5-27-2009

Chlorine Contribution to Quantitative Structure and Activity Relationship Models of Disinfection By-Products' Quantum Chemical Descriptors and Toxicities

Fang Wang
Florida International University, fangwang2006@gmail.com

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

CHLORINE CONTRIBUTION TO QUANTITATIVE STRUCTURE AND ACTIVITY RELATIONSHIP MODELS OF DISINFECTION BY-PRODUCTS' QUANTUM CHEMICAL DESCRIPTORS AND TOXICITIES

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Fang Wang

2010

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Fang Wang, and entitled Chlorine Contribution to Quantitative Structure and Activity Relationship Models of Disinfection By-Products' Quantum Chemical Descriptors and Toxicities, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Hector R. Fuentes

_____
Zhenmin Chen

_____
Fernando R. Miralles-Wilhelm, Co-Major Professor

_____
Walter Z. Tang, Co-Major Professor

Date of Defense: May 27, 2009

The dissertation of Fang Wang is approved.

_____
Dean Amir Mirmiran
College of Engineering and Computing

_____
Interim Dean Kevin O'Shea
University Graduate School

Florida International University, 2010

ACKNOWLEDGMENTS

This study was carried out at the Florida International University, Department of Civil and Environmental Engineering during the years 2006-2010 and I am grateful to the department for providing such good opportunity and excellent working facilities.

Foremost, I would like to express my deepest gratitude to my Ph.D advisor Dr. Walter Z. Tang. He has always been a great resource for ideas and solutions, and his encouragement and support have made difficult time during my research much less frustrating. I am also most grateful to him for helping me to overcome the language barrier as non-native speaker of English. His understanding, patience, and encouragement have been the most important factors in building my confidence and when I took my first faltering steps on the long road to the skills needed of an independent researcher.

I would also like to thank Dr. Fernando R. Miralles-Wilhelm, Dr. Hector R. Fuentes, Dr. Zhenmin Chen for serving in my committee. Their insights and suggestions have been greatly helpful in my research. I have the sincerely thanks to Dr. Miralles for his help to provide the financial support for my Ph.D study. Also thanks Dr. Fuentes for his enthusiasm and endless ideas. Additionally, I also appreciate many discussions with him about many things including, but not limited to, general research principles, the nature of science et al. Last, but by no means least, thanks go to Dr. Chen who mentored me in the finer points of QSAR and statistical analysis. His patient help and guidance was invaluable towards all statistical analysis perspectives in this thesis. Finally, and always, my personal special thanks my family. My parents, Wang Chun-ming and Zhou Sheng-ping, thank you for always support and encourage me in all the endeavors I

undertook. I thank my sister, Wang Yan, for her always being there cheering me up and stand by me. Even though you are far away from me, you are always in my heart.

ABSTRACT OF THE DISSERTAITON

CHLORINE CONTRIBUTION TO QUANTITATIVE STRUCTURE AND ACTIVITY RELATIONSHIP MODELS OF DISINFECTION BY-PRODUCTS' QUANTUM CHEMICAL DESCRIPTORS AND TOXICITIES

by

Fang Wang

Florida International University, 2010

Miami, Florida

Professor Walter Z. Tang, Co-Major Professor

Professor Fernando R. Miralles-Wilhelm, Co-Major Professor

Quantitative Structure-Activity Relationship (QSAR) has been applied extensively in predicting toxicity of Disinfection By-Products (DBPs) in drinking water. Among many toxicological properties, acute and chronic toxicities of DBPs have been widely used in health risk assessment of DBPs. These toxicities are correlated with molecular properties, which are usually correlated with molecular descriptors. The primary goals of this thesis are: 1) to investigate the effects of molecular descriptors (e.g., chlorine number) on molecular properties such as energy of the lowest unoccupied molecular orbital (E_{LUMO}) via QSAR modelling and analysis; 2) to validate the models by using internal and external cross-validation techniques; 3) to quantify the model uncertainties through Taylor and Monte Carlo Simulation. One of the very important ways to predict molecular properties such as E_{LUMO} is using QSAR analysis. In this study, number of chlorine (N_{Cl}) and number of carbon (N_C) as well as energy of the highest occupied molecular orbital (E_{HOMO}) are used as molecular descriptors.

There are typically three approaches used in QSAR model development: 1) Linear or Multi-linear Regression (MLR); 2) Partial Least Squares (PLS); and 3) Principle Component Regression (PCR). In QSAR analysis, a very critical step is model validation after QSAR models are established and before applying them to toxicity prediction. The DBPs to be studied include five chemical classes: chlorinated alkanes, alkenes, and aromatics. In addition, validated QSARs are developed to describe the toxicity of selected groups (i.e., chloro-alkane and aromatic compounds with a nitro- or cyano group) of DBP chemicals to three types of organisms (e.g., Fish, T. pyriformis, and P.pyosphoreum) based on experimental toxicity data from the literature.

The results show that: 1) QSAR models to predict molecular property built by MLR, PLS or PCR can be used either to select valid data points or to eliminate outliers; 2) The Leave-One-Out Cross-Validation procedure by itself is not enough to give a reliable representation of the predictive ability of the QSAR models, however, Leave-Many-Out/K-fold cross-validation and external validation can be applied together to achieve more reliable results; 3) E_{LUMO} are shown to correlate highly with the N_{Cl} for several classes of DBPs; and 4) According to uncertainty analysis using Taylor method, the uncertainty of QSAR models is contributed mostly from N_{Cl} for all DBP classes.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1 INTRODUCTION

1.1 Introduction

During water treatment processes, the disinfection is commonly used to destroy pathogenic organisms and prevent the outbreak of waterborne infectious diseases. Although the benefits of water disinfection are well documented, there is an undesirable side effect of producing various Disinfection By-Products (DBPs) when disinfectants such as chlorine react with natural inorganic and organic matters in the water.

Accurate estimation of toxicological properties of DBPs has been a challenging task for establishing DBP standards of drinking water. To set up standards of DBPs, various toxicological properties such as the acute and chronic toxicity of DBPs have been used in health risk assessments. The major challenge is that more than 500 of DBPs could be present in drinking water disinfected by chlorine. The U.S. EPA has set up the regulation for chemicals with the highest occurrence in drinking water in the Stage 1 DBPR. However, hundreds of other DBPs, using various disinfectants such as chlorine, have been identified. In addition, there are many unidentified DBPs, as evidenced by measurements of total organic halides compared with known halogenated DBPs. Since toxicity tests of DBPs could be very costly and time consuming, Quantitative Structure-Activity Relationship (QSAR) analysis is an economic and efficient way to unveil the relationships between the toxicity of DBPs and their chemical properties. QSAR can be used to predict the toxicity of untested DBPs of known molecular properties. It can also be used to better characterize the potential health effects by setting

priority of toxicity testing of different DBPs in establishing maximum contaminant level goals (MCLGs) for drinking water standards.

Because there are many classes of DBPs present in chlorinated drinking water, QSAR models could be used to predict the toxicities of chemicals from physical and chemical descriptors such as hydrophobilicity properties (i.e., logP), and electronic properties (i.e., E_{LUMO} and E_{HOMO}). QSAR methodology is a cost-effective tool for toxicity prediction for hazard identification, setting of testing priorities, and providing scientific support for decisions. Therefore, in this work, five chemical classes of DBPs are studied, compared within the framework of QSAR along with different comparative statistical modeling methods, model validation and statistically model uncertainty analysis. Please see sections 1.3, 1.4 and 1.5 for objectives, significance of study and the thesis outline in details.

1.2 Quantitative Structure-Activity Relationship (QSAR)

QSAR analysis is a promising tool based on the assumption that the biological activities of new, untested and even non-synthesized chemicals have the correlation with molecular structure, or properties of similar compounds. To develop a QSAR model, three elements are needed: i) biological data for a set of chemicals, ii) descriptors e.g. for physical or chemical properties of the chemicals, and iii) a statistical method to relate the biological activity and the descriptor(s) (Walker et al., 2003). The two main fundamental assumptions of QSAR are: i) the same molecule, under the same conditions, is expected to generate the same toxicological response, and highly similar molecules are expected to generate similar toxicological responses (Mallakin et al., 2005), and ii) differences in

reaction rates for this common rate-limiting step will give rise to observed differences in activity or quantitative potency (Schultz et al., 2003).

Although recognition of the relationship between chemical activity and structure began a long time ago, the use of formal structure-activity relationships started with the pioneering work of Hammett in the 1930s, Taft in the 1950s, Hansch in the 1960s and Tang in 2003. QSAR methodology was developed and has been used most extensively in the areas of drug and pesticide research. In the 1970s, spurred by the burgeoning number of chemicals being released to the environment, QSAR methodology began to be applied to environmental toxicology. The primary focus in the area of environmental toxicology has been bio-concentration and toxic effects on fish and other aquatic life. Some work has been done in relating chemical structure characteristics to toxicity in bacteria of environmental interest.

For risk assessment, Blum and Speece (1990) reported their research that QSAR can reveal the relationship between the toxicity of a compound and its structural descriptors. Moreover, the benefit in the development of property/toxicity data is that they allow estimation of toxicity to an organism based on easily measured or calculated molecular descriptors such as E_{LUMO} or E_{HOMO}. This quick method saves tremendous time and money in determining the toxicity tests of thousands of DBPs. There are also large numbers of relevant examples in QSAR studies depending on quantum chemical descriptors (Baj and David, 1994; Lewis, 1989; Nevalainen and Kolehaminen, 1994; Mekenyan et al., 1994; Xu et al., 1994; Dai, 1998), because quantum chemical descriptors such as E_{LUMO} and E_{HOMO} could provide meaningful insight into toxic mechanisms.

In 2000, QSAR analysis of the toxicity of 14 heterocyclic nitrogen compounds, which are extensively used as intermediates in the manufacturing of pesticides and herbicides, has been reported by Xu et al. (2000). Mallakin et al. (2000) have applied QSAR to model the photoinduced toxicity of anthracene and oxygenated anthracenes. Woo et al. (2002) reported mechanism-based structure–activity relationships analysis in carcinogenic for drinking water DBPs. The QSAR analysis of hypoglycemic agents also has been conducted by using the topological indices (Murcia-Soler et al., 2001). In regards to QSAR model process and model validation, linear QSAR regression models have been studied for the prediction of bio-concentration factors by physicochemical properties, and structural theoretical molecular descriptors are studied by Papa et al. (2007). Eriksson et al. (2000) selected training set in environmental QSAR analysis when compounds are clustered in QSAR analysis. Figure 1.1 illustrates the process of developing a QSAR model which involves several basic steps.

In relation to the general scheme for the development of QSAR model, as illustrated in figure 1.1, building a QSAR model begins by collecting and organizing the property or (biological) activity, called “endpoint,” followed by identification of the chemical group for which the model will be developed. Commonly, the endpoint is determined in accordance with an experimental protocol, and in the case of an endpoint of regulatory interest with a test guideline, which is listed in table 1.1.

If the number of chemicals is sufficiently large, they can be split into a training set and a test set. The training set is to be selected to cover the chemical domain of the model to develop the model, while the test set is used to validate the model. The next step is to eliminate data by selecting the pertinent descriptors from a large set variable that

correlates with the activity of interest using computer software. The descriptors can be physicochemical, electronic or steric (molecular volume, molecular weight) (Walker et al., 2003). Examples of commonly used descriptors and the toxicological characteristic they reflect are shown in table 1.2.

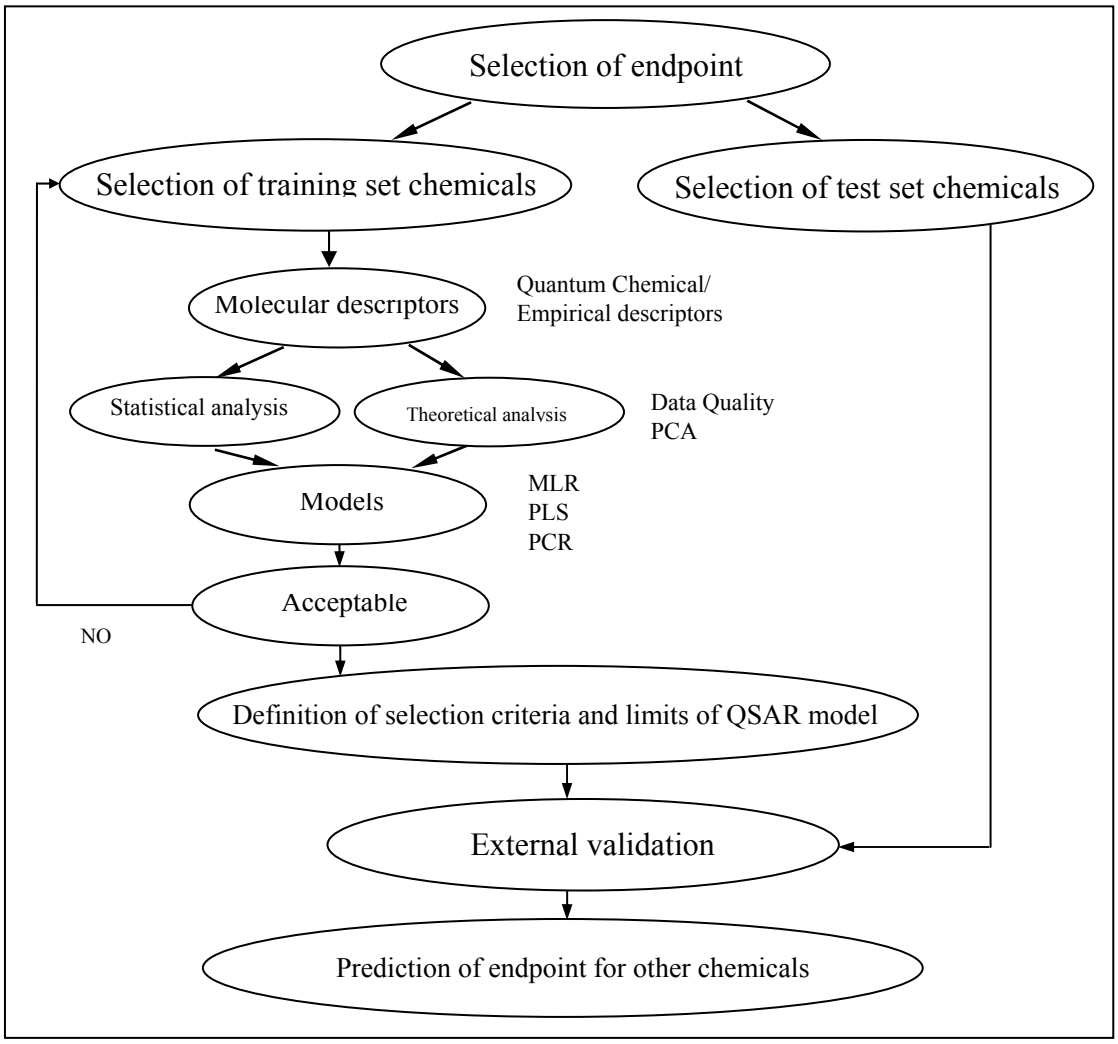


Figure 1.1 Basic scheme for the development of QSAR models

Table 1.1 Endpoints associated with EU and OECD methods (Worth et al., 2005)

Physicochemical Properties:
Melting Point;
Boiling Point;
Vapour Pressure;
K octanol/water partition coefficient;
K_{oc} organic carbon/water partition coefficient;
Water Solubility.
Ecological Effects:
Acute Fish;
Long-term Toxicity;
Acute Daphnid;
Alga;
Terrestrial toxicity.
Human Health Effects:
Acute Oral;
Acute Inhalation;
Skin Irritation;
Eye Irritation;
Skin Sensitization;
Repeated Dose Toxicity;
Genotoxicity (in vitro, bacterial or mammalian cells);
Genotoxicity (in vivo).

Table 1.2 Examples of descriptors and the relevant toxicological characteristics

Calculated descriptor	Relevant toxicological characteristic	Reference(s)
octanol water partition coefficient; $\log P=\log(C_{org}/C_{water})$	hydrophobicity / lipophilicity	Zvinavashe et al., 2008
energy of the highest occupied molecular orbital; E_{HOMO}	ionization potential, ease of oxidation, nucleophilic reactivity	Benigni et al., 2000
energy of the lowest unoccupied molecular orbital; E_{LUMO}	oxidation potential, ease of reduction, electrophilic reactivity	Zhang et al., 2007 Cronin et al., 2001
molecular weight, molecular volume, molecular surface area	size and polarizability of a molecule fragment	Sixt et al., 1995
dipole moment	charge separation in a molecule	Wang et al., 2004

C_{org} =concentration of the non-ionised solute in the organic phase;
C_{water} =concentration of the non-ionised solute in the water phase.

The most commonly used physicochemical descriptor is the octanol-water partition coefficient, logP, which reflects the ability of organic compounds to passively partition and accumulate in organisms. The importance of hydrophobicity in explaining the toxicity for a large set of 133 PCB congeners was shown by Padmanabhan et al. (2006). On the other hand, some quantum parameters often used in QSAR studies are the energies of frontier orbitals such as E_{LUMO} and E_{HOMO}, which determine the nucleophilic and electrophilic reactivities of a compound, respectively.

The energies of the frontier orbitals e.g. the lowest unoccupied molecular orbital (E_{LUMO}) and the highest occupied molecular orbital (E_{HOMO}) determine the electrophilic or nucleophiic reactivity of a compound towards its toxicological receptor (Fleming, 1976). Reactivity between an electrophile and a nucleophile increases when i) the E_{HOMO} is increased or ii) the E_{LUMO} is decreased (Fleming, 1976). Given that the toxicological receptor is constant for a series of chemicals to be modeled by a QSAR, the relative reactivity and thus toxicity of a series of chemicals may be modeled by looking at their relevant frontier orbital without the requirement for knowledge on the orbital characteristics of the toxicological receptor (Zvinavashe, 2008).

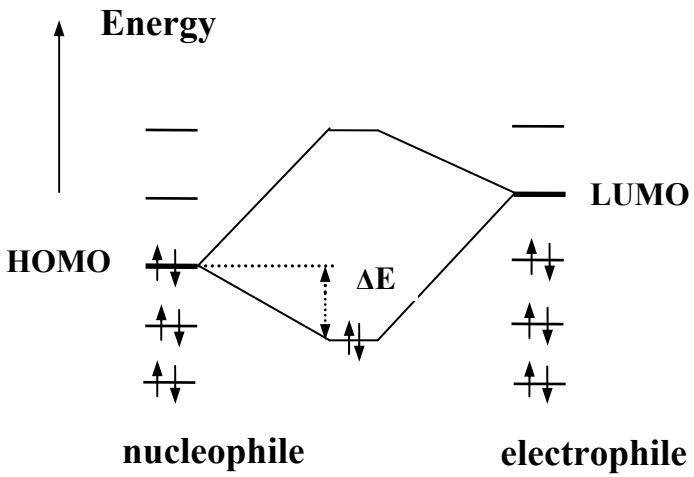


Figure 1.2 molecular orbital diagram for the reaction between an electrophile and nucleophile of a reactive toxic intermediate and its toxicological receptor (Soffers et al., 2001)

Traditional QSARs use experimentally derived descriptors such as logP (P, distribution coefficient), cavity surface area (CSA) and Hammetts constant (σ), among others to quantify physicochemical characteristics. However, due to the non-sufficient large data sets of experimentally derived parameters, QSARs have been developed based on descriptors derived from quantum mechanical computation because they are not restricted to closely related compounds and can be easily obtained. Also, they can explain the clearly mechanistic meaning of toxicology in QSAR studies by Sixt et al. (1995), Schmitt et al. (2000), Cronin et al. (2002), Hatch and Colvin (1997).

The correlation between the chosen descriptor(s) and the endpoint is often analyzed with statistical software. There are many statistical techniques appropriate to the development of QSAR for acute toxicity (Livingstone et al., 1995). These techniques include linear (e.g. regression based) and non-linear methods (Cronin and Schultz, 2001).

The most commonly used correlative method is regression analysis due to its simplicity. Three techniques, namely: 1) Multiple Linear Regression (MLR), 2) Partial Least Squares (PLS), and 3) Principle Component Regression (PCR), are used in these aspects. To assess quality, it is important that different modeling techniques are compared so that their strengths and weaknesses may be evaluated (Cronin et al., 2002).

In the next step, the reliability or quality of the developed QSAR model can be estimated by comparing the outcomes the model predicts to the experimentally determined endpoint values in the training set. If the predictions are poor, one can restart the model developed by using different descriptors. If the predictions are good, one can define the selection criteria and the limits of QSAR models and then make the model fit other chemical classes based on the same selection criteria.

Currently, QSAR models have been used in regulatory assessment of chemical safety in many countries for many years; however, few systemic studies have been completed in the development of QSAR models in DBPs area and there were no universal principles for their regulatory applicability. In 2004, member countries of the Organization for Economic Cooperation and Development (OECD) agreed on the principles for developing and validating QSAR models for their use in regulatory assessment of chemical safety (OECD, 2004). In 2007, the OECD published a "Guidance Document on the Validation of (Q)SAR Models," which provided detailed criteria in five categories: i) a defined endpoint, ii) an unambiguous algorithm, iii) a defined domain of applicability, iv)appropriate measures of good-of-fit, robustness, and predictivity, and v) a mechanistic, with the aim of providing guidance on how specific QSAR models can be evaluated with respect to the OECD principles.

1.3 Objectives

The purpose of this work is to use a multivariate regression method to develop QSAR models for DBP compounds. The research problem has two main facets:

1. Develop Quantitative Structure-PROPERTY Relationship (E_{LUMO} vs. N_{Cl}, E_{HOMO}, and N_C) models for various DBP classes.

The specific aims are:

(1) To select appropriate molecular properties followed by identification representative DBP chemical classes;

(2) To determine the surrogate molecular descriptors (such as N_{Cl}, E_{HOMO}, and N_C) to estimate the molecular properties;

(3) To model the relationship between E_{LUMO} and molecular descriptors using multivariate statistics;

(4) To evaluate the contribution of each molecular descriptor to QSAR models based on the mechanism principle of different DBP classes;

(5) To validate the models by using internal and external cross validation techniques;

(6) To quantify the model uncertainties through the Bootstrapping and the Taylor methods.

Through the comprehensive QSAR models predicting molecular property (E_{LUMO}) for the five DBP classes, this research also contributed additional knowledge that was not previously available for the DBP study.

2. Determin Quantitative Structure-TOXICITY Relationship (toxicity vs. logP, E_{LUMO}, and N_{Cl}) for various organisms (i.e., fish, T. pyriforims, and P. phosphoreum) of DBPs.

We collected or tested data for the toxicity and molecular properties of a broad range of DBP chemicals including chloroalkanes, chloroaromatics, and chloroaromatic compounds with a nitro- and cyano group. According to this main research facet, the following objectives emerge:

(1) To obtain information about molecular properties that influences the toxicity of DBPs with regard to their proposed mode of toxic action. The number of chlorine and number of carbon utilized in this research represent a realistic and typical example of the type of molecular descriptor for explaining the toxic activity;

(2) To find the outlier to be present in the models and were removed to facilitate model development and explain the reason the outliers were numerically distant from the rest of the data;

(3) To understand and compare different toxic mechanisms according to different contribution of N_{Cl}, N_C, and E_{LUMO} to the developed QSAR models.

1.4 Significance of Research

The study is significant for three reasons. First, systematic study of effects of N_{Cl} and/or N_C on molecular properties for various DBP classes, our QSAR models are valuable to practicing engineers for predicting the molecular property (E_{LUMO}) of untested chemicals. These chemicals are related to our test chemicals using E_{HOMO}, N_{Cl}, and N_C as molecular descriptors.

Second, this dissertation shows that three regression methods in QSAR analysis are performed and compared for the estimating characteristics of DBPs. Otherwise, model validation and uncertainty quantification are used as the critical steps before QSAR model can be applied to predict and estimate the molecular properties and the toxicity of untested chemicals.

Third, there have been many studies to develop QSAR using numerous descriptors for the prediction of toxicity of chlorinated compounds. However, few studies have been reported to investigate QSAR analysis in various chemicals toxicities for the possibility of estimating the toxicity mechanism of DBP. This research is believed to be the first attempt to link two atom descriptors (N_{Cl} and N_C) to explain and predict the toxicity of chlorinated compounds.

1.5 Outline of Dissertation

The aim of this dissertation is to develop a computational chemistry-based QSAR approach that enables identification of priorities within various selected groups of DBP chemicals. Validated QSAR models for molecular propertities and acute toxicity of selected groups of DBP chemicals were developed and taken into account. **Chapter 1** gives a general introduction of the subjects that are relevant within the context of the present dissertation. **Chapter 2** describes the selected data analysis methods and other correlated approaches for this research, such as Multi-linear regression (MLR), Partial Linear Regression (PLS) and Principal Component Regression (PCR) for developing QSAR model, Leave-One-Out (LOO) and K-fold Cross Validation for validating the reliability of the model, and also Taylor and Monte Carlo Simulation for estimating the

model uncertainty. In the next three chapters, QSAR models were developed to focus on the effect of the number of chlorine on electronic molecular properties such as E_{LUMO} and E_{HOMO} for five different classes of chlorinated DBPs, namely, chlorinated alkanes (**chapter 3**), chlorinated alkenes (**chapter 4**) and chlorinated aromatics (**chapter 5**). Three descriptors were investigated for their molecular properties in modeling the physicochemical activity (such as E_{LUMO}) of the chemicals in the five groups. These were:

(i) Energy of the highest occupied molecular orbital (E_{HOMO}), which models the nucleophilic nature of the chemicals;

(ii) Number of chlorine (N_{Cl});

(iii) Number of carbon (N_C).

In **chapter 6**, using experimental literature data sets on the acute toxicity of chlorinated alkanes, benzenes, anilines, phenols, nitro-phenols, and other substituted compounds on fish, *T. pyriformis*, and *photobacterium phosphoreum* to establish quantum chemistry-based QSARs were investigated. The logP is an important descriptor in explaining the toxicity of chlorinated compounds with additional electronic descriptors, E_{LUMO}, with N_{Cl} and/or N_C being required for the targeted test system.

Finally, the overall conclusion and a general discussion of this thesis are presented (**chapter 7**). A summary of the QSAR models for molecular properties and acute toxicity developed in this thesis, their applicability domains and the organisms for which the QSARs are shown in figure 1.3.

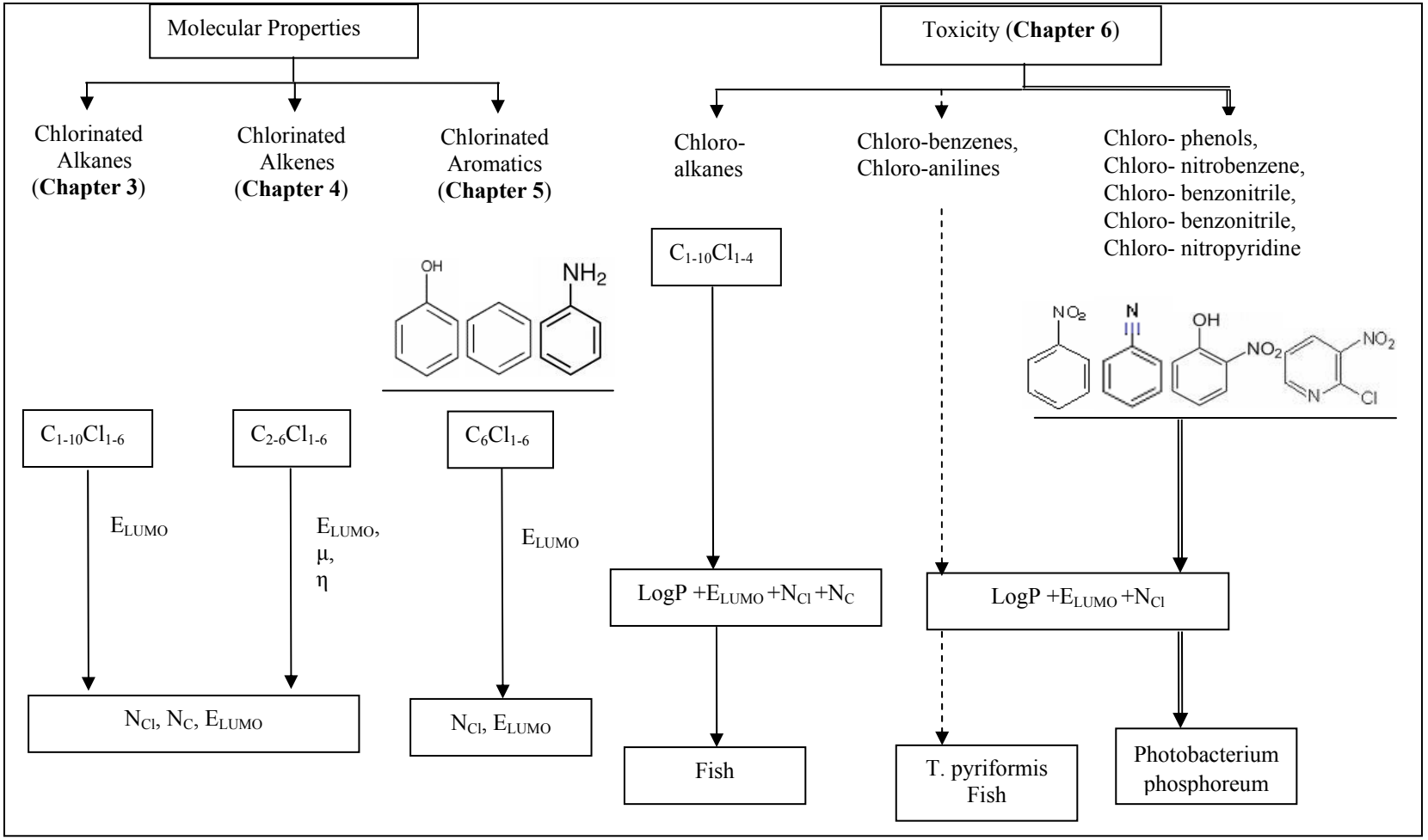


Figure 1.3 A summary of the QSAR for molecular properties and acute toxicity developed in this thesis.

CHAPTER 2 PRINCIPLES AND METHODOLOGIES IN QSAR

The exemplification of QSAR is that variation in structurally and electronically inherent properties of molecular similar compounds, reflects the variation in a given biological or physicochemical activity.

The behavior of organic compounds is closely related to the inherent molecular properties of the compound descriptive to environment partitioning and transport processes within and between different phases, as well as toxicological response of living organisms. Generally, three properties governing molecular activity, e.g., hydrophobic, electronic and structural inherent properties of the compounds are used to estimate the predominant parameters and toxicity response in models (Thomsen, 2001):

$$\text{Endpoint} = F\left(P_{\text{hydrophobicity}}, P_{\text{electronic}}, P_{\text{structural}}, P_x\right) + e$$

$P_{\text{hydrophobicity}}$ — hydrophobicity, is related to the individual compound affinity for partition to a biological membrane;

$P_{\text{electronic}}$, $P_{\text{structural}}$ — Electronic and structural, related to the ability to pass through the membrane, and bind to a receptor or specific sorption site;

P_x — accounts for underlying known or unknown effects, which influence the measured endpoint.

Historically, many statistical techniques are used in QSAR analyses. The predominant method is the linear regression technique because it is the method of choice for QSAR analysis. To describe these statistical QSAR modeling techniques in details, this chapter

starts with the introduction in section 2.1 and follows the model building in section 2.2. In section 2.2, the review of statistical analysis methods for linear regression follows the chronological progression, starting with the relatively simple Multiple Linear Regression (MLR) and progressing through the principal component based methods, such as Principal Component Analysis (PCA) and Partial Least-Squares (PLS), to the model validation and the model uncertainty estimation (Korhonen, 2007).

2.1 Introduction

In QSAR, molecular descriptors (X) are always correlated with one or more response variables (y). The conclusions drawn from a regression analysis are dependent on the assumption of the regression model (Myers, 1997). The model expresses the value of a regressor variable as a linear function of one or more variables and an error term, a general model might be expressed as:

$$y_i = b_o + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_m x_{i,m} + e_i \qquad (2.1)$$

In equation 2.1, the b_0 is regression constant, b_m is unknown coefficient on the m^{th} predictor, and m is the total number of predictors. The eq. 2.1 is estimated by minimal least square, which yields parameter estimates such that the sum of squares of errors is minimized. The resulting prediction equation is:

$$\hat{y}_i = \hat{b}_o + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + \ldots + \hat{b}_m x_{i,m} \qquad (2.2)$$

where the variables are defined as in eq. 2.1 except the "^" denotes estimated values. Throughout this thesis, the lower case italic characters *i, j, k, l* and *m* will be used as

running indices, where i=1,…,I; j=1,…,J; k=1,…,K; l=1,…L and m=1,…M. It is assumed that all vectors are column vectors.

The error term in equation 2.1 is unknown because the true model is unknown. In case the model has been estimated, the regression residuals are defined as:

$$\hat{e}_i = y_i - \hat{y}_i \tag{2.3}$$

2.2 Model Building

2.2.1 Multiple Linear Regression (MLR)

MLR is the earliest and simplest of linear regression techniques. However, it is still quite useful in classical QSAR analysis with a small number of variables. When the endpoint needs to be modeled using more than one descriptor, then multivariate techniques are applied, a relationship between y and X in figure 2.2 is established. The basic MLR model is shown in eq.2.4, which models a response variable, y, as a linear combination of X-variables, with the coefficient b. The deviations between the data (y) and the model (Xb) are called residuals, and are denoted by e.

$$\begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} \quad \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

Figure 2.2 A typical QSAR data set for MLR method

$$y = Xb + e \tag{2.4}$$

The limitations of MLR are: (1) MLR requires normally distributed, independent and 100% relevant descriptors. This means that each descriptor is assumed the 100% relevant for the explanation of the "cause" of the measured endpoint. (2) When the number of variables is greater than the number of observations, the MLR will not yield a unique solution but rather a set of possible solutions (Korhonen, 2007). Topliss and Edwards (1979) recommended that the ratio of compounds to variables should be at least five.

To assess goodness-of-fit, the coefficient of multiple determination R^2 is used (eq. 2.5). R^2 is often described as the proportion of the variation of y that is explained by the regression.

$$R^2 = \frac{SS_{\mathrm{Re},g}}{SS_T} = \frac{\left(SS_T - SS_{\mathrm{Re},s}\right)}{SS_T} = 1 - \frac{SS_{\mathrm{Re},s}}{SS_T} \qquad (2.5)$$

- $SS_T = \sum_i \left(y_i - \bar{y}\right)^2$ =total sum of squares;
- y_i = observed dependent variable;
- $SS_{\mathrm{Re},s} = \sum_i \left(y_i - \hat{y}\right)^2$ =residual sum of squares;
- $\hat{y}$ = calculated dependent variable.
- $SS_{\mathrm{Re},g} = \sum_i \left(\hat{y} - \bar{y}\right)^2$ =sum of squares;
- $\bar{y}$ = mean value of the dependent variable;

Generally, the R^2 value can be greater when adding extra descriptors to the model, even if these added descriptors do not contribute to reducing the variance of the dependent variable. In order to avoid overfitting, another statistical parameter, R^2_{adj}, was taken into consideration.

$$R_{adj}{}^2 = 1 - \frac{SS_{\mathrm{Re},s}/(n-p-1)}{SS_T/(n-1)} = 1 - (1-R^2)\cdot\frac{(n-1)}{(n-p-1)} \qquad (2.6)$$

2.2.2 Partial Least Square (PLS)

PLS is an advanced regression methodology, which was first introduced by Wold et al. (1984, 1993) and has also been extensively utilized in chemometric applications. In PLS analysis, the relationship is sought between an X-block of p predictors and a single y response (PLS1) or a Y-block of r responses (PLS2). Where X is an n*p matrix (n is the number of chemical compounds included in the model, and p is the number of descriptors). In QSAR study, only one Y-variable is considered and therefore Y is an n*1 matrix. The method is especially suitable when the descriptors of X are intercorrelated (Thomsen, 2001).

The PLS method overcomes the disadvantages of the MLR method in: (1) PLS is insensitive to the collinearity among the variables; (2) PLS offers the advantage of handling data sets where the number of variables is greater than the number of observations; (3) PLS is minimizing the probability of obtaining chance correlations since it is determined by cross-validation.

2.2.3 Principal Component Analysis / Regression (PCA/PCR)

PCA is a method for reducing data dimensionality by applying mathematical techniques. In PCA, the independent block X (figure 2.3) is replaced by Principle Components (PCs) which are linear combinations of the columns in X. The methodology of PCA is to decompose the X- data matrix into the following bilinear form:

$$X = t_1 p_1' + t_2 p_2' + \ldots + t_a p_a' + E \qquad (2.7)$$

where t_i comprises the score values of samples and p_i' containing the loadings of variables. Thus, the purpose of PCA is simply to decompose X into A component score vectors T and loading vectors P where A< J. In QSAR, the X-variables are generally mean-centered and often scaled before PCA is applied.

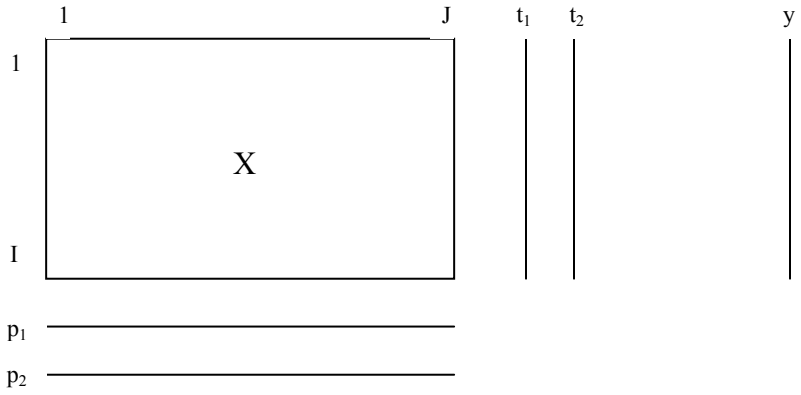


Figure 2.3 A graphical representation of the first two PCs (Nillson, 1998)

To perform Principal Component Regression (PCR), one must derive a matrix P, collecting the loading vectors where each column corresponds to an original loading vector, from the results of the PCA. Similarly, it can be shown that the matrix T is created to represent the scores. The regression coefficients can easily be derived using equation 2.10. In QSAR analysis, Equation 2.10 can be used for external predictions but may not be utilized for interpretation purposes.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1a} \\ p_{21} & p_{22} & \cdots & p_{2a} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{ma} \end{bmatrix} \tag{2.8}$$

$$T=XP \tag{2.9}$$

$$\hat{B}_{PCR} = (T'T)^{-1} T'Y \tag{2.10}$$

2.3 Working with Outliers

The chemical domain of applicability is an important issue when the reliable predictivity of QSAR model is assessed. Typically, applicability domain is a theoretical region in space defined by nature of the chemicals in the training set, and can be characterized in various ways. The Williams plot of the regression allows a graphical detection of both the outliers for the response and the structurally influential chemicals in a model.

Williams Plot is the plot of standardized residuals verses leverages. In this graphic method, the horizontal and vertical straight lines indicate the limits of normal values, i.e. plot of standardized residuals (y-axis) versvs leverages (x-axis) for each compound of the training set. Each standardized (cross-validated) residual is divided by its standard deviation, which is calculated without the ith observation. A simple formula for the standardized residual is shows in eq. 2.12. Leverage values can be calculated for both training compounds and new compounds where the leverage h_i of a compound measures its influence on the model. The leverage of a compound in the original variable space is defined as eq. 2.13:

$$r_i' = \frac{r_i}{s\sqrt{1 - h_{ii}}} \tag{2.12}$$

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \quad \text{(i=1, …, n)} \tag{2.13}$$

where x_i is the descriptor vector of the considered compound and X is the $n \times k - 1$ model matrix derived from the training set descriptor values. The warning leverage h* is defined as eq. 2.14. A leverage greater than the warning leverage h* is outside the chemical domain of the training set and, therefore, may not be reliable.

$$\bar{h}^* = 3 \times h = 3 \times \sum_i h_i / n = 3 \times p' / n \quad (i=1,\ldots, n) \tag{2.14}$$

Where n is the number of training compounds and p' is the number of model parameters.

An outlier of a QSAR model is a data point that is not well predicted. As part of the validation process, we should use the information that is generated about outliers to remove them from the QSAR equation, and then recalculate the equation until the results are satisfied. Before working with outliers, we must have a validated QSAR equation. The validation process identifies outliers and generates diagnostic data that helps us make decisions.

2.4 Model Validation

After a QSAR model is developed, it is essential to develop some quantitative measure of the predictive power and goodness of the fit of the new model for the training set. When estimating the predictive ability of a QSAR model, it is necessary to distinguish two types of predictive power, the internal and external predictivity, as illustrated in figure 2.4 (Worth, et al., 2005). The internal predictivity measures the accuracy of the model to predict the set of observations during building the statistical model in the training set, while the external predictivity is to measure the model's predictive power for compounds.

In order to estimate the predictive power of the model, one needs to have more complex scores for the quantitative models. The predictability is quantified as: squared correlation coefficient (R^2, eq. 2.15, range: 0-1.0), the Prediction Error Sum of Squares (PRESS, eq. 2.16, range: 0-∞), Residual Sum of Squares (RSS), Standard Deviation Error in Calculation (SDEC, eq. 2.17, range: 0-∞) and standard deviation error in prediction (SDEP, eq. 2.18, range: 0-∞).

$$R^2 = \frac{\left(\sum (y_{obs} - \bar{y}_{obs})(y_{pred} - \bar{y}_{pred})\right)^2}{\sum (y_{obs} - \bar{y}_{obs})^2 \sum (y_{pred} - \bar{y}_{pred})^2} \tag{2.15}$$

$$PRESS = \sum (y_{obs} - \bar{y}_{pred})^2 \tag{2.16}$$

$$SDEC = \sqrt{\frac{RSS}{n}} \tag{2.17}$$

$$SDEP = \sqrt{\frac{PRESS_{ex}}{n}} \tag{2.18}$$

The most commonly used cross-validation technique for "internal predictivity" is **Leave-One-Out** Cross Validation (LOOCV). LOOCV means one candidate is excluded from the training set at a time and used as the internal test set described earlier, and then a regression is carried out. As this process is repeated for all samples, the results obtained from the excluded values can be used to estimate the external predictivity of the model. However, there is a compelling problem for LOOCV where this approach is not sufficient to assess robustness and predictivity, the estimated Q^2 being too similar to R^2. It means that LOOCV often causes over-fitting, and on average, it gave an under-estimation of the true predictive error. The reason for LOOCV having such a deficiency is that many data

sets have a considerable structural redundancy, meaning that it tends to include unnecessary components into the model and make the model larger than it should be which can rigorously compromise the reliability of the LOOCV. Therefore, the model with the number of components determined by LOOCV often performs good in calibration, but poor in prediction. On the other hand, much attention has been paid to CV with more than one example left out at each step for validation, such as **Leave-Many-Out** (LMO) or synonymous **Leave-Group-Out** (LGO) Cross-Validation techniques, where the training set is divided into large subgroups, each containing a fixed proportion (typically, up to 50%) of samples which are in turn excluded just as in LOOCV. LMOCV is generally repeated a number of times, due to the large number of possible combinations of training sets generated by leaving out a fixed proportion of objects from the original data set.

Another important cross-validation statistical technique is **K-fold Cross-Validation**, In K-fold CV, the training set is randomly divided into K approximately equal parts (called folds). Recommended values are 5 or 10 groups. Each observation is randomly allocated to belong to one of the K groups.

For the cross-validation model, the following parameters are homologous to the parameters obtained from the non-cross-validation model, and are as a measure of the goodness of internal predictivity: cross-validated standard error of prediction (S_{PRESS} , eq. 2.19, range: 0-∞) and cross-validated squared correlation coefficient (Q^2, eq. 2.20, range: -∞-1.0).

$$S_{PRESS} = \sqrt{\frac{PRESS_{CV}}{n - NPC - 1}} \tag{2.19}$$

$$Q^2 = 1 - \frac{\sum (y_i - \bar{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{PRESS_{CV}}{SSY_{CV}} \qquad (2.20)$$

where the n is the number of samples, the NPC is the number of principal components extracted, if the analysis is not based on principle components NPC=1. Cross-validated scores Q^2 and S_{PRESS} are also used to decide the necessary number of principal components for PCR ad PLS models due to its role in determining the predictive ability of a QSAR model.

In contrast, the fitting parameter R^2, which always improves when more descriptors are added, while the value of Q^2 increases only with the useful predictors added. Variations to both Q^2 and R^2 are suggested in the literature (Cruciani et al., 1992; Baroni et al., 1989) but throughout this thesis, cross-validated Q^2, predicted Q^2 and R^2 are used for the presentation of cross-validations, external predictions and model calibration, respectively.

2.5 Model Uncertainty Analysis

2.5.1 Taylor Method

The functional relationship between the measured Y and the input quantities X_i is given by:

$$Y = F(X_1, X_2, ..., X_N) \qquad (2.17)$$

The function F includes not only corrections for systematic effects, but also accounts for sources of variability. The partial derivatives are computed at the mean values $\bar{x}_i$, and this is acceptable provided that the uncertainties in x_i are small and all values of x_i are close to $\bar{x}_i$.

The standard deviation $\sigma(x_i)$ is referred to, by the Guide to the expression of Uncertainty in Measurement (GUM), as the standard uncertainties associated with the input estimate x_i. The standard uncertainty in y and can be obtained by Taylor (1997):

$$u^2(y)=\sigma^2(y)=\frac{1}{N}\sum_{i=1}^{N}(y_i-y)^2=\left(\frac{\partial f}{\partial x_1}\right)^2\sigma(x_1)^2+\left(\frac{\partial f}{\partial x_2}\right)^2\sigma(x_2)^2+\ldots+\left(\frac{\partial f}{\partial x_N}\right)^2\sigma(x_N)^2 \\ +2\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_k}\right)\sigma(x_i x_k) \tag{2.18}$$

This equation gives the uncertainty as a standard deviation irrespective of whether or not the measurement of x_i is independent of the nature of the probability distribution. eq. 2.18 can be written in terms of the correlation coefficient, $\rho_{x_i x_k}$

$$u(y)=\sigma(y)=\sqrt{\sum_{i=1}^{N}\left(\frac{\partial f}{\partial x_i}\right)^2\sigma(x_i)^2+2\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_k}\right)\rho_{x_i x_k}\sigma(x_i)\sigma(x_k)} \tag{2.19}$$

The partial derivatives $(\partial f/\partial x_i)$ are called sensitivity coefficients, which give the effects of each input quantity on the final results.

The term "expanded uncertainty" is used in GUM to express the percent confidence interval about the measurement result within which the true value of the measurand is believed to lie and is given by:

$$U(y)=tu(y) \tag{2.20}$$

where t is the coverage factor on the basis of the confidence expressed as $y\pm U(y)$. For a level of confidence of approximately 95%, the value of t is two times the standard deviation. In other words, y is between $y\pm 2\sigma(y)$ with a 95% confidence interval. For a detailed analysis of the subject, please refer to Coleman and Steele (1995).

2.5.2 Monte-Carlo Simulation Method

Monte Carlo Simulation is defined as the numerical simulation of a QSAR model using probability approach. It *iteratively* evaluates a deterministic model using sets of random numbers as inputs. The major steps in MCS are: 1) an input is described with a distribution, thereby yielding a distribution for output; 2) the distribution of the input has to be determined either through statistic analysis or assumption. Uncertainty ranges and shapes of the Probability Density Function (PDFs) have to be quantitatively defined; 3) the model will be calculated at least 10,000 times using the random input variable predefined by a PDF using software such as Crystal Ball; and 4) software such as Crystal Ball will summarize the statistics of output such as mean, standard distribution and confidence interval. Histograms, cumulative distribution functions, and sensitivity of each input variable can also be analyzed.

For example, the slope and intercept of a QSAR model are generated by randomly sampling predefined error distribution populations, and adding these errors to a predefined true value. Monte Carlo simulation allows a test of statistical significance of the data with relatively simple calculations (Bevington and Robinson, 1992). "True" values for each variable in the data reduction equation were selected and the "true" value for the result was calculated. The word "true" is emphasized to indicate that it represents the actual physical quantity of the parameter, if it could be measured without any bias or precision error, which is always unobtainable. The 2σ bias limits for each error source were then assigned assuming that individual error sources were normally distributed, and random values for each error source were found using a Gaussian random deviate

generator subroutine with the assigned 2σ bias limit for each error source. When the elemental errors for the variables were correlated, the same elemental error value was used for each variable. The individual random elemental error values were then summed and added to the true value for each variable. These variable values were then used in the data reduction equation to obtain the random value of the result.

Monte Carlo simulation can be interpreted as representing what would happen if a molecular descriptor was used to predict a molecular property such as E_{HOMO} or E_{LUMO} with each of the elemental error sources estimated at a 95% confidence value. Since each of the elemental error sources is specified at 95% confidence, the uncertainty in the result is also desired to have 95% confidence. Calculation process starts by defining the probability distributions to the variables of a QSAR model. After that, the probability distribution of parameters is calculated by inserting the generated distributions of variables to the QSAR models. MCS is based on a large number M of trials, the r^{th} of which takes a random sample from the PDF for the value of each X_i and forms the corresponding model value y_r. A graphical illustration of the concept is given in figure 2.4. Three input quantities influence the measured data. The resulting PDF for the measure is obtained by combining "through" the model all possible combinations of values for the input quantities.

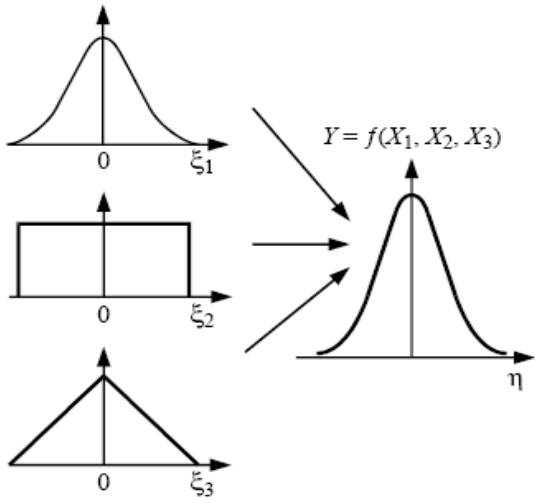


Figure 2.4 Demonstration of principle of the propagation of distributions

Based on the framework of statistical QSAR modeling described in this chapter, QSAR models will be developed with focus on the effect of number of chlorine on electronic molecular properties for different classes of chlorinated DBPs in a comparative way of study, namely, chlorinated alkanes (**chapter 3**), chlorinated alkenes (**chapter 4**) and chlorinate aromatics (**chapter 5**).

CHAPTER 3 QSAR STUDY OF CHLORINE EFFECTS ON E_{LUMO} OF CHLORALKANES

Summary

QSAR is developed between molecular properties such as E_{LUMO} and molecular descriptors such as number of chlorine atoms and number of carbon contained in chlorinated alkanes, which is one of the major classes of DBPs in drinking water. After QSAR models are established and before being applied to toxicity prediction, three model validation methods are used to validate the models. For example, 1) Linear or Multi-linear Regression (MLR); 2) Partial Least Squares (PLS); and 3) Principle Component Regression (PCR) are used to investigate the effects of chlorine number on molecular descriptors such as E_{LUMO} and E_{HOMO} of DBPs. The LOOCV procedure by itself is not enough to give reliable representation of predictive ability of the QSAR models. However, K-fold Cross-Validation and external validation can be applied together to achieve much more reliable results. According to the results from uncertainty analysis using the Taylor method, the uncertainty of the intercept of QSAR model is more sensitive than the slope.

3.1 Introduction

Chlorinated Alkane (CA) is a major class of DBPs formed during chlorination of water. Generally, CAs can be classified as chlorinated n-alkanes with carbon chain lengths ranging from C_1 to C_{10}. Chlorine content of the products varies between 30 and 70% by weight. Typical CA compounds such as Trihalomethanes (THMs) are regulated by the US EPA and have the Maximal Contaminant Level (MCL) of 60 μg/L. Major toxicological effects caused by chlorinated alkanes are their mutagenic, carcinogenic, and reproductive toxicity. Chlorinated alkanes are an important group of chemicals with widespread use, large production volumes, and thus a large potential for environmental pollution.

It has been reported that the more chlorine atoms a chlorinated compound contains, the more toxic is the chemical. The energy of lowest unoccupied molecular orbital (E_{LUMO}) has shown to be correlated to toxicity of DBPs, but as far as we are aware, there exists no QSAR models to predict the relationship between E_{LUMO} and N_{Cl}, E_{HOMO}, and/or N_C of chlorinated alkanes. Molecular prosperity tests were performed for a relatively large set of chlorinated alkanes across a wide range of electronic values and carbon chain lengths (C_1-C_{10}), allowing for the examination of quantitative relationships between physical properties based on carbon chain length and degree of chlorination for these compounds.

3.2 Data Set

Data was collected for thirty-six derivatives of chlorinated alkanes. Figure 3.1 presents the structure of these compounds. Due to their structural similarity, it is to be expected that they all fall into a similar mode of toxic action, and thus can be modeled by the same

QSAR model. In table 3.1, the first 31 compounds constitute a so-called training set to develop the QSAR model. These compounds differ in the number of chlorine and carbon on the carbon chain. The rest of 5 different compounds were used as the prediction set to test the models as well as the algorithms developed in this study.

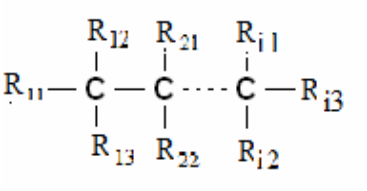


Figure 3.1 Chemical structures of chlorinated alkanes used in this study

Table 3.1 Molecular properties of 36 chlorinated alkane congeners

	Compounds	N_{Cl}	N_C	E_{HOMO}	E_{LUMO}		Compounds	N_{Cl}	N_C	E_{HOMO}	E_{LUMO}
1	2-chlorobutane	1	4	-0.422	0.212	19	1,2-dichlorobutane	2	4	-0.428	0.192
2	1-chlorohexane	1	6	-0.418	0.216	20	2,3-dichlorobutane	2	4	-0.425	0.180
3	2-chloro-2-methyl-butane	1	5	-0.419	0.205	21	1,2-dichloro-2-methylbutane	2	5	-0.427	0.179
4	1-chloroheptane	1	7	-0.416	0.216	22	1,1,1-trichloroethane	3	2	-0.447	0.131
5	Chloromethane	1	1	-0.432	0.217	23	1,1,2-trichloroethane	3	2	-0.445	0.156
6	2-chloropropane.	1	3	-0.423	0.210	24	1,1,1-trichloropropane	3	3	-0.445	0.133
7	1-chloropentane	1	5	-0.420	0.216	25	1,1,2-trichloropropane	3	3	-0.432	0.161
8	1-chlorooctane	1	8	-0.415	0.217	26	1,1,3-trichloropropane	3	3	-0.432	0.159
9	1-chlorodecane	1	10	-0.411	0.217	27	1,2,2-trichloropropane	3	3	-0.438	0.153
10	2-chlorohexane	1	6	-0.416	0.214	28	1,1,2,2-tetrachloroethane	4	2	-0.441	0.148
11	3-chlorohexane	1	6	-0.417	0.215	29	carbon tetrachloride	4	1	-0.467	0.095
12	Chloroethane	1	2	-0.426	0.215	30	Pentachloroethane	5	2	-0.448	0.118
13	1,2-dichloropropane	2	3	-0.429	0.188	31	Hexachloroethane	6	2	-0.454	0.110
14	1,4-dichlorobutane	2	4	-0.425	0.203	32	1-chlorobutane	1	3	-0.422	0.215
15	1,5-dichloropentane	2	5	-0.422	0.203	33	1-chloropropane	1	4	-0.425	0.211
16	Dichloromethane	2	1	-0.443	0.166	34	1,2-dichloroethane	2	2	-0.439	0.189
17	1,1-dichloroethane	2	2	-0.440	0.168	35	2,2-dichloropropane	2	3	-0.432	0.169
18	trans-1,2-dichlorocyclohexane	2	6	-0.421	0.180	36	1,2,3-trichloropropane	3	3	-0.436	0.160

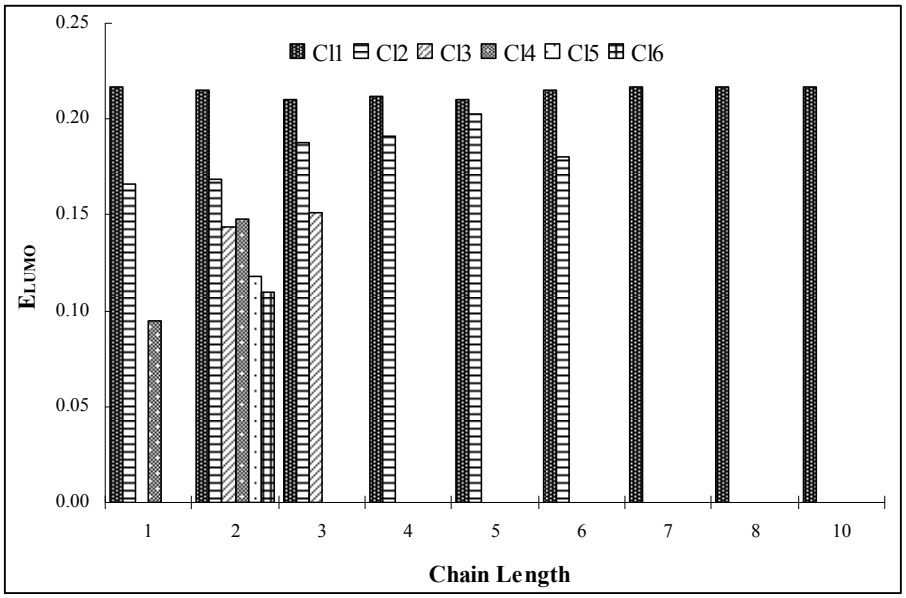


Figure 3.2 Typical congener and homologue group patterns of E_{LUMO} and number of chlorine with the chain length from C1-10 for CAs

Figure 3.2 shows typical distributions of congener groups and homologues in chlorinated alkane data sets. E_{LUMO} varies not significantly when the one atom chlorine substitute to carbon-chain, where the number of carbon is distributed from 1 to 10. The relationship between the number of chlorine and E_{LUMO} is visualized in this figure, that is, E_{LUMO} decreases as chlorine content in an alkane increases.

3.3 Results and Discussion

This chapter evaluates the contribution of number of chlorine, number of carbon, and E_{HOMO} to the value of E_{LUMO}. Statistic methods are developed to select the molecular descriptors which can form robust QSAR model in predicting E_{LUMO}. The aims of this research are four fold: 1) Statistical methods will be established to select the most influential molecular descriptors among E_{HOMO}, N_{Cl}, and N_C; 2) The most robust QSAR model will be developed to predict E_{LUMO}; 3) The developed QSAR model will be

validated; and 4) Uncertainty of the developed QSAR model will be quantitatively defined.

3.3.1 Evaluation of Molecular Descriptors

Model 1: $E_{LUMO}=a_1*N_{Cl}+k_1$

Figure 3.3 is a Williams plot which shows that the majority of compounds of the training set are inside of this square area. However, compound 31 has a leverage greater than threshold h*, and shows standard deviation values greater than the limit (±2σ), which implies that it can be considered as an outlier or influential chemical, respectively. Otherwise, two responding outliers can be identified in the training set: 1,1,1-trichloroethane and carbon tetrachloride. By removing these three outliers, R^2 value is improved to 0.8694 from 0.854. Those three points will be discarded in the following analysis.

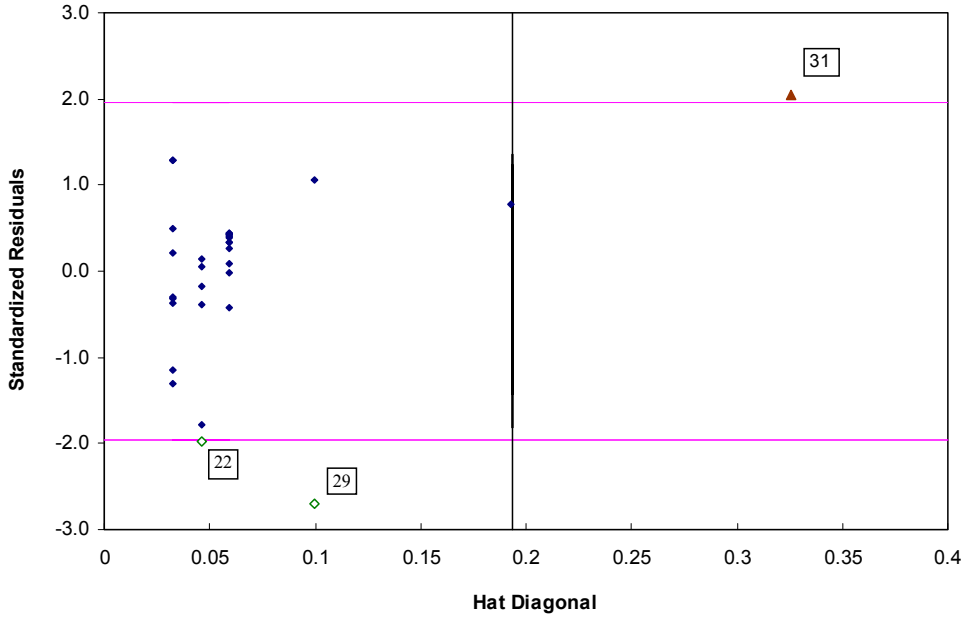


Figure 3.3 Outlier detection of model 1 for alkane

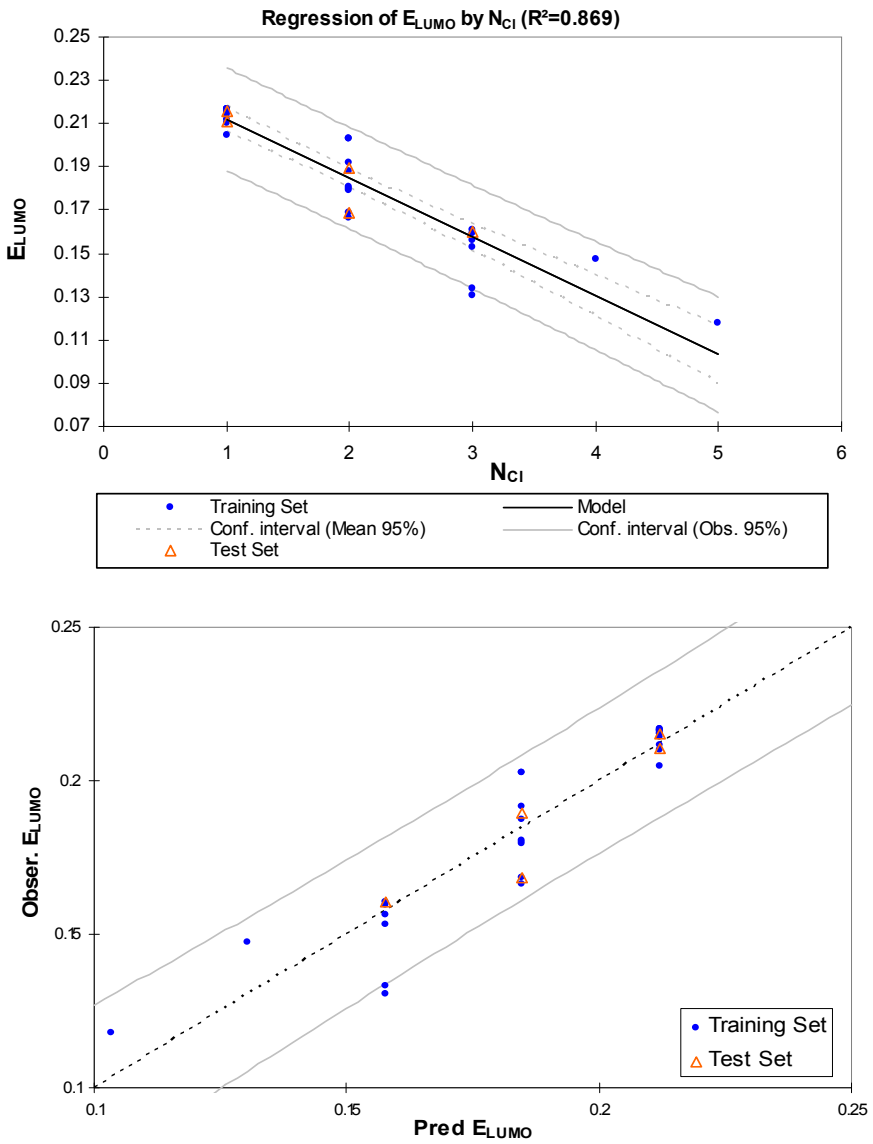


Figure 3.4 (A) The trend of N_{Cl} and E_{LUMO} of model 1, (B) Relationship between observed and predicted endpoint data

After refitting the model, the relationship between N_{Cl} and E_{LUMO} is plotted in figure 3.4 A, which shows that the descriptor number of chlorine is negativly interrelated to the E_{LUMO}. As E_{LUMO} decreases, the ability of a compound to undergo reduction increases; therefore, as chlorine content in an alkane increases, the reductive potential of the molecule increases. E_{LUMO} represents 88.5% of the variance in the linear regression

equation; the slope of the regression model is -0.0272, which is the decreasing rate of E_{LUMO} with number of chlorine atom. The correlation indicates that, as the number of chlorine increases from 1 to 6 in an alkane compound, E_{LUMO} will decrease 0.2166. E_{LUMO} is an electrophilicity parameter, and it appears as directly proportional to the electronic affinity of the compounds. The lower the E_{LUMO} values, the stronger the electrophilicity.

As shown by figure 3.4B, for the chlorinated alkanes study, the correlation between observed and predicted E_{LUMO} values is very significant (r=-0.9407, P<0.0001). The low residual values reveal the importance of the chlorine number as a predictive descriptor for E_{LUMO}. The relationship between number of chlorine and E_{LUMO} was determined.

$$E_{LUMO}=-0.02717\ N_{Cl}+0.2391 \quad (3.1)$$

N= 28, R^2 = 0.8855, F=201.00, RMSE= 0.0101, P= 0.000

Model 2: $E_{LUMO}=a_2*N_{Cl}+b_2*N_C+k_2$

Following the same model development procedure as in model 1 and table 3.1, compounds 9 and 31 have the leverage value greater than threshold h*, and only one response outlier can be identified in the training set: carbon tetrachloride. After removing carbon tetrachloride from the data set and refitting the model, PLS model for the relationship between E_{LUMO} and two descriptors (N_{Cl} and N_C) is shown as eq. 3.2. It shows that the number of chlorine has a negative effect to E_{LUMO}, and number of carbon has a positive effect to E_{LUMO}.

$$E_{LUMO}=-0.02717\ N_{Cl}+0.00218\ N_C+0.22168 \quad (3.2)$$

N= 30, R^2 = 0.8840, F= 102.869, RMSE= 0.01170, P= 0.0000

Model 3: $E_{LUMO}=c_3 * E_{HOMO}+k_3$

E_{LUMO} is also correlated with E_{HOMO} as listed in table 3.1. Figure 3.5 shows that slope of the QSAR model is 2.591. The correlation indicates that as the E_{LUMO} increases from 0.1098 to 0.2166 in a chlorinated alkane compound, that E_{HOMO} will decrease by 0.2166. Correlation between E_{LUMO} and E_{HOMO} as descriptors provides a reasonably good coefficient of determination r^2=0.8271.

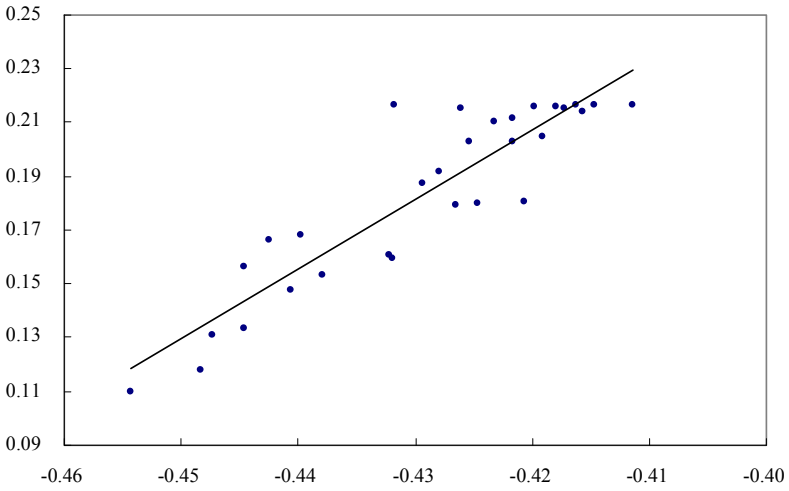


Figure 3.5 The trend of E_{HOMO} and E_{LUMO} of model 3

E_{LUMO}=2.5906 E_{HOMO}+1.295 (3.3)

N= 30, R^2 = 0.8271, F= 133.95, RMSE= 0.01403, P= 0.0000

The significance of E_{HOMO}, E_{LUMO} and N_{Cl} relationship (figure 3.6) indicated that number of chlorine is an effective modeling descriptor for predicting molecular properties of compounds. E_{HOMO} and E_{LUMO} are global molecular properties that describe the electrophilicity of a compound in general terms, and a measurement of the ability of the molecule to lose or accept an electron, respectively. The correlation between the number of chlorines and E_{HOMO}, or E_{LUMO} shows that E_{HOMO} and E_{LUMO} decrease as the

number of chlorines increase with the correlation coefficient r = -0.8389 and -0.9410. Overall, E_{LUMO} correlates well with number of chlorine atoms for the chlorinated alkane compound.

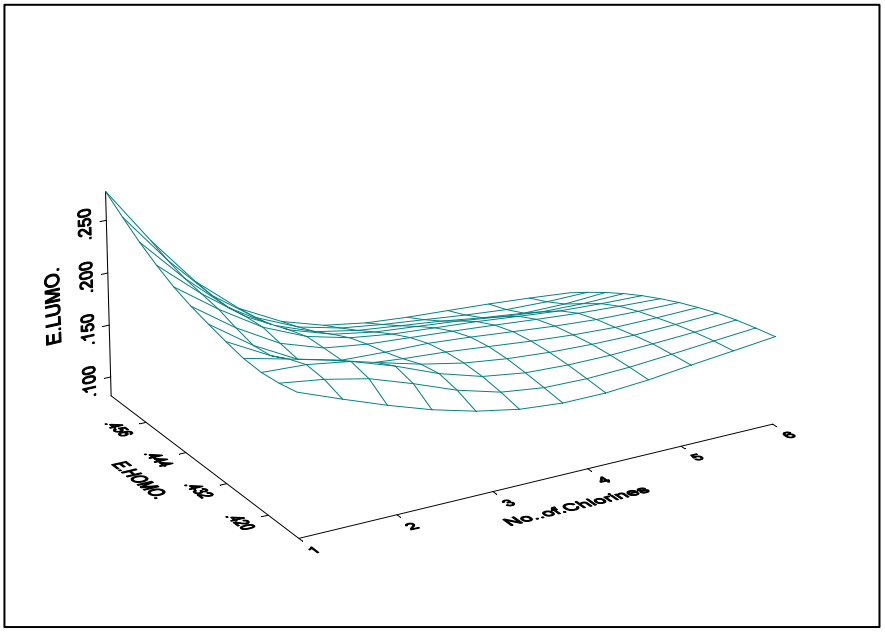


Figure 3.6 3D plot for E_{HOMO}, E_{LUMO}, and N_{Cl}

3.3.2 Development of QSAR Model

Activities of molecules in the biological systems are highly influenced by their inherent electronic properties. Hence, E_{HOMO} along with N_{Cl} and N_C were selected as candidate parameters for E_{LUMO} simulation and prediction. Using all three descriptors in the QSAR model, for the combined set of 31 alkane congeners, maximum values for the coefficient of correlation and lowest root mean square error were obtained. The final model with three molecular descriptors is as follows:

$$E_{LUMO} = a_4 * N_{Cl} + b_4 * N_C + c_4 * E_{HOMO} + k_4$$

The QSAR for the complete data set were examined further to identify statistical outliers. Compound 31, hexachloroethane, with high residual values was identified and excluded as an outlier. By removing this outlier, R^2 value is improved to 0.956 from 0.951. Multiple Linear Regression (MLR) Model:

$$E_{LUMO}=-0.1474\ N_{Cl}-0.003766\ N_C+1.9528\ E_{HOMO}+1.0664 \quad (3.4)$$

N= 30, R^2=0.956, F=188.498, RMSE =0.00762, P=0.000

The results for the corresponding PLS method are showed in figure 3.7. The data analysis resulted in a QSAR with R^2X=0.861, R^2Y=0.819, and Q^2Y=0.848, which are excellent statistics considering that the response is handled simultaneously. Figure 3.7A is the plot interpretation one considers the distance to the plot origin. The further away from the plot origin an X- or Y-variable lies, the stronger the model impact that particular variable has. It indicates that all X-variables load strongly in the model, and that E_{HOMO}, number of carbons (N_{Cl}), and number of chlorine (N_C) are closely related. Overall, N_{Cl} and E_{HOMO} are the most important X-variables. 1-chlorodecane, has the least number of chlorine, but the highest number of carbon. Therefore, it has the least E_{LUMO}. In addition, we must also consider the sign of the PLS loading, which informs about the correlation among the variables. Figure 3.7B shows the model scores, the ellipse indicates the model applicability domain as defined by Hotelling's T^2. It provides a check for compounds adhering to multivariate normality (Jackson 1991). There are no outliers in the score space because all compounds lie inside the elliptic 95% tolerance volume depicted in the plot. Hotelling's T^2 is a multivariate generalization of Student's t-test.

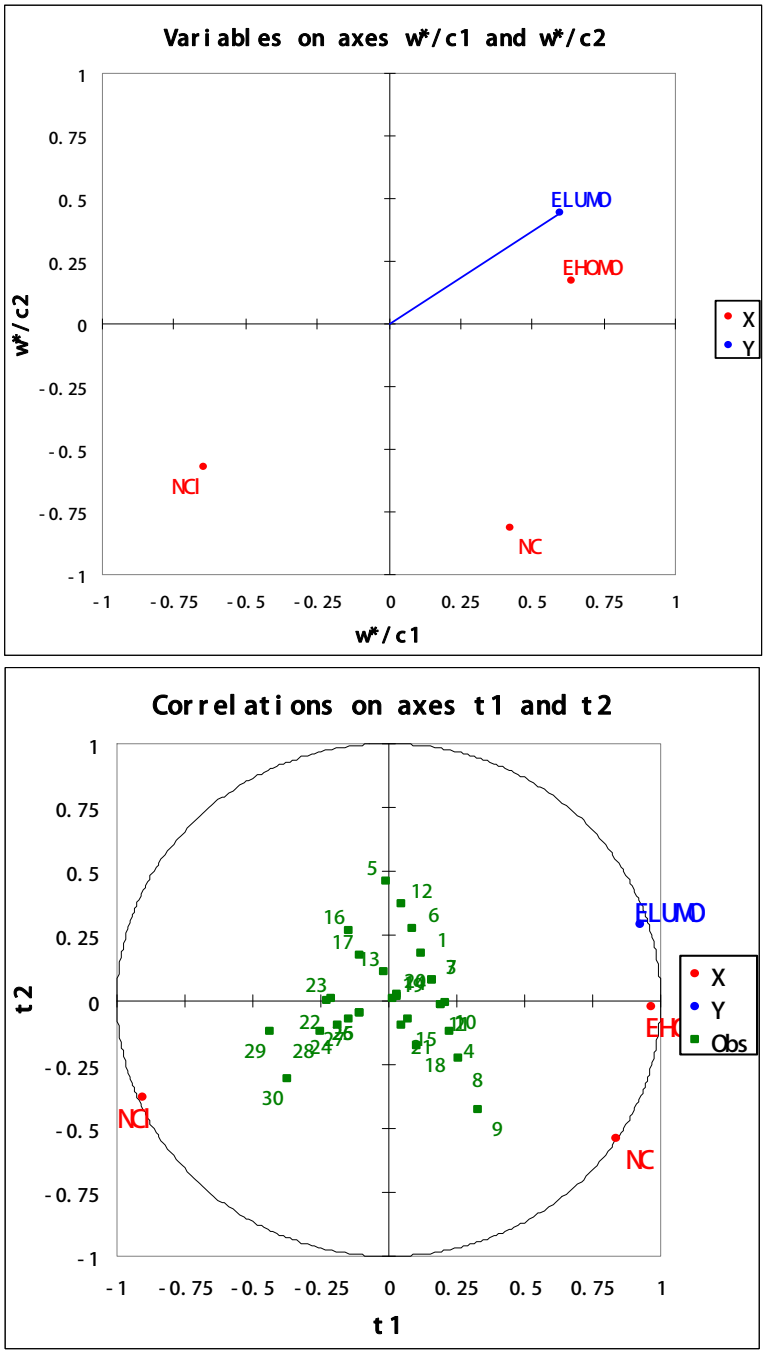


Figure 3.7 (A) PLS loading plot, (B) PLS scores plot of first two PC

Partial Linear Squares (PLS) Selected Model:

E_{LUMO}=-0.0147 N_{Cl}-0.003766 N_C+1.9528 E_{HOMO}+1.066 (3.5)

N= 30, R^2=0.956, F=188.498, RMSE =0.00709, P=0.000

Another suitable Multivariate technique is Principle Component Regression (PCR). This method is similar to MLR and PLS in that they all use a variant of principal component extraction to overcome the problems of correlated descriptors. Figure 3.8A shows 2D biplot displays for the first two PCs (PC_1: E_{HOMO}, PC_2: N_{Cl}). It shows that compound 20 is significantly positive relative to N_{Cl}, and 1-chlorodecane was strongly correlated with N_C. Biplot analysis allowed the confirmation of the relationship between different variables as well as to define groups of strains.

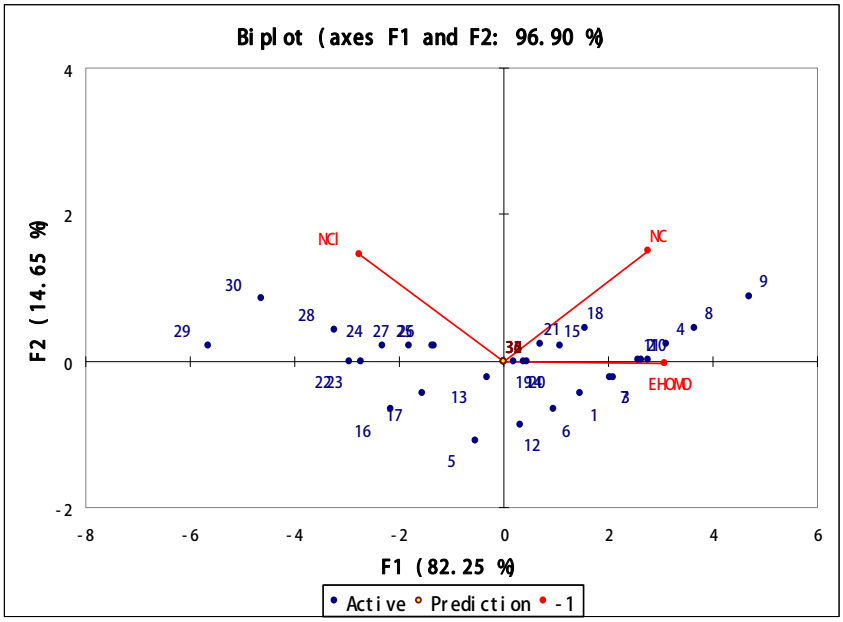


Figure 3.8 Biplot of F1(82.25%) vs. F2 (14.65%)

Table 3.2 Result comparison of model 4 using three different calibration methods

Calibration methods	Model 4	R^2	RMSE
MLR	E_{LUMO}=-0.1474 N_{Cl}-0.003766 N_C+1.9528 E_{HOMO}+1.066	0.956	0.00762
PLS	E_{LUMO}=-0.0147 N_{Cl}-0.003766 N_C+1.9528 E_{HOMO}+1.066	0.9560	0.00709
PCR	E_{HOMO}=-0.01474 N_{Cl}-003766 N_C+1.9528 E_{LUMO}+1.066	0.9560	0.00762

Table 3.3 Experimental and calculated values of E_{LUMO} for the model 4

No.	Compounds	Descriptors			E_{LUMO}[a] values		
		E_{HOMO}	N_{Cl}	N_C	Calculated	Predicted	Residual[b]
1	2-chlorobutane	-0.4217	1	4	0.2116	0.2127	-0.0011
2	1-chlorohexane	-0.4179	1	6	0.2161	0.2126	0.0035
3	2-chloro-2-methyl-butane	-0.4191	1	5	0.2045	0.2140	-0.0095
4	1-chloroheptane	-0.4163	1	7	0.2164	0.2121	0.0043
5	Chloromethane	-0.4319	1	1	0.2166	0.2042	0.0124
6	2-chloropropane.	-0.4233	1	3	0.2102	0.2134	-0.0032
7	1-chloropentane	-0.4199	1	5	0.2160	0.2126	0.0034
8	1-chlorooctane	-0.4146	1	8	0.2165	0.2115	0.0050
9	1-chlorodecane	-0.4114	1	10	0.2167	0.2103	0.0063
10	2-chlorohexane	-0.4156	1	6	0.2142	0.2171	-0.0029
11	3-chlorohexane	-0.4173	1	6	0.2151	0.2138	0.0013
12	Chloroethane	-0.4261	1	2	0.2152	0.2117	0.0034
13	1,2-dichloropropane	-0.4293	2	3	0.1875	0.1869	0.0006
14	1,4-dichlorobutane	-0.4253	2	4	0.2027	0.1910	0.0117
15	1,5-dichloropentane	-0.4217	2	5	0.2027	0.1942	0.0084
16	Dichloromethane	-0.4425	2	1	0.1662	0.1688	-0.0026
17	1,1-dichloroethane	-0.4398	2	2	0.1684	0.1703	-0.0019
18	trans-1,2-dichlorocyclohexane	-0.4206	2	6	0.1803	0.1927	-0.0123
19	1,2-dichlorobutane	-0.4280	2	4	0.1915	0.1858	0.0056
20	2,3-dichlorobutane	-0.4246	2	4	0.1801	0.1923	-0.0123
21	1,2-dichloro-2-methylbutane	-0.4265	2	5	0.1793	0.1850	-0.0057
22	1,1,1-trichloroethane	-0.4473	3	2	0.1306	0.1409	-0.0103
23	1,1,2-trichloroethane	-0.4446	3	2	0.1562	0.1462	0.0101
24	1,1,1-trichloropropane	-0.4446	3	3	0.1334	0.1425	-0.0091
25	1,1,2-trichloropropane	-0.4323	3	3	0.1605	0.1665	-0.0060
26	1,1,3-trichloropropane	-0.4319	3	3	0.1593	0.1671	-0.0079
27	1,2,2-trichloropropane	-0.4379	3	3	0.1532	0.1554	-0.0023
28	1,1,2,2-tetrachloroethane	-0.4406	4	2	0.1476	0.1393	0.0083
29	carbon tetrachloride	-0.4667	4	1	0.0947	0.0920	0.0026
30	Pentachloroethane	-0.4484	5	2	0.1178	0.1094	0.0085
32	1-chloropropane	-0.4246	1	3	0.2106	0.2109	-0.0003
33	1-chlorobutane	-0.4217	1	4	0.2154	0.2127	0.0027
34	1,2-dichloroethane	-0.4385	2	2	0.1892	0.1728	0.0164
35	2,2-dichloropropane	-0.4322	2	3	0.1687	0.1813	-0.0127
36	1,2,3-trichloropropane	-0.4361	3	3	0.1604	0.1590	0.0014

[a] With N_{Cl}, N_C, E_{HOMO} as descriptors. [b] The residual is the difference between calculated and predicted E_{LUMO} values.

One of the initial aims of developing QSAR models is finding a good fit of the model to the data set. Table 3.2 shows a comparison of the quality of MLR, PLS and PCR models as represented by their R^2 and RMSE. A comparison of the slopes and intercepts indicates that they are not significantly different. The PLS approach has a slightly better statistical quality and can predict the activity of the compounds with lower RMS error than the other two approaches in developing QSAR models. This conclusion is in agreement with the findings of Verhaar and Eriksson (1994), who demonstrated that PLS, among other techniques, should show a better predictiveness than subset regression techniques (Henk et al., 1994). Table 3.3 lists the experimental and the calculated E_{LUMO} values for the selected set of 31 alkane congeners obtained using all three selected descriptors.

The range of compounds for which the model is valid was determined by taking into account the minimum and maximum values of the (i) carbon chain length (C_1-C_{10}), (ii) the chlorine atoms (Cl_1-Cl_5), and (iii) the E_{HOMO} values [(-0.4484)-(-0.4114)] of compounds included in the training set. Taking these criteria into consideration, our QSAR models are thus applicable to chlorinated alkanes with up to 10 carbon and five chlorine atoms and the E_{HOMO} values between -0.4484 and -0.4114. E_{LUMO} tended to decrease with increasing carbon chain length and degree of chlorination according to the correlation for model 4.

Regression models for the training set of 31 chlorinated alkane congeners with calculated E_{LUMO} values were 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 (E_{LUMO}) energy as independent variables, are presented in table 3.4. This table gives an overview of the PLS models for the endpoint E_{LUMO} that have the lowest RMSE values. It can be seen from table 3.4 that, according to the rule that r^2 should be greater than 0.6 in a good model, all models in this table are significant and most models can be considered good models. All these QSARs were developed after removing the outliers. The outliers, and possible reasons for these compounds being outliers, are listed in table 3.5.

Table 3.4 Regression models for E_{LUMO} using various descriptors for CAs

Model	Regression Equations	R^2	RMSE	Outlier*
1	E_{LUMO}=-0.02717 N_{Cl} +0.2391	0.8855	0.0101	22, 29, 31
2	E_{LUMO}=-0.0224 N_{Cl}+0.00218 N_C -0.2217	0.8840	0.0117	29
3	E_{LUMO}=2.5422 E_{HOMO} +1.2731	0.8271	0.0140	29
4	E_{LUMO}=-0.0147 N_{Cl}-0.003766 N_C+1.9528 E_{HOMO}+1.066	0.9560	0.0071	31

* Outliers had already been removed before refitting the model

Table 3.5 Outliers and potential reasons for these compounds being outliers

Outliers	Potential reasons for outliers
Outliers to model 1 alone : 1,1,1-trichloroethane	• Non-polar solvent; • Three Chlorine atoms lie on the same side; • Slightly polar.
Outlier to model 1, 2, and 3: Carbon tetrachloride	• Four chlorine atoms are positioned symmetrically; • Symmetrical geometry; • No net dipole moment; • Non-polar.
Outlier to model 1 and 4 : Hexachloroethane	• Electron deficiency on the carbon atoms; • Susceptible to reduce reaction.

Some possible combinations of parameters were considered.The best equation was selected among other equations by considering the various statistical criteria. The results showed that there was one best equation included E_{HOMO}, N_{Cl}, and N_C for the activities against E_{LUMO}.

3.3.3 QSAR Model Validation

In order to judge the validity of the predictive power of the QSAR, a cross-validation method was applied to the original data set for models 1 to 4. Internal predictabilities of the models are characterized by r^2_{cv} and root mean squares errors of cross-validation (RMS_{CV}) are given. LOO-CV and three-fold CVs with k=10, 5, and 2 were calculated.

In table 3.6, d=1 indicates LOO-CV, it is seen that, for model 4, RMS_{CV} based on the LOO-CV is greater than the true RMS by about 7.3% ((0.0076-0.0071)/0.0071), and RMS_{CV} based on K-fold are greater than true RMS in the range 6.6%-9.2%. The difference between RMS_{CV} and true RMSE reaches its minimum at k=10. For models 1, 2 and 3, in the case d=1, CV estimates the r^2_{cv} and RMS_{CV} values with satisfactory accuracy.

Table 3.6 Results of LOO and K-fold Cross-Validation test for alkanes

Model No.	d=1 (LOOCV)		k=2		k=3		k=5		k=10	
	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}
Model 1	0.8856	0.0101	0.8588	0.0116	0.8862	0.0098	0.8821	0.0101	0.8876	0.0100
Model 2	0.8844	0.0117	0.8901	0.0114	0.9013	0.0104	0.8753	0.0116	0.8864	0.0114
Model 3	0.8269	0.0140	0.7775	0.0155	0.7837	0.0149	0.8085	0.0140	0.8129	0.0143
Model 4	0.9561	0.0076	0.9546	0.0077	0.9534	0.0077	0.9505	0.0075	0.9559	0.0075

Compared with k=2, 3, 5 and 10 show clearly that K-fold gives an excellent correlaiton coefficient when k>2. However, when k=10, the results may not be very reliable, since there are 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. These results have been tested by Breiman and Spector (1992), and Zhang

(1993), who did not reveal any statistical advantages of using 10-fold CV over 5-fold CV. Additionally, Zhang (1993) summarized that twofold CV would lead to the worst prediction errors.

In the present study, cross validation did confirm model 4 as the best QSAR model to predict E_{LUMO} of any compound in the class of Chlorinated alkanes. The cross-validation r^2_{cv} values had a maximum at a three-term model. Cross validation results using N_{Cl}, or E_{HOMO}, as the variable were less good than those including three parameters (N_{Cl}, E_{HOMO}, and N_C). The three parameters model was very stable, leading to cross-validated values in the range between 0.9505-0.9559, whereas the one-parameter models gave cross-validated value between 0.8588-0.9035. This showed that the three-term model indicated higher predictive ability, as shown by cross validation. On the other hand, the greater the number of variables tested, the greater the role chance will play in the observed correlation. Another two-term model was significant, but it has lower r_{cv} value and higher RMS_{CV} value than the three-term model.

3.3.4 Uncertainty Analysis

From the regression equation discussed in section 3.3, coefficients of parameters and the standard deviations are estimated in table 3.7 using bootstrap analysis. Fitting the logistic regression model by bootstrap with 5000 iterations gives the following coefficient estimates and their standard errors:

Table 3.7 Summary of coefficients and the standard deviations for model 1-4

Model No.	a	b	c	k	σ_a	σ_b	σ_c	σ_k
1	-0.0269	-	-	0.2389	0.00223	-	-	0.00354
2	-0.02297	0.00226	-	0.2222	0.00290	0.00119	-	0.00963
3	-	-	2.5484	1.2758	-	-	0.1474	0.06324
4	-0.01494	-0.00385	1.9578	1.0690	0.00291	0.00126	0.34145	0.14592

The expression for the uncertainty in E_{LUMO} determined from the regression model 4 at a measured or specified value of X is found by equation 3.6. Here, we did not consider the correlated uncertainties between any two of these variables and the uncertainty of number of chlorine and number of carbon are zero, and then all terms involving correlated uncertainties in eq.3.6 will be simplified as following equation 3.7.

$$\begin{aligned} U^2_{E_{LUMO}} &= \left(\frac{\partial(E_{LUMO})}{\partial a}\right)^2 U_a^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_{Cl}}\right)^2 U^2_{N_{Cl}} \\ &+ \left(\frac{\partial(E_{LUMO})}{\partial b}\right)^2 U_b^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_C}\right)^2 U^2_{N_C} \\ &+ \left(\frac{\partial(E_{LUMO})}{\partial c}\right)^2 U_c^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_C}\right)^2 U^2_{E_{HOMO}} + \left(\frac{\partial(E_{LUMO})}{\partial k}\right)^2 U_k^2 \end{aligned} \tag{3.6}$$

where, $\frac{\partial(E_{LUMO})}{\partial a} = N_{Cl}$, $\frac{\partial(E_{LUMO})}{\partial N_{Cl}} = a$, $\frac{\partial(E_{LUMO})}{\partial b} = N_C$, $\frac{\partial(E_{LUMO})}{\partial N_C} = b$, $\frac{\partial(E_{LUMO})}{\partial c} = E_{HOMO}$,

$\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})} = c$, $\frac{\partial(E_{LUMO})}{\partial k} = 1$.

$$U^2_{E_{LUMO}} = {N_{Cl}}^2 U_a^2 + {N_C}^2 U_b^2 + {E_{HOMO}}^2 U_c^2 + c^2 U^2_{E_{HOMO}} + U_k^2 \tag{3.7}$$

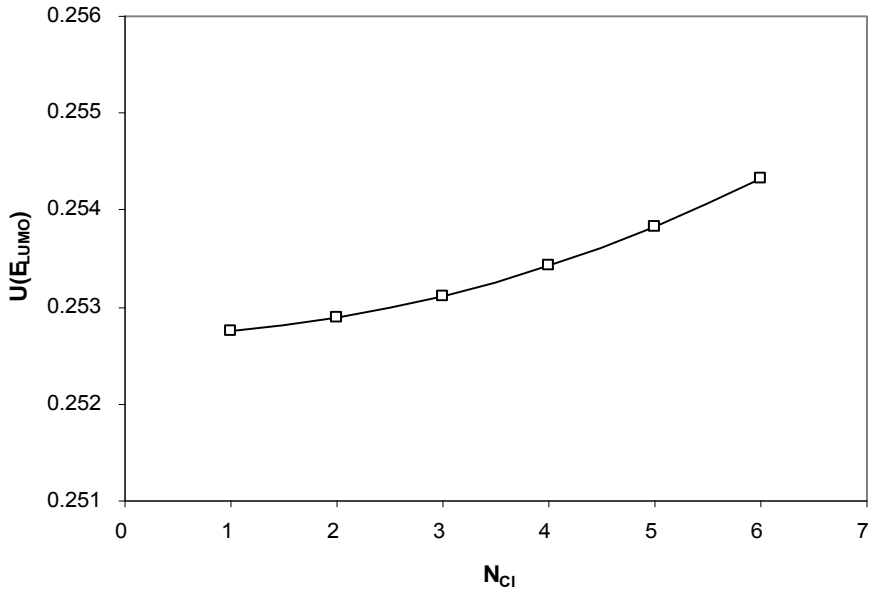


Figure 3.9 Relationship between N_{Cl} and uncertainty in E_{LUMO} for model 4

Figure 3.9 shows the change of uncertainty in E_{LUMO} with number of chlorine. It is clearly shown that number of chlorine has a slight effect on the uncertainty of E_{LUMO}. For example, the uncertainty of E_{LUMO} will increase from 0.2528 to 0.2543 when N_{Cl} increases from one to six. Figure 3.10 shows the relationship between uncertainty in E_{LUMO} and number of carbon. The uncertainty in E_{LUMO} does not change significantly from 0.2527 to 0.2537 when the number of carbon increases from 1 to 10. Similarly, figure 3.11 indicated the impact of E_{HOMO} on uncertainty of E_{LUMO} follows the same pattern as shown in figures 3.9 and 3.10. E_{HOMO} has obvious effects on the relative error on E_{LUMO}, and the uncertainty of E_{LUMO} will increased from 0.34 to 0.37 when randomly distributed E_{HOMO} values decrease from -0.41 to -0.47.

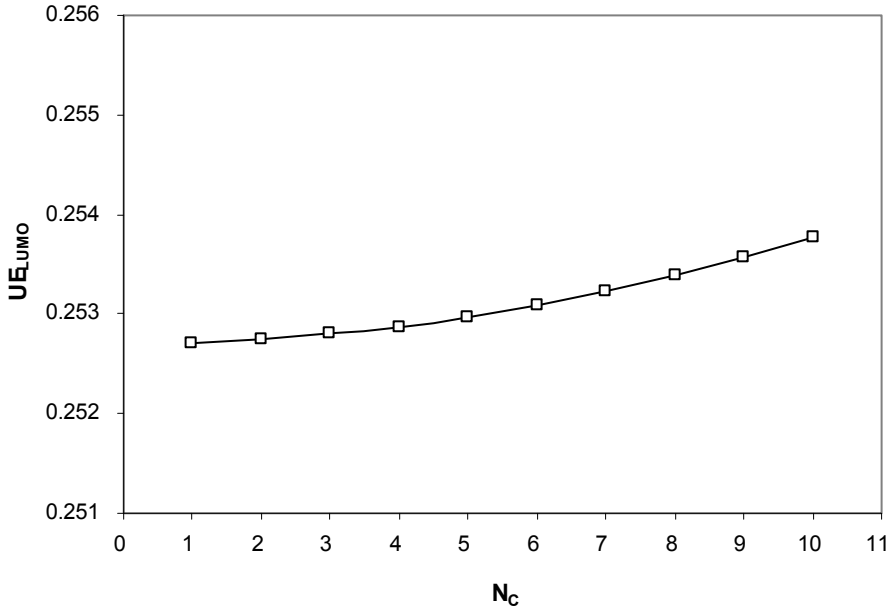


Figure 3.10 Relationship between N_C and uncertainty in E_{LUMO} for model 4

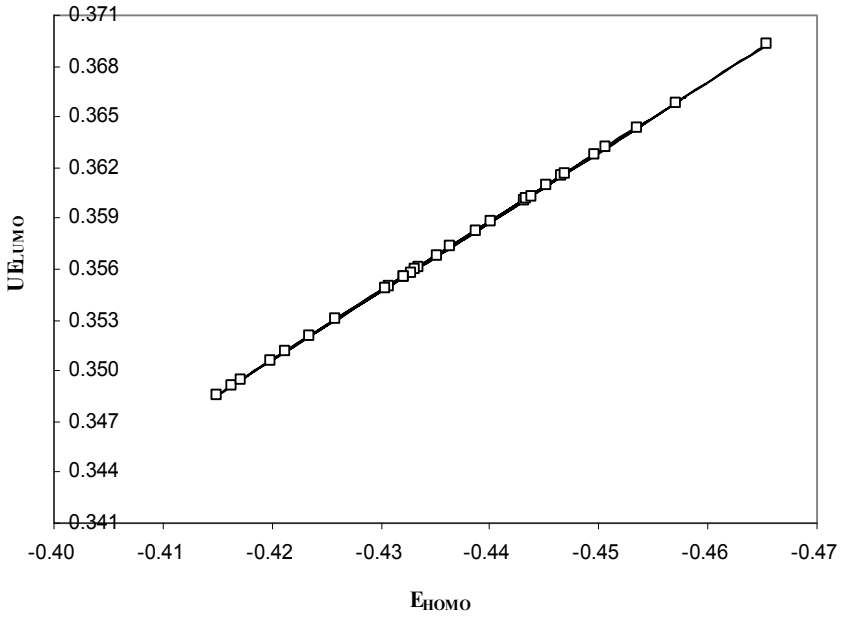


Figure 3.11 Relationship between E_{HOMO} and uncertainty in E_{LUMO} for model 4

3.4 Conclusion

Chlorinated alkanes are an important group of DBPs found in drinking water. Chlorinated alkanes are built from straight chains of carbon and hydrogen with varying numbers of hydrogen atoms replaced by chlorine atoms. The introduction of chlorine atoms into the hydrocarbon chain alters molecular properties such as E_{LUMO}. In this study, the developed QSAR model is applicable to chlorinated alkanes with up to 10 carbon atoms, up to six chlorine atoms, and E_{HOMO} values lying within the range from -0.4667 to -0.4114. E_{LUMO} lies within the range of 0.1098 to 0.2167.

Linear regression methods (MLR, PLS and PCR) can describe the molecular properties and are suitable for prediction (based on Leave-One-Out CV, K-fold CV and external validation results). Limitations and advantages in the use and informational content of QSAR based on simple linear regression and PCR/PLS have been addressed. In all models studied, E_{LUMO} has been shown to correlate highly with number of chlorine and number of carbon in a specific class of DBPs. For a set of 31 chlorinated alkanes, a PLS approach yields better results for building a model from a set of descriptors than the corresponding MLR approach. These better results are reflected in, on average, a better fit of the model to the measured values, as shown by the individual r^2 values, as well as lower RMSE values. These results stress that the most important descriptor is the number of chlorine atoms contained in chlorinated alkanes.

The model validation step also suggests that the most important descriptor for predicting the molecular property of alkane is the number of chlorine. It has been shown that, using the entire data set, N_{Cl}, N_C and E_{HOMO} as descriptors provide a reasonably

good coefficient of determination and RMS value indicating the significance of the developed model.

In summary, model selection and ascertaining the prediction ability of the model are the central tasks in modeling and predicting the problem. Simple molecular descriptors such as number of chlorine and carbon for a given class of DBP can be used to predict molecular properties such as E_{LUMO}.

CHAPTER 4 QSAR STUDY OF CHLORINE EFFECTS ON E_{LUMO} OF CHLORALKENES

Summary

QSAR models predicting molecular property, E_{LUMO} for chlorinated alkene as a subclass of Disinfection By-products were developed due to toxicological interest in risk assessment of DBPs. The QSAR models were statistically validated for predictivity of E_{LUMO}. Molecular descriptors such as N_{Cl}, N_C and E_{HOMO} have been used in order to take into account relevant information provided by molecular features and physicochemical properties. The best model was selected using PLS and MLR which led to models with satisfactory predictive ability for a data set of 15 compounds with E_{LUMO} ranging from 0.0317 to 0.1616. All these models have been statistically validated using both LOO and K-fold CV. The higher r^2_{cv} of cross validations manifest good predictive ability, which demonstrates the practical value of the final QSAR model for screening and priority testing of DBPs. It also examines the uncertainties of the parameters and the models based on conventional methods. These models can be applied to chlorinated alkenes on which toxicity was not tested and even for those not yet synthesized, because theoretical molecular descriptors might be easily and rapidly calculated.

4.1 Introduction

Chlorine substitution in aliphatic compounds results, by its electron attracting effect, in a destabilization in alkanes and the stabilization in alkenes. Thus, in alkenes, the stability of the molecule increases with the number of chlorine substitutions. 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 depend on their structure. Among various properties, E_{LUMO} is of critical importance for describing the ability to gain electrons from other sources.

The objective of this study is to develop QSAR models for the prediction of E_{LUMO} of chlorinated alkenes using new externally predictive MLR. Models have been developed including N_{Cl} as a molecular descriptor together with other theoretical descriptors, such as N_C and E_{HOMO}. The other model is to test the robustness of the obtained model through some statistical methods. A model without the uncertainty test would be confined in practical prediction, and it cannot be used to interpret the toxicity behavior with known uncertainty.

4.2 Data Set and Material

4.2.1 Theoretical Background

As defined by Pearson (1986), the operational definitions of chemical potential, μ, and absolute hardness, η, are:

$$\mu = -\frac{(I + A)}{2} \tag{4.1}$$

$$\eta = \frac{I - A}{2} \tag{4.2}$$

where I and A are the vertical ionization potential and electron affinity of any chemical system, atom, ion, molecule, or radical. "I" is the change of energy when an electron is removed from the system, while "A" is the variation of the energy when an electron is added to the system (Iczkowski and Margrave, 1961).

Within the validity of Koopmans' theorem (1934), the frontier orbital energies are given by

$$-E_{HOMO} = I \text{ and } -E_{LUMO} = A \tag{4.3}$$

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

4.2.2 Data Set

Data was collected for fifty derivations of chlorinated alkene. From this dataset (table 4.1), 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 to 80% of the compounds in the training set. These compounds differ in the number of chlorine and length of the carbon chain. Regression models are developed for the alkene congeners to predict their E_{LUMO} values; number of chlorine (N_{Cl}), number of carbon (N_C) and E_{HOMO} were used as independent variables. Figure 4.1 shows the structure of alkene compounds.

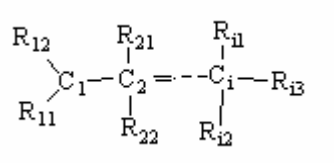


Figure 4.1 Molecular structures of chlorinated alkenes

Table 4.1 Molecular properties 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

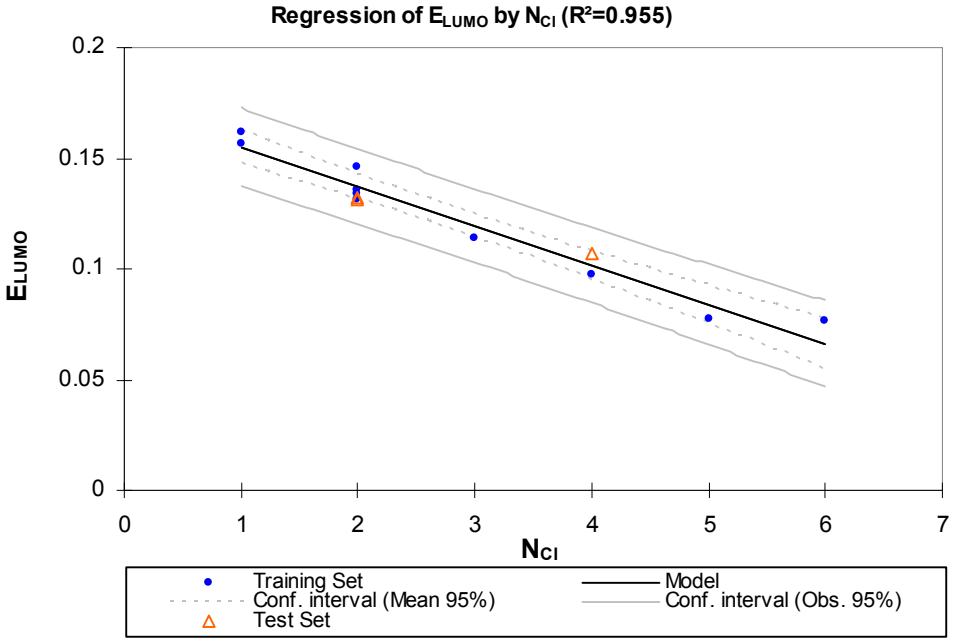
Multiple linear regression analysis and variable selection were performed by the SAS software using the partial least square regression (PLS) and principle component regression (PCR). The acceptable linear models were subjected to a Cross-Validation analysis by Leave-One-Out and K-fold procedures to ensure that the models were not overfitted or underfitted.

4.3 Results and Discussion

4.3.1 Evaluation of Molecular Descriptors

A 1+1 predictor fit, intercept and 1 predictor, is examined first. In model 1, the single response variable is E_{LUMO}, and potential predictor variable is N_{Cl}. $E_{LUMO}=a_1 * N_{Cl} + k_1$

According to leverage plot method, the majority of compounds of the training set are inside of the square area. However, two chemicals (hexachlorocyclohexene and hexachlorocyclopentadiene) have leverages greater than the cutoff value, and their 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 QSAR development processing.



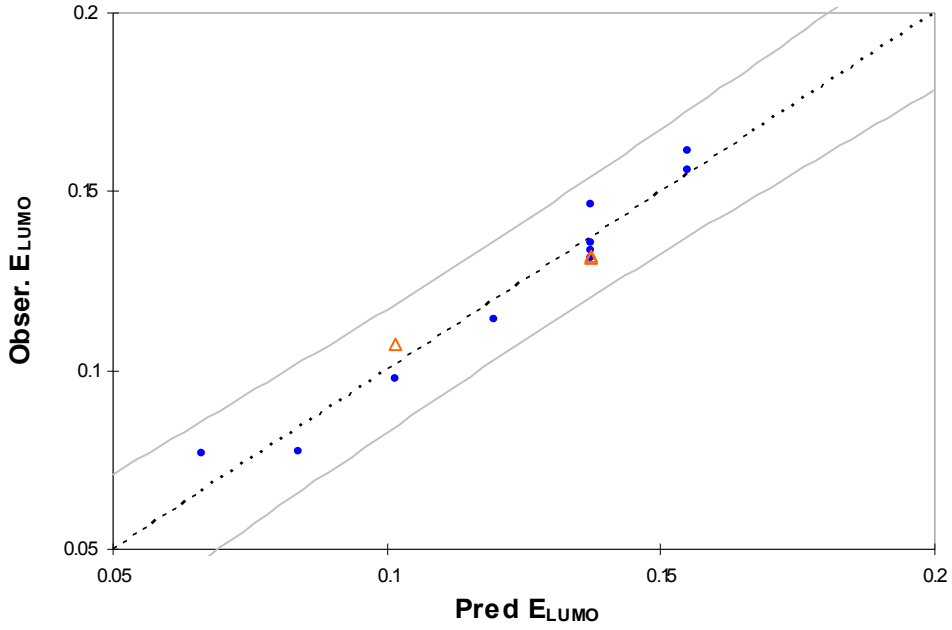


Figure 4.2 (A) The trend of N_{Cl} and E_{LUMO} of model 1, (B) Relationship between observed and predicted alkenes data

The model was rerun on a training set of 10 compounds and validated with 3 external compounds. With linear regression method, a regression equation consisting of coefficients is produced. For chlorinated alkene compounds, figure 4.2A demonstrated that the descriptor, N_{Cl}, is negativly interrelated to E_{LUMO}; that means, E_{LUMO} will decreases as the number of chlorine increases. 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 -0.0206 and the intercept is 0.179. The correlation indicates that, as the number of chlorine increases from 1 to 6 in an alkene compound, E_{LUMO} will decrease 0.13. The probability of the alkene getting a correlation of -0.985 for a sample size of 10 is less than 0.01%. Figure 4.2B shows a plot

of observed E_{LUMO} against fitted values for model 1; the correlation of the scatter is -0.9849.

Linear Regression Model:

E_{LUMO}=-0.02056 N_{Cl}+0.1790 (4.7)

N= 10, R^2 = 0.9701, F= 169.197, RMSE= 0.00485, P= 0.0000

Two-predictor models uses N_{Cl} and N_C to predict E_{LUMO}, which is: $E_{LUMO}=a_2*N_{Cl}+b_2*N_C+k_2$

For model 2, it is important to note that chemical, hexachlorocyclopentadiene (compound 11, table 4.1), can be identified as the outlier with the standardized residual value greater than the cutoff value in the training set. Eq.4.8 indicates that, in the introduction of N_C into the QSAR model, there is a better correlation coefficient than 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 N_{Cl}+0.00592 N_C+0.1641 (4.8)

N= 11, R^2 = 0.9956, F= 102.869, RMSE= 0.00188, P= 0.0000

The next two-predictor models use N_{Cl} and N_C to predict μ and η, respectively. Therefore, the models are: $\mu=a_3*N_{Cl}+b_3*N_C+k_3$ and $\eta=a_4*N_{Cl}+b_4*N_C+k_4$.

The Williams plot verified the presence of an outlier, and this responding outlier can be identified in the training set for model 3 is 1,1-dichloropropene and for model 4 is hexachlorocyclopentadiene, respectively. PLS model for the relationship between μ, η and two descriptors (N_{Cl} and N_C) is shown as follows:

μ= -0.0101*N_{Cl}+ 0.00053*N_C-0.0961 (4.9)

N= 11, R^2 = 0.96772, F= 119.907, RMSE= 0.00373, P= 0.0000

η= -0.00995*N_{Cl}+ 0.00494*N_C+0.2592 (4.10)

N= 11, R^2 = 0.9268, F= 50.661, RMSE= 0.0043, P= 0.0000

4.3.2 Development of QSAR Model

Activities of molecules in the biological systems are highly influenced by their inherent electronic properties. Hence, E_{HUMO} energy along with N_{Cl} and N_C were selected as molecular descriptors to predict E_{LUMO}. Using all three descriptors in the QSAR model, for the combined set of 12 alkene congeners, maximum values for the coefficient of correlation and lowest root mean square errors were obtained. Full Model with 3 predictors is reported as: $E_{LUMO}= a_5 * N_{Cl}+b_5 * N_C +c_5 * E_{HOMO}+ k_5$

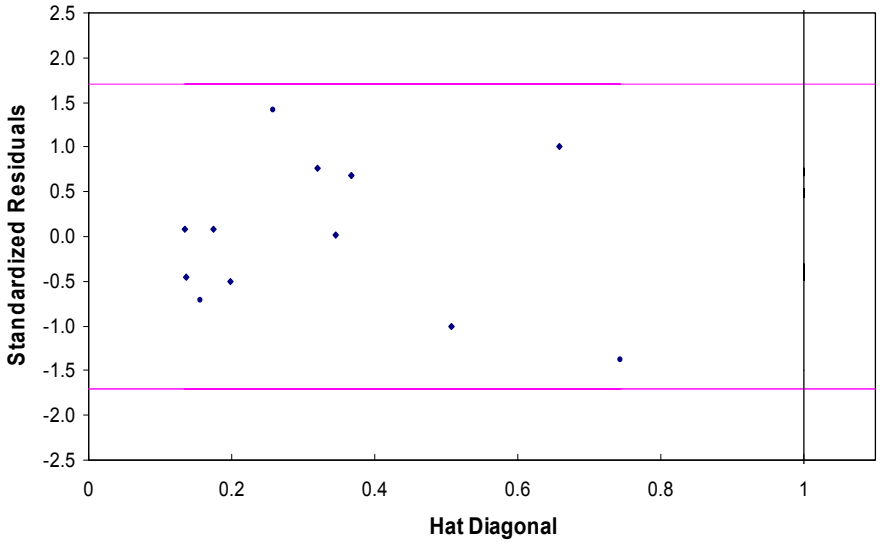


Figure 4.3 Outlier detection of model 5 for alkenes

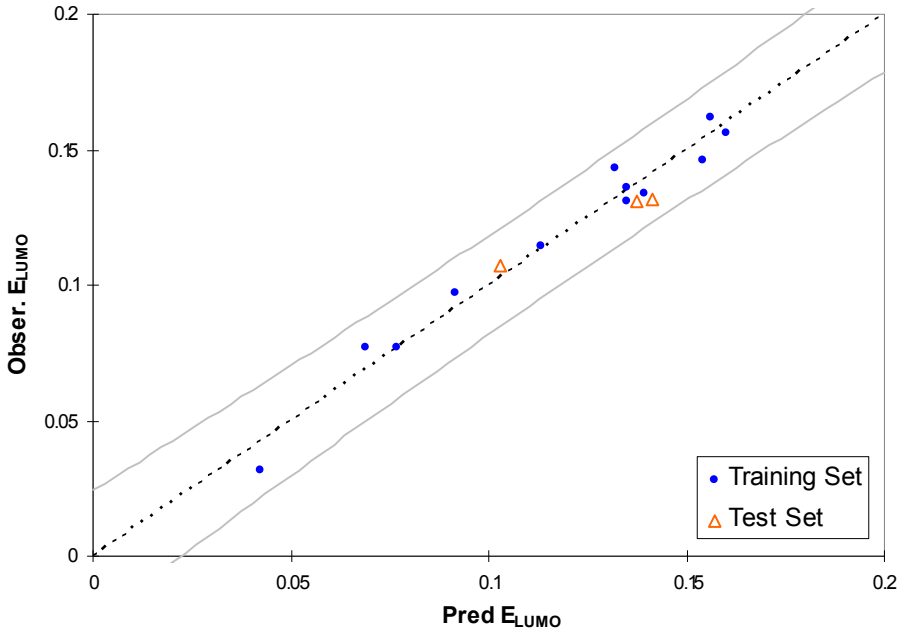
Figure 4.3 explains that neither statistical nor obvious visual outliers were observed. With straightforward MLR, a regression equation consisting of coefficients is produced.

Table 4.2 presents the correlation matrix, where it is clear that the three selected descriptors are not highly correlated, while the descriptors (N_{Cl} and N_C) are correlated 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. 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 4.2 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**

Figure 4.4A starts with repeating observed E_{LUMO} versus fitted E_{LUMO} for the full model. The regression coefficients are plotted in figure 4.4B. In fact, since molecular descriptors do not have equal variance, their relative importance in the model is measured better than standardized regression coefficients.



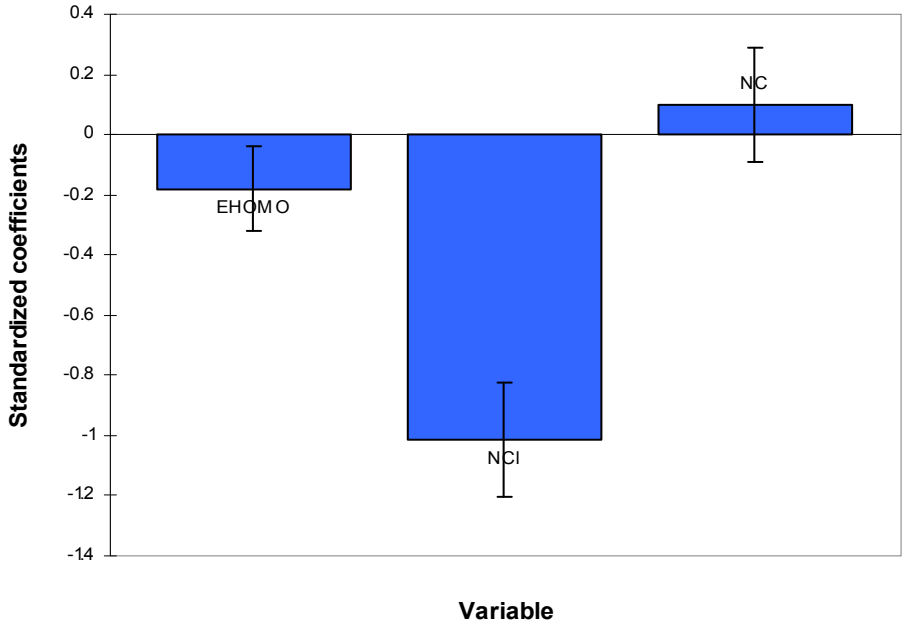
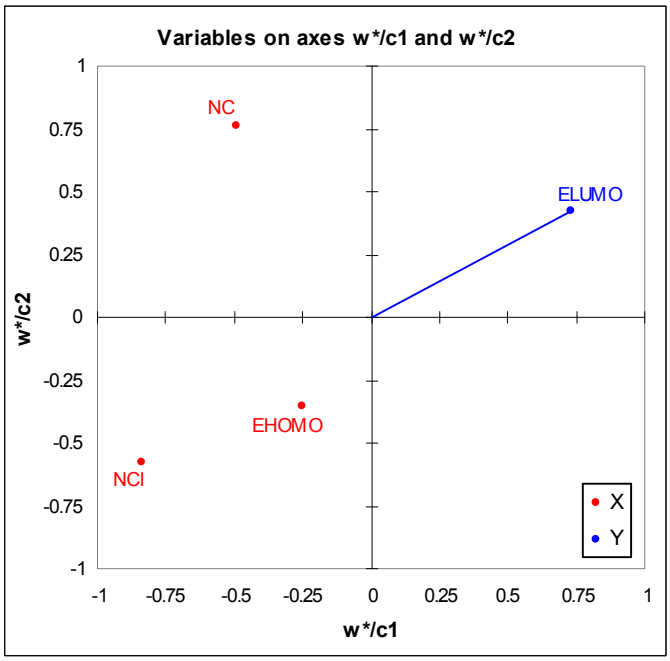


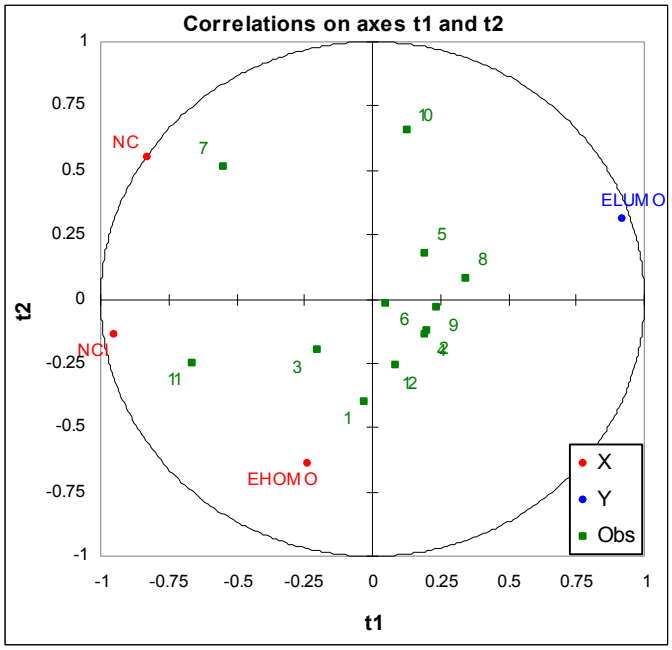
Figure 4.4 (A) Relationship between observed and predicted data for model 5, (B) Regression coefficients of scaled and centered variables.

The other interpretation is very interesting compared to other MLR alternatives. PLS regression makes it possible to calculate the applicability domain of a QSAR model. Figure 4.5A indicates that all X-variables such as E_{HOMO}, N_{Cl} and N_C load strongly in the model and are 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 $Q^2Y=0.700$, which are excellent performance statistics considering that the response is handled simultaneously.

Figure 4.5B 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 (figure 4.5C) 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 new samples. Altogether, hexachlorocyclohexene is the highest toxic compound because it contains the highest number of chlorines and carbons.





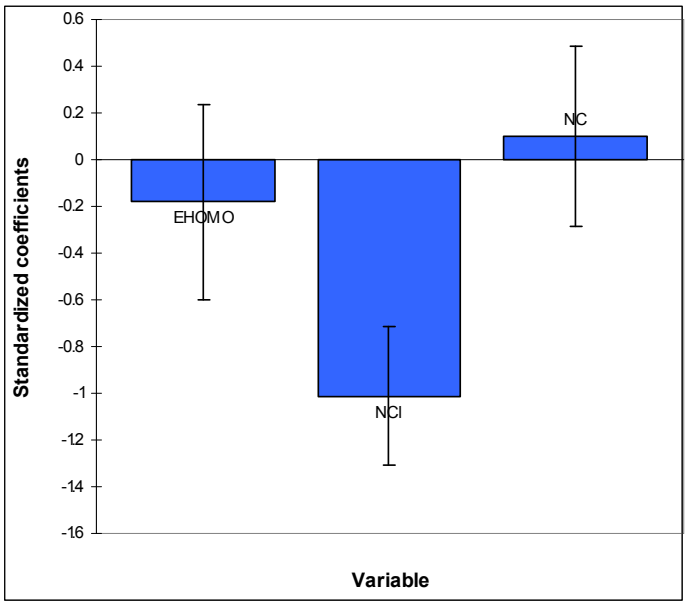


Figure 4.5 (A) PLS loading plot, (B) PLS scores plot, (C) PLS coefficients plot

Another suitable multivariate technique is Principal Component Regression (PCR). PCR is a reduced regression that uses derived inputs, based on principal components, of less than or equal dimension than the original inputs. Figure 4.6 shows the combined plot of scores and loadings in the space defined by the first two principle components (PC_1: N_{Cl}; 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 (PC1 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 with the most informative principal component PC1, 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 significantly positive relative to N_{Cl}, and compound 7 is strongly correlated with N_C.

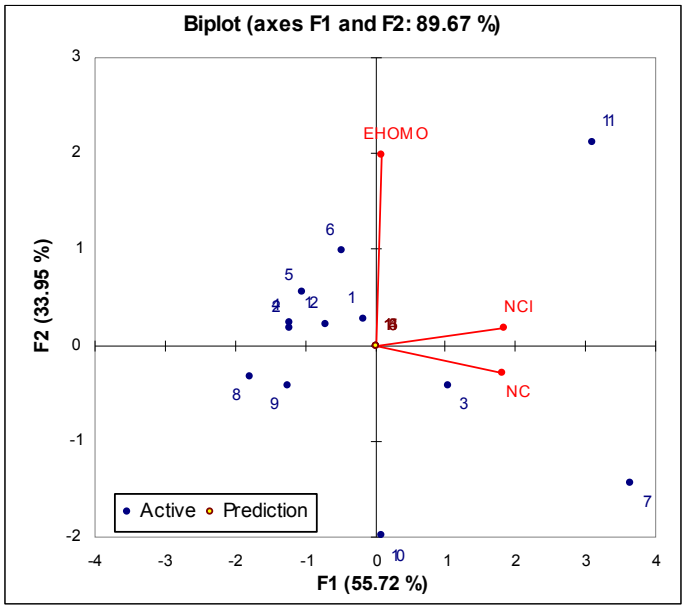


Figure 4.6 Bioplot of F_1 (55.72%) vs. F_2 (33.95%)

Table 4.3 Result comparison for model 5 using three calibration methods

Model calibration methods	Model 5	R^2	RMSE
MLR	E_{LUMO}=-0.02187 N_{Cl}+0.002896 N_C-0.5701 E_{HOMO}-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 4.3 compares the quality of MLR, PLS and PCR models as represented by r^2 and RMSE. It shows that the number of variables is the same for PLS and MLR, but the former method shows a lower RMSE (0.00638 verses 0.00781) compared to the latter one. Similar improvements can be seen for 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 N_{Cl}, N_C, and E_{HOMO} as independent variables are presented in table 4.4. Table 4.5 gives an overview of the PLS models that have the lowest RMSE values for the endpoint E_{LUMO}, μ and η. According to the rule that R^2 should be greater than 0.6 in a good model, all models in table 4.5 are significant and most can be considered as good ones. All QSARs were developed after removing the outliers. The outliers, and possible reasons for these compounds being outliers, are listed in table 4.6.

Table 4.4 Experimental and calculated values of E_{LUMO} for the model 5

No.	Compounds	Descriptors			E_{LUMO} values		
		E_{HOMO}	N_{Cl}	N_C	Calculated	Predicted	Residual
1	tetrachloroethylene	0.0975	4	2	0.09745	0.09154	0.00591
2	cis-1,2-dichloroethylene	0.1358	2	2	0.13581	0.13518	0.00063
3	1,1,2,3,3-pentachloropropene	0.0771	5	3	0.07710	0.07696	0.00014
4	1,2-dichloroethylene	0.1312	2	2	0.13123	0.13474	-0.00350
5	2-chloropropene	0.1616	1	3	0.16159	0.15625	0.00534
6	1,1-dichloropropene	0.1430	2	3	0.14295	0.13183	0.01111
7	hexachlorocyclohexene	0.0768	6	6	0.07676	0.06886	0.00790
8	chloroethylene	0.1562	1	2	0.15621	0.16012	-0.00392
9	1,1-dichloroethylene	0.1337	2	2	0.13369	0.13930	-0.00560
10	3,4-dichloro-1-butene	0.1463	2	4	0.14625	0.15417	-0.00792
11	hexachlorocyclopentadiene	0.0317	6	5	0.03166	0.04244	-0.01078
12	trichloroethylene	0.1142	3	2	0.11417	0.11347	0.00070
13	trans-1,2-dichloroethylene	0.1313	2	3	0.13125	0.13763	-0.00638
14	1,3-dichloropropene	0.1318	2	3	0.13184	0.14139	-0.00955
15	tetrachlorocyclopropene	0.1072	4	3	0.10720	0.10275	0.00445

Table 4.5 Summary of the models for alkenes

Model	Regression Equations	N	R^2	RMSE	Outlier*
1	E_{LUMO}=-0.02056 N_{Cl}+0.1790	10	0.970	0.0049	7, 11
2	E_{LUMO}=-0.0205 N_{Cl}+0.00592 N_C+0.1641	11	0.996	0.0019	11
3	μ=-0.0101 N_{Cl}+ 0.00053 N_C-0.0961	11	0.968	0.0037	6
4	η=-0.00995 N_{Cl}+ 0.0049 N_C+0.2592	11	0.927	0.0043	11
5	E_{LUMO}=-0.0219 N_{Cl}+0.0029 N_C-0.57 E_{HOMO}-0.034	12	0.971	0.0064	-

* Outliers were already been removed before refitting the model

Table 4.6 Outliers and potential reasons for these compounds being outliers

Outliers	Potential reasons for outliers
Outliers to model 1,2 and 4: hexachlorocyclopentadiene	• Two chlorine atoms in a position allylic to two double bonds; • Slightly polar; • Poorly volatile.
Outlier to model 1 alone: hexachlorocyclohexene	• Hydrogen rich surface
Outlier to model 3: 1,1-dichloropropene	• Two chlorine atoms at the end of the double bond in molecule

Some of the possible combinations of parameters were considered. The best equation was selected among other equations by considering the various statistical criteria. For alkenes, E_{HOMO} slightly increases as the number of chlorines increase, and its linear regression model inclines in a horizontal line. E_{HOMO} represents only 0.18% of the variance in the linear regression equation. The slope of this regression model is 0.000297, and the intercept is about -0.3667. Therefore, E_{HOMO} is not a good parameter to predict E_{LUMO} of the chlorinated alkenes. The results also show that there was one best equation using N_{Cl} and N_C as molecular descriptors to predict E_{LUMO}. It is again important to note 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}.

4.3.3 QSAR Model Validation

The robustness of the models and their internal predictive ability was evaluated by cross-validation. A cross-validation method was applied to the original data set for models 1 to 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. r^2 and r^2_{cv} values are good tests 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, independent of the applied splitting (Gramatica and Papa, 2005). The results of LOO, 2-fold, 3-fold, 5-fold, and 10-fold for models 1-5 are reported in table 4.7. The reported validation parameters r^2_{cv} and RMS_{CV}, as expected, indicating that the model has very good descriptive and predictive performances.

Table 4.7 Results of LOO and K-fold Cross-Validation test for alkene

Model No.	d=1 (LOO-CV)		k=2		k=3		k=5		k=10	
	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}	RMS_{CV}	r^2_{cv}
1	0.0048	0.9687	0.0051	0.9550	0.0051	0.9641	0.0051	0.9635	0.0051	0.9636
2	0.0022	0.9956	0.0022	0.9934	0.0020	0.9968	0.0021	0.9963	0.0021	0.9961
3	0.0037	0.9688	0.0021	0.9913	0.0038	0.9706	0.0041	0.9652	0.0039	0.9677
4	0.0043	0.9281	0.0030	0.9635	0.0040	0.9355	0.0044	0.9276	0.0044	0.9238
5	0.0074	0.9747	0.0073	0.9880	0.0054	0.9874	0.0070	0.9783	0.0080	0.9723

Table 4.7 indicates that RMS_{CV} value for model 2 based on the LOOCV is greater than the true RMS by 17.55%, and RMS_{CV} based on K-fold are greater than true RMS in the range 4.43%-15.43%. The difference between RMS_{CV} and true RMS 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 0.04%-1.37%. In practice, true RMS error usually means that RMS was estimated based on the model since the true RMS error is not known.

K-fold CV 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 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.

In summary, for the chlorinated alkene data set, K-fold CV performs better than LOO-CV and the full model with respect to model size, model complexity and, most importantly, predictive power.

4.3.4 Uncertainty Analysis

From the regression equation discussed in sections 4.3.1 and 4.3.2, coefficients of parameters and the standard deviations are estimated in table 4.8 using bootstrap analysis. The bootstrapping is repeated 5000 times for each validated model and gives the following parameter estimates and their standard errors.

The expression of the uncertainty in E_{LUMO} determined from the regression model 5 at a measured or specified value of X is given by equation 4.11. Here, we do not consider the correlated uncertainties between any two of these variables, and the uncertainty of number of chlorine and number of carbon are zero, then all terms involving correlated uncertainties in eq. 4.11 will be simplified as the following equation 4.12.

Table 4.8 Summary of coefficients and the standard deviations for alkenes

Model No.	a	b	c	k	σ_a	σ_b	σ_c	σ_k

1	-0.0207	-	-	0.1792	0.0013	-	-	0.0034
2	-0.0205	0.0060	-	0.1638	0.0007	0.0010	-	0.0025
3	-0.0101	0.0005	-	-0.0960	0.0012	0.0017	-	0.0030
4	-0.0098	0.0047	-	0.2593	0.0011	0.0023	-	0.0052
5	-0.0219	0.0029	-0.5701	-0.0340	0.0018	0.0024	0.1936	0.0706

$$U^2_{E_{LUMO}} = \left(\frac{\partial(E_{LUMO})}{\partial a}\right)^2 U_a^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_{Cl}}\right)^2 U^2_{N_{Cl}} + \left(\frac{\partial(E_{LUMO})}{\partial b}\right)^2 U_b^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_C}\right)^2 U^2_{N_C} + \left(\frac{\partial(E_{LUMO})}{\partial c}\right)^2 U_c^2 + \left(\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})}\right)^2 U^2_{E_{HOMO}} + \left(\frac{\partial(E_{LUMO})}{\partial k}\right)^2 U_k^2 \quad (4.11)$$

where, $\frac{\partial(E_{LUMO})}{\partial a} = N_{Cl}$, $\frac{\partial(E_{LUMO})}{\partial N_{Cl}} = a$, $\frac{\partial(E_{LUMO})}{\partial b} = N_C$, $\frac{\partial(E_{LUMO})}{\partial N_C} = b$, $\frac{\partial(E_{LUMO})}{\partial c} = E_{HOMO}$,

$\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})} = c$, $\frac{\partial(E_{LUMO})}{\partial k} = 1$.

$$U^2_{E_{LUMO}} = N_{Cl}{}^2 U_a^2 + N_C{}^2 U_b^2 + E_{HOMO}{}^2 U_c^2 + c^2 U^2_{E_{HOMO}} + U_k^2 \quad (4.12)$$

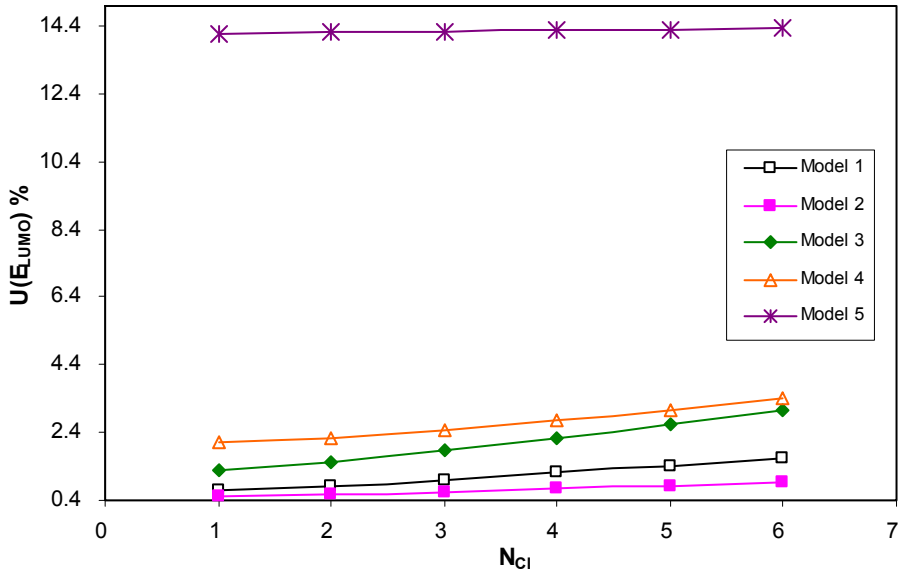


Figure 4.10 Relationship between N_{Cl} and uncertainty in E_{LUMO}

Figure 4.10 shows the change of uncertainty in E_{LUMO} with number of chlorine. It is clearly shown that number of chlorine has a slight 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} increases from one to six for model 1. Figure 4.11 shows the relationship between uncertainty in E_{LUMO} and number of carbon. The uncertainty in E_{LUMO} for model 2-4 does not change significantly from 0.6338 to 2.9347 with the number of carbon increased from 2 to 6. Similarly, figure 4.12 indicated the impact of E_{HOMO} on uncertainty of E_{LUMO} follows the same pattern as shown in figures 10 and 11. E_{HOMO} has obvious effects on the relative error on E_{LUMO}, and the uncertainty of E_{LUMO} for models 2-4 will increase from 1.0304 to 1.4077 when randomly distributed E_{HOMO} values decrease from -0.33 to -0.39. In figures 4.10, 4.11, and 4.12, three variables that affect the uncertainty of E_{LUMO} for each model are following: Model 5 > model 4 > model 3 > model 2.

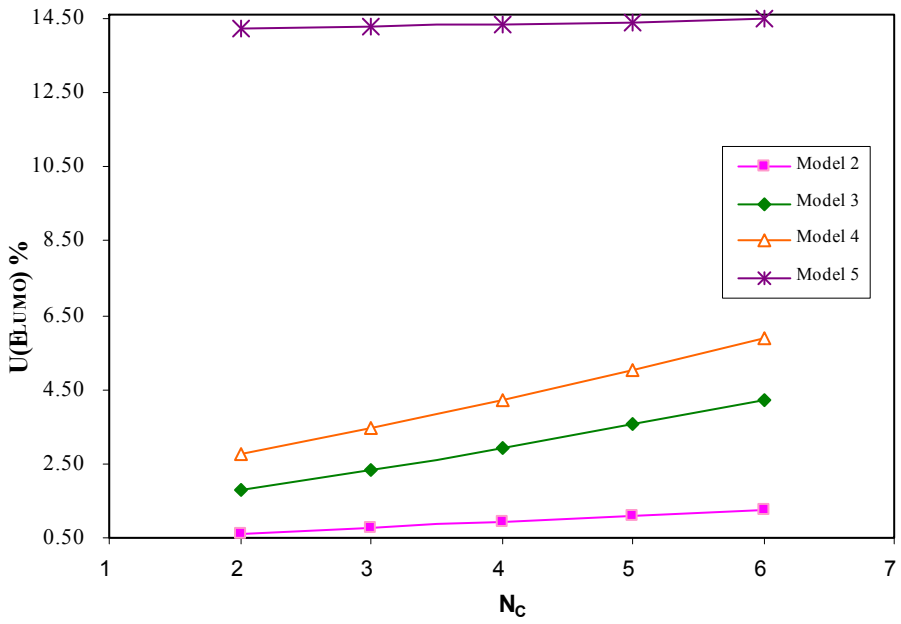


Figure 4.11 Relationship between N_C and uncertainty in E_{LUMO}

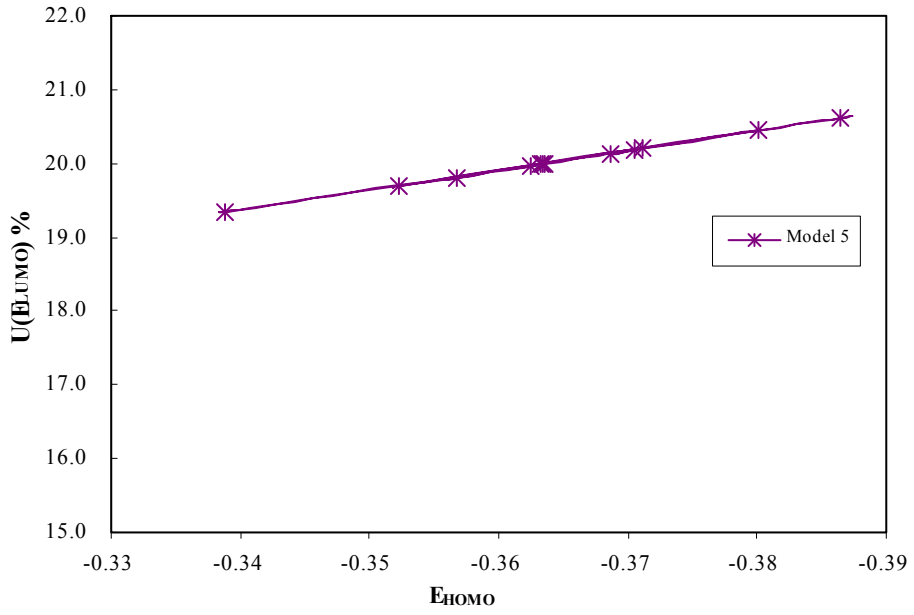


Figure 4.12 Relationship between E_{HOMO} and uncertainty in E_{LUMO}

4.4 Conclusion

From the results and discussion, it can be concluded that the model is homogenous and stable from models 1 to 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}.

The developed QSAR models are demonstrated to ensure the continued applicability of proposed models, and are presented as evidence that different theoretical molecular descriptors with similar chemical interpretability can be useful and interchangeable. It important to note that, 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, $r^2_{5\text{-fold}}$=0.9963), its RMS values ($RMS_{(training\ set)}$=0.0019, $RMS_{(cross\text{-}val.\ set)}$=0.0021) are all smaller than in the 3-descriptor model ($RMS_{(training\ set)}$= 0.0064, $RMS_{(cross\text{-}val.\ set)}$=0.00696). At the same

time, the 3-variables model, obtained by introducing E_{HOMO} into the model (r^2=0.9709, r^2_{LOO}=0.9747, $r^2_{5\text{-}fold}$=0.9783), did not significantly increase the predictive performance of the nested model, as is evident when comparing the internal and external r^2_{cv} values.

In summary, model selection, ascertaining the prediction ability and uncertainty of the model, are the central tasks in modeling and predicting problems. Simple molecular descriptors such as N_{Cl}, N_C, and E_{HOMO} for a given class of alkene can be used to predict molecular properties such as E_{LUMO}.

CHAPTER 5 QSAR STUDY OF CHLORINE EFFECTS ON E_{LUMO} OF CHLORAROMATICS

Summary

The energy of the lowest unoccupied molecular orbital (E_{LUMO}) is an important property of various chlorinated compounds for predicting toxicity. QSAR analysis was developed for 53 chlorinated aromatic compounds, including chloro- phenols, anilines, and benzenes, using the conceptual density functional theory based global reactivity parameter such as N_{Cl} along with E_{HOMO} as descriptors. The one-descriptor (N_{Cl}) and two-descriptor (N_{Cl} and E_{HOMO}) QSAR models were developed to predict E_{LUMO} for each subset. The six equations were found to fit well. After the variable selection step, MLR with LOO and K-fold Cross-Validation were used for building and validating the QSAR models. The cross-validated R^2_{CV} values for the ideal QSAR models is in the range between 0.9659 and 0.995, indicating a good predictive capability for E_{LUMO} values for chlorinated aromatics. The QSAR results show that the main factor affecting E_{LUMO} values is the number of chlorine.

5.1 Introduction

Chlorinated phenols, anilines, and benzenes belong to chlorinated aromatic DBPs. There have been many reports on the prediction of E_{LUMO} for chlorinated aromatics, because the compounds have serious, harmful, ecological effects and are implicated as potential carcinogens (Safe, 1990). E_{LUMO} is an important property to predict toxicity of chlorinated aromatic DBPs. Several methods have been described in the literature for the estimation of E_{LUMO}. It has been reported that number of chlorine is an important descriptor to express the mechanism of toxicity for CP, CB and CA molecules. It is expected that these properties may explain the toxicities of the compounds. The aim of this study is to develop QSAR model to predict E_{LUMO} of 22 chorophenols, 15 chloroanilines, and 16 chlorobenzenes based on number of chlorine substituents (N_{Cl}) and the highest occupied molecular orbital (E_{HOMO}) energy as descriptors.

5.2 Data Set

Three well-studied data sets, namely, chloro- phenol, aniline, and benzene were used. A structural highly heterogeneous data set of 53 compounds with calculated E_{LUMO} and E_{HOMO} values for chlorinated aromatic was organized in the Excel datasheet. The data set covers a wide range of E_{LUMO} and E_{HOMO} values (E_{LUMO} ranging from 0.0684 to 0.134; E_{HOMO} ranges from -0.361 to -0.292) with number of chlorine ranging from 1 to 6, and is highly representative of DBPs in these classes.

QSAR studies were carried out for the chlorinated phenols, anilines and benzenes in table 5.1, 5.2 and 5.3. Regression models are developed for the chlorinated aromatic

congeners with calculated E_{LUMO}; N_{Cl} and E_{HOMO} were taken as independent variables. Each subset was divided into training and prediction sets.

5.3 Results and Discussion

Linear and multiple linear regression analysis were performed by software SAS using the forward stepwise regression method. The robustness and internal predictivity of the models was firstly evaluated by both Leave-One-Out (R^2_{LOO}) and K-fold ($R^2_{k\text{-}fold}$) CV. In the last procedure, K n-dimensional groups are generated by a ramdomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample and then R^2 is calculated for each model. The proposed models are also checked for the descriptive, predictive and modeling powers.

5.3.1 Model Selection and Validation

5.3.1.1 Chlorinated Phenols

Modeling of chlorinated phenols (data set P) comprises a total of 22 compounds, structurally highly heterogeneous, with a range of E_{LUMO} between 0.072 and 0.13 with the molecular properties parameters, which shows a rather good result. Determining the right form of the model in order to reduce the model mismatch errors is accomplished during the model construction phase, whereas determining the correct model parameters can be achieved at the model selection and validation phase. For the 17 tested phenol compounds, the following correlation equations were established:

$$E_{LUMO}=-0.0137\ N_{Cl}+0.136 \qquad (5.1)$$

($N_{training}$=17, $N_{pred.}$=5, R^2=0.9705, $RMSE_{training}$=0.00278, R^2_{LOO}=0.9709, $R^2_{5\text{-}fold}$=0.9659, F=493.751, P=0.0001)

Table 5.1 Observed, predicted and residual values of 22 phenol compounds

No.	Compounds	Descriptors			eq. 5.1		eq. 5.2	
		N_{Cl}	E_{HOMO}	E_{LUMO}	Predicted	Residual	Predicted	Residual
1	3-chlorophenol	1	-0.3280	0.1219	0.1223	-0.0004	0.1200	0.0019
2	4-chlorophenol	1	-0.3168	0.1238	0.1223	0.0015	0.1226	0.0012
3	2,4-dichlorophenol	2	-0.3254	0.1073	0.1085	-0.0012	0.1090	-0.0017
4	2,5-dichlorophenol	2	-0.3343	0.1060	0.1085	-0.0025	0.1069	-0.0009
5	2,6-dichlorophenol	2	-0.3346	0.1065	0.1085	-0.0020	0.1068	-0.0003
6	2,3-dichlorophenol	2	-0.3338	0.1083	0.1085	-0.0003	0.1070	0.0013
7	3,4-dichlorophenol	2	-0.3269	0.1088	0.1085	0.0002	0.1086	0.0001
8	3,5-dichlorophenol	2	-0.3395	0.1055	0.1085	-0.0030	0.1057	-0.0002
9	2,3,4-trichlorophenol	3	-0.3329	0.0961	0.0948	0.0013	0.0957	0.0005
10	2,3,5-trichlorophenol	3	-0.3445	0.0930	0.0948	-0.0018	0.0929	0.0001
11	2,3,6-trichlorophenol	3	-0.3376	0.0928	0.0948	-0.0020	0.0946	-0.0017
12	4-chloro-3-methylphenol	1	-0.3131	0.1300	0.1223	0.0077	0.0950	-0.0024
13	2,4,6-trichlorophenol	3	-0.3355	0.0926	0.0948	-0.0022	0.0949	-0.0001
14	3,4,5-trichlorophenol	3	-0.3360	0.0948	0.0948	0.0000	0.0806	0.0003
15	2,3,5,6-tetrachlorophenol	4	-0.3476	0.0809	0.0811	-0.0002	0.0820	-0.0003
16	2,3,4,6-tetrachlorophenol	4	-0.3416	0.0818	0.0811	0.0007	0.0694	0.0022
17	pentachlorophenol	5	-0.3461	0.0716	0.0674	0.0042	0.1200	0.0019
18	2-chlorophenol	1	-0.3280	0.1219	0.1223	0.00004	0.1212	0.0011
19	2,4,6-tribromophenol	3	-0.3168	0.1238	0.0948	0.0015	0.0973	-0.0010
20	2-bromo-4,6-dichlorophenol	2	-0.3254	0.1073	0.1085	-0.0150	0.1074	-0.0139
21	2,3,4,5-tetrachlorophenol	4	-0.3343	0.1060	0.0811	0.0019	0.0820	0.0010
22	2,4,5-trichlorophenol	3	-0.3346	0.1065	0.0948	-0.0010	0.0952	-0.0014

For model 5.1, 2,3,4,6-tetrachlorophenol can be tested have a leverage greater than h*(=0.353), but standard deviation values within the 2σ limit, which implies that it is not to be considered as an outlier but an influential chemical. Figure 5.1A shows the relationship between the number of chlorine and E_{LUMO}. For model 5.1, E_{LUMO} decreases as the number of chlorine increases. The probability of getting a correlation of -0.9646 for a sample size of 17 is less than 0.01%. The correlation indicates that, as the number of chlorine increases from 1 to 5 in a phenol compound, E_{LUMO} will decrease 0.0585. The

plot of the observed versus predicted E_{LUMO} values for the above presented model is shown in figure 5.1 B. It shows that these two E_{LUMO} values give a good correlation coefficient (r) of 0.9851. The low residual values reveal the importance of the selected descriptors in QSAR analysis on DBPs. In figure 5.1 A and B, the compound 20, 2-bromo-4,6-dichlorophenol, in test set is out of the critical line.

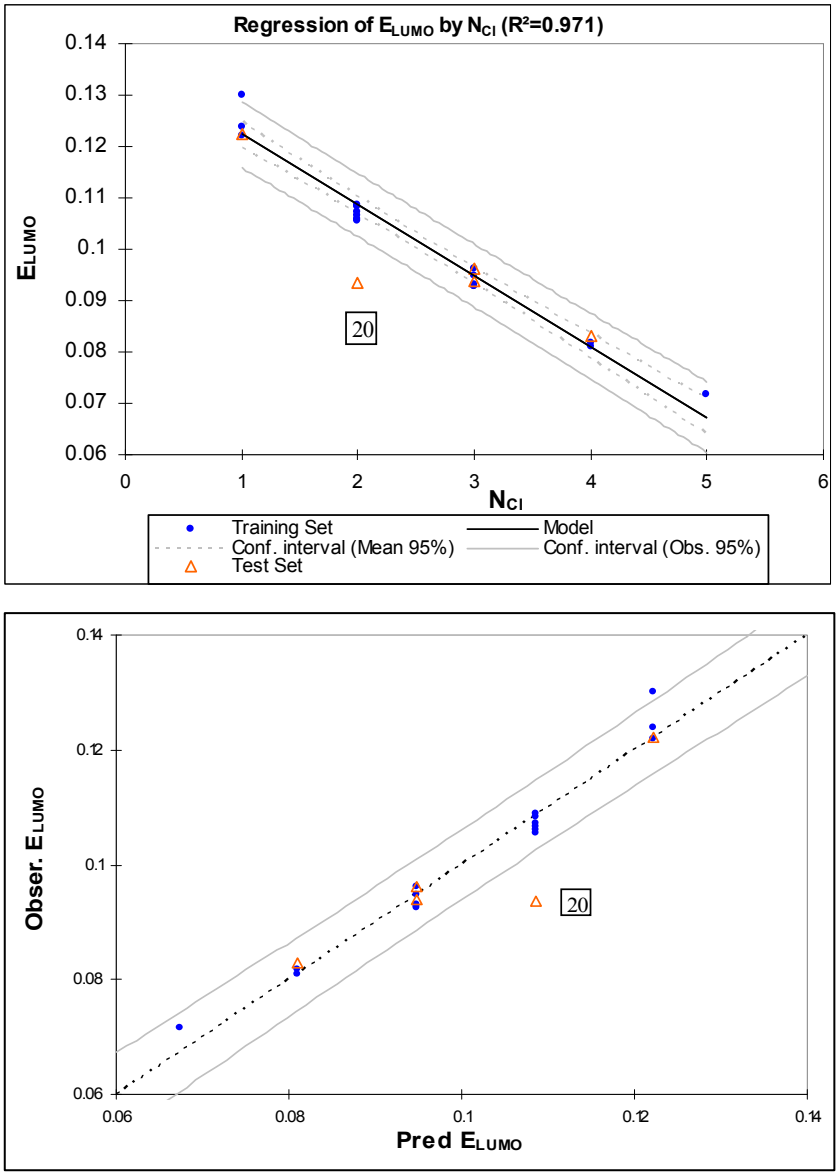


Figure 5.1 (A) N_{Cl} as descriptor for predicting E_{LUMO} in model 5.1, (B) Relationship between observed E_{LUMO} and predicted E_{LUMO} values

The next two-predictor model use E_{HOMO} and the number of chlorine as the independent variables; E_{LUMO} acted as the dependent variable. For model 5.2, 4-chloro-3-methylphenol can be identified as the outlier, and it should be noted that this compound has the smallest energy gap. A training set of 16 compounds was then refitted and shown in the following eq. 5.2:

$$E_{LUMO} = -0.01158\ N_{Cl} + 0.2356\ E_{HOMO} + 0.2088 \quad (5.2)$$

($N_{training}$=16, $N_{pred.}$=5, R^2=0.9919, R^2_{LOO}=0.9920, $R^2_{5\text{-}fold}$=0.9921, $RMSE_{training}$=0.00124, F=799.316, P=0.0001)

5.3.1.2 Chlorinated Anilines

For the 12 tested aniline compounds, the following correlation equations were established:

$$E_{LUMO} = -0.01279\ N_{Cl} + 0.1439 \quad (5.3)$$

($N_{training}$=12, $N_{pred.}$=5, R^2=0.9883, $RMSE_{training}$=0.00181, $R^2_{5\text{-}fold}$=0.9862, F=841.997, P<0.0001)

For model 5.3, outlier detection process shows that all compounds of the training set are inside of the area, and the relationship between N_{Cl} and E_{LUMO} are plotted in figure 5.2A, which shows that the descriptor N_{Cl} is negatively interrelated to the E_{LUMO}. For anilines, E_{LUMO} represents 98.83% of the variance in the linear regression equation, and the probability of getting a correlation of -0.995 for a sample size of 12 is less than 0.01%. The correlation indicates that, as the number of chlorine increases from 1 to 5 in an aniline compound, E_{LUMO} will decrease 0.052. Figure 5.2B shows that E_{LUMO} values were calculated by Spartan and the E_{LUMO} values give a good correlation coefficient of 0.995.

$$E_{LUMO} = -0.01091\ N_{Cl} + 0.3301\ E_{HOMO} + 0.2410 \quad (5.4)$$

($N_{training}$=12, $N_{pred.}$=5, R^2=0.9955, R^2_{LOO}=0.9955, $R^2_{5\text{-}fold}$=0.995, $RMSE_{training}$=0.00119, F=984.03, P<0.0001)

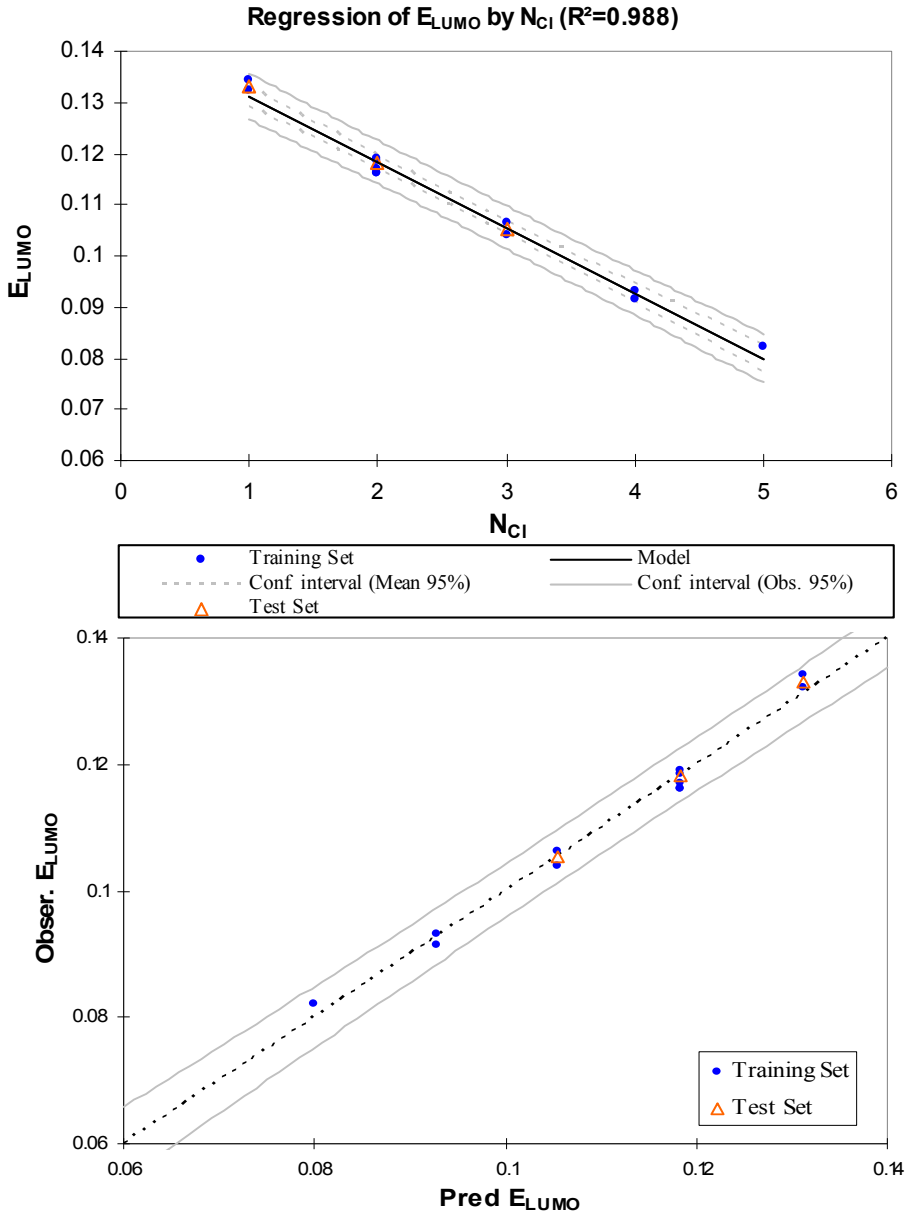


Figure 5.2 (A) N_{Cl} as descriptor for predicting E_{LUMO} in model 5.3. (B) Relationship between observed and predicted endpoint data

The next two-predictor model (eq. 5.4) uses N_{Cl} and E_{HOMO} to predict E_{LUMO} employed by the multi-linear PLS method. For the interpretation of this QSAR model, we may

consider the model coefficients (scores and loadings) to see how the compounds and the X- and Y- variables are interrelated.

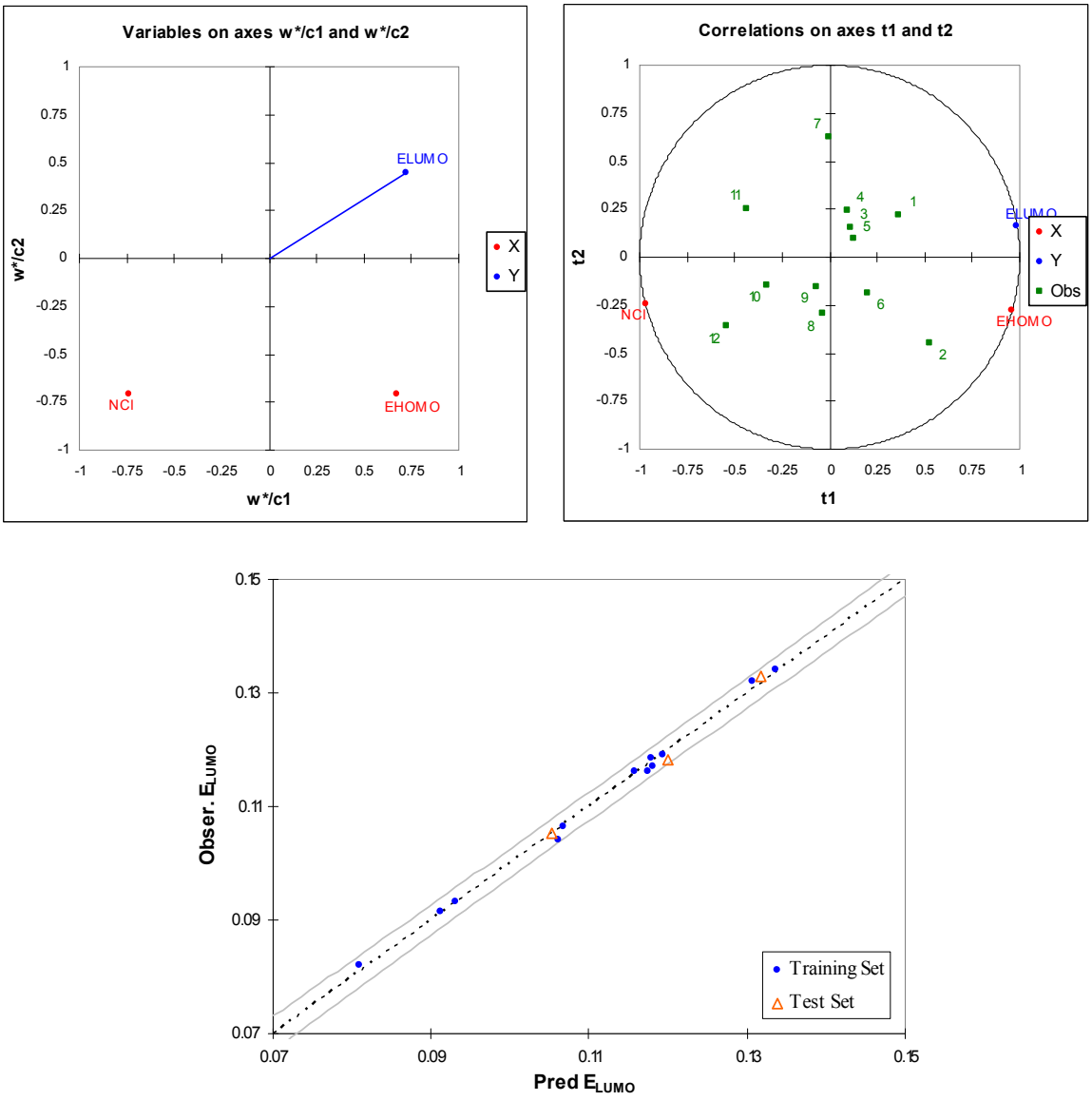


Figure 5.3 (A) PLS loading plot for equation 5.4, (B) PLS scores plot, (C) Observed E_{LUMO} vs. predicted E_{LUMO}.

Figure 5.3A indicates that all X-variables load strongly in the model, and that N_{Cl} and E_{LUMO} are closely related. Overall, N_{Cl} is the most important X-variable. Altogether,

pentachloroaniline is the highest toxic compound to these aniline organisms, and it also has highest number of chlorine. Figure 5.3 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. In figure 5.3, the data analysis resulted in a QSAR with R^2X=0.93, R^2Y=0.94, and Q^2Y=0.84, which are excellent performance statistics considering that four responses are handled simultaneously.

Table 5.2 Observed, predicted and residual values of 15 aniline compounds

No.	Compounds	Descriptors			eq. 5.3		eq. 5.4	
		N_{Cl}	E_{HOMO}	E_{LUMO}	Predicted	Residual	Predicted	Residual
1	3-chloroaniline	1	-0.3014	0.1322	0.1311	0.0010	0.1306	0.0016
2	4-chloroaniline	1	-0.2924	0.1342	0.1311	0.0031	0.1336	0.0007
3	2,3-dichloroaniline	2	-0.3071	0.1186	0.1184	0.0002	0.1178	0.0008
4	2,5-dichloroaniline	2	-0.3082	0.1161	0.1184	-0.0022	0.1174	-0.0013
5	2,6-dichloroaniline	2	-0.3062	0.1171	0.1184	-0.0013	0.1181	-0.0010
6	3,4-dichloroaniline	2	-0.3025	0.1192	0.1184	0.0008	0.1194	-0.0002
7	3,5-dichloroaniline	2	-0.3134	0.1160	0.1184	-0.0023	0.1157	0.0003
8	2,3,4-trichloroaniline	3	-0.3076	0.1064	0.1056	0.0008	0.1068	-0.0004
9	2,4,5-trichloroaniline	3	-0.3094	0.1040	0.1056	-0.0016	0.1061	-0.0022
10	2,3,4,5-tetrachloroaniline	4	-0.3161	0.0933	0.0928	0.0005	0.0930	0.0003
11	2,3,5,6-tetrachloroaniline	4	-0.3214	0.0915	0.0928	-0.0013	0.0913	0.0003
12	pentachloroaniline	5	-0.3198	0.0821	0.0800	0.0021	0.0809	0.0012
13	2-chloroaniline	1	-0.2980	0.1331	0.1311	0.0019	0.1317	0.0013
14	2,4-dichloroaniline	2	-0.3007	0.1181	0.1184	-0.0002	0.1199	-0.0018
15	3,4,5-trichloroaniline	3	-0.3119	0.1054	0.1056	-0.0002	0.1053	0.00004

5.3.1.3 Chlorinated Benzenes

For the 12 tested benzene compound in model 5.5, one compound (4-chloroaniline) is considered as an outlier. The refitting correlation equation for number of chlorine and E_{LUMO} was established:

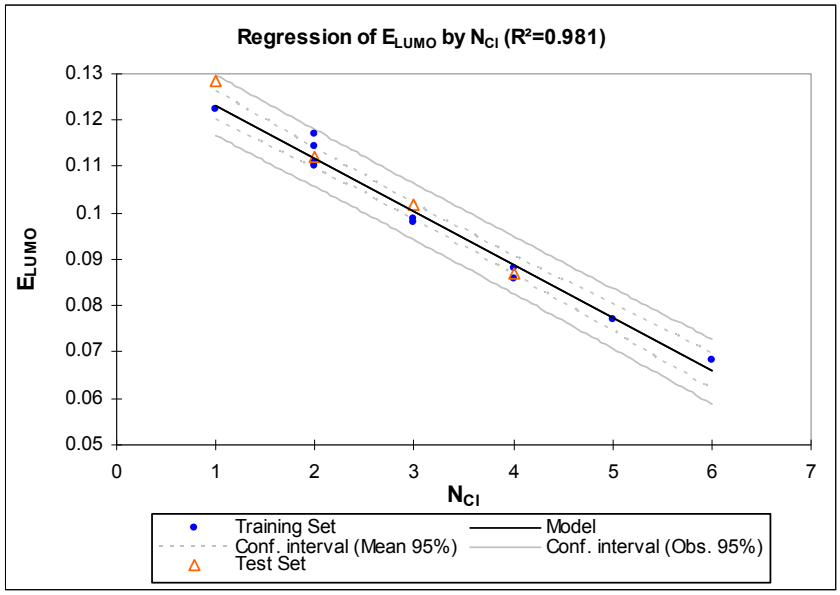
$$E_{LUMO}=-0.01148\ N_{Cl}+0.1347 \qquad (5.5)$$

($N_{training}$=11, $N_{pred.}$=4, R^2=0.981, $RMSE_{training}$=0.0026, R^2_{LOO}=0.981, $R^2_{5\text{-}fold}$=0.977, F=406.295, P<0.0001)

The relationship between number of chlorine and E_{LUMO} are plotted in figure 5.4. For benzenes, the probability of getting a correlation of -0.9904 for a sample size of 11 is less than 0.01%. In figure 5.4 A, we can see the correlation indicates that, as the number of chlorine increases from 1 to 6 in a benzene compound, that E_{LUMO} will decrease 0.054. E_{LUMO} represents 98.08% of the variance in the linear regression equation. It has been inferred from figure 5.4 B that the direct correlation analyses are carried out between the E_{LUMO} values which are calculated by Spartan and the E_{LUMO} values are predicted by equation. A plot between these two E_{LUMO} values gives a good correlation coefficient (r) of -0.9904.

$$E_{LUMO} = -0.0116\ N_{Cl} - 0.02037\ E_{HOMO} + 0.1281 \qquad (5.6)$$

($N_{training}$=11, $N_{pred.}$=4, R^2=0.9809, R^2_{LOO}=0.9851, $R^2_{5\text{-}fold}$=0.9772, $RMSE_{training}$=0.00231, F=205.644, P=0.0001)



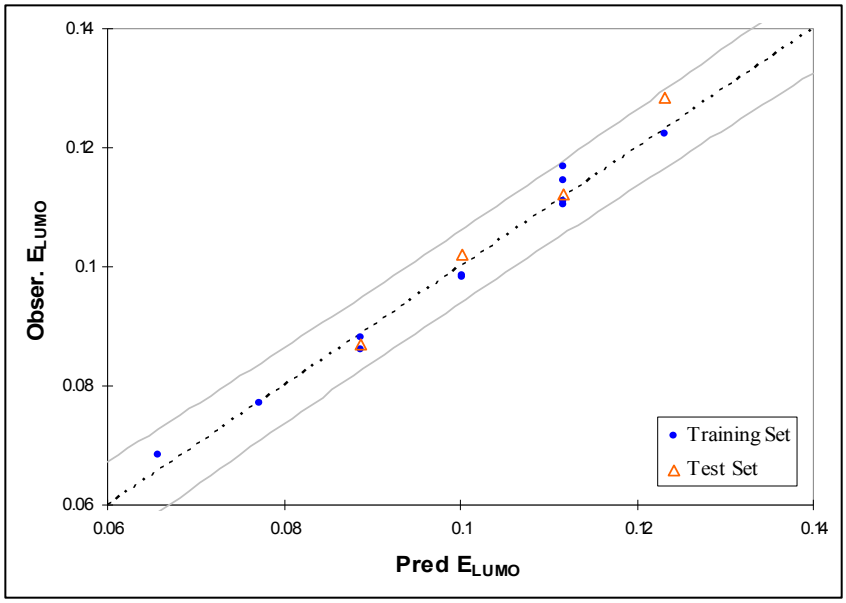


Figure 5.4 (A) N_{Cl} as descriptor for predicting E_{LUMO} in model 5.5. (B) Relationship between observed and predicted endpoint data

Table 5.3 Observed, predicted and residual values of 16 benzene compounds

No.	Compounds	Descriptors			eq. 5.5		eq. 5.6	
		N_{Cl}	E_{HOMO}	E_{LUMO}	Predicted	Residual	Predicted	Residual
1	benzyl chloride	1	-0.3407	0.1224	0.1232	-0.0008	0.1234	-0.0010
2	(2-chloroethyl)-benzene	1	-0.3091	0.0939	0.1118	0.0027	0.1118	0.0026
3	1,2-dichlorobenzene	2	-0.3422	0.1144	0.1118	-0.0014	0.1117	-0.0013
4	1,4-dichlorobenzene	2	-0.3363	0.1104	0.1118	-0.0008	0.1113	-0.0003
5	4,6-dichloro-1,3-benzenediol	2	-0.3170	0.1110	0.1118	0.0051	0.1118	0.0051
6	dichloroethylbenzene	2	-0.3397	0.1169	0.1003	-0.0017	0.1005	-0.0020
7	1,3,5-trichlorobenzene	3	-0.3574	0.0986	0.1003	-0.0021	0.1003	-0.0021
8	1,2,4-trichlorobenzene	3	-0.3438	0.0981	0.0888	-0.0006	0.0888	-0.0006
9	1,2,3,4-tetrachlorobenzene	4	-0.3499	0.0882	0.0888	-0.0028	0.0887	-0.0027
10	1,2,4,5-tetrachlorobenzene	4	-0.3476	0.0860	0.0773	-0.0002	0.0773	-0.0002
11	pentachlorobenzene	5	-0.3546	0.0771	0.0658	0.0026	0.0658	0.0026
12	hexachlorobenzene	6	-0.3614	0.0684	0.1232	-0.0008	0.1234	-0.0010
13	chlorobenzene	1	-0.3365	0.1284	0.1232	0.0052	0.1233	0.0051
14	1,3-dichlorobenzene	2	-0.3440	0.1123	0.1118	0.0005	0.1119	0.0004
15	1,2,3-trichlorobenzene	3	-0.3522	0.1019	0.1003	0.0016	0.1004	0.0015
16	1,2,3,5-tetrachlorobenzene	4	-0.3525	0.0869	0.0888	-0.0018	0.0888	-0.0019

Considering the relatively wide range of physicochemical properties for the selected benzene, equation 5.6 indicates that E_{LUMO} correlates with quantum chemical descriptors very well. The observed values, fitted values and the residuals for the selected E_{LUMO} are also presented in table 5.3. Furthermore, the t-test method was used to test the correlation of each independent variable, and the student t-values for partial correlation coefficients in equation 5.6 are -14.431 and -0.202 for N_{Cl} and E_{HOMO}, respectively. This indicates that the number of chlorine is the most important factor for E_{LUMO} prediction. Considering Eqs. 5.5 and 5.6, the higher N_{Cl} and E_{HOMO}, the lower the E_{LUMO}.

Table 5.4 displays the QSAR models and their statistical properties. Regression models for the training set of chlorinated aromatics with observed E_{LUMO} taken as dependent variables and the combination of the two descriptors (N_{Cl} and E_{HOMO}) as independent variables are presented. According to the rule that r^2 should be greater than 0.8 in a good model, all models in table 5.4 are significant and most models can be considered good models.

Table 5.4 Summary of the models for chlorinated aromatics

eq.	Regression Equations	N*	R^2	RMSE
5.1	E_{LUMO}=-0.0137 N_{Cl}+0.136	17	0.9705	0.00278
5.2	E_{LUMO} =-0.01158 N_{Cl} +0.2356 E_{HOMO} +0.2088	16	0.9919	0.00124
5.3	E_{LUMO}=-0.01279 N_{Cl}+0.1439	12	0.9883	0.00181
5.4	E_{LUMO} = -0.01091 N_{Cl}+0.3301 E_{HOMO} +0.2410	12	0.9955	0.00119
5.5	E_{LUMO}=-0.01148 N_{Cl}+0.1347	11	0.9808	0.00256
5.6	E_{LUMO}= -0.0116 N_{Cl} -0.02037 E_{HOMO}+0.1281	11	0.9809	0.00231

* No outlier is included.

In order to test the robustness of obtained model, the cross-validation method was applied to test the data set where a random number of observations were deleted at a time,

and the regression was refit for the other observed values. The overall results of the cross-validation study are summarized in table 5.5.

Table 5.5 Results of LOO and K-fold Cross-Validation test for chloroaromatic

eq.	LOO-CV			5-fold CV		
	av. R^2	av. R^2_{adj}	RMS_{CV}	av. R^2	av. R^2_{adj}	RMS_{CV}
P_1	0.9871	0.8952	0.0019	0.9866	0.9849	0.0018
P_2	0.9955	0.9944	0.0021	0.9950	0.9936	0.0012
A_1	0.9659	0.9674	0.0026	0.9709	0.9682	0.0028
A_2	0.9920	0.9907	0.0014	0.9921	0.9905	0.0013
B_1	0.9809	0.9785	0.0026	0.9766	0.9732	0.0026
B_2	0.9811	0.9757	0.0027	0.9772	0.9696	0.0121

We observed that all models predict better than chance and can be considered statistically significant. Satisfactory internal stability can be verified for the 6 models, calculated Δ (R_{LOO}^2-$R_{5\text{-fold}}^2$) ranges from 0.01% to 0.43%.

5.3.2 Model Quality Evaluation

The equations and statistics for the QSAR models can be adjusted to the format suggested by Sagrado and Cronin (2006). Table 5.6 shows some conventional (r^2, RMSE) and those used in this work (Dp, Pp, and Mp) statistics related to QSAR models in table 5.5. Dp and Pp are diagnostic statistics (0-100% range) that reflect the descriptive and predictive power of the model, respectively. Mp (= $f_{Dp}D_p$ + (1-f_{Dp}) P_p) represents the overall model quality; all models are selected independently on the f_{Dp} value when it is equal to 0.5 (Sagrado and Cronin, 2006).

Table 5.6 some statistics related to QSARs in table 5.5

eq.	r^2	RMSE	Dp	Pp	Mp
5.1	0.9705	0.00278	94.5924	70.7782	82.6853
5.2	0.9919	0.00124	71.2371	83.1233	77.1802
5.3	0.9883	0.00181	94.4515	70.8273	82.6394
5.4	0.9955	0.00119	58.2466	87.2545	72.7505
5.5	0.9808	0.00256	85.8752	46.4514	66.1633
5.6	0.9809	0.00231	-	-	-

Model 5.6 is of no interest in discrimination between wholly unacceptable models with negative Dp values. In this regard, it is clearly seen for model 5.6 that the use of variable E_{HOMO} is not appropriate, since the confidence interval of b, 100 U(b)/b >100%, is unacceptable. This suggests that the model should be simplified. Model 5.4 has the Dp<60% (algorithm automatically set the limit), which means that this model is less significant than model 5.3.

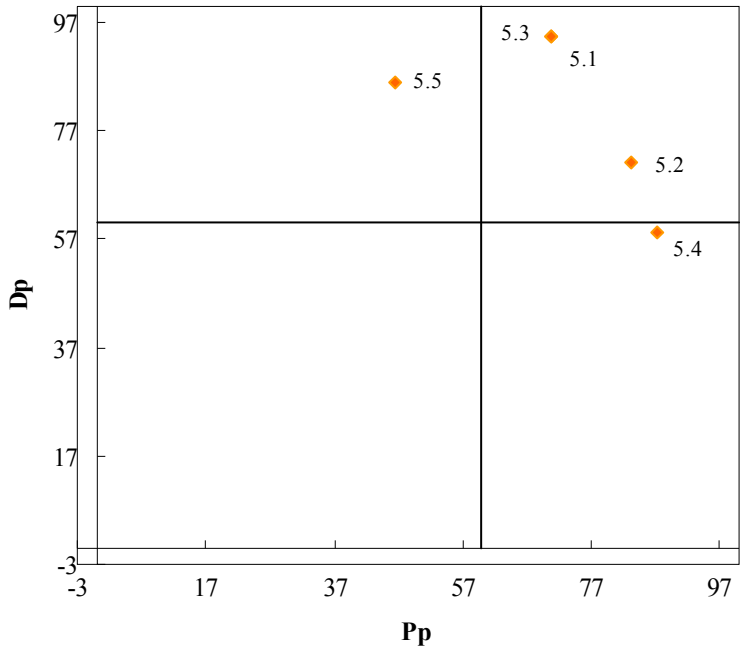


Figure 5.5 Graphical comparison of models by the modeling power plot, based on the descriptive power (Dp) and the predictive power (Pp)

Figure 5.5 shows the global modeling power plot for all the models. As can be observed from this figure, except model 5.6, there is a notable predictive ability for other QSAR models, if the criterion to make this conclusion is that Pp values must be approximately equal to, or larger than, 60% (Sagrado and Cronin, 2006). In contrast, all models have some descriptive ability except models 5.4 and 5.6 (Model 5.4 has the Dp= 58.25%, slightly smaller than 60%) using the same criterion based on 60% limit.

It can be easily concluded that using N_{Cl} also gives results that are more adequate (based on the higher Dp and Pp values of this model). This indicates that these linear models will always result as being the best models when optimizing both the descriptive and the predictive power. Therefore, N_{Cl} containing in a chlorinated aromatic compound is the determining factor of its electronic property E_{LUMO}.

5.3.3 Uncertainty Analysis

From the regression equation discussed in section 5.3.1, coefficients of parameters and the standard deviations are estimated in table 5.7 using bootstrap analysis. The bootstrapping is repeated 5000 times for each validated model and gives the following parameter estimates and their standard errors:

Table 5.7 Summary of coefficients and the standard deviations for aromatic models

eq.	a	b	k	σ_a	σ_b	σ_k
5.1	-0.0137	-	0.1360	0.0006	-	0.0016
5.2	-0.0116	0.2356	0.2088	0.0007	0.0647	0.0205
5.3	-0.0128	-	0.1439	0.0006	-	0.0016
5.4	-0.0109	0.3301	0.2410	0.0007	0.0843	0.0247
5.5	-0.0115	-	0.1347	0.0006	-	0.0021
5.6	-0.0116	-0.0204	0.1281	0.0014	0.2517	0.0830

The expression for the uncertainty in E_{LUMO} determined from the regression model at a measured or specified value of X is found by equation 5.7. Here, we do not consider the correlated uncertainties between any two of these variables, and the uncertainty of number of chlorine and number of carbon are zero. Then, all terms involving correlated uncertainties in equation 5.7 will be simplified as the following equation 5.8.

$$\begin{aligned} U^2_{E_{LUMO}} &= \left(\frac{\partial(E_{LUMO})}{\partial a}\right)^2 U_a^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_{Cl}}\right)^2 U^2_{N_{Cl}} \\ &+ \left(\frac{\partial(E_{LUMO})}{\partial b}\right)^2 U_b^2 + \left(\frac{\partial(E_{LUMO})}{\partial N_C}\right)^2 U^2_{N_C} \\ &+ \left(\frac{\partial(E_{LUMO})}{\partial c}\right)^2 U_c^2 + \left(\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})}\right)^2 U^2_{E_{HOMO}} + \left(\frac{\partial(E_{LUMO})}{\partial k}\right)^2 U_k^2 \end{aligned} \quad (5.7)$$

where, $\frac{\partial(E_{LUMO})}{\partial a} = N_{Cl}$, $\frac{\partial(E_{LUMO})}{\partial N_{Cl}} = a$, $\frac{\partial(E_{LUMO})}{\partial b} = N_C$, $\frac{\partial(E_{LUMO})}{\partial N_C} = b$, $\frac{\partial(E_{LUMO})}{\partial c} = E_{HOMO}$,

$\frac{\partial(E_{LUMO})}{\partial(E_{HOMO})} = c$, $\frac{\partial(E_{LUMO})}{\partial k} = 1$.

$$U^2_{E_{LUMO}} = {N_{Cl}}^2 U_a^2 + {N_C}^2 U_b^2 + {E_{HOMO}}^2 U_c^2 + c^2 U^2_{E_{HOMO}} + U_k^2 \quad (5.8)$$

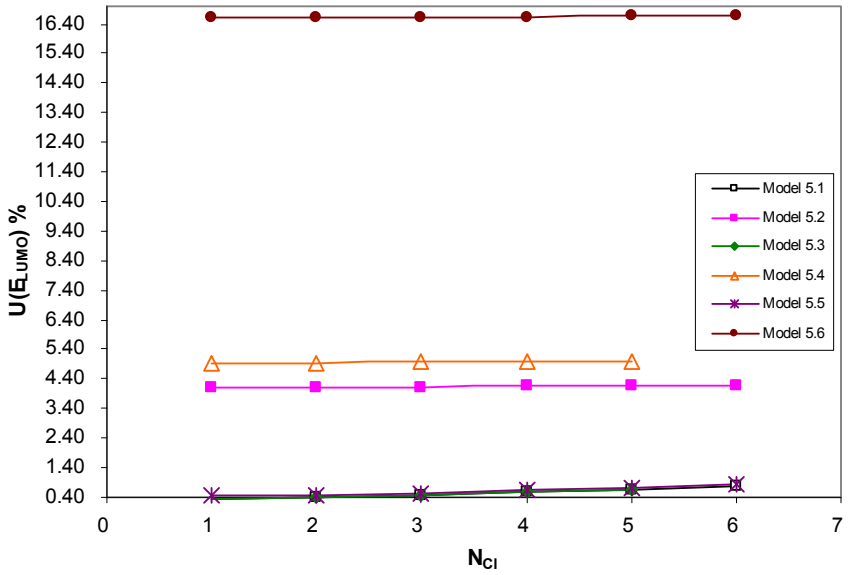


Figure 5.6 Relationship between N_{Cl} and uncertainty in E_{LUMO}

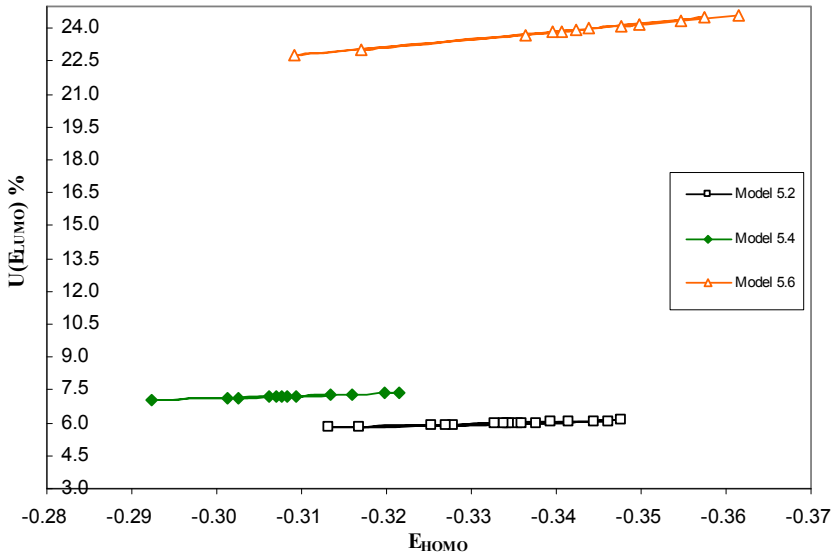


Figure 5.7 Relationship between E_{HOMO} and uncertainty in E_{LUMO}

Figure 5.6 shows the change of uncertainty in E_{LUMO} with number of chlorine. It is clearly shown that number of chlorine has a slight effect on the uncertainty of E_{LUMO} for all models except model 5.6. For example, the uncertainty of E_{LUMO} will increase from

0.34 to 4.99 when N_{Cl} increases from one to six. Similarly, figure 5.7 indicates the impact of E_{HOMO} on the uncertainty of E_{LUMO} follows the same pattern, as shown in figure 5.6. E_{HOMO} has an obvious effect on the relative error on E_{LUMO}, and the uncertainty of E_{LUMO} for model 5.6 will be increased from 22.75 to 24.63 when randomly distributed E_{HOMO} values decrease from -0.31 to -0.36. In figures 5.6 and 5.7, two variables effect on the uncertainty of E_{LUMO} for each model are following: Model 5.6 > model 5.4 > model 5.2 > model 5.5 > model 5.3= model 5.1.

5.4 Conclusion

The success of any QSAR model depends on the selection of appropriate descriptors. Exploring the usefulness of descriptors, especially, conceptual density function theory based descriptors along with other descriptors and analyzing their applicability could lead to a drastic improvement in QSAR models. Based on this fact, quantitative structure-activity relationship for the data set containing 53 choro- phenols, anilines, and benzenes congeners on the energy of lowest unoccupied molecular orbital (E_{LUMO}) is analyzed. Traditional regression procedures along with cross-validation are carried out and the predictive, descriptive and modeling powers are evaluated for the developed models. It has been shown that, using the entire data set, the number of chlorine substituents (N_{Cl}) with the highest occupied molecular orbital (E_{HOMO}) energy as descriptors provides a reasonably good coefficient of determination ($0.9705<r^2<0.9955$) and the cross-validated squared correlation coefficient ($0.9659<r^2_{cv}<0.9659$) value indicates the significance of the developed model.

CHAPTER 6 QSAR MODELS FOR TOXICITY ANALYSIS OF CHLORINATED COMPOUNDS

Summary

Five sets of acute toxicity data of chlorinated compounds against various species such as fish (chloroalkanes), T. pyriformis (chlorophenols and aromatic compound with a nitro- or cyano group) and P. phosphoreum (chlorobenzenes and chloroanilines) were used to establish QSAR models of toxicity of chlorinated DBPs. The logarithm of the octanol/water partition coefficient (logP), E_{LUMO}, and N_{Cl} were used as molecular descriptors. Suitable QSAR models ($0.514 < r^2 < 0.992$) to predict acute toxicity of chlorinated compounds to fish, T. pyriformis, and P. phosphoreum have been derived. The logP was an important descriptor with the additional E_{LUMO} and N_{Cl} descriptors being required for all cases. Based on these results, an advisory tool has been developed that directs users to the appropriate QSAR model to apply to various types of organisms within specified logP, E_{LUMO} and N_{Cl} ranges. Using this tool, it is possible to obtain a good indication of the toxicity of a large set of DBP chemicals and newly developed chlorinated compounds to the different organisms without the need for additional experimental testings.

6.1 Introduction

The toxicological assessment of chlorinated compounds is essential for environmental risk assessment purposes during the establishment of DBP standards such as Maximum Contaminant Level (MCL). There have been many attempts to develop QSAR for the prediction of toxicity of chlorinated compounds. By far, the largest numbers of QSARs are developed based on the hydrophobicity property as a descriptor since hydrophobicity can be expressed by the octanol-water partition coefficient (K_{ow} or P) in eq. 6.1. Values of P may be several orders of magnitude, but it is usually expressed in the logarithmic form as logP (Ribeiro and Ferreira, 2003).

$$K_{ow} or P = \frac{C_{org}}{C_{water}} \tag{6.1}$$

where C_{org}=concentration of the non-ionised solute in the organic phase, C_{water}=concentration of the non-ionised solute in the water phase. LogP > 0, means greater solubility in the organic phase; and if logP < 0, it means greater solubility in the aqueous phase.

The logP is essential for understanding the transport mechanisms and distribution of compounds into the end point. For example, the mechanism that a substance to accumulate in the liquid phase of biomembranes. Padmanabhan et al. (2006) shows the relevance of the logP term to the toxicity for *polychlorinated biphenyls* compounds for 133 PCB. The inclusion of a connectivity or 'shape' term with logP produced acceptable results. Even in a group of 16 halo- and methyl-phenols and anilines with known respiratory uncoupling ability a reasonable correlation (r^2=0.85, n=16) between toxicity and logP was found. In contrast, Blaha et al. (1998) and Sixt et al. (1995) found that logP

was a poorer descriptor of the toxic potential than other 'size of the molecular' terms by studying the larger chlorinated aliphatic and aromatic compounds.

It is clear from these studies that logP alone is not sufficient to model the toxicity of chlorinated aliphatic and aromatic compounds. Other molecular descriptors which reflect electronic property of a chemical must be included. Molecular descriptors derived from quantum chemical computation have obvious advantages, because first they are not restricted to closely related compounds and can be easily obtained; secondly, they can clearly describe defined molecular properties. For these reasons, there have been many examples of the use of quantum chemical molecular descriptors (Lu et al., 2001; Zhang et al., 2007; and Wang et al., 2004).

In previous chapters, electrophilicity parameter (e.g., E_{LUMO}), it is directly proportional to the electronic affinity of the compound, and has been demonstrated to correlate with many physico-chemical properties for chloro-aliphatic and chloro-aromatic including N_{Cl}, N_C (for aliphatic) and E_{HOMO} (for chlorinated alkane and aromatic compounds). Hence, it is logical to examine the relationship between the toxicity and electrophilicity parameters with hydrophobicity parameter and number of chlorine. In this study, 50% effective inhibition concentration (EC_{50}) or 50% growth inhibition concentration (IGC_{50}) of chlorinated alkane, aromatic and amide compounds were used as dependent variables in the development of QSAR models. Suitable QSAR models were developed using logP and E_{LUMO}, with N_{Cl} and/or N_C as molecular descriptors for an extensive series of chloro-alkanes, benzenes, anilines, phenols, nitrophenols, and other chlorinated compounds. The applicability and limits of the QSARs were also identified by detecting certain types of compounds that are outliers and the applicable domain for the QSAR models.

6.2 Data Set and Molecular Descriptors

In this chapter, a broad range of toxicity data of chemicals, including chlorinated alkanes, phenols, benzenes, anilines and aromatic compounds with nitro/cyano group was collected. The experimental toxicity predictors (E_{LUMO}, logP, E_{HOMO}, $logEC_{50}$ and/or $logIGC_{50}$) for each chlorinated data set were taken from different sources, which are listed in table 6.1. Table 6.2 gives an overview of these parameters with a short description.

Table 6.1 Descriptors and reference in various classes

Data set	N	Toxicity predictor	Descriptors	Test System	source
alkane	13	in vivo $logEC_{50}$, in vitro $logEC_{50}$	logKow, E_{LUMO}, N_{Cl}, N_C	Fish	Zvinavashe et al. (2008)
phenol	37	$Log(1/IGC_{50})$	logP, E_{LUMO}, N_{Cl}	T. pyriformis	Schultz et al. (1997)
benzene	10	$logEC_{50}$	logP, E_{LUMO} (E_{HOMO}), N_{Cl}	Photobacterium phosphoreum	Kaiser et al. (1994)
aniline	15	$logEC_{50}$	logP, E_{LUMO} (E_{HOMO}), N_{Cl}	Photobacterium phosphoreum	Kaiser et al. (1994)
aromatic compound	42	$Log(1/IGC_{50})$	logP, E_{LUMO}, N_{Cl}	T. pyriformis	Cronin et al. (2001)

Table 6.2 Theoretical physico-chemical parameters

No.	Abbreviation	Definition
1	EC_{50}	the 50% effective inhibition concentration
2	IGC_{50}	The 50% growth inhibition concentration
3	LogP ($logK_{ow}$)	1-octanol/water partition coefficient
4	E_{LUMO}	Energy of Lowest unoccupied molecular orbital
5	N_{Cl}	Number of chlorine
6	N_C	Number of carbon

6.3 Results

There have been many reports on the prediction of the relationship between logP and toxicity for various chlorinated compounds, because the compounds have serious

ecological harmful effects and are implicated as potent carcinogens. In addition, E_{LUMO} has also been reported as a useful molecular descriptor to predict the toxicity. It has been reported, moreover, that N_{Cl} is also an important descriptor to express the mechanism of toxicity. It is expected that these properties may explain the new partition and electronic properties of the compounds. In the present study, the toxicological problems posted by the different chlorinated compounds were explained together with evidence on their mechanisms of action. The available QSAR models from the experimental results on the toxicity properties of the chlorinated compounds were compared with the predicted data. QSAR models for toxicity of homogenous group of chemical compounds based on logP, accompanied with E_{LUMO} and N_{Cl} substituents as descriptors are developed using multiple linear regression method.

6.3.1 Chloro- alkanes

Chloroalkanes are one class of DBPs and are also widely introduced into the environment (Freitag et al, 1994). In general, the genotoxic potential is dependent on the nature, number, and position of chlorine(s) and the molecular size of the compounds. Short-chain monochloroine alkanes are potential direct-acting alkylating agents, particularly if the chlorine is at the terminal end of the carbon chain (Woo et al., 2002). E_{LUMO} could be used as the descriptor to assess the ability of a chemical to accept an electron. Clearly, the lower the E_{LUMO} of the organochlorine, the easier it is for the organochlorine can accept electrons, as there is a smaller gap to jump (Gallagher, 2001). Fully chlorinated alkanes tend to act by free radical or nongenotoxic mechanisms or

undergo reductive dechlorogenation to yield chloroalkenes that in turn could be activated to epoxides (Woo et al., 2002).

The experimental toxicity data sets from the literature were obtained using the direct exposure method to fish. For 13 chlorinated alkane compounds, in vivo and in vitro $logEC_{50}$, logP and the semi-empirical quantum chemical descriptors E_{LUMO}, N_{Cl} and N_C are listed in table 6.3. The EC_{50} values of the chlorinated alkanes were calculated using a Microsoft Excel plug-in, Life Sciences Workbench (LSW) Data Analysis Toolbox Version 1.1.1 and logP values calculated using the software CLogP version 4.0. Initially, Zvinavashe et al. (2008) performed a linear regression analysis for 18 chloroalkanes using experimental $logK_{ow}$ as independent variable and $logEC_{50}$ as dependent variable on the training set, and MMT test revealed a good correlation (r^2=0.883, s=0.193). In this study, two QSAR models (eq. 6.1 and 6.2) were analyzed with in vivo or in vitro $logEC_{50}$, which acted as the dependent variables; chemical descriptors (logP, E_{LUMO}, N_{Cl} and N_C) acted as the independent variables. For in vivo EC_{50}, the QSAR model is as follows:

$$\text{in vivo } logEC_{50} = 2.964-0.144\ logP+6.273\ E_{LUMO}-0.160\ N_{Cl}-0.244\ N_C \quad (6.1)$$

($N=12$, $R^2=0.88$, $R^2_{5\text{-fold}}=0.8674$, $R^2_{LOO}=0.8741$, RMSE=0.262, F=12.798, P>0.002)

For in vitro EC_{50}, the QSAR model is as follows:

$$\text{in vitro } logEC_{50} = 3.527-0.464\ logP+3.30\ E_{LUMO}-0.0078\ N_{Cl}-0.0496\ N_C \quad (6.2)$$

($N=13$, $R^2=0.865$, $R^2_{5\text{-fold}}=0.8908$, $R^2_{LOO}=0.8898$, RMSE=0.281, F=12.78, P>0.002)

Table 6.3 Chloroalkanes present in the training set in of the present study

No.	DBP Compounds	logP	E_{LUMO}	N_{Cl}	N_C	in vitro $logEC_{50}$			in vivo $logEC_{50}$		
						obs.	pred.	res.	obs.	pred.	res.
1	1-chlorohexane	3.05	0.2161	1	6	2.24	2.52	-0.28	2.05	2.26	-0.21
2	1-chlorooctane	4.64	0.2165	1	8	1.43	1.69	-0.26	1.30	1.54	-0.24
3	1-chlorodecane	5.70	0.2167	1	10	1.44	1.10	0.34	1.31	0.90	0.41
4	1,2-dichloropropane	1.99	0.1875	2	3	3.03	3.06	-0.03	3.01	2.80	0.21
5	1,2-dichlorobutane	2.52	0.1915	2	4	2.61	2.78	-0.17	2.39	2.51	-0.12
6	1,1,2-trichloroethane	2.05	0.1562	3	2	3.04	2.97	0.07	2.85	2.68	0.17
7	pentachloroethane	3.63	0.1178	5	2	2.17	2.09	0.08	1.87	1.89	-0.02
8	1-chlorobutane	2.52	0.2154	1	4	3.27	2.86	0.41	3.02	2.82	0.20
9	1-chloroheptane	4.11	0.2164	1	7	1.73	1.98	-0.25	1.58	1.86	-0.28
10	1,5-dichloropentane	2.77	0.2027	2	5	2.82	2.65	0.17	-	2.47	-0.02
11	1,2,3-trichloropropane	1.98	0.1604	3	3	3.12	2.97	0.15	2.45	2.38	-0.04
12	1,1,2,2-tetrachloroethane	2.64	0.1476	4	2	2.44	2.66	-0.22	2.34	2.26	-0.06
13	carbon tetrachloroethane	2.88	0.0946	4	1	2.40	2.42	-0.02	2.20	2.26	-0.21

It can be inferred from Eqs. 6.1 and 6.2 that the regression model developed using all four descriptors as independent variables is capable of explaining variation in data (r^2=0.88 and 0.865) with Leave-One-Out and 5-fold cross-validated squared correlation coefficient $r^2_{LOO\text{-}CV}$=(0.874 and 0.889) and $r^2_{5\text{-}fold}$=(0.867 and 0.89), respectively.

Table 6.4 Pearson correlation coefficient of models for chloroalkanes

	logP	E_{LUMO}	N_{Cl}	N_C	in vitro $logEC_{50}$
logKow	1.000				
E_{LUMO}	0.345	1.000			
N_{Cl}	-0.352	-0.948	1.000		
N_C	0.805	0.806	-0.798	1.000	
in vitro $logEC_{50}$	-0.924	-0.225	0.240	-0.695	1.000

	logP	E_{LUMO}	N_{Cl}	N_C	in vivo $logEC_{50}$
logKow	1.000				
E_{LUMO}	0.367	1.000			
N_{Cl}	-0.360	-0.952	1.000		
N_C	0.816	0.808	-0.797	1.000	
in vivo $logEC_{50}$	-0.908	-0.215	0.204	-0.703	1.000

The results of Pearson matrix correlations for the examination of relationships between the variables are shown in table 6.4. The highest relationship (r=-0.924) was obtained between logP and in vitro $logEC_{50}$. For in vivo $logEC_{50}$, Pearson regression tests indicated logP also has a good correlation (r= -0.908). On the other hand, the medium level of correlations (r= -0.695 and -0.703) were obtained between number of carbon and in vitro and in vivo $logEC_{50}$. Several studies reported that there is a linear relationship between logP and $logEC_{50}$, as was the case with our present study. However, only limited research has been done on analyzing DBPs toxicity with atom descriptors such as N_{Cl} and N_C for chloroalkanes. Zvinavashe et al. (2008) have shown that the number of carbon has the positive effect to cytotoxicity, that is, with an increase in chain length for single chlorinated compounds from C_4-C_{10}, the toxicity will increase for 1-chloropentane, 1-chlorohexane, and 1-chlorodecane. For those compounds with the same hydrocarbon, an additional chlorine atom was associated with an increase in toxicity (the chlorine atoms Cl_1-Cl_5). Toxicity results of chlorianted alkane compounds reported in the published study (Zvinavashe et al., 2008) are comparable to this study, Eqs. 6.1 and 6.2 have no significant improvement on the correlation coefficient. Thus, hydrophobicity has been confirmed as an important parameter to describe the toxicity of the chlorinated alkanes, increasing hydrophobicity leads to easier passage through membranes and greater distribution into the organisms, thus resulting in greater toxicity.

6.3.2 Chloro-phenols

Chlorophenols (CPs) are commonly found in drinking water as disinfection by-products (DBPs) due to chlorination (Czaplicka, 2004; Peller et al., 2003). CPs are

also ubiquitous pollutants, which enter the environment as by-products from the leaching of wood products, synthesis during bleaching operations and emissions from operating facilities (Puhakka and Melin, 1996). For example, pentachorophenol is commonly used as a wood preservative, fungicide, and herbicide (Hoos, 1978); the trichlorophenols are used as bactericides and fungicides (Fragiadakis, 1981); 2,4-dichlorophenol (2,4-DCP) is used in the manufacture of 2,4-dichlorophenoxyacetic acid, an herbicide (Rappe, 1980). Furthermore, many chlorinated hydrocarbons are extremely stable in the environment and prone to bioaccumulation (Wang et al., 1999; Makinen et al., 1993). Therefore, it is not surprising that CPs are classified as priority pollutants by the US EPA. Many chlorophenol contaminated sites have been targeted for cleanup.

Toxicological assessment of chlorophenol compounds is essential for risk assessment purposes. QSARs for chlorophenols have enabled predictions of toxicity to be made for untested compounds. Moreover, they offer insight into the mechanisms of toxic action. To achieve this, an historical data set of chlorine substituted phenolic ring toxicity values was used and detailed structural criteria has been previously described (Schultz et al., 1997). Cronin and Schultz (1996) developed a two-term (logP and E_{LUMO}) QSAR model with good correlation coefficient (n=120, R^2=0.90, s=0.26) for the toxicity for the selected compounds of this data set.

The objective of this investigation was to develop QSAR analyses of chlorophenols toxicity data to *T. pyriformis* with three calculated physico-chemical predictors using multiple linear regression techniques (table 6.5). The 37 chlorophenols utilized in this study were structurally and mechanistically diverse, with some compounds having well established mechanisms of action (i.e. polar narcosis). For QSAR derivation, $\log(1/IGC_{50})$

acted as the dependent variable and chemical descriptors (logP, E_{LUMO} and N_{Cl}) acted as the independent variables. logP and E_{LUMO} were calculated using the ACD/Labs software and Chem-X version 2000.1.

The following relationship was found between the toxicity of the phenols to *T. pyriformis* in all subsequent analyses:

$$\log(1/IGC_{50}) = -0.265+0.404\ \text{logP}-0.352\ E_{LUMO}+0.0903\ N_{Cl} \quad (6.3)$$

(N=37, R^2=0.638, RMSE=0.402, F=19.411, P<0.0001)

Since the correlation coefficient is 0.638, the statistical fit to model 3 is poor. The possibilities that the group of compounds were poorly predicted are: (1) The residual shows that some compounds were outliers due to phenol substituted in the 2- or 4-position by an amino or a nitro group; (2) Phenols substituted with three or more chlorines; (3) hydroquinones (Cronin et al., 2002). Such compounds are associated with the weak respiratory uncoupling mechanism of toxic action (Terada, 1990). Thus, the leverage test was utilized to verify the presence of outliers (chlorohydroquinone and 2,6-dichloro-4-nitrophenol) and their removal resulted in the following improved QSAR:

$$\log(1/IGC_{50}) = -0.446+0.471\ \text{LogP}-0.488\ E_{LUMO}+0.0487\ N_{Cl} \quad (6.4)$$

(N=35, R^2=0.748, R^2_{CV}=0.75, RMSE=0.341, $RMSE_{CV}$=0.347, F=30.66, P<0.0001)

Table 6.5 Chlorophenol toxicity to *T. pyriformis* and physicochemical descriptors

	NAMES	SMILES	MOA[a]	Toxicity	logP	E_{LUMO}	N_{Cl}
1	2-chlorophenol	Clc(ccc1)c(c1)O	polar narc[b]	0.18	2.04	0.030	1
2	2-chloro-5-methylphenol	Oc1c(Cl)ccc(C)c1	polar narc	0.39	2.50	0.019	1
3	4-chlorophenol	Clc(ccc1O)cc1	polar narc	0.55	2.43	0.095	1
4	2-chloro-4,5-dimethylphenol	c1(C)c(C)cc(Cl)c(O)c1	polar narc	0.69	2.96	0.053	1
5	4-chloro-2-methylphenol	Clc(ccc1O)cc1C	polar narc	0.70	2.89	0.080	1
6	2,6-dichlorophenol	Clc(ccc1)c(O)c1Cl	polar narc	0.74	2.61	-0.259	2
7	3-chloro-5-methoxyphenol	c1(O)cc(Cl)cc(OC)c1	polar narc	0.76	2.66	0.027	1
8	4-chloro-3-methylphenol	Clc(ccc1O)c(c1)C	polar narc	0.80	2.89	0.133	1
9	2,4-dichlorophenol	Clc(ccc1O)cc1Cl	polar narc	1.04	2.99	-0.245	2
10	4-chloro-3-ethylphenol	c1c(Cl)c(CC)cc(O)c1	polar narc	1.08	3.42	0.141	1
11	2,5-dichlorophenol	Clc(ccc1Cl)cc1O	polar narc	1.13	2.88	-0.325	2
12	2,4,6-trichlorophenol	Clc(cc(Cl)c1O)cc1Cl	polar narc	1.41	3.58	-0.502	3
13	3,5-dichlorosalicylaldehyde	C(=O)c1c(O)c(Cl)cc(Cl)c1	polar narc	1.55	3.52	-0.893	2
14	3,5-dichlorophenol	Clc(cc(Cl)c1)cc1O	polar narc	1.57	3.33	-0.285	2
15	3,4-dichlorophenol	Clc(ccc1O)c(Cl)c1	polar narc	1.75	3.22	-0.236	2
16	4-chloro-2-isopropyl-5-methylphenol	Clc(cc(c1O)C(C)C)c(c1)C	polar narc	1.85	4.22	0.114	1
17	2,3,5-trichlorophenol	Oc1c(Cl)c(Cl)cc(Cl)c1	polar narc	2.37	3.69	-0.578	3
18	4-chlororesorcinol	Oc1cc(O)c(Cl)cc1	polar narc	0.13	1.67	-0.008	1
19	3-chlorophenol	Clc(ccc1)cc1O	polar narc	0.87	2.40	0.019	1
20	4,6-dichlororesorcinol	Oc1cc(O)c(Cl)cc1(Cl)	polar narc	0.97	2.58	-0.263	2
21	2,3-dichlorophenol	Clc(ccc1)c(Cl)c1O	polar narc	1.28	2.83	-0.262	2
22	2,4,5-trichlorophenol	Clc(cc(Cl)c1O)c(Cl)c1	polar narc	2.10	3.71	-0.555	3
23	2-amino-4-chlorophenol	Clc(ccc1O)cc1N	pro-elec[c]	0.78	1.67	0.043	1
24	4-chlorocatechol	Oc1c(O)cc(Cl)cc1	pro-elec	1.06	2.15	0.001	1
25	chlorohydroquinone	Oc1c(Cl)cc(O)cc1	pro-elec	1.26	1.52	-0.111	1
26	tetrachlorocatechol	Oc1c(O)c(Cl)c(Cl)c(Cl)c1(Cl)	pro-elec	1.70	4.75	-0.830	4
27	2,6-dichloro-4-nitrophenol	Clc(cc(c1)N(=O)=O)c(O)c1Cl	resp unc[d]	0.63	2.88	-1.441	2
28	2,4-dichloro-6-nitrophenol	c1(O)c(Cl)cc(Cl)cc1(N(=O)=O)	resp unc	1.75	3.41	-1.579	2
29	pentachlorophenol	Clc(c(Cl)c(Cl)c1O)c(Cl)c1Cl	resp unc	2.05	4.78	-0.978	5
30	2,3,5,6-tetrachlorophenol	Clc(cc(Cl)c1Cl)c(Cl)c1O	resp unc	2.22	4.06	-0.817	4
31	2,3,4,5-tetrachlorophenol	Clc(c(Cl)c(Cl)c1O)c(Cl)c1	resp unc	2.71	4.39	-0.752	4
32	2-chloromethyl-4-nitrophenol	c1(O)c(CCl)cc(N(=O)=O)cc1	soft elec[e]	0.75	1.84	-1.195	1
33	2-amino-4-chloro-5-nitrophenol	Clc(cc(c1O)N)c(c1)N(=O)=O	soft elec	1.17	2.63	-0.960	1
34	2-chloro-4-nitrophenol	Clc(cc(c1)N(=O)=O)c(O)c1	soft elec	1.59	2.22	-1.264	1
35	4-chloro-6-nitro-m-cresol	Clc(cc(c1O)N(=O)=O)c(c1)C	soft elec	1.64	3.21	-1.346	1
36	4-chloro-2-nitrophenol	Clc(ccc1O)cc1N(=O)=O	soft elec	2.05	2.75	-1.388	1
37	tetrachlorohydroquinone	Oc1c(Cl)c(Cl)c(O)c(Cl)c1Cl	pro-redox[f]	2.11	3.79	-0.928	4

[a] MOA: mechanism of toxic action; [b]polar narcotic; [c] respiratory uncoupler; [d] pro-electrophile; [e] soft electrophile; [f] pro-redox cycler.

Table 6.6 Correlation matrix between the variables included in eq. 6.4

Variables	LogP	E_{LUMO}	N_{Cl}	$\log(1/IGC_{50})$
LogP	**1.000**			
E_{LUMO}	-0.304	**1.000**		
N_{Cl}	0.784	-0.373	**1.000**	
$\log(1/IGC_{50})$	0.733	-0.600	0.692	**1.000**

The inter-correlations between the variables in eq. 6.4 are listed in table 6.6. There are no significant correlations between variables. Table 6.6 shows that, among the three descriptors selected, logP is capable of providing maximum variation in data (r=0.733) compared to the other two descriptors. It is possible that increasing hydrophobicity leads to a greater uptake of the higher chlorinated phenols and therefore their greater toxicity. It is also interesting to note that the lower the E_{LUMO} value, the greater the toxicity interactions by chloro-phenols and the smaller the $\log(1/IGC_{50})$ values. Again, the trend of increasing toxicity with increasing chlorination is evident and can be expressed as follows:

PCP > 2,3,4,5-TCP > 2,3,5-TCP > 2,4-DCP > 2-CP > phenol

These results are consistent with explanations for the toxicity of the chlorophenols that invoke their interference with electron transport and/or the proton movements that accompany phosphorylation events (Dedonder and Van Sumere, 1971). Thus, increasing chlorination of the phenol molecule would result in the increasing ability to accept the electron and in the increasing toxicity of the chlorinated phenols.

6.3.3 Chloro-benzenes

The compounds involved in this section and the values of toxicity were found in the study from the Computox Database (Kaiser et al., 1994). The Microtox is defined as the negative logarithm of the concentration that causes a 50% reduction of bioluminescence ($\log(1/EC_{50})$ in mmol/L) of the *Photobacterium phosphoreum* after a certain time of exposure. The aim of the present study was to predict the toxicity of 10 selected chlorobenzene compounds listed in table 6.7, from their structures, without reference to exact toxicity mechanisms of individual compounds. To this end, two multiple regressions reported were performed using the SAS program, and the stepwise regression equations are:

$$\text{Log}(1/EC_{50}) = 0.623-0.478\ \text{logP}+9.767\ E_{LUMO}+0.626\ N_{Cl} \quad (6.5)$$

$N=10$, $R^2=0.601$, $R^2_{cv}=0.613$, RMSE=0.229, $RMSE_{cv}=0.217$, F=3.012, P=0.116

$$\log(1/EC_{50}) = 8.061+0.639\ \text{logP}+25.858\ E_{HOMO}\ -0.0271\ N_{Cl} \quad (6.6)$$

$N=10$, $R^2=0.673$, $R^2_{cv}=0.641$, RMSE=0.207, $RMSE_{cv}=0.209$, F=4.112, P=0.067

Both eqs. 6.5 and 6.6 indicate that the relationship between toxicity and the quantum descriptors is still uncertain, which probably results from the great differences in the molecular structures of the selected compounds. In contrast to the result obtained for the chlorophenols, table 6.8 seems to indicate that, for chlorobenzene compounds, hydrophobicity is relatively not important (r=0.735) for toxicity potency while the atom predictor (expressed by N_{Cl}) comes into play (r=0.761). In eq. 6.5, the descriptors logP and N_{Cl} are negatively and positively correlated to the toxicity respectively. Since all of the selected compounds have polar functional groups such as –Cl, and these groups tend

to form a hydrogen bond with water molecules, and reduce the sensitivity of logP for distinguishing the differences in toxicity.

Table 6.7 Chlorobenzenes with Microtox, logP, N_{Cl}, and E_{HOMO} as predictors

	DBP Compounds	logP	N_{Cl}	E_{HOMO}	E_{LUMO}	$Log(1/EC_{50})$	eq. 6.5		eq. 6.6	
							pre.	res.	pre.	res.
1	1,2,3,4-tetrachlorobenzene	4.46	4	-0.350	0.088	1.73	1.86	-0.13	1.76	-0.03
2	1,2,3,5-tetrachlorobenzene	4.65	4	-0.353	0.087	1.94	1.75	0.19	1.81	0.13
3	1,2,4,5-tetrachlorobenzene	4.52	4	-0.348	0.086	1.68	1.81	-0.13	1.85	-0.17
4	1,2,3-trichlorobenzene	4.11	3	-0.352	0.102	1.76	1.53	0.23	1.50	0.26
5	1,3,5-trichlorobenzene	4.15	3	-0.357	0.099	1.11	1.48	-0.37	1.39	-0.28
6	1,2,4-trichlorobenzene	4.02	3	-0.344	0.098	1.66	1.54	0.12	1.66	0.00
7	1,2-dichlorobenzene	3.38	2	-0.342	0.114	1.39	1.38	0.01	1.32	0.07
8	1,3-dichlorobenzene	3.38	2	-0.344	0.112	1.46	1.36	0.10	1.27	0.19
9	1,4-dichlorobenzene	3.39	2	-0.336	0.110	1.44	1.33	0.11	1.48	-0.04
10	chlorobenzene	2.84	1	-0.337	0.128	1.00	1.15	-0.15	1.15	-0.15

Table 6.8 Correlation matrix of molecular descriptors for eq. 6.5 and 6.6

	logP	E_{LUMO}	N_{Cl}	$Log(1/EC_{50})$
logP	1			
E_{LUMO}	-0.984	1		
N_{Cl}	0.991	-0.993	1	
$Log(1/EC_{50})$	0.735	-0.748	0.761	1

	logP	E_{HOMO}	N_{Cl}	$Log(1/EC_{50})$
logP	1			
E_{HOMO}	-0.79	1		
N_{Cl}	0.991	-0.725	1	
$Log(1/EC_{50})$	0.735	-0.357	0.761	1

For compounds considered in eq. 6.5, a multiple regression between toxicity and molecular descriptors such as logP, E_{LUMO}, and N_{Cl} has a correlation coefficient R^2 of 0.673. It is, therefore, possible to replace E_{LUMO} by E_{HOMO}. According to model 6.6, $logEC_{50}$ does not depend on the electronic properties (E_{HOMO}) and hydrophobicity (logP) but on number of chlorine (r=0.761). This study shows that number of chlorine was a better descriptor of the toxic potential than other 'size of the molecule' terms for

chlorinated benzenes. Generally, an increase in the number of chlorine atoms substituted on a benzene ring, known as a "heavy atom effect", decreases the lifetime of the excited state and accelerates intersystem crossing (Uchimura, 2005).

6.3.4 Chloro-anilines

The chlorinated anilines are one of the chemical classes in which the structural and molecular basis of toxicity is most clearly understood. Exposure to anilines occurs in different industrial and agricultural activities as well as in the textile industry. The substitution of a chloro group to the aniline often enhances activity, and chloroanilines are found to be common contaminates in several working environments, including the chemical and mechanical industries.

Toxicity data for chloroanilines in this study are obtained from the Computox Database (Kaiser et al., 1994). The aim of the present study was to predict the toxicity of 15 selected chloroaniline compounds, as in earlier studies by Gombar et al. (1988 and 1989) and Ribo et al. (1984), pentachloroaniline was found to be an outlier. Therefore, pentachloroaniline was excluded in the following investigations.

Table 6.9 shows that the toxicity of the dichloroanilines seems to be correlated to the distance of the chlorine atoms to the amino group. The toxicity of chloroaniline fits into the sequence (3,5<2,4<2,5<2,3<2,6) except for 3,4-dichloroaniline since this compound acts as another mechanism and is more toxic than the 2,6-dichloroaniline. High toxic values may be due to some oxidative actions of the amino group. The toxic effects of these compounds were not only controlled by the electronic affinity factors, but also by parameters that characterize the oxidative tendency.

Table 6.9 Chloroanilines with Microtox, logP, N_{Cl}, and E_{HOMO} as predictors

	Compounds	Log(1/EC_{50})	logP	E_{LUMO}	N_{Cl}	E_{HOMO}
1	2-chloroaniline	0.91	1.90	0.1330625	1	-0.298019
2	3-chloroaniline	0.96	1.88	0.1321811	1	-0.301362
3	4-chloroaniline	1.40	1.83	0.1342397	1	-0.292434
4	2,3-dichloroaniline	1.77	2.71	0.1185907	2	-0.307084
5	2,4-dichloroaniline	1.54	2.78	0.1181457	2	-0.300741
6	2,5-dichloroaniline	1.63	2.75	0.1161306	2	-0.308247
7	2,6-dichloroaniline	1.97	2.20	0.1170938	2	-0.306245
8	3,4-dichloroaniline	2.40	2.69	0.1191548	2	-0.302458
9	3,5-dichloroaniline	1.19	2.90	0.1160324	2	-0.313406
10	2,3,4-trichloroaniline	1.92	3.46	0.1063855	3	-0.307558
11	2,4,5-trichloroaniline	2.12	3.45	0.1039936	3	-0.309408
12	3,4,5-trichloroaniline	1.77	3.32	0.1053502	3	-0.311935
13	2,3,4,5-tetrachloroaniline	2.37	4.33	0.09329177	4	-0.316051
14	2,3,5,6-tetrachloroaniline	2.16	4.24	0.09152912	4	-0.321421
15	pentachloroaniline	1.35	4.59	0.08212995	5	-0.319792

$$\text{Log}(1/EC_{50}) = -1.926 - 0.711\ \text{logP} + 24.591\ E_{LUMO} + 1.261\ N_{Cl} \quad (6.7)$$

$N=14$, $R^2=0.636$, $R^2_{cv}=0.5976$, RMSE=0.331, $RMSE_{cv}=0.3509$, F=5.818, P=0.014

Table 6.10 Correlation matrix of descriptors for eq. 6.7

	logP	E_{LUMO}	N_{Cl}	Log(1/EC_{50})
logP	1			
E_{LUMO}	-0.976	1		
N_{Cl}	0.981	-0.994	1	
log(1/EC_{50})	0.697	-0.745	0.758	1

In a study of 14 chlorinated aniline compounds, QSAR model using logP, E_{LUMO} and N_{Cl} as descriptors has a poor correlation coefficient R^2 of 0.636. This was attributed to different reactivity mechanisms for these compounds. Table 6.10 shows that the correlation coefficient of each descriptor to toxicity is follow: logP (r=0.697) <E_{LUMO} (r=0.745) <N_{Cl} (r=0.758), the most significant parameter, again is the number of chlorine.

6.3.5 Chlorinated Aromatics Containing A Nitro- or Cyano Group

The aim of this section was to determine which descriptor best parametrized the essentiality of aromatic compounds with regard to their acute toxicity. The experimental toxicity data was obtained from the literature based upon the Microtox test in the 40-h *Tetrahymena pyriformis* population growth impairment assay (Cronin et al., 2001) expressed as $\log(1/IGC_{50})$ values. Cronin et al. (2001) previously have utilized two different descriptions such as E_{LUMO} and logP for 203 substituted aromatic compounds containing a nitro- or cyano group. Chemicals that contain multiple functional groups deserve special attention. Such chemicals might exhibit enhanced effects as a result of synergism or even exhibit a different Mode of Action (MOA) and are likely to be the outlier to well established QSAR models. In this section, all 42 data points including $\log(1/IGC_{50})$, logP and additional molecular physico-chemical descriptors (E_{LUMO} and N_{Cl}) were calculated and listed in table 6.11 together with the CAS registry numbers of each compound. Initially, QSAR models were developed based on various chemical groups to analyze if the significant descriptors across several classes shared the same properties (table 6.12).

Table 6.11 Toxicity and molecular descriptors of 47 monoaromatic homologues

No.	Compounds	CAS	log (1/IGC_{50})	logP	E_{LUMO}	N_{Cl}
1	3-chlorobenzonitrile	766-84-7	-0.06	2.29	-0.6767	1
2	4-chlorobenzonitrile	623-03-0	0.00	2.29	-0.7351	1
3	2-chlorobenzonitrile	873-32-5	0.28	2.16	-0.6704	1
4	2-amino-5-chlorobenzonitrile	5922-60-1	0.44	1.79	-0.7918	1
5	4-chloro-3-nitrobenzonitrile	939-80-0	1.71	1.83	-1.7138	1
6	4-chloro-3,5-dinitrobenzonitrile	1930-72-9	2.66	1.37	-2.3008	1
7	2-chloro-4-methyl-3-nitropyridine	23056-39-5	0.29	1.48	-1.2233	1
8	2-chloro-4-methyl-5-nitropyridine	23056-33-9	0.42	1.68	-1.6062	1
9	2-chloro-5-nitropyridine	4548-45-2	0.80	1.26	-1.6847	1
10	2-chloro-3-nitropyridine	5470-18-8	0.87	1.06	-1.4344	1
11	2-chloro-6-methoxy-3-nitropyridine	38533-61-8	1.36	1.74	-1.4300	1
12	2-chloro-3,5-dinitropyridine	2578-45-2	2.64	0.84	-2.4456	1
13	2,6-dichloronitropyrimidine	16013-85-7	2.03	1.73	-1.7643	2
14	4,6-dichloro-5-nitropyrimidine	4316-93-2	2.12	0.44	-1.7643	2
15	4-chloronitrobenzene	100-00-5	0.43	2.39	-1.3440	1
16	2-chloronitrobenzene	88-73-3	0.68	2.52	-1.0753	1
17	3-chloro-2-methylnitrobenzene	83-42-1	0.68	3.09	-1.2187	1
18	3-chloro-4-fluoronitrobenzene	350-30-1	0.80	2.74	-1.5495	1
19	methyl-4-chloro-2-nitrobenzoate	42087-80-9	0.82	2.41	-1.5542	1
20	5-chloro-2-methylnitrobenzene	89-59-8	0.82	3.05	-1.2255	1
21	3-chloronitrobenzene	121-73-3	0.84	2.47	-1.2869	1
22	2,3-dichloronitrobenzene	3209-22-1	1.07	3.05	-1.2288	2
23	2,5-dichloronitrobenzene	89-61-2	1.13	3.03	-1.2939	2
24	3,5-dichloronitrobenzene	618-62-2	1.13	3.09	-1.4892	2
25	3,4-dichloronitrobenzene	99-54-7	1.16	3.12	-1.5249	2
26	2,4,6-trichloronitrobenzene	18708-70-8	1.43	3.69	-1.3404	3
27	2,3,5,6-tetrachloronitrobenzene	117-18-0	1.47	4.38	-1.4192	4
28	2,3,4-trichloronitrobenzene	17700-09-8	1.51	3.61	-1.4777	3
29	2,4,5-trichloronitrobenzene	89-69-0	1.53	3.47	-1.5435	3
30	2,3,4,5-tetrachloronitrobenzene	879-39-0	1.78	3.93	-1.6539	4
31	4-chloro-1,3-dinitrobenzene	97-00-7	1.98	2.14	-2.0613	1
32	2,4,6-trichloro-1,3-dinitrobenzene	Not known	2.19	2.97	-2.0382	3
33	1,2-dichloro-4,5-dinitrobenzene	6306-39-4	2.21	2.93	-2.2399	2
34	3,5-dichloro-1,2-dinitrobenzene	28689-08-9	2.42	2.85	-2.0925	2
35	1,3-dinitro-2,4,5-trichlorobenzene	2678-21-9	2.60	3.05	-2.1277	3
36	2,3,5,6-tetrachloro-1,4-dinitrobenzene	20098-38-8	2.74	3.44	-2.2138	4
37	2,6-dichloro-4-nitrophenol	618-80-4	0.66	2.94	-1.4418	2
38	2-chloromethyl-4-nitrophenol	2973-19-5	0.75	2.42	-1.1947	1
39	4-chloro-3-nitrophenol	610-78-6	1.27	2.46	-1.3407	1
40	2-chloro-4-nitrophenol	619-08-9	1.59	2.33	-1.2623	1
41	4-chloro-3-methyl-6-nitrophenol	7147-89-9	1.63	2.93	-1.1938	1
42	4-chloro-2-nitrophenol	89-64-5	1.67	2.47	-1.2296	1

All data sets in table 6.12 can be fitted with a combination of logP, E_{LUMO}, and N_{Cl} (r^2=0.714, F=35.76). It is notable that E_{LUMO} was the 'best' overall descriptor for all aromatic compounds. E_{LUMO} has often been considered to be more successful than logP and N_{Cl} in describing electrophilicity with regard to the toxicity of aromatic chemicals and N_{Cl} is retained as the second important descriptor. Figure 6.1 shows that, for each of the 4 chemicals groups, E_{LUMO} is a good descriptor in explaining the toxicity for most aromatic compounds except chlorinated nitrophenol. Similarly, Lu et al. (2001) investigated the toxicity of nitrobenzenes to P. phosphoreum and developed QSARs with E_{LUMO} and $logK_{ow}$, and concluded also that the toxicity of substituted nitrobenzenes is controlled mainly by electronic factors (E_{LUMO}). Model 6.9 indicates that the relationship between $log(1/IGC_{50})$ and independent predictors is quite good and is capable of explaining variation in data (R^2=0.92). But if the electronic reactivity of nitrobenzene is very weak, then the biological concentration in the organism may be the main factor controlling the toxicity, as stated by Yan et al. (2005).

Table 6.12 QSARs of the full and reduced data sets for chlorinated compounds

Model	Subsets	eq. $\log(1/IGC_{50})$ =	N	R^2	RMS	F
8	All data set	- 0.849-0.0306 logP-1.245 E_{LUMO}+0.22 N_{Cl}	47	0.714	0.404	35.76
9	Nitrobenzene	-1.134-0.0624 logP-1.383 E_{LUMO}+0.26 N_{Cl}	22	0.92	0.208	68.56
10	Benzonitrile	0.331-0.504 logP-1.321 E_{LUMO}	5	0.992	0.147	119.68
11	Nitrophenol	-1.168+0.719 logP-1.681 E_{LUMO}-1.355 N_{Cl}	6	0.514	0.501	0.71
12	Nitropyridine	-2.341-0.092 logP-1.681 E_{LUMO+}0.775 N_{Cl}	8	0.818	0.487	6.01

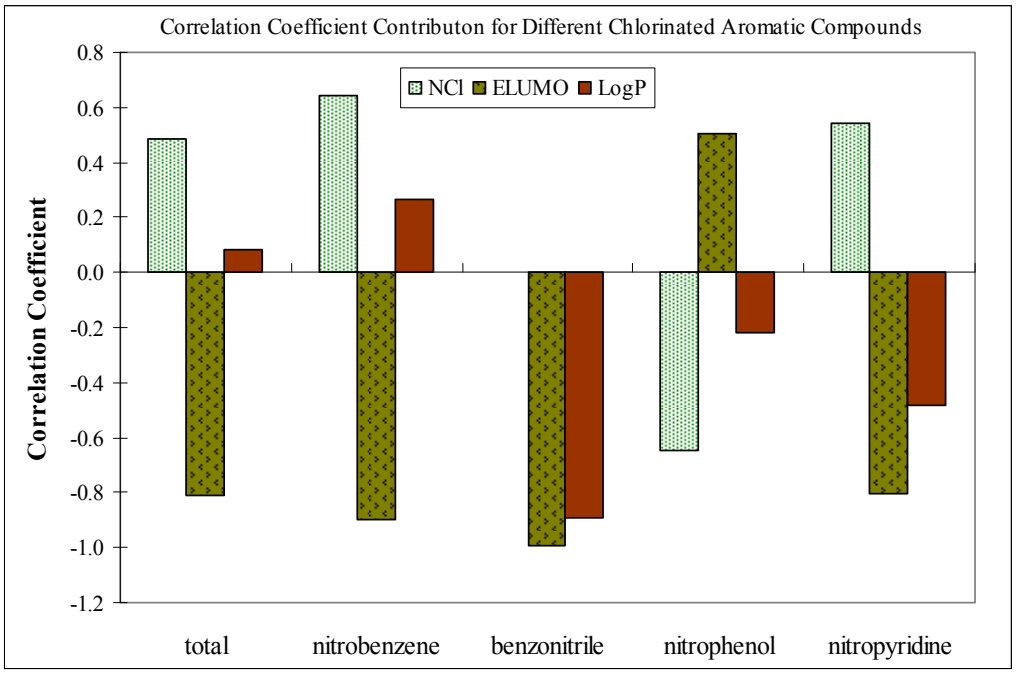


Figure 6.1 Correlation coefficient contributions to different chlorinated aromatic compounds

6.4 Discussion

The present results provide several examples of how computer-based quantum chemical calculated parameters have been used to define QSAR for analyzing experimental data on toxicity. Within the data set, a wide variety of toxic mechanisms of action is represented ranging from polar narcosis (e.g., chloronitrobenzenes) to chemicals capable of acting by electrophilic interactions with biological molecules. It is not possible to assign a definitive mechanism of action to each chemical in the data set because this knowledge is not currently available. It is this lack of knowledge concerning mechanisms of toxic action and the difficulty in assigning a mechanism for a novel chemical which makes mechanism of action-based QSAR impractical for prediction of many compounds (Cronin et al., 1998; Schultz and Mekenyan, 2001). Therefore, there has been

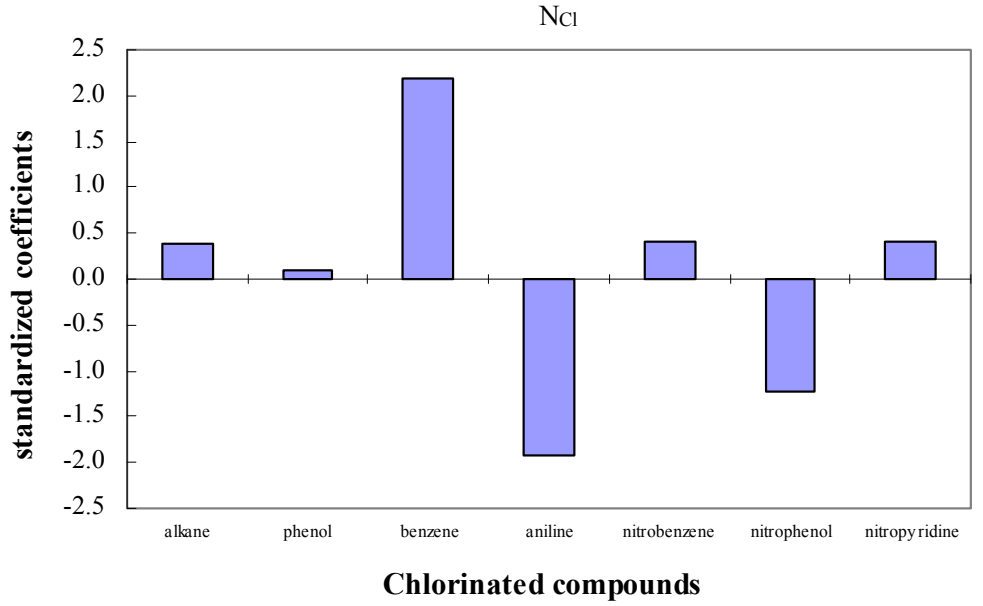
considerable interest in the development of QSAR that do not require a prior assignment of mechanism of action. Despite the lack of knowledge regarding specific mechanisms of toxic action for some compounds, it is recognized that, while it is not easy to quantify, electrophilicity is an important property governing the toxicity of these compounds.

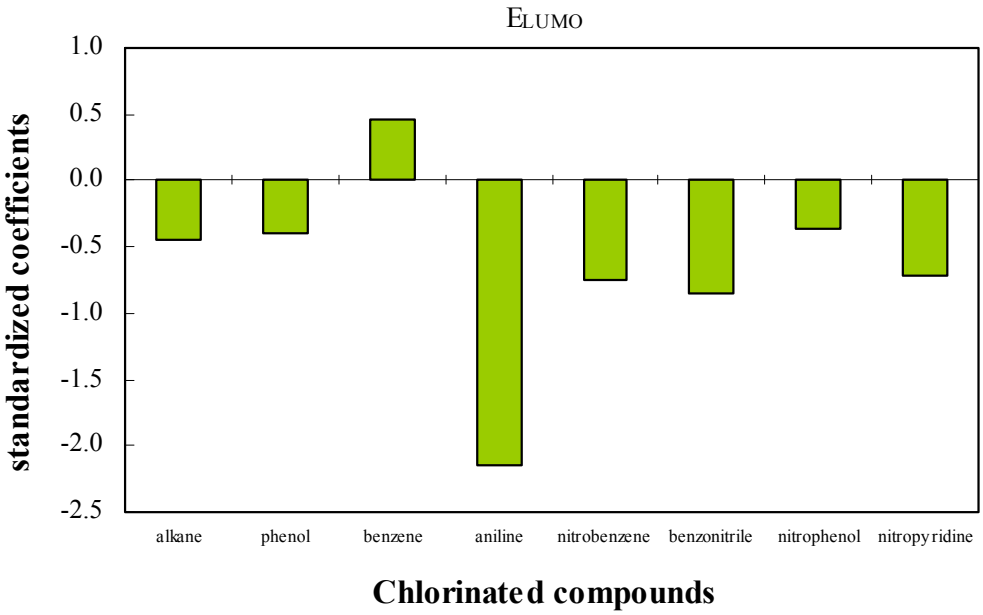
Table 6.13 The effect of N_{Cl} on correlation coefficient of QSAR models

Compounds	Equation	N	R^2	F	RMS
Chloro-alkane	in vivo $\log(1/EC_{50}) = -3.41-0.495\ \log P -1.923\ E_{LUMO}$	12	0.840	23.59	0.266
	in vivo $\log(1/EC_{50}) = -2.96+0.14\ \log P-6.27\ E_{LUMO}+0.16\ N_{Cl}+0.24\ N_C$	12	0.880	12.80	0.262
	in vitro $\log(1/EC_{50}) = -3.82+0.535\ \log P-1.599\ E_{LUMO}$	13	0.863	31.55	0.253
	in vitro $\log(1/EC_{50}) = -3.53+ 0.46\ \log P-3.3\ E_{LUMO}+0.01\ N_{Cl} +0.05\ N_C$	13	0.865	12.78	0.281
Chloro-phenol	$\log(IGC_{50})^{-1} = -0.38+0.495\ \log P-0.381\ E_{LUMO}$	35	0.629	28.79	0.401
	$\log(IGC_{50})^{-1} = -0.446+0.471\ \log P-0.488\ E_{LUMO}+0.0487\ N_{Cl}$	35	0.748	30.66	0.341
Chloro-benzene	$\log(1/EC_{50}) = 3.22-0.0082\ \log P -16.34\ E_{LUMO}$	10	0.559	4.436	0.223
	$\log(1/EC_{50}) = 0.623-0.478\ \log P+9.767\ E_{LUMO}+0.626\ N_{Cl}$	10	0.601	3.012	0.229
Chloro-aniline	$\mathrm{Log}(1/EC_{50}) = -8.35+0.37\ \log P+48.43\ E_{LUMO}$	14	0.574	7.406	0.341
	$\mathrm{Log}(1/EC_{50}) = -1.926-0.711\ \log P+24..591\ E_{LUMO}+1.261\ N_{Cl}$	14	0.636	5.818	0.331
Chloro-nitrobenzene	$\log(IGC_{50})^{-1} = -2.45-0.388\ \log P-1.69\ E_{LUMO}$	22	0.898	83.35	0.228
	$\log(IGC_{50})^{-1} = -1.134-0.0624\ \log P-1.383\ E_{LUMO}+0.26\ N_{Cl}$	22	0.92	68.56	0.208
Chloro-benzonitrile	$\log(IGC_{50})^{-1} = 0.331-0.504\ \log P-1.321\ E_{LUMO}$	5	0.992	119.7	0.147
Chloro-nitrophenol	$\log(IGC_{50})^{-1} = 4.39 -0.089\ \log P+2.27\ E_{LUMO}$	6	0.256	0.515	0.507
	$\log(IGC_{50})^{-1} = -1.168+0.719\ \log P-1.681\ E_{LUMO}-1.355\ N_{Cl}$	6	0.514	0.706	0.501
Chloro-nitropyridine	$\log(IGC_{50})^{-1} = -1.245-0.264\ \log P-1.737\ E_{LUMO}$	8	0.658	4.807	0.598
	$\log(IGC_{50})^{-1} = -2.341-0.092\ \log P-1.681*E_{LUMO+}0.775\ N_{Cl}$	8	0.818	6.008	0.487

Furthermore, to assess the effect of number of chlorine, two sets of QSAR models are developed with- or without N_{Cl}. The details are shown in the following table 6.13. This table clearly describes that, whenever N_{Cl} is introduced, the correlation coefficient is increased. This allowed for a comparison of the descriptors for the explanation of toxicity. The standardized coefficients of N_{Cl}, E_{LUMO} and logP were plotted by comparing variable

chlorinated chemicals (figure 6.2) and showed that the model derived from the structural parameters of a single component can be used successfully to predict the toxicity of compounds contained in a nitro-, cyano group.





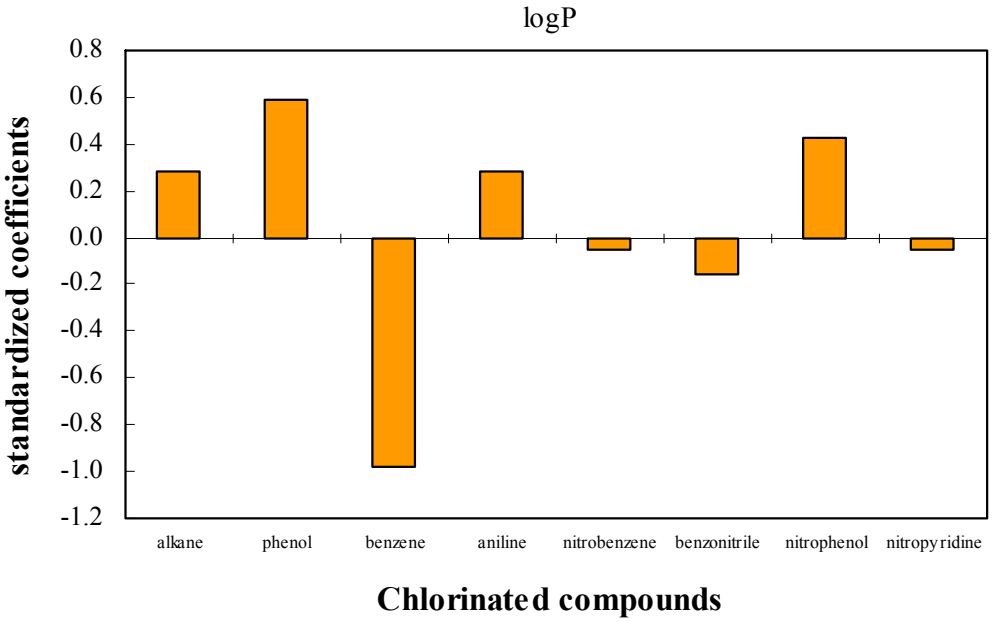


Figure 6.2 Contribution of N_{Cl}, E_{LUMO}, and LogP to toxicity for DBPs

Table 6.14 significant descriptors in QSAR models for various DBP chemicals

Compounds	Significant descriptors
Chloroalkane	logP
Chlorophenols	logP
Chloroanilines	E_{LUMO}, N_{Cl}
Chlorobenzenes	N_{Cl}, logP
Chloroamides	logP
Chloro-nitrobenzenes	E_{LUMO}, N_{Cl}
Chloro-benzonitriles	E_{LUMO}, N_{Cl}
Chloro-nitrophenols	E_{LUMO}, N_{Cl}
Chloro-nitropyridines	E_{LUMO}, N_{Cl}

The distribution of the values for each calculated variable can be analyzed to determine which parameters are significantly important. The obtained models revealed that the significances of descriptors are mainly related to the structures of chlorinated chemicals and toxicity mechanisms. For chloro- alkanes, phenols, benzenes and amides, the models

indicated that toxicities are mainly related to the hydrophobicity. LogP is a hydrophobicity parmater; the higher the logP, the stronger the hydrophobicity and the easier for the compounds to accumulate in an organism. Otherwise, the origin of toxicity of chlorinated compounds (chloro- aniline, aromatic containing a nitro/cyano group) has been attributed to the electron-accepting nature in charge of transfer complex with a receptor in living cells. Toxicity increases with greater negative ΔE. That is, the smaller the value of ΔE, the easier the electron transfers from HOMO orbital to LUMO orbital and the stronger the toxicity. Moreover, atom descriptor (N_{Cl}) also plays a good contribution to toxicity, that is, the chlorine substitution in DBP chemicals results, by its electron attracting effect, in the increase of the toxicities.

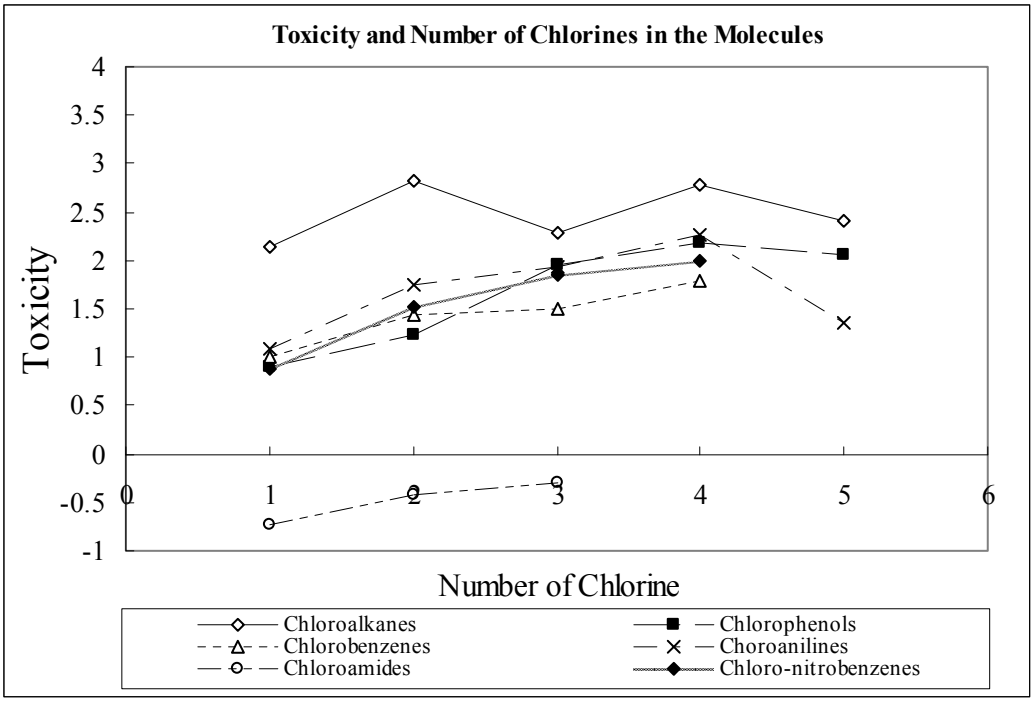


Figure 6.3 Relationship between number of chlorine and toxicity

Figure 6.3 presents the average toxicity values versus the number of chlorine atoms in the various chemical classes tested. Generally, the toxicity of compounds increases with

increasing number of chlorine atoms in the molecules. The position for each compound in the toxicity ranking tables 6.3, 6.5, 6.7, 6.9, and 6.11 were linked. From the ranking structure of chlorinated compounds, the more chlorine atoms concentrated at one C-atom, the higher the toxicity of the compound. CH_3-groups replacing chlorine increase toxicity as well for chlorinated alkanes. The structure influence on the toxic effects of chlorinated compounds is qualitatively recognizable.

6.5 Conclusion

QSAR models depend on the selection of appropriate descriptors. This study explores the usefulness of descriptors, especially conceptual molecular properties descriptors along with other descriptors, and analyzes the descriptors' applicability to drastic improvement in QSAR models. Based on this fact, structure-activity relationships for each data set including chloro- alkane, aromatic, amide and aromatic compounds containing a nitro- or cyano group on the toxicological behavior were analyzed. Traditional regression procedures along with cross-validation are carried out to evaluate the predicting power of the developed model. It has been shown that using the entire data set, logP with E_{LUMO} and a number of chlorine substituents as descriptors for chloro-alkane (r^2=0.88), nitrobenzene (r^2=0.92), benzonitrile (r^2=0.992), nitropyridine (r^2=0.818) provide a reasonably good coefficient of determination and cross-validated squared correlation coefficient values indicating the significance of the developed model.

CHAPTER 7 CONCLUSIONS AND RECOMMENDATION FOR FUTURE RESEARCH

To this point, we have discussed in details QSAR modeling for different classes with focus on a certain group of descriptors. In a comparative way of study, the results are summarized and presented in this chapter comparatively for different classes. The future work and possible research directions in this research area are also outlined.

7.1 Conclusions

In this thesis, QSAR models were developed for chlorinated alkanes (**chapter 3**), chlorinated alkenes (**chapter 4**), and chlorinated aromatics (**chapter 5**). Three descriptors were investigated for their molecular properties in modeling the physicochemical activity (such as E_{LUMO}, energy of the lowest unoccupied molecular orbital) of the chemicals in the five groups. These were:

a) Energy of the highest occupied molecular orbital (E_{HOMO}), which models the nucleophilic nature of the chemicals.

b) Number of chlorine (N_{Cl});

c) Number of carbon (N_C).

The initial task was to identify suitable software packages, descriptor calculation, and statistical techniques for use in the thesis. This was first done using a group of compounds called chlorine substituted alkanes as they had sufficient data for modeling purposes (**chapter 3**). Chlorinated alkanes are built from straight chains of carbon and hydrogen with varying numbers of hydrogen atoms replaced by chlorine atoms. The

introduction of chlorine atoms into the hydrocarbon chain alters molecular properties such as E_{LUMO}; the developed QSAR model is applicable to chlorinated alkanes with up to 10 carbon atoms, up to six –Cl, and E_{HOMO} values lying within the range from -0.4667 to -0.4114. Out of SAS software packages, number of chlorine atoms (N_{Cl}) was identified as the most suitable one for predicting molecular property, for instance, E_{LUMO}, a conclusion based on the highest correlation (r^2=0.956) of three-terms model between experimental and predicted E_{LUMO} values for a set of chloroalkanes. A two-stage model uncertainty composed of applying bootstrapping and an algorithm's identified noise sensitivity, subsequently. Three-terms QSAR model was assigned to estimate the corresponding uncertainty for every descriptor. Using the entire data set, N_{Cl}, N_C, and E_{HOMO} as descriptors provides a reasonably good coefficient of determination and RMSE value indicating the significance of the developed model. The QSARs were valid for neutral substituted alkanes with no other substitutents such as, –OH, -COOH, or –CN attached directly to the carbon chain.

In the next step, QSAR models were developed for chlorinated alkenes, a group of chemicals which is likely to be carcinogenic to humans by all routes of exposure. Cl-alkenes' are similar in structure to other chlorinated organics that are known to cause liver and kidney damage. It is known that Cl-alkenes are absorbed by all routes of exposure, and therefore, have a large potential for environmental pollution. Whereas in **chapter 3** there is sufficient data for substituted alkanes available for developing QSAR model, this was not the case for the chlorinated alkenes. Therefore, in **chapter 4**, chemical potential μ and chemical hardness η were calculated according to density function theory in order to generate QSAR models and eventually by correlated to the

limited available data to describe their acute molecular properties. It is important to note that, 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, $r^2_{5\text{-fold}}$=0.9963), its RMS values ($RMS_{(\text{training set})}$=0.00188, $RMS_{(\text{cross-val set})}$=0.00207)are all higher than in the 3-descriptors model 2. On the contrary, the 3-variables model, obtained by introducing E_{HOMO} in QSAR model (r^2=0.9709, r^2_{LOO}=0.9747, $r^2_{5\text{-fold}}$=0.9783), did not significantly increase the predictive performance of the nested model, as is evident when comparing the internal and external r^2_{cv} values. The developed QSAR models are applicable to chlorinated alkenes with up to 6 C-atoms, up to six Cl-atoms, and E_{LUMO} values lying within the range from 0.0768-0.147, and they cover 15 DBP chemicals.

The QSAR methodologies applied to chlorinated alkanes (**chapter 3**) and chlorinated alkenes (**chapter 4**) were extended to a third group of chemicals, chlorinated aromatic (**chapter 5**). Structure-activity relationship for the data set containing 53 choro- phenols, anilines, and benzenes congeners on the energy of lowest unoccupied molecular orbital (E_{LUMO}) is analyzed. The six QSAR models were developed and validated either internally or externally. The number of chlorine substituents (N_{Cl}) with the highest occupied molecular orbital (E_{HOMO}) energy as descriptors provides the reasonably good coefficients of determination ($0.9705<r^2<0.9955$) and cross-validated squared correlation coefficients ($0.9659<r^2_{cv}<0.9659$) value indicate the significance of the developed model.

Based on the previous study for chlorinated alkanes (**chapter 3**), chlorinated alkenes (**chapter 4**), and chlorinated aromatic (**chapter 5**), electrophilicity parameter with number of chlorine, number of carbon (for aliphatic), and the energy of the highest occupied molecular orbital (for chlorinated alkane and aromatic compounds) and it is

directly proportional to the electronic affinity of the compound, has been demonstrated to correlate with many physico-chemical properties for chloro-aliphatic and chloro-aromatic. Hence, it is logical to examine the relationship between the topology-based E_{LUMO} models and ecotoxicological properties.

In **chapter 6**, using experimental literature data sets on the acute toxicity of chlorinated alkanes, benzenes, anilines, phenols, nitro-phenols, and other substituted compounds to fish, *T. pyriformis*, and *photobacterium phosphoreum* to establish quantum chemistry-based QSARs were investigated. logP is an important descriptor in explaining the toxicity of chlorinated compounds with additional electronic descriptors, E_{LUMO}, with N_{Cl} and/or N_C being required for the targeted test system. Suitable QSAR models were derived for chloro- alkane (r^2=0.88), nitrobenzene (r^2=0.92), benzonitrile (r^2=0.992), nitropyridine (r^2=0.818) to provide a reasonably good coefficient of determination and cross-validated squared correlation coefficient values indicating the significance of the developed model. The obtained models revealed that toxicity of the most chlorinated chemicals was related mainly to their hydrophobicity (e.g. logP) and electronic properties (e.g. E_{LUMO}). Moreover, the chlorine substitution in DBP chemicals results, by its electron attracting effect, in the increase of the toxicities.

7.2 Recommendation for Research

QSAR models are expected to play an important role in the risk assessment of chemicals of DBPs. The results of this thesis reveal that, (i) number of chlorine atoms is the special descriptor for explaining the molecular activity; furthermore, it is responsible for the mechanisms of toxicity. (ii) despite the fact that individual QSAR may often each

cover only limited (i.e. less that 1% of the DBP compounds), QSAR approaches have the potential to cover more DBP compounds due to their properties of being analogous to assessed ones.

In general, carcinogenicity has been the primary driving force behind drinking water regulations, and it is likely that carcinogenicity will continue in this role, although other health effects end points may also be of concern. Toxicology end points are the subject of the current EPA investigation, and the area of other end points for human health effects could be an interesting area of DBP research for the future. Another area of future DBP research is in the 60% or so of the halogenated material that is not part of the identifiable classes of compounds (i.e., HAAs, THMs, haloacetonitriles) (USEPA, 2001).

One clearly positive aspect on the development of new QSAR methodologies is the ever-increasing computing capacity that will allow a great increase in the flexibility and power of QSAR techniques in the future. In my opinion, one of the great challenges in developing the new generation of QSAR methodologies lies in the incorporation of the dynamic nature of the molecules.

Further research in several areas will enhance the development of high-quality correlations. The effects of combining toxicants need to be researched. New advances in analytical chemistry may complement the use of QSAR to provide data to future enlarged QSARs and interspecies correlations. This method can be cheaper and faster than traditional animal toxicological studies.

REFERENCES

Baj, S., David, M., 1994. Correlation between the chemistry structures of dialkyl peroxides and their retention in reversed-phase higherformance liquid thromatography, J. Liquid Chromat. 17, pp. 3933-3949.

Baroni, M., Clementi, S., Crucinani, G., Costantino, G., Riganelli, D., Oberrauch, E., 1989. Predictive Ability of Regression Models. Part II: Seleciton of the Best Predictive PLS Model. J. of Chemometrics, 6, pp. 347-356.

Benigni, R., Giuliani, A., Franke, R., Gruska, A., 2000. Quantitative Structure-Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. Chem. Rev. 100, pp. 3697-3714.

Bevington, P. R. and Robinson, D. K., 1992. Data Reduction and Error Analysis for the Physical Sciences 2nd edn (Boston, MA: McGraw-Hill) pp. 328

Blaha, L., Damborsky, J. and Nemec, M., 1998. QSAR for acute toxicity of saturated and unsaturated halogenated aliphatic compounds, Chemosphere 36, pp. 1345-1365

Blum, D.J.W., Speece, R.E., 1990. Determining chemical toxicity to aquatic species. Environ. Sci., Technol., 63, pp. 198-207.

Breiman, L. and Spector, P., 1992. Submodel selection and evaluation in regression: the X-random case. Int. Stat. Rev., 60, pp. 291-319.

Coleman, H.W., Steele, W.G., 1995. Engineering application of experimental uncertainty analysis, AIAA J. 33 (10), pp. 1888-1896.

Cronin, M.T.D., Aptula, A.O., Duffy, J.C., NETZEVA, T.I., Rowe, P.H., Valkova, I.V., Schultz, T.W., 2002. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 49, pp. 1201-1221.

Cruciani, G., Baroni, M., Costantino, G., Riganelli, D., Skagerberg, B., 1992. Predictive Abiity of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP). J. of Chemometrics, 6, pp. 335-346.

Cronin, M.T.D., Manga, N., Seward, J.R., Sinks, G.D., Schultz, T.W., 2001. Parametrization of Electrophilicity for the Prediction of the Toxicity of Aromatic Compounds. Chem. Res. Toxico. 14, pp. 1498-1505.

Cronin, M.T.D., Schultz, T.W., 1996. Structure-toxicity relationships for phenols to Tetrahymena pyriformis. Chemosphere 32, (8), pp. 1453-1468.

Czaplicka, M., 2004. Sources and transformations of chlorophenols in the natural environment. Science of the Total Environment 322, pp. 21-39.

Dai, J.Y., 1998. Prediction of water solubility and toxicity of substituted indoles to Photobacterium phosphoreum by using molecular connectivity indices and quantum chemical parameters. Bull. Environ. Contam. Toxicol. 61, pp. 591-599.

Dedonder, A., Van Sumere, C.F., 1971. The effect of phenolics and related compounds on the growth and the respiration of Chlorella vulgaris. Z. Pflanzenphysiol. 65, pp. 70-80.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003. Methods for Reliability and uncertainty assessment and for applicability evaluations of classification- and regression- based QSARs, Environ. Health Perspect., 111 (10), pp. 1361-1375.

Eriksson, L., Johansson, E., Muller, M., Wold, S., 2000. On the selection of training set in environmental QSAR analysis when compounds are clustered. J. Chem. 14, pp. 599-616.

Eriksson, L., Johansson, J.S., Worth A. P., and Cronin M. T. D., 2003. Methods for Reliability and Uncertainty Assessment, and Applicability Evaluations of QSARs: Existing Methods, Environ. Health Perspect., In Press.

Feng, C.X.J., Yu, Z.G.S., Emanuel, J.T., Li, P.G., Shao, X.Y., Wang, Z.H., 2008. Threefold verse fivefold cross-validation an individual versus average data in predictive regression modeling of machining experimental data, International J. Comp. Integ. Manufacturing, 21 (6), pp. 702-714.

Fleming, I. 1976. Molecular Orbitals and Frontier Orbitals., In Frontier Orbitals and Organic Chemical Reactions, John Wiley & Sons, New York.

Fragiadakis, A., Sotiriou, N., Korte, F., 1981. Absorption, balance and metabolism of 14C-2,4,6-trichlorophenol in hydroponic tomato plants. J. Chem. 10, pp. 1315-1320.

Freitag, D.,Ballhorn, L., Korte,S., Korte, F, 1990. Bioaccumulation and degradation of some nitroalkanes. In: Practical applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology (Eds.: W. Karcher, H. Devillers). Brussel, Luxembourg: ECSC, EEC, EAEC, pp. 371-388.

Freitag, D., Ballhorn, L., Behechti, A., Fischer, K., Thumm, W., 1994. Structural Configuration and Toxicity of Chlorinated alkanes, J. Chemosphere, 28 (2), pp. 253-259.

Gallagher, D.A., 2001. Predicting environmental remediation rates correlation analysis offers a means for estimating the reaction rates of organic pollutants, J. Chemist, 10 (10), pp. 57-60.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., Tropsha, A., 2003. Rational Selection of Training Sets for the Development of Validated QSAR Models. J. Comput. Aided Mol. Des. 17, pp. 241-253.

Golbraikh, A., Tropsha, A., 2002. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training Set Selection. J. Comput. Aided Mol. Des. 16, pp. 357-369.

Gombar, V.K. and Enslein, K., Blake, B.W., 1988/89. In Vitro Toxicol. 2, 117-127.

Gramatica, P., Papa, E., 2005. An update of the BCF QSAR model based on theoretical molecular descriptors, QSAR Comb. Sci. 24, 8, pp. 953-960.

Gramatical, P., Pilutti, P., 2004. Evaluation of Different Statistical Approaches for the Validation of Quantitative Structure-Activity Relationships, Joint Research Centre (JAC) Contract ECVA-CCR.496576-Z.

Hammett, L.P., 1937. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivative, J.Am. Chem. Soc., 59 (1), pp. 96-103.

Hansch, C., Fujita, T., 1964. ρ-σ-π analysis. A method for the correlation of biological activity and chemistry structure. J. Am. Chem. Soc. 86, pp. 1616-1626.

Hatch, F.T., Colvin, M.E., 1997. Quantitative structure-activity (QSAR) relationships of mutagenic aromatic and heterocyclic amines. Mutat Res 376, pp. 87-96.

Henk J. M., Verhaar, L., Eriksson, M. S., 1994. Modeling the Toxicity of Organophosphates: a Comparison of the Multiple Linear Regression and PLS Regression Methods, Quant. Struct.-Act. Relat., 13, pp. 133-143

Hoos, R.A.W., 1978. Patterns of pentachlorophenol usage in Canada-An overview. In K.R. Rao. ed., Pentachlorophenol. Chemistry, Pharmacology and Environmental Toxicology. Plenum, New York, NY, USA, pp. 3-12.

Jackson, J. E., 1991. A User's Guide to Principal Components, New York: John Wiley & Sons, Inc.

Kaiser, K.L.E.; McKinnon, M.B., 1994. Computox Toxicity Database, Version 4.01; National Water Research Institute: Burlington.

Koopmans, T., 1934. Physica 1, pp. 104-110.

Korhonen, S. P., 2007. FLUFF-BALL, a Fuzzy Superposition and QSAR Technique-Towards an Automated Computational Detection of Biologically Active Compounds Using Multivariate Methods. Kuopio University Pubicaitons C. National and Environmental Sciences 206, 154 p.

Kottegoda, N.T., Rosso, R., 1998. Statistics, probability and reliability for civil and environmental engineers, *McGraw-Hill.*

Lewis, O.F.V., 1989. The calculation of molar polarizabilities by the CNDO/3 method: Correlation with the hydrophobic parameter LogP. J. Comput. Chem. 10, pp. 145-151.

Livingstone, D.J., 1995. Data Analysis for Chemists: Application to QSAR and Chemical Product Design. Oxford University Press, Oxford.

Lu, G.H., Yang, X., Zhao, Y. H., 2001. QSAR study on the toxicity of substituted benzenes to the algae (Scenedesmus obliquus), Chemosphere 44, pp. 437-440.

Makinen, P.M., Theno, T.J., Ferguson, J.F., Ongerth, J.E., Puhakka, J.A., 1993. Chlorophenol toxicity removal and monitoring in aerobic treatment: recovery from process upsets. Environmental Science and Technology 27, pp. 1434-1439.

Mallakin, A., Dixon, D.G., Greenberg, B.M., 2000. Pathway of anthracene modification under simulated solar radiation. Chemosphere, 40, pp. 1435-1441.

Mallakin, A., Mezey, P.G., Zimpel, Z., Berenhaut, K.S., Greenberg, B.M., Dixon, D.G., 2005. Use of Quantitative structure-activity relationship to model the photoinduced toxicity of anthracene and oxygenated anthracene. QSAR & Comb. Sci., 24, pp. 844-852.

Melek, T.S., Inel, Y., 1995. Application of the characteristic root index model to the estimation of N-octanol/water partition coefficients. polychlorinated biphenyls, Chemosphere 30, pp. 39–50.

Murcia-Soler, M., Perez-Gimenez, F., Nalda-Molina, R., Salabert-Salvador, M.T., Garcia-March, F.J., Cercos-del-Pozo, A., Garrigues, T.M., 2001. QSAR Analysis of Hypoglycemic Agents Using the Topological Indices. J. Chem. Inf. Comput. Sci., 41, 5, pp. 1345-1354.

Netzeva, T.I., Schultz T.W., Aptula A.O., Cronin M.T., 2003. Partial Least Squares Modeling of the Acute Toxicity of Aliphatic Compounds to Tetrahymena pyriformis, SAR and QSAR in Environmental Research, 14, pp. 265-283.

Nevalainen, T., Kolehaminen, E., 1994. New QSAR models for polyhalogenated aromatics. Environ. Toxicol. Chem. 13, pp. 1699-1706.

Oberg, T., 2005. A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application, and prediction, *Atmospheric Environment*, 39, pp. 2189-2200

OECD, 2004. The report from the expert group on (Quantitative) Structure-Activity Relationship ((Q)SARs) on the principles for the validation of (Q)SARS. ENV/JM/MONO, 24., In OECD Series on testing and assessment: number 49.

OECD, 2007. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships (QSAR) Models. Environment Directorate Joint Meeting of The Chemicals Committee and The working Party on Chemicals, Pesticides and Biotechnology, ENV/JM/MONO 2

Padmanabhan, J., Parthasarathi, R., Chattaraj, P.K., 2006. QSPR models for polychlorinated biphenyls: n-Octanol/water partition coefficient. Bioorganic & Medicinal Chemistry 14, pp. 1021-1028.

Papa, E., Dearden, J.C., Gramatica, P., 2007. Linear QSAR regression models for the prediction of bioconcentraiton factors by physicochemical properties and structural theoretical molecular descriptors, Chemosphere 67, pp. 351-358

Peller, J., Wiest, O., Kamat, P.V., 2003. Synergy of combining sonolysis and photocatalysis in the degradation and mineralization of chlorinated aromatic compounds. Environmental Science and Technology 37, pp. 1926-1932.

Puhakka, J., Melin, E., 1996. In: Crawford, R., Crawford, D. (Eds.), Bioremediation: Principles and Applications. Cambridge University Press, Cambridge, MA, USA, pp. 254–299.

Rappe, C., 1980. Chloroaromatic compounds containing oxygen: Phenols, diphenyl ethers, dibenzo-p-dioxins and dibenzofurans. In O. Hutzinger, ed., The Handbook of Environmental Chemistry. Springer-Verlag, Berlin, Germany, pp. 157-179.

Ribeiro, F.A.L., and Ferreira, M.M.C., 2003. QSPR models of boiling point, octanol-water partition coefficient and retention time index of polycyclic aromatic hydrocarbons. J. Mole. Struc. 663, pp. 109-126.

Ribo, J. M., Kaiser, K.L.E., 1984. Toxicities of Chloroanilines to Photobacterium phosphoreum and Their Correlations with Effects on Other Organisms and Structural Parameters In QSAR in Environmental Toxicology; K.L.E. Kaiser, Ed; Reidel Publishing Company: Dordrecht, pp. 319-336.

Pearson, P.G., 1986. Absolute electronegativity and hardness correlated with molecular orbital theory. Proc. Natl. Acad. Sci. 83, pp. 8440-8441.

Safe, S., 1990. Polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PC-DDs), dibenzofurans (PCDFs) and related compounds: Environmental and mechanistic considerations which support the development of toxic equivalency factors (TEFs). *Crit. Rev. Toxicol.* 21, pp. 51–88.

Sagrado, S., Cronin, M.T.D., 2006. Diagnostic Tools to Determine the Quality of "Transparent" Regression-Based QSARs: The "Modelling Power" Plot, *J.Chem. Inf. Model.* 46, 3, pp. 1523-1532.

Schmitt, H., Altenburger, R., Jastorff, B., Schuurmann, G., 2000. Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds. Chem. Res. Toxicol. 13, pp. 441-450.

Schultz, T.W., Cronin, M.T.D., Walker, J.D., Aptula, A.O., 2003. Quantitative structure-activity relationship (QSAR) in toxicology: a historical perspective. Journal of Molecular Structure: THEOCHEM 622, pp. 1-22.

Schultz, T.W., Sinks, G.D., Cronin, M.T.D., 1997. Identification of mechanisms of toxic action of phenols to Tetrahymena pyriformis from molecular descriptors. In: Chen, F., Schuurmann, G. (Eds.), Quantitative Structure–Activity Relationships in Environmental Sciences–VII. SETAC Press, Pensacola, FL, USA, pp. 329–342.

Schultz, T.W., Sinks, G.D., Cronin, M.T.D., 2000. Effect of substituent size and dimensionality on potency of phenolic xenoestrogens evaluated with a recombinant yeast assay. Environ. Toxicol. Chem. 19 (11), pp. 2637–2642.

Schultz, T.W., Yarbrough, J.W. and Koss, S.K., 2006. Identification of reactive toxicants: Structure-activity relationships for amides, Cell Biol Toxicol, 22:339-349

Sixt, S., Altschuh, J., Bruggemann, R., 1995. Quantitative Structure-Toxicity Relationships for 80 Chlorinated Compounds Using Quantum Chemical Descriptors. Chemosphere, Vol. 30, No. 12, pp. 2397-2414.

Sjostrom, M., Eriksson, L., 1995. Applications of Statistical Experimental Design, in Chemometric Methods in Molecular Design. Van de Warerbeemd H., VCH, 2, pp. 63-90.

Soffers, A.E.M.F., Boersma, M.G., Vaes, W.H.J., Vervoort, J., Tyrakowska, B., Hermens, J.L.M. and Rietjens, I.M.C.M., 2001. Computer-modeling-based QSARs for analyzing experimental data on biotransformation and toxicity. Toxicoogy in Vitro 15, pp. 539-551.

Taft, R.W., 1956. Separation of Polar, Steric and Resonance Effects in Reactivity, in Steric Effects in Organic Chemistry, M.S. Newman (Ed.), John Wiley, New York, 556.

Taylor, J.R., 1997. An introduction to error analysis. The study of uncertainties in physical measurements (2nd ed), Uncertainty Science Books, [ISBN 0-935702-42-3].

Terada, H., 1990. Uncouplers of oxidative phosphorylation. Environ. Health Perspect. 87, pp. 213–218.

Thomsen, M., 2001. QSARs in Environmental Risk Assessment. Interpretation and Validation of SAR/QSAR Based on Multivariate Data Analysis. PhD thesis.

Topliss, J.G., Edwards, R.P., 1979. Chance Factors in Studies of QSAR. J. Med. Chem. 22, pp. 1238-1244.

Uchimura, T., Deguchi, T., and Imasaka, T., 2005. Development of a Narrow-Band Tunable Picosecond Dye Laser and Its Application to Exicted-State Lifetime Measurement of a Chlorinated Aromatic Hydrocarbon, Analytical Sciences, Vol. 21.

USEPA, 2001. Controlling Disinfection By-Products and Microbial Contaminants in Drinking Water. EPA-600-R-01-110. Office of Research and Development, Washington DC.

Verhaar, H.J.M., Eriksson, L. and Sjostrom, M., 1994. Modeling the toxicity of organophosphates-A comparison of the Multiple Linear-Regression and PLS regression methods Quantitative Structure-Activity Relationships, Quant.Struct. Act. Relat. 13, pp. 133-143.

Walker, J.D., Jaworska, J., Comber, M.H.I., Schultz, T.W. and Dearden, J.C., 2003. Guidelines for developing and using Quantitative Structure Activity Relationships. Environmental Toxicology and Chemistry 22, pp. 1653-1665.

Wang, K.H., Hsieh, Y.H., Chou, M.Y., Chang, C.Y., 1999. Photocatalytic degradation of 2-chloroand 2-nitrophenol by titanium dioxide suspensions in aqueous solution. Applied Catalysis B 21, pp. 1-8.

Wang, Z.Y., Zhai, Z.C., Wang, L.S., 2004. Quantitative Structure-Activity Relationship of Toxicity of Alkyl(1-phenylsulfonyl) Cycloalkane-carboxylates Using MLSER Model and Ab initio. QSAR & Comb. Sci., 24, pp. 211-217.

Wold, S., Johansson, E., Cocchi, M., 1993. PLS: Partial Least Square Projections to Latent Structures, in H. Kubinyi (ed.), 3D-QSAR in Drug Design: Theory, Methods, and Applications, pp. 523-550, ESCOM Science, Leiden, The Netherlands.

Wold, S., Ruhe, A., Wold, H., 1984. The Collinearity Problem in Linear Regression, the Partial Least Square (PLS) Approach to Generalized Inverses. SIAM J. of Sci. Stat. and Comput., 5, pp. 735-743.

Woo, Y.T., Lai, D., McLain, J.L., Manibusan, M.K., Dellarco, V., 2002. Use of Mechanism-based Structure-Activity Relationships Analysis in Carcinogenic Potential Ranking for During Water Disinfection By-Products. Environ. Health Perspect, 110 (1), pp. 75-87.

Worth, A.P., Bassan A., Gallegos A., Netzeva T.I., Patlewicz G., Pavan M., Tsakovska I., Vracko M., 2005. The Characterisation of (Quantitative) Structure-Activity Relationship: Preliminary Guidance. EUR 21866 EN, European Commission Directorate General Joint Research Center.

Xu, S., Li, L., Tan, Y., Feng, J., Wei, Z., Wang, L., 2000. Prediction and QSAR Analyisis of Toxicity to Photobacterium phosphoreum for a Group of Heterocyclic Nitrogen Compounds. Bull. Environ. Contam. Toxicol. 64, pp. 316-322.

Yan, X.F., Xiao, H.M., Gong, X.D., Ju, X.H., 2005. Quantitative structure-activity relationships of nitroaromatics toxicity to the algae(Scenedesmus obliquus). Chemosphere 59, pp. 467-471

Zhang, L., Zhou, P.J., Yang, F., Wang, Z. D., 2007. Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives. Chemosphere 67, pp. 396-401.

Zhang, P., 1993. Model selection via multifold cross-validation. Annls Statistics, 21, pp. 299-311.

Zvinavashe, E., Berg, H.V.D., Soffers, A.E.M.F., Vervoort, J., Freidig, A., Murk, A.J., Rietjens, I.M.C.M., 2008. QSAR Models for Predicting in Vivo Aquatic Toxicity of Chlorinated Alkanes to Fish. Chem. Res. Toxicol. 21, pp. 739-745.

VITA

FANG WANG

May 31, 1980	Born, Shanxi Province, P.R. China
2003	B.A., Environmental Engineering Taiyuan Uvinersity of Technology Shanxi Province, P.R. China
2006	M.S., Environmental Engineering Taiyuan Uvinersity of Technology Shanxi Province, P.R. China
2006-2009	Teaching Assistant and Research Assistant Florida International University Miami, Florida
2010	Doctoral Candidate in Civil Engineering Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Fang Wang and Zengzhang Wang, "Research on the Application of the Techniques for Treating the High-concentration Organic Wastewater", SCL/TECH Information, Development & Economy, vol. 15, no.9, pp. 1005-1033.

Walter Z. Tang, Emile Damisse, Fang Wang, and Fernando Miralles-Wilhelm, "Uncertainty Analysis of Flow Rating Equation for Uncontrolled Submerged Flow over a Spillway", ISSAT RQD conference, 2009.

Emile Damisse, Juan Gonzalez, Fang Wang, Fernando Miralles-Wilhelm, and Walter Z. Tang, "Uncertainty Analysis of Rating Equation of Submerged Orifice Flow at Gated Spillways", ISSAT RQD conference, 2009.

Fang Wang and Walter Z. Tang, "Quantitative Structure Activity Relationship (QSAR) of Chlorine Effects on E_{LUMO} of Disinfection By-Product: Chlorinated Alkanes", Chemosphere, 2009.

Fang Wang and Walter Z. Tang, "Quantitative Structure Activity Relationship (QSAR) Study of Chlorine Effects on E_{LUMO} of Chlorinated Alkenes", SAR and QSAR in Environmental Research, 2009. (Under review)

Walter Z. Tang, Fang Wang, and Fabius D. Foti, "Photocatalytic Oxidation of Halogen Substituted Meta-Phenols by UV/TiO_2 in Acidic and Basic Aqueous Solutions", Chemosphere, 2009. (Under review)

Oral Presentation, Crystal Ball Software Presentation, South Florida Water Management District, West Palm Beach, Florida, United States, 2008.

Oral Presentation, Quantitative Structure Activity Relationship (QSAR) Study of number of Chlorine Effects on Molecular Properties of Disinfection By-products. Florida International University, Miami, Florida, United States, 2008.