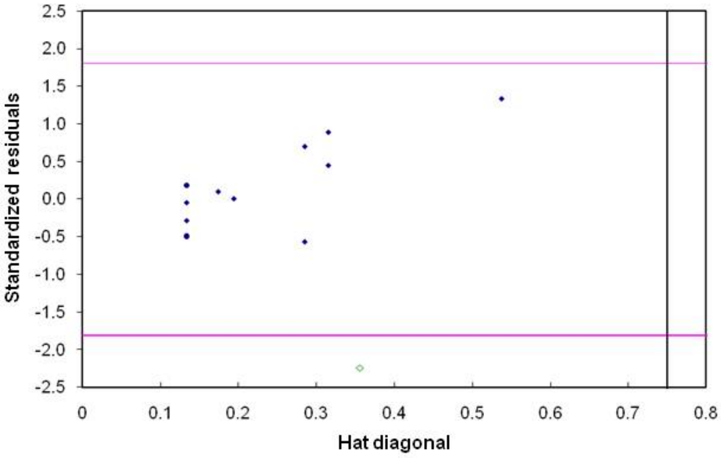

Fig. 6(a) Outlier detection of Model 4 for alkenes

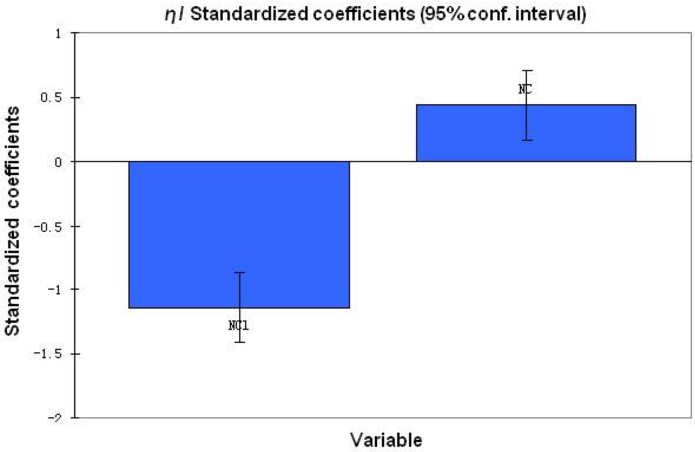

Fig. 6(b) Regression coefficients of scaled and centered variables for Model 4

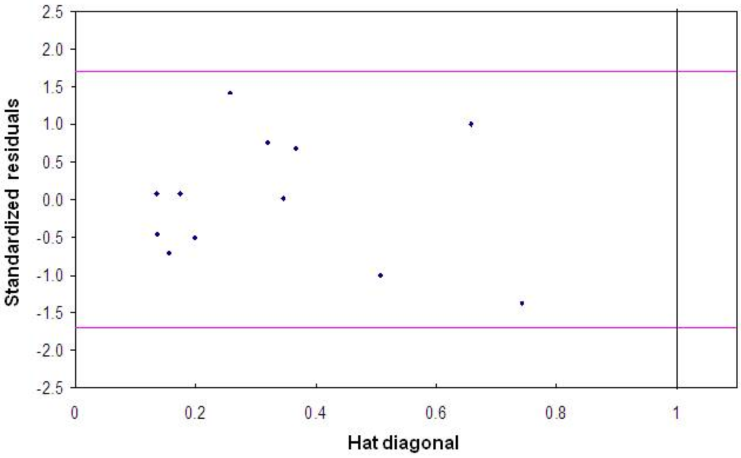

Fig. 7 Outlier detection of Model 5 for alkene

and N_C were selected as molecular descriptor to predict E_{LUMO}. Using all three descriptors in the QSAR model, for the combined set of 12 alkane congeners, maximum values for the coefficient of correlation and lowest root mean square error were obtained. By 3 predictors, $E_{LUMO} = a_5*N_{Cl}+b_5*N_C +c_5*E_{HOMO}+ k_5$. Fig. 7 shows that all of compounds of the training set are inside this area, and all compound are within the cut-off value limit, which implies that they are not be consider outliers but influential chemicals. The presence of outliers and chemicals very structurally influential in determining model parameters was verified by the Williams plot.

With straightforward MLR, a regression equation consisting of coefficients is produced. These coefficients are intuitively simple and are summarized in Tables 4-6, Table 4 presents the correlation matrix, where it is clear that the three selected descriptors are not highly correlated, while the pair of descriptors, (N_{Cl} and N_C) is correlated each with each other. In order to examine the importance of each descriptor and answer the question which of the independent variables has a great effect on the dependent variable in the multiple regression analysis, the standardized regression coefficients were also calculated. Results from an ANOVA for the alkene are summarized in Table 5. The contribution from these three factors to E_{LUMO} can thus be described by a simple linear model, accounting for 97.09% of the variance.

Table 4 Correlation matrix for the three selected descriptors

	E_{HOMO}	N_{Cl}	N_C
E_{HOMO}	**1.000**		
N_{Cl}	0.096	**1.000**	
N_C	-0.066	0.671	**1.000**

Table 5 Analysis of variance

Source	*DF*	*SS*	*MS*	*F*	*Pr > F*
Model	3	0.0163	0.0054	88.9681	< 0.0001
Error	8	0.0005	0.0001		
Corrected total	11	0.0168			
Factors					
Intercept	1	0.0005	0.0005	11.573	0.643
E_{HOMO}	1	0.00137	0.00137	22.408	0.001
N_{Cl}	1	0.01483	0.01483	243.031	< 0.0001
N_C	1	0.00009	0.00009	1.465	0.261

Table 6 Goodness of fit statistics

Observations	12	MSE	0.00006
Sum of weights	12	RMSE	0.00781
DF	8	MAPE	6.45885
R^2	0.97090	Press RMSE	0.01942
Adjusted R^2	0.95999		

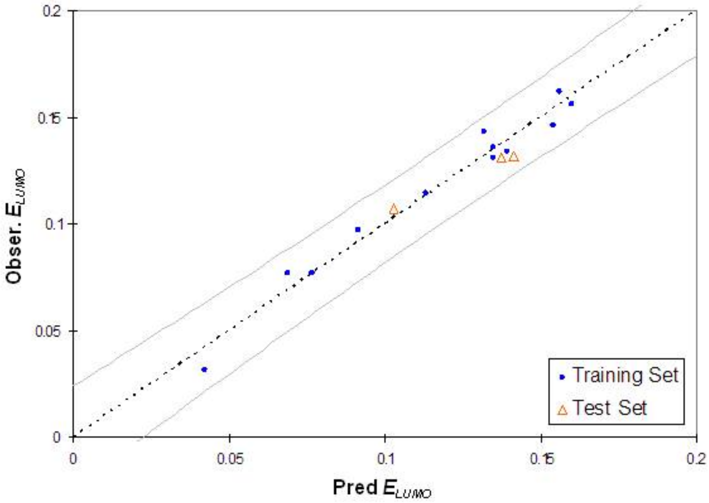

Fig. 8(a) Relationship between observed and predicted endpoint data

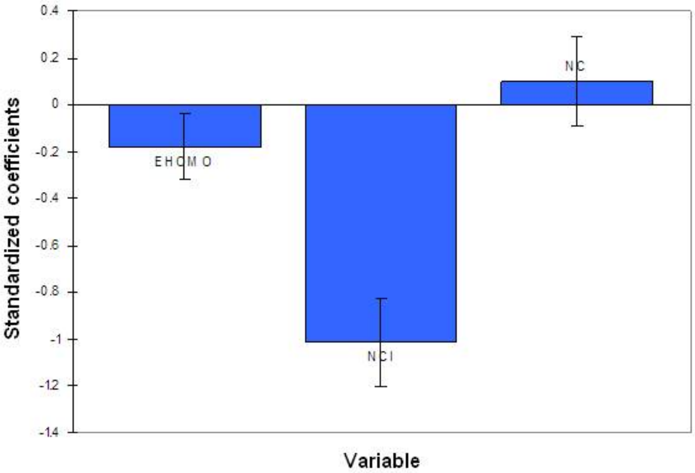

Fig. 8(b) Regression coefficients of scaled and centered variables. The full regression equation is Model 8. The error bars represent 95% confidence interval

Fig. 8(a) starts with repeating observed versus fitted and residuals over fitted for the full model. The regression coefficients are plotted in Fig. 8(b). Since molecular descriptors do not have equal variance, their relative importance in the model is measured better by standardized regression coefficients. The errors of the regression coefficients have also been reported for each equation.

The other interpretation is very interesting compared to other MLR alternatives; PLS regression is possible to calculate the applicability domain of a QSAR model, that is, the range within which it "tolerates" a new molecule. This specification can be made regarding both the *X*- and the *Y*- data as long as not all initial variance is used in the model.

Fig. 9(a) indicates that all *X*-variables such as E_{HOMO}, number of carbons (N_{Cl}), and number of chlorine (N_C) load strongly in the model and closely related. Overall, N_{Cl} and E_{HOMO} are the most important *X*-variables. The data analysis resulted in a QSAR with $R^2X = 0.545$, $R^2Y = 0.846$, and

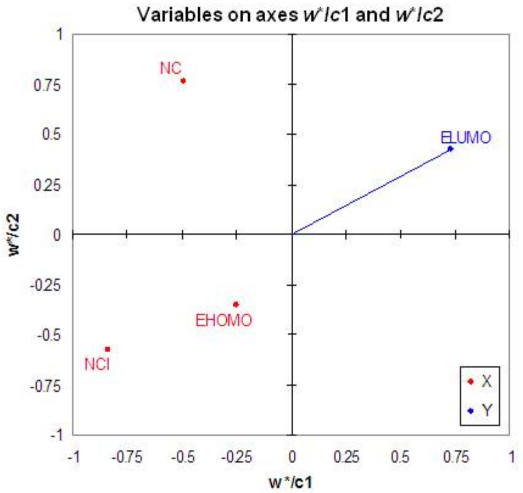

Fig. 9(a) PLS loading plot of the data set showing the relationships between the three *X*- and the one *Y*-variable at the same time

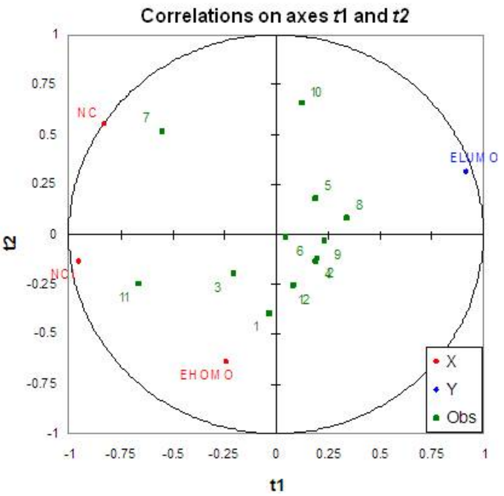

Fig. 9(b) Distribution of the 12 compounds in the latent variable space, defined by the scores of the first two latent variables

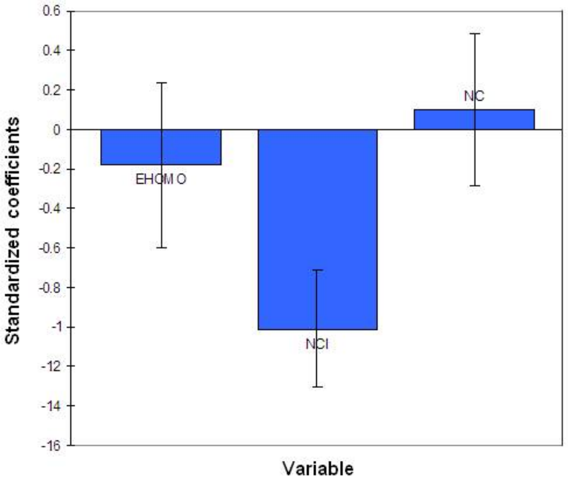

Fig. 9(c) PLS coefficients plot for Model 5. These show that number of chlorine is more important than E_{HOMO}

$Q^2Y = 0.700$, which are excellent performance statistics considering that four responses are handled simultaneously.

Fig. 9(b) shows the model scores. There are no outliers in the score space because all compounds lie inside the elliptic 95% tolerance volume depicted in the plot. We also plot coefficients using PLS to simplify comparison with MLR (Fig. 9(c)) since the sizes and signs of the coefficients (β_{PLS}) predict the relative importance of the variables and are basically needed for revealing and interpreting of new samples. Altogether, hexachloro-cyclohexene (Figs. 9(a) and (b), point 7) is the highest toxic compound because it contains the high number of chlorine and carbon.

Principal Component Regression (PCR) is a reduced regression that uses derived inputs, based on principal components, of less than or equal dimension than the original inputs. Fig. 10(a) shows the combined plot of scores and loadings in the space defined by the first two principle components (PC_1: NC_l, PC_2: E_{HOMO}) of the studied chemicals which are represented by the response variable (E_{LUMO}). The explained variance of these two components is 89.67% of the total information (PC_1 explained variance = 55.72%). The loading plot (the lines in the figure) reveals the relevance of each variable in each of the first two principle components. All the variables are oriented in the same direction along the most informative principal component PC_1, which is evidence of their satisfactory correlation and is consistent with the results of previous pair-wise correlation analyses. It shows that compound 11 is significant positive relative to N_{Cl}, and compound 7 was strongly correlated with N_C. Fig. 10(b) is the regression coefficient plot.

Table 7 compares the quality of MLR, PLS and PCR model as represented by r^2 and RMSE. It shows that the number of variables is same for PLS and MLR, but the former model shows a lower RMSE (0.00638 verses 0.00781) compared to the latter one. Similar improvements can be seen for the Models 1-4.

Regression models for the training set of 12 chlorinated alkene congeners with calculated E_{LUMO} values taken as dependent variables and all possible combinations of the three descriptors, such as number of chlorine (N_{Cl}), number of carbon (N_C), and the highest occupied molecular orbital energy (E_{LUMO}) as independent variables are presented in Table 9. Table 9 summarizes the PLS models for the endpoint E_{LUMO}, μ and η that have the lowest RMSE values. According to the rule that to count as a good model when r^2 should be greater than 0.8, all models in Table 9 are significant and that most models can be considered good models.

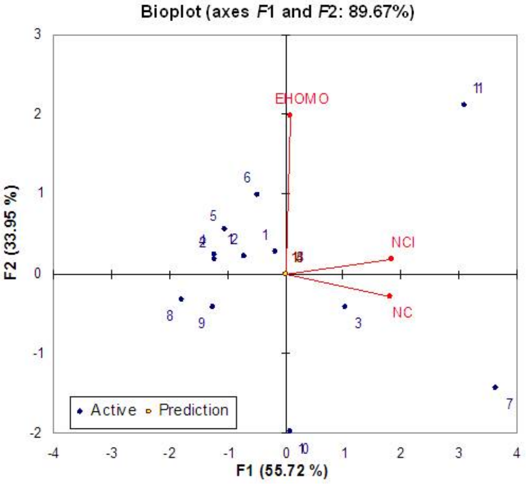

Fig. 10(a) Bioplot of F1 (55.72%) vs. F2 (33.95%)

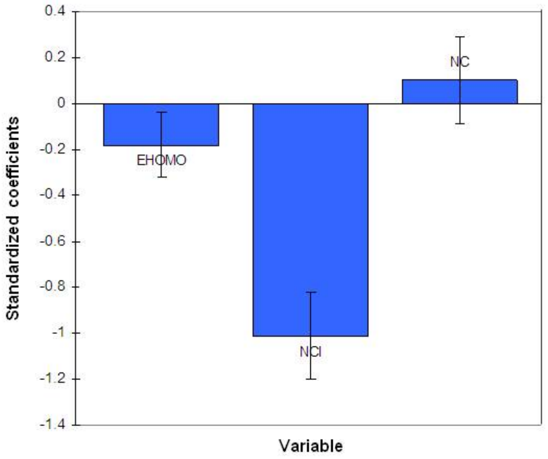

Fig. 10(b) Coefficient plot for Model 5

Table 7 Comparison for chlorinated alkenes Model 4 using three different calibration methods

Model calibration methods	Model 4	R^2	RMSE
MLR	E_{LUMO} = -0.02187 N_{Cl}+0.002896 N_C-0.5701 E_{LUMO}-0.03399	0.9709	0.00781
PLS	E_{LUMO} = -0.02187 N_{Cl}+0.002896 N_C-0.5701 E_{HOMO}-0.03399	0.9709	0.00638
PCR	E_{LUMO} = -0.02187 N_C+0.002896 N_C+0.5701 E_{HOMO}-0.03399	0.9709	0.00781

Table 8 Experimental and calculated values of E_{LUMO} for the model training (1-12) and test (13-15) set

No.	Compounds	Descriptors			E_{LUMO}[a] values		
		E_{HOMO}	N_{Cl}	N_C	Calculated	Predicted	Residual[b]
1	2-chlorobutane	0.09745	4	2	0.09745	0.09154	0.00591
2	1-chlorohexane	0.13581	2	2	0.13581	0.13518	0.00063
3	2-chloro-2-methyl-butane	0.07710	5	3	0.07710	0.07696	0.00014
4	1-chloroheptane	0.13123	2	2	0.13123	0.13474	-0.00350
5	Chloromethane	0.16159	1	3	0.16159	0.15625	0.00534
6	2-chloropropane	0.14295	2	3	0.14295	0.13183	0.01111
7	1-chloropentane	0.07676	6	6	0.07676	0.06886	0.00790
8	1-chlorooctane	0.15621	1	2	0.15621	0.16012	-0.00392
9	1-chlorodecane	0.13369	2	2	0.13369	0.13930	-0.00560
10	2-chlorohexane	0.14625	2	4	0.14625	0.15417	-0.00792
11	3-chlorohexane	0.03166	6	5	0.03166	0.04244	-0.01078
12	Chloroethane	0.11417	3	2	0.11417	0.11347	0.00070
13	1,2-dichloropropane	0.13125	2	3	0.13125	0.13763	-0.00638
14	1,4-dichlorobutane	0.13184	2	3	0.13184	0.14139	-0.00955
15	1,5-dichloropentane	0.10720	4	3	0.10720	0.10275	0.00445

[a] With N_{Cl}, N_C, E_{HOMO} as descriptors

[b] The residual is the difference between calculated and predicted E_{LUMO} values

Table 9 Summary of the models for chlorinated alkenes

Model No.	Regression equations	N*	R^2	RMSE
1	E_{LUMO} = -0.02056 N_{Cl}+0.1790	10	0.9701	0.00485
2	E_{LUMO} = -0.0205 N_{Cl}+0.00592 N_C+0.1641	11	0.99559	0.00188
3	μ = -0.0101*N_{Cl}+ 0.00053*N_C-0.0961	11	0.96772	0.00373
4	η = -0.00995*N_{Cl}+ 0.00494*N_C+0.2592	11	0.92682	0.00427
5	E_{LUMO} = -0.02187 N_{Cl}+0.002896 N_C-0.5701 E_{HOMO}-0.03399	12	0.9709	0.00638

*None outlier are included

Possible combinations of parameters were considered among other equations using various statistical criteria. The results showed that there was one best equation using N_{Cl}, and N_C as molecular descriptor to predict E_{LUMO}. The negative regression coefficient of this descriptor is due to its influence in decreasing E_{LUMO}. That the dimensional descriptor (N_{Cl}) in the QSAR model was negative in sign, as was expected, while the descriptor, N_C, was positive. This demonstrates that an increase in chemical size leads to a decrease in energy of E_{LUMO}; on the contrary the presence of halogen groups tends to increase E_{LUMO}.

3.3 QSAR model validation

The robustness of the models and their internal predictive ability was evaluated by cross-

Table 10 Results of LOO and *K*-fold cross-validation test for alkene

Model No.	$d = 1$ (LOO-CV)		$k = 2$	
	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}
1	0.004848	0.96874	0.005120	0.95495
2	0.002210	0.99557	0.002170	0.99335
3	0.003679	0.96876	0.002105	0.99130
4	0.004254	0.92809	0.003000	0.96350
5	0.007358	0.97469	0.007250	0.98795

Model No.	$k = 3$		$k = 5$		$k = 10$	
	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}
1	0.005127	0.96410	0.005134	0.96350	0.005091	0.96356
2	0.001963	0.99680	0.002072	0.99628	0.002092	0.99605
3	0.003813	0.97063	0.004146	0.96518	0.003891	0.96770
4	0.003993	0.93550	0.004364	0.92764	0.004410	0.92384
5	0.005430	0.98743	0.006960	0.97832	0.008022	0.97230

validation. A cross-validation method was applied to the original data set for Models 1-5. In particular, the LOO-CV and two-fold CVs with $k = 3$, 5, and 10 were utilized for the evaluation of the QSAR models and compared the statistics results with PLS method, because r^2 and r^2_{cv} values are a good test for evenly distributed data, but they are not always reliable for unevenly distributed data sets; RMSE provides a more reliable indication of the robustness of the model, independently of the applied splitting (Gramatica and Papa 2005). The reported fitting and validation parameters have, as expected, very high values, indicating that the model has very good descriptive and predictive performance. The quality of r^2_{cv} and the small RMSE values confirm the high predictivity of these models. A group of 2, 3, 5, and 10 compounds was randomly selected from the training set. The results of LOO, 2-fold, 3-fold, 5-fold, and 10-fold for Models 1-5 are reported in Table 10. The consistency in the statistics of r^2_{cv} ($r^2_{cv} > 0.5$) and RMS_{CV} reveals the stability and robustness for different model.

Cross validation did confirm the Model 2 as the best QSAR model to predict E_{LUMO} of any compound in the class of Chlorinated alkenes. The cross-validation r^2_{cv} values had a maximum at three-term equation. Cross validation results using N_{Cl} and N_C as the variable were much better those including three parameters N_{Cl}, E_{HOMO}, and N_C. This two-parameter equation was very stable, leading to cross-validated values in the range between 0.9935-0.9968, whereas the three-parameter equations gave cross-validated value between 0.9723-0.9879. This showed that two-parameter equation indicated higher predictive ability, as shown by cross validation. It was again verified that the best single molecular descriptor is N_{Cl} ($r_{cv}{}^2 = 95.5\%$-96.87%). It is important to note that the 2-variables model (Model 2) gives satisfying fitting and prediction performances ($r^2 = 0.9956$, $r^2_{\text{LOO}} = 0.99557$, $r^2_{\text{5-fold}} = 0.9963$) are higher than in the 3-descriptors model, its RMSE values ($\text{RMSE}_{(\text{training, set})} = 0.00188$, $\text{RMSE}_{(\text{cross-val, set})} = 0.002072$) are all lower than in the 3-descriptors model. On the contrary the 3-variables model ($r^2 = 0.9709$, $r^2_{\text{LOO}} = 0.9747$, $r^2_{\text{5-fold}} = 0.9276$) did not significantly increase the predictive performance of the nested model, as is evident on comparing the internal and external r^2_{cv} values. The RMSE value (RMSE$_{\text{(training, set)}}$= 0.00638, RMSE$_{\text{(cross-val, set)}}$= 0.00696) also show that this last model had a worse internal performance and an external predictivity lower than in Model 2.

Table 10 shows that for Model 2 RMS$_{CV}$ based on the LOO-CV is greater than the true RMSE by about 17.55% ((0.00221-0.00638)/0.00638), and RMS$_{CV}$ based on *K*-fold are greater than true RMSE in the range 4.43%-15.43%, the difference between RMS$_{CV}$ and true RMSE reaches its maximum at $d = 1$. For Models 1, 3 and 4, in the case $d = 1$, CV estimates the r^2_{cv} and RMS$_{CV}$ values with satisfactory accuracy. In addition, the RMS$_{CV}$ based on the selected model is very close to the true RMS error in these cases. The difference between them is about 0.0412%-1.37%. In practice, true RMS error usually means that RMSE estimated based on the model since the true RMS error is not known.

Compared with *K*-fold cross-validation with $k = 2$ yields unsatisfactory results, but for all values of $k > 2$ good models are obtained. However, when $k = 10$, the results may not be very reliable due

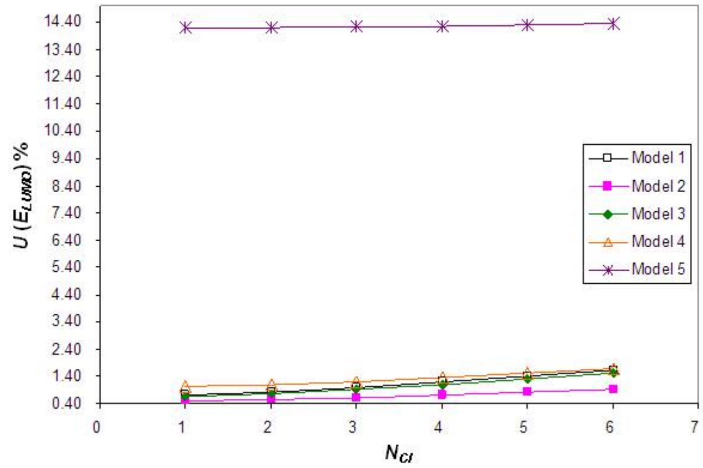

Fig. 11 Relationship between N_{Cl} and uncertainty in E_{LUMO}

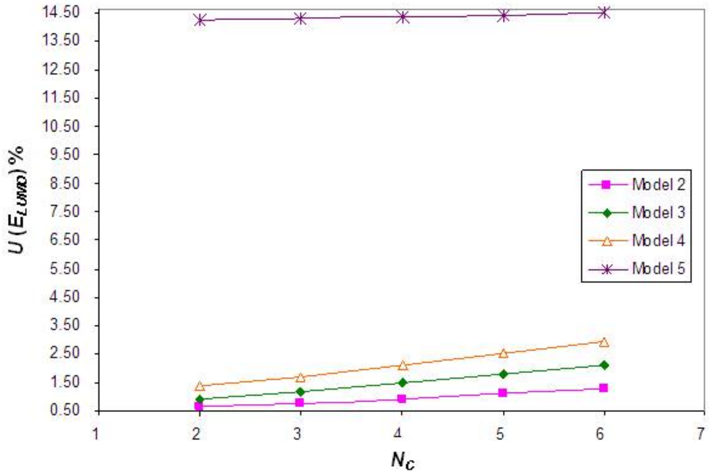

Fig. 12 Relationship between number of carbon and uncertainty in E_{LUMO}

to very few observations per predictor. Consequently, fivefold instead of tenfold CV may be used to reduce the computational cost in predicting experimental data of modeling.

For chlorinated alkenes, *K*-fold cross-validation performs better than LOO-CV and the full model with respect to model size, model complexity and, most importantly, predictive power. *K*-fold is comparatively insensitive to the parameter *k*, since all runs with $k > 2$ resulted in good final models.

3.4 Uncertainty analysis

From the regression equation discussed in the section 3.2, coefficients of parameters and the standard deviations are estimated in Table 9 using bootstrap analysis. The bootstrapping is repeated 5000 times for each validated model and gives the following parameter estimates and their standard errors (Table 11).

The expression for the uncertainty in E_{LUMO} determined from the regression Model 5 at a measured or specified value of *X* is found by Eq. (13). Here we don't considered the correlated uncertainties between any two of these variables, and the uncertainty of number of carbon are zero, then all terms involving correlated uncertainties in Eq. (16) will be simplified as following Eq. (14).

Table 11 Summary of parameter coefficients and the standard deviations for Models 1-5

Model No.	a	b	c	k	σ_a	σ_b	σ_c	σ_k
1	-0.02069	-	-	0.1792	0.00125	-	-	0.00339
2	-0.02048	0.005972	-	0.16384	0.00066	0.00098	-	0.00249
3	-0.01011	0.000525	-	-0.09602	0.00117	0.00167	-	0.00304
4	-0.00982	0.004708	-	0.2593	0.00113	0.00229	-	0.00515
5	-0.02187	0.00290	-0.5701	-0.03399	0.00178	0.00239	0.19364	0.07062

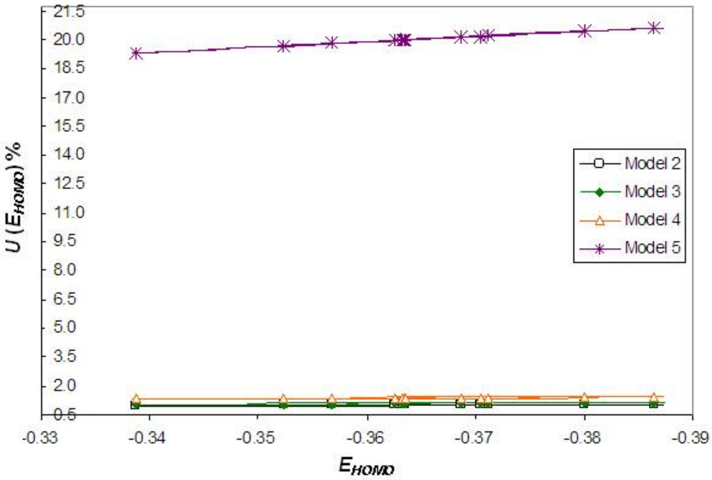

Fig. 13 Relationship between E_{HOMO} and uncertainty in E_{LUMO}

$$
\begin{aligned}
U^{2}_{E_{LUMO}} &= \left(\frac{\partial(E_{LUMO})}{\partial a}\right)^2 U_a^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_{Cl}}\right)^2 U_{N_{Cl}}^2 + \left(\frac{\partial(E_{LUMO})}{\partial b}\right)^2 U_b^2 \\
&+ \left(\frac{\partial(E_{LUMO})}{\partial N_C}\right)^2 U_{N_C}^2 + \left(\frac{\partial(E_{LUMO})}{\partial c}\right)^2 U_C^2 + \left(\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})}\right)^2 U_{E_{HOMO}}^2 + \left(\frac{\partial(E_{LUMO})}{\partial k}\right)^2 U_k^2
\end{aligned}
\qquad (13)
$$

where, $\partial(E_{LUMO})/\partial a = N_{Cl}$, $\partial(E_{LUMO})/\partial N_{Cl} = a$, $\partial(E_{LUMO})/\partial b = N_C$, $\partial(E_{LUMO})/\partial N_C = b$, $\partial(E_{LUMO})/\partial c = E_{HOMO}$, $\partial(E_{LUMO})/\partial(E_{HOMO}) = c$, $\partial(E_{LUMO})/\partial k = 1$.

$$U_{E_{LUMO}}^2 = N_{Cl}{}^2 U_a^2 + N_C{}^2 U_b^2 + E_{HOMO}{}^2 U_c^2 + c^2 U_{E_{HOMO}}^2 + U_k^2 \qquad (14)$$

Fig. 11 shows the change of uncertainty in E_{LUMO} with number of chlorine. It is clearly show that number of chlorine has slightly effect on the uncertainty of E_{LUMO} for Models 1-5. For example, the uncertainty of E_{LUMO} will increase from 0.5152 to 1.7028 when N_{Cl} increase from 1 to 6. Fig. 12 shows the relationship between uncertainty in E_{LUMO} and number of carbon. The uncertainty in E_{LUMO} for Models 2-4 does not change significantly from 0.6338 to 2.9347 with the number of carbon when the N_C is increase from 2 to 6. Similarly, Fig. 13 indicated the impact of E_{HOMO} on uncertainty of E_{LUMO} follows the same pattern as showed in Figs. 11 and 12. E_{HOMO} has obvious effect on the relative error on E_{HOMO}, and the uncertainty of E_{LUMO} for Models 2-4 will be increased from 1.0304 to 1.4077 when randomly distributed E_{HOMO} values decrease from -0.33 to -0.39. In Figs. 12-13, three variables effect to the uncertainty of E_{LUMO} for each model are following: Model 5 > Model 4 > Model 3 > Model 2.

4. Conclusions

The QSAR models are homogenous and stable from Models 1-5, since the cross-validated r^2_{cv} was not affected very much when larger groups of molecules were left out each time. Each cross-validation experiment was repeated and, accordingly, reported as the average r^2_{cv}. In QSAR modeling, decision-making has to take into account the models' uncertainty. Even if the 2-variables model, nested on the previous one, gives satisfying fitting and prediction performances (r^2 = 0.9956, r^2_{LOO} = 0.9956, $r^2_{\text{5-fold}}$ = 0.9963), its RMSE values (RMSE$_{\text{(training set)}}$ = 0.00188, RMSE$_{\text{(cross-val set)}}$= 0.00207) are all higher than in the 3-descriptors Model 2. On the contrary the 3-variables model, obtained by replacing E_{HOMO} with E_{LUMO} (r^2= 0.9709, r^2_{LOO}= 0.9747, $r^2_{\text{5-fold}}$= 0.9783), did not significantly increase the predictive performance of the nested model, as is evident on comparing the internal and external r^2_{cv} values. The RMSE values (RMSE$_{\text{(training set)}}$= 0.00638, RMSE$_{\text{(cross-val set)}}$ = 0.00696) also show that this last model has a worse internal performance and an external predictivity lower than in Model 2.

Acknowledgements

This work has been supported by the National Institute of Environmental Health Sciences.

Contents of this publication do not necessarily reflect the views and the policies of the U.S. NIH, nor does mention of trade names and commercial products constitute their endorsement by the U.S. Government. In addition, financial support from the Dean's Infrastructure Award for the College of Engineering and Computing at Florida International University to WT is greatly appreciated.

References

Baumann, K. von Korff, M. and Albert, H. (2002), "A systematic evaluation of the benefits and hazards of variable selection in latent variable regression, Part II; Practical applications", *J. Chemometrics*, **16**(7), 351-360.

Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2!", *J. Mol. Graph Model*, **20**(4), 269-276.

Gramatica, P. and Papa, E. (2005), "An update of the BCF QSAR model based on theoretical molecular descriptors", *QSAR Sci.*, **24**(8), 953-960.

Iczkowski, R.P. and Margrave, J.L. (1961), "Electronegativity", *J. Am.Chem. Soc.*, **83**(17), 3547-3551.

Kottegoda, N.T. and Rosso, R. (1998), *Statistics, Probability and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York.

Martens, H. and Naes, T. (1989), *Multivariate Calibration*, Wiley, New York.

Öberg, T. (2005), "A QSAR for the hydroxyl radical reaction rate constant: Validation, domain of application, and prediction", *Atmospheric Environment*, **39**(12), 2189-2200.

Sagrado, S. and Cronin, M.T.D. (2006), "Diagnostic tools to determine the quality of "Transparent" regression-based QSARs: The "Modeling Power" plot", *J. Chem. Inf. Model.*, **46**(3), 1523-1532.