PREDICTION OF AQUATIC TOXICITY OF PESTICIDES BY USING LINEAR AND NONLINEAR TECHNIQUES

by

Gülçin Tuğcu

B. S. in Mathematics, Hacettepe University, 1995

M. S. in Applied Mathematics, Georgia Southern University, 2004

Submitted to the Institute of Environmental Sciences in partial fulfillment of the requirements for the degree of

Master of Science

in

Environmental Sciences

Boğaziçi University

2011

PREDICTION OF AQUATIC TOXICITY OF PESTICIDES BY USING LINEAR AND NONLINEAR TECHNIQUES

APPROVED BY:

Prof. Dr. Melek Türker Saçan ……
(Thesis Supervisor)

Prof. Dr. Ferhan Çeçen ……

Prof. Dr. Safiye Erdem ……

DATE OF APPROVAL: 03 February 2011

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor, Prof. Dr. Melek Türker Saçan for her continuous guidance and support for this thesis. Without her help, success in this study would never have been possible. I also thank my jury members Prof. Dr. Ferhan Çeçen and Prof. Dr. Safiye Erdem for their help and advices.

I would like to thank research assistant Doğa Ertürk for his help and support. I would also like to thank our TÜBİTAK project partners Marjan Vracko, Marjana Novic, and Nikola Minovski from National Institute of Chemistry, Ljubljana, Slovenia for their help and support.

I thank numerous people from Boğaziçi University Computer Engineering Department for their help and support.

Finally, I thank my family members for their patience, help, and support throughout my study.

This work was partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under project number 108Y119.

PREDICTION OF AQUATIC TOXICITY OF PESTICIDES BY USING LINEAR AND NONLINEAR TECHNIQUES

Economical and environmental considerations for assessing toxicity of chemicals have led to a considerable amount of studies on the computational techniques. Pesticides allocate a significant part in these chemicals, mainly for their toxic effects on nontarget organisms. In the present study, the toxicities of 91 organic compounds including pesticides to freshwater algae, *Chlorella vulgaris*; and the toxicities of a set of 34 pesticides to *Oncorhynchus mykiss* were modeled employing Counter Propagation Neural Network (CPNN) and Multiple Linear Regression (MLR). The analyses were performed with about 1500 descriptors calculated using Dragon 5.4, Spartan 06, and Codessa 2.2 software. Additionally, we used the Characteristic Root Index (*CRI*) which was proved to be a significant descriptor in previous QSPR/QSTR studies. Descriptor selection was made by Heuristic Method. Kohonen network was used for splitting the data set into training and test sets. Linear and nonlinear 3, 4 and 5-descriptor models were compared according to their statistics such as squared correlation coefficient and Root Mean Squared Error (*RMSE*). All models were validated externally by using test sets. BLTD48 from Dragon, electrophilicity from Spartan, and the *CRI* appeared to be significant for the developed QSTR models of *Chlorella vulgaris*. *Oncorhynchus mykiss* model underscores the Dragon descriptors. The statistical quality of the models for *Chlorella vulgaris* is compared to those of the previously published models using the same experimental data and found to be superior to those models. *Oncorhynchus mykiss* models are compared to literature models in terms of chemical classes, mechanism of action, and statistical tools and fits. Linear and nonlinear methods were found to be comparable for both species.

LİNEER VE LİNEER OLMAYAN YÖNTEMLERLE PESTİSİTLERİN SUDAKİ TOKSİSİTELERİNİN TAHMİN EDİLMESİ

Ekonomik ve çevresel faktörler kimyasalların toksisitelerinin belirlenmesi konusundaki çalışmaların büyük bir çoğunluğunu hesaplama yöntemlerine yöneltmiştir. Pestisitler, hedef olmayan canlılar üzerindeki zararlı etkileri nedeniyle, kimyasallar içinde önemli bir yer tutarlar. Bu çalışmada, pestisit de içeren 91 organik kimyasalın tatlı su algi *Chlorella vulgaris*'e ve 34 pestisitin *Oncorhynchus mykiss*'e olan toksisiteleri Counter Propagation Neural Network (CPNN) ve Çoklu Doğrusal Regresyon (MLR) ile modellendi. Analizler Dragon 5.4, Spartan 06 ve Codessa 2.2 programları kullanılarak hesaplanan yaklaşık 1500 tanımlayıcı ile yapıldı. Ek olarak, daha önceki QSPR/QSTR çalışmalarda önemli bir tanımlayıcı olduğu kanıtlanan Karakteristik Kök İndisi (*CRI*)'ni de kullandık. Tanımlayıcı seçimi Heuristic Yöntem'le yapıldı. Veri setini eğitim ve test setlerine ayırmada Kohonen ağları kullanıldı. Doğrusal ve doğrusal olmayan 3, 4 ve 5 tanımlayıcılı modeller korelasyon katsayısının karesi ve ortalama hatanın karekökü (*RMSE*) gibi istatistiklerine göre karşılaştırıldı. Bütün modellerin validasyonu test setler kullanılarak yapıldı. Dragon'dan BLTD48, Spartan'dan elektrofilisite ve *CRI*'ın *Chlorella vulgaris* için geliştirilen QSTR modellerinde önemli olduğu görüldü. *Oncorhynchus mykiss* modeli Dragon tanımlayıcılarını öne çıkardı. *Chlorella vulgaris* için olan modellerin istatistiksel kalitesi aynı veri seti kullanılarak yayınlanmış modellerle karşılaştırılmış ve bu modellerden daha üstün olduğu görülmüştür. *Oncorhynchus mykiss* modelleri literatür modelleriyle kimyasal sınıf, etki mekanizması, istatistiksel yöntemler ve uygunluk açısından karşılaştırıldı. Lineer ve lineer olmayan yöntemlerin her iki canlı türü için de karşılaştırılabilir olduğu görüldü.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ ABBREVIATIONS

Symbol	Explanation	Unit
AAE	Average Absolute Error	
AChE	Acetyl cholinesterase	
AD	Applicability Domain	
AE	Absolute Error	
ANN	Artificial Neural Networks	
BLTD48	Daphnia base-line toxicity from MLOGP (mM)	
CAS	Chemical Abstracts Service	
CPNN	Counter Propagation Neural Networks	
CRI	Characteristic Root Index	
DDD	Dichlorodiphenyl dichloroethane	
DDE	Dichlorodiphenyl dichloroethylene	
DDT	Dichloro diphenyl trichloroethane	
E	Energy	eV
EC_{50}	Concentration of a compound that causes 50% effect on test organism relative to a control.	mM
ECOTOX	ECOTOXicology database	
E_{HOMO}	Energy of the highest occupied molecular orbital	eV
E_{LUMO}	Energy of the lowest unoccupied molecular orbital	eV
EPA	Environmental Protection Agency	
F	Fischer statistic	
h^*	Critical hat value	
HATS6m	Leverage-weighted autocorrelation of lag 6/ weighted by atomic masses	
HM	Heuristic Method	
K_{ow}	Octanol-water partition coefficient	

LC_{50}	Concentration of a compound that causes 50% lethality of the test organisms in a batch assay.	mM
LOO	leave-one-out statistical procedure	
LMO	leave-many-out statistical procedure	
MLR	Multiple Linear Regression	
MOA	Mechanism of Action	
Mor27u	3D-MoRSE-signal 27/ unweighted	
MSE	Mean Squared Errors	
n	Number of compounds	
*p*T	Negative logarithm of toxic concentration	
QSAR	Quantitative Structure-Activity Relationship	
QSPR	Quantitative Structure-Property Relationship	
QSTR	Quantitative Structure-Toxicity Relationship	
R	Correlation coefficient	
RDF080p	Radial Distribution Function – 8.0 /weighted by atomic polarizabilities	
RMSE	Root Mean Squaered Error	
SE	Standard Error	
SOM	Self-Organizing Maps	
SR	Standardized Residual	
SSE	Sum of Squared Errors	
VIF	Variance Inflation Factor	
X1sol	Solvation connectivity index (χ^{-1})	
ω	Electrophilicity	

1. INTRODUCTION

The increasing number of chemicals around us raises the problem of characterization, prediction, and evaluation of their consequences to human health and the environment (Gini et al., 2004). Since all pesticides are assumed to be hazardous to environment, their properties and toxicities should be known before they are used. The toxicity of new chemicals (pesticides) can be determined in two ways: (i) conducting experiments; (ii) predicting their values via modeling.

Experimental investigations can be carried out to collect toxicity values of pesticides. However, the data collection procedure is extremely time consuming (Tao et al.; 2002). Additionally, the cost of *in vivo* testing is prohibitive and weighs heavily on the final price of chemicals. Beside economic constraints, ethical considerations and public pressure work to reduce tests on animals (Mazzatorta et al., 2005).

Utilization of Quantitative Structure-Activity/Property/Toxicity Relationships (QSAR/QSPR/QSTR) can be an alternative way to predict toxic effects (Wang et al., 2000, Tao et al., 2002, Bermudez-Saldana and Cronin, 2006). Application of QSAR/QSTR to the aquatic toxicology field has started drawing attention in the late 1970's (Kaiser, 2003). Since then, QSAR has been used in regulatory assessment of chemicals. The New Chemicals Policy of The European Commission, Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation, also strongly recommends the use of QSARs. QSAR/QSPR technique is mainly employed in the fields of prediction of physico-chemical properties, environmental fate, ecotoxicity, and other activities related to human health (Roy and Ghosh, 2007; Porcelli et al., 2008). The toxicity of the 24 phenol derivatives to *Rana japonica* was modeled by Wang et al. (2000). They reported that separating toxicants into subsets according to their mechanism of toxic action and deriving separate QSARs for each mechanism of toxic action increases the success of QSAR. Tao et al. (2002) developed a multivariate regression model to predict organic chemical toxicities to *Daphnia magna*. They also tested the accuracy

of their model based on coefficient of determination for the regression and associated residual values. Zvinavashe et al. (2009) proposed a QSAR model that can be used to predict fish toxicity with *Daphnia magna* toxicity values. Walker (2003) summarized the applications of QSARs by the US Government organizations. Saçan et al. (2005) developed models to predict water solubility, log K_{ow}, and Henry's Law constant of PCDD/PCDF and phthalate esters. They concluded that the *CRI* which reflects branching and global molecular properties of a chemical was an important descriptor in these predictions.

QSAR can be used for analyzing the toxicities of organophosphorous and carbamate compounds, which are widely used as insecticides. Bermudez-Saldana and Cronin (2006) modeled the toxicity of these pesticides to the rainbow trout. They stated that linear models which have datasets basing on the same mechanism of action (MOA) are more successful than those have heterogeneous datasets in terms of producing models for predictive purposes. Zvinavashe et al. (2009) developed linear models of the toxicity of organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. They used quantum chemical molecular descriptors namely log K_{ow}, E_{LUMO}, and E_{HOMO}. They also proposed a relationship between the toxicities of *Daphnia magna* and fish, and found a high correlation between them. A QSAR study of pesticide toxicity of rainbow trout was made by Mazzatorta et al. (2005). They developed linear and nonlinear models with chemical descriptors using the OpenMolGRID system. They also compared linear and nonlinear approaches for ecotoxicological QSARs. Ultimately, they found the Genetic Algorithm (GA)/CPNN combination is the most suitable one among the methods tested.

Historically, the first studies on QSAR models included homolog series of chemicals. Eventually, diverse datasets has been started to use in models to cover as much as possible chemicals. Of the methods used in QSAR/QSTR studies, MLR relates the dependent variable, y (toxicity), to a number of independent (descriptors) variables, $x_{i,}$ by using linear equations. Principal component analysis (PCA) is based on linear combinations of the variables and used for dimension reduction. PCA is able to arrange a good training set. However, the model has a poor ability to predict the test set. Additionally, it is difficult to discriminate outliers while applying PCA. Partial least squares (PLS) aims to explain the variance in the descriptors and

engages in obtaining a good correlation between activity and the descriptors. The primary advantage of PLS is that it is very useful when collinearity exists in the independent variables. Likewise PCA, there is outlier identification problem. An artificial neural network (ANN) is a mathematical model that "learns" from data just like learning of human brain via neurons. In every epoch of training, network adjusts the weights of each neuron. *k*-nearest Neighbor (*k*NN) is a clustering algorithm that assesses to which class an object belongs to. The class of the object is determined according to the class of the nearest neighbors of the object, typically by weighing based on the distances between the object and its neighbors. GA, inspired from the theory of evolution, attempts to solve problems or develop control strategies. It differs from most other artificial intelligence techniques due to its ability to develop many solutions in parallel. Each solution is regarded as an individual of the population, and its fitness to the environment is evaluated by a domain-specific function. GA improves the overall fitness of the population by applying natural selection of the individuals based on fitness, crossbreeding by recombination of parts of existing good solutions, and mutation (OECD, 2007).

Toxicities of organic chemicals to a specific aquatic biota can be obtained by using QSTRs. Nonlinear models such as counter propagation artificial neural networks (CPNN) are usually more powerful than linear ones, but are often considered "black boxes" because they do not formulize the relationship between variables and response in clear numbers or coefficients. Mazzatorta et al. (2005) stated that linear methods are not capable of solving complex problems such as toxicities of a diverse database. Cronin et al. (2004) chose a nonlinear method, *k*NN, for modeling of organic chemicals toxicity to *Chlorella vulgaris*. They concluded that method selection in QSAR is task dependent and more complicated methods should be preferred only in clear need.

1.1. Purpose of the Study

Most of the studies stated above Wang et al. (2000), Tao et al. (2002), Saçan et al. (2005), Bermudez-Saldana and Cronin (2006), Zvinavashe et al. (2009) are based on linear statistical methods such as multiple linear regression. Although they are useful models, they suffer from the limitation that in some cases the relationship between a molecular descriptor

and toxicity may be intrinsically nonlinear. In such cases, the use of linear statistical may not result in best models. Additionally, some models (Roy and Gosh, 2007) have high number of descriptors which are not preferable in QSAR studies. The first condition of model validity deals with the ratio of the number of chemicals over the number of selected descriptors which is called as Topliss ratio. The recommended Topliss ratio should have a value of at least 5 (OECD, 2007). Therefore, in this study, considering this ratio we carried out two comparative studies of multiple regression vis-à-vis neural net methods in predicting the toxicity of two different data sets to two aquatic organisms.

To look for more useful QSTR models in ecotoxicology will contribute to this field. The main aims of the present study are as follows: 1) to elaborate QSTR models for the prediction of toxicity values of two diverse sets of chemicals to *Chlorella vulgaris* and *Oncorhynchus mykiss* (Rainbow trout) by using a large number of theoretical descriptors generated from Spartan 06, Dragon 5.4, Codessa 2.2 software, and the *CRI*, 2) to compare the performance of linear (MLR) and nonlinear (CPNN) models, 3) to evaluate the validity of the developed models according to the OECD principles and 4) to compare the predictive ability of the proposed models with literature QSTR models for the same toxicity data.

2. THEORETICAL BACKROUND

2.1. Use of Pesticides

The need for food is dramatically increasing owing to the world population is increasing rapidly. On the other hand, the arable land area is fixed. Moreover, erosion and construction of new buildings cause a decrease in total arable land area. But the more crucial point is that a number of crops are depleted by pests. In order to increase food production to feed six billion people, protecting the crops from pests is inevitable. The leading solution for fighting against pests is plant protection products, namely, pesticides.

Pesticide chemicals have served humans for over 3000 years to control pests such as insects, weeds, and fungi (Crosby, 1998). Each year, 5 billion tons of pesticides are applied worldwide. Major proportion of this usage is for agricultural activities, and others are used in heavy metal industry (Wright and Welbourn, 2002). When we compare pesticide usage of Turkey with those of developed countries, it is observed that relatively small dosages are applied in Turkey. For instance; pesticide application per hectare is 0.63 kg in Turkey while it reaches 3.5 kg in the USA, 17.5 kg in Holland, and 4.4 kg in France (Dag et al., 2000).

2.2. Types of Pesticides

There are multiple ways of classifying pesticides. They can be categorized in three broad groups with their respective percentages according to the type of pest they control: herbicides (70%), insecticides (20%), and fungicides (10%) (Wright and Welbourn, 2002). Pesticides can also be classified as synthetic pesticides and biological pesticides (biopesticides).

Additionally, pesticides can be classified according to the pests they control as follows:

- Algicides or algaecides for the control of algae
- Avicides for the control of birds
- Bactericides for the control of bacteria
- Fungicides for the control of fungi and oomycetes
- Herbicides (e.g. glyphosate) for the control of weeds
- Insecticides (e.g. organochlorines, organophosphates, carbamates, and pyrethroids) for the control of insects - These can be ovicides (substances that kill eggs), larvicides (substances that kill larvae) or adulticides (substances that kill adults)
- Miticides or acaricides for the control of mites
- Molluscicides for the control of slugs and snails
- Nematicides for the control of nematodes
- Rodenticides for the control of rodents
- Virucides for the control of viruses (e.g. H5N1)

Chemically-related pesticides can be grouped as follows:

- Organophosphate pesticides
- Carbamate pesticides
- Organochlorine insecticides
- Pyrethroid pesticides

(EPA, 2010; Miller, 2004).

2.2.1. Organophosphates

Organophosphates constitute one of the main classes of insecticides acting on the acetylcholinesterase enzyme (AChE). Starting at the mid 1940s, the first developed organophosphate pesticide was tetraethylpyrophosphate (TEPP). Eventually parathion came into use. Parathion is a phosphorothionate ester which is metabolically converted to paraoxon. Paraoxon attaches to the active site of the AChE, which makes it extremely toxic to non-target organisms. Their toxicity is augmented by their ability to be absorbed through the skin. It is, therefore, important to develop organophosphorous compounds as pesticides that can be

metabolized by mammals to less toxic forms. Malathion and dichlorvos are the best known examples (Wright and Welbourn, 2002).

The general chemical structure of organophosphorous pesticides is shown in Figure 2.1. They are classified into two main groups, organophosphates (P=O) and organothiophosphates (P=S) depending on whether oxygen or sulphur forms a double bond with the central phosphorous atom (Zvinavashe et al., 2009).

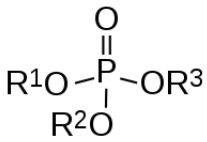


Figure 2.1. General chemical structure of organophosphorous pesticides. The atom that forms a double bond to P is either O (organophosphate; P=O) or S (organothiophosphate; P=S)

The AChE inhibition of some organophosphorous pesticides can take place through metabolic activation. For example, the P=S bond of the organothiophosphate pesticides is converted to oxon form (P=O bond) because oxygen is more electronegative than sulphur, thus the P=O bond is more polarized than the P=S bond. After then, the pesticide becomes being able to inhibit AChE. Organothiophosphates are manufactured more than organophosphates as they are considered to be safer and more selective due to the biotransformation step to the organothiophosphate that is necessary to exhibit full toxicity (Bermudez-Saldana and Cronin, 2006; Zvinavashe et al., 2009).

2.2.2. Carbamates

The carbamate pesticides are esters of carbamic acid. Similar to organophosphate pesticides, they act as AChE inhibitors and are toxic to non-target organisms. They also affect the immune system. Unlike organophosphates, carbamate pesticides have lower dermal toxicity (Wright and Welbourn, 2002). Most carbamates are narrow spectrum insecticides,

unlike organochlorines and organophosphates, which are toxic to only a few types of insects. General chemical structure of carbamates is given in Figure 2.2.

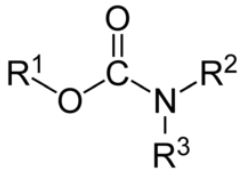
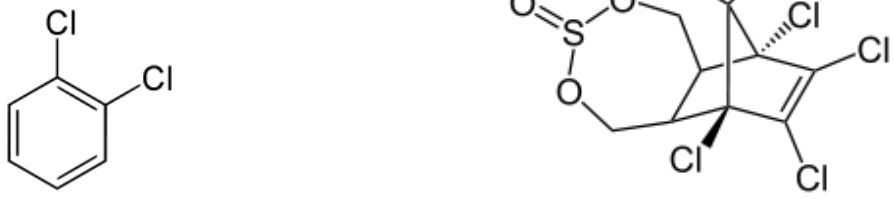


Figure 2.2. Chemical structure of carbamates

2.2.3. Organochlorines

Organochlorines are organic compounds containing at least one covalently bonded chlorine atom. Organochlorine insecticides are stable solids of limited vapor pressure, very low water solubility, and high lipophilicity. Some of them are highly persistent in their original form or as stable metabolites. The most known organochlorine pesticide is the insecticide DDT (dichloro diphenyl trichloroethane). Launched in the 1940's, DDT was widely used in agriculture around the world for many years. DDT was banned in many countries in the 1970s because of their unacceptably slow degradation and subsequent bioaccumulation linking DDT with damage to wildlife. Since then, agricultural uses of DDT have been outlawed worldwide (Wright and Welbourn, 2002; Walker et. al, 2006). The chemical structure of organochlorine pesticides 1,2-dichlorobenzene and endosulfan are shown in Figure 2.3 (a) and (b), respectively.

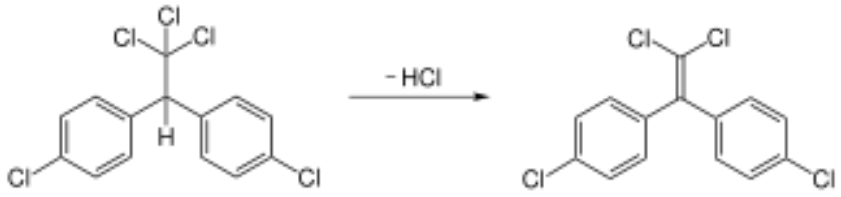


(a) (b)

Figure 2.3. Chemical structure of (a) 1, 2-dichlorobenzene and (b) endosulfan

DDT and other chlorinated hydrocarbons are very resistant to metabolic breakdown. However, in animals and humans, DDT is degraded to DDE (dichlorodiphenyl dichloroethylene) or DDD (dichlorodiphenyl dichloroethane or rhothane). A limited conversion of DDT to DDE occurs in human subjects. The conversion is catalyzed by DDT dehydrogenase, and the resultant DDE is a stable metabolite (Yu, 2001). Formation of DDD and DDE from DDT are shown in Figure 2.4. (a) and (b), respectively. DDD and DDE are in our rainbow trout data set.

(a) Reductive dechlorination of DDT to form DDD



(b) Degradation of DDT to form DDE by an elimination of HCl

Figure 2.4. Metabolites of DDT (a) DDD and (b) DDE

2.3. Concerns on Pesticide Use: Distribution and Environmental Fate

Pesticides are supposed to have following properties. An ideal pesticide kills only the target organism, has no health effects on non-target organisms, is broken down into harmless chemicals in a fairly short time, is unlikely to develop genetic resistance in target organisms, and is economical to use comparing to having no action. However, there is no such pesticide (Corrigan et al., 1997). Beginning their first uses in the history, chemical industry has been

working on new chemicals which are effective on target organisms, but less noxious to environment.

Beside their effective use in agriculture against pests, pesticides also have their negative aspects. Essential effects are listed as their persistence in the biosphere and their chronic toxicity to non-target species such as birds, fish, even humans (Wright and Welbourn, 2002).

Once a pesticide is used, it inevitably disperses in all phases via leaching, runoff, volatilization, and precipitation. Besides staying as it is, a pesticide may be transformed to another substance. The fate of pesticides comprises hydrolysis, aqueous photolysis, photodegradation on soil, volatility, sorption, and bioaccumulation in living organisms. There are numerous studies in the literature about the distribution and environmental fate of pesticides (Goldsborough and Crumpton, 1998; Centofanti et al. 2008; Acero 2008; Atasoy et al. 2009; Ozcan and Aydin, 2009). The fate of pesticides in the environment is summarized in Figure 2.5.

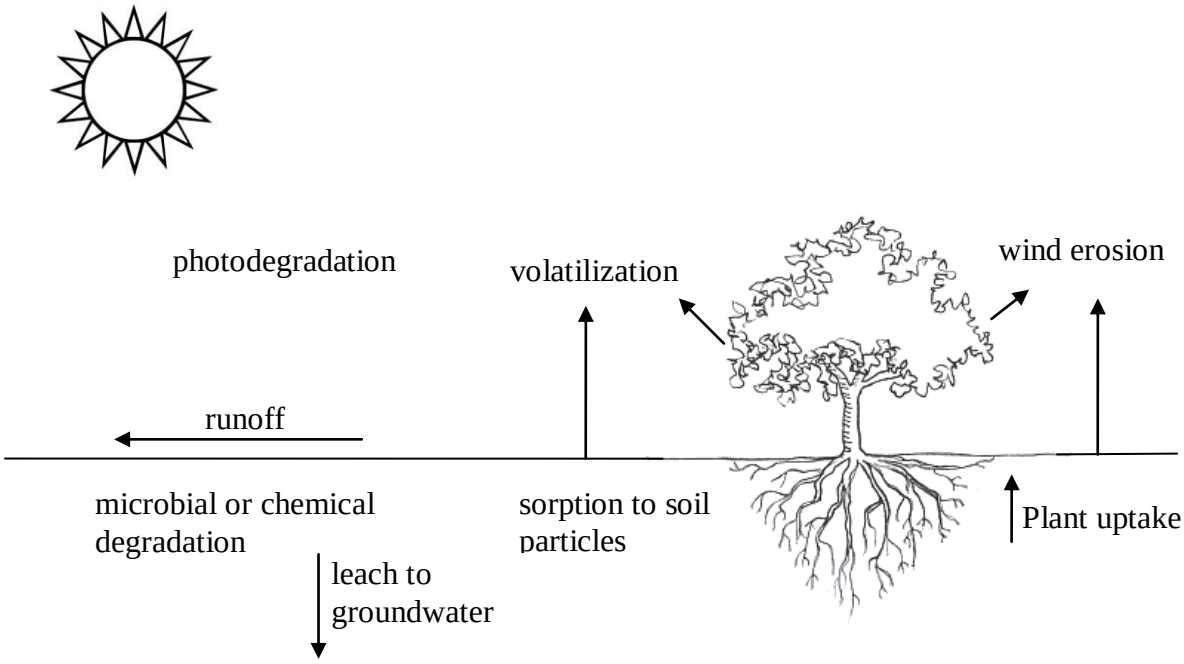


Figure 2.5. The fate of pesticides in the environment

In order to interpret the existence of the pesticide concentrations and their effects to the environment, research has focused on their measurements, estimations, and compilation of the data. In a study reported by Vijver et al. (2008), many pesticide measurements from surface waters of the Netherlands were compiled. A freely available tool, Dutch pesticides atlas, is also provided to promote the use of monitoring data in the process of risk mapping. The atlas can be used to present all measurements of pesticides in surface water on the level of individual active ingredients in a spatial framework. Predictions can also be made where a pesticide might be exceeding environmental standards. Laabs et al. (2002) studied on monitoring 29 pesticides and 3 metabolites in surface water, sediment, and rainwater during the main application season of northeastern Pantanal basin in Brasil. They provided information on pesticide distribution and dynamics in the area.

Even though the bans have taken place and usage has been restricted, DDT and their metabolites have still been used and there are numerous studies about their existence and effects around the world. For instance, Sibali et. al. (2009) studied the determination of DDT and metabolites in surface water and sediment of Jukskei River in South Africa. They suggested that there is a recent contamination of the river by DDT metabolites. Dörr and Liebezeit (2009) analyzed bivalve mollusks of Wadden Sea for organochlorine compounds. They found these contaminants in all samples and evidenced that there is an increase in levels of contaminants compared to last 6 years. DDT is used for indoor residual spraying in South Africa to control malaria. Findings of Barnhoorn et al. (2009) show that water and food may be major routes of human exposure to DDT and metabolites, thus expose people to adverse health effects. Measurements of organochlorine pesticides concentrations including DDTs in sediments of Haihe River of China between 2006 and 2008 were made by Chi (2009). The study showed that there is a significant decrease in levels of hexachlorocyclohexanes (HCHs) and DDTs comparing to the levels in 2003. The author states that this may be due to the decrease in production and usage of these chemicals and biodegradation of HCHs and DDTs in sediment.

Since the most affected people by pesticides are farmers, research has been done to determine concentrations in human body fluids. Exposure of farmers' children to pesticides

has been measured as an EPA project (Lu et al., 2006). The study demonstrated that saliva can be used to assess exposures to diazinon in pesticide applicators and their children.

Short-term pollution events via runoff are typical of streams in agricultural areas. Berenzen et al. (2005) developed a model that can predict runoff-related pesticide concentrations in many streams on a landscape level. They validated their model by predicting the pesticide load of 18 small lowland streams and compare against measured concentrations obtained by runoff-triggered sampling. The authors suggested that the presented model is suitable for use in routine exposure assessment of pesticides on a landscape level.

As research being done and the public awareness increases, bans and restrictions on pesticides are on the agenda. At the beginning of 2008, cosmetic pesticides usage in lawns, gardens, school yards and parks has been banned in Canada. The ban came into effect in order to protect families, especially children from toxic chemicals (OME, 2009). Formerly safe-called pesticide endosulfan has been banned by the Australian Pesticides and Veterinary Medicines Authority. This decision follows a recent assessment of new information by the Department of Sustainability, Environment, Water, Population and Communities (DSEWPC) that the prolonged use of endosulfan is likely to lead to adverse environmental effects via spray drift and run-off (APVMA, 2010).

2.4. QSTR Studies

Preliminary studies of QSAR/QSTR go back to first half of 19th century. Blake noted that “salts of isomorphous bases have a similar action” in 1841. This finding appears to be the first example of an attempt to relate the activity of some compound to a physical property. At the International Congress in 1860, Canizzaro showed that Avogadro’s law of combining volumes could be used to derive the correct empirical formula for simple compounds. Horsford reported in 1851 that the taste of some compounds could be related to their composition. Pelikan in 1854 observed that toxic effects depended on composition. Borodin in 1858 expressed that toxicological property of compounds and their chemical compositions are

closely interrelated. Additionally, Borodin stated that similar substances or substances taking part in similar chemical reactions exert similar actions on the organism. Between 1858 and 1870, Kekule, Couper and Crum-Brown developed the concept of molecular structure. Crum-Brown and Fraser proposed the existence of a mathematical relationship between structure and bioactivity in 1868 (Charton, 2008).

Quantitative-structure toxicity relationship (QSTR) studies are expected to reduce the cost and the number of organisms used for toxicity testing and to fill the existing data gaps within the REACH regulatory framework in the EU. Many QSTR studies in ecotoxicology are carried out with different types of descriptors using statistical methods like regression analysis (Cronin et al., 2004, Saçan et al., 2007), PLS (Roy and Gosh, 2006 and 2007) and ANN (Roy and Roy, 2009) for diverse set of chemicals. Pavan et al. (2006) used multivariate linear regression and GA for modeling the toxicity of heterogenous chemicals to fathead minnow. Kahn et al. (2007) compared the best multilinear regression approach and the heuristic back-propagation neural networks (BPNN) for modeling the toxicity of chemicals to the organism *Tetrahymena pyriformis*. Papa et al. (2005) modeled toxicity of organic chemicals to fathead minnow with Dragon descriptors. Saçan et al. (2007), modeled the toxicity of aromatic compounds to the algae *Scenedesmus obliquus* using the quantum-chemical descriptor (the energy of the lowest unoccupied molecular orbital (E_{LUMO})) calculated by Spartan 06, together with the *CRI*.

In developing QSAR/QSTR models, the approach begins with the compilation of available endpoint data sets for a variety of chemicals. If endpoint data are available for a sufficient number of chemicals, the data set is often divided into a training set used in the model development, and a test set containing chemicals not used in the derivation of the model but used to evaluate the model. The method used for splitting the data set should be clear in proposed model. Methods available include those based on similarity analysis, for example, D-optimal distance (Hasegawaa and Funatsub, 1998), Kohonen map or self-organizing map (SOM) (Vracko et al., 2006), the *k*-means cluster analysis (Caballero and Fernandez, 2006), sphere-exclusion algorithms (Golbraikh et al., 2003) or random selection through activity sampling (Gramatica, 2007). The selection of variables in the model, referred to as

predictors/descriptors, can be performed by one of the techniques such as Heuristic Method (HM), GA, PCA or Factor Analysis (FA).

2.5. Literature Studies on *Chlorella vulgaris* and *Oncorhynchus mykiss*

Adverse effects of chemicals on aquatic organisms, especially algae, are of special concern. For the assessment of the environmental impact of toxicants, fresh water algae *Chlorella vulgaris* is particularly important because they are habitants of freshwaters that are threatened by a variety of pollutants, their experiments are economical, and green algae respond to chemicals very rapidly. Cronin et al. (2004) developed QSTR models to predict the 15-min toxicity values of a diverse set of chemicals to *Chlorella vulgaris* using both MLR and *k*NN with three descriptors, namely, hydrophobicity, log K_{ow}, electrophilicity expressed by E_{LUMO}, and first order Δ valence connectivity index ($\Delta^1\chi^v$). In another study, Roy and Gosh (2007) used four different statistical techniques in modeling the same algal toxicity data used by Cronin et al. (2004). They highlighted the importance of extended topochemical atom (ETA) and non-ETA descriptors in their QSTR models.

For an aquatic risk assessment, the toxicity of pesticides to non-target organisms is assessed to evaluate the toxic potential of the compounds in an aquatic environment. The environmental hazard on vertebrates in aquatic systems is evaluated by performing acute and chronic fish experiments. The most widely performed test is the acute fish toxicity test (Knauer et al., 2007). The rainbow trout is a preferred species to meet this requirement since it is sensitive, and considered as a representative cold water fish species by regulatory bodies (Mazzatorta et al., 2005). Rainbow trout is also the most frequently used species among the studied fish in ECOTOX database of EPA (Hrovat et al., 2009). Hrovat et al. (2009) also stated that life stage differences and test conditions such as temperature, pH, and water hardness can have a clear influence on the LC_{50} test results. Capkin et al. (2006) have studied toxicity of endosulfan to *Oncorhynchus mykiss.* They also examined the effects of fish size temperature, alkalinity, and hardness to the toxicity tests. They ultimately found that all the items except hardness have an effect on rainbow trout toxicity tests.

2.6. Artificial Neural Networks

Artificial Neural Network (ANN) analysis, which is an accepted nonlinear technique in QSAR studies, was adopted to investigate nonlinear patterns in the data. ANN is an information processing paradigm that resembles biological nervous systems, such as the brain. The functioning of a neural network also resembles human brain in that it learns by examples. The whole network is an interconnected group of artificial neurons that uses a mathematical model. Each neuron receives an input signal which has the total information from other neurons, processes it locally through an activation function and produces a transformed output signal to other neurons or external outputs as represented in Figure 2.6 (Zhang et al., 1998).

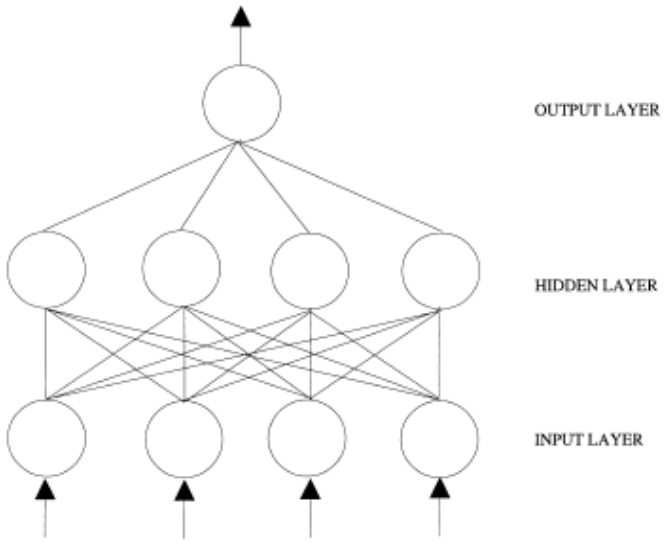


Figure 2.6. A typical feed-forward neural network

ANNs typically start out with randomized weights for all their neurons. This means that they do not "know" anything and must be trained to solve the particular problem for which they are used. The training process is usually as follows: first, examples of the training set are entered into the input nodes. The activation values of the input nodes are weighted and accumulated at each node in the first hidden layer. The total is then transformed by an activation function into the node's activation value. It in turn becomes an input into the nodes in the next layer, until eventually the output activation values are found. The training

algorithm is used to find the weights that minimize some overall error measure such as the sum of squared errors (SSE) or mean squared errors (MSE). Hence the network training is actually an unconstrained nonlinear minimization problem (Zhang et al., 1998).

The appropriate training and test algorithms were chosen in the course of the model development process, as the structure of the models (which is difficult to predict beforehand) is the most important factor behind the decisions regarding algorithm selection.

The CPNN models generally have two layers, the input (Kohonen) layer and the output layer. CPNNs are built up from two layers of neurons arranged in 2D rectangular matrices. The Kohonen layer receives the input variables. Afterwards, it converts 3D input into 2D map such that similar compounds (having similar descriptors) are located in the same neuron. The output layer, which has the same topological arrangement of neurons as the input layer, receives the target (toxicity) values during the learning process. Architecture of a CPNN, together with Kohonen network, is shown in Figure 2.7 (Mazzatorta et al., 2003). Details about Kohonen maps and architecture and learning strategy of CPNN can be found in numerous text books and articles (Devillers, 1996; Zupan and Gasteiger, 1999; Vracko, 2005). We tested different network architectures and different number of learning steps (epochs), roughly more than hundred models in total, to obtain each of the models.

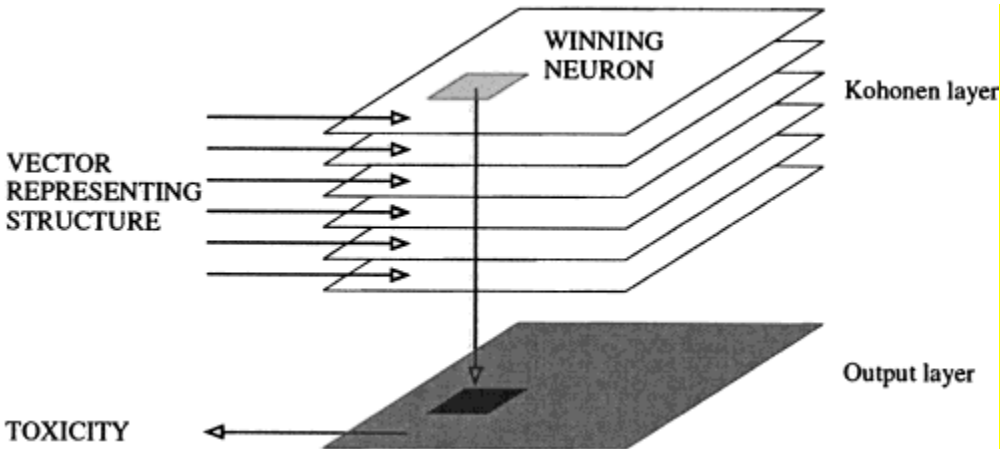


Figure 2.7. Architecture of a typical CPNN

3. MATERIALS AND METHODS

We followed the workflow for the development of linear and nonlinear models for the toxicity of different classes of chemicals including pesticides to fish and algae (Figure 3.1). The details of each step are given in corresponding subtitles.

3.1. Data Sets

A structurally heterogeneous data set on acute toxicity of 91 organics from different chemical classes such as, cresols, monohalogenated and mononitro-substituted benzenes, anilines and phenols, and 10 pesticides to algae *Chlorella vulgaris* were taken from Cronin et al. (2004). Toxicity (*p*T) is defined as the negative logarithm of the concentration which affects 50% of the population ($pT = -\log (EC_{50})(mM)$) in 15 min static monitoring test.

The second data set of 34 chemicals includes organophosphates, carbamates, and organochlorines (Bermudez-Saldana et al., 2005). Authors have compiled 96 h *p*T ($-\log LC_{50}$) experimental values (mM) of *Oncorhynchus mykiss* from ECOTOX database from U.S. EPA.

3.2. Molecular Descriptors

A large number of molecular descriptors were calculated for each of these data sets using three software packages, namely, Dragon 5.4 (Talete, 2006), Spartan 06 (Wavefunction, 2006) and Codessa 2.2 (Semichem, 1996), and the *CRI*. Before the calculation step, the structures of the compounds were sketched using the Spartan 06 software package and geometrically optimized employing the semi-empirical PM3 method. The molecular geometries corresponding to the lowest energy conformer were selected for the calculation of the molecular descriptors. Calculated Spartan descriptors were saved as a text file format for further descriptor selection step using the Codessa 2.2 software package. Molecular structures prepared as MDL mol files were loaded into Codessa software for the calculation of pre-

integrated Codessa descriptors (Katritzky, 1994). MDL mol files were also used for the generation of descriptors employing Dragon 5.4 software package. The total pool of 1356 Dragon descriptors, 125 Codessa descriptors and 5 Spartan descriptors were computed.

The Dragon 5.4 software yields 20 descriptor groups, which cover 1664 descriptors. The descriptor group details and the number of descriptors in each group are provided in Table 3.1.

The Codessa descriptor set includes constitutional, topological, geometrical, and electrostatic descriptors (Katritzky, 1994).

The Spartan descriptors used are dipole moment (μ), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the energy of the highest occupied molecular orbital (E_{HOMO}), the gas phase energy (E), CPK volume (V), and area (A). The rest of the calculated descriptors such as, E_{LUMO}-E_{HOMO} gap, hardness, electronegativity, softness, and electrophilicity (ω) were calculated from the energies obtained from Spartan 06 and using the formula reported for each by LoPachin et al. (2007). The parameters, molecular volume and molecular area belong to the calculated descriptors very frequently used in many QSAR studies (Netzeva et al., 2004; Aptula et al., 2005). Since the use of variables with different scales may weight the variables with larger scale, the ratio of volume to area obtained from CPK model (a molecular model in which atoms are represented by spheres, the radii of which correspond to van der Waals radii) in Spartan 06 was calculated and designated as (V/A), instead of using autoscaling or standardized values of these descriptors. Aqueous phase energy of molecules containing phosphorous could not be calculated due to the limitations of Spartan 06 software. Therefore, only gas phase energies were calculated for all molecules.

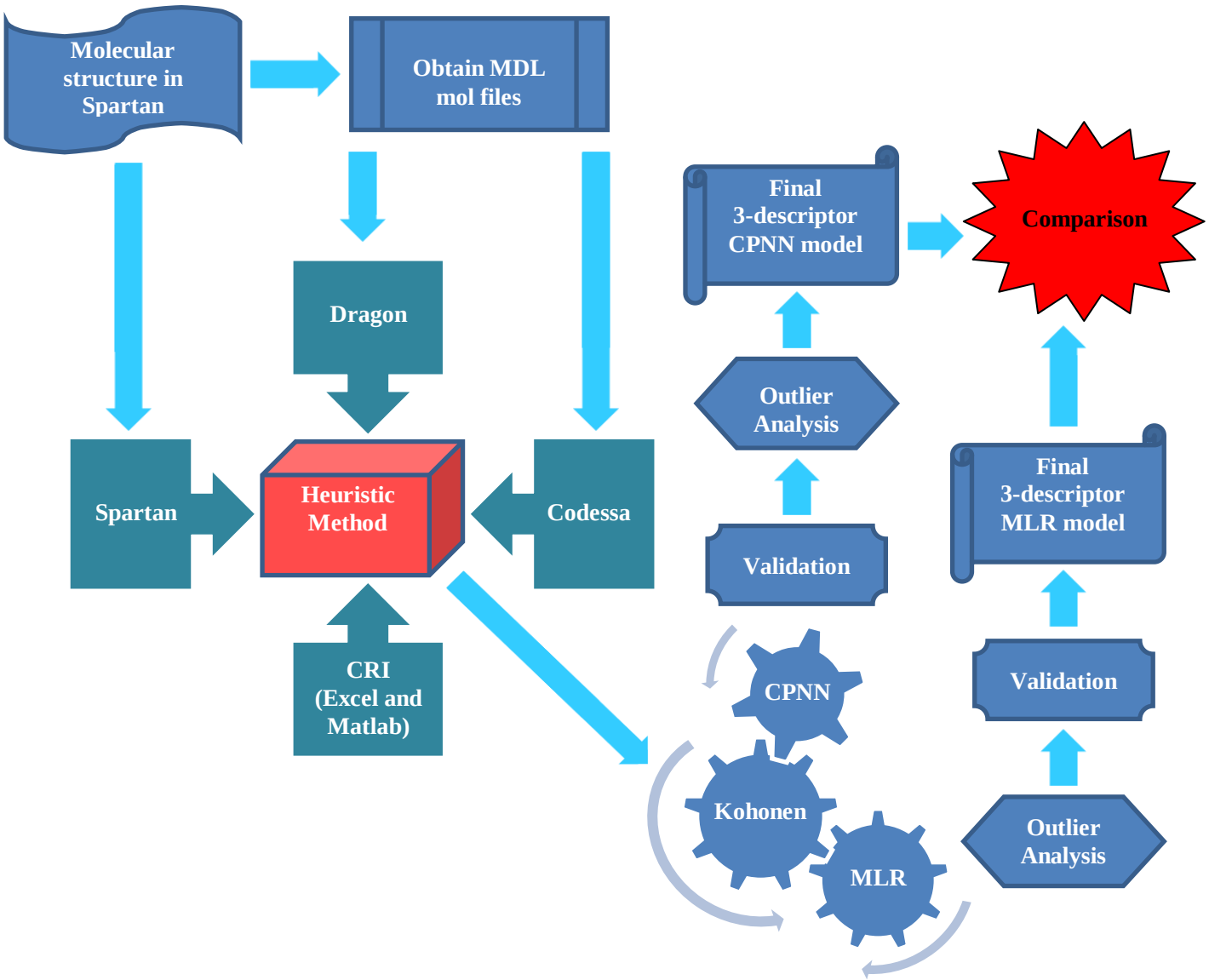


Figure 3.1. Workflow of present study

Table 3.1**.** Descriptor groups in the Dragon 5.4 software

Dimension	Descriptor Blocks	Number of descriptors
0D	Constitutional descriptors	48
1D	Functional group counts	154
1D	Atom-centered fragments	120
2D	Topological descriptors	119
2D	Connectivity indices	33
2D	2D autocorrelations	96
2D	Burden eigenvalues	64
2D	Eigenvalue-based indices	44
2D	Walk and path counts	47
2D	Information indices	47
2D	Edge adjacency indices	107
2D	Topological charge indices	21
3D	Randic Molecular Profiles	41
3D	RDF descriptors	150
3D	WHIM descriptors	99
3D	Geometrical descriptors	74
3D	3-D MoRSE descriptors	160
3D	GETAWAY descriptors	197
others	Charge descriptors	14
others	Molecular Properties	29

The *CRI* parameter which is an eigenvalue-based descriptor and has been shown to be effective in modeling various properties of chemicals including the toxicity (Saçan et al., 2007) was also included in the descriptor pool. The procedure to calculate the *CRI* and relevant references were reported previously (Saçan and Inel, 1993; Saçan and Inel, 1995; Saçan and Balcioglu, 1996; Saçan et al., 2004; Saçan et al., 2007). The only difference

between the new *CRI* and the previous one is the eigenvalues of the characteristic matrix stored in Microsoft Excel were calculated with Matlab 6 (Mathworks, 2000), instead of Scientific Workplace.

In order to reduce the number of descriptors, HM algorithm running in Codessa 2.2 was performed. The HM selects the descriptors according to the following criteria. The program calculates all correlations between individual descriptors and property (toxicity) and eliminates descriptors with *F*-test's value is less than 1, correlation coefficient is less than the set value (0.1) and the *t*-value is less than the set value (0.1). It selects only one of the highly intercorrelated descriptors. Additionally, descriptors with low variance (<0.1) and with variance inflation factor (VIF) values close to five after heuristic analysis were eliminated from the descriptor list.

A significant drawback of obtaining chance correlation in MLR models was eliminated by using the HM. This phenomenon occurs especially when the ratio of the number of molecules to the number of descriptors is very low. Thus, there is a very low probability for obtaining chance models. Nevertheless, the developed models were evaluated for over-fitting, generalization and predictivity by cross-validation as well as using an external test set.

We also forced to use the *CRI* parameter in the derived correlations because it was proved to be a prevailing descriptor in our previous studies involving algal toxicity (Saçan et al., 2007) and several physicochemical/biological properties (Saçan and Inel, 1993; Saçan and Inel, 1995; Saçan and Balcioglu, 1996; Saçan and Balcioglu, 1998; Saçan et al., 2004; Saçan et al., 2005; Saçan et al., 2007). Therefore, the potential of the *CRI* in QSTR modeling can be further verified for different organisms.

3.3. Model Development and Validation

QSTR models for each organism were developed using both MLR and CPNN considering the OECD principles.

The OECD principles of QSAR validation give five basic elements for a reliable model.

1. a defined endpoint;
2. an unambiguous algorithm;
3. a defined domain of applicability;
4. appropriate measures of goodness-of-fit, robustness, and predictivity;
5. a mechanistic interpretation, if possible.

According to Principle 4, a QSAR model should have appropriate measures of goodness-of-fit, robustness, and predictivity. While the internal performance of a model determined by using a training set, the predictivity is determined by using an appropriate test set (OECD, 2007). Therefore, before performing the modeling procedure, the data set was divided into training and test set using Kohonen Neural Network alias Self-Organizing Maps (SOM). SOM are able to select a meaningful training set and a representative test set. Kohonen networks have been adequately explained by Devillers (1996), and Zupan and Gasteiger (1999). Kohonen networks project multi-dimensional space onto 2D array of neurons. The projection, which is called learning of network, runs in two steps. In the first step, an object (represented by a vector) is presented to all neurons and the algorithm selects the neuron that is most similar to it. The selected neuron is called “winning neuron”. In the second step, the weights of the winning neuron are modified to the vector values and in the same time the neighboring neurons are modified to become similar to it (Vracko, 2005). We used different networks for each developed model and approximately 70-75% of the data set was allocated for training set. In order to compare the performance of the linear with that of nonlinear models, the same training and test set of compounds were used.

3.3.1. Multiple Linear Regression

MLR models were obtained using the Statistical Package for Social Scientists (SPSS® 17.0) for Windows (SPSS Inc., 2008). In MLR, as a rule of thumb, the number of independent variables (descriptors) should not be higher than the number of observations (chemicals) otherwise the results would be biased, and there is a chance correlation.

For robustness of MLR models, number of compounds (*n*), squared correlation coefficient (R^2), adjusted (for degrees of freedom) squared correlation coefficient (R^2_{adj}), Fischer statistics (*F*) and standard error of the model (*SE*) are calculated. For training and test sets of all models root mean square error (*RMSE*) and average absolute error (*AAE*) are calculated. Formula for *RMSE* is given in Equation 3.1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{3.1}$$

where, *n* is the number of compounds, $\hat{y}_i$ is predicted and y_i is observed toxicity value

Internal validation of all models was tested with the leave-one-out (LOO) procedure and cross-validation correlation coefficient (R^2_{cv}) was calculated. However, leave-many-out (LMO) procedure was performed only for MLR models using Weka 3.6.1 (2009) software.

Correlations between the independent variables in each of the models were examined. Collinearity was tested with respect to the VIF value for each of the independent variables, defined as $1/ (1-R^2)$, where *R* is the correlation coefficient for one independent variable against others. Large VIF values (over five) imply that there is multicolinearity.

The reliability of MLR models was also tested by response randomization (Y-scrambling) procedure. For model randomization, the dependent variables of the training set are shuffled and new correlation coefficients are calculated. The process is repeated several times. Kiralj and Ferreira (2009) demonstrated that it is sufficient to perform 10-25 y-randomization runs for a model validation. The significantly low correlation coefficients of the new models indicate that the originally proposed model was not obtained by chance correlation. Y-scrambling procedure was run in MDM 2010.2 (Molegro Data Modeler, 2007-2010).

3.3.2. Counter Propagation Neural Network

Prior to CPNN modeling, *p*T and descriptor values were normalized in the range from 0 to 1 as stated by Vracko et al. (2006) using the PREDATA program developed by National Institute of Chemistry Ljubljana. The performance of the network was checked by various network size epochs. Correlation coefficient (R^2), cross-validation coefficient (R^2_{cv}), and RMSE of the model played an important role in selecting the best network-epoch combination. Over-training was abstained by selecting different training/test set divisions, networks, and epochs. After selection of the best models and monitoring the outliers, the normalized values were converted into re-normalized values to obtain comparable predicted toxicity, residual, *RMSE* and *AAE* values to those of MLR models.

3.3.3. Validation

The statistical quality of the MLR models was judged by the parameters like the square of correlation coefficient (R^2), standard error of the estimate (*SE*) and variance ratio (*F*) at specified degrees of freedom. Leave-one-out (LOO) cross-validation statistics (R^2_{cv}) and leave-many-out cross-validation was applied on the final model.

In order to obtain compounds for external validation, the available set of chemicals was divided into a training set and a test set. Dividing was carried out by self-organized maps explained above.

With the best developed equations, toxicities of chemicals outside the sample set were predicted and compared with the reported literature data and models. Furthermore, to compare the predictive performance of the model developed in this study with those of the literature models, both *AAE* of and *RMSE* values were also reported for model comparison.

Consonni et al. (2009) formulated a novel external correlation coefficient (Q^2_{F3}) for the test set based on sum of squares (SS) referring to mean deviations of observed values from the

training set mean over the training set instead of the external evaluation set. They concluded that correlation coefficients using either training set activity mean or test set activity mean have drawbacks. Therefore, the external predictive ability of the models should have information about the whole data set. They proposed Equation 3.2 to test the external predictive ability of the models,

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{test}} (\hat{y}_i - y_i)^2\right] / n_{test}}{\left[\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2\right] / n_{tr}} \qquad (3.2)$$

where, $\hat{y}_i$ is the predicted test set compound, y_i is the observed value, $\bar{y}_{tr}$ is the mean of training set compounds, n_{test} is the number of compounds in test set and n_{tr} is the number of compounds in training set.

We adopted the criteria of Golbraikh et al. (2003) which correspond to OECD principle no 4 (OECD, 2007). Models were considered acceptable, if they satisfied all of the following conditions:

I. $R^2_{cv} > 0.5$

II. $R^2 > 0.6$,

III. R^2_0 or R'^2_0 close to R^2.

i.e.: **(a)** $(R^2 - R^2_0)/R^2 < 0.1$ and $0.85 \le k \le 1.15$ or

(b) $R^2 - R'^2_0)/R^2 < 0.1$ and $0.85 \le k' \le 1.15$

IV. $\left| R^2_0 - R'^2_0 \right| < 0.3$,

where R^2 is predicted vs. observed, R'^2 is observed vs. predicted, k and k' are slopes, R^2_0 and R'^2_0 are squared correlation coefficients (without intercept).

3.3.4. Outliers and Applicability Domain

The presence of outliers in any model can significantly change its predictive power. There are a variety of methods to highlight outliers including identification of those compounds with significantly high standard residuals from regression-based techniques. We identify a compound as a response outlier in MLR models if its predicted value is higher than 2.5 standard residuals. Standardized residuals are calculated as given in Equation 3.3.

$$SR_i = \frac{\hat{y}_i - y_i}{sd} \tag{3.3}$$

where SR_i is the standardized residual of the i^{th} compound, $\hat{y}_i$ is the predicted and y_i is the observed toxicity value, and *sd* is the standard deviation of the errors.

Additionally, chemicals structurally very influential in determining model parameters, i.e., creating leverage effect, were demonstrated in the Williams plot. The leverage of a chemical provides a measure of the distance of the chemical from the centroid of its training set. This graph is obtained by plotting hat values (*h*) versus standardized residuals to verify prediction reliability. In this approach, if the hat value of a test set compound is greater than the critical hat value (*h**), then the compound is identified as structural outlier. Critical hat value is set at $3p/n$, where *p* is the number of descriptors plus one and *n* is the number of compounds in the model (Papa, et al., 2007). If the vector of observed values is denoted by *y* and the vector of estimated values by $\hat{y}$, then $\hat{y} = H\,y$, where *H* is the hat matrix. Diagonal elements of the hat matrix are the leverage values. The hat matrix is given as $H = X\,(X^T X)^{-1} X^T$ where *X* is the design matrix consist of descriptors (Egan and Morgan, 1998).

Compounds with normalized absolute error greater than 0.1 for training set and greater than 0.2 for the test set are considered as possible outliers in the CPNN models. These compounds with high absolute error require further processing as described by Mazzatorta et al. (2003).

Not a single model is expected to predict toxicity/property/activity of all the chemicals. In fact, predictions for chemicals of a well defined domain can be reliable. Therefore, defining an Applicability Domain (AD) of a model is necessary. The AD of MLR models was verified by using the ranges of descriptors and toxicity values, and the leverage approach. First, the range of the variables is defined, and then the leverage values were calculated for tuning of AD of linear models. Outliers having (i) high leverage and low discrepancy do not affect the regression line but tend to increase R^2 and reduce the standard error; (ii) low leverage and high discrepancy tend to influence the intercept but not the slope of the regression or R^2, while usually inflating the standard error; and (iii) both a high leverage and a high discrepancy influence the slope, the intercept, and the R^2 value. Compounds in the external data set that are predicted due to overextrapolation of the model (i.e. fall outside the applicability domain) are detected when their leverage values are far from h^*. Nonlinear model ADs are defined by only descriptor and toxicity ranges.

4. RESULTS AND DISCUSSION

The aim of present study was to compare linear and nonlinear techniques to derive predictive models for a freshwater algae and a fish toxicity of two diverse data sets of chemicals. After preprocessing of descriptors with the HM, a number of models that involve 3-5 descriptors were obtained for these two organisms. For each MLR model a separate nonlinear model was developed. In other words, CPNN models were trained by the descriptors appeared in corresponding linear models. Best 10 models for *Chlorella vulgaris* are given in section 4.1. and one linear and one nonlinear model are given for the *Oncorhynchus mykiss* data set in Section 4.2. The total number of descriptors presented in all resulted models is 15. Table 4.1 presents the abbreviations and the full description of parameters previously selected.

4.1. MLR and CPNN Models for *Chlorella Vulgaris*

Molecular descriptors obtained from the same software modeled with the toxicity separately or in combination with the other software descriptors provide 3-5 descriptor models with comparable statistical quality. Ten linear and nonlinear models were developed for *Chlorella vulgaris*. They were given in Table 4.2 and 4.3, respectively. The predictivity of MLR models was compared to CPNN models in terms of internal and external validation parameters. In general, CPNN models have higher correlation coefficient than MLR models for training sets. However, MLR models have higher cross-validation coefficients. If one considers test set correlation coefficient for predictive ability of the models, obviously MLR models have similar or superior R^2_{test} values than that of CPNN models. On the other hand, CPNN models have lower *AAE*s and *RMSE* for training sets compared to MLR models (Table 4.2.). Furthermore, MLR and CPNN models produced using Spartan 06 descriptors together with the *CRI* had lower correlation coefficients compared to other models. Phenylazophenol appeared as a response outlier in five out of seven MLR models (Model no: 2, 3, 6, 7, and 10). All models were subjected to the test for the criteria of external validation as recommended by Golbraikh et al. (2003) (Table 4.4). All of the models are presented in Table 4.2 and Table 4.3

without any improvement (e.g., elimination of outliers) fulfilled all the criteria given in section3.3.3.It is interesting to note that BLTD48 is a common descriptor in all models when Dragon descriptors are used. Furthermore, electrophilicity (ω) appears as a common descriptor when Spartan descriptors are included in models.

As it was stated in Section 3.2, the *CRI* appeared to be an important descriptor in our previous studies. Therefore, we forced the *CRI* to obtain models together with Spartan descriptors (model no: 3, 4, and 5) and Spartan and Dragon descriptors (model no 8 and 9). For the latter models, we also forced electrophilicity and BLTD48, which are frequently appearing descriptors from Spartan and Dragon software respectively, in model development.

Table 4.1. The abbreviations and full names of descriptors selected by the HM

Descriptor	Meaning of descriptor [a]	Type
DRAGON 5.4 [b]		
BLTD48	Verhaar model of Daphnia base-line toxicity from MLOGP (mM)	Molecular property
X1sol	Solvation connectivity index (χ^{-1})	Connectivity index
HOMT	HOMA total	Geometrical descriptor
piPC07	Molecular multiple path count of order 07	Walk and path count
ATS1m	Broto-Moreau autocorrelation of a topological structure –lag 1 / weighted by atomic masses	2D autocorrelation
EEig05r	Eigenvalue 05 from edge adj. matrix weighted by resonance integrals	Adjacency index
R4e	R autocorrelation of lag 4/ weighted by atomic Sanderson electronegativities	GETAWAY descriptor
RDF080p	Radial Distribution Function – 8.0 /weighted by atomic polarizabilities	RDF descriptor
HATS6m	Leverage-weighted autocorrelation of lag 6/ weighted by atomic masses	GETAWAY descriptor
Mor27u	3D-MoRSE-signal 27/ unweighted	3D-MoRSE descriptor
SPARTAN 06 [c]		
E	Gas-phase energy	Semi-empirical quantum chemical descriptor
E_{HOMO}	Highest occupied molecular orbital energy	Semi-empirical quantum chemical descriptor
Electrophilicity (ω)	Electronegativity^2/(E_{LUMO}-E_{HOMO}) [d]	Semi-empirical quantum chemical descriptor
V/A	CPK Volume/CPK Area	Spatial descriptor
CODESSA 2.2 [e]		
Max Partial Charge for a Carbon Atom	Maximum partial charge for a carbon atom	Electrostatic descriptor
EXCEL 2003 & MATLAB 6 [f]		
CRI	Characteristic root index	Eigenvalue-based descriptor

[a] Todeschini and Consonni (2000)); [b] Talete (2006); [c] Wavefunction (2006); [d] LoPachin (2007); [e] Semichem (1996); [f] Mathworks (2000)

Table 4.2. Statistical summary of MLR models for *Chlorella vulgaris*. The models are numbered as cited in the text.

				Training Set				**Test Set**		
Model No	**Model name**	**Descriptors**	**Training / test set**	R^2	R^2_{cv}	***RMSE***	***AAE***	R^2	***RMSE***	***AAE***
1	D3	ATS1m, BLTD48, piPC07	64/27	.916	.901	.486	.299	.954	.450	.355
2	D4	ATS1m, BLTD48, EEig05r, R4e	63/28	.939	.929	.463	.270	.937	.444	.348
3	SCRI 3	V/A, ω, CRI	71/20	.736	.694	.767	.560	.888	.569	.442
4	SCRI 4	V/A, ω, CRI, E	68/23	.818	.780	.667	.462	.886	.614	.494
5	SCRI 5	V/A, E, CRI, ω, E_{HOMO}	70/21	.842	.801	.612	.436	.926	.476	.383
6	**DS 3**	**BLTD48, X1sol, ω**	**64/27**	**.928**	**.914**	**.474**	**.307**	**.930**	**.427**	**.320**
7	DS 4	BLTD48, X1sol, ω, HOMT	61/30	.926	.906	.450	.301	.958	.357	.293
8	DS-CRI 3	BLTD48, ω, CRI	63/28	.916	.904	.543	.306	.919	.580	.470
9	DS-CRI 4	BLTD48, HOMT, ω, CRI	66/25	.923	.906	.439	.301	.955	.352	.300
10	DCS 4	BLTD48, X1sol, MPC for a C, ω	65/26	.923	.912	.433	.289	.965	.359	.299

Bold model indicates the best 3-descriptor model. The abbreviations used for model names are as follows: D: Dragon descriptors; SCRI: Spartan descriptors and the *CRI*; DS: A combination of Dragon and Spartan descriptors; DS-CRI: Dragon and Spartan descriptors and the *CRI*; DCS: Dragon, Codessa, and Spartan descriptors; DCS-CRI: Dragon, Codessa and Spartan descriptors and the *CRI*. Note that the training and test sets are the same for the two methods.

Table 4.3. Statistical summary of CPNN models for *Chlorella vulgaris*. The models are numbered as cited in the text.

				Training Set					Test Set		
Model No	**Model name**	**Descriptors**	**Training / test set**	**Network size and epochs**	R^2	R^2_{cv}	***RMSE***	***AAE***	R^2	***RMSE***	***AAE***
1′	D3	ATS1m, BLTD48, piPC07	64/27	8x8 200	.964	.757	.253	.159	.915	.633	.485
2′	D4	ATS1m, BLTD48, EEig05r, R4e	63/28	7x7 100	.955	.774	.309	.222	.929	.529	.433
3′	SCRI 3	V/A, ω, CRI	71/20	8x8 100	.937	.757	.325	.252	.893	.636	.474
4′	SCRI 4	V/A, ω, CRI, E	68/23	9x9 100	.958	.740	.235	.199	.896	.652	.493
5′	SCRI 5	V/A, E, CRI, ω, E_{HOMO}	70/21	9x9 100	.950	.757	.274	.224	.857	.654	.520
6′	**DS 3**	**BLTD48, X1sol, ω**	**64/27**	**8x8 200**	**.964**	**.846**	**.270**	**.178**	**.937**	**.474**	**.386**
7′	DS 4	BLTD48, X1sol, ω, HOMT	61/30	8x8 800	.969	.774	.242	.133	.926	.506	.375
8′	DS-CRI 3	BLTD48, ω, CRI	63/28	8x8 200	.966	.828	.242	.138	.918	.702	.517
9′	DS-CRI 4	BLTD48, HOMT, ω, CRI	66/25	8x8 800	.968	.774	.247	.151	.946	.414	.340
10′	DCS 4	BLTD48, X1sol, MPC for a C, ω	65/26	9x9 100	.960	.792	.233	.200	.886	.613	.460

Bold model indicates the best 3-descriptor model. The abbreviations used for model names are as follows: D: Dragon descriptors; SCRI: Spartan descriptors and the *CRI*; DS: A combination of Dragon and Spartan descriptors; DS-CRI: Dragon and Spartan descriptors and the *CRI*; DCS: Dragon, Codessa, and Spartan descriptors; DCS-CRI: Dragon, Codessa and Spartan descriptors and the *CRI*. Note that the training and test sets are the same for the two methods.

Table 4.4. Additional statistics for the test set of MLR and CPNN models for *Chlorella vulgaris*

Model No	Name	Te R_o^2	Te $R_o'^2$	$(R^2 - R_0^2)/R^2$	k	$(R^2 - R_0'^2)/R^2$	k'
				MLR			
1	DRG 3	0.923	0.930	0.032	0.913	0.026	1.045
2	DRG 4	0.912	0.914	0.026	0.848	0.024	1.105
3	SCRI 3	0.886	0.888	0.002	*0.838**	0.000	1.060
4	SCRI 4	0.886	0.886	0.000	0.932	0.000	0.951
5	SCRI 5	0.913	0.917	0.014	0.902	0.010	1.027
6	**DS 3**	**0.918**	**0.921**	**0.013**	**0.925**	**0.010**	**1.005**
7	DS 4	0.950	0.952	0.009	0.903	0.007	1.061
8	DS-CRI 3	0.885	0.890	0.037	0.928	0.031	0.991
9	DS-CRI 4	0.950	0.952	0.005	0.971	0.003	0.983
10	DCS 4	0.958	0.960	0.007	0.882	0.005	1.095
				CPNN			
1	DRG 3	0.843	0.865	0.078	0.884	0.055	1.035
2	DRG 4	0.866	0.873	0.068	0.915	0.061	1.015
3	SCRI 3	0.831	0.857	0.070	*0.820*	0.040	1.089
4	SCRI 4	0.865	0.871	0.035	*0.819*	0.028	1.094
5	SCRI 5	0.834	0.844	0.026	0.854	0.015	1.003
6	**DS 3**	**0.906**	**0.912**	**0.034**	**0.983**	**0.027**	**0.953**
7	DS 4	0.890	0.900	0.039	0.877	0.028	1.055
8	DS-CRI 3	0.817	0.837	0.111	0.830	0.089	1.105
9	DS-CRI 4	0.926	0.933	0.021	0.933	0.014	1.013
10	DCS 4	0.858	0.870	0.032	*0.847*	0.018	1.046

*The italic values are outside the criteria limits.

The *CRI* and X1sol are highly correlated descriptors. X1sol is also a measure of branching of the molecules as the *CRI*. On this account, we employed the *CRI* instead of X1sol in model 6 (given in Equation 4.1a) and used the same training and test sets to inspect performance of the *CRI* (Equation 4.1b) in toxicity modeling. Phenylazophenol appeared again as a response outlier with a high standardized residual (>3). The 95% confidence intervals are given in parentheses. All the β-coefficients are significant at 95% level.

$$pT = -5.583\ (\pm 0.225) - 0.949\ (\pm 0.060)\ BLTD48 + 0.272\ (\pm 0.042)\ X1sol + 0.298\ (\pm 0.073)\ \omega$$

$$n = 64, \quad R^2 = 0.928, \quad F_{3,60} = 257, \quad SE = 0.397 \qquad (4.1a)$$

$$pT = -5.756\ (\pm 0.260) - 0.923\ (\pm 0.086)\ BLTD48 + 0.342\ (\pm 0.091)\ CRI + 0.594\ (\pm 0.060)\ \omega$$

$$n = 64, \quad R^2 = 0.902, \quad F_{3,60} = 183, \quad SE = 0.463 \qquad (4.1b)$$

Table 4.5 Statistical summary of models with X1sol and the *CRI*

	Training Set				Test Set		
Model	R^2	R^2_{adj}	R^2_{cv}	*RMSE*	R^2	*RMSE*	Q^2_{F3}
Equation 4.1a	0.928	0.924	0.914	0.474	0.930	0.427	0.911
Equation 4.1b	0.902	0.897	0.887	0.449	0.938	0.461	0.896
After removal of outlier							
Equation 4.2a	0.937	0.934	0.926	0.354	0.935	0.424	0.893
Equation 4.2b	0.920	0.915	0.906	0.401	0.939	0.468	0.890

Removal of outlier in Equation 4.1a and Equation 4.1b resulted in Equations 4.2a and 4.2b, respectively. Their correlation coefficients are comparable (Table 4.5).

$pT = -5.597\ (\pm 0.208) - 0.956\ (\pm 0.056)\ BLTD48 + 0.246\ (\pm 0.040)\ X1sol + 0.333\ (\pm 0.068)\ \omega$

$n = 63, \quad R^2 = 0.937, \quad F_{3,59} = 294, \quad SE = 0.365$ (4.2a)

$pT = -5.743\ (\pm 0.232) - 0.913\ (\pm 0.077)\ BLTD48 + 0.334\ (\pm 0.081)\ CRI + 0.599\ (\pm 0.053)\ \omega$

$n = 63, \quad R^2 = 0.920, \quad F_{3,59} = 225, \quad SE = 0.414$ (4.2b)

Our next attempt was to inspect CPNN models with the same descriptors and training/test set division and compare it with MLR model. To find the best performance for 3-descriptor CPNN model, the trials for network and epoch combinations together with their statistical parameters are given in Table 4.6.

Replacing X1sol with the *CRI* in CPNN model resulted in a comparable correlation coefficient (Table 4.6). A comparison of the overall statistical parameters of 3-descriptor MLR and CPNN models including the *CRI* with those of 3-descriptor MLR and CPNN models including X1sol showed that the latter models have better statistical parameters than the former models, although their correlation coefficients are comparable.

Model 6 of MLR models from Table 4.2 is important because it gives high R^2 (0.928) with a few variables and hence is considered relatively simple compared to other model, from which we get still high R^2 (model 2, $R^2 = 0.939$) but with 4 descriptors. Similarly, in CPNN models (Table 4.3), only 4-descriptor models have higher correlation coefficients than model 6 has.

Table 4.6. Comparison of the best performance of two different 3-descriptor CPNN models for *Chlorella vulgaris*

Training set				Test set		
R	R_{cv}	**Network**	**Epochs**	***R***	R^2	***RMSE***
BLTD48- ω- X1sol						
0.97	0.91	7x7	100	0.95624	0.91440	0.07480
0.97	0.91	7x7	200	0.95300	0.90820	0.07700
0.97	0.92	7x7	400	0.95276	0.90775	0.07716
0.97	0.92	7x7	800	0.95275	0.90773	0.07712
0.98	0.91	7x7	1000	0.95275	0.90773	0.07711
0.97	0.92	8x8	100	0.97100	0.94285	0.06330
0.98*	**0.92**	**8x8**	**200**	**0.96812**	**0.93725**	**0.06622**
0.98	0.91	8x8	400	0.95275	0.90773	0.07712
0.99	0.90	8x8	800	0.94910	0.90079	0.08142
0.98	0.90	8x8	1000	0.95141	0.90519	0.07803
BLTD48- ω-*CRI*						
R	R_{cv}	**Network**	**Epochs**	***R***	R^2	***RMSE***
0.98	0.87	8x8	100	0.96338	0.92810	0.07022
0.99	0.89	8x8	200	0.96493	0.93109	0.06933
0.98	**0.88**	**8x8**	**400**	**0.96500**	**0.93123**	**0.06959**

*The compared networks are written in bold.

Having known that the use of few descriptors has important advantages when constructing regression equations, we performed a MLR analysis on the entire data set of 91 compounds to develop a model with three descriptors appearing in model 6. The linear 3-descriptor model and corresponding statistics is shown below (Equation 4.3).

$$pT = -5.640\ (\pm 0.183) - 0.988\ (\pm 0.048)\ BLTD48 + 0.226\ (\pm 0.033)\ X1sol + 0.371\ (\pm 0.060)\ \omega$$

$$n = 91, \quad R^2 = 0.928, \quad R^2_{adj} = 0.926, \quad R^2_{cv} = 0.921, \quad F_{3,87} = 374, \quad SE = 0.400 \qquad (4.3)$$

Equation 4.3 is a three variable equation with 92.1 % predicted variance and 92.6% explained variance. The 95% confidence intervals are given in parentheses. All the β-coefficients are significant at 95% level. Additionally, the small difference between R^2_{cv} and R^2 indicates a stable model. However, one response outlier was apparent in this model. This could be due to the unique structure of phenylazophenol. It should be noted that it is the only compound studied with an azo (-N=N-) group. Removal of this outlier resulted in a statistically more robust model given in Equation 4.4.

$$pT = -5.642\ (\pm 0.173) - 0.988\ (\pm 0.046)\ BLTD48 + 0.211\ (\pm 0.031)\ X1sol + 0.394\ (\pm 0.057)\ \omega$$

$$n = 90, \quad R^2 = 0.935, \quad R^2_{adj} = 0.932, \quad R^2_{cv} = 0.927, \quad F_{3,86} = 410, \quad SE = 0.378 \qquad (4.4)$$

The splitting of the original data set in a representative training set of 64 and a validation set of 27 was obtained by applying SOM. After many tries with different network architectures and number of epochs (Table 4.7), the Kohonen network constructed of 10x10 neurons and 100 epochs were selected to obtain this ratio of training and test sets. The statistical parameters of the MLR model (model 6) developed based on these three variables are highlighted in Table 4.2.

Phenylazophenol appeared in the list of training set compounds and has a large standardized residual (>2.5×s). The removal of this outlier resulted in Equation 4.2a. Finalized linear model do not have any response outliers.

Table 4.7. Kohonen division trials of data set for *Chlorella vulgaris*

Architecture	Network size	Epochs	Number of compounds in training/ test set
1	**10x10**	**100**	**64/27**
2	10x10	200	64/27
3	11x11	100	69/22
4	11x11	200	66/25
5	12x12	100	68/23
6	12x12	200	74/17

$pT = -5.597\ (\pm 0.208) - 0.956\ (\pm 0.056)\ BLTD48 + 0.246\ (\pm 0.040)\ X1sol + 0.333\ (\pm 0.068)\ \omega$

$n_{training} = 63,\ R^2 = 0.937,\ R^2_{adj} = 0.934,\ R^2_{cv} = 0.926,\ F_{3,59} = 294,\ SE = 0.365;$

$n_{test} = 27,\ R^2 = 0.935,\ R^2_0 = 0.919,\ SE = 0.440,\ Q^2_{F3} = 0.893$ (4.2a)

The $[(R^2\text{-}R_0^2)/R^2]$ and k values for Equation 4.2a are found to be within the acceptable range with values being equal to 0.014 and 0.916, respectively. The *t*-values for partial correlation coefficients in Equation 4.2a are -17.208, 6.193 and 4.876 for the BLTD48, X1sol and ω, respectively. On the basis of the *t*-values for the BLTD48, X1sol and ω, it can be concluded that BLTD48 explains the toxicity more than the others. An empirical descriptor BLTD48 is related to hydrophobicity and computed with Verhaar model of Daphnia base-line toxicity (48-h) from MLOGP (mM). X1sol is the solvation connectivity index, which represents the linear fragment of one carbon atom that is defined in order to model solvation entropy and to describe dispersion interactions occurring in solutions. This index coincides with the Randic connectivity index 'χ for the hydrocarbons (Todeschini and Consonni, 2000) and has a direct relationship with toxicity. Electrophilicity (chemical potential) is related to electron affinity (Todeschini and Consonni, 2000) and tells us about the global reactivity of

the studied molecules thereby having a direct relationship with toxicity as stated by Roy et al. (2005).

The squared correlation coefficients of the training and test sets were 0.937 and 0.935, respectively. The R^2 value of 0.935 for the prediction set indicates that the MLR model can explain 93% of the variances. A chance model has low ability to reproduce y-variable of the external test set molecules. The model shown in Equation 4.2a represents high external prediction ability, similar to self prediction ability. It is interesting to note that removing 27 molecules as prediction set from the original data did not change the statistical quantity of Equation 4.2a and also slight changes were observed in the coefficients. This is a result of obtaining a representative training set with Kohonen network algorithm.

For training set of 63 compounds, 7, 9, 21, and 63 fold cross validations were run using Weka 3.6.1 (Waikato, 2009). The overall results of random deletion study statistics are summarized in Table 4.8.

Table 4.8. Leave-many-out cross validation results for Equation 4.2a.

Number of compounds deleted	**Average** R^2_{LMO}	**Average** ***RMSE***
1	0.926	0.385
3	0.927	0.381
7	0.929	0.376
9	0.929	0.376

Moreover, the model robustness was also checked by response randomization (Y-scrambling). The toxicity values were shuffled randomly between the molecules and regression models were developed. Random shuffling of response was repeated several times (25) for Equation 4.2a. R^2 values were between 0.002 and 0.173, and the average R^2 was

0.044. The results confirm that the proposed model is well founded and not just the result of a chance correlation.

The predicted vs. observed toxicity values of the training and test set compounds obtained from Equation 4.2a is shown in Figure 4.1. The outlier of the training set, phenylazophenol, is marked as a triangle in Figure 4.1.

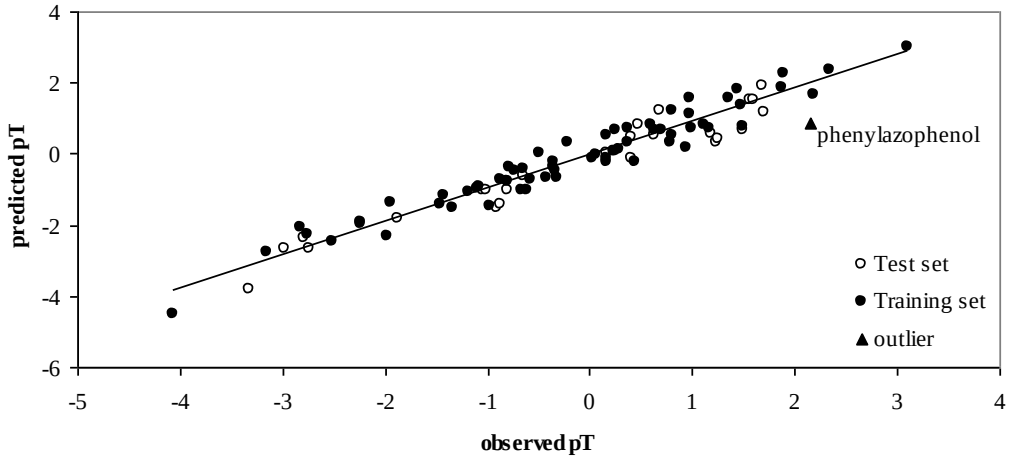


Figure 4.1. Scatter plot of predicted vs. observed toxicity of the proposed MLR model for *Chlorella vulgaris*

Outliers, in other words, compounds with high leverage values were shown in the Williams plot (Figure 4.2). Two structural outliers were apparent from the MLR model. Methidathion and piperine have higher leverage values than the critical hat value which is at 0.190. Methidathion is the only sulfur-containing aromatic heterocyclic compound in the data set. Piperine has a saturated heterocyclic ring called piperidine. These compounds are from the training set and can be called influential chemicals which are important in developed model. However, compounds having high leverage and low discrepancy do not affect the regression line but tend to increase R^2 and reduce the standard error and accepted as good leverage. Two test set compounds are appeared to have high leverage values: phosmet and malathion, however these compounds have very low residuals. For the test set, a leverage greater than h^* means that the prediction is the result of substantial extrapolation of the model and could not

be reliable. It is interesting to note that the three of the four chemicals with high leverage values belong to the pesticides.

For the comparison of linear and nonlinear models, a CPNN model was developed by using all of three descriptors appeared in Equation 4.1a using the same training and test compounds. Training set of 64 compounds was trained according to the parameters defined in Table 4.9. Different network and epochs (Table 4.10) were tried to find best architecture. Generally, a network having number of neurons (8x8 = 64) close to the number of compounds of the training set (64) is appropriate. As network size and epoch increase, possibility of overtraining (very high correlation coefficient) appears (Architecture no: 4, 9, and 10). Although $\boldsymbol{R_{cv}}$ of architecture no 6 is higher than that of no 2, its test set parameters are worse. Taking into account of best training and test set statistics simultaneously, i.e., high correlation coefficient and low *RMSE* while abstaining overtraining, the best performance was obtained using 8x8 network and 200 epochs. The scatter plot of predicted vs. observed values is illustrated in Figure 4.3. The squared correlation coefficients of the training and test sets are 0.964 and 0.937, respectively (CPNN model 6′ in Table 4.3), which are higher than the corresponding linear model. The resulted model has cross-validation coefficient R_{cv}^2 = 0.846 and predictive Q_{F3}^2 = 0.890. Three descriptors, BLTD48, X1sol, and electrophilicity were found to be important and sufficient in nonlinear modeling of algal toxicity.

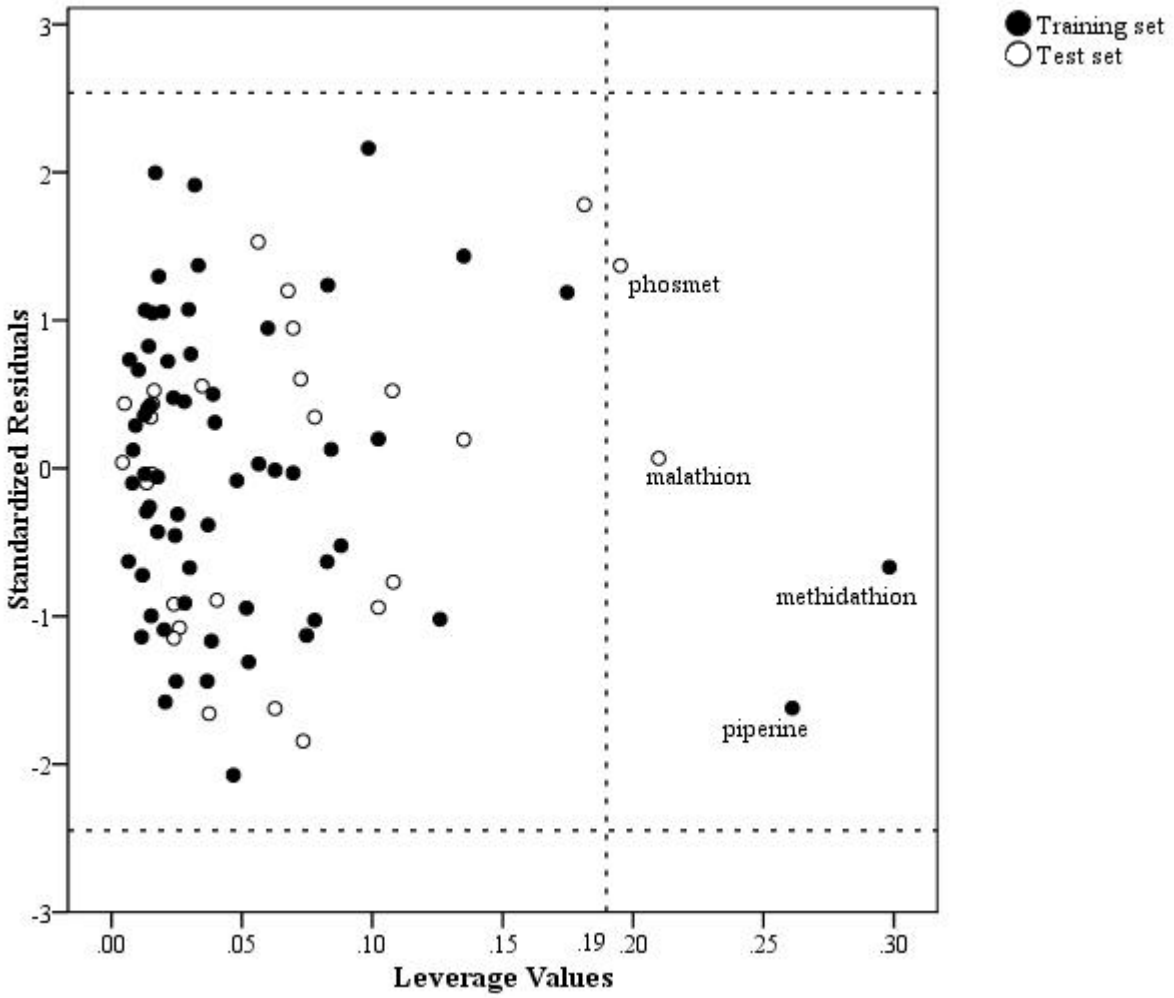


Figure 4.2. Williams plot of *Chlorella Vulgaris* data set. Filled circles are training set compounds, empty circles are test set compounds. (h^* = 0.19)

Table 4.9. Parameters used for CPNN model of *Chlorella vulgaris*

Parameter	**Value**	**Range**
random number for initialization	1234	>0
number of neurons in x direction	8	1-35
number of neurons in y direction	8	1-35
number of weights in each neuron	4	1-1400
toroid boundary conditions	no	yes; no
type of neighborhood correction	triangular	flat; triangular; chef hat function; Mexican hat function
furthest neuron for corrections	8	1-8
maximal correction factor	0.50	0.1-0.9
minimal correction factor	0.01	0.00-maximal correction factor
epochs	200	1-∞

Table 4.10. Trials to find the best performance of 3-decriptor CPNN model for *Chlorella vulgaris**

Architecture	Architecture		Training set		Test set	
no	**Network**	**Epochs**	***R***	R_{cv}	R^2	***RMSE***
1	8x8	100	0.97	0.92	0.943	0.0633
2	**8x8**	**200**	**0.98**	**0.92**	**0.937**	**0.0662**
3	8x8	400	0.98	0.91	0.907	0.0771
4	8x8	800	0.99	0.90	0.901	0.0814
5	8x8	1000	0.98	0.90	0.905	0.0780
6	9x9	100	0.98	0.93	0.921	0.0734
7	9x9	200	0.98	0.92	0.913	0.0720
8	9x9	400	0.98	0.93	0.905	0.0740
9	9x9	800	0.99	0.93	0.904	0.0752
10	9x9	1000	0.99	0.92	0.904	0.0753

* Statistics are obtained from normalized *p*T and descriptor values.

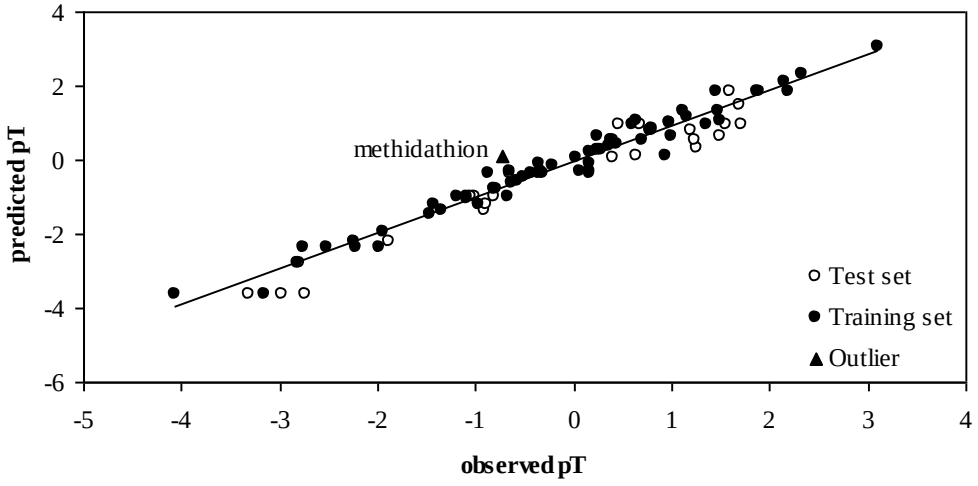
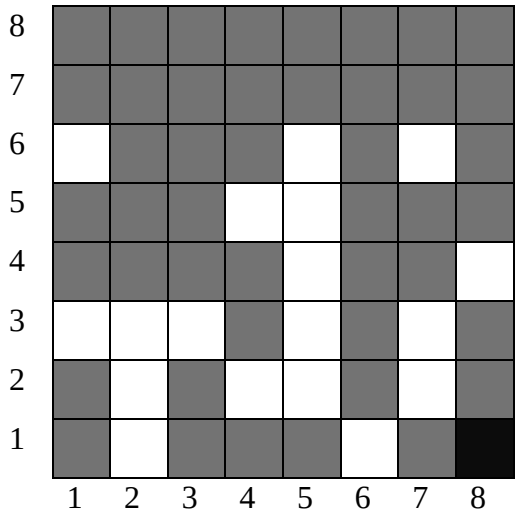
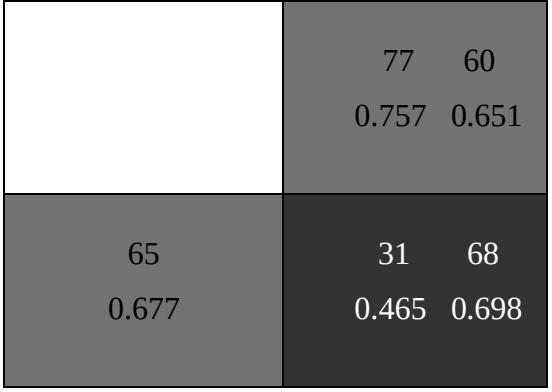


Figure 4.3. Scatter plot of predicted vs. observed toxicity values for the 3-descriptor CPNN model for *Chlorella vulgaris* (Model 6′ in Table 4.3)

Compounds of the training set with an absolute error (*AE*) greater than 0.1 were analyzed as possible outliers of the model. Based on this criteria, possible outliers of 3-descriptor CPNN model (Model 6′ in Table 4.3) methidathion (no. 31) and thiometon (no. 68) located in neuron (8, 1) (Figure 4.4 (a)). These compounds are located in the same neuron because they have very similar descriptor values. If these compounds have the similar toxicity values, then we have a confirmation of our model. If their toxicity values differ essentially, there is an implication of real outliers. Selection of these outliers is made by inspecting toxicity values of the neuron containing possible outliers and surrounding neurons. Outlier selection procedure is simulated in Figure 4.4 (b). Since the compound no 31 (methidathion) has a very different toxicity than the other neighboring compounds, it is identified as an outlier and shown in Figure 4.3. The same compound has been reported as an outlier in *k*NN modeling of the same data set by Cronin et al. (2004). Reason for being an outlier of the model could be due to the fact that methidathion is the only compound with a sulfur-containing aromatic heterocyclic structure in the pesticide sub-set. None of the test set compounds had *AE* greater than 0.2, therefore test set do not have any outliers.





(a)

Figure 4.4. Outlier analysis for CPNN model for *Chlorella vulgaris* (a) Top map of the training algorithm. Grey cells are occupied neurons, black neuron have the possible outlier(s). (b) In each cell, normalized toxicity value is under the compound number

Methidathion was removed from the training set and using the same parameters as in Table 4.9, a new model was developed. We obtained R^2 = 0.983 and R^2_{cv} = 0.846 for training set and R^2 = 0.912 and *RMSE* = 0.547 for the test set. Removing the outlier from training set resulted in higher R^2, however test set R^2 was decreased and *RMSE* was increased. It may be concluded that methidathion act as an influential chemical in the determination of model descriptors. Therefore, we proposed the final model including methidathion.

AD of the proposed models (Equation 4.2a for MLR and model 6' for CPNN) is defined by the following limits given in Table 4.11.

Table 4.11. Boundaries of the proposed MLR and CPNN models for *Chlorella vulgaris*

	Training set		**Test set**	
	min	**max**	**min**	**max**
*p*T	-4.06	3.10	-3.32	1.71
Electrophilicity	1.04	5.37	0.993	5.17
BLTD48	-5.05	-1.16	-5.68	-0.55
X1sol	1.41	10.33	1.00	10.87

Pesticides, the CAS numbers, descriptor values, their observed and predicted toxicity values to *Chlorella vulgaris*, and residuals obtained from both methods are given in Table 4.12.

Table 4.12 The CAS numbers, descriptors, observed and predicted toxicity values obtained from MLR (Equation 4.2a) and CPNN (Model no 6') for *Chlorella vulgaris*

Comp no	CAS number	Compound name	ω	BLTD48	X1sol	pT_{obs}	pT_{pred}			
							MLR		CPNN	
							from Eq 4.2a	res	from Model no 6'	res
Training set										
1	67-56-1	Methanol	0.993	-0.55	1.000	-4.06	-4.49	0.43	-3.62	-0.44
3	75-65-0	2-methyl-propan-2-ol	1.107	-2.08	2.000	-3.16	-2.75	-0.41	-3.62	0.46
5	868-77-9	2-hydroxyethyl methacrylate	2.559	-1.73	4.181	-2.82	-2.06	-0.76	-2.79	-0.03
7	96-33-3	Methyl acrylate	2.830	-1.77	2.808	-2.75	-2.27	-0.48	-2.37	-0.38
9	78-93-3	Butanone	2.115	-1.94	2.270	-2.51	-2.48	-0.03	-2.37	-0.14
10	80-62-6	Methyl methacrylate	2.529	-2.16	3.181	-2.24	-1.91	-0.33	-2.21	-0.03
11	96-22-0	Pentan-3-one	2.084	-2.33	2.808	-2.23	-1.98	-0.25	-2.37	0.14
12	4170-30-3	Crotonaldehyde	2.757	-1.84	2.414	-1.98	-2.33	0.35	-2.37	0.39
13	6728-26-3	Trans-2-hexenal	2.756	-2.59	3.414	-1.94	-1.36	-0.58	-1.95	0.01
15	108-95-2	Phenol	2.085	-2.75	3.394	-1.46	-1.44	-0.02	-1.46	0.00
16	96-05-9	Allyl methacrylate	2.295	-2.77	4.181	-1.42	-1.16	-0.26	-1.20	-0.22
17	62-53-3	Aniline	1.858	-2.75	3.394	-1.34	-1.51	0.17	-1.35	0.01
18	110-43-0	2-heptanone	2.101	-3.03	3.770	-1.18	-1.07	-0.11	-0.98	-0.20
19	100-66-3	Anisole	2.044	-3.09	3.932	-1.09	-0.99	-0.10	-0.98	-0.11
20	367-12-4	2-fluorophenol	2.398	-3.18	3.394	-1.08	-0.92	-0.16	-1.07	-0.01
23	150-76-5	4-methoxyphenol	2.017	-2.50	4.326	-0.97	-1.47	0.50	-1.20	0.23
26	87-62-7	2,6-dimethylaniline	1.817	-3.40	4.215	-0.87	-0.70	-0.17	-0.36	-0.51
27	100-52-7	Benzaldehyde	2.900	-3.00	3.932	-0.81	-0.80	-0.01	-0.81	0.00
30	98-95-3	Nitrobenzene	3.637	-3.11	4.305	-0.78	-0.35	-0.43	-0.77	-0.01
31	950-37-8	Methidathion	5.172	-1.07	9.621	-0.73	-0.49	-0.24	0.12	-0.85
32	106-44-5	4-cresol	2.004	-3.09	3.788	-0.66	-1.04	0.38	-0.98	0.32
34	104-87-0	4-tolualdehyde	2.818	-3.32	4.326	-0.65	-0.42	-0.23	-0.30	-0.35
35	94-71-3	2-ethoxyphenol	2.093	-2.82	4.843	-0.62	-1.01	0.39	-0.62	0.00

Table 4.12 (continued)

Comp no	CAS number	Compound name	ω	BLTD48	X1sol	pT_{obs}	pT_{pred} MLR from Eq 4.2a	MLR res	pT_{pred} CPNN from Model no 6'	CPNN res
36	24964-64-5	3-cyanobenzaldehyde	3.454	-2.65	4.864	-0.57	-0.72	0.15	-0.57	0.00
37	99-08-1	3-nitrotoluene	3.510	-3.45	4.698	-0.50	0.03	-0.53	-0.49	-0.01
38	106-48-9	4-chlorophenol	2.216	-3.34	4.076	-0.42	-0.66	0.24	-0.37	-0.05
39	97-02-9	2,4-dinitroaniline	4.013	-2.65	6.020	-0.36	-0.25	-0.11	-0.34	-0.02
40	106-41-2	4-bromophenol	2.348	-3.50	4.365	-0.35	-0.39	0.04	-0.09	-0.26
41	106-40-1	4-bromoaniline	2.150	-3.50	4.365	-0.33	-0.46	0.13	-0.33	0.00
42	108-42-9	3-chloroaniline	2.102	-3.34	4.076	-0.31	-0.70	0.39	-0.37	0.06
43	2495-37-6	Benzyl methacrylate	2.323	-3.78	6.198	-0.21	0.32	-0.53	-0.16	-0.05
44	618-87-1	3,5-dinitroaniline	4.306	-2.65	6.003	0.03	-0.15	0.18	0.04	-0.01
45	89-98-5	2-chlorobenzaldehyde	2.999	-3.57	4.631	0.06	-0.05	0.11	-0.30	0.36
46	540-38-5	4-iodophenol	2.536	-3.66	4.654	0.16	-0.11	0.27	-0.09	0.25
48	58-27-5	2-methyl-1,4-naphthoquinone	3.927	-3.44	6.198	0.16	0.52	-0.36	0.19	-0.03
49	88-69-7	2-isopropylphenol	2.049	-3.70	4.715	0.17	-0.22	0.39	-0.36	0.53
50	626-43-7	3,5-dichloroaniline	2.321	-3.91	4.759	0.24	0.09	0.15	0.26	-0.02
51	603-71-4	1,3,5-trimethyl-2-nitrobenzene	3.045	-4.07	5.520	0.25	0.67	-0.42	0.62	-0.37
53	88-18-6	2-tert-butyl phenol	2.022	-3.99	5.016	0.29	0.13	0.16	0.26	0.03
54	95-50-1	1,2-dichlorobenzene	2.453	-4.62	4.382	0.37	0.72	-0.35	0.36	0.01
55	99-65-0	1,3-dinirtobenzene	4.737	-3.11	5.609	0.38	0.33	0.05	0.54	-0.16
58	99-61-6	3-nitrobenzaldehyde	4.009	-2.85	5.236	0.45	-0.25	0.70	0.43	0.02
60	298-00-0	Methylparathion	4.753	-2.84	8.659	0.60	0.83	-0.23	0.97	-0.37
62	99-30-9	2,6-dichloro-4-nitroaniline	3.540	-3.76	6.097	0.64	0.68	-0.04	1.06	-0.42
64	121-14-2	2,4-dinitrotoluene	4.526	-3.45	6.020	0.70	0.69	0.01	0.54	0.16
65	2636-26-2	Cyanophos	4.379	-2.55	8.287	0.79	0.34	0.45	0.78	0.01

Table 4.12 (continued)

Comp no	CAS number	Compound name	ω	BLTD48	X1sol	pT_{obs}	pT_{pred} MLR from Eq 4.2a	MLR res	pT_{pred} CPNN from Model no 6'	CPNN res
66	3531-19-9	6-chloro-2,4-dinitroaniline	4.150	-3.24	6.719	0.80	0.54	0.26	0.79	0.01
67	99-28-5	2,6-dibromo-4-nitrophenol	3.851	-4.06	6.674	0.81	1.21	-0.4	0.82	-0.01
68	640-15-3	Thiometon	5.026	-2.24	7.852	0.94	0.15	0.79	0.12	0.82
69	89-61-2	2,5-dichloronitrobenzene	3.680	-4.27	5.686	0.97	1.11	-0.14	0.98	-0.01
70	94-62-2	Piperine	3.046	-3.77	10.327	0.97	1.56	-0.59	1.02	-0.05
71	939-97-9	4-tert-butylbenzaldehyde	2.814	-4.18	5.537	1.00	0.70	0.3	0.62	0.38
72	634-93-5	2,4,6-trichloroaniline	2.415	-4.47	5.475	1.11	0.83	0.28	1.30	-0.19
73	83-42-1	2-chloro-6-nitrotoluene	3.355	-4.02	5.415	1.17	0.70	0.47	1.16	0.01
77	2463-84-5	Dicapthon	4.791	-3.40	9.359	1.36	1.55	-0.19	0.97	0.39
78	128-37-0	2,6-di-tert-butyl-4-methyl phenol	1.897	-5.29	7.032	1.45	1.82	-0.37	1.83	-0.38
79	3481-20-7	2,3,5,6-tetrachloroaniline	2.671	-4.76	6.191	1.48	1.37	0.11	1.30	0.18
80	609-89-2	2,4-dichloro-6-nitrophenol	3.821	-3.76	6.097	1.50	0.77	0.73	1.06	0.44
86	89-69-0	1,2,4-trichloro-5-nitrobenzene	3.882	-4.83	6.386	1.88	1.88	0.00	1.86	0.02
87	6284-83-9	1,3,5-trichloro-2,4-dinitrobenzene	4.783	-4.57	7.724	1.89	2.27	-0.38	1.86	0.03
89	90134-10-4	4-(dibutylamino) benzaldehyde	2.376	-4.62	8.312	2.18	1.66	0.52	1.83	0.35
90	117-18-0	2,3,5,6-tetrachloronitrobenzene	3.922	-5.12	7.102	2.34	2.35	-0.01	2.33	0.01
91	608-71-9	Pentabromophenol	3.421	-5.68	8.351	3.10	3.03	0.07	3.05	0.05
88	1689-82-3	Phenylazophenol	2.918	-3.86	7.343	2.16	0.87	1.29	2.12	0.04
Test set										
2	64-17-5	Ethanol	1.041	-1.16	1.414	-3.32	-3.79	0.47	-3.62	0.30
4	78-92-2	Butan-2-ol	1.122	-2.08	2.270	-2.98	-2.68	-0.30	-3.62	0.64
6	818-61-1	2-hydroxyethyl acrylate	2.985	-1.37	3.808	-2.79	-2.36	-0.43	-2.79	0.00
8	71-36-3	Butan-1-ol	1.058	-2.08	2.414	-2.73	-2.66	-0.07	-3.62	0.89

Table 4.12 (continued)

Comp no	CAS number	Compound name	ω	BLTD48	X1sol	pT_{obs}	pT_{pred} MLR from Eq 4.2a	MLR res	pT_{pred} CPNN from Model no 6'	CPNN res
14	1576-87-0	Trans-2-pentenal	2.756	-2.23	2.914	-1.88	-1.83	-0.05	-2.20	0.32
21	348-54-9	2-fluoroaniline	2.130	-3.18	3.394	-1.05	-1.01	-0.04	-0.98	-0.07
22	108-39-4	3-cresol	2.004	-3.09	3.788	-1.01	-1.04	0.03	-0.98	-0.03
24	95-55-6	2-hydroxyaniline	1.868	-2.65	3.805	-0.91	-1.51	0.60	-1.35	0.44
25	90-05-1	2-methoxyphenol	2.181	-2.50	4.343	-0.88	-1.41	0.53	-1.20	0.32
28	95-48-7	2-cresol	2.066	-3.09	3.805	-0.81	-1.02	0.21	-0.98	0.17
29	90-02-8	2-hydroxybenzaldehyde	2.879	-2.90	4.343	-0.80	-0.80	0.00	-0.81	0.01
33	95-65-8	3,4-dimethylphenol	1.999	-3.40	4.198	-0.65	-0.65	0.00	-0.37	-0.28
47	4748-78-1	4-ethylbenzaldehyde	2.838	-3.62	4.864	0.16	0.01	0.15	-0.30	0.46
52	608-31-1	2,6-dichloroaniline	2.267	-3.91	4.793	0.26	0.07	0.19	0.27	-0.01
56	51-28-5	2,4-dinitrophenol	4.387	-2.65	6.020	0.40	-0.12	0.52	0.04	0.36
57	100-25-4	1,4-dinitrobenzene	5.078	-3.11	5.609	0.41	0.45	-0.04	0.54	-0.13
59	732-11-6	Phosmet	4.860	-2.29	10.691	0.47	0.84	-0.37	0.97	-0.50
61	121-75-5	Malathion	5.371	-1.79	10.543	0.64	0.50	0.14	0.12	0.52
63	86-50-0	Methyl azinphos	5.073	-2.57	10.868	0.69	1.22	-0.53	0.97	-0.28
74	5388-62-5	4-chloro-2,6-dinitroaniline	4.245	-3.24	6.719	1.19	0.57	0.62	0.79	0.40
75	528-29-0	1,2-dinitrobenzene	4.722	-3.11	5.626	1.23	0.33	0.90	0.54	0.69
76	100-00-5	1-chloro-4-nitrobenzene	3.780	-3.70	4.987	1.25	0.43	0.82	0.31	0.94
81	83-38-5	2,6-dichlorobenzaldehyde	3.082	-4.13	5.331	1.50	0.69	0.81	0.62	0.88
82	55-38-9	Fenthion	3.961	-3.69	9.255	1.56	1.53	0.03	0.97	0.59
83	96-76-4	2,4-di-tert-butylphenol	1.942	-5.05	6.621	1.60	1.51	0.09	1.82	-0.22
84	87-86-5	Pentachlorophenol	2.952	-5.03	6.907	1.69	1.89	-0.20	1.45	0.24
85	122-14-5	Fenitrothion	4.603	-3.14	9.070	1.71	1.17	0.54	0.97	0.74

The descriptors used in the present study with those reported by Cronin et al. (2004) for describing the toxicity of the same compounds are coherent. They have reported 3-descriptor model in which they have selected two descriptors (log K_{ow} and E_{LUMO}) empirically as a result of their experience in modeling acute toxicity. The third descriptor appeared in their model was $\Delta^1\chi^v$. As demonstrated in the literature (Pontolillo and Eganhouse, 2001), the octanol–water partition coefficient of a given compound could be subject to high variability due to the applied experimental procedure or the selected calculation method. Thus, the accuracy and quality of a QSAR model are often greatly affected by the specific log K_{ow} used. Since BLTD48 appearing in our models is a hydropobicity related parameter like log K_{ow} and calculated directly from the molecular structure, the use of this descriptor instead of log K_{ow} will eliminate the drawbacks related to log K_{ow}. Electrophilicity in our toxicity model coincides with E_{LUMO} which quantifies the electrophilic potency of biochemically reactive compounds, whereas X1sol which is a type of connectivity index coincides with the third descriptor, $\Delta^1\chi^v$, reported by Cronin et al. (2004).

It is of our interest to compare the results of the optimized model with those of recently published studies in which QSTR models were developed using the same data set and using same number of descriptors (Table 4.13). Considering the whole data set, our MLR model reveals a better performance than its counterparts. 3-descriptor CPNN model seems to perform the best in terms of R^2 for the training set. Of the results reported in Table 4.13, all of our models have lower errors than the other models. Additionally, in our MLR model R^2_{cv} and R^2 of test set are better than those of reported by Cronin et al. (2004). The reason for this can be due to the different training/test set ratio. On the other hand, their nonlinear model parameters for the test set are slightly better than ours. It should be noted that their test set compounds is less than ours. The best 4-descriptor model developed in this study was compared with the 4-descriptor model utilized by Roy and Gosh (2007). Both 4-descriptor MLR and CPNN models are superior to the 4-descriptor PLS model in terms of statistical parameters.

Table 4.13. Comparison of different QSTR models of algae *Chlorella vulgaris*

	3-descriptor Models					4-descriptor Models		
	Cronin et al., (2004)		Roy and Gosh (2007)	Present Study*		Roy and Gosh (2007)	Present Study	
Technique	**MLR**	***k*NN**	**MLR (ETA descriptors)**	**MLR**	**CPNN**	**PLS (ETA descriptors) Model no 21**	**MLR**	**CPNN**
Full set				**Eq. 4.3**				
R^2	0.890	0.824	0.849	**0.928**	-			
R^2_{cv}	0.875	0.623	0.832	**0.921**	-			
SE	0.494	-	0.580	**0.400**	-			
F	235	-	163	**374**	-			
R^2_{adj}	-	-	0.844	**0.926**	-			
N	91	91	91	**91**				
Training set				**Eq. 4.2a**	**Model 6'**		**Model 2**	**Model 2'**
R^2	0.892	0.824	-	**0.937**	**0.964**	0.915	**0.939**	**0.955**
R^2_{cv}	0.878	-	-	**0.926**	**0.846**	0.897	**0.929**	**0.774**
SE	0.496	0.623	-	**0.365**	-			
F	189	-	-	**294**	-	169	**225**	
N	73	73		**63**	**64**	68	**63**	**63**
Test set								
R^2	0.901	0.941	-	**0.935**	**0.937**	0.812	**0.937**	**0.929**
R^2_0	0.860	0.937	-	**0.919**	**0.906**	-		
n	18	18		**27**	**27**	23	**28**	**28**

*Results of present study are written in bold.

4.2. MLR and CPNN Models for *Oncorhynchus Mykiss*

The 3-descriptor MLR model developed for a small and heterogeneous data set without an outlier is given in Equation 4.5. HM selected three descriptors from Dragon software namely, HATS6m, Mor27u, and RDF080p and yielded the best MLR model for rainbow trout. The 95% confidence intervals are given in parentheses. All the β-coefficients are significant at 95% level.

$$pT = -0.491\ (\pm 0.134) - 2.534\ (\pm 0.231)\ \text{HATS6m} - 3.296\ (\pm 0.585)\ \text{Mor27u} + 0.161\ (\pm 0.045)\ \text{RDF080p}$$

$$n = 34, \quad R^2 = 0.837, \quad R^2_{adj} = 0.820, \quad R^2_{cv} = 0.760, \quad F_{3,21} = 51.23, \quad SE = 0.461 \tag{4.5}$$

Data set was split into training and test sets using Kohonen network. We wrote in bold the selected 6x6 network and 1000 epochs combination in Table 4.14.

Table 4.14. Kohonen division trials of data set for *Oncorhynchus mykiss*

Architecture	Network size	Epochs	Number of compounds in training/ test set
1	6x6	100	23/ 11
2	6x6	200	22/ 12
3	6x6	800	23/ 11
4	7x7	100	27/ 7
5	**6x6**	**1000**	**25/ 9**

For the training set selected by Kohonen network described above, the following MLR model developed without an outlier.

pT = -0.484 (±0.172) – 2.456 (±0.264) HATS6m – 2.605 (±0.796) Mor27u + 0.139 (±0.056) RDF080p

$n_{\text{training}} = 25$, $R^2 = 0.819$, $R^2_{adj} = 0.794$, $R^2_{cv} = 0.696$, $F_{3, 21} = 31.75$, $SE = 0.486$;

$n_{\text{test}} = 9$, $R^2 = 0.867$, $R_0^2 = 0.867$, $SE = 0.508$, $Q^2_{F3} = 0.817$ (4.6)

The *t*-values for partial correlation coefficients in Equation 4.6 are 9.312, -3.274, and 2.467 for the HATS6m, Mor27u and RDF080p, respectively. On the basis of the *t*-values, it can be concluded that HATS6m explains the toxicity significantly more than the others.

The predicted vs. observed toxicity values of the training and test set compounds obtained from Equation 4.6 are shown in Figure 4.5.

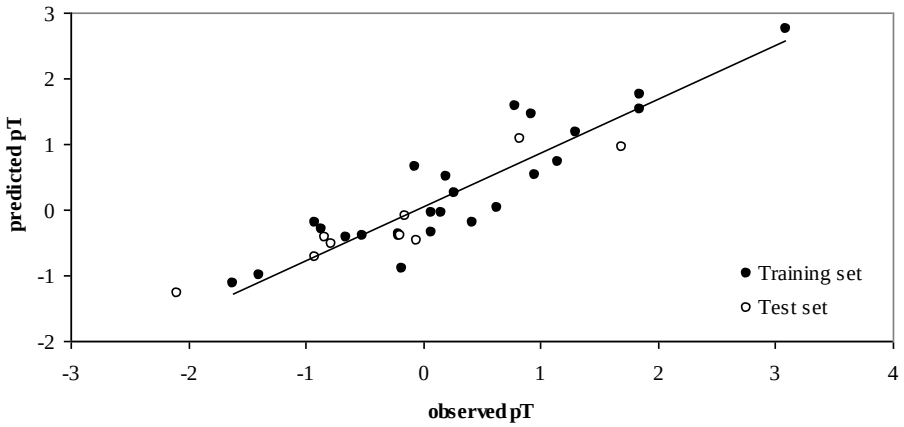


Figure 4.5. Scatter plot of predicted vs. observed toxicity of the proposed MLR model for *Oncorhynchus mykiss*

Descriptors appearing in proposed MLR model are all Dragon descriptors. Mor27u is a 3D-MoRSE-signal 27/unweighted descriptor which is calculated by summing atom weights viewed by a different angular scattering function. MoRSE (Molecule Representation of Structures based on Electron diffraction) descriptors are used in many

QSAR study. For instance, Mor27u was used by Gramatica et al. (2003) in a mutagenicity model. Their model pointed out the importance of structural descriptors. HATS6m is a GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptor obtained from Dragon. This descriptor is leverage-weighted autocorrelation of lag 6/ weighted by atomic masses (*lag* is a topological distance equal to seven). Consonni et al. (2002) reported that the GETAWAY descriptors have an overall good modeling capability, proving their usefulness in QSAR/QSPR studies. HATS6m was successfully used in mitochondrial toxicity modeling with support vector machine (SVM) method combined with GA by Zhang et al. (2009). RDF descriptors are calculated from the radial distribution function of an ensemble of N atoms that can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *r*. RDF080p is a radial distribution function descriptor weighted by atomic polarizabilities and take into account the atoms inside virtual spheres of 8.0 Å spheres.

The statistical parameters of MLR models were compared to those of CPNN models in terms of internal and external validation parameters (Table 4.15).

Table 4.15. Statistical summary of MLR and CPNN models for *Oncorhynchus mykiss*

		Training Set				**Test Set**		
Network size and epochs		R^2	R^2_{cv}	*RMSE*	*AAE*	R^2	*RMSE*	*AAE*
				MLR				
		0.819	0.696	0.445	0.386	0.867	0.448	0.376
				CPNN				
5x5	200	0.945	0.497	0.239	0.108	0.830	0.521	0.460

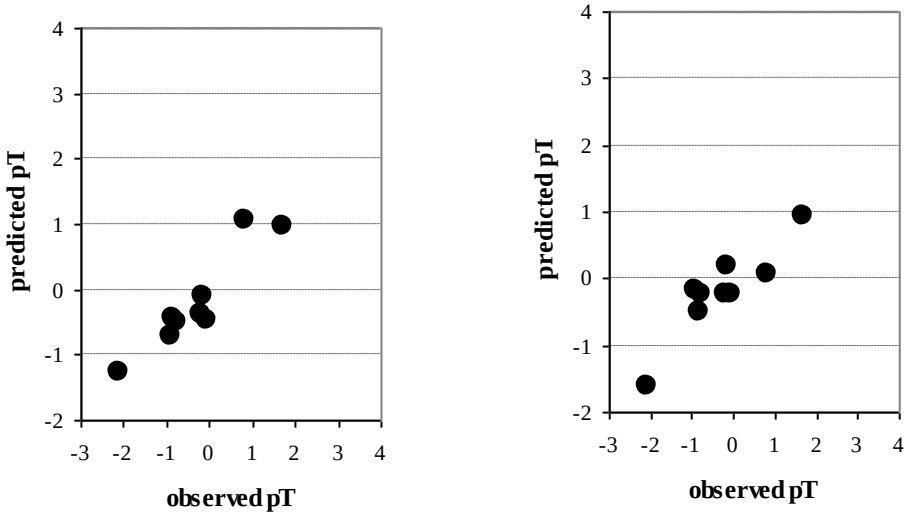
The Golbraikh's criteria results indicate a trend of underestimation of the toxicity for the test set compounds (Table 4.16). When inspecting the test set graphs of predicted vs. observed values both for MLR and CPNN (Figure 4.6 (a) and (b), respectively), it is

obvious that predicted values are compressed to about between -1 and 1 for both models. This may be due to the small size of test set.

Table 4.16. Additional statistics for the test set of MLR and CPNN models for *Oncorhynchus mykiss*

	Te R_o^2	Te $R_o'^2$	$(R^2 - R_0^2)/R^2$	k	$(R^2 - R_0'^2)/R^2$	k'
MLR	0.867	0.867	0.000	*0.651**	0.000	*1.331*
CPNN	0.827	0.830	0.000	*0.568*	0.000	*1.462*

* The italic values are outside the criteria limits.



(a) (b)

Figure 4.6. Scatter plot of predicted vs. observed toxicity values of the test set compounds of proposed (a) MLR model (b) CPNN model

For training set of 25 compounds, 5 and 8 fold cross validations were run using Weka 3.6.1 (Waikato, 2009). The overall results of random deletion study statistics are summarized in Table 4.17.

Table 4.17. Leave-many-out cross validation results for MLR model of *Oncorhynchus mykiss*

Number of compounds deleted	**Average** R^2_{LMO}	**Average *RMSE***
1	0.696	0.581
3	0.667	0.608
5	0.711	0.566

Endosulfan seems to be an influential compound of the training set with a high leverage value as seen in the Williams plot (Figure 4.7). However, this compound was well predicted with a low residual. In this case, this compound is labeled as "good leverage" (Gramatica, 2007). On the other hand, although, endosulfan had higher leverage (h = 0.51) and yet closer to h^* of 0.48, therefore, it is not considered structurally the most influential in determination of the model descriptors. The toxicity value of endosulfan is much higher than the compounds in the data set because endosulfan has significantly higher water solubility than other organochlorines (Walker et. al, 2006). It is likely that high toxicity of this compound is due to its water solubility. Being a high leverage compound is consistent with the fact that cyclodiene structure of endosulfan is very different than the structures of the other compounds in the aromatic chlorines set.

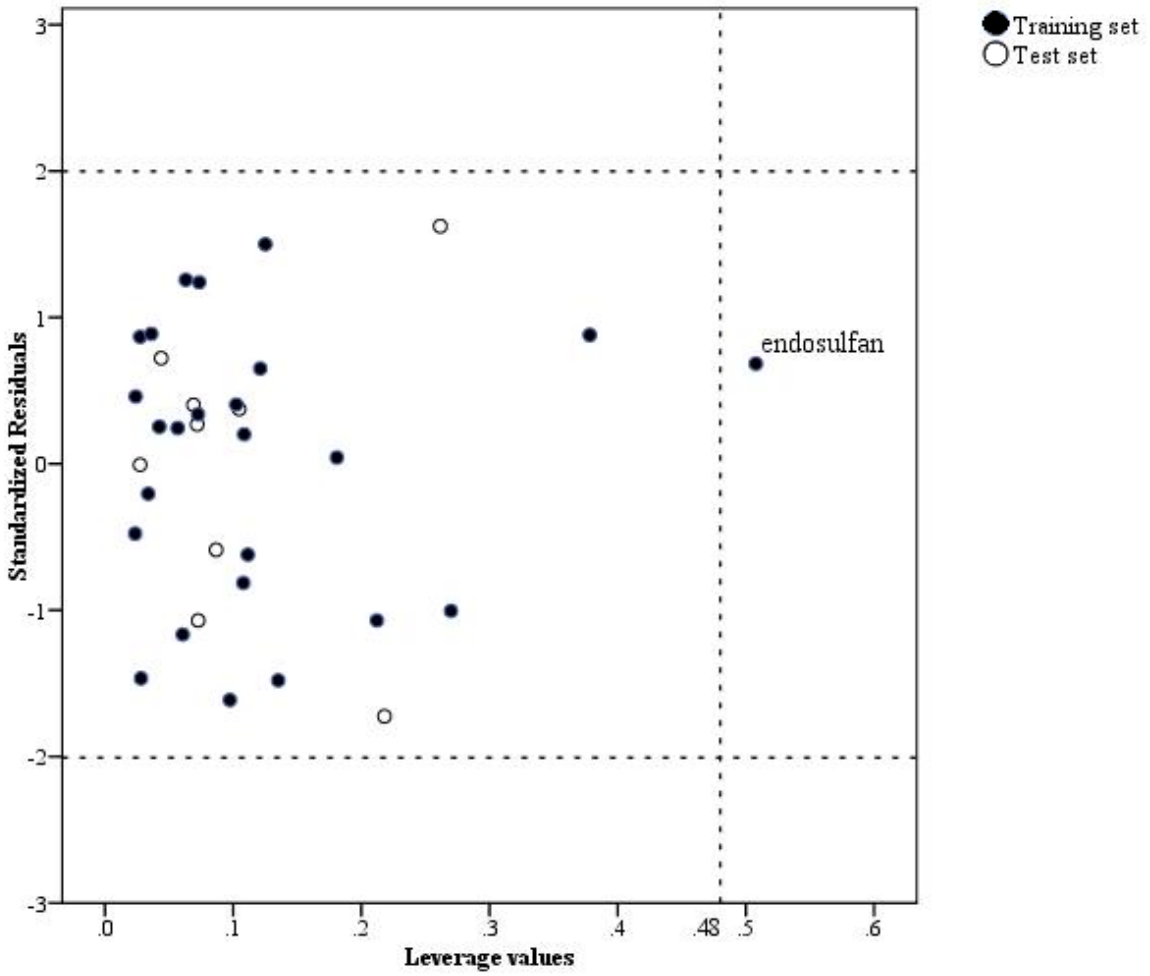


Figure 4.7. Williams plot of *Oncorhynchus mykiss* data set. Filled circles are training set compounds, empty circles are test set compounds ($h^* = 0.48$)

Moreover, the model robustness was also checked by response randomization (Y-scrambling). The toxicity values were shuffled randomly between the molecules and regression models were developed. Random shuffling of response was repeated several times (25) for MLR equation. R^2 values were between 0.007 and 0.302, and the average R^2 was 0.078. The results reveal that the proposed model is well founded and not just the result of a chance correlation.

A CPNN model was developed by using three descriptors appeared in Equation 4.6. Training set of 25 compounds was trained according to the parameters defined in

Table 4.18. The best performance was obtained using 5x5 network architecture and 200 epochs (Table 4.19). The squared correlation coefficients of the training and test sets are 0.945 and 0.830, respectively. The resulted model has cross-validation coefficient R^2_{cv} = 0.497 and predictive Q^2_{F3} = 0.753. Predicted vs. observed values are illustrated in Figure 4.8. Although statistical quality is sufficient, it is not superior to MLR model.

Table 4.18. Parameters used for CPNN model of *Oncorhynchus mykiss*

Parameter	**Value**	**Range**
random number for initialization	1234	>0
number of neurons in x direction	5	1-35
number of neurons in y direction	5	1-35
number of weights in each neuron	4	1-1400
toroid boundary conditions	no	yes; no
type of neighborhood correction	triangular	flat; triangular; chef hat function; Mexican hat function
furthest neuron for corrections	8	1-8
maximal correction factor	0.50	0.1-0.9
minimal correction factor	0.01	0.00-maximal correction factor
epochs	200	1-∞

Table 4.19. Trials to find the best performance of CPNN model for *Oncorhynchus mykiss* *

Architecture no	Architecture		Training set		Test set	
	Network	Epochs	R	R_{cv}	R^2	RMSE
1	5x5	100	0.92	0.68	0.805	0.1009
2	**5x5**	**200**	**0.97**	**0.70**	**0.830**	**0.1006**
3	5x5	400	0.95	0.69	0.847	0.1130
4	5x5	800	0.97	0.63	0.829	0.1006
5	5x5	1000	0.96	0.70	0.536	0.1404
6	6x6	100	0.98	0.60	0.826	0.1107
7	6x6	200	0.98	0.64	0.859	0.0969
8	6x6	400	0.98	0.64	0.774	0.1127
9	6x6	800	0.98	0.65	0.796	0.1100
10	6x6	1000	0.98	0.67	0.726	0.1009

* Statistics are obtained from normalized *p*T and descriptor values.

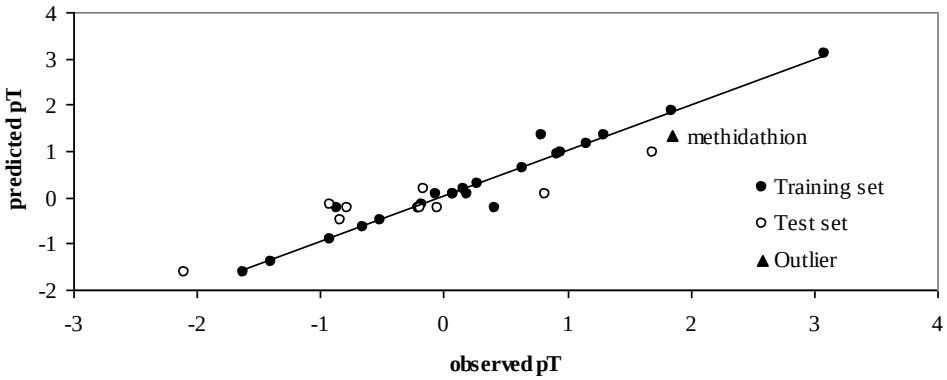


Figure 4.8. Scatter plot of predicted vs. observed toxicity values of the proposed CPNN model for *Oncorhynchus mykiss*

Determination of outliers was done in three steps. First, we analyzed the absolute errors. Absolute errors greater than 0.1 for the training set and greater than 0.2 for the test

set were labeled as possible outliers. Compounds of the training set with an absolute error higher than 0.1 are listed in Table 4.20.

Table 4.20. Possible outliers for the training set (*AE* > 0.1)

ID	name	*AE*	nx	ny
4	methidathion	0.102	3	5
34	dicofol	0.103	3	5
22	carbofuran	0.124	5	1
25	pebulate	0.125	5	1

In the second step, we analyzed if two or more compounds are located on the same neuron. These compounds are located in the same neuron because they have very similar descriptor values. Outlier selection procedure is simulated in Figure 4.9.

Selection of these outliers is made by inspecting toxicity values of the neuron containing possible outliers and surrounding neurons. For the objects associated with neuron (5, 1) (Figure 4.9 (b)), it was not possible to determine an outlier because the neighboring neurons do not present a clear trend. Therefore, in this case, all the objects were kept. However, in neuron (3, 5) (Figure 4.9 (c)), compound no 4 (methidathion) has a very different toxicity value comparing with its neuron-mate and neighboring neurons. Thus, methidation is identified as outlier of this model. It is seen that this compound was underestimated by model. This might be due to the metabolite of methidathion or having different MOA. Another reason could be the structure of this compound because methidathion is the only compound with a sulfur-containing five-membered aromatic heterocyclic ring in the data set. The model may not be representing this compound with these descriptors. None of the test set compounds had *AE* greater than 0.2, therefore test set do not have any outliers.

Methidathion was removed from the training set and using the same parameters as in Table 4.18, a new model was developed. We obtained R^2 = 0.951 and R^2_{cv} = 0.429 for

training set and R^2 = 0.858 and *RMSE* = 0.501 for the test set. Removing the outlier from training set resulted in higher R^2; however cross-validation correlation coefficient dropped well below the acceptable limit (0.5). Although test set statistics are improved, because of the low R^2_{cv} value, this model was not accepted.

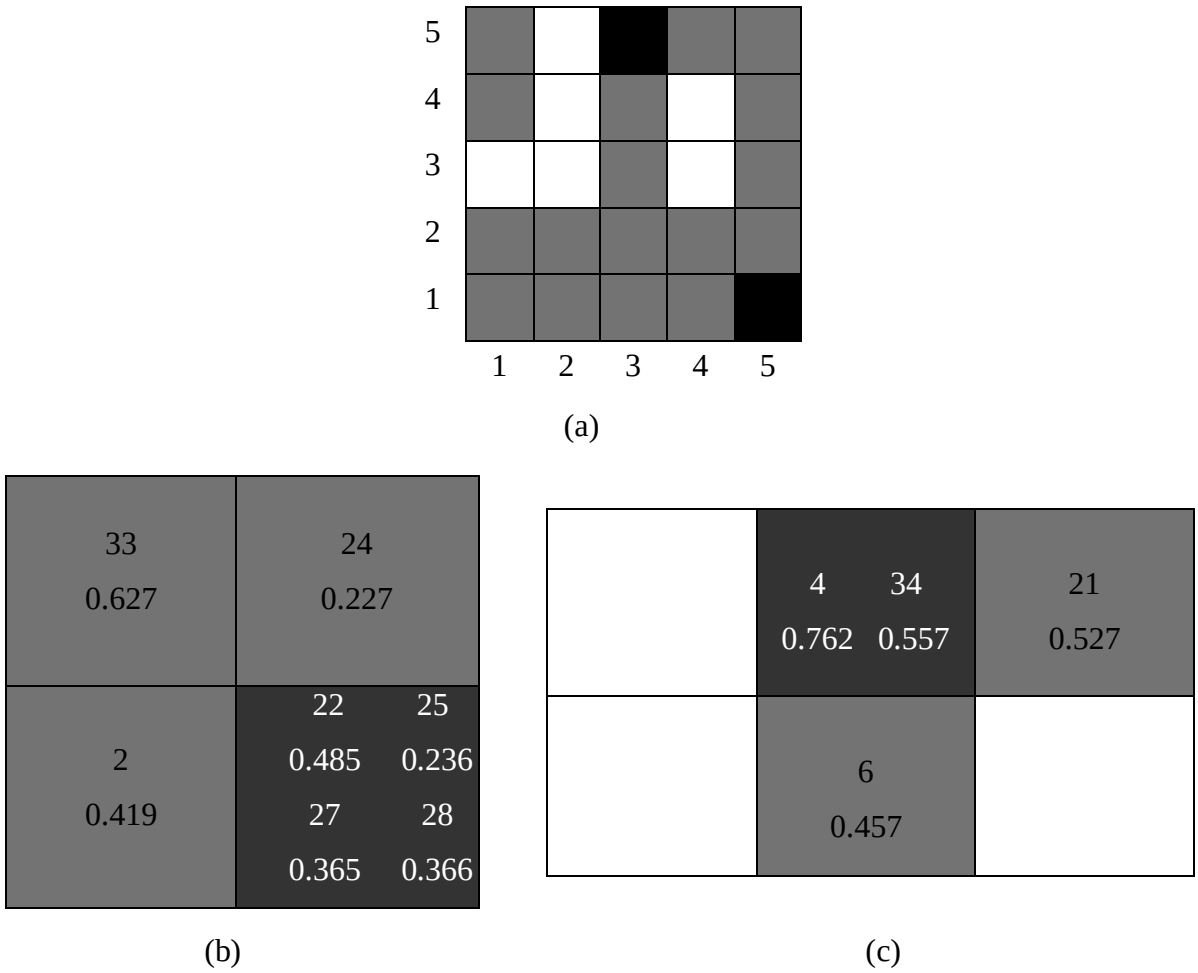


Figure 4.9. Outlier analysis for CPNN model for *Oncorhynchus mykiss* (a) Top map of the training algorithm. Grey cells are occupied neurons; black neurons have the possible outlier(s). Neurons near neuron (5,1) (b) and (3,5) (c). In each cell, normalized toxicity value is under the compound number

Since MLR and CPNN models have the same training and test sets, both models have the same AD. AD of models is defined by the following limits given in Table 4.21.

Pesticides, the CAS numbers, descriptor values, their observed and predicted toxicity values to *Oncorhynchus mykiss*, and residuals obtained from both methods are given in Table 4.22.

Table 4.21. Boundaries of the proposed MLR and CPNN models for *Oncorhynchus mykiss*

	Training set		**Test set**	
	min	**max**	**min**	**max**
pT	-1.62	3.09	-2.10	1.70
RDF080p	0.00	6.92	0.00	6.31
HATS6m	0.00	1.66	0.00	0.38
Mor27u	-0.11	0.39	-0.11	0.45

Studies done with the same species and similar compounds were inspected. The statistical fits of these models are much lower than our models have. Furthermore, they generally lack external validation.

Bermudez-Saldana et al. (2005) studied the toxicity in fish of pesticides. They estimated retention factors of pesticides to use them as descriptors in toxicity prediction models. Aquatic toxicity of pesticides having only one type of MOA (non-polar narcosis) excluding endosulfan was modeled. They demonstrated that retention factors are useful parameters in fish toxicity estimation of non-polar narcosis compounds. However, this finding was not validated by an external test set. Another drawback of their model is the retention factors. This descriptor is obtained by an experiment and not available for all chemicals. Although this descriptor seems to be effective, this parameter has a drawback of being an experimental parameter, and it is preferably obtained from the same lab to reduce the experimental errors. Their model covering only one type of MOA is expected to have high statistical quality. On the other hand, our models with a more diverse data set (including both carbamates and organophosphorous pesticides), are more valuable. The compounds in our data set have two types of MOA, non-polar narcosis and AChE inhibitors.

Table 4.22 The CAS numbers, descriptors, observed and predicted toxicity values obtained from MLR and CPNN models for *Oncorhynchus mykiss*

Training set							pT_{pred}			
							MLR		CPNN	
Comp no	**CAS number**	**Compound name**	**RDF080p**	**HATS6m**	**Mor27u**	**pT_{obs}**	**from Eq 4.6**	**res**	**Proposed model**	**res**
1	10265-92-6	Methamidophos	0.000	0.000	0.199	-1.3979	-1.002	0.395	-1.399	0.000
2	52-68-6	Trichlorfon	0.000	0.222	0.040	0.0758	-0.043	-0.119	0.078	0.000
4	950-37-8	Methidathion	3.164	0.775	0.040	1.8539	1.756	-0.098	1.327	-0.529
5	121-75-5	Malathion	6.920	0.160	0.133	0.9547	0.527	-0.428	0.954	0.000
6	732-11-6	Phosmet	4.206	0.369	0.289	0.2764	0.255	-0.021	0.275	0.000
7	298-00-0	Parathion-methyl	0.788	0.159	0.163	-0.5077	-0.408	0.099	-0.508	0.000
8	29232-93-7	Pirimiphos-methyl	2.725	0.134	0.109	0.1646	-0.059	-0.224	0.166	0.000
10	5598-13-0	Chlorprifos-methyl	0.797	1.095	0.336	0.9208	1.441	0.520	0.923	0.000
11	333-41-5	Diazinon	2.869	0.125	-0.106	0.1972	0.499	0.302	0.068	-0.130
12	55-38-9	Fenthion	0.666	0.153	0.127	0.0757	-0.346	-0.422	0.073	-0.005
14	56-72-4	Coumaphos	3.890	0.147	-0.091	-0.0627	0.656	0.719	0.068	0.130
15	2921-88-2	Chlorpyrifos	1.886	0.675	-0.036	1.8468	1.530	-0.316	1.845	0.000
16	23135-22-0	Oxamyl	1.080	0.080	0.107	-0.6477	-0.416	0.232	-0.648	0.000
17	1646-88-4	Aldoxycarb	0.944	0.091	0.386	-1.6232	-1.135	0.489	-1.622	0.000
18	16752-77-5	Methomyl	0.790	0.070	0.270	-0.1761	-0.905	-0.729	-0.176	0.000
21	17804-35-2	Benomyl	3.850	0.099	0.100	0.6383	0.035	-0.603	0.638	0.000
22	1563-66-2	Carbofuran	0.552	0.070	-0.017	0.4202	-0.191	-0.611	-0.223	-0.643
24	2212-67-1	Molinate	1.429	0.053	0.020	-0.9191	-0.207	0.712	-0.917	0.000
25	1114-71-2	Pebulate	0.440	0.049	0.000	-0.8692	-0.302	0.567	-0.223	0.648
27	99-30-9	Dicloran	0.000	0.064	0.016	-0.2041	-0.369	-0.164	-0.223	-0.021
28	95-50-1	1,2-Dichlorobenzene	0.000	0.000	-0.034	-0.1987	-0.395	-0.197	-0.223	-0.026
30	33213-65-9	Endosulfan	0.000	1.661	0.322	3.0889	2.756	-0.332	3.089	0.000
32	3547-04-4	DDE	2.961	0.601	0.087	1.301	1.178	-0.123	1.327	0.026
33	72-54-8	DDD	1.279	0.438	0.018	1.1549	0.723	-0.432	1.151	-0.005
34	115-32-2	Dicofol	3.651	0.776	0.136	0.7922	1.576	0.784	1.327	0.534

Table 4.22 (continued)

Test set							pT_{pred}			
							MLR		CPNN	
Comp no	**CAS number**	**Compound name**	**RDF080p**	**HATS6m**	**Mor27u**	pT_{obs}	**from Eq 4.6**	**res**	**Proposed model**	**res**
3	60-51-5	Dimethoate	0.811	0.197	0.207	-0.8337	-0.427	0.407	-0.508	0.326
9	2642-71-9	Azinphos-ethyl	6.305	0.182	-0.045	1.6990	0.957	-0.742	0.954	-0.745
13	2310-17-0	Phosalone	2.499	0.379	-0.105	0.8239	1.068	0.244	0.070	-0.754
19	23103-98-2	Pirimicarb	1.555	0.070	0.448	-2.0934	-1.263	0.830	-1.621	0.473
20	114-26-1	Propoxur	0.638	0.113	0.231	-0.9138	-0.720	0.194	-0.178	0.736
23	63-25-2	Carbaryl	2.697	0.056	0.045	-0.1537	-0.089	0.065	0.190	0.344
26	108-90-7	Chlorobenzene	0.000	0.000	0.012	-0.7724	-0.515	0.257	-0.223	0.550
29	106-46-7	1,4-Dichlorobenzene	0.000	0.000	0.000	-0.0492	-0.484	-0.435	-0.223	-0.174
31	120-82-1	1,2,4-Trichlorobenzene	0.000	0.000	-0.035	-0.1847	-0.393	-0.208	-0.223	-0.038

Mazzatorta et al. (2005) modeled 274 chemicals with 319 descriptors employing MLR, PLS, and BPNN. These models cannot comparable to ours because of the number of descriptors. Another model with seven descriptors was developed using GA-CPNN combination. Their training and test set correlation coefficients and *RMSE*s for training and test sets are 0.81, 0.79, 0.68, and 0.73, respectively. Our CPNN model reveals better statistical fit and predictive ability with the corresponding results as 0.945, 0.830, 0.239 and, 0.521.Their data whose details were not elucidated may explain this statistical difference.

Bermudez-Saldana and Cronin (2006) modeled toxicity of 75 pesticides covering organophosphates and carbamates. They started with a set of 75 compounds and remove five of them for several reasons (eg. volatility and degradation possibilities, low water solubility, etc.). MLR modeling of these 70 compounds with five descriptors resulted in four outliers and a poor statistical output. The final model has somewhat acceptable statistical fit; n = 66, R^2_{adj} = 0.71, R^2_{cv} = 0.69, and SE = 0.68 with four descriptors. They, then, divide the data set into two groups according to their MOA: specifically acting compounds and non-specifically acting compounds. MOA based models had relatively better statistical fits. For instance, specifically acting pesticides have the model: n = 49, R^2_{adj} = 0.73, R^2_{cv} = 0.69 and SE = 0.70 with three descriptors. In their study, a test set for external validation to measure predictive ability is not available.

Slavov et al. (2008) modeled 125 aromatic compounds obtained form ECOTOX database of EPA. The results include 2D and 3D QSAR analyses. They obtained a multilinear QSAR equation using 96 compounds out of 125. The forward stepwise multilinear regression resulted in a 3-descriptor model which has a low correlation coefficient. Moreover, scrutiny of predictive ability of model was not performed with a test set. They concluded that the electrostatic interactions are of much lesser importance for the aquatic toxicity than the steric interactions. Due to the moderate quality of the 2D-QSAR model, they split the data set into training and test sets to apply CoMFA (Comparative molecular field analysis) for 3D analysis. The PLS analysis for the steric interactions resulted in much higher statistical significance. They again concluded that

steric interactions play more important role than the electrostatic one for the aquatic toxicity. They emphasize that the CoMFA produces highly predictive models. Nevertheless, drawback of their model is that it is good for only data sets of similar compounds. Our data set has both aromatic and aliphatic compounds; therefore, it is more diverse than their data set.

5. CONCLUSIONS

The quality of prediction is sensitive to the dependence of various parameters such as response variable being considered on structural specificity/ diversity present in the data set, data set size, type of descriptors, statistical techniques used, and the range of response values. Therefore, in this thesis, all these parameters were kept constant in modeling toxicity of freshwater algae *Chlorella vulgaris* and *Oncorhynchus mykiss* except the statistical techniques used.

The contribution of this study is to obtain a very large number of descriptors with only knowledge of the three- dimensional structure of chemicals available through various specialized software packages to enable an efficient variable selection procedure like heuristic; and a training-test set splitting methodology like Kohonen networks and to compare the outputs of linear and nonlinear modeling applied to these data sets. MLR (linear) and CPNN (nonlinear) were used to model the structure and toxicity relationships. Both methods with three descriptors resulted in useful MLR and CPNN models with good generalization and prediction ability as it was measured by cross-validation and application of it to predict the toxicity of compounds in the test set. These models have a conclusive mechanistic interpretation since the descriptors used in the model are considered as relevant to toxicity. Their mechanistic meanings reflect the same mechanisms as stated in the literature. The proposed models have been proved to fulfill the fundamental points set down by OECD principles for regulatory QSAR acceptability. These models could be used to predict reliable toxicities for only those compounds with unknown toxicity belonging to the AD of the models.

Data sets which are more homogeneous are expected to have better statistical fit. Although *Chlorella vulgaris* data set is more diverse, *Chlorella vulgaris* models have much better statistical fits than *Oncorhynchus mykiss* models. This might be probably due to the source of the data set. Chemicals of *Chlorella vulgaris* models came from the same laboratory, yet *Oncorhynchus mykiss* data set was compiled from ECOTOX. As apprised

in Section 2, bioassay results from different labs may have different test conditions and should be undertaken attentively.

Proposed models for *Chlorella vulgaris* have various descriptors from the used software packages. However, in *Oncorhynchus mykiss* models, Dragon descriptors suppressed the other descriptors. The *CRI* is a significant descriptor in *Chlorella vulgaris* models, whereas it did not appear in *Oncorhynchus mykiss* models.

Phenylazophenol was an outlier in the proposed MLR model for *Chlorella vulgaris*. The reason might be its unique structure in the data set. Methidathion and piperine with a high leverage value in the training set can be classified as influential chemicals in determination of the model descriptors, but they are accepted as good leverages since they are better predicted with small prediction residues. Note that, methidathion is a pesticide. Additionally, malathion and phosmet are high leverage compounds of the test set again are pesticides. However, none of the compounds in the external validation data had leverage values far from methidathione. Predicted values for these compounds are not obtained due to over extrapolation, but also their absolute residuals were small.

Methidathion appeared to be a response outlier in both CPNN models of *Chlorella vulgaris* and *Oncorhynchus mykiss* data sets. Therefore, care should be taken while using CPNN models for the prediction of toxicity of compounds having a structural similarity to methidathion. Endosulfan is the high leverage compound of *Oncorhynchus mykiss* training set. Although endosulfan has a much higher toxicity than other chemicals in the organochlorines subset, both of our models are able to make precise estimations. Additionally, endosulfan had higher leverage (h = 0.51) and yet closer to h^* of 0.48, therefore, it may not be considered structurally the most influential in determination of the model descriptors.

We investigated the linear and nonlinear models of *Chlorella vulgaris* and *Oncorhynchus mykiss* data sets. Training sets of CPNN models generally have higher correlation coefficients and lower *RMSE*s than MLR models. On the other hand, MLR models have higher cross-validation values. It can be concluded that correlation

coefficients and cross-correlation coefficients are not proportional in both methods. Additionally, test sets of MLR models generally have higher correlation coefficients. This may be interpreted as MLR models have higher predictive abilities. If one considers models in the outlier aspect, MLR and CPNN models have somewhat similar outlooks. However, unlike MLR models, removing outliers did not improve the statistical quality of the CPNN models.

Nonlinear modeling methods such as CPNN are effective in complex and diverse data sets with high number of compounds which are from various chemical classes and different MOAs. The *Chlorella vulgaris* data set is relatively more diverse data set than fish data set. In this aspect we can say that CPNN successfully covered phenylazophenol compound which is an outlier in MLR model.

REFERENCES

Acero, J. L., Benítez, F. J., Real, F. J., González, M., 2008. Chlorination of organophosphorus pesticides in natural waters. Journal of Hazardous Materials, 153, 320-328.

Aptula, A. O., Jeliazkova, N. G., Schultz, T W., Cronin, M. T. D., 2005. The better predictive model: High q^2 for the training set or low root mean square error of prediction for the test set? QSAR and Combinatorial Science, 24, 385-396.

APVMA Australian Pesticides and Veterinary Medicines Authority http://www.apvma.gov.au/news_media/media_releases/2010/mr2010-12.php accessed December, 2010.

Atasoy, A. D., Mermut, A. R., Kumbur, H., İnce, F., Arslan, H., Avcı, E. D., 2009. Sorption of alpha and beta hydrophobic endosulfan in a Vertisol from southeast region of Turkey. Chemosphere, 74, 1450-1456.

Barnhoorn, I.E.J., Bornman, M.S., Jansen van Rensburg, C., Bouwman, H., 2009. DDT residues in water, sediment, domestic and indigenous biota from a currently DDT-sprayed area. Chemosphere, 77, 1236-1241.

Berenzen, N., Lentzen-Godding, A, Probst, M., Schultz, H., Schultz, R., Liess M., 2005. A comparison of predicted and measured levels of runoff-related pesticide concentrations in small lowland streams on a landscape level. Chemosphere, 58, 683-691.

Bermudez-Saldana, J. M., Escuder-Gilabert, L., Medina-Hernandez, M. J., Villanueva-Camanas, R. M., Sagrado, S., 2005. Chromatographic evaluation of the toxicity in fish of pesticides. Journal of Chromatography B, 814, 115-125.

Bermudez-Saldana, J. M., Cronin, M. T. D., 2006. Quantitative structure-activity relationships for the toxicity of organophosphorus and carbamate pesticides to the rainbow trout *Onchorhyncus mykiss*. Pest Management Science, 62, 819-831.

Caballero, J., Fernandez, M., 2006. Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. Journal of Molecular Modeling, 12, 168-181.

Capkin, E., Altinok I., Karahan, S., 2006. Water quality and fish size affect toxicity of endosulfan, an organochlorine pesticide, to rainbow trout. Chemosphere, 64, 1793–1800.

Centofanti, T., Hollis, J.M., Blenkinsop, S., Fowler, H.J., Truckell, I., Dubus, I.G., Reichenberger, S., 2008. Development of agro-environmental scenarios to support pesticide risk assessment in Europe. Science of the Total Environment, 407, 574-588.

Charton, M., 2008. Philip S. Magee: a life in QSAR. Journal of Computer Aided Molecular Design, 22, 335-337.

Chi, J., 2009. Vertical fluxes and accumulation of organochlorine pesticides in sediments of Haihe River, Tianjin, China. Bulletin of Environmental Contamination and Toxicology, 82, 510-515.

CODESSA 2.20 (1994-1996), Semichem Inc., Shawnee Mission, USA.

Consonni, V., Todeschini,R., Pavan, M., Gramatica, P., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. The Journal of Chemical Information and Computer Sciences, 42, 693-705.

Consonni V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q^2 parameter for QSAR validation. The Journal of Chemical Information and Modeling, 49, 1669–1678.

Corrigan, R., Owens, J., Bennett G., 1997. Truman's Scientific Guide to Pest Control Operations, Fifth Edition, Advanstar Communications, New York.

CPNN software package for counter propagation modeling, kctrf.f ©, National Institute of Chemistry, Ljubljana, Slovenia.

Cronin, M. T. D., Netzeva, T. I., Dearden, J. C., Edwards, R., Worgan, A. D. P., 2004. Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database. Chemical Research in Toxicology, 17, 545–554.

Crosby, D. G., 1998. Environmental Toxicology and Chemistry, Oxford University Press, New York, NY.

Dag, S., Aykac, V., Gunduz A., Kantarci M., and Sisman N., 2000. Pesticide industry in Turkey and its future, V. Technique Congress of Turkish Agriculture Engineering, Ankara, Vol: 2.

Devillers, J. (Ed.), 1996. Neural Networks in QSAR and Drug Design, Academic Press Limited, London, UK.

Dörr, B. and Liebezeit, G., 2009. Organochlorine compounds in blue mussels, mytilus edulis, and pacific oysters, crassostrea gigas, from seven sites in the lower saxonian Wadden Sea, Southern North Sea. Bulletin of Environmental Contamination and Toxicology, 83, 874-879.

Egan, W. J., Morgan, S. L., 1998. Outlier detection in multivariate analytical chemical data. Analytical Chemistry, 70, 2372-2379.

EPA (U. S. Environmental Protection Agency) Types of Pesticides
http://www.epa.gov/opp00001/about/types.htm#type (accessed November 2010)

Gini, G., Craciun, M. V., König, C., Benfenati, E., 2004. Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. Journal of Information and Computer Sciences, 44, 1897-1902.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y., Lee K., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. Journal of Computer-Aided Molecular Design, 17, 241–253.

Goldsborough, L. G., Crumpton, W. G., 1998. Distribution and environmental fate of pesticides in prairie wetlands. Great Plains Research, 8, 73-95.

Gramatica, P., Consonni, V., Pavan, M., 2003. Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. SAR and QSAR in Environmental Research, 14, 237-250.

Gramatica, P., 2007. Principles of QSAR models validation: Internal and external. QSAR and Combinatorial Science, 26, 694–701.

Hasegawaa, K., Funatsub, K., 1998. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. Journal of Molecular Structure (Theochem), 425, 255-262.

Hrovat, M., Segner, H., Jeram, S., 2009. Variability of *in vivo* fish acute toxicity data. Regulatory Toxicology and Pharmacology, 54, 294–300.

Kahn, I., Sild, S., Maran, U., 2007. Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using heuristic Multilinear Regression and heuristic back-propagation neural networks. The Journal of Chemical Information and Modeling, 47, 2271-2279.

Kaiser, K. L. E., 2003. The use of neural networks in QSARs for acute aquatic toxicological endpoints. Journal of Molecular Structure, 622, 85-95.

Katritzky, R., Lobanov, V. S., Karelson, M., 1994. CODESSA: Comprehensive Descriptors for Structural and Statistical Analysis, version 2.2.1. Reference Manual. University of Florida, Gainesville, Florida, U.S.A.

Kiralj, R., Ferreira, M. M. C., 2009. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application. Journal of the Brazilian Chemical Society, 20, 770-787.

Knauer, K., Lampert, C., Gonzales-Valero, J., 2007. Comparison of *in vitro* and *in vivo* acute fish toxicity in relation to toxicant mode of action, Chemosphere, 68, 1435-1441.

Laabs, V., Amelung, W., Pinto, A. A., Wantzen, M., da Silva, C. J., Zech, W., 2002. Pesticides in surface water, sediment, and rainfall of the northeastern Pantanal Basin, Brazil. Journal of Environmental Quality, 31, 1636-1648.

LoPachin, R. M., Gavin, T., Geohagen, B. C., Das, S., 2007. Neurotoxic mechanisms of electrophilic type-2 alkenes: Soft-soft interactions described by quantum mechanical parameters. Toxicological Sciences, 98, 561-570.

Lu, C., Rodriguez T., Funez, A., Irish, R.S., Fenske, R.A., 2006. The assessment of occupational exposure to diazinon in Nicaraguan plantation workers using saliva biomonitoring. Annals of the New York Academy of Sciences: 1076, 355-365.

Matlab Student Version 6.0, Copyright © 2000 by the MathWorks.

Mazzatorta P., Vracko M., Jezierska A., Benfenati E., 2003. Modeling toxicity by using supervised Kohonen neural networks. Journal of Chemical Information and Computer Sciences, 43, 485-492.

Mazzatorta, P., Smiesko, M., Lo Piparo, E., Benfenati, E., 2005. QSAR Model for predicting pesticide aquatic toxicity. Journal of Chemical Information and Modeling, 45, 1767-1774.

Miller, G. T., 2004. Sustaining the Earth, 6th edition. Thompson Learning, Inc. Pacific Grove, California.

Molegro Data Modeller (MDM 2010.2) 2007-2010, Molegro ApS.

Netzeva, T. I., Dearden, J. C., Edwards, R., Worgan, A. D. P., Cronin, M. T. D., 2004. QSAR Analysis of the toxicity of aromatic compounds to *Chlorella* v*ulgaris* in a novel short-term assay. Journal of Chemical Information and Computer Sciences, 44, 258-265.

OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. ENV/JM/MONO (2007)2.

OME Ontario Ministry of the Environment
http://www.ene.gov.on.ca/en/land/pesticides/index.php accessed December, 2010.

Ozcan, S., Aydin, M. E., 2009. Polycyclic aromatic hydrocarbons, polychlorinated biphenyls and organochlorine pesticides in urban air of Konya, Turkey. Atmospheric Research, 93, 715-722.

Papa, E., Villa, F., Gramatica, P., 2005. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). Journal of Chemical Information and Modeling, 45, 1256-1266.

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

Pavan, M., Netzeva, T. I., Worth, A. P., 2006. Validation of a QSAR model for acute toxicity. SAR and QSAR in Environmental Research, 17, 147–171.

Pontolillo J., Eganhouse R. P., 2001. The search for reliable aqueous solubility (*Sw*) and octanol-water partition coefficient (*Kow*) Data for hydrophobic organic compounds: DDT and DDE as a case study. U.S. Geological Survey, Water-Resources Investigations Report 01-4201.

Porcelli, C., Roncaglioni, A., Chana, A., Benfenati, E., 2008. A comparison of DEMETRA individual QSARs with an index for evaluation of uncertainty. Chemosphere, 71, 1845-1852.

Roy, D. R., Parthasarathi, R., Maiti, B., Subramanian, V., Chattaraj, P. K., 2005. Electrophilicity as a possible descriptors for toxicity prediction. Bioorganic & Medicinal Chemistry Letters, 13, 3405 – 3412.

Roy K., Gosh, G., 2006. QSTR with extended topochemical atom (ETA) indices. 8.a QSAR for the inhibition of substituted phenols on germination rate of *Cucumis sativus* using chemometric tools. QSAR and Combinatorial Sciences, 25, 846 – 859.

Roy, K., Ghosh, G., 2007. QSTR with extended topochemical atom (ETA) indices. 9. Comparative QSAR for the toxicity of diverse functional organic compounds to *Chlorella vulgaris* using chemometric tools. Chemosphere, 70, 1-12.

Roy, K., Roy, P. P., 2009. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. European Journal of Medicinal Chemistry, 44, 2913–2922.

Saçan, M.T., Inel, Y., 1993. Prediction of aqueous solubility of PCBs related to molecular structure. Turkish Journal of Chemistry**,** 17,188-195.

Saçan, M.T., Inel, Y., 1995. Application of the Characteristic Root Index model to the estimation of n-Octanol water partition coefficients. Polychlorinated biphenyls. Chemosphere, 30, 39-50.

Saçan, M.T., Balcioglu, I. A., 1996. Prediction of soil sorption coefficient of organic pollutants by the Characteristic Root Index model. Chemosphere**,** 32, 1993-2001.

Saçan, M.T., Balcioglu, I. A., 1998. Estimation of liquid vapor pressures for low-volatility environmental chemicals. Chemosphere, 36, 451-460.

Saçan, M. T., Erdem, S. S., Özpınar, G. A., Balcioglu, I. A., 2004. QSPR Study on the bioconcentration factors of nonionic organic compounds in fish by Characteristic Root Index and semiempirical molecular descriptors. Journal of Chemical Information and Computer Sciences**,** 44, 985-992.

Saçan, M. T., Ozkul, M., Erdem., S. S., 2005. Physico-chemical properties of PCDD/PCDFs and phthalate esters. SAR and QSAR in Environmental Research, 16, 443-459.

Saçan, M. T., Özkul, M., Erdem, S. S., 2007. QSPR analysis of the toxicity of aromatic compounds to the algae (*Scenedesmus obliquus*). Chemosphere, 68, 695-702.

Sibali, L.L., Okonkwo, J.O., Zvinowanda, C., 2009. Determination of DDT and metabolites in surface water and sediment using LLE, SPE, ACE and SE. Bulletin of Environmental Contamination and Toxicology, 83, 885-891.

Slavov, S., Gini, G., Benfenati, E., 2008. QSAR trout toxicity models on aromatic pesticides. Journal of Environmental Science and Health Part B, 43, 633–637.

SPARTAN 06 Copyright © 2006 by Wavefunction, Inc., Irvine, USA

SPSS 17.0 for Windows (SPSS® 2008) Statistical Package for Social Scientists by SPSS, Inc.

Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.4 – 2006 – http://www.talete.mi.it/

Tao, S., Xi, X., Xu, F., Li, B., Cao, J., Dawson, R., 2002. A fragment constant QSAR model for evaluating the EC_{50} values of organic chemicals to *Daphnia magna*. Environmental Pollution, 116, 57-64.

Todeschini, R. and Consonni, V., Handbook of Molecular Descriptors, WILEY-VCH, Germany, 2000.

Vijver, M. G., Van't Zelfde, M., Tamis, W. L. M., Musters, K. J. M., De Snoo, G. R., 2008. Spatial and temporal analysis of pesticides concentrations in surface water: Pesticides atlas. Journal of Environmental Science and Health, 43, 665 - 674.

Vracko, M., 2005. Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. Current Computer-Aided Drug Design, 1, 73-78.

Vracko, M., Bandelj, V., Barbieri, P., Benfenati, E., Chaudhry, Q., Cronin, M., Devillers, J., Gallegos, A., Gini, G., Gramatica, P., Helma, C., Mazzatorta, P., Neagu, D., Netzeva, T., Pavan, M., Patlewicz, G., Randic, M., Tsakovska, I., Worth, A., 2006. Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR and QSAR in Environmental Research, 17, 265–284.

Walker, C. H., Hopkin, S. P., Sibly, R. M., Peakall, D. B., 2006. Principles of Ecotoxicology, CRC Press Taylor & Francis Group, USA.

Walker, J. D., 2003. Applications of QSARs in toxicology: a US Government perspective. Journal of Molecular Structure, 622, 167-184.

Wang, X., Dong, Y., Xu, S., Wang, L., Han, S., 2000. Quantitative structure-activity relationships for the toxicity to the tadpole *Rana japonica* of selected phenols. Bulletin of Environmental Contamination and Toxicology, 64, 859-865.

WEKA 3.6.1: Waikato Environment for Knowledge Analysis © 1999-2009 http://www.cs.waikato.ac.nz/~ml/weka/, March 2010.

Wright, D. A. and Welbourn, P., 2002. Environmental Toxicology, Cambridge University Press, UK.

Yu, M-H, 2001. Environmental Toxicology, Lewis Publishers, USA.

Zhang, G., Patuwo, B. E., Hu, M. Y., 1998. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14, 35-62.

Zhang, H., Chen, Q-Y., Xiang, M-L., Maa, C-Y., Huang, Q., Yang, S-Y., 2009. *In silico* prediction of mitochondrial toxicity by using GA-CG-SVM approach. Toxicology in Vitro, 23, 134–140.

Zupan, J. and Gasteiger, J., 1999. Neural Networks in Chemistry and Drug Design, WILEY-VCH, Germany.

Zvinavashe E., Du T., Griff T., van den Berg H. H. J., Soffers A. E. M. F, Vervoort J., Murk A. J, Rietjens I. M. C. M., 2009. Quantitative structure-activity relationship modeling of the toxicity of organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. Chemosphere, 75, 1531-1538

Appendix A Significant descriptors of *Chlorella vulgaris* data set

Comp No	HOMT	Max Partial Charge for a C Atom	*CRI*	ATS1m	R4e	EEig05r	piPC07	*E*	ω	E_{HOMO}	*Z*
1	0.000	0.000	0.447	0.847	0.000	0.000	0.000	-2.250	0.993	-11.138	0.666
2	0.000	0.000	0.924	1.204	0.375	0.000	0.000	-2.547	1.041	-11.127	0.724
3	0.000	0.000	1.623	1.674	2.402	0.000	0.000	-3.091	1.107	-11.277	0.795
4	0.000	0.000	1.583	1.674	1.506	0.000	0.000	-3.058	1.122	-11.231	0.793
5	-0.875	0.082	1.965	2.335	1.323	0.000	0.000	-5.283	2.559	-10.738	0.837
6	-0.353	0.081	1.583	2.233	0.819	-0.052	0.000	-4.933	2.985	-10.795	0.826
7	-0.411	0.080	1.136	1.945	0.911	-0.734	0.000	-2.922	2.830	-11.067	0.790
8	0.000	0.000	1.580	1.674	1.245	0.000	0.000	-3.020	1.058	-11.011	0.807
9	0.000	0.033	1.403	1.674	1.060	0.000	0.000	-2.510	2.115	-10.654	0.789
10	-0.984	0.080	1.504	2.079	1.684	-0.301	0.000	-3.266	2.529	-10.754	0.829
11	0.000	0.034	1.816	1.846	1.198	-0.618	0.000	-2.723	2.084	-10.534	0.828
12	-0.290	0.037	1.161	1.674	0.756	0.000	0.000	-1.255	2.757	-10.485	0.785
13	-0.290	0.038	1.955	1.990	1.163	-0.215	0.000	-1.704	2.756	-10.490	0.819
14	-0.317	0.038	1.585	1.860	0.869	-0.717	0.000	-1.469	2.756	-10.500	0.805
15	5.897	0.026	1.353	2.200	0.705	0.000	0.000	-0.940	2.085	-9.175	0.858
16	-1.059	0.081	1.990	2.302	1.631	0.000	0.000	-2.389	2.295	-10.379	0.831
17	5.894	0.000	1.404	2.100	1.130	0.000	0.000	0.924	1.858	-8.609	0.853
18	0.000	0.033	2.548	2.120	1.415	0.098	0.000	-3.225	2.101	-10.625	0.848
19	5.894	0.024	1.607	2.268	0.747	0.185	2.784	-0.628	2.044	-9.140	0.876
20	5.832	0.061	1.369	2.294	0.772	0.375	0.000	-2.837	2.398	-9.432	0.856
21	5.778	0.063	1.447	2.277	1.288	0.222	2.151	-0.945	2.130	-8.776	0.853
22	5.879	0.026	1.760	2.233	1.241	0.432	0.000	-1.341	2.004	-8.951	0.866
23	5.828	0.029	1.719	2.398	0.895	0.587	2.784	-2.558	2.017	-8.710	0.868
24	5.749	0.032	1.541	2.251	1.341	0.233	2.151	-0.948	1.868	-8.382	0.853
25	5.797	0.039	1.764	2.398	1.126	0.578	3.169	-2.521	2.181	-9.140	0.866
26	5.732	0.000	2.176	2.319	1.439	0.653	2.784	0.230	1.817	-8.478	0.884
27	5.106	0.040	1.546	2.233	0.822	0.280	3.446	-0.462	2.900	-10.050	0.871
28	5.828	0.027	1.755	2.233	1.336	0.243	2.151	-1.314	2.066	-9.062	0.874

Appendix A (continued)

Comp No	HOMT	Max Partial Charge for a C Atom	*CRI*	ATS1m	R4e	EEig05r	piPC07	*E*	ω	E_{HOMO}	*Z*
29	5.534	0.044	1.651	2.367	0.972	0.666	3.663	-2.539	2.879	-9.351	0.884
30	5.921	0.040	1.527	2.423	0.905	0.254	4.123	0.630	3.637	-10.602	0.855
31	0.000	0.091	2.758	3.718	1.445	1.518	3.761	-4.672	5.172	-9.552	0.861
32	5.872	0.026	1.734	2.233	1.302	0.000	0.000	-1.341	2.004	-8.951	0.866
33	5.838	0.026	2.111	2.335	1.396	0.450	2.151	-1.712	1.999	-8.894	0.880
34	5.143	0.040	1.902	2.335	1.393	0.733	3.446	-0.877	2.818	-9.756	0.877
35	5.776	0.039	2.231	2.485	0.837	0.920	3.524	-2.786	2.093	-9.012	0.886
36	4.896	0.040	1.798	2.442	0.869	1.368	4.627	1.091	3.454	-10.348	0.880
37	5.777	0.040	1.950	2.508	1.308	0.596	4.407	0.220	3.510	-10.275	0.863
38	5.930	0.035	1.805	2.423	0.721	0.000	0.000	-1.231	2.216	-9.009	0.861
39	5.620	0.051	1.874	2.816	1.654	2.000	5.139	0.204	4.013	-9.867	0.859
40	5.931	0.028	2.325	2.707	0.703	0.000	0.000	-0.613	2.348	-9.312	0.865
41	5.889	0.020	2.398	2.696	1.173	0.000	0.000	1.245	2.150	-8.781	0.860
42	5.893	0.034	1.889	2.409	1.194	0.318	0.000	0.625	2.102	-8.760	0.856
43	4.908	0.081	2.748	2.708	1.589	2.000	4.095	-2.028	2.323	-9.742	0.885
44	5.815	0.048	1.908	2.816	1.493	2.000	5.316	0.314	4.306	-9.872	0.845
45	4.933	0.045	2.013	2.508	0.903	0.567	3.663	-0.693	2.999	-9.706	0.881
46	5.843	0.027	2.682	2.939	0.681	0.000	0.000	-0.016	2.536	-8.840	0.874
47	5.147	0.040	2.380	2.428	1.367	1.372	4.095	-1.076	2.838	-9.830	0.882
48	1.036	0.047	2.527	2.751	1.370	2.040	5.507	-1.373	3.927	-10.255	0.929
49	5.805	0.027	2.532	2.428	1.607	0.692	3.663	-1.722	2.049	-9.046	0.908
50	5.872	0.040	2.371	2.644	1.178	0.473	0.000	0.339	2.321	-8.910	0.859
51	5.847	0.040	2.707	2.658	1.468	0.990	4.343	-0.405	3.045	-10.040	0.872
52	5.719	0.044	2.406	2.644	1.356	0.473	2.784	0.361	2.267	-8.738	0.872
53	5.747	0.027	2.868	2.512	2.228	0.713	3.992	-1.838	2.022	-9.012	0.930
54	5.864	0.048	2.187	2.557	0.574	0.160	2.151	0.483	2.453	-9.294	0.871
55	5.733	0.045	1.772	2.744	1.109	2.000	5.094	0.400	4.737	-11.470	0.851
56	5.732	0.059	1.809	2.826	1.241	2.000	5.139	-1.523	4.387	-10.948	0.841

Appendix A (continued)

Comp No	HOMT	Max Partial Charge for a C Atom	*CRI*	ATS1m	R4e	EEig05r	piPC07	*E*	ω	E_{HOMO}	*Z*
57	5.706	0.043	1.696	2.744	1.205	2.000	5.440	0.440	5.078	-11.304	0.850
58	4.829	0.042	1.818	2.611	1.032	1.542	4.808	-0.807	4.009	-10.869	0.864
59	6.194	0.069	3.154	3.717	1.664	2.389	5.514	-5.734	4.860	-9.360	0.889
60	5.846	0.042	2.701	3.480	1.586	2.000	5.220	-6.097	4.753	-9.809	0.870
61	0.000	0.085	4.121	3.688	1.095	1.956	3.219	-11.573	5.371	-9.706	0.839
62	5.692	0.051	2.578	2.909	1.578	1.157	4.766	-0.014	3.540	-9.416	0.865
63	7.487	0.076	3.391	3.726	1.600	2.433	5.586	-2.160	5.073	-9.515	0.887
64	5.642	0.046	2.084	2.806	1.597	2.000	5.139	0.088	4.526	-11.190	0.859
65	5.863	0.032	2.672	3.413	1.247	1.553	5.094	-4.209	4.379	-9.568	0.893
66	5.181	0.056	2.407	2.979	1.856	2.203	5.388	0.000	4.150	-9.784	0.867
67	5.751	0.053	2.084	3.255	1.037	1.099	4.766	-0.517	3.851	-10.288	0.872
68	0.000	0.000	2.352	3.536	1.208	1.220	1.946	-4.732	5.026	-9.378	0.848
69	5.879	0.060	2.448	2.844	1.445	0.826	4.495	0.259	3.680	-9.778	0.866
70	5.640	0.056	4.597	3.264	1.672	2.480	5.715	-2.692	3.046	-8.769	0.939
71	5.132	0.040	3.030	2.590	2.304	1.603	4.766	-1.424	2.814	-9.813	0.915
72	5.823	0.049	2.939	2.834	1.357	0.473	2.784	0.087	2.415	-8.719	0.873
73	5.858	0.045	2.368	2.723	1.531	0.981	4.577	0.140	3.355	-9.944	0.873
74	4.867	0.056	1.987	2.979	2.052	2.000	5.388	0.055	4.245	-9.464	0.874
75	5.143	0.057	1.811	2.744	1.148	2.000	4.868	1.076	4.722	-11.323	0.872
76	5.814	0.043	1.972	2.655	0.919	0.598	4.123	0.366	3.780	-10.221	0.858
77	5.815	0.054	3.170	3.567	1.587	2.235	5.360	-6.260	4.791	-9.820	0.886
78	5.557	0.029	4.727	2.853	2.290	2.285	5.124	-2.945	1.897	-8.699	0.955
79	5.800	0.063	3.163	2.994	1.364	1.055	3.446	-0.112	2.671	-8.911	0.884
80	5.460	0.063	2.548	2.918	1.083	1.110	4.766	-1.867	3.821	-9.545	0.872
81	4.809	0.050	2.447	2.724	0.988	0.629	3.841	-0.845	3.082	-9.523	0.893
82	5.731	0.032	3.219	3.542	1.279	1.532	4.915	-6.026	3.961	-8.689	0.901
83	5.747	0.027	4.291	2.793	2.412	2.152	4.868	-2.838	1.942	-8.844	0.951
84	5.556	0.078	3.699	3.139	0.869	1.220	3.841	-2.094	2.951	-9.136	0.891

Appendix A (continued)

Comp No	HOMT	Max Partial Charge for a C Atom	*CRI*	ATS1m	R4e	EEig05r	piPC07	*E*	ω	E_{HOMO}	*Z*
85	5.718	0.043	3.099	3.510	1.466	2.145	5.309	-6.401	4.603	-9.775	0.888
86	5.855	0.063	2.863	3.002	1.464	1.072	4.577	0.043	3.882	-9.793	0.872
87	5.416	0.079	3.192	3.195	1.378	2.444	5.265	0.304	4.783	-10.299	0.897
88	11.612	0.028	2.840	2.892	0.613	2.000	5.469	1.962	2.918	-8.974	0.905
89	5.199	0.040	4.719	2.936	1.586	2.119	4.990	-2.086	2.376	-8.572	0.933
90	5.033	0.081	3.481	3.139	1.077	1.220	4.888	0.020	3.922	-9.524	0.893
91	5.356	0.059	6.391	3.728	0.758	1.161	3.841	1.066	3.421	-9.650	0.911

Appendix B Significant descriptors of *Oncorhynchus mykiss* data set

Comp No	*CRI*	Volume	Area	Dipole moment	E_{LUMO}	E_{HOMO}	*E*	$E_{LUMO} - E_{HOMO}$	Hardness	Electro-negativity	*Z*	Softness	ω
1	1.213	122.580	161.359	3.505	-1.355	-9.493	-5.485	8.138	4.069	5.424	0.760	0.246	3.615
2	3.184	187.317	227.829	0.389	-0.327	-10.535	-9.898	10.208	5.104	5.431	0.822	0.196	2.890
3	2.270	202.544	244.823	2.839	-2.568	-9.647	-5.898	7.078	3.539	6.107	0.827	0.283	5.270
4	2.758	246.187	285.929	6.079	-2.514	-9.552	-4.672	7.038	3.519	6.033	0.861	0.284	5.172
5	4.121	303.986	362.256	3.758	-2.627	-9.706	-11.573	7.079	3.539	6.166	0.839	0.283	5.371
6	3.154	279.444	314.506	7.914	-2.330	-9.360	-5.734	7.030	3.515	5.845	0.889	0.284	4.859
7	2.701	226.118	256.389	7.351	-2.166	-9.823	-6.130	7.657	3.829	5.994	0.882	0.261	4.693
8	4.489	298.380	337.107	3.582	-1.650	-8.793	-6.438	7.142	3.571	5.222	0.885	0.280	3.817
9	4.388	312.271	348.293	7.895	-2.350	-9.382	-2.342	7.031	3.516	5.866	0.897	0.284	4.894
10	3.696	238.924	278.764	4.174	-1.802	-9.144	-6.180	7.341	3.671	5.473	0.857	0.272	4.080
11	4.797	302.433	340.628	4.400	-1.677	-9.391	-6.873	7.714	3.857	5.534	0.888	0.259	3.970
12	3.219	259.280	287.912	4.576	-1.776	-8.689	-6.026	6.913	3.457	5.233	0.901	0.289	3.961
13	3.822	317.114	353.795	3.471	-2.471	-9.314	-6.502	6.843	3.421	5.892	0.896	0.292	5.074
14	5.142	316.629	346.214	1.978	-2.027	-9.169	-9.224	7.142	3.571	5.598	0.915	0.280	4.387
15	4.807	274.675	310.546	5.738	-1.749	-9.132	-6.634	7.383	3.691	5.441	0.884	0.271	4.010
16	2.812	212.141	253.129	3.217	-0.602	-9.384	-2.487	8.781	4.391	4.993	0.838	0.228	2.839
17	3.290	209.604	245.084	1.903	-0.514	-10.094	-4.239	9.580	4.790	5.304	0.855	0.209	2.937
18	1.988	159.155	196.611	2.633	-0.214	-9.160	-1.407	8.946	4.473	4.687	0.809	0.224	2.455
19	3.950	255.106	289.765	1.401	-0.256	-8.701	-2.738	8.445	4.223	4.479	0.880	0.237	2.375
20	3.429	223.767	250.344	1.611	0.117	-9.259	-4.307	9.377	4.688	4.571	0.894	0.213	2.228
21	4.534	296.416	327.113	2.906	-0.434	-9.099	-3.757	8.664	4.332	4.767	0.906	0.231	2.622
22	3.571	231.482	258.494	2.385	0.094	-9.043	-4.361	9.138	4.569	4.474	0.896	0.219	2.191
23	2.987	211.520	229.702	4.615	-0.839	-9.093	-1.427	8.253	4.127	4.966	0.921	0.242	2.988
24	3.318	202.799	226.709	1.565	-0.344	-9.327	-2.349	8.983	4.491	4.835	0.895	0.223	2.603
25	4.043	234.820	269.143	1.353	-0.362	-9.450	-3.122	9.088	4.544	4.906	0.872	0.220	2.648
26	1.343	112.322	129.989	0.954	0.063	-9.389	0.723	9.452	4.726	4.663	0.864	0.212	2.300
27	2.574	157.556	182.067	6.025	-1.302	-9.416	-0.014	8.114	4.057	5.359	0.865	0.246	3.540
28	2.187	125.269	143.752	1.351	-0.168	-9.294	0.483	9.126	4.563	4.731	0.871	0.219	2.453
29	2.191	125.472	144.876	0.000	-0.242	-9.237	0.438	8.994	4.497	4.739	0.866	0.222	2.497
30	5.433	261.886	272.522	7.021	-0.421	-9.404	-4.087	8.983	4.491	4.912	0.961	0.223	2.686

Appendix B (continued)

Comp No	*CRI*	Volume	Area	Dipole moment	E_{LUMO}	E_{HOMO}	E	$E_{LUMO} - E_{HOMO}$	Hardness	Electro-negativity	Z	Softness	ω
31	2.651	138.403	158.605	0.666	-0.435	-9.241	0.208	8.807	4.403	4.838	0.873	0.227	2.658
32	4.792	266.850	285.082	0.015	-0.515	-9.099	1.819	8.584	4.292	4.807	0.936	0.233	2.692
33	4.989	271.525	288.113	0.714	-0.365	-9.429	0.895	9.064	4.532	4.897	0.942	0.221	2.645
34	5.509	290.686	299.618	1.334	-0.491	-9.460	-0.718	8.969	4.485	4.976	0.970	0.223	2.760