QUANTUM MODELLING OF THE TOXICITIES OF SOME POLYCHLORINATED AROMATIC COMPOUNDS

BY

HASSAN SAMUEL

DEPARTMENT OF CHEMISTRY

FACULTY OF PHYSICAL SCIENCES

AHMADU BELLO UNIVERSITY

ZARIA, NIGERIA

FEBRUARY, 2017

QUANTUM MODELLING OF THE TOXICITIES OF SOME POLYCHLORINATED AROMATIC COMPOUNDS

BY

HASSAN SAMUEL
M.Sc/SCI/24062/2012-2013

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, AHMADU BELLO UNIVERSITY, ZARIA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTERS OF SCIENCE IN PHYSICAL CHEMISTRY

DEPARTMENT OF CHEMISTRY
FACULTY OFPHYSICAL SCIENCES
AHMADU BELLO UNIVERSITY,
ZARIA, NIGERIA.

FEBRUARY, 2017

DECLARATION

I declare that the work in this dissertation entitled "Quantum Modelling of the Toxicities of some Polychlorinated Aromatic Compounds" has been carried out by me in the Department of Chemistry, Ahmadu Bello University, Zaria, under the supervision of Prof. A. Uzairu, Prof. P.A.P. Mamza and Dr. O. J. Okunola. The information derived from the literature has been duly acknowledged in the Dissertation and a list of references provided. No part of this Dissertation was previously presented for any degree or diploma at any Institution.

Hassan Samuel

| --- | --- | --- |
| Name of Student | Signature | date |

CERTIFICATION

This dissertationtitled “Quantum Modelling of theToxicitiesof Some Polychlorinated Aromatic Compounds” by Hassan Samuel meets the regulation governing the award of masters of Science in Physical Chemistry of Ahmadu Bello University Zaria, and is approved for its contribution to knowledge and literary presentation.

Prof. A. Uzairu

Chairman, Supervisory Committee Department of Chemistry Ahmadu Bello University Zaria	Signature	Date

Prof. P.A.P Mamza

Member, Supervisory Committee Department of Chemistry Ahmadu Bello University Zaria	Signature	Date

Dr. O. J. Okunola

Member, Supervisory Committee Department of Applied Chemistry Federal University, Dutsin-ma, Katsina	Signature	Date

Prof. A.O. Oyewale

Head of Department Department of Biochemistry Ahmadu Bello University Zaria	Signature	Date

Prof. A. Z. Abubakar

Dean of Postgraduate studies Ahmadu Bello University Zaria	Signature	Date

DEDICATION

I dedicate this work to The Almighty GodWho is the source of my strength and also to my parents, Mr. and Mrs. Hassan Ali for their love and support.

ACKNOWLEDGEMENT

I sincerely express my heartfelt gratitude to my distinguished supervisors, Prof. A. Uzairu, Prof. P. A Mamza and Dr O. J. Okunolafor their scholarly ideas, constructive criticism and open dialogue towards the improvement of this research.

I also wish to express my sincere gratitude to Dr G.A Shalangwa for his guidance and encouragement throughout the course of the work.

Finally, my wonderful parent, Mr and Mrs Hassan Ali, my siblings and friends are acknowledged for their encouragement, prayers, patience and goodwill throughout the period of the study.

ABSTRACT

Polychlorinated aromatic compounds represent a large group of industrial and byproduct compounds which are resistant to chemical and biological degradation and highly toxic. QSAR analysis was performed on 74 molecules of three classes of polychlorinated aromatic compounds (polychlorinated dibenzo-p-dioxin (PCDDs), polychlorinated dibenzofuran (PCDF) and polychlorinated biphenyl (PCB)). QSAR models were also developed using each of the three series. Genetic function algorithm (GFA) approach was used to generate 5 models, for the three series all together and for each of the three series. The models with the highest statistical significance were selected as the best. For the three series all together: (Model-1: R^2 = 0.9673, $R^2_{adjusted}$ = 0.9592, R^2_{cv} = 0.9402, R^2_{pred}. = 0.7209, F-test = 118.48, LOF = 0.4377). For the series PCDDs: (Semi empirical- LOF = 0.5881, R^2= 0.9516, R^2_{adj}.= 0.9389, R^2_{cv}= 0.9090, F-value = 74; DFT- LOF = 0.3877, R^2 = 0.9680, R^2_{adj}= 0.9596, R^2_{cv} = 0.9518, F-value = 115.07). For the series PCDFs: (LOF = 0.2955, R^2 = 0.9229, R^2_{adj}. = 0.9091, R^2_{cv} = 0.8885, F-value = 67.03).For the series PCBs: (LOF = 0.4582, R^2 = 0.9186, R^2_{adj}. = 0.8734, R^2_{cv} = 0.8086, F-value = 20.31). From the models generated using the three series all together, it seems that polarizability, SP-7, ETA_Epsilon_5, GRAVH_3, and MOMI-R contributed positively to the toxicity of these compounds while MaxHBint5, ETA_dApha_B, ETA_Epsinlon-2, n5Ring and GRAV_2 contribute negatively. From the comparison of the models generated using DFT and semi-empirical and based on their statistical parameters, semi-empirical (AM1) has slightly better predictive power than DFT (BLYP/6-31G*). The robustness and applicability of the models were established by internal and external validation techniques. These validated modelsbring important insight to aid the prediction and identification of other toxic polychlorinated aromatic compounds.

TABLE OF CONTENT

Cover page i

Title page ii

Declaration iii

Certification iv

Dedication v

Acknowledgement vi

Abstract vii

List of Tables xiv

List of Figures xvi

List of Abbreviations xix

CHAPTER ONE 1

1.0INTRODUCTION 1

1.1.0 Sources of Polychlorinated Aromatic Compounds 2

1.1.1 Incineration sources 3

1.1.2 Combustion sources 6

1.1.3 Industrial sources 7

1.1.4. Reservoir sources 8

1.2.0 Source for Human Exposure 11

1.3.0 Toxicity of PolychlorinatedAromatic Compounds 11

1.4.0 Economic Impact of Polychlorinated Aromatic Compounds 13

1.5 Statement of the Problem 14

1.6 Justification to the Research 15

1.7 Research Question: 17

1.8 Research Hypothesis 17

1.9 Theoretical Frame Work 17

1.10. Aim and Objectives 19

1.7 Scope and Limitations of the Study 20

CHAPTERTWO 22

2.0LITERERATUREREVIEW 22

2.1 Number of Possible Isomers 24

2.2 Toxicity of Polychlorinated Aromatic Compounds 29

2.2.1 The concept of toxic equivalency factors 30

2.2.2 Evolution of the toxicity equivalence methodology 31

2.2.3 Toxicological Effects on experimental animals and in vitro test systems 39

2.2.4 Carcinogenic effects on experimental animal 40

2.2.5 Toxicological effects in humans 41

2.2.6 Carcinogenic effects in Human 41

2.2.7 Population groups at higher exposures 42

2.2.8 Critical organs, tissues and effects 43

2.3.0 QSAR 44

2.3.1 Short history of QSAR and molecular descriptors 45

2.3.2 Chemometrics and QSAR modeling 48

2.3.3 Specific QSAR Approaches 53

2.3.3 Molecular descriptors 57

2.4 Previous QSAR work on polychlorinated aromatic compounds 70

2.4.1 Docking and 3D-QSAR studies on the Ah receptor binding affinities of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs) 71

2.4.2 Binding affinities (AhR) of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs) study combining DFT and QSAR Results 72

2.4.3 QSAR for toxicities of polychlorodibenzofurans, polychlorodibenzo-1, 4-dioxins, and polychlorobiphenyls 73

2.4.4 QSAR Studies on polychlorinated aromatic compounds using topological descriptors 75

CHAPTERTHREE 77

3.0 MATERIALSANDMETHODS 77

3.1 Materials 77

3.1 Methods 77

3.2.0 Data Collection 78

3.3 Molecular Optimization 79

3.3.1 Optimization with spartan 14. V.1.1.0 97

3.4 Calculation of molecular descriptors 97

3.4.1 Computing molecular descriptors using Spartan 14v.1.1.0 software 100

3.4.2 Computing molecular descriptors using Padel descriptor tool kit. 110

3.5 Calculation of Molecular Descriptors using Mathematical Formulae 111

3.6 Selection and Elimination of Molecular Descriptors 121

3.7 Data Set Division 122

3.8 Correlation Analysis using Material Studio 7.0 126

3.9 Model Building using Genetic Functional Approximation (GFA) 126

3.9.1 Procedure for Building GFA Model 132

3.10 Model Validation 130

3.10.1 Internal validation parameters 136

3.10.2 External validation parameters 142

3.11 Multicollinearity among Descriptors: Variance Inflation Factor (VIF) 144

3.12 Euclidean Based Applicability Domain (AD) 144

3.13 Procedure for Computing Euclidean Distance using DTC-Euclidean Program 145

CHAPTER FOUR..... 150

4.0 **RESULTS** 150

4.1 Results of QSAR Modeling of pEC50 for (PCDDs, PCDFs and PCBs)..... 150

4.1.1 GFA Generated models of pEC50 for PCDDs, PCDDFs and PCBs all together..... 150

4.1.2 Statistical/Validation parameters for the generated models..... 150

4.1.3 Comparison of the observed and predicted pEC50..... 151

4.1.4 Univariate analysis of the toxicity (pEC50) data 154

4.1.5 External validation parameters 160

4.1.5.2 Euclidean base applicability domain for the selected model (model-1) 160

4.1.5.3 Golbraikh and Tropsha's parameters for the selected model (model-1)..... 160

4.1.6 Linear relationship of observed and the predicted pEC_{50} values for all the series 160

4.1.7 Plot of residual versus actual toxicity (pEC_{50}) values 160

4.2.0 Results of QSTR modeling for pEC_{50} for all the series together (PCDDs, PCDFs, and PCBs). 161

4.2.1 Five (5) generated models by GFA..... 152

4.2 Results of QSAR modeling of pEC50 for (PCDDs)..... 168

4.2.1 GFA Generated models of pEC50 for PCDDs Using DFT and SE approaches..... 168

4.2.2 Correlation matrix for DFT and SE approaches 168

4.2.3 Statistical/Validation parameters for the five models generated for DFT and SE approaches 168

4.2.4 Comparison of the observed and predicted pEC50..... 172

4.2.5 Univariate analysis of the of the toxicities (pEC50) data For DFT and SE approaches.... 175

4.2.4 Linear relationship of observed pEC_{50} values with the predicted values 175

4.2.6 Plot of residual versus actual toxicity values for DFT and SE calculated descriptors 175

4.3 Results of QSAR modeling of pEC50 for (PCDDs)..... 175

4.4.0 Results of QSAR modeling of pEC_{50} for PCDFs 183

4.4.1 GFA Generated models of pEC_{50} for PCDFs 183
4.4.2 Statistical/Validation parameters for the generated models 183
4.4.3 Comparison of the observed and predicted pEC50 183
4.4.4 Univariate analysis of the toxicity (pEC_{50}) data 183
4.4.6 Linear relationship of observed and the predicted pEC_{50} values for PCDFs 183
4.4.7 Plot of residual versus actual toxicity (pEC_{50}) values 184
4.5.0 Results of QSAR modeling of pEC_{50} for PCBs 185
4.5.1 GFA Generated models of pEC_{50} for PCBs 185
4.5.2 Statistical/Validation parameters for the generated models 185
4.5.3 Comparison of the observed and predicted pEC50 185
4.5.4 Univariate analysis of the toxicity (pEC_{50}) data 192
4.5.6 Linear relationship of observed and the predicted pEC_{50} values for PCBs 192
4.5.7 Plot of residual versus actual toxicity (pEC_{50}) values 192
CHAPTERFIVE 199
5.0DISCUSSIONOFRESULTS 199
5.1 Geometry Optimization and Calculation of Molecular Descriptors 199
5.2 GFA Generated Model for PCDDs, PCDDFs and PCBs all together 199
5.2.1 Significance of the descriptors in model 1 200
5.3 GFA Generated Models for the pEC_{50} of PCDDs 201
5.3.1 Significance of the descriptors in the models for PCDDs 202
5.4 GFA Generated Models for the pEC_{50} of PCDFs 203
5.4.1 Significance of the descriptors in the models for PCDFs 204
5.5 GFA Generated Models for the pEC_{50} of PCBs 204
5.5.1 Significance of the descriptors in the models for PCBs 206

CHAPTERSIX 207

6.0 **SUMMARY**, **CONCLUSIONANDRECOMMENDATIONS**..... 207

6.1 Summary of the Findings..... 207

6.2 Conclusion 208

6.3 Recommendation 208

REFERENCES 210

LIST OF TABLES

2.1: Permutations of the substituents in polychlorinated-biphenyls 24

2.2: Permutations of the substituents in polychlorodibenzo-1,4-dioxins 26

2.3: The possible permutations for PCDFs 27

2.4: Number of polychlorinated dioxin, furan, and biphenyl congeners32

2.5: World Health Organization TEFs for mammals, birds, and fish 37-38

3.1: Experimental Biological Activities in pEC_{50} of the compounds,(PCDDs, PCDFs, PCBs)

79-94

3.2 Various types of descriptors (0D, 1D, 2D and 3D) calculated 113-119

3.3: Criteria for selection of acceptable model146

4.1: Statistical/validation parameters for the generated model 146

4.2: Chemical names, observed and calculated toxicity (pEC_{50})for the training set compounds

151-153

4.3: Chemical names, observed and calculated toxicity (pEC_{50})for the test setcompounds \

154-155

4.4:Univariate analysis of the toxicity data 156

4.5: Euclidean based applicability domain for model-1 (training set) 158-159

4.6 Euclidean based applicability domain for model-1 (training set) 160

4.7: Golbraikh and Tropsha external validation for model-1 161

4.8: Five generated models using DFT and SE approaches 166-67

4.9: The best model for DFT and SE for PCDDs 168

4.10: Correlation matrix for DFT calculated descriptors 169

4.11: Correlation matrix for SE calculated descriptors 170

4.12: Statistical/validation parameters for DFT and SE 171

4.13: Actual and predicted toxicity values for DFT and SE 173-174

4.14: Univariate analysis for the toxicity data for DFTand SE 175

4.15: Statistical/validation parameters for PCDFs 183

4.16: Comparison of the observed and predicted pEC_{50} for PCDFs 184-185

4.17: Results for univariate analysis for PCDF186

4.18:Statistical/ validation parameters for PCBs191

4.19: The actual, predicted and residual pEC_{50}values for PCBs 192

4.20: Results for univariate analysis for PCBs 193

LIST OF FIGURES

1.1: Emission sources of polychlorinated aromatic compounds 10

1.2: The Significance of QSAR 17

2.1: General scheme of the QSAR/QSPR strategy 49

2.2: General scheme of the different sources of molecular descriptors 58

2.3: A lattice of grid point with an embedded molecule 59

3.1: Requirements for QSAR (selection of data set) 77

3.2: General structural formulae for the compounds (PCDDs, PCDFs and PCBs) 78

3.3: The interface of the software spartan 14v1.1.0 96

3.4 .Atomic fragments for molecular drawing 97

3.5: The Spartan interface showing the minimized molecules 98

3.6: Chart for the methods of calculation of molecular descriptors in QSAR 99

3.7. *Calculations* dialog box using Spartan 14 software 100

3.8. File browser for the optimized molecules 101

3.9. Notification that calculation has started. 102

3.10 Notification that calculation is completed. 103

3.11. Dialog box showing molecular properties (descriptors) 105

3.12: Padel-Descriptor interface 106

3.13. Descriptor's file 107

3.14. Calculation of 1D, 2D and 3D descriptors 108

3.15. Collection of the calculated descriptors using all the methods. 109

3.16: The interface for dataset division software 120

3.17: “Choose file” dialog box 121

3.18: “Save file” dialog box 122

3.19: Dialog box for imputing the percent of the training set 124

3.20: Material Studio’s work space window and project window 127

3.21: Material Studio’s *Study table* window 126

3.22: Highlighted data for the correlation analysis 127

3.23: Square correlation matrix between $pEC_{50}$and the descriptors 128

3.24: Material Studio’s work space window and project window 131

3.25: Material Studio’s *document* window 132

3.26: Material Studio’s *Study table* window 133

3.27. Imported descriptors showing the highlighted column of the pEC50 134

3.28. Material Studio’s Genetic function approximation wind 135

3.29: The folder of Euclidean-Applicability domain 144

3.30: Euclidean-Applicability domain training / test set dialog box 145

4.1: Linear relationship of observed and predicted toxicity data of the training set 162

4.2: Linear relationship of observed and predicted toxicity data of the test set 163

4.3: Plot of residual values versus actual (pEC50) values 164

4.4: Linear relationship between actual and the predicted (pEC50) using DFT approach 176

4.5: Linear relationship between actual and the predicted (pEC50) using SE approach 177

4.6: Plot of residual versus actual toxicity for DFT calculated descriptors 178

4.7: Plot of residual versus actual toxicity for SE calculated descriptors 179

4.8: Linear relationship of observed and predicted pEC50 value for PCDFs 187

4.9: Plot of residual versus actual toxicity value for PCDFs 188

4.10: Linear relationship of observed and predicted pEC50 value for PCBs 194

4.11: Plot of residual versus actual toxicity value for PCBs 194

LIST OF ABBREVIATIONS

0D, 1D, 2D, 3D	Zero,one, two and three dimensional descriptor respectively.
AhR	aryl hydrocarbon receptor
AM1	Austin Model 1
ANNs	Artificial neural networks
B3LYP	Becke's three-parameter hybrid functional
CDD	tetrachlorodibenzo- p- dioxin
ClPAHs	Chlorinated polycyclic aromatic hydrocarbons
DFT	Density functional theorem
DLC	dioxin-like compounds
GFA	Genetic function approximation
I-TEF	international scheme of toxic equivalency factors
MLR	multiple linear regression
MMFF	Molecular mechanics force field
NRC	National Research Council
OLS	Ordinary least squares
PACs	polychlorinated aromatic compounds
PCA	principal component analysis
PCBs	polychlorinated biphenyls

PCDDs	polychlorinated dibenzo–p-dioxins
PCDFs	polychlorinated dibenzofurans
PLS	partial least squares
QSA	Quantitative structure activity relationship
QSTR	Quantitative structure toxicity relationship
SE	Semi-empirical
TEFs	toxic equivalency factors
TEQs	toxic equivalents

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background to the study

Since the middle of the 20th century there has been an increasing concern about the exposure ofhumans and wildlife to certain xenobiotic that were released into the environment due to diverse anthropogenic activities. One group of environmental toxicants that is of particular interest with regard to potential environmental health effects are polychlorinated aromatic compounds (PACs). These ubiquitous compounds are hydrophobic, lipophilic and resistant to biological and chemical degradation, properties that impart persistency and a propensity to bio-accumulate and biomagnify to concentrations that can cause harmful effects (Eichbaum *et al.,* 2014; Giesy *et al.,* 1994; Larsson *et al.,* 2013; Poland *et al.,* 1982; Song *et al.,* 2006; Vries *et al.,* 2006).

Polychlorinated aromatic compounds are a group of compounds comprising polycyclic aromatic hydrocarbons with two or more aromatic rings and one or more chlorine atoms attached to the ring system. Polychlorinated aromatic compounds can be divided into two groups: chloro-substituted Polycyclic Aromatic Hydrocarbons(PAHs), which have one or more hydrogen atoms substituted by a chlorine atom, and chloro-addedChlorinated Polycyclic Aromatic Hydrocarbons(ClPAHs), which have two or more chlorine atoms added to the molecule(Nilsson*etal*., 1993).They are products of incomplete combustion of organic materials. They have many congeners, and the occurrences and toxicities of the congeners differ (Kitazawa*et al*., 2016). ClPAHs are hydrophobic compounds and their persistence within ecosystems is due to their low water solubility (Cernigli*a etal*., 1992). They are structurally

similar to other halogenated hydrocarbons such as polychlorinated dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), and polychlorinated biphenyls (PCBs). ClPAHs in the environment are strongly susceptible to the effects of gas/particle partitioning, seasonal sources, and climatic conditions (Ohura*et al*., 2008).

Polychlorinated aromatic compounds such as polychlorinated dibenzo–p-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), polychlorinated biphenyls (PCBs), as well as other related halogenated aromatic compounds are trace amounts of undesired impurities (by products) in the manufacture of other chemicals such as chlorinated phenols and their derivatives, chlorinated biphenyl ethers, and polychlorinated biphenyls (PCBs), and combustion of chlorine containing materials under some conditions. These compounds are also referred to as unintentionally produced Persistent Organic Pollutants (UPOP) (WHO, 1988).

1.2 Sources of PolychlorinatedAromatic Compounds

Polychlorinated aromatic compounds are generated by combustion of organic compounds. They enter the environment from a multiplicity of sources and tend to persist in soil and in particulate matter in air. Environmental data and emission sources analysis for polychlorinated aromatic compounds reveal that the dominant process of generation is by reaction of polycyclic aromatic hydrocarbon (PAHs) with chlorine in pyrosynthesis (Ohura, 2007).

Polychlorinated aromatic compounds (PCBs and Dioxins) have commonly been detected in tap water, fly ash from an incineration plant for radioactive waste, emissions from coal combustion and municipal waste incineration, automobile exhaust, snow, and urban air (Nilsson

et al., 1993). PCBs and Dioxins have also been detected in electronic wastes, workshop-floor dust, vegetation, and surface soil collected from the vicinity of an electronic waste (e-waste) recycling facility and in surface soil from a chemical industrial complex (comprising a coke-oven plant, a coal-fired power plant, and a chloro-alkali plant), and agricultural areas in central and eastern China (Ma*et al*., 2009). In addition, the combustion of polyvinylchloride and plastic wrap made from polyvinylidene chloride result in the production of polychlorinated aromatic compounds, suggesting that combustion of organic materials including chlorine is a possible source of environmental pollution (Wang*et al*., 2003).

1.2.1 Incineration sources

Incinerationis the largest source of dioxinrelease into the environment. According to National Renewable Energy Laboratory (1999), the majority of dioxins are predominantly from municipal solid waste incineration.

From literature, PVC produces HCl upon combustion that is almost quantitatively related to its chlorine content (Steiglitz *et al*., 1988). However, extensive studies in Europe indicate that the chlorine found in emitted dioxins is not derived from HCl in the flue gases, instead, most dioxins arise in the condensed solid phase by the reaction of inorganic chlorides with graphitic structures in char-containing ash particles and the reaction was catalyzed by copper (Steiglitz *et al*., 1988).

Studies of household waste burning indicate consistent increases in dioxin generation with increasing PVC concentrations (Costner*et al*., 2005). According to the EPA dioxin inventory, landfill fires are likely to represent an even larger source of dioxin to the environment. A survey of international studies consistently identifies high dioxin concentrations in areas

affected by open waste burning and a study that looked at the homologue pattern found the sample with the highest dioxin concentration was typical for the pyrolysis of PVC (Costner*et al*., 2005). Other EU studies indicate that PVC likely accounts for the overwhelming majority of chlorine that is available for dioxin formation during landfill fires(Costner*et al*., 2005).

The next largest sources of dioxin in the EPA inventory are medical and municipal waste incineraton (*Beychok* , 1987). For instance a study of commercial-scale incinerators showed no relationship between the PVC content of the waste and dioxin emissions (Rigo*et al*., 1995). Other studies have shown a clear correlation between dioxin formation and chloride content and indicate that PVC is a significant contributor to the formation of both dioxin and PCB in incinerators (Katami*et al*., 2002; Wagner*et al*., 1993; Thornton, 2002).

Several researchers have described their mechanism of formation dioxins and dioxin-like compounds. Overall, it is observed that the emission of dioxins and furans into the environment can be explained mainly by two principal surface catalytic processes:

i) Formation from precursors and

ii) formation by de novo synthesis (Altwicker*et al.*, 1996).

An informative review of the formation and mechanism of dioxins from municipal solid waste incineration was reported (Tuppurainen*et al.*, 1998). It was observed that several past studies demonstrated the presence of significant quantities of dioxins and dioxin precursors in municipal solid waste around 50 ng(Toxic equivalent) I-TEQ/kg (Abad*et al.*, 2002).

Gullett *et al*. (1990a; 1990b; 1991a; 1991b; 1992) studied the formation mechanisms through extensive combustion research at EPA, and verified the observations of Vogg *et al*. (1987). It was proven that CDDs and CDFs could be ultimately produced from low temperature

reactions (i.e., 350°C) between Cl_2 and a phenolic precursor, combining to form a chlorinated precursor, followed by oxidation of the chlorinated precursors (catalyzed by a copper catalyst such as copper chloride) as in examples (1) and (2), below.

(1) The initial step in forming dioxin is the formation of chlorine from HCl in the presence of oxygen (the Deacon process), as follows (Vogg *et al*., 1987; Bruce *et al*., 1991):

$$2HCl + \frac{1}{2}O_2 \rightarrow H_2O + Cl_2$$

(2) Phenolic compounds adsorbed on the fly ash surface are chlorinated to form the dioxin precursor, and the dioxin is formed as a product from the breakdown and molecular rearrangement of the precursor. The reaction is promoted by copper chloride acting as a catalyst (Vogg et al., 1987; Dickson and Karasek, 1987; Gullett et al., 1992):

(a) $Phenol + Cl_2 \rightarrow chlorophenol$ (*dioxin precursor*)

(b) 2-chlorophenol + ½ O_2 ————> dioxin + Cl_2

On the other hand, Eklund *et al*. (1986) observed the high temperature formation of a large variety of chlorinated toxic compounds, including CDDs and CDFs, from precursors during a simple experiment in which phenol was oxidized with HCl at 550°C. One milligram of phenol was placed in a quartz tube reactor with an aqueous solution (10μL) of HCl and heated at 550°C for 5 minutes. Trichlorobenzene, dichlorophenol, dichlorobenzofuran, tetrachlorobenzene, trichlorophenol, and tetrachlorophenol were identified as major products formed. Monochlorobenzene, chlorophenol, dichlorobenzene, tetrachloropropene, pentachloropropene, trichlorobenzofuran, tetrachlorodibenzofuran, trichlorodibenzodioxin, tetrachlorodibenzodioxin,

hexachlorodibenzodioxin, hexachlorodibenzofuran, pentachlorobenzene, pentachlorobiphenyl, and pentachlorodihydroxycylohexane were seen as minor products. Trace species formed included: monochlorodibenzofuran, pentachlorodibenzofuran, pentachlorodibenzodioxin, octachlorodibenzofuran, and octachlorodibenzodioxin. (Eklund *et al*. 1986) hypothesized that chlorinated organic compounds can be produced from phenols, acids, and any chlorine source in the hot post-combustion region (e.g., just exit to the furnace). The reaction was seen as very sensitive to HCl concentration

1.2.2 Combustion sources

Cement kilns

The switch to burning hazardous waste as fuels for cement kilns has created problems for individuals and organizations. About 16% of the facilities burn hazardous wastes as auxiliary fuel. Some limited data suggest that PCDD/PCDF levels in clinker dust and stack emissions of these kilns may be significantly higher than the kilns which do not burn hazardous waste (Abad *et al.,* 2004; Eduljee *et al.,* 1999).

Wood burning

A number of studies have found dioxins in the emissions and ash/soot from wood fires in non-industrial situations (Stanmore *et al.,* 2004). According to the European Emission Inventory, wood combustion is at present one of the most important air emission sources of dioxins (Quaß*et al.,* 2000). In an appealing review paper, it is reported that the dioxin emission from wood burning is about 945 g (Toxic equivalent) I-TEQ/year (Lavric *et al.*,2004).

Diesel vehicles

A very scanty literature available on this source showed that dioxins could be emitted from diesel vehicles. In fact, researchers from Sweden and Norway have shown that dioxins are emitted from diesel vehicles (Marklund*et al.,* 1990; Oehme *et al.,* 1991). As these studies depend on the fuel used in a particular country more studies are required in order to reach a conclusive estimate.

Crematoria

Crematoria procedures can be a ready source of organic material and chlorine, and hence are possible source of dioxin emission (Alcock*et al.,* 1999). Inventory estimates rate this source as 0.3% of European output and 0.24% of US output (USEPA, 1998).

Coal-fired utilities

Although emission of dioxins compared to the wood burning are less, they are numerous, large in size and their high stacks indicate that they could impact verylarge areas (Chen*et al.,* 2004; Harrad*et al.,* 1991). Considering the large scale usage the importance of these facilities is very much unknown.

1.2.3 Industrial sources

Pulp and paper mills

The manufacture of bleached pulp and paper has in the past resulted in dioxin releases to water, land and paper products. These compounds can be formedthrough the chlorination of naturally occurring phenolic compounds such as those present in wood pulp (Rappe *et al.,* 1987). It is

reported that the waste generated from a pulp mill of China produces dioxin concentration of 300 pg/l I-TEQ (Zheng *et al.,* 2001).

Metals industry

The metallurgical processes such as high temperature steel production, smelting operations, and scrap metal recovery furnaces are found to be typical sources of dioxins (Anderson *et al.,* 2002). Processes in the primary metals industry, such as sintering of iron ore, have also been identified as potential sources (Ciepliket *al.,* 2003; Wang*et al.,* 2003). In several countries the annual release of dioxins is estimated to be 500–4000 g I-TEQ (Anderson *et al.,* 2002).

Chemical manufacturing

PCDDs and PCDFs can be formed as by-products from the manufacture of chlorinated compounds such as chlorinated phenols, PCBs, phenoxy herbicides, chlorinated benzenes, chlorinated aliphatic compounds, chlorinated catalysts and halogenated diphenyl ethers (Öberg*et al*., 1992; Oberg *et al.,* 1993; Sidhu *et al.,* 2002).

1.2.4 Reservoir sources

The persistent and hydrophobic nature of these compounds causes them to accumulate in soils, sediments, landfill sites, vegetation and organic matter. They have potential for redistribution and circulation of dioxins in the environment. The dioxin compounds in the reservoirs‖ can be redistributed and circulated in the environment by dust or sediment suspension and transport (Kjelle*et al.,* 1995; Rotard *et al.,* 1994). The major reservoir sources include:

a. Biological processes: The action of microorganisms on chlorinated phenolic compounds results in the formation of dioxins under certain environmental conditions (Siewers*et al.,* 1994).

b. Photochemical processes: Dioxins like OCDD (1,2,3,4,5,6,7,8,9-octachloro-dibenzodioxin), HPCD (1,2,3,4,5,6,7,8-heptachloro-dibenzodioxin) formation occurs by photolytic radical reactions of pentachlorophenol (Baker *et al.,* 2000; Tysklind*et al.,* 1993).

c. Accidental sources: The incidents of dioxin release at Seveso, Italy and Yusho Japan could be considered as an accidental release of dioxins into the atmosphere. Further, forest fires and volcanoes also come under this category (Clement*et al.,* 1991; Ruokojärvi *et al.,* 2000).

d. Miscellaneous sources: Miscellaneous sources include formation of dioxins in FBC (Fluidized Bed Combustion) boilers, thermal oxygen cutting of scrap metal at demolition sites, power generation, PVC in house fires, Kraft liquor boilers, laboratory waste, drum and barrel reclaimers, tire combustors, carbon reactivation furnaces and scrap electric wire recovery facilities, etc. (Anthony*et al.,* 2001; Carroll*et al.,* 1996; Menzel*et al.,* 1998). The summary of the various sources of polychlorinated aromatic compounds is presented in Figure 1.1 according to Weber *et al.* (2008).

The main new (de novo) sources of PCDD/Fs are combustion processes, such as burning of waste and metal smelting,and refining. In Europe, the Baltic Sea is an important sink ofPCBs and dioxins. However, recent studies have revealed a major problem at localized spots,

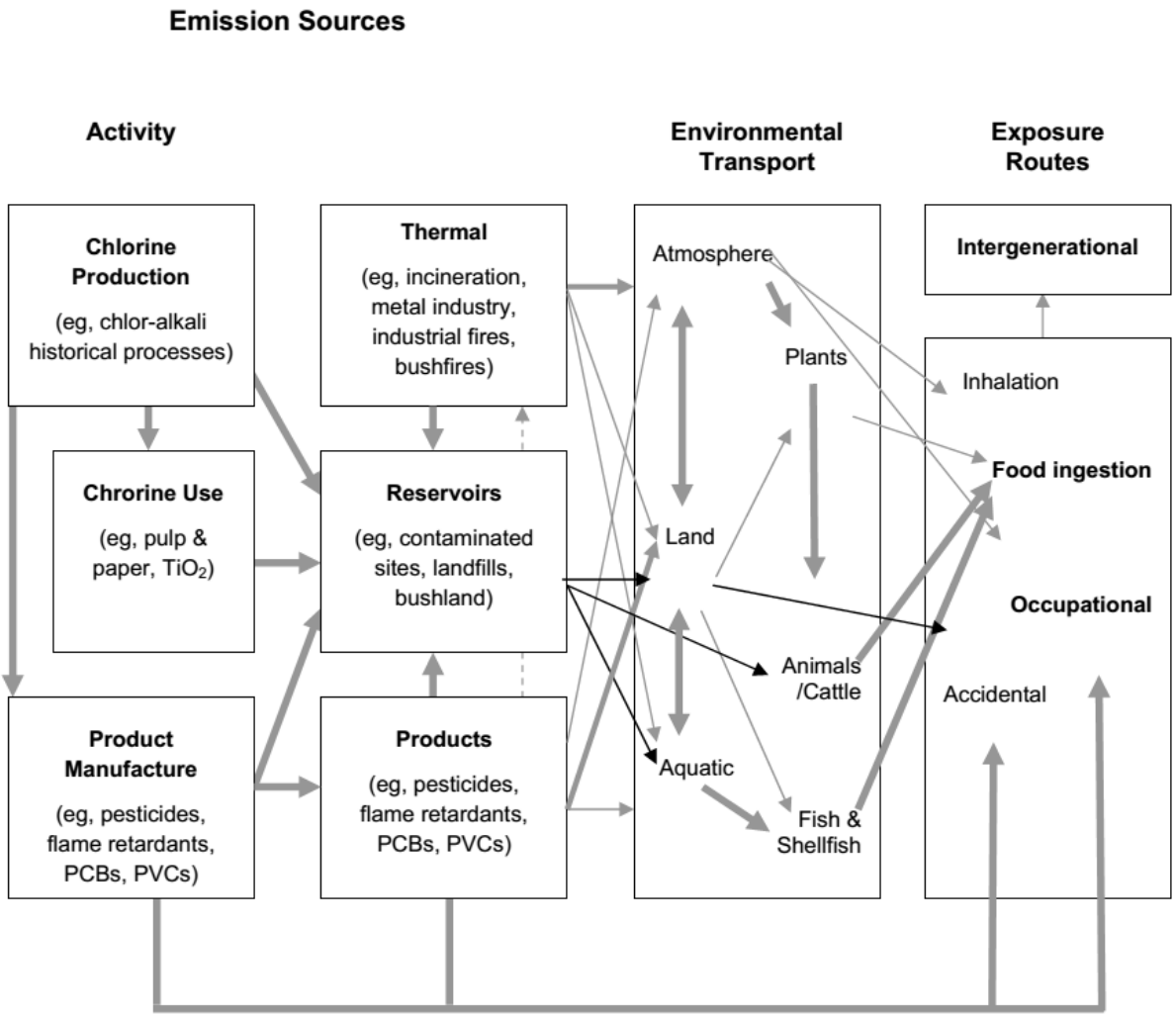


Figure 1.1 Emission sources of polychlorinated aromatic compounds

\

due to theproduction and use of chlorophenolsfor impregnation of timber. In the most contaminated regions the concentration of PCDDs and PCDFs in soil and sediments appears to be incredibly high. An unpredictable source is old transformers and capacitors, each of which may contain several kilograms of PCBs and hundreds of milligrams of PCDD/Fs.

1.2.5 Source for Human Exposure

Food is the major source for human exposure to PCBs and dioxins, especially fatty foods: dairy products (butter, cheese, fatty milk), meat, egg, and fish. Food of animal origin accounts for 95 % of total exposure. The current average body burden of dioxins is about 5–50 ng/kg (as WHOTEq in fat; pg/g = ng/kg) or 100–1000 ng (WHO-TEq) per person which is close to the lowest concentrations possibly causing health effects. Some subgroupswithin the society (e.g., nursing babies and people consuming plenty of fish) may be exposed to higher than average amounts of these compounds and are thus at greater risk. Dioxin concentrations have been screened in five WHO international studies, and in Central Europe the concentrations have decreased in breast milk from about 40 ng/kg (as TEq in milk fat) in 1987 to below 10 ng/kg in 2006. PCBs have decreased at about the same rate. The decrease in environmental concentrations is due to cessation of PCB use and improved incineration technology.

1.2.6 Toxicity of PolychlorinatedAromatic Compounds

When several Polychlorinated Aromatic compounds appear in mixture (Fiedler *et al.*, 1990) (as they occur in the environment), and when as usual they have the same mechanism of action, their cumulated toxic effect could be found by a relationship involving partial toxicities, reminiscent of Dalton's Law of partial pressures: total equivalent toxicity is the sum of the

product of their concentration with their Toxic Equivalent Factor (TEF) value (Van der Berg *et al.,* 1994).

Polychlorinated aromatic compounds (Dioxin-like compounds) is one of the most toxic chemicals known. A report released for public comment on September 1994 by the US Environmental Protection Agency (USEPA, 1994) clearly describes dioxin and dioxin-like compounds as a serious public health threat in the 1960s. According to the EPA report, not only does there appear to be not safelevel of exposure to dioxin, but levels of dioxin and dioxin -like chemicals have been found in the general US population that is at or near levels associated with adverse health effects. The EPA report confirmed that dioxin is a cancer hazard to people, that exposure to dioxin could also cause severe reproductive and developmental problems (at levels 100 times lower than those associated with its cancer causing effects); and that dioxin could cause immune system damage and interfere with regulatory hormones.

These organic pollutants constitute a group of lipophilic chemically, stable environmental contaminants with low volatility which have been identified in fatty tissues of animals and humans. Several of these polychlorinated aromatic compounds have been shown to exert a number of common toxic responses similar to those observed for 2, 3, 7, 8 tetrachlorodibenzo- p-dioxin (TCDD). These include thermal toxicity, immunotoxicity, endocrine toxicity and carcinogenicity/tumour promotion.

Some of the effects of polychlorinated aromatic compounds biological effects including hepatoxicity, endocrine effects, immunotoxicity, body weight loss, teratogenecity,carcinogenicity and the induction of enzymes such as ary hydrocarbon

hydroxylate [AHH] and 7-ethoxyresorufin o-deethylase[EROD] in various organisms (chovoncova*et al.,* 2005).

There is strong evidence suggesting a common mechanism of action of 2, 3, 7, 8 – TCDD and related compounds, based on a binding of these compounds to a specific receptor (the aryl hydrocarbon – receptor) (Walker *et al* 1991). Some of polychlorinated biphenyl (PCB) also have dioxin–like properties and are included as part of dioxin and dioxin–like compounds with 1 to 10 chlorine atoms attached to biphenyl molecule. There are about 209 possible PCB congeners although only 130 were found in commercial PCB mixture (Van den Berg *et al*, 2006).

1.2.7 Economic Impact of Polychlorinated Aromatic Compounds

Information on effects in humans predominantly comes from high exposure in industrial settings or accidents and intoxication incidents, e.g. the contamination of rice oil with PCBs that contained comparatively high concentrations of PCDFs in Japan and Taiwan in 1968 and 1979, respectively. The effect observed are generally in agreement with what could be expected from animal experimentation, although the most typical human effect of high exposure to PCDDsand PCDFs seems to be the dermatological response chlorace Subtle endocrinological effect e.g. modulation of thyroid hormone and testosterone levels in plasma, and decreased glucose tolerance and subtle neurological effect have been observed in humans, but their clinical significance is currently unknown (Huisman *et al.,* 1995).

The production and use of PCBs have been discontinued in most countries, but large amounts remain in electrical equipment, plastic products, buildings (e.g. plastic carpeting, sealing materials), and in the environment. Because PCBs are considered problem waste, their disposal is expensive, and may sometimes lead to attempts to dispose of them by mixing them to

other waste products. Two well-established environmental accidents leading to human suffering have occurred, called Yusho (Japan) and YuCheng (Taiwan) poisonings. In both cases rice oil was contaminated andcaused a number of health effects. Later contamination incidents have not lead to clear health consequences.

1.3 Statement of the Problem

Polychlorinated aromatic compounds are persistent and widespread environmental contaminants which can cause a great diversity of biological effects including hepatoxicity, endocrine effects, immunotoxicity, body weight loss, teratogenecity,carcinogenicity and the induction of enzymes such as ary hydrocarbon hydroxylate [AHH] and 7-ethoxyresorufin o-deethylase[EROD] in various organisms. (chovoncova*et al.,* 2005).

These environmental pollutants, which persist worldwide and are implicated in a variety of adverse health effects, are metabolized by cytochrome P450s to their hydroxylated forms.
The formation of the hydroxylated form of these environmental pollutants may lead to their detoxication, however, these metabolites have been shown to be retained in blood, liver and adipose tissues (safe *et al*, 1985).

Polychlorinated aromatic compounds are ubiquitously environmental contaminants .Some of them, notably TCDD (2,3,7,8-tetrachlorodibenzo-p-dioxin) belong to the most toxic synthetic compounds known. They are very stable against chemical and microbiological degradation and therefore persistent in the environment. They are fat-soluble and thus tend to bioaccumulate in tissues, lipid and in the food chain. These factors increase their potential harzards to humans and animals. Polycholorinated aromatic compounds (PCBs) which are

considered problem waste are present in large amount in electrical equipments, plastic products, buildings (plastic carpenting, sealing materials) and in the environment.

The Quantitative Structure-Toxicity Relationship (QSTR) models for predicting the toxicity of polychlorinated compounds can be used to screen out the most toxic compounds of polychlorinated compounds and other related compounds.

1.4 Justification to the Research

Because of the persistent and toxic effect of polychlorinated aromatic compounds which are released into the enviroment by chemical manufacture, industrial and agricultural uses, and disposal following domestic application, information regarding their toxicity and fate is required.

However, assessing the fate and toxicity of large number of these chemicals using conventional methods using classical in vivo animal testing becomes expensive and time consuming, hence, there is need for an alternative method. Consequently, in the area of computer – aided toxicity prediction, quantitative structure activity relationships (QSARs) have been seen asan attractive method for toxicity and fate assessment. (Knight *et al.*, 2002).Therefore, this research is initiated to help:

a. reduce the requirement for lengthy and expensive animal tests,
b. reduces animal use,
c. reduces /eliminates pains and discomfort to animals,
d. promotes savings in the cost of product development and
e. ultimately promotes greener chemistry to increase efficiency and eliminate waste.

The summary of the need for this research to be carried out is presented in Figure 1.2.

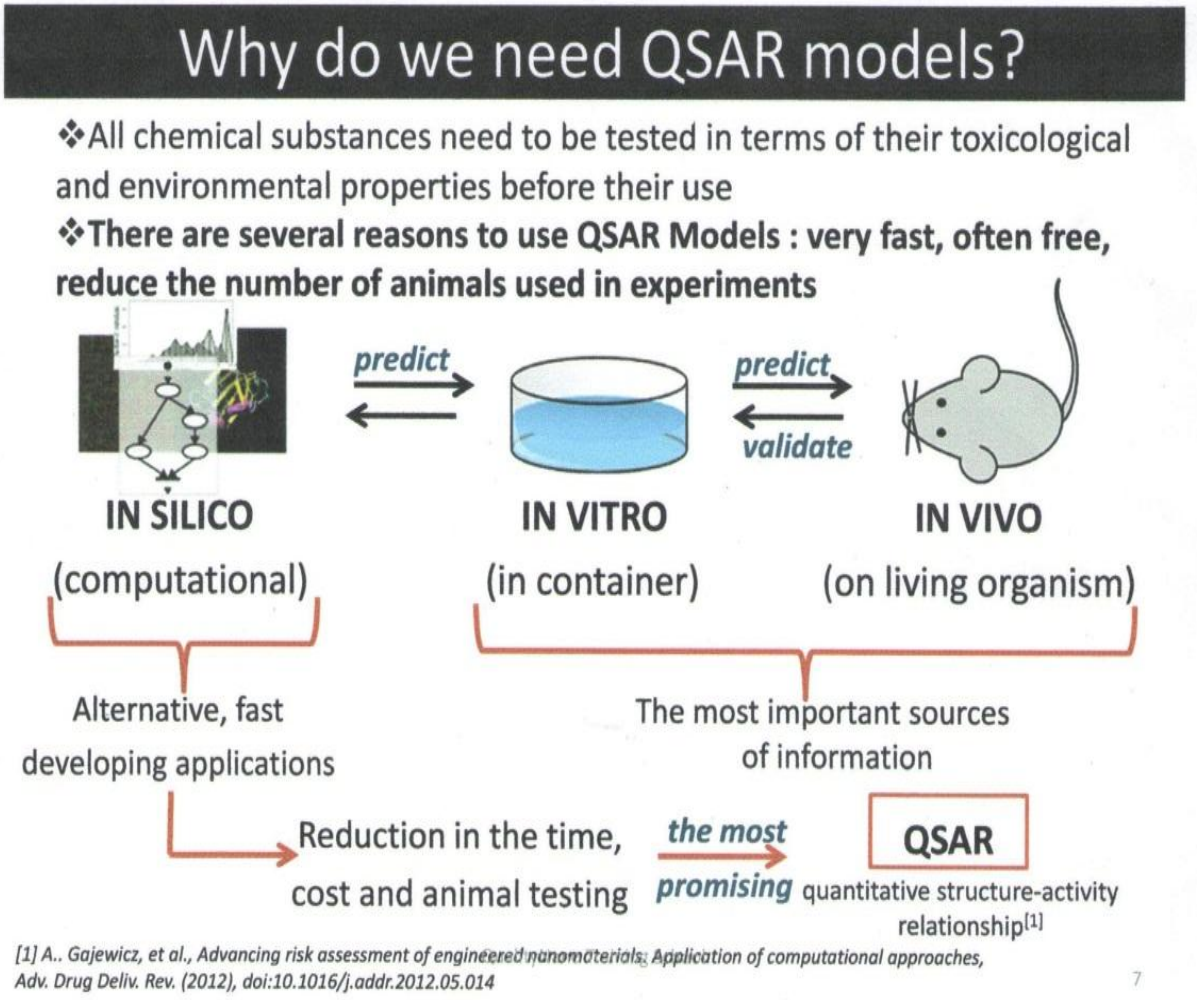


Figure1.2: The Significance of QSAR (Gajewicz *et al*, 2012).

1.5 Research Question

Structural features of molecules (molecular descriptors) are function of the activity/toxicity of the molecules. This research intends to address the questions:

(i) What are the structural features (molecular descriptors) responsible for the toxicity of polychlorinated aromatic compounds.

(ii) How could tocixity of polychlorinated dibenzo-p-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), polychlorinated biphenyls (PCBs) and other related compounds be predicted using QSAR approach?

1.6 Research Hypothesis

.This research is guided by the following null and alternative hypotheses.

Null hypothesis (H_0): The biological activities of new untested and even non-synthesized chemicals (polychlorinated aromatic compounds) are dependent of their structural properties (descriptors).

Alternative Hypothesis (H_1): The biological activities of new untested and even non-synthesized chemicals (polychlorinated aromatic compounds) are independent of their structural properties (descriptors).

1.7 Theoretical Frame Work of the study

Quantitative Structural-Activity Relationship (QSAR) is a mathematical model used to relate known activity of a congeneric series of compounds to their structure or properties to

predict other compounds' unknown activity. QSAR is based on the assumption that the biological activity of a new or untested chemical can be inferred from the molecular structure, or properties of similar compounds whose activities have already been assessed. When the relationship is developed with;non continuous or categoramrical data, it is called Structural Activity relationship (SAR) and continuous or quantitative data it is called a (QSAR) (Schultz *et al.,* 2003).

The study of the quantitative relationship between toxicity and molecular structure (QSTR/QSAR) is an important research area in computational chemistry and has been widely used in the prediction of toxicity and other biological activities of organic compounds (Katritzky *et.al*, 2000 and Katritzky *et.al*, 2001). This kind of study develops a method for the prediction of the activity under investigation of new compounds that have not been synthesized. It can also identify and describe important structure features of moleculesthat are relevant to variations in molecular properties, thus gain some insight into structural factors affecting molecular properties. To develop a QSAR model, the following steps are usually involved; data collection, molecular geometry optimization, molecular descriptors generation, descriptors selection, model development and finally model performance evaluation.

The reliability of the developed QSAR model can be tasted by comparing the end point values of the model to predict experimentally determined end point values of similar chemicals (training set chemicals). If the predictions are poor, one can restart the model development by using different descriptors, or noting training set compounds, whose predicted values deviate greatly from the experimental values. This can help to identify compounds that act by a different mechanism of action. If the predictions are good, one can define the selection criteria and limits

of the model and then make toxicity predictions to other chemicals that meet the selection criteria, and were not used during model development (Walker *et al.,* 2003).

1.8 Aim and Objectives

The aim of this research is to develop a Quantitative Structure-Toxicity relation(QSTR) model using Genetic function approximation (GFA) for the prediction of the toxicity of polychlorinated aromatic compounds (PCDDs, PCDF and, PCBs) solely on basis of chemical structures.The aim of this research was achieved through the following objectives:

(i) Selecting the data set of the molecules (polychlorinated aromatic compounds)

(ii) Drawing the molecular structure of the compounds and determining the energy of minimization and optimization for the molecules using Density Functional Theory and Semi empirical method;

(iii) Computation of molecular descriptors (numerical values that characterize properties of the molecules) using the softwares: Spartan "14", PaDel descriptor and theoretical calculations.

(iv) Conducting elimination and selection techniques of variables (descriptors).

(v) Dividing the data set into training set and test set.

(vi) Modeling the relationship between molecular descriptors and the activities of the training set using Genetic function approximation (GFA);

(vii) Validating the models using internal and external validation techniques

(vii) Selecting the best model that can give correct prediction.

1.9 Scope and Limitations of the Study

In this study, polychlorinated dibenzo-p-dioxins (PCDDs), polychlorinated dibenzofurans(PCDFs), and polychlorinated biphenyls (PCBs) are considered to represent a large group of industrial and byproduct compounds which are resistance to chemical and biological degradation and highly toxic. This research involves subjecting series of PCDDs, PCDFs, and PCBs compounds to Quantitative Structural-TOXICITY Relationship (QSTR) studies computationally using different molecular descriptors. The dependent variable in this study represents log of (EC_{50}) values (EC_{50} - median effective concentration during bioassay). Multivariate regression analysis will show how much well the toxicity can be modeled. This work will therefore contribute in the identification of those compounds which are very toxic and polluting our environment.

The limitations of this research are as follows:

- If there is a measurement error in the experimental data, it is very likely that false correlations may arise
- If the training dataset is not large enough, the data collected may not reflect the complete property space. Consequently the QSAR results cannot be used to confidently predict the most likely compounds of the best toxicity.
- There are many successful applications of this method but one cannot expect the QSAR approach to work well all the time. QSAR models are easy to build also very easy to get it wrong.

Multiple Linear Regression analysis (MLR) has the following limitations:

- MLR required normally distributed, independent and 100% relevance descriptors. This means that each descriptor is assumed the 100% relevance for explanation of the ‘cause’ of the measured endpoint.
- When the number of variables is greater than the number of observations, the MLR will not yield a unique solution (Korhonen, 2007).

CHAPTER TWO

2.0 LITERERATURE REVIEW

2.1 Polychlorinated Aromatic Compounds

Polychlorinated aromatic compounds are a group of compounds comprising polycyclic aromatic hydrocarbons with two or more aromatic rings and one or more chlorine atoms attached to the ring system. ClPAHs can be divided into two groups: chloro-substituted PAHs, which have one or more hydrogen atoms substituted by a chlorine atom, and chloro-added ClPAHs, which have two or more chlorine atoms added to the molecule (Nilsson *et al.*, 1993). They are products of incomplete combustion of organic materials. They have many congeners, and the occurrences and toxicities of the congeners differ (Kitazawa *et al.*, 2006). ClPAHs are hydrophobic compounds and their persistence within ecosystems is due to their low water solubility (Cerniglia, 1992). They are structurally similar to other halogenated hydrocarbons such as polychlorinated dibenzo-p-dioxins (PCDDs), dibenzofurans(PCDFs), and polychlorinated biphenyls (PCBs). ClPAHs in the environment are strongly susceptible to the effects of gas/particle partitioning, seasonal sources, and climatic conditions (Ohura *et al.*, 2008).

Chlorinated polycyclic aromatic hydrocarbons (ClPAHs) are generated by combustion of organic compounds. ClPAHs enter the environment from a multiplicity of sources and tend to persist in soil and in particulate matter in air. Environmental data and emission sources analysis for ClPAHs reveal that the dominant process of generation is by reaction of PAHs with chlorine in pyrosynthesis (Cerniglia, 1992). ClPAHs have commonly been detected in tap water, fly ash from an incineration plant for radioactive waste, emissions from coal combustion and municipal waste incineration, automobile exhaust, snow, and urban air (Nilsson *et al.*,1993). They have also been detected in electronic wastes, workshop-floor dust, vegetation, and surface soil

collected from the vicinity of an electronic waste (e-waste) recycling facility and in surface soil from a chemical industrial complex (comprising a coke-oven plant, a coal-fired power plant, and a chlor-alkaliplant), and agricultural areas in central and eastern China (Ma *et al.*, 2009). In addition, the combustion of polyvinylchloride and plastic wrap made from polyvinylidene chloride result in the production of ClPAHs, suggesting that combustion of organic materials including chlorine is a possible source of environmental pollution (Wang *et al.*, 2003).

Specific class ofChlorinated polycyclic aromatic hydrocarbons(ClPAHs), polychlorinated naphthalenes (PCNs), are persistent, bioaccumulative, and toxic contaminants that have been reported to occur in a wide variety of environmental and biological matrixes. ClPAHs with three to five rings have been reported to occur in air from road tunnels, sediment, snow, and kraft pulp mills (Horii *et al.*, 2008).

Recently, the occurrence of particulate ClPAHs has been investigated. Results have shown that most particulate ClPAH concentration detected in urban air tended to be high in colder seasons and low in warmer seasons. This study also determined through compositional analysis that relatively low molecular weight ClPAHs dominated in warmer seasons and high molecular weight ClPAHs dominated in colder seasons (Ohura *et al.*, 2008).

ClPAHs are hybrids of dioxins and Polycyclic Aromatic Hydrocarbons (PAHs), they are suspected of having similar toxicities (Cerniglia, 1992). These types of compounds are known to be carcinogenic, mutagenic, and teratogenic. Toxicological studies have shown that some ClPAHs possess greater mutagenicity, aryl-hydorcarbon receptor activity, and dioxin-like toxicity than the corresponding parent PAHs (Kitazawa *et al.*, 2006).

The relative potency of three ring ClPAHs was found to increase with increasing degree of chlorination as well as with increasing degree of chlorination. However, the relative potencies of even the most toxic ClPAHs have been found to be 100,000-fold lower than the relative potency of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) (Horii, *et al.,* 2009).Even though ClPAHs aren't as toxic as TCDD, it has been determined using recombinant bacterial cells that the toxicities of exposure to ClPAHs based on AhR activity were approximately 30-50 times higher than that of dioxins (Ohura *et al.,* 2008). ClPAHs demonstrate a high enough toxicity to be a potential health risk to human populations that come into contact with them.

2.2 Number of Possible Isomers

The number of all possible isomers of the above three classes of polychlorinated aromatic compounds can be found easily by means of Polya's Theorem. First, one has to take into account the cycle index of the molecule, depending on the automorphisms (symmetry operations involving proper axes of rotation). For PCBs, one has to consider all possible rotamers of the substituted biphenyl molecule. The cycle index leads to the figureure counting series, which is a polynomial in x whose degree equals the number ofhydrogen or halogen atoms in the molecule. Each subscript 'a' for x_a indicates the order of the permutation, and each superscript 'b' for xab the number of permutations or the number of unchanged positions. Then, according to Polya's Theorem, by performing the substitution:

$$x_a^b = (y^a + 1)^b$$

one ends up with the generating function, which is a polynomial in y whose coefficients indicate the numbers of isomers. For PCBs, the possible permutations are indicated in Table 2.1

Table 2.1: Permutations of the substituents in polychlorobiphenyls

1	2	3	4	5	6	7	8	9	10	
1	2	3	4	5	6	7	8	9	10	x^{10}
5	4	3	2	1	10	9	8	7	6	$x^2x_2^4$
1	2	3	4	5	10	9	8	7	6	$x^6x^2_2$
5	4	3	2	1	6	7	8	9	10	$x^6x^2_2$
6	7	8	9	10	1	2	3	4	5	x^5_2
10	9	8	7	6	5	4	3	2	1	x^5_2
6	7	8	9	10	5	4	3	2	1	$x_2x^2_4$
10	9	8	7	6	1	2	3	4	5	$x_2x^2_4$

Figure counting series: $(x^{10}+ 2x_2^5+ 2x^6x_2^2+ 2x^2x_4^2+ x^2x_2^4)/8$.

Generating function: $y^{10}+ 3y^9+ 12y^8+ 24y^7+ 42y^6+ 46y^5+ 42y^4+ 24y^3+ 13y^2+ 3y +1$.

Therefore, there are 3 possible monochlorobiphenyls, 12 dichlorobiphenyls, 24 tri-, 42 tetra-, 46 penta-, 42 tetrachlorobiphenyls, etc. Many of the 209 possible polychlorobiphenyls have been synthesized and characterized chemically, physically, and toxicologically. Of these, fourteen have a pronounced toxic effect, similar to that of tetrachlorodibenzodioxins(Safe *et al.,* 1990; Waller *et al.,* 1992). For PCDDs, the possible permutations are shown in Table 2.2

Figure counting series: $(x^8+ 3x_2^4)/4$.

Generating function: $y^8+ 2y^7+ 10y^6+ 14y^5+ 22y^4+ 14y^3+ 10y^2+ 2y + 1$.

Therefore, there are 2 possible monochlorodibenzodioxins, 10 dichloro-,14 tri-, 22 tetra-, 14 pentachlorodibenzodioxins, etc. Many of the 75 possible polychlorodibenzodioxins have been synthesized and characterized chemically, physically, and toxicologically. Of these, seven compounds have a high toxicity, and one of them is the most toxic (Waller *et al.,* 1992; Safe *et al.,* 1990; Mason *et al.,*1986; Tysklind *et al.,* 1992). This is the infamous and extremely toxic 2,3,7,8-tetrachlorodibenzo-1,4-dioxin or TCDD (LD_{50}= 45 µg/kg in rats), formed as a low-yield byproduct during the syntheses of 2,4,5-trichlorophenoxyacetic acid (herbicide and defoliating agent)and hexachlorophene (germicide) from chloroacetic acid and sodium 2,4,5-trichlorophenoxide. In the Seveso accident in Italy, a large amount of soil contaminated with 2,3,7,8-tetrachlorodibenzo-1,4-dioxin had to be removed at high cost. During the Vietnam War, the defoliant (Agent Orange) used by the US Army contained small amounts of 2,3,7,8-tetrachlorodibenzo-1,4-dioxin, and US veterans were compensated financially for the resulting health problems.For PCDFs, the possible permutations are presented in Table 2.3.

Table2.2: Permutations of the substituents in polychlorodibenzo-1,4-dioxins

1	2	3	4	5	6	7	8	
1	2	3	4	5	6	7	8	X_4^8
4	3	2	1	8	7	6	5	X_4^2
8	7	6	5	4	3	2	1	X_4^2
5	6	7	8	1	2	3	4	X_4^8

Table 2.3: The possible permutations for PCDFs

1	2	3	4	5	6	7	8	
1	2	3	4	5	6	7	8	X^8
8	7	6	5	4	3	2	1	X_2^4

Figure counting series: $(x^8 + x2^4)/2$.

Generating function: $y^8 + 4y^7 + 16y^6 + 28y^5 + 38y^4 + 28y^3 + 16y^2 + 4y + 1$.

Therefore, there are 4 possible monochlorodibenzofurans, 16 dichloro-28 tri-, 38 tetra-, 28 pentachlorodibenzofurans, etc.Many of the 135 possible polychlorodibenzofurans have been synthesized and characterized chemically, physically, and toxicologically. Ten of them have high toxicity (Waller *et al.,* 1992; Safe *et al.,* 1990; Tysklind *et al.,* 1992; Poiger *et al.,* 1989; Bandiera *et al.,* 1984; Mason *et al.,* 1985). Polyhalogenated biphenyls have been used in the past as dielectric liquids in electrical transformers, as heat transfer agents (along withpolyhalogenated diphenyl ethers), as lubricating agents, and as additives for pesticides or plastics. Since 1970, however, their use has been prohibited, but owing to their persistence they are still polluting the environment (Espandiari *et al.,* 2001; Pomarentz *et al.,* 1978; Lang *et al.,* 1992; Safe, 1994; Schafer *et al.,* 2002; Wood *et al.,* 1997;Erickson…; Oberg, 1997).

2.3 Toxicity of Polychlorinated Aromatic Compounds

Polychlorinated aromatic compounds have high hydrophobicity and persistence (Öberg*et al.,* 1997; Dimitrov*et al.,* 2002).They are,therefore widespread environmental contaminants, and they accumulate in fatty tissues of animals. The biological activityof polyhalogenated aromatic compounds is characterized by their multiple effects on mammals: hepatotoxicity, porphyria, dermal lesions including chloracne, endocrine effects, immunotoxicity, gastric lesions, lung edema, thymic atrophy, body weight loss, teratogenicity, reproductive lesions, and carcinogenicity (Safe, 1989; Silberhorn*et al.,* 1990; Silberhorn, 1982).

The in vitro induction of two enzymes, have been associated withthese compounds, namely: aryl hydrocarbon hydroxylase (AHH) and 7-ethoxyresorufin O-deethylase (EROD).

All these effects are believed to result via a specific protein complex, the intracellular cytosolic aryl hydrocarbon receptor (AhR) that is highly conserved in the evolutionary scale and acts similarly in humans and in laboratory animals. The receptor binding affinitydepends on steric, electrostatic, hydrophobic, hydrogen bonding factors, with dispersion forces and electrostatic potentials playing important parts (Bandiera*et al.,* 1984; Tuppurainen*et al.,*2000).

2.3.1 The concept of toxic equivalency factors

Owing to the fact that dioxin-like compounds normally exist in environmental and biological samples as complex mixtures of congeners, the concept of toxic equivalents (TEQs) has been introduced for risk assessment and regulation. In applying this concept, relative toxicities of dioxin-like compounds in relation to TCDD (i.e. toxic equivalency factors, TEFs) are determined from in vitroand in vivostudies. This approach is based on the evidence that there is a common, receptor-mediated mechanism of action for these compounds. However, it has limitations owing to a number of simplifications, the most important being that the combined toxic effects of the components of a given mixture are assumed to be additive, neglecting possible synergism, antagonism or differences in toxicokinetics. The criteria for including a compound in a TEF scheme have been discussed; the following criteria should be met.

I. It should show structural relationship to the PCDDs and PCDFs.

II. It should bind to the aryl hydrocarbon (Ah) receptor.

III. It should elicit dioxin-specific biochemical and toxic responses.

IV. It should be persistent and accumulate in the food-chain.

An international scheme of toxic equivalency factors (I-TEF) recommended by the North Atlantic Treaty Organization Committee on the Challenges of Modern Society and covering the 2,3,7,8-substituted PCDDs and PCDFs normally occurring in various compartments of theenvironment has been widely used. An international expert group, convened by WHO in 1997, revisited the existing TEFs for human risk assessment and recommended TEFs for wildlife

2.3.2 Evolution of the toxicity equivalence methodology

In the 1970s and 1980s, human health risk assessments of complex mixtures of PCDDs and PCDFs were generally performed including only 2,3,7,8-TCDD or assuming that all dioxin-like chemicals were equally potent to 2,3,7,8-TCDD (U.S. EPA, 1987, 1989). A review of the scientific information currently available clearly demonstrates that both of these assumptions were inaccurate. While many PCDD and PCDF congeners act through a common mechanism of action (binding and activation ofthe AHR) and induce similar biochemical and toxicological effects, the relative potency of individual dioxin-like chemicals to induce such effects has been shown to vary. The first use of a toxicity equivalence-like method for risk assessment purposes was described by Eadon*et al*. (1986) as a means to estimate potential human health risks associated with a PCB transformer fire in Binghamton, New York. In an examination of the initial human health risk assessment methodologies designed toaddress the emission of dioxins and furans from waste incinerators, EPA also concluded that TEFs were the best available interim scientific 4 policy for dealing with complex mixtures of these contaminants. Hence, in 1987, EPA adopted an interim procedure, based on TEFs, for estimating the hazard and dose-response of complex mixtures containing PCDDs and PCDFs in addition to 2,3,7,8-TCDD (U.S. EPA, 1987). Following adoption of the toxicity equivalence methodology in the United States

and Canada, the North Atlantic Treaty Organization/Committee on the Challenges of Modern Society (NATO/CCMS) examined the methodology and concluded that it was the best available interim method for PCDD/PCDF human health risk assessment (NATO, 1988a, b).

The TEFs proposed for the different dioxin-like chemicals were refined by the NATO/CCMS based on inclusion of more recent data sets, resulting in a greater number of the TEFs being based on toxicity observed in vivo. The NATO/CCMS panel assigned TEFs to octachlorinated dibenzo-p-dioxin (OCDD) and octachlorinated dibenzofuran (OCDF), and removed TEFs for all congeners lacking chlorine in the 2,3,7,8-positions. Although it was indicated that, theoretically, it may be possible to detect nearly all of the 210 PCDD/PCDF isomers inthe environment, only the seventeen 2,3,7,8-substituted PCDD and PCDF congeners were known to significantly bioaccumulate (Table 2.4). EPA officially adopted the revised TEFs in 1989 (TEFs-NATO89), with the caveat that the methodology remain interim and continued revisions be made (U.S. EPA, 1989; Kutz*et al.*,1990). The use of the toxicity equivalence methodology for human health risk assessment and risk management purposes has since been formally adopted by a number of other countries (e.g.,Canada, Germany, Italy, the Netherlands, Sweden, and the United Kingdom) (Yrjänheiki, 1992).

During the initial development of the toxicity equivalence methodology for PCDDs/PCDFs, a number of researchers were also examining the structure-activity relationships for PCBs(Leece*et al.,* 1985; Safe 1990; 1994). These studies revealed that only PCB congeners substituted in the meta and para-positions (Figure 2.4) were approximate stereo isomers of 2,3,7,8-TCDD and induced dioxin-like biochemical and toxicological effects. PCBs with a single chlorine substitution in an ortho-position on the biphenyl (mono-ortho) have diminished dioxin-like activity. In some classesorganism, (e.g.,fish), the reduction in dioxin-like activity is

substantial. However, in 1991, EPA convened a workshop to consider TEFs for PCBs (Barnes*et al.,* 1991; U.S. EPA, 1991).From the workshop it was concluded that a small subset of the PCBs displayed dioxin-like activity and met the criteria forinclusion in the methodology.It was also noted that the PCBs not included in the toxicity equivalence methodology (i.e.,the non-dioxin-like PCBs) are not a single class of chemicals and have multiple toxicities with separate structure-activity relationships (Barnes*et al.,* 1991).

In the years since the initial adoption of the toxicity equivalence methodology, additional data have accumulated on the toxicological potency of individual PCDDs, PCDFs, and PCBs relative to 2,3,7,8-TCDD. A joint project to harmonize toxicity equivalence methodologies for dioxin-like chemicals, conducted by the WHO-ECEH and the International Programme on Chemical Safety (IPCS), resulted in the development of a database consisting of all relevant toxicological data for dioxin-like chemicals available through 1993. Following a review of almost 1,200 peer-reviewed publications, 146 were selected and analyzed to derive TEFs for PCBs (TEFs-WHO94). Based on the reported results for 14 different biological and toxicological parameters from a total of 60 articles, a panel of experts from eight countries recommended interim TEFs for 13 dioxin-like PCBs (Ahlborg*et al.,* 1994). Application of this methodology in human health risk assessment was reaffirmed in EPA's Supplementary Guidance for Conducting Health Risk Assessment of Chemical Mixtures (U.S. EPA, 2000a). At a second WHO-ECEH consultation in 1997, the TEFs for PCDDs, PCDFs, and PCBs were reviewed and the toxicity equivalence methodology expanded, based on availability of additional data, to include class-specific TEFs for mammals, birds, and fish. TEFs for seven PCDD, 10 PCDF and 12 PCB congeners for mammals, birds, and fish (TEFs-WHO98; Table 2.5) were included in the resulting report (Van den Berg*et al.,* 1998).

Table 2.4: Number of polychlorinated dioxin, furan, and biphenyl congeners

Chemical Class	Number of Congeners	Dioxin-like Chemicals
Dioxins (PCDDs)	75	7
Furans (PCDFs)	135	10
Biphenyls (PCBs)	209	12

At the WHO-ECEH consultation, the TEFs previously assigned to PCB 170 and PCB 180 were withdrawn and a TEF for PCB 81 was established, such that the number of PCB congeners with TEFs assigned was reduced from 13 to 12 (Van den Berg*et al.,* 1998). It should be noted that the species and endpoints examined for assignment of TEFs varied among individual dioxin-like chemicals. Van den Berg*et al*. (1998) also provide greater documentation on how the expert panel at the WHOECEH consultation selected studies for consideration, derived relative potency factors from individual studies, and developed TEFs from the existing database. Although a number of uncertainties associated with the toxicity equivalence methodology have been identified (Van den Berg*et al.,* 1998), it was the conclusion of the WHO-ECEH consultation that an additive toxicity equivalence methodology remained the most appropriate risk assessment method for assessing complex mixtures of dioxin-like PCDDs, PCDFs, and PCBs. In 1998, EPA and DOI sponsored a meeting entitled: "Workshop on the Application of 2,3,7,8 -TCDD Toxicity Equivalency Factors to Fish and Wildlife." The major objective of the workshop was to address uncertainties associated with the use of the toxicity equivalence methodology in ecological risk assessment. Thirty-one experts from academia, government, industry, and environmental groups participated in the workshop. General conclusions regarding application of the toxicity equivalence methodology in ecological risk assessment included:

I. The toxicity equivalence methodology is technically appropriate for evaluating risks to fish, birds, and mammals associated with AHR agonists and it can support risk analyses beyond screening-level assessments.

II. The methodology entails less uncertainty and isless likely to underestimate risks than are methods based on single chemicals. Specifically, because the methodology takes into account the possible effects of the suite of dioxin-like chemicals found in

complex environmental mixtures, it is less likely to underestimate risk than methods based on only one of these chemicals (i.e.,2,3,7,8-TCDD). Further, because total PCBs in the environment can be comprised of many chemicals that vary in concentration and relative potency as AHR agonists, the toxicity equivalence methodology provides a means for accounting for their variable potency.

III. The uncertainties associated with using the methodology are not thought to be larger than other sources of uncertainty within the ecological risk assessment process (e.g.,dose-response assessment, exposure assessment, and risk characterization). For a thorough understanding of the technical issues discussed and conclusions drawn from the EPA/DOI workshop, refer to U.S. EPA (2001a). In 2005, the WHO International Programme on Chemical Safety held another expert meeting during which the 1998 mammalian TEFs-WHO98were re-evaluated. Preceding this meeting, a one-day public hearing was convened at which members of the expert panel discussed various aspects of the TEF concept with stakeholders and interested parties. The 2005 WHO reevaluation relied extensively on the mammalian TEF database recently published by Haws*et al*. (2006); however, the expert panel used all available RePs, including those from studies published since 1997, in making their assessments. Changes made to the mammalian TEFsWHO98are reflected in Table 2.5 and designated as TEFs-WHO05. This expert panel also concluded that additively, an important pre-requisite of the TEF concept, was further confirmed by recent in vivo mixture studies by Walker*et al.* (2005) and Van den Berg *et al*. (2006). In 2006, the National Research Council of the National Academies (NRC) re-affirmed the scientific basis and credibility of the use of the toxicity equivalence methodology in risk assessment.

Table 2.5: World Health Organization TEFs for mammals, birds, and fish

Congener	TEF		
	Mammals1	**Birds**	**Fish**
Dioxins			
2,3,7,8-TCDD	1	1	1
1,2,3,7,8-PeCDD	1	1	1
1,2,3,4,7,8-HxCDD	0.1	0.05	0.5
1,2,3,6,7,8-HxCDD	0.1	0.01	0.01
1,2,3,7,8,9-HxCDD	0.1	0.1	0.01
1,2,3,4,6,7,8-HpCDD	0.01	<0.001	0.001
OCDD	0.0003	0.0001	<0.0001
Furans			
2,3,7,8-TCDF	0.1	1	0.05
1,2,3,7,8-PeCDF	0.03	0.1	0.05
2,3,4,7,8-PeCDF	0.3	1	0.5
1,2,3,4,7,8-HxCDF	0.1	0.1	0.1
1,2,3,6,7,8-HxCDF	0.1	0.1	0.1
1,2,3,7,8,9-HxCDF	0.1	0,1	0.1
2,3,4,6,7,8-HxCDF	0.1	0.1	0.1
1,2,3,4,6,7,8-HpCDF	0.01	0.01	0.01
1,2,3,4,7,8,9-HpCDF	0.01	0.01	0.01
OCDF	0.0003	0.0001	<0.0001

Non-orthoPCBs			
3,3',4,4'-TCB (77)	0.0001	0.05	0.0001
3,4,4',5-TCB (81)	0.0003	0.1	0.0005
3,3',4,4',5-PeCB (126)	0.1	0.1	0.005
3,3',4,4',5,5'-HxCB (169)	0.03	0.001	0.00005
Mono-orthoPCBs			
2,3,3',4,4'-PeCB (105)	0.00003	0.0001	<0.000005
2,3,4,4',5-PeCB (114)	0.00003	0.0001	<0.000005
2,3',4,4',5-PeCB (118)	0.00003	0.00001	<0.000005
2',3,4,4',5-PeCB (123)	0.00003	0.00001	<0.000005
2,3,3',4,4',5-HxCB (156)	0.00003	0.0001	<0.000005
2,3,3',4,4',5'-HxCB (157)	0.00003	0.0001	<0.000005
2,3',4,4',5,5'-HxCB (167)	0.00003	0.00001	<0.000005
2,3,3',4,4',5,5'-HeCB (189)	0.00003	0.00001	<0.000005

As part of their evaluation of the draft Exposure and Health Reassessment of 2,3,7,8-TCDD and Related Compounds (U.S. EPA 2003a), the NRC reviewed EPA's use of the Toxicity Equivalence Methodology in assessing risks from dioxin-like compounds. The NRC concluded that, "...the toxic equivalency factor methodology provides a reasonable, scientifically justifiable, and widely accepted method to estimate the relative potency of DLCs" (DLC = dioxin-like compounds; NRC, 2006). In addition, the NRC Committee examined a number general of issues that have been raised regarding the assumptions underlying the use of the toxicity equivalence methodology. One important conclusion of the Committee that is particularly relevant to this document is that addressing the additively assumption. The Committee concluded that "from an overall perspective, this assumption appears valid, at least in the context of risk assessment" (NRC, 2006).

2.3.3 ToxicEffects of Polychlorinated Aromatic Compoundson Experimental animals

Most of the toxic and biochemical effects caused by PCDDs and PCDFs are probably mediated by the Ah receptor. TCDD is the prototype compound for structurally similar polychlorinated aromatic hydrocarbons that act through the Ah receptor mechanism (Whitlock *et al.,* 1993). The initial step by which TCDD causes biological effects is through binding stereospecifically and with high affinity to the cytosolic receptor. Endogenous ligands for the receptor have not been identified. After initial binding, the TCDD-receptor complex dimerizes with the Ah-receptor nuclear translocation protein (Arnt). The TCDD-receptor-Arnt complex binds to dioxin response elements (DREs) on DNA resulting in induction of target gene (e.g. CYP1A1) transcription. The Ah receptor binding is thought to be necessary but not sufficient to produce toxicity. The genes whose expression is increased or decreased as a result of TCDD

action, which are required for the tissue and species specific toxic responses have not been identified. Biological responses to TCDD and related compounds are observed in animal species expressing Ah receptor. The signs of toxicity caused by TCDD and related compounds that bind to the Ah receptor can include lethality, body weight loss, immune-suppression, oedema, dermal, developmental, reproductive, neurobehavioural, liver and endocrine toxicity, alterations in lipid metabolism and gluconeogenesis, modulation of responsiveness to hormones and growth factors,alterations in circulating levels of hormones, thymic atrophy, cancer, tumour promotion and induction of various enzyme activities including aryl hydrocarbon hydroxylase (AHH) and EROD activities that are associated with the induction of cytochrome P-4501A1 (Birnbaum, 1994).

2.3.4 Carcinogenic Effect of Polychlorinated Aromatic Compounds on Experimental animal

Many studies have been conducted to evaluate the carcinogenicity of TCDD in rodents. This information has been reviewed (WHO, 1989). TCDD has been shown to be carcinogenic in rats, mice and hamsters. Tumours have been demonstrated in several organs but predominantly in the liver. WHO (Ahlborg *et al.,* 1994). Concluded that the NOAEL in these studies was 1 ng/kg bw per day for the induction of hepatic adenomas and that, at this dose, the level of TCDD in the rat liver was 540 ng/kg on a wet weight basis. A mixture of 1,2,3,6,7,8- and 1,2,3,7,8,9-HxCDDs has also been demonstrated to cause liver tumours in rodents, albeit at higher dosages. Oral administration of 2,7-dichloro-dibenzo-p-dioxin to rats (110 weeks) and mice (90 weeks) did not lead to increases in tumour incidence (NTP US., 1979).

Several studies have demonstrated that TCDD is a very potent tumour promoter in two-stage promoter models (91–95). Several other PCDDs and PCDFs, e.g. 1,2,3,7,8-PeCDD, 2,3,4,7,8-PeCDF and 1,2,3,4,6,7,8-HpCDD have also been shown to be potent tumour promoters (Schrenk *et al.,* 1993). WHO concluded that the tumour-promoting action of TCDD showed a clear-cut threshold and also that, at lower doses, "protective" effects against tumour induction were observed. It should be noted that at these doses, other noncarcinogenic effects may occur. (Ahlborg *et al.,* 1992).

2.3.5 Toxicological Effects of Polychlorinated Aromatic Compoundsin Humans

Information on effects in humans predominantly comes from high exposure in industrial settings or accidents and intoxication incidents, e.g. the contamination of rice oil with PCBs that contained comparatively high concentrations of PCDFs in Japan and Taiwan in 1968 and 1979, respectively. The effects observed are generally in agreement with what could be expected from animal experimentation, although the most typical human effect of high exposure to PCDDs and PCDFs seems to be the dermatological response chloracne. Subtle endocrinological effects, e.g. modulation of thyroid hormone and testosterone levels in plasma, and decreased glucose tolerance and subtle neurological effects have been observed in humans, but their clinical significance is currently unknown (Sweeney *et al.,* 1992).

2.3.6 Carcinogenic Effects Polychlorinated Aromatic Compounds in Human

Numerous epidemiological studies have been performed on occupationally or accidentally highly exposed groups. WHO concluded that there is consistent evidence that

workers, who were involved in high-level exposure, demonstrated an increased risk of cancer, although the statistical power of these studies is low (Ahlborg *et al.,*1992). Data on high-dose exposure (exhibiting chloracne) were found to demonstrate that for those workers who were observed for a period of 20 years after exposure and beyond, there was a small increased risk (standardized mortality ratio (SMR) 2.01) for the category of "all cancers combined". Several recent epidemiological studies support the association between high levels of dioxin exposure and cancer (Kogevinas *et al.,* 1993). The International Agency for Research on Cancer evaluated all available evidence in 1997 and classified 2,3,7,8-TCDD as a human carcinogen (Group 1). Other polychlorinated dibenzo-p-dioxins were not classifiable as to their carcinogenicity for humans (IARC, 1997).

2.3.7 Population Groups at Higher Exposures

PCDDs and PCDFs are found in human breast milk at comparatively high concentrations. The infant receiving breast milk as the only or dominating food source will consequently have a much higher intake than adults. Measurements of the concentrations of 2,3,7,8-substituted congeners in human milk samples in the WHO-coordinated second round of exposure studies on human milk, in which 19 countries participated, showed that the lowest TEQ levels were found in Albania, Hungary, Pakistan and the less industrialized regions in Croatia, Norway and the Russian Federation (4–10 ng/kg of milk fat) (WHO, 1996). The highest TEQ levels, around 20–30 ng/kg of milk fat, were found in samples from Belgium, Canada, Finland, the Netherlands and Spain. This would correspond to a daily TEQ intake ranging from about 20 up to about 130 pg/kg bw. These figures are on average about 35% lower than those that were found in the first WHO exposure study (WHO, 1989), although such a decrease could not be observed in all

countries. When the contribution from dioxin-like PCBs is taken into account, the intake for the breastfed infant is about 50–100% higher. However this increase is highly variable from country to country. The factors determining the levels found in humans include both lifestyle and the degree of industrialization. Monitoring programs have indicated a declining trend in PCDD/PCDF concentrations in human breast milk during the last 5–15 years, probably reflecting the relatively recent emission control measures (Alder *et al.,* 1994).

2.3.8 Critical Organs, Tissues and Effects

The toxic potency of PCDDs and PCDFs is a function of the number and positions of chlorine substituents. PCDDs and PCDFs with a decreased number of lateral chlorine atoms (positions 2, 3, 7 and 8) or increased non-lateral chlorine atoms are less toxic than TCDD (Safe, 1990). However, most of these 2,3,7,8-substituted congeners are toxic and the pattern of the responses within animals of the same species, strain, sex and age will be similar to that of TCDD (McConnell *et al.,* 1979).TCDD exposure mainly affects epithelial tissues (Poland *et al.,* 1982). Lesions vary between tissues and include proliferative as well as metaplastic and atrophic changes. PCDDs, PCDFs and related compounds cause a number of immune responses and are clearly immunotoxic in experimental animals including non-human primates. The information on effects on the immune system in humans is limited, and a definite immunotoxic effect of these compounds in humans has not been shown. Animal data show that, in susceptible species, the fetus and young animal are especially sensitive to the immunotoxic effect. PCDDs and PCDFs may be immunotoxic at some levels in humans, the impact on the immune system and implications for characterizing risks are largely unknown at present. Exposure to TCDD during pregnancy causes prenatal mortality in primates and rodents. Gestational TCDD exposure has

been shown to cause subcutaneous oedema, gastrointestinal haemorrhage and decreased fetal growth.

PCDDs and PCDFs can produce developmental effects in the offspring of experimental animals following in uteroand lactational exposure, and these effects, which may persist into adulthood, include neurobehavioural effects, developmental reproductive effects and neurochemical effects (Gray *et al.,* 1997). Although the dose–response data available to date are limited, the LOAELs, based on TEQ body burden in monkeys and rodents are close to the background human body burden. WHO estimated human daily intakes corresponding with body burdens similar to those associated with LOAELs for the most sensitive responses in animal. These intakes ranged from 14–37 pg/kg bw/day (Van Leeuwen *et al.,* 2000).

2.4 Quantitative Structure Activity Relationship (QSAR)

The discovery of relationships among different concepts, in particular concepts provided by different scientific fields, represents the most important way to develop new scientific knowledge and transform isolated information into a deeper theoretical knowledge. The concepts of molecular structure, its representation by theoretical molecular descriptors, and its relationship with experimental properties of molecules are an interdisciplinary network, where a lot of theories, knowledge, and methodologies and their interrelationships are present, leading to a new scientific research field with a relevant follow-up in several practical applications

2.4.1 Short History of QSAR and Molecular Descriptors

The history of QSAR and molecular descriptors is closely related to the history of what can be considered one of the most important scientific concepts of the last part of the nineteenth century and the whole of twentieth century, that is, the concept of molecular structure.

The years between 1860 and 1880 were characterized by a strong dispute about the concept of molecular structure, arising from the studies on substances showing optical isomerism and the studies of Kekule (1861–67) on the structure of benzene. The concept of the molecule thought of as a three-dimensional (3D) body was first proposed by Butlerov (1861–65), Wislicenus (1869–73), Van't Hoff (1874–75), and Le Bel (1874). The publication in French of the revised edition ofLa chimie dans l'e´space by Van't Hoff in 1875 is considered a milestone of the 3D conception of the chemical structures.

QSAR history started a century earlier than the history of molecular descriptors, being closely related to the development of the molecular structure theories. QSAR modeling was born in toxicology field. Attempts to quantify relationships between chemical structure and acute toxic potency have been part of the toxicological literature for more than 100 years. In the defense of his thesis entitled 'Action de l'alcohol amylique sur l'organisme' at the Faculty of Medicine, University of Strasbourg, France, on 9 January 1863, Cros noted that a relationship existed between the toxicity of primary aliphatic alcohols and their water solubility. Chemometrics in QSAR 131 Author's personal copyrelationship demonstrated the central axiom of structure–toxicity modeling, that is, the toxicity of substances isgoverned by their properties, which are determined in turn by their chemical structure. Therefore, there areinterrelationships among structure, properties, and toxicity. Crum-Brown and Fraser (1868–69)

Crum-Brown(1864) andCrum-Brown(1866) proposed the existence of a correlation between biological activity of different alkaloids and their molecular constitution. More specifically, the physiological action of a substance in a certain biological system (∅) was defined as a function (f) of its chemical constitution (C):

$$(\emptyset) = f(C)$$

Thus, an alteration in chemical constitution, ΔC, would be reflected by an effect on biological activity Δ∅.This equation can be considered the first general formulation of a QSAR.

A few years later, a hypothesis on the existence of correlations between molecular structure and physicochemical properties was reported in the work of Ko¨rner (1874), (Korner,1874) which dealt with the synthesis of disubstituted benzenes and the discovery of ortho, meta, and para derivatives. The different colors of disubstituted benzenes were thought to be related to their differences in molecular structure. Ten years later, Mills (1884) (Mills,1884) published a study 'On Melting Point and Boiling Point as Related to Composition' in the Philosophical Magazine.

The quantitative property–activity models, commonly referred to mark the beginning of systematic QSAR/QSPR studies, (Richet, 1893) have come out from the search for relationships between the potency of local anesthetics and the oil/water partition coefficient between narcosis and chain length, and narcosis and surface tension (Traube, 1904). In particular, the concepts developed by Meyer and Overton are often referred to as the Meyer–Overton theory of narcotic action (Meyer, 1899; Overton, 1901; Smolenskii,1964; Spialter, 1964; Balaban, 1968).

The first theoretical QSAR/QSPR approaches date back to the end of 1940s and are those that relatebiological activities and physicochemical properties to theoretical numerical indices derived from the molecular structure (Wiener, 1947).

On the basis of the graph theory, the Wiener index and the Platt number, (Platt, 1947) proposed in 1947 to model the boiling point of hydrocarbons, were the first theoretical molecular descriptors based on the graph theory. In the early 1960s, new molecular descriptors were proposed, giving the start to systematic studies on the molecular descriptors, mainly based on the graph theory (Fujita*et al.*, 1964; Gordon*et al.*, 1964).

The use of quantum-chemical descriptors in QSAR studies date back to early 1970s, (Kier, 1971) although quantumchemical descriptors were defined and used a long time before in the framework of quantum-chemistry. During 1930–60, the milestones were the works of Pauling (Pauling, 1932; Pauling, 1937) and Coulson (Coulson, 1939) on the chemical bond, (Sanderson, 1952) on electronegativity, and (Fukui et al. 1994) and (Mulliken, 1995) on electronic distribution.

Once the concept of molecular structure was definitively consolidated by the successes of quantum chemistry theories and the approaches to the calculation of numerical indices encoding molecular structure information were accepted, all the constitutive elements for the take-off of QSAR strategies were available. Based on the Hammett equation, (Hammett, 1935; Hammett, 1937), the seminal work of Hammett gave rise to the $'\sigma - \rho'$ culture in the delineation of substituent effects on organic reactions, whose aim was the search for linear free energy relationships (LFERs): steric, electronic, and hydrophobic constants were defined, becoming a basic tool for modeling properties of molecules (Hammett, 1938).

In the 1950s, the fundamental works of Taft (Taft, 1952; Taft, 1953; Taft, 1953) in physical organic chemistry were the foundation of relationships between physicochemical properties and solute–solvent interaction energies (linear solvation energy relationships, LSERs), based on steric, polar, and resonance parameters for substituent groups in congeneric

compounds. In the mid-1960s, led by the pioneering works of Hansch, (Fujita*et al.,* 1964; Hansch*et al.,* 1962; Hansch*et al.,* 1963) the QSAR/QSPR approach began to assume its modern look. In 1962, (Hansch *et al.,* 1992) published their study on the structure–activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity. Using the octanol/water system, a whole series of partition coefficients were measured, and thus, a new hydrophobic scale was introduced for describing the attitude of molecules to move through environments characterized by different degrees of hydrophilicity such as blood and cellular membranes. The delineation of Hansch models led to explosive development in QSAR analysis and related approaches (Hansch*et al.,* 1995). In the same years, Free and Wilson developed a model of additive substituent contributions to biological activities, giving a further push to the development of QSAR strategies (Free*et al.,* 1964).

They proposed to model a biological response on the basis of the presence/absence of substituent groups on a common molecular skeleton (Free*et al.,* 1964). This approach, called ‘de novo approach’ when presented in 1964, was based on the assumption that each substituent gives an additive and constant effect to the biological activity regardless of the other substituents in the rest of the molecule (Kubinyi*et al.,* 1988).

2.4.2 Chemometrics and QSAR modeling

The development of QSAR/QSPR models is a quite complex process, as outlined inFigure 2.1. Once theresearch goal has been clearly defined, which in most cases means defining the property to be modeled, that is, the endpoint, the decision to be made concerns how much general the final model should be. This entails the selection of the set of molecules the modeling procedure is applied to. For a long time, QSAR models were developed on sets of

congeneric compounds, that is, molecules with a common parental structure and different substituent groups. Later, the interest in producing tools for quick molecular property estimations moved forward more general QSAR models suitable for diverse molecules belonging to different chemical classes, that is, not congeneric sets. The final decision in defining the molecule set mainly depends on the foreseen use of the model and availability of experimental data.

In this phase of the QSAR process, it is of primary concern to gain an exhaustive knowledge about the compounds in analysis with specific regard to the endpoint of interest. This obviously implies the acquisition of reliable experimental data regarding the endpoint and possibly already existing models. Data of the chemicals can be produced experimentally or retrieved from literature. In both cases, accuracy should be carefully evaluated: the limiting factor in the development of QSAR/QSPR models is the availability of high quality experimental data, because the accuracy of the property estimated by a model cannot exceed the degree of accuracy of the input data. Moreover, when data are collected from literature, to avoid an additional variability into the data because of different sources of information, data should be taken just from one source or from almost comparable sources. Another important phase of the QSAR process is the definition of a reliable chemical space; in other words, the selection of those structural features is thought to be the most responsible for modeling the endpoint in analysis.This implies the selection of proper molecular descriptors but, in most cases, there is noa priori knowledge about which molecular descriptors are the best. Then, the tendency is to use a huge number of descriptors, which hopefully include the candidate variables for modeling, and later apply a variable selection technique. Two basic strategies can be adopted: (1) the use of algorithms to select the optimal subset(s) of descriptors and (2) the use of chemometric methods

(e.g., principal component analysis (PCA) or partial least squares (PLS)) able to condense the large amount of available chemical information into a few principal variables.

The next step is the selection of the validation procedure, which, in addition to the fitting performance of the model, allows the evaluation of the model prediction ability. The latter is usually considered the most important characteristic for an acceptable QSAR model. The predictive ability of the models is evaluated by dividing the compounds into the training set, that is, the set by which the model is calculated, and the test set, that is, the set of compounds by which the model predictive ability is evaluated. The partition into training/test sets is performed in different ways, depending on the validation procedure.

Exploratory data analysis is a common preliminary step in all the QSAR/QSPR studies. In particular, PCAand clustering methods (both hierarchical and nonhierarchical) are the most commonly used. A wide importance has been gaining in these last years by the clustering approach based on the Kohonen maps (or selforganizing maps, SOMs), which is an artificial 2D neural network providing easy interpretable information about similarity/diversity among objects (Kohonen, 1989).

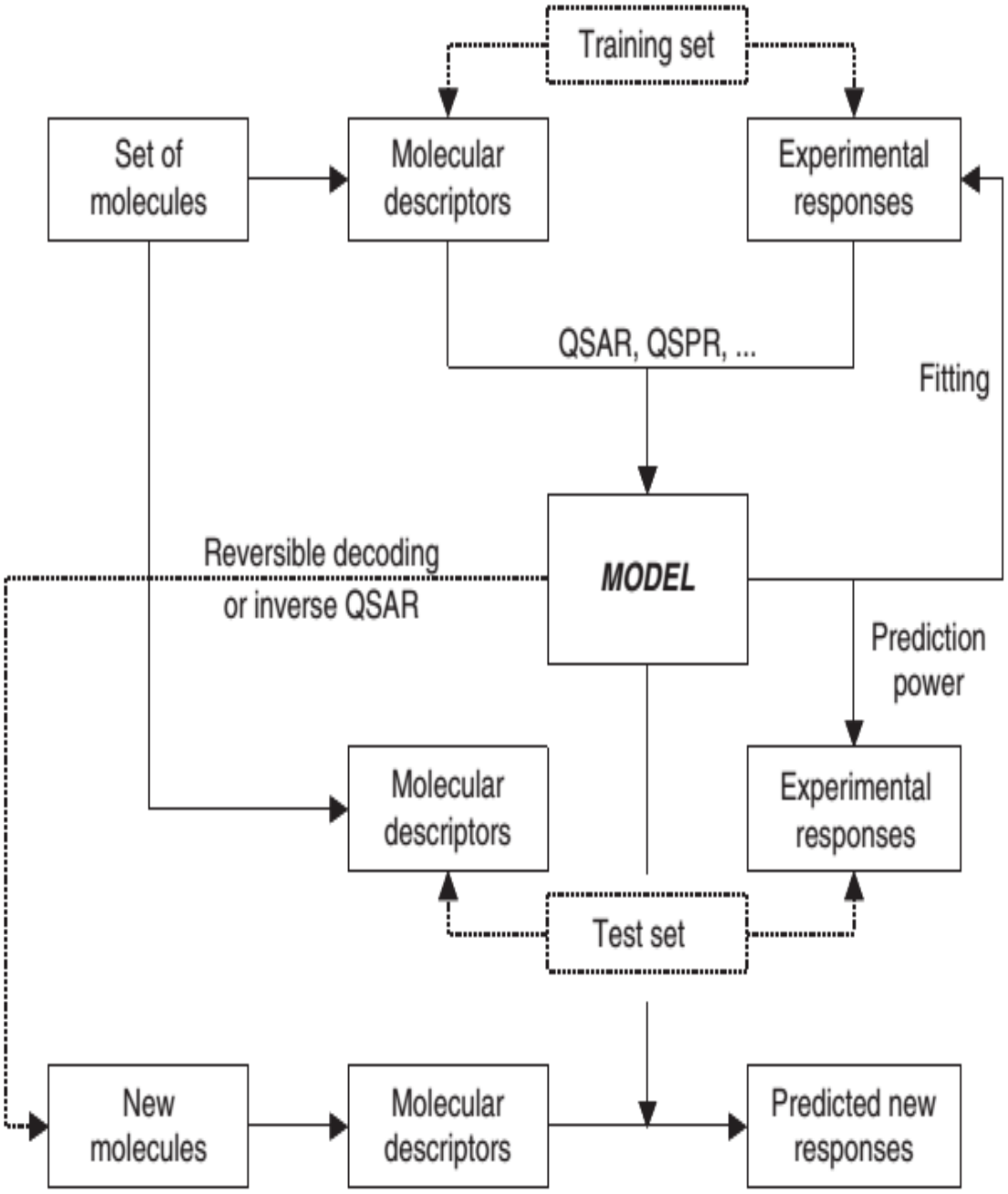


Figure 2.1 General scheme of the QSAR/QSPR strategy.

By exploratory analysis, the QSAR expert can evaluate whether the chosen molecular descriptors are suitable for describing the compounds in analysis and the chemical space is sufficiently represented. Moreover, the tendency observed nowadays is to build a reference chemical space for large categories ofchemicals for which molecular properties are known by using methods such as PCA on molecular fingerprints. Then, this chemical space is used to analyze similarities among groups of chemicals showing, for example, groups of different biological activity, and to find which regions in the chemical space require to be more explored by designing new molecular structures (Zupan*et al.*, 1995).

The majority of the QSAR strategies aimed at building models are based on regression and classification methods, depending on the studied problem. For continuous properties, the typical QSAR/QSPR model is defined as

$$P = f(x_1, x_2 \dots \dots \dots \; x_P)$$

Where *P* is the molecular property/activity,$x_1 x_P$, are the molecular descriptors, and f is a function representing the relationship between response and descriptors. In most of the cases, the function f is not a prioriknown and needs to be estimated (Todeschini*et al.*, 2009)

Ordinary least squares (OLS) regression, also called multiple linear regression (MLR), is the most common regression technique used to estimate the quantitative relationship between molecular descriptors and the property. PLS regression is widely applied especially when there are a large number of molecular descriptorswith respect to the number of training compounds, as it happens for 4D-QSAR methods such as GRID andCoMFA (Todeschini*et al.*, 2009).

Several other methods play a fundamental role in QSARs such as principal component regression (PCR), k-nearest neighbor (k-NN) regression, and stepwise regression (SWR), the last

being the most applied method to select a descriptor subset from a not too large set of candidate variables.

Regression techniques based on the artificial neural networks (ANNs) are also frequently used, (Livingstone, 1992) Such as backward propagation (BP) and radial basis functions (RBFs), and on ensemble approaches such as random forests (Breiman*et al.,* 2001; Svetnik*etal*., 2003)

2.4.3 Specific QSAR approaches

There are a lot of QSAR approaches in the literature, often reported with a large variety of names, which make difficult to rationalize them into a well-defined classification system. An attempt to classify QSAR approaches might be made by considering the objective to be reached by a QSAR approach, the type of molecular property is modeled on, the type of molecular descriptors the model is composed of, and the mathematical method or computational algorithm used to estimate the model parameters. Therefore, focusing on the objective of the analysis, it is possible to distinguish, for example, among drug design, high-throughput screening, and molecular similarity analysis. Paying more attention to the property, terms such as ADME (absorption, distribution, metabolism, and elimination properties) analysis, environmental QSAR, LSERs, and binary-QSAR are commonly used. 2D-QSAR, 3D-QSAR, and 4D-QSAR namely refer to the type of molecular descriptors they are based on. Terms such as group contribution method (GCM), structural analysis, and grid-based QSAR technique mainly derive from the specific applied methods. In the following, some QSAR approaches are explained more in detail (Todeschini*et al.,* 2009).

Hanschapproach

There is a consensus among current predictive toxicologists that Corwin Hansch is the founder ofmodern QSAR. In the classic article it was illustrated that, in general, biological activity for a group of 'congeneric' chemicals can be described by a comprehensive model:

$$\log\left(\frac{1}{C_{50}}\right) = a\pi + b\varepsilon + cS + d \ldots\ldots\ldots\ldots\ldots (2.2)$$

whereC, the toxicant concentration at which an endpoint is manifested (e.g., 50% mortality or effect), is related to a hydrophobicity term, π an electronic term, ε(originally the Hammett substituent constant, σ), and a steric term,S(typically Taft's substituent constant, Es), beingda general additional term depending on the kind of property to be modeled (Todeschini*et al.*, 2009).

In particular, the parameter,π which is the relative hydrophobicity of a substituent, was defined as

$$\pi = logP_x - logP_H$$

Where$P_x$and$P_H$represent the partition coefficients of a derivative and the parent molecule, respectively. This is a substituent constant denoting the difference in hydrophobicity between a parent compound and a substituted analog and is usually replaced by the more general molecular term the log of the 1 – octanol/ water partition coefficient, $logK_{ow}$ or $logP$

The analysis was performed on the following descriptors: logP, the octanol/water partition coefficient; the Hammett electronic constant of Chemometrics in QSAR Author's personal copy the substituent;Ip, an indicator variable that takes the value 1 for p-derivatives and σ for the other compounds; Es, the Taft steric constant for o-derivatives; andR, the electronic parameter for o-derivatives (Balaban, 1982). The Hansch QSAR models with four, five, and six independent variables are

$$\log\left(\frac{1}{ED_{50}}\right) = 2.280 + 0.265(logP)^2 + 1.222(logP) - 0.161\sigma - 0.079I_P$$

$$n = 30 \quad R^2 = 0.490 \quad S = 0.228 \quad F = 5.99$$

$$\log\left(\frac{1}{ED_{50}}\right) = 2.311 + 0.290(logP)^2 + 1.309(logP) - 0.135\sigma - 0.057I_P + 0.404E_0$$

$$n = 30 \quad R^2 = 0.640 \quad S = 0.195 \quad F = 10.05$$

$$\log\left(\frac{1}{ED_{50}}\right) = 2.478 + 0.276(logP)^2 + 1.229(\text{logP}) - 0.353\,\sigma - 0.223I_P + 0.278E_s + 0.621R$$

$$n = 30 \quad R^2 = 0.731 \quad S = 0.172 \quad F = 7.83$$

Note that these models contain a nonlinear term $(logP)^2$ and their predictive ability is not considered. The contributions of Hammett and Taft together laid the basis for the development of the QSAR paradigm by Hansch and Fujita, which combined the hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms(Tarko *et al.*, 2001).

Free–Wilson approach

The Free–Wilson approachis based on the assumption that a biological response can be modeled by additivesubstituent contributions, that is, the substituent effects are considered independent of each other, thecompound's congenericity being also another basic requirement(Free*et al.*, 1964).Once a common skeleton for the chemical analogs is defined, regression analysis is performed, considering a numbers of substitution sites Rs (s = 1,S), and for each site a numberNs of different substituents. Hydrogen atoms are also considered as substituents if present in a substitution site of some compounds. The Free–Wilson descriptors of theith compound are indicator variables$I_{i,k,s}$ where$I_{i,k,s}$ =1 ifthe kth substituent is present in the sth site and$I_{i,k,s}$ =0 otherwise.

The Free–Wilson model is defined as

$$\widehat{y_i} = b_0 + \sum_{s=1}^{s} \sum_{K=1}^{Ns} b_{ks} I_{i,k,s}$$

whereb_0 is the intercept of the model corresponding to the average biological response calculated from the data set and b_{ks} are the regression coefficients. The biological responseyis usually used in the form log (1/C), whereC is the concentration achieving a fixed effect. The regression coefficients b_{ks} of the Free–Wilson model give the importance of each kth substituent in each sth site in increasing/decreasing the response with respect to the mean response, that is, the activity contribution of the substituent(Tarko *et al.*, 2001).

Read-Across approach

Recently adopted to feel in gaps on data of chemicals, read-acrossis a nonformalized approach in which endpoint information for one chemical (called a 'source chemical') is used to

make a prediction of the endpoint for another chemical (called a 'target chemical'), which is considered to be similar in some way (usually on the basis of structural similarity). In principle, read-across can be applied to characterize physicochemical properties, fate, human health effects, and ecotoxicity, and it may be performed in a qualitative or quantitative manner (Worth*et al.*, 2008).

2.4.3 Molecular descriptors

In the last decades, several scientific researchers have been focused on studying how to catch and convert – by a theoretical pathway – the information encoded in the molecular structure into one or more numbers used to establish quantitative relationships between structures and properties, biological activities or other experimental properties. Molecular descriptors are formally mathematical representations of a molecule obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure: ''The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment'' (Todeschini*et al.*, 2000).

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, toxicology, ecotoxicology, health research, and quality control. Evidence of the interest of the scientific community in the molecular descriptors is provided by the huge number of descriptors proposed up today: more than 3000 descriptors 112 derived from different theories and approaches are actually defined and computable by using dedicated software tools (Todeschini*et al.*, 2000).

Each molecular descriptor takes into account a small part of the whole chemical information contained in the real molecule, and, as a consequence, the number of descriptors is continuously increasing with the increasing request of deeper investigations on chemical and biological systems. Different descriptors are different ways or perspectives to view a molecule, taking into account the various featuresof its chemical structure. By now molecular descriptors have become one among the most important variables used inmolecular modeling, and, consequently, managed bystatistics, chemometrics, and chemoinformatics(Tarko *et al.*, 2001).

The availability of the molecular descriptors has not only been a new opportunity to search for new relationships but also been a great change of the research paradigm in this field: in effect, the use of the molecular descriptors – calculated by theories – has permitted for the first time to link experimental knowledge to theoretical information arising from the molecule structure. While until 1960s–70s molecular modeling mainly consisted in searching for mathematical relationships between experimentally measured quantities, nowadays it is mainly performed searching for relationships between a measured property and molecular descriptors able to catch structural chemical information. A general consideration about the use of molecular descriptors in modeling problems concerns their information content. This depends on the kind of molecular representation used and the defined algorithmfor its calculation. There are simple molecular descriptors derived by counting some atom types or structural fragments in the molecule, as well as physicochemical and bulk properties such as molecular weight, number of hydrogen bond donors/acceptors, and number of OH-groups.Other molecular descriptors are derived from algorithms applied to a topological representation and usually called topological or 2D descriptors. Other molecular descriptors are derived from the spatial (x, y, z) coordinates of the molecule, usually called geometrical or 3D descriptors; another class of molecular

descriptors, called 4D descriptors, is derived from the interaction energies between the molecule, imbedded into a grid, and some probe(Tarko *et al.*, 2001).

InFigure 2.2, a (very) simplified scheme of the major classes of molecular descriptors is shown.It is true that geometrical 3D/4D descriptors have higher information content than other simpler descriptors, such as counting descriptors or topological descriptors, which often show relevant levels ofdegeneracy. Then, several people think that it is better to use the most informative descriptors in all modeling processes. This thinking is incorrect because the 'best descriptors' are those whose information content is comparable with the information content of the response for which the model is searched for. In effect, too high information in the independent variables (the descriptors) with respect to the response is often seen as noise on behalf of the model, thus giving instable or not predictive models. For example, a property whose values are equal or similar for isomeric structures is better modeled by a simple descriptor with degenerate values for isomeric structures. In this case, descriptors able to discriminate among the isomeric structures have redundant information which cannot be integrated in the model. In conclusion, it can be stated that the best descriptor(s) valid for all the problems does not exist.

In general, molecular descriptors, besides the trivial invariance properties, should satisfy some basic requirements. List of desirable requirements of chemical descriptors suggested by Randic (Randic, 1996).

A lot of software calculates wide sets of different theoretical descriptors, from SMILES, 2D graphs to 3D x, y, zspatial coordinates. Some of the most popular software are mentioned here: ADAPT, 114 OASIS, (Mekenyan*et al.*,1990).

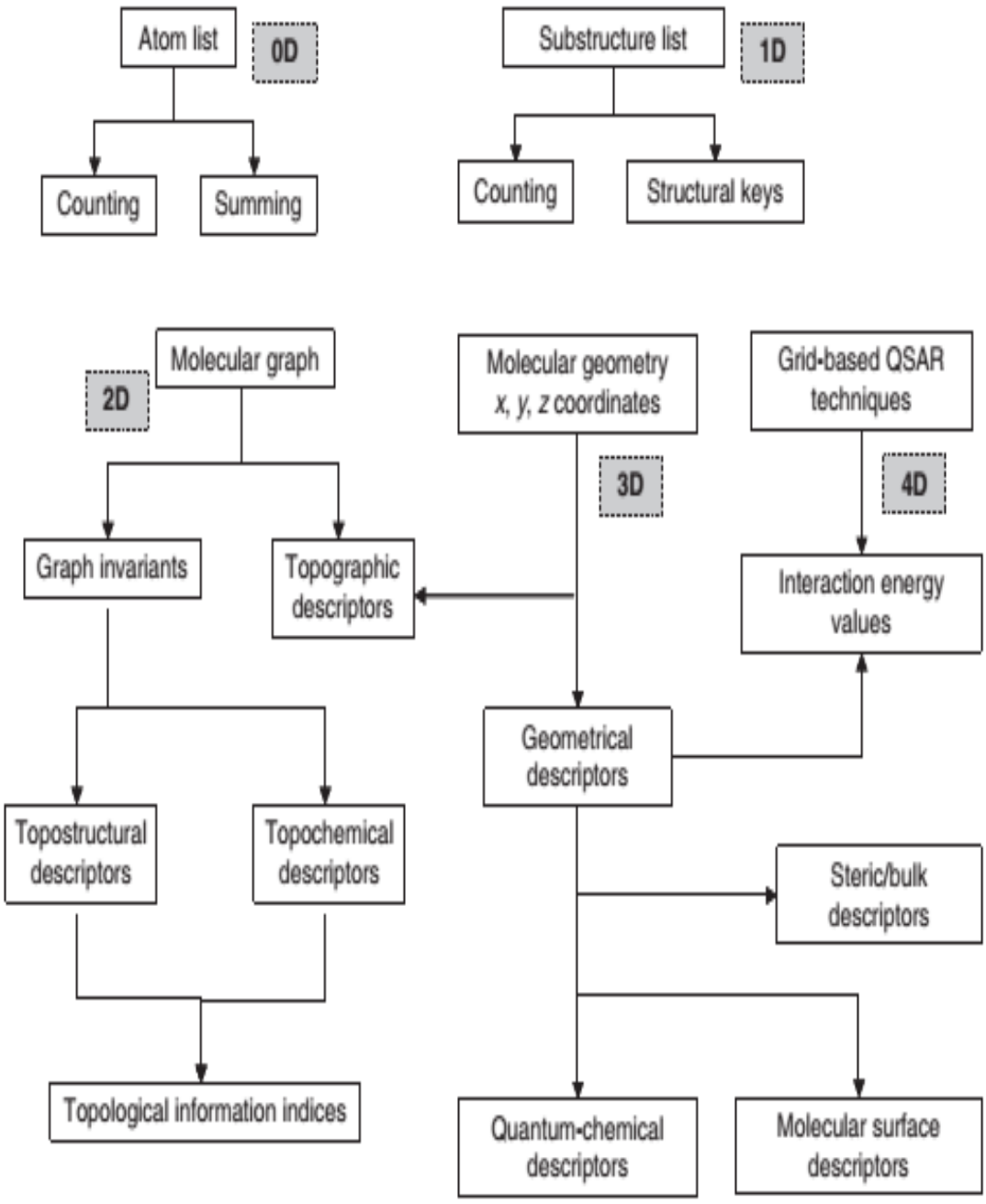


Figure2.2: General scheme of the different sources of molecular descriptors.

0D Descriptors or Count Descriptors

All the molecular descriptors for which no information about molecular structure and atom connectivity is needed belong to the class of 0D descriptors. Atom and bond counts, as well as sum or average of the atomic properties are typical of this class of descriptors. These descriptors can be always easily calculated, are naturally interpreted, do not require optimization of the molecular structure, and are independent of any conformationalproblem.

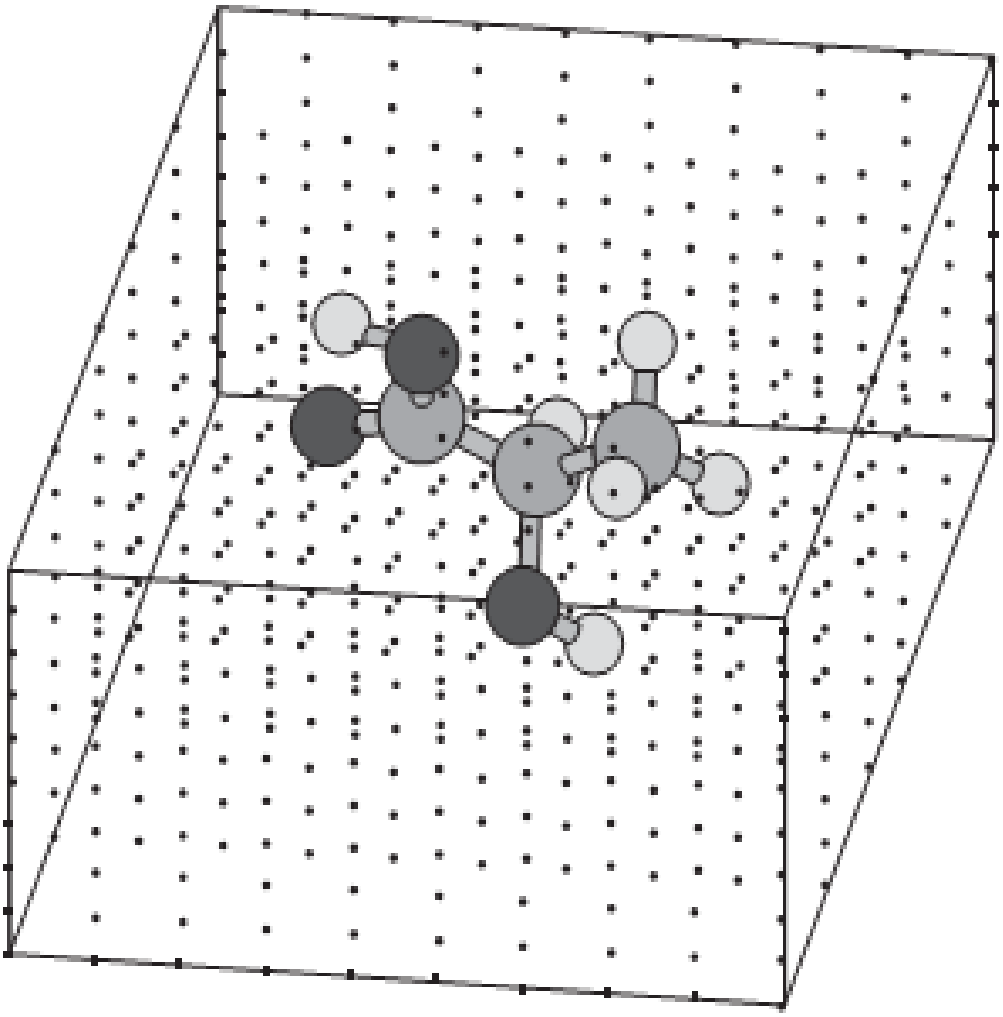


Figure. 2.3:A lattice of grid point with an embedded molecule.

They usually show a very high degeneration, that is, they have equal values for several molecules,such as isomers. Their information content is low, but nevertheless they can play an important role in modelingseveral physicochemical properties or take a part into more complex models(Tarko *et al.*, 2001).

2.4.3.2: 1D Descriptors or fingerprints

All the molecular descriptors that can be calculated from sub-structural information about the molecule belongto the 1D descriptors. Counting of functional groups and substructure fragments, as well as atom-centered descriptors, are the most known 1D descriptors. These descriptors are often presented as fingerprint, that is, a binary vector where 1 indicates the presence of the defined sub-structure and 0 its absence. A relevant advantage in describing molecules by fingerprints is the possibility to perform quick calculations for molecule similarity/ diversity problems.Like 0D descriptors, these descriptors can be usually easily calculated, are naturally interpreted, do not require optimization of the molecular structure, and are independent of any conformational problem. They usually show a medium-high degeneration and are often very useful in modeling both physicochemical and biological properties (Tarko *et al.*, 2001).

2.4.3.32D Descriptors or topological descriptors

TIs are molecular descriptors based on a graph representation of the molecule and represent graph–theoretical properties that are preserved by isomorphism, that is, properties with identical values for isomorphic graphs. A graph invariant may be a characteristic polynomial, a sequence of numbers, or a single numerical index obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. TIs are usually derived from a H-depleted molecular graph. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry,

branching, and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. In fact, TIs are usually divided into two categories: topostructural and topochemical indices.Topostructural indices encode only information on the adjacency and distance of atoms in the molecular structure; topochemical indices quantify information on topology but also specific chemical properties of atoms such as their chemical identity and hybridization state (Basak*et al.,* 1997).

Topological information indices are graph invariants, based on information theory and calculated as information content of specified equivalence relationships on the molecular graph. In general, TIs do not uniquely characterize molecular topology; different structures may have some of the same TIs. A consequence of TIs nonuniqueness is that they do not, in general, allow reconstructing molecule. There are several ways to obtain topological descriptors. Simple TIs consist in the counting of some specific graph elements; examples are the Hosoya Z index, (Hosoya1997) path counts, (Randic*et al.,* 1979)walk counts, self-returning walk counts, (Harary, 1969)Kier shape descriptors, (Kier, 1985) path/walk shape indices (Randic, 2001).However, the most common TIs are derived by applying some algebraic operators (e.g., the Wiener operator) to a matrix representation of the molecular structure, such as adjacency and distance matrices. Among them are the Wiener index, (Wiener, 1947)spectral indices, (Ivanciuc*et al.,* 1999) and Harary indices (Ivanciuc*et al.*, 1993).

Molecular matrices are the most common mathematical tool to encode structural information of molecules. Very popular molecular matrices are the graph–theoretical matrices, a huge number of which were proposed in the last decades in order to derive TIs and describe molecules from a topological point of view. Graph– theoretical matrices are matrices derived

from a molecular graphG(often froma H-depleted molecular graph). A comprehensive collection of graph–theoretical matrices is reported by (Janezˇic *et al.*, 2007).

Vertex matrices are undoubtedly the graph–theoretical matrices most frequently used for characterizing a molecular graph. The matrix entries encode different information about pairs of vertices such as their connectivities, topological distances, sums of the weights of the atoms along the connecting paths; the diagonal entries can encodechemical information about the vertices. From vertex matrices a huge number of TIs were proposed.Other topological molecular descriptors can be obtained by using suitable functions applied to local vertexinvariants (LOVIs), these being numerical representations of the atoms derived from molecular graphs. The most common functions are atom and/or bond additives, resulting in descriptors that correlate well physicochemical properties that are atom and/or bond additives themselves. For example, Zagreb indices, (Gutman*et al.*, 1972) Randic connectivity index, (Randic, 1975)related higher-order connectivity indices, (Kier, 1977)and Balaban distance connectivity indicesare derived according to this approach (Balaban, 1982).

Particular TIs are derived from weighted molecular graphs where vertices and/or edges are weighted by quantities representing some 3D features of the molecule, like those obtained by computational chemistry. The graph invariants obtained in this way encode both information on molecular topology and molecular geometry. BCUT descriptorsare an example of these topological descriptors. Graph invariants have been successfully applied in characterizing the structural similarity/diversity of molecules and in QSAR/QSPR modeling (Burden, 1997).

3D Descriptors or geometrical descriptors

Another class of molecular descriptors, called geometrical or 3D descriptors, is derived from a geometrical representation of the molecule, that is, fromx–y–zCartesian coordinates of the molecule atoms. Some of the most known geometrical descriptors are here shortly presented. WHIM56 descriptors are molecular descriptors based on statistical indices calculated on the projections of the atoms along principal axes of the molecule. They are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The algorithm consists in performing a PCA on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the atoms. Foreach weighting scheme, a set of statistical indices is calculated on the atoms projected onto each principalcomponent, that is, the scores(Tarko *et al.,* 2001).

Gravitational indices (Katritzk*et al.,* 1996)are geometrical descriptors reflecting the mass distribution in a molecule, defined as

$$G_1 = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{m_i m_j}{r_{ij}^2}$$

$$G_2 = \sum_{b=1}^{B} \left(\frac{m_i m_j}{r_{ij}^2}\right)_b$$

Wheremi and mj are the atomic masses of the considered atoms,rijthe corresponding interatomic distances, AandBthe number of atoms and bonds of the molecule, respectively. TheG1index takes into account all atom pairs in the molecule, whereas theG2index is restricted to pairs of bonded atoms. These indices are related to the bulk cohesiveness of the molecules accounting,

simultaneously, for both atomic masses (volumes) and their distribution within the molecular space.

EVA descriptorswere proposed to extract chemical structural information from mid- and near-infrared spectra. The approach is to use, as a multivariate descriptor, the vibrational frequencies of a molecule, a fundamental molecular property characterized reliably and easily from the potential energy function. The EVA descriptor is a function of the eigenvalues obtained from the normal coordinate matrix; it corresponds to the fundamental vibrational frequencies of the molecule, which can be calculated using standard quantum or molecular mechanical methods from computational chemistry (Fergusone*t al.*, 1997).

The EEVA descriptorsare analogous to the EVA descriptors, but semi-empirical molecular orbital energies,that is, the eigenvalues of the Schrodinger equation, are used instead of the vibrational frequencies of the molecule (Tuppurainen, 1999).3D-MoRSE descriptors (Schuu*et al.*, 1998)are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. The derived expression is the following

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i w_j \frac{\sin(sr_{ij})}{sr_{ij}}$$

WhereI(s) is the scattered electron intensity,wan atomic property (e.g., the atomic number),rij the interatomic distance between theith and the jth atoms, and Athe number of atoms. Radial distribution function (RDF) descriptors (Raevsky*et al.*, 2000)are based on the distance distribution in the geometrical representation of a molecule and constitute a RDF code that shows certain characteristics in common with the 3D-MoRSE descriptors. The GETAWAY 61

descriptors are derived from the molecular influence matrix (H), which is a representation of the molecular structure, defined as

$$H = M \times (M^T \times M)^{-1} \times M^T$$

whereMis the molecular matrix consisting of the centered Cartesian coordinates, x, y, zof the molecule atoms in a chosen conformation. Atomic coordinates are assumed to be calculated with respect to the geometrical center of the molecule in order to obtain translational invariance. The molecular influence matrix is a symmetric $A \times A$matrix, whereArepresents the number of atoms, and shows rotational invariance with respectto the molecule coordinates, thus resulting independent of molecule alignment rules. The diagonal elements$b_{ii}$of this matrix range from 0 to 1 and encode atomic information related to the 'influence' of each molecule atom in determining the whole shape of the molecule; in effect, mantle atoms always have higher b_{ii} values than atoms near the molecule center. GETAWAY descriptors are obtained by using double-weighted autocorrelation functions, where one weighting scheme is the leverage and the other an atomic property (e.g., atomic mass).

As a geometrical representation involves the knowledge of the relative positions of the atoms in 3D space, that is, the (x, y, z) atomic coordinates of the molecule atoms, geometrical descriptors usually provide more information and discrimination power also for similar molecular structures and molecule conformations than topological descriptors. Despite their high information content, geometrical descriptors usually show some drawbacks. They require geometry optimization and therefore the overhead to calculate them. Moreover, for flexible molecules, several molecule conformations are available: on one hand, new information is available and can be exploited, but, on the other hand, the problem complexity can significantly increase(Tarko *et al.*, 2001).

For these reasons, topological descriptors, fingerprints based on fragment counts, and other simple descriptors are usually preferred for the screening of large databases of molecules. On the contrary, searching for relationships between molecular structures and complex properties, such as biological activities, can often efficiently be performed by using geometrical descriptors, exploiting their large information content. Moreover, it is important to remember that the biologically active conformation of the studied chemicals is seldom known. Some authors overcome this problem by using a multi-conformation dynamic approach (Mekenyan*et al.,* 1994; Mekenyan*et al.,* 2003).

4D Descriptors or grid-based descriptors

GRIDand CoMFAapproaches were the first methods based not uniquely on the molecular structure but on the calculation of the interaction energy between molecule and probe. The focus of these approaches is to identify and characterize quantitatively the interactions between the molecule and the receptor's active site. They place the molecules in a 3D lattice constituted by several thousands of evenly spaced grid points and use a probe (steric, electrostatic, hydrophilic, etc.) to map the surface of the molecule on the basis of the molecule interaction with the probe (Goodford, 1985; Cramer*et al.,* 1988).

Besides the two most popular methods GRID and CoMFA, the other known methods based on this approach are CoMSIA, (Klebe*et al.,* 1994) Compass, (Jain*et al.,* 1994) G-WHIM descriptors, (Todeschini*et al.,* 1997) Voronoi Field Analysis, (Chuma*et al.,* 1998) SOMFA, (Robinson*et al.,* 1999) VolSurf descriptors, (Crucian*et al.,*2000) and GRIND(Pastor*et al.,* 2000)Although these descriptors are often called 3D descriptors, they can be more properly called 4D descriptors (or grid-based descriptors) because to geometrical information is added another source of information given by the interaction energy with a specific probe at each point

of a 3D grid embedding the molecule. Therefore, the molecular descriptors are the MIFs generated by probes. These scalar fields can be efficiently visualized and used to think visually about new drug candidates, thus resulting very helpful in the drug discovery process (Buolamwini*et al.,* 2002; Xu*et al.,* 2002).

An advantage of these approaches is that final results show where and how to modify the compounds to reachthe desirable values of the studied molecular property. On the contrary, a drawback is the need of molecularalignment in order to achieve molecular comparability and the selection of the most appropriate conformation. The alignment determines to what extent the descriptors differ from one molecule to the next. Consequently, it substantially influences the results of the evaluation. Hence, significant and relevant results can only be expected if the alignment was carried out properly and unambiguously. Often, the need for an alignment limits the application of certain descriptors to homogeneous data sets, and even then the alignment is not always easily performed. As a consequence, different research groups started to develop alignmentindependent molecular descriptors. The first set of descriptors based on scalar fields but alignment independent were G-WHIM descriptors, (Todeschini*et al.,* 1999)based on the theoretical principles of the WHIM descriptors 56 but applied to the MIFs. VolSurf and GRINDdescriptors are also independent of any previous alignment of the molecules (Todeschini*et al*., 1994; Crucian*et al.,* 2000; Pastor *et al.,* 2000).

2.5 Previous QSAR Work on Polychlorinated Aromatic Compounds

Other works were carried on polychlorinated aromatic compounds and other related compounds using different and techniques and software from the ones used in this work

2.4.1 Docking and 3D-QSAR Studiesof (PCBs), (PCDDs) and(PCDFs)

(Huifeng *et al.,* 2011) reported that polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs) binding with the aryl hydrocarbon receptor (AhR) have been correlated with many toxic responses. Hence, it is very necessary to study the interactions between these ligands and AhR for further understanding of the mechanism of toxicity. In this study, an integrated molecular docking and 3D-QSAR approach was employed to investigate the binding interactions between PCBs, PCDDs, PCDFs and AhR. From molecular docking, hydrogen-bonding and hydrophobic interactions were observed to be characteristic interactions between compounds and AhR. Based on the mechanism of interactions, an optimum 3D-QSAR model with good robustness (Q2 CUM=0.907) and predictability (Q2 EXT=0.863) was developed by partial least squares. Additionally, the developed QSAR model indicated that the molecular size, shape profiles, polarizability and electropological states of compounds were related to thebinding affinities to AhR.PLS regression with pEC50 as the dependent variable and the selected molecular structural parameters as predictorvariables resulted in the following optimal QSAR model:

pEC50 =7.05+5.51E1p+0.633RGyr−0.331RDF065u−0.610Mor14u−0.176SEig+3.10E1

n (training set) = 52, A = 3, Q2 CUM = 0.907, R2 =0.922, RMSE=0.430 (training set)

n (validation set) = 13,Q2 EXT=0.863, RMSE=0.446 (validation set),$p<0.0001$.

whereAis the number of PLS components andpis the significance level.

The 3D crystal structure of AhR was homologously modeled, and docking analysis showed that hydrogen bonding and hydrophobic interactions between compounds and AhR governed the binding affinities. Based on the mechanism of interactions, a QSAR was

established to characterize the interactions and to model the relative binding affinity for PCBs, PCDDs, and PCDFs. Molecular size, shape profiles, polarizability and electropological were important factors for the binding interactions between compounds and AhR. The developed QSAR model had good robustness, predictive ability and mechanism interpretability, which could be applied to predict the binding affinity of other compounds.

2.4.2 Binding affinities (AhR) of PCBs, PCDDs and PCDFs and QSAR (DFT)

Quantitative Structure-Activity Relationship (QSAR) study according to Tahar*et al.* (2014) was applied to the prediction of the binding interactions between PCBs, PCDDs, PCDFs, and AhR. The concentration of the test chemicals reducing specific binding of [3H] 50% pEC50 of 65 compounds were modeled using the parameters from quantum - chemical calculations at density functional theory level (B3LYP/6-31G (d,p)) were applied. It was expected that the main contribution to the degradation rate was given by the total energy ET. Principal Component Analysis (PCA) method, the Multiple Linear Regression method (MLR), Partial Multiple Non Linear R egression method (MNLR) and the Artificial Neural Network (ANN), showed a correlation coefficient more than 0,9. The prediction results were in excellent agreement with the experimental values; It is expected to be beneficial to predict the activity of other compounds in the same groups.

Many attempts have been made to develop a relationship with the indicator variable of mean pEC50, but the best relationship obtained by this method is only one corresponding to the linear combination of several descriptors: the total energy, energy EHOMO, energy ELUMO, activation energy Ea, the dipole moment μ and the factor of oscillation f(SO)

$$PEC_{50} = 6.409 - 2.402 * 10^{-4}E_T - 0.416E_{HOMO} + 0.184E_{LUMO} - 0.136\mu - 1.726E_a + 1.366 * 10^{-2}\lambda_{max} + 2.437f_{so}$$

N = 65 R2 = 0.51 S= 1.1537 F = 1.0585 RMSE = 1.0408

Finally, we can conclude that one studied descriptors (ET, EHOMO, ELUMO, ΔE, Ea,λ_{max}, μ and f(SO).), which are sufficiently rich in chemical and electronic information to encode the structural features, may be used with other topological descriptors for development of predictive QSAR models.

2.4.3 QSAR for toxicities of PCDF, PCDDs and PCBs

According the research carried out by (Adrian *et al.,* 2004), the toxic equivalency factors (TEF values) for 12 polychlorobiphenyls (PCBs), 14 polychlorodibenzo-1,4-dioxins (PCDDs), and 24 polychlorodibenzofurans (PCDFs) have been correlated with the extent and position of chloro-substitution using a variety of molecular descriptors., with Todeschini's program Dragon 4.0 and the CODESSA program of Katritzky and coworkers.

(i) Polychlorodibenzofurans (PCDFs):

A. Dragon program (24 compounds). Two PCDFs from the original set were left out.

logTEF = –5.013(±0.653)·DP05 + 0.774(±0.180)·AGDD –39.36

n = 24; r = 0.925; s = 0.472; F = 62.43; Q = 1.959

DP05: Randic molecular profile number 05.

AGDD: Average geometric distance degree.

(i)**Polychlorodibenzo-1,4-dioxins**

A. Dragon program with three descriptors

logTEF = 22.11(±13.59) Hy96.15(±50.36) ·BLTF96+86.62(±41.46)· BLTA96+6.684(±1.247)· GGI3 – 40.30

n = 14; r = 0.936; s = 0.451; F = 16.17; Q = 2.075

Hy: Hydrophilic factor

BLTF96: Verhaal model of Fish base-line toxicity from MLOGP (mmol/L)

BLTA96: Verhaal model of Algae base-line toxicity from MLOGP (mmol/L)

GGI3: Topological charge index of order 3

(i) Polychlorobiphenyls

A. Dragon program with four descriptors

logTEF = –6.167(±0.963)DP05 + 0.295(±0.204)ATS3e – 2.130(±0.920)MLOGP– 7.407(±3.106)RDF015v + 83.10

n = 12; r = 0.927; s = 0.437; F = 10.62; Q = 2.121

DP05: Randic molecular profile number 05

ATS3e: Broto-Moreau autocorrelation of topological strucuture-lag3/weighted by atomic Sanderson

MLOGP: Moriguchi octanol-water partition coefficient (log P)

RDF015v: Radial distribution function-1.5/ weighted by atomic van der Waals volumes

Di-, tri-, or tetra-parametric linear correlations between toxicity values (logTEF) of three classes of polychloroaromatics, namely polychlorobiphenyls (PCBs), polychlorodibenzo-1,4-

dioxins (PCDDs), and polychlorodibenzofurans (PCDFs) have been presented. This QSAR study has used the Dragon program of Todeschini and the CODESSA program ofKatritzky and their coworkers. The latter programwas slightly superior. Only about 6% of all possible PCBs, and 19% of all possible PCDDs and PCDFs were available with such toxicity data, but since 3.5–4 orders of magnitude of toxicities were involved, one may confidently assume that mostmajor structural factors could manifest themselves in these data. However, many substitution patterns are not present in the experimentally available toxicity data, so thatsome of these patterns may escape detection. Qualitatively, one may infer that the highest toxicity is observed with 3,3',4,4'-tetrachlorobiphenyls (possibly withsupplemental 3' or 5' chlorosubstituents) but devoid of ortho-chloro substituents. Apart from 2,3,7,8-tetrachloro-1,4-dibenzodioxin with the highest known toxicity of polychloroaromatics so far, any dioxin or dibenzofuran with at least one or two chlorine substituents on each benzenoid ring in positions 2,3,6,or 7 has a significant toxicity.

2.4.4 QSAR Studies on polychlorinated aromatic compounds using topological descriptors

According to Kumar*et al. (*2013), polychlorinated dibenzo-p-dioxins (PCDDs) and Polychlorinated dibenzofurans (PCDFs) represent a large group of industrial and byproduct compound which is resistance to chemical and biological degradation and highly toxic. Series of 18 PCDD and PCDFcompounds is subjected to QSTR studies by using different molecular descriptors. The dependent variable used in this study represents log(EC50) values (EC50-median effective concentration during a bioassay). The regression analysis of the data has shown that the toxicity of the compound can be modeled excellently in multi-parametric model.Its resulting model exhibited good R2 value up to 0.8938.

Software: Allmolecular modeling studies were carried out using HYPERCHEM(version 7.5) and DRAGON software. The structures of molecules were drawn using Chemsketch software. NCSS Inc. is a leading worldwide provider of predictive analytics software and solutions.

We have carried out stepwise multiple regression analysis for modeling of Compound no 18.

$$\text{Log } (EC_{50}) = -1.618(\pm 0.225)\alpha + 26.940(\pm 3.518)J + 0.02800(\pm 0.0047)W - 23.945$$

On the basis of above observation it leads to the conclusion that the activity logEC50 of the present set of compounds can be successfully modelled using molecular descriptors. It was also observed that out of the molecular descriptors used α, β; W and J are most useful for this purpose. The MFA equation suggested that (-ve) sign of descriptor is disfavour the activity while (+ve) sign of W and J parameters indicated that they favoured activity. Our resultsopen very interesting perspectives regarding polychlorinated aromatic compounds with toxicity.

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 Materials

The materials used in this research include both software and hardware. The hardware used includes the following:

- HP Pavilion (Pentium (R) Dual-core CPU), 3.00GB RAM with 4.00GHz processor
- Printer

The softwares used include the following:

- Spartan 14v.1.1.0 software
- Material Studio software version 7.0
- Chem3D Pro12.0.1v
- Padel descriptor tool kit
- datasetdivisionGUI1.2 software
- Chemsketch, and
- Microsoft office Excel 2013.

3.2 Methods

Quantitative Structure-Toxicity Relationship studies of the 76 polychlorinated aromatic compounds (polychlorinated dibenzo-p-dioxins PCDDs, polychlorinated dibenzofurans PCDFs, and polychlorinated biphenyls PCBs) was carried out using one of the QSAR approach (Hansch's approach) which was developed by the founder of QSAR, Corwin Hansch.. In this approach, biological activity for a group of 'congeneric' chemicals can be described by a comprehensive model:

$$\log\left(\frac{1}{C_{50}}\right) = a\pi + b\varepsilon + cS + d \ldots\ldots\ldots\ldots\ldots\ldots 3.1$$

whereC, the toxicant concentration at which an endpoint is manifested (e.g., 50% mortality or effect), is related to a hydrophobicity term, π an electronic term, ε (originally the Hammett substituent constant, σ), and a steric term,S(typically Taft's substituent constant, Es), beingda general additional term depending on the kind of property to be modeled (Todeschini*et al.*, 2009).

Different types ofstructural properties of the molecules were calculated and these properties (descriptors) are correlated with biological activities (toxicities) using regression analysis. In order to reduce the dispersion of data set and to get data fitting and linear respond,the experimentally determined toxicities (EC_{50}) of the compounds was converted to log EC_{50}. Figure3.1 presents the steps involve in this work.

3.2.1Data Collection

The chemical names and experimentally determined toxicities (EC_{50}) data set of 74 molecules (25 PCDDs, 34 PCDFs and 15 PCBs) were taken from the literature (Azeddine *et al.*, 2014, Adrian *et al.*, 2004, Tudoran *et al.*, 2012). The data set was chosen based on the requirement presented in Figure 3.1.

The structures of the compounds were drawn using Chemsketch and chemdraw softwares. The general structural formulae of the three series of polychlorinated aromatic compounds are presented in Figure3.2 and (IUPAC names and toxicity data) of all the compounds are listed in Table.3.1.

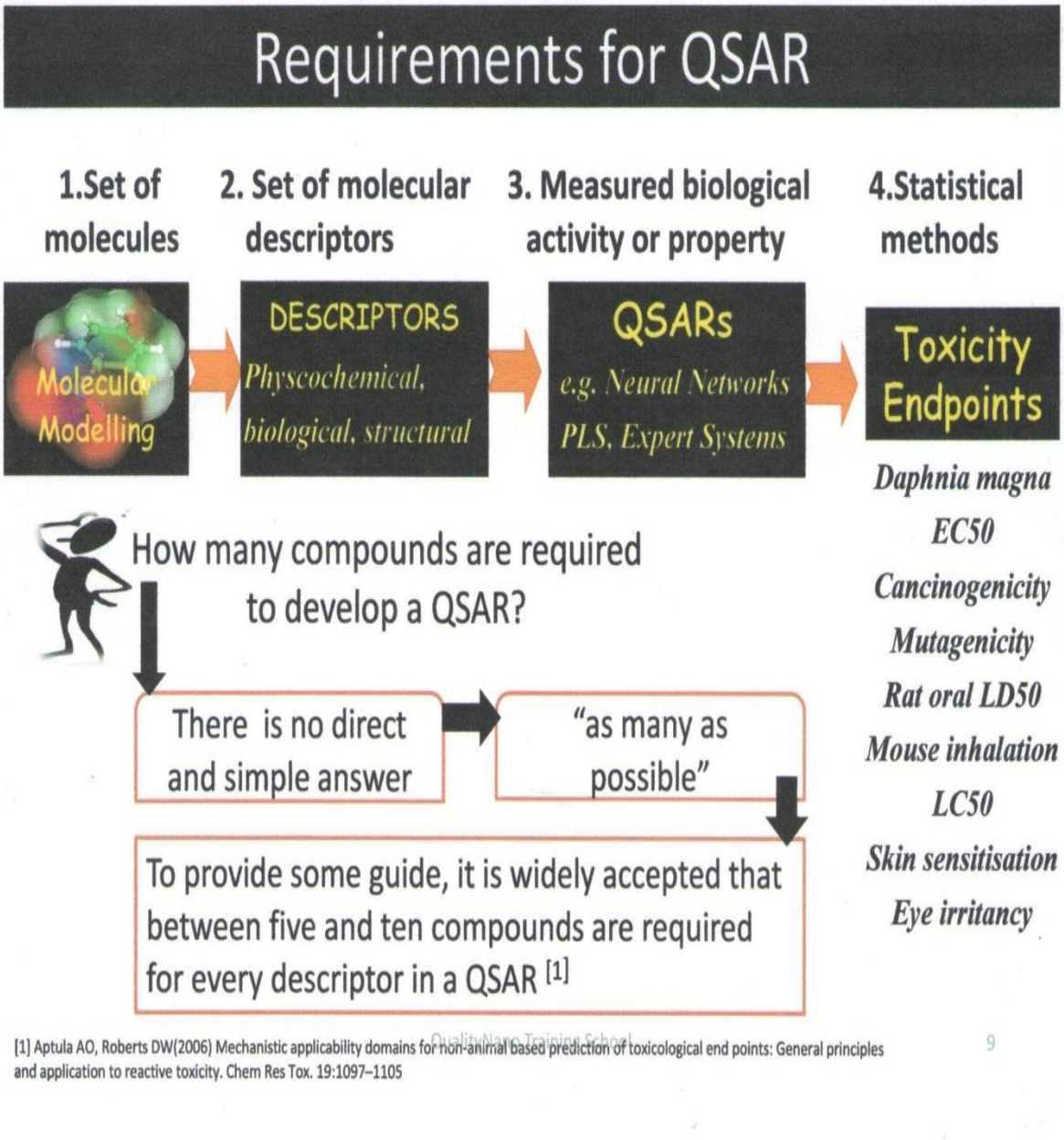


Figure 3.1: Requirements for QSAR (selection of data set)(Atula *et al.*, 2006).

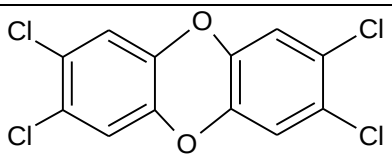
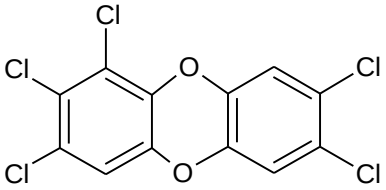
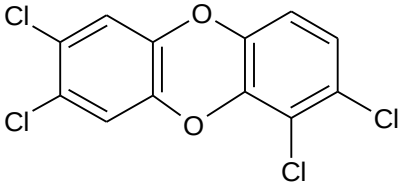
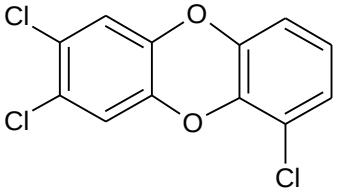
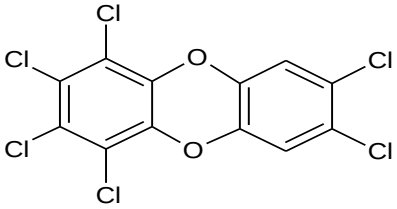
Polychlorinated dibenzo-P-dioxins (PCDDs)

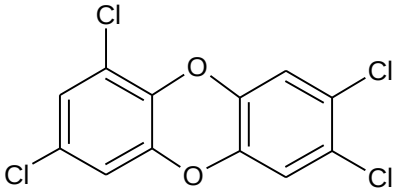
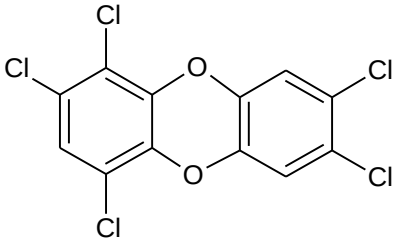
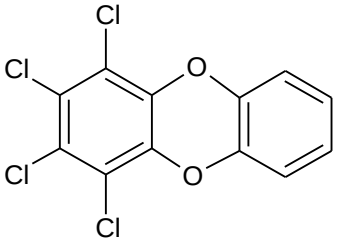
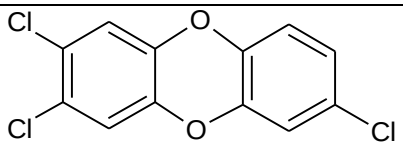
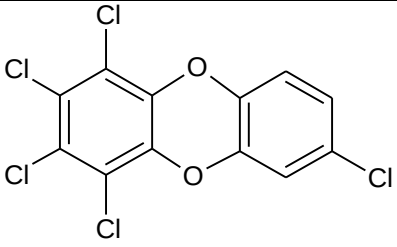
Polychlorinated biphenyls (PCBs)

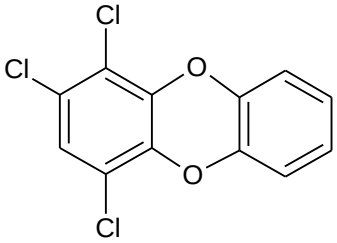
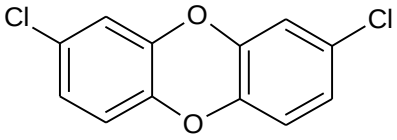
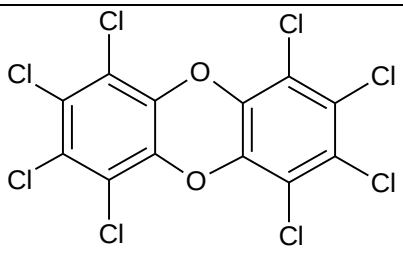
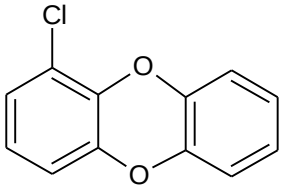
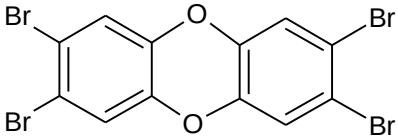
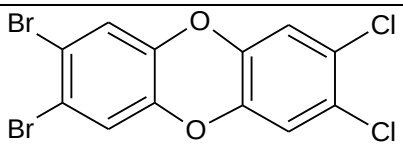
Polychlorinated dibenzofurans (PCDFs)

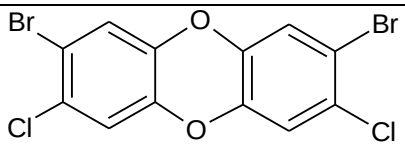
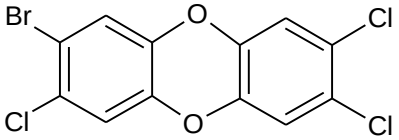
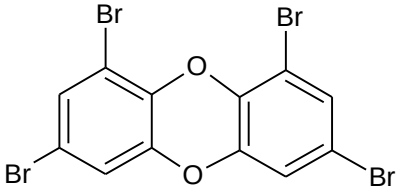
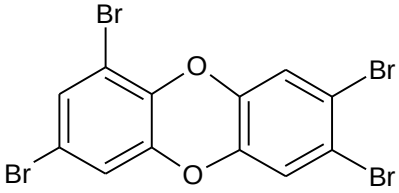
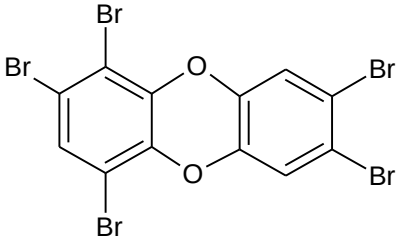
Figure.3.2: General structural formulae for the compounds (PCDDs, PCDFs and PCBs)

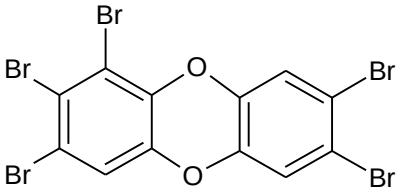
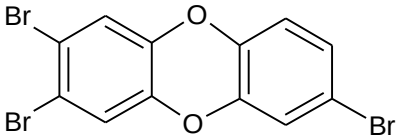
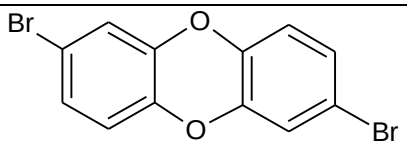
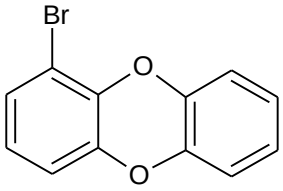
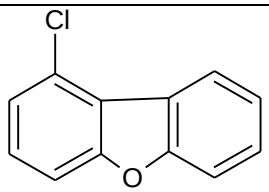
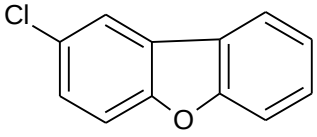
Table 3.1: Experimental Biological Activities(pEC_{50}) of the compounds, (PCDDs, PCDFs, PCBs)

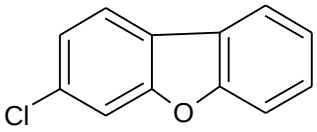
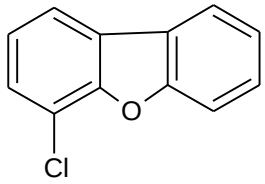
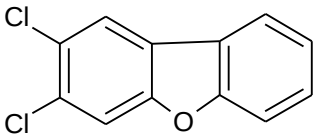
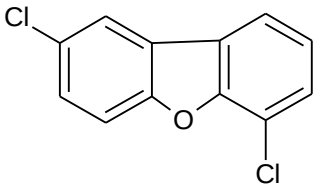
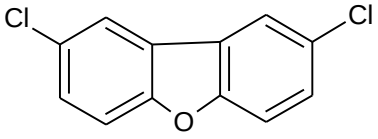
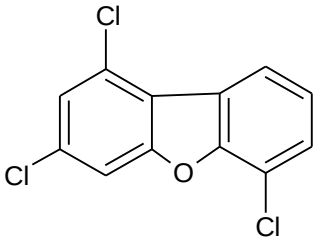
S/N	IUPAC NAME	pEC_{50}
1	 2,3,7,8 Tetrachlorodibenzo-p-dioxin	8.00
2	 1,2,3,7,8 pentachlorodibenzo-p-dioxin	7.10
3	 2,3,6,7 Tetrachlorodibenzo-p-dioxin	6.80
4	 2,3,6 Trichlorodibenzo-p-dioxin	6.66
5	 1,2,3,4,7,8Hexachlorodibenzo-p-dioxin	6.55

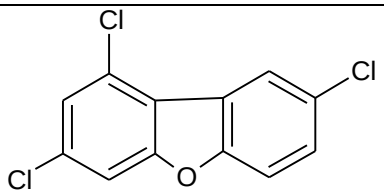
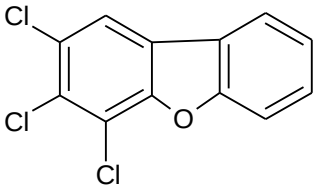
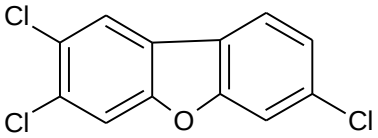
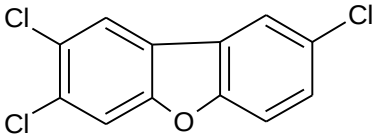
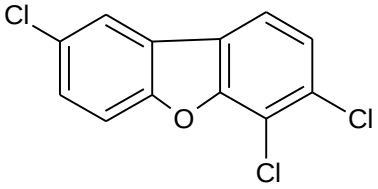
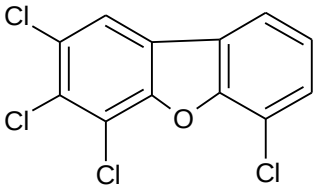
6	 1,3,7,8-Tetrachlorodibenzo-p-dioxin	6.10
7	 1,2,4,7,8-pentachlorodibenzo--dioxin	5.96
8	 1,2,3,4-Tetrachlorodibenzo-p-dioxin	5.89
9	 2,3,7-Trichlorodibenzo-p-dioxin	7.15
10	 1,2,3,4,7-pentachlorodibenzo-p-dioxin	5.19

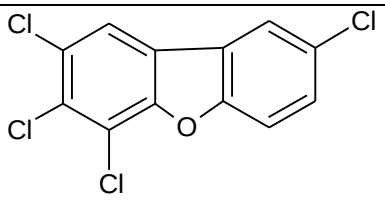
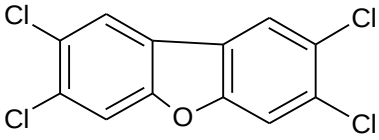
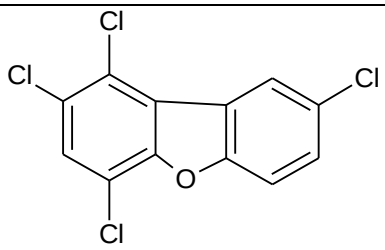
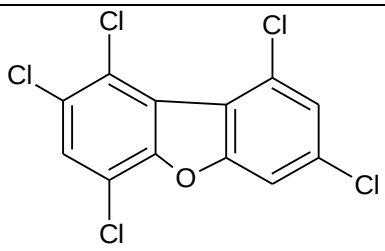
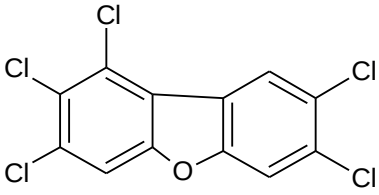
11	 1,2,4-Trichlorodibenzo-p-dioxin	4.89
12	 2,8-dichlorodibenzo-p-dioxine	5.49
13	 1,2,3,4,6,7,8,9-Octachlorodibenzo-o-dioxin	5.00
14	 1-chlorodibenzo-p-dioxin	4.00
15	 2,3,7,8-Tetrabromodibenzo-p-dioxin	8.82
16		8.83

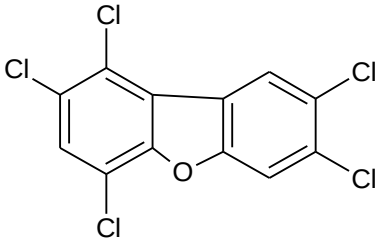
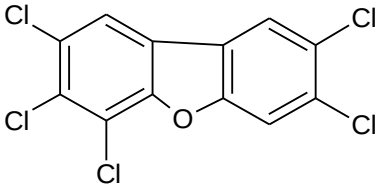
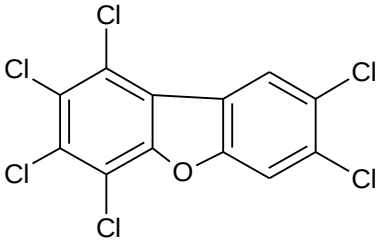
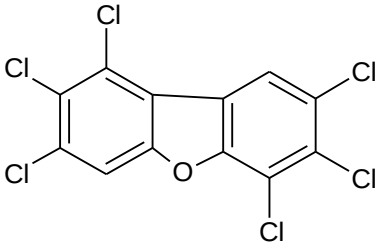
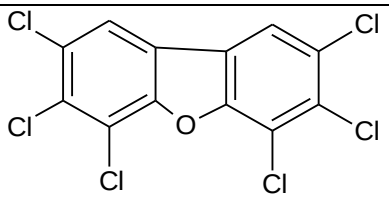
	2,3- Dibromo 7,8-chlorodibenzo-p-dioxin	
17	 2,8- Dibromo -3,7-dichlorodibenzo-p-dioxin	9.35
18	 2,Bromo-3,7,8-trichlorodibenzo-p-dioxin	7.94
19	 1,3,7,9-Tetrabromodibenzo-p-dioxin	7.03
20	 1,3,7,8-Tetrabromodibenzo-p-dioxin	8.70
21	 1,2,4,7,8-Pentabromodibenzo-p-dioxin	7.77

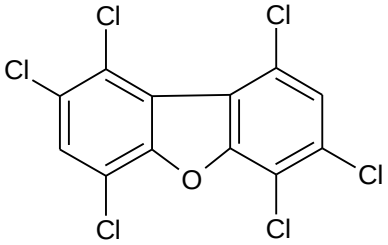
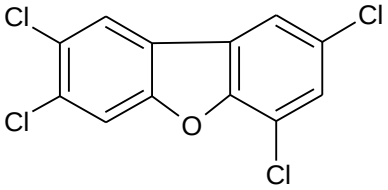
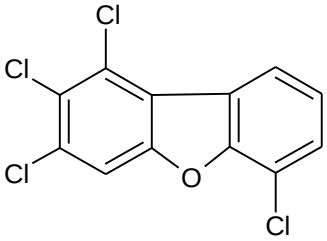
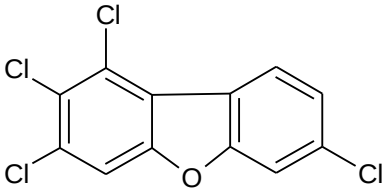
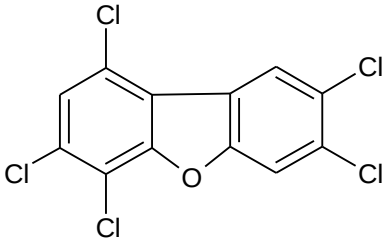
22	 1,2,3,7,8-Pentabromodibenzo-p-dioxin	8.18
23	 2,3,7-Tribromodibenzo-p-dioxin	8.93
24	 2,7-Dibromodibenzo-p-dioxin	7.81
25	 2-Bromodibenzo-p-dioxin	6.53
26	 1-chlorodibenzofuran	4.53
27	 2-chlorodibenzofuran	3.55

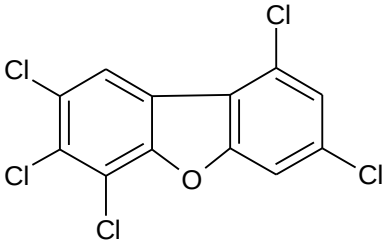
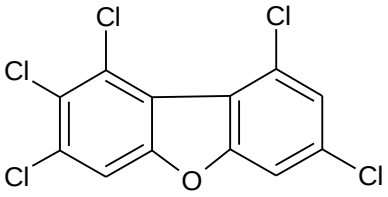
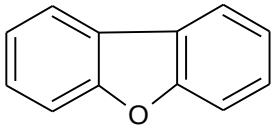
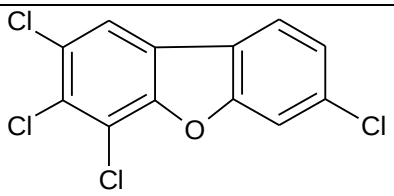
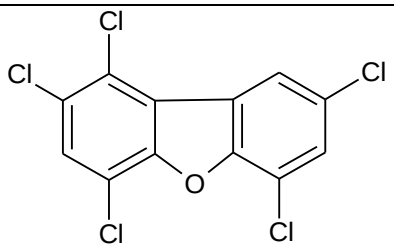
28	 3-chlorodibenzofuran	4.38
29	 4-chlorodibenzofuran	3.00
30	 2,3-Dichlorodibenzofuran	5.36
31	 2,6-dichlorodibenzofuran	3.61
32	 2,8-Dichlorodibenzofuran	5.05
33	 1,3,6-Trichlorodibenzofuran	5.36

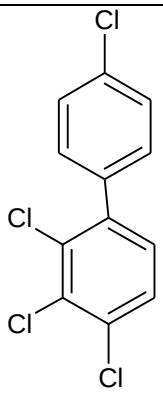
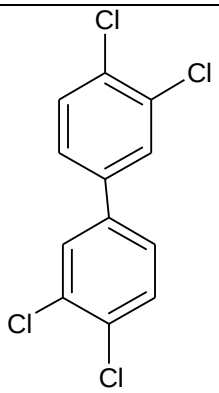
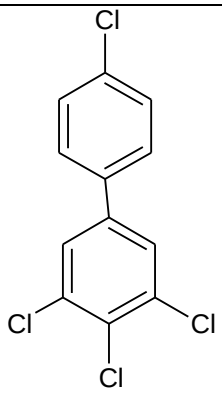
34	 1,3,8-Trichlorodibenzofuran	4.07
35	 2,3,4-Trichlorodibenzofuran	4.72
36	 2,3,7-Trichlorodibenzofuran	7.10
37	 2,3,8-Trichlorodibenzofuran	6.00
38	 2,6,7 Trichlorodibenzofuran	6.35
39	 2,3,4,6 Tetrachlorodibenzofuran	6.46

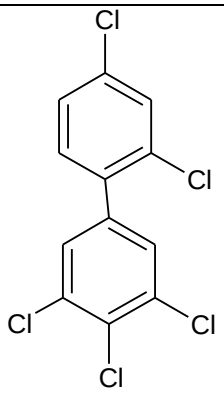
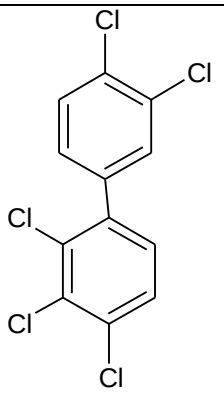
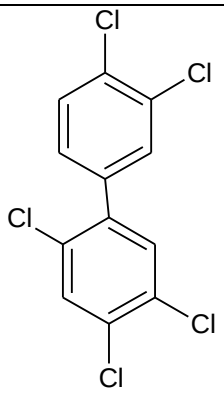
40	 2,3,4,8-Tetrachlorodibenzofuran	6.70
41	 2,3,7,8-Tetrachlorodibenzofuran	7.39
42	 1,2,4,8-Tetrachlorodibenzofuran	5.00
43	 1,2,4,7,9-Pentachlorodibenzofuran	4.70
44	 1,2,3,7,8-Pentachlorodibenzofuran	7.13

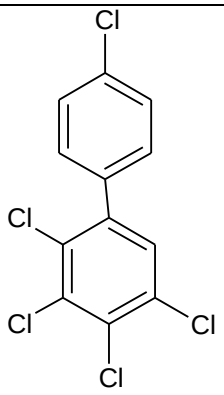
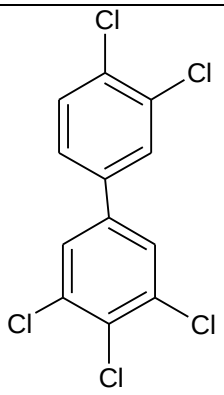
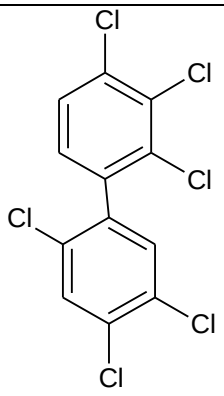
45	 1,2,4,7,8-Pentachlorodibenzofuran	5.89
46	 2,3,4,7,8-Pentachlorodibenzofuran	7.82
47	 1,2,3,4,7,8-Hexachlorobenzofuran	6.64
48	 1,2,3,6,7,8-Hexachlorobenzofuran	6.57
49	 2,3,4,67,8-Hexachlorobenzofuran	7.33

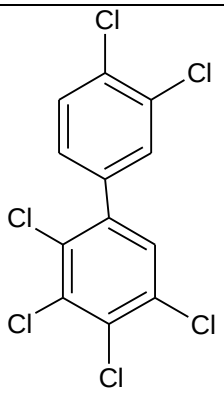
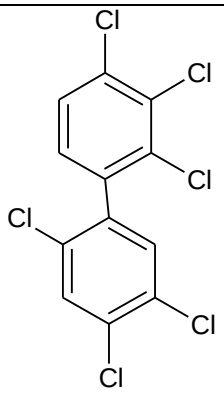
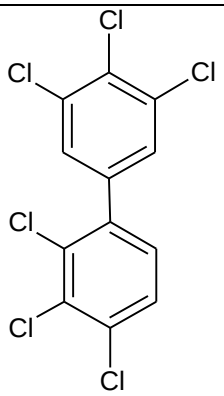
50	 1,2,4,6,7,9-Hexachlorodibenzofaran	5.08
51	 2,3,6,8-Tetrachlorodibenzofuran	6.66
52	 1,2,3,6-Tetrachlorodibenzofuran	6.46
53	 1,2,3,7-Tetrachlorodibenzofuran	6.96
54	 1,3,4,7,8-Pentachlorodibenzofuran	6.70

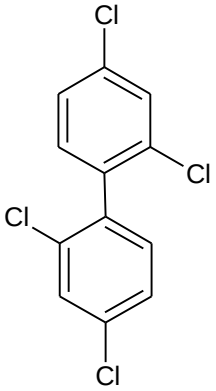
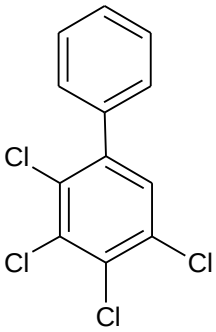
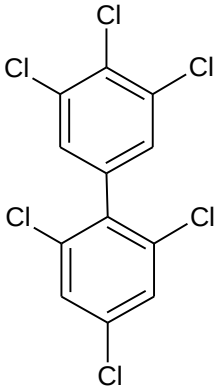
55	 2,3,4,7,9-Pentachlorodibenzofuran	6.70
56	 1,2,3,7,9-Pentachlorodibenzofuran	6.40
57	 Dibenzofuran	3.00
58	 2,3,4,7-Tetrachlorobiphenyl	7.60
59	 1,2,4,6,8-Pentachlorobiphenyl	5.51

60	 2,3,4,4’-Tetrachlorobiphenyl	4.94
61	 3,3’,4,4’-Tetrachlorobiphenyl	6.15
62	 3,4,4’,5-Tetrachlorobiphenyl	4.55

63	 2’,3,4,4’,5-Pentachlorobiphenyl	4.85
64	 2,3,3’,4,4’-Pentachlorobiphenyl	5.37
65	 2,3’,4,4’,5-Pentachlorobiphenyl	5.04

66	 2,3,4,4’,5-Pentachlorobiphenyl	5.39
67	 3,3’4,4’5-Pentachlorobiphenyl	6.92
68	 2,2’4,4’,5,5’-Hexachlorobiphenyl	4.26

69	 2,3,3’,4,4’,5-Hexachlorobiphenyl	5.15
70	 2,3’,4,4’,5,5’-Hexachlorobiphenyl	4.80
71	 2,3,3’,4,4’,5’-Hexachlorobiphenyl	5.30

72	 2,2’,4,4’-Tetrachlorobiphenyl	
73	 2,3,4,5-Tetrachlorobiphenyl	3.85
74	 2,3’,4,4’,5’,6-Hexachlorobiphenyl	4.00

3.2.2 Molecular Optimization

Molecular optimization is simply the process of finding the lowest energy geometry of the molecules. The chemical structure of each molecule of the 74 compounds was drawn using Chemdraw V12.0 and was saved as cdx file. Geometry optimization was performed using the molecular modeling program Spartan '14' V1.1.0 on HP Pavilion on Microsoft windows 7 Ultimate operating system.

3.2.3 Optimization with spartan 14. V.1.1.0

The detail procedures involved in geometry optimization of molecules using the software Spartan 14 V1.1.0 are heighted below:

The software icon of Spartan "14" V.1.1.0 located on the desktop of the computer was launched by double clicking on it. The user interface of the software Spartan "14" V.1.1.0 is displayed on the screen of the computer. This interface is presented in Figure 3.3. By double clicking on the file menu, then *New build,* the organic model kit appeared on the left side of the user interface as presented in Figure 3.4. The selection of the atomic fragments corresponding to the elements commonly found in the organic molecule, (say 2,3,7,8 Tetrachlorodibenzo-p-dioxin) was made. The fragments, (ring) was first selected and the group *benzene*was chosen and displayed on the interface of the software by clicking on it. Double clicking on the interface again displays another benzene group. The fragment (-O-) was chosen from the organic model kit and was used with the help of bond making fragment at the lower part of the model kit by clicking on the bond making and then on the two point to be joined between the benzene and the fragment *oxygen* at both ends. The fragment (-Cl) was chosen and placed on the positions (2,3,7,8) of the molecule. After all the molecules have been built, by clicking on the icon

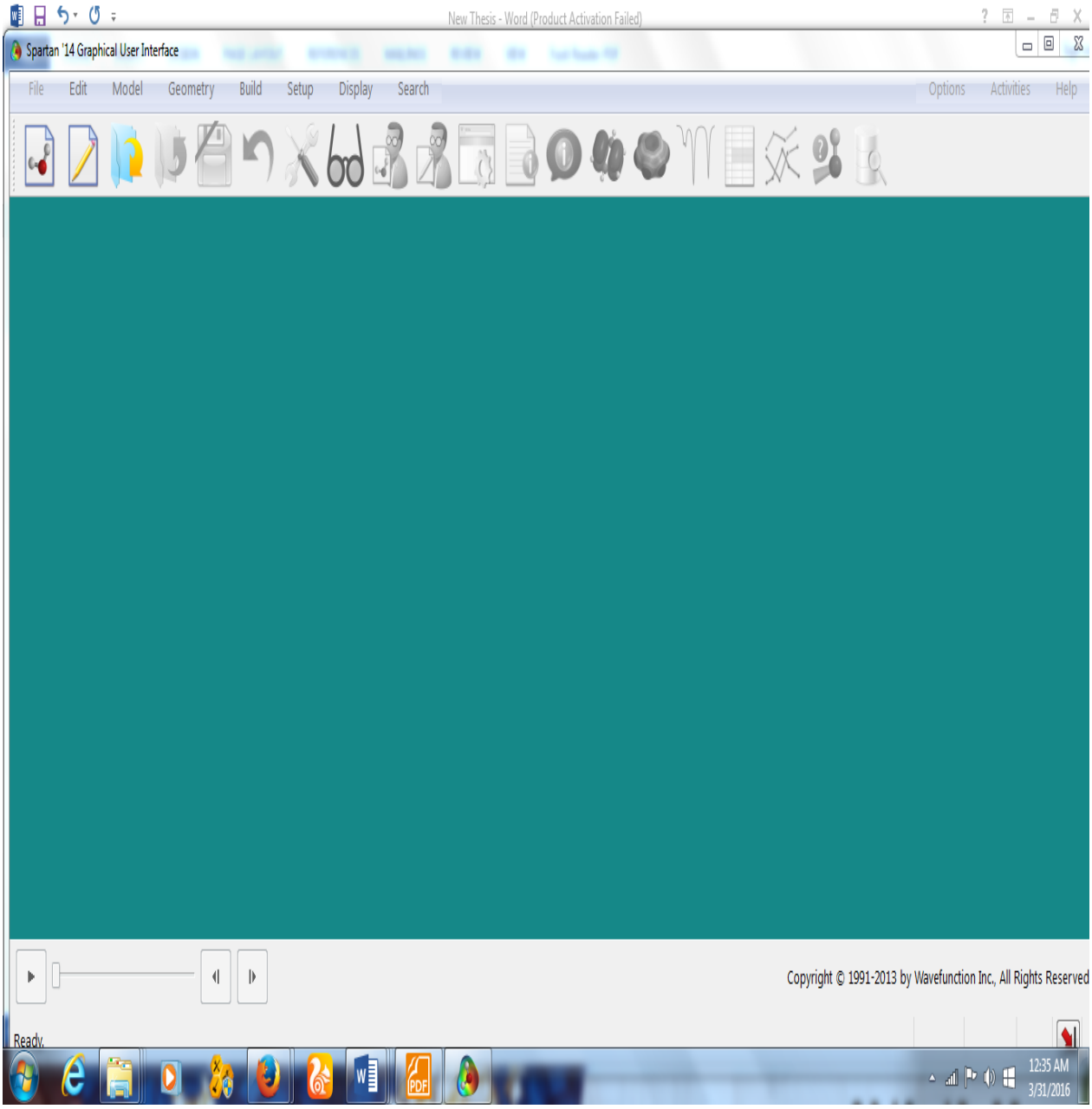


Figure 3.3: The interface of the software spartan 14v1.1.0

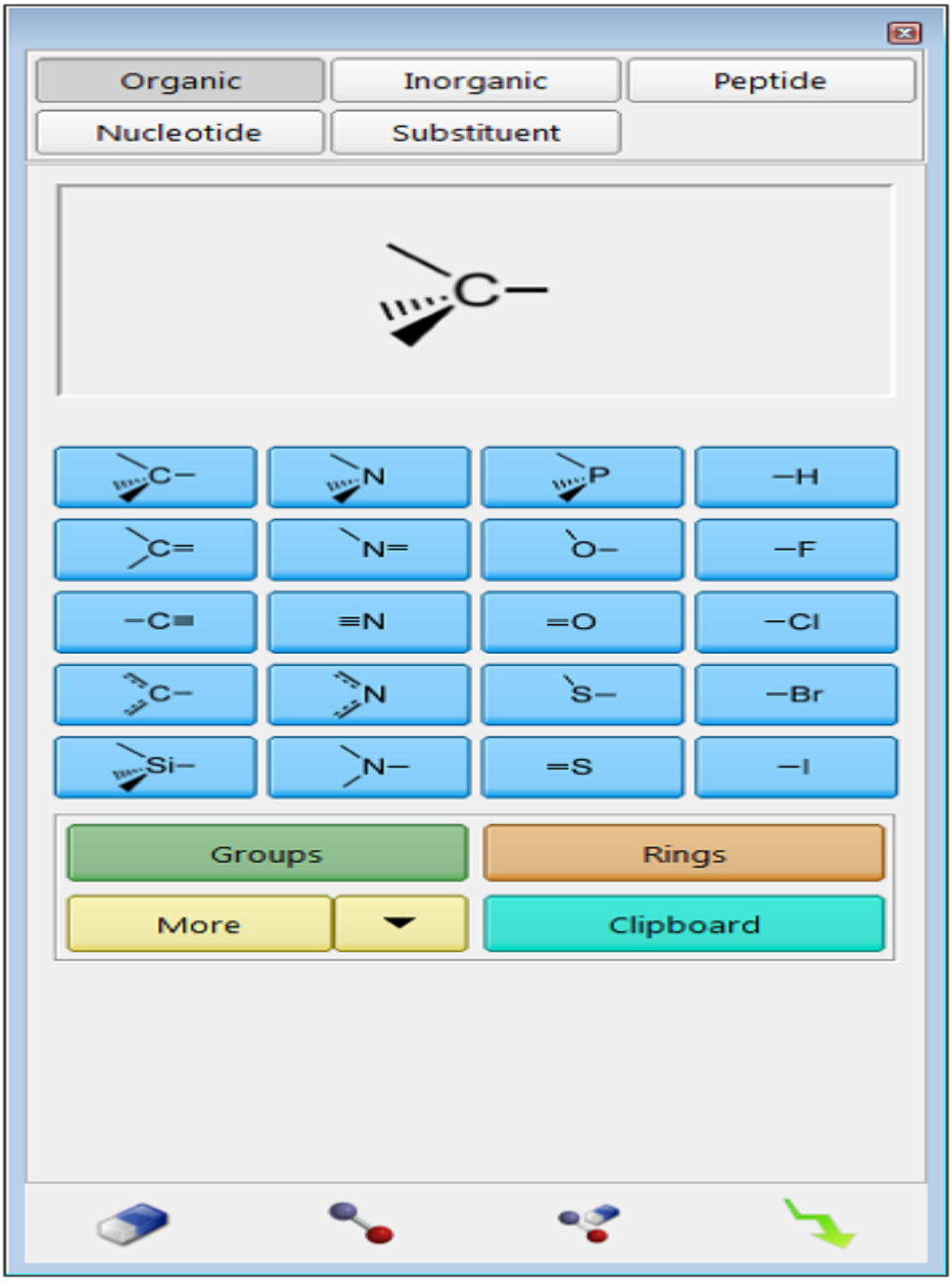


Figure 3.4 .Atomic fragments for molecular drawing.

“minimize’ located at bottom of the right hand side of the interface, strain energy of the molecule was removed. This is presented in Figure. 3.5. This procedure was applied to all the other molecules in each of the data set.

3.4 Calculation of Molecular Descriptors

About 1700 different types of molecular descriptors were calculated using Padel descriptor tool kit, Spartan 14 software and mathematical formulae. Calculation of molecular descriptors in QSAR using computational algorithm and mathematical formulae is presented in Figure 3.6.

3.4.1 Computing molecular descriptors using Spartan 14v.1.1.0 software

From the optimized molecules presented in Figure.3.5, the following procedures were employed in calculating molecular descriptors. By clicking on the *set-up* menu and selecting the option *calculations*, dialog box appeared in which the type of the calculations, molecular mechanics (MMFF) was checked, the level of theory (semi-empirical (AM1) or DFT (B3LYP)/6-31G*) was specified and the properties of interest (such as QSAR, IR, orbital & energies, thermodynamics, charges and bond order) were checked. The whole job was submitted by selecting the *submit* option at bottom end of the dialog box. This is presented in Figure. 3.7.File name and location (File browser) where to store the calculation results was then chosen on the PC from a pop up window that appeared after the job was submitted. This is presented in Figure. 3.8. By clicking the *save* option, a notification that the *calculation has started* was displayed as presented in Figure.3.9. Also, the progress of each calculation was monitored by selecting the option, *monitor* and then clicking the option *calculations* in the upper window that was displayed. Eventually, a notification that the *calculation has completed* appears (Figure.3.10).The output of the calculation was accessed by click on the *display menu* on the appeared on the top of the screen

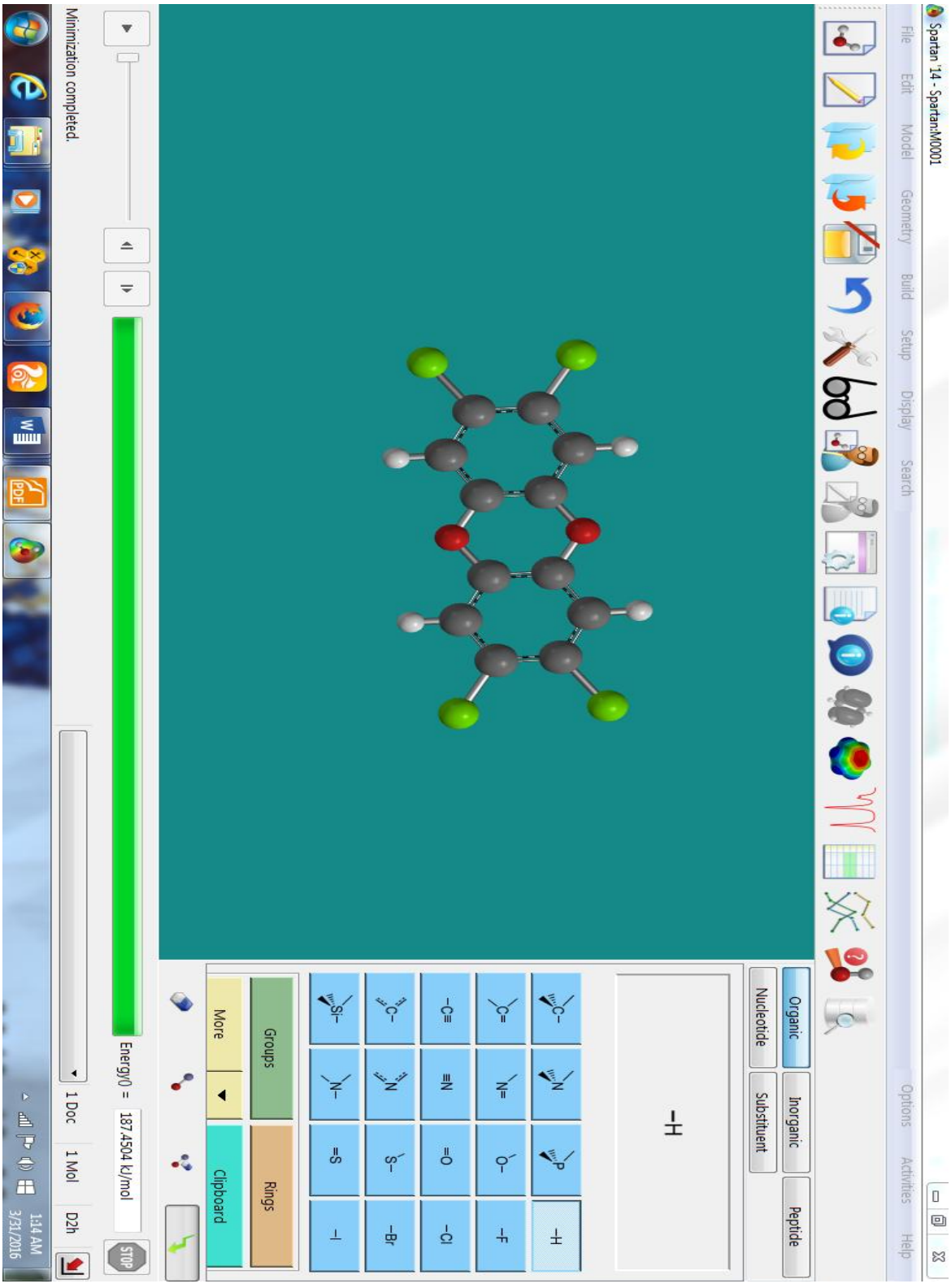


Figure 3.5 The Spartan interface showing the minimized molecules.

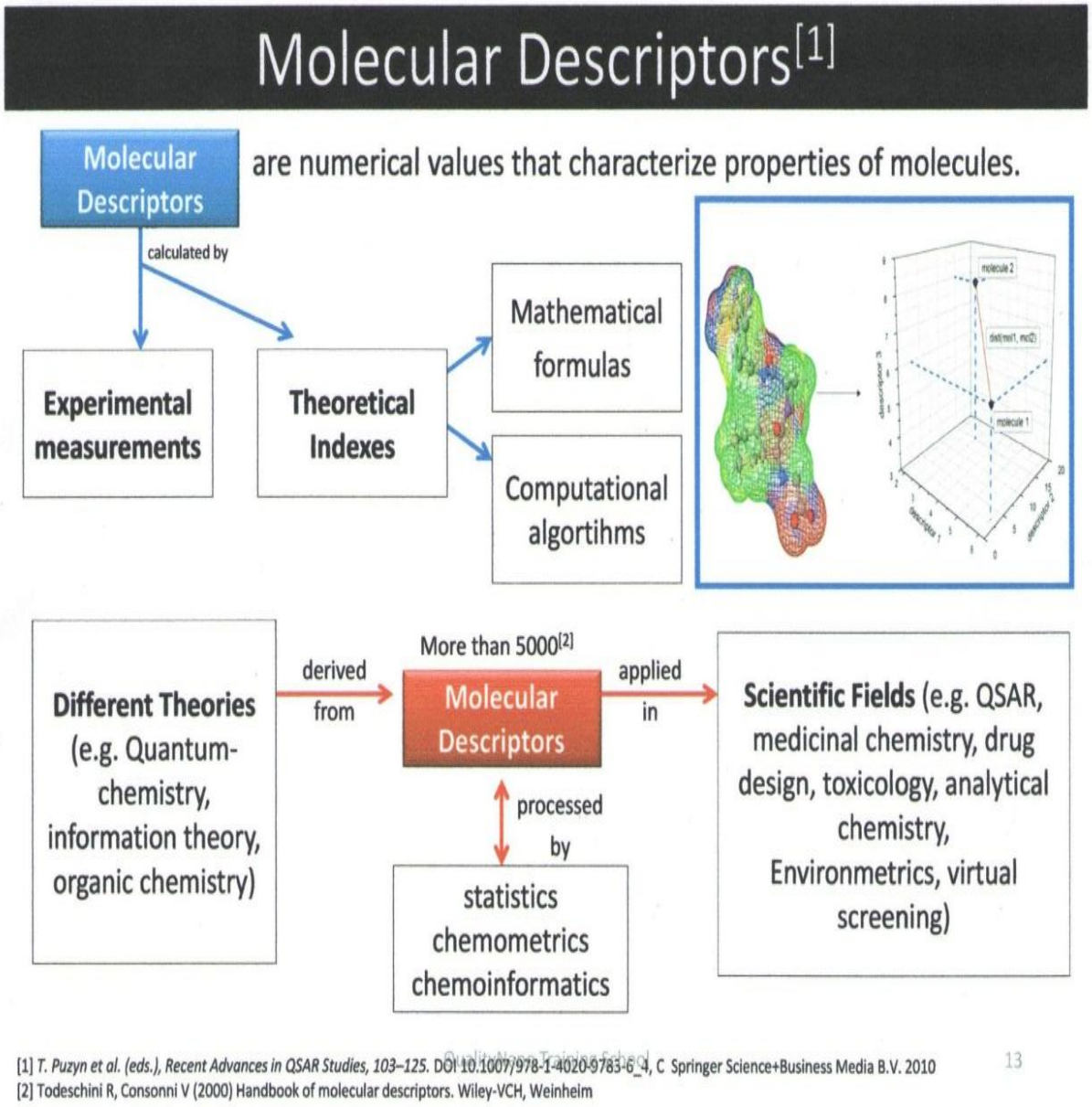


Figure3.6 Chart for the methods of calculation of molecular descriptors in QSAR

(Puzyn *et al.*, 2010)

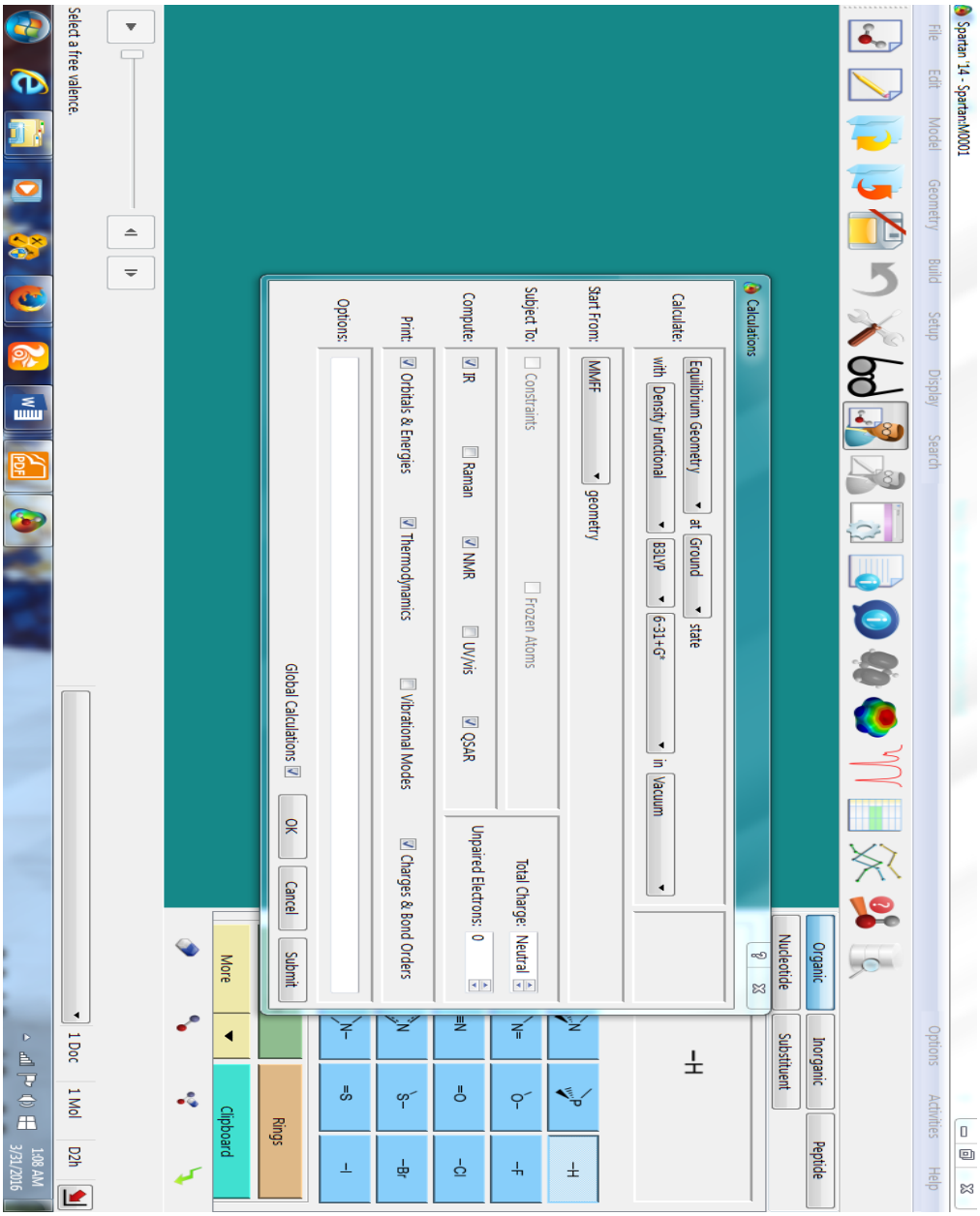


Figure 3.7 *Calculations* dialog box using Spartan 14 software

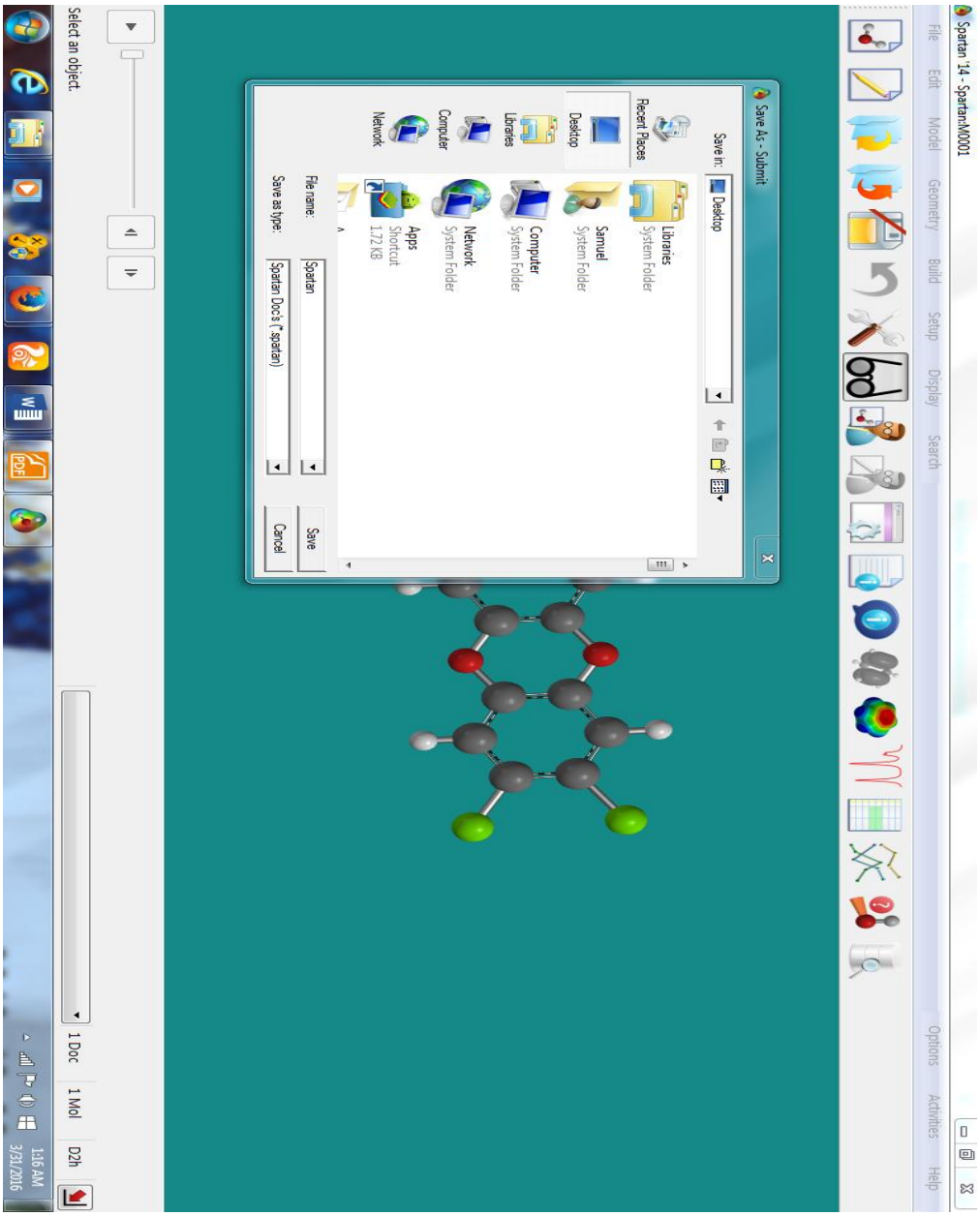


Figure 3.8 File browser for the optimized molecules.

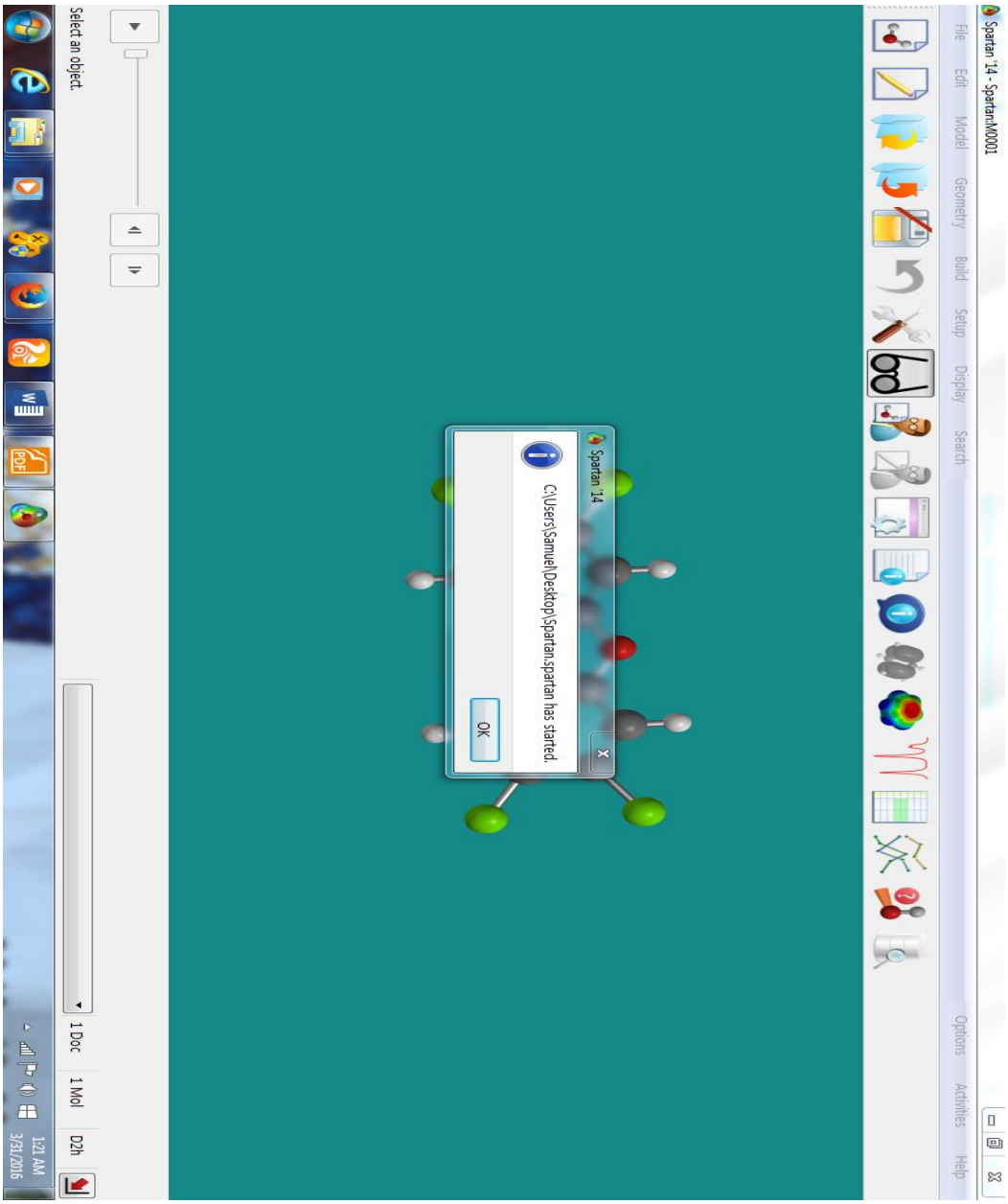


Figure 3.9. Notification that calculation has started.

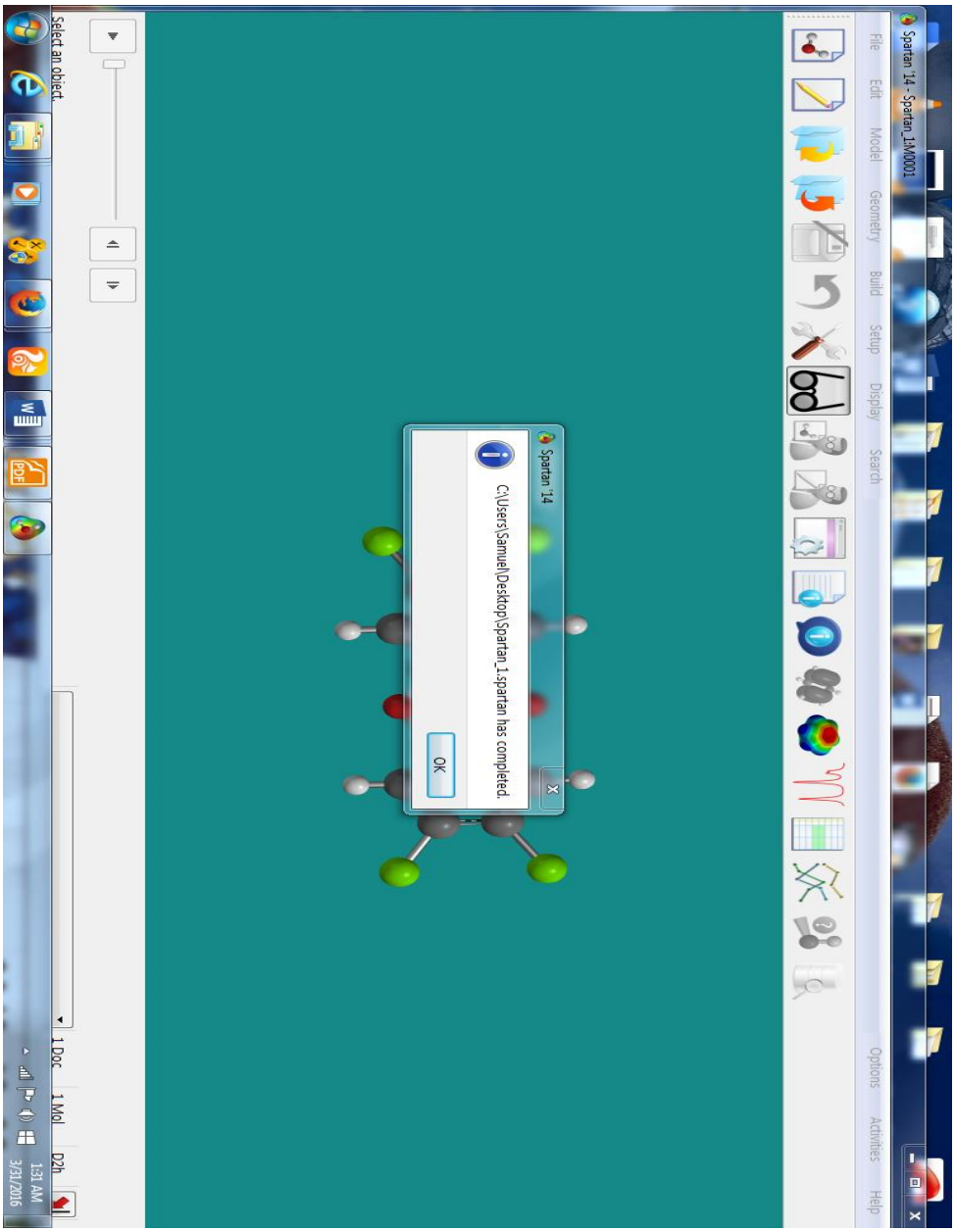


Figure 3.10 Notification that calculation is completed.

and selecting the option *Properties*. The*molecule Properties dialog* appeared. It is divided into four parts designated by tabs. Molecule provided the energy and other information relating to the isolated molecule, QSAR provided quantities that may be used as QSAR descriptors, Thermodynamics provided the enthalpy, heat capacity, Gibbs energy and zero-point energy and Acidity/Basicity provides information about the energies of deprotonation and protonation in the gas phase. Make certain that the *Molecule*tab is selected. These are presented in Figure. 3.11.

3.4.2 Computing molecular descriptors using Padel descriptor tool kit

The Padel software was launched by double clicking on the file *Padel* on the desktop of the computer and its interface was displayed as shown in Figure 3.12.

On the interface, a directory containing the molecular structural files was selected from a prop up dialog box shown in Figure 3.13 by clicking the *molecules directory/file* option on the window. Through the *descriptor output file* menu, the folder where the result of the calculations were saved as *CSV*file type was created from a file browser (as presented in Figure 3.8) that propped up as soon as the *descriptor output file* menu was clicked. By clicking the options- *1D, 2D,*and*3D*on the Padel window shown in Figure 3.14 and marking appropriate boxes against each

descriptors, the descriptors of choice were selected, the molecules were also standardized by checking the options such as *retain 3D coordinates*, *convert to 3D, remove salt, detect aromaticity*, Standardize tautomers and calculations commenced via the ***start***option at the bottom right of the interface as shown in Figure.3.12. The results of the calculations were written in the output file specified on Microsoft excel sheet (Figure 3.15).

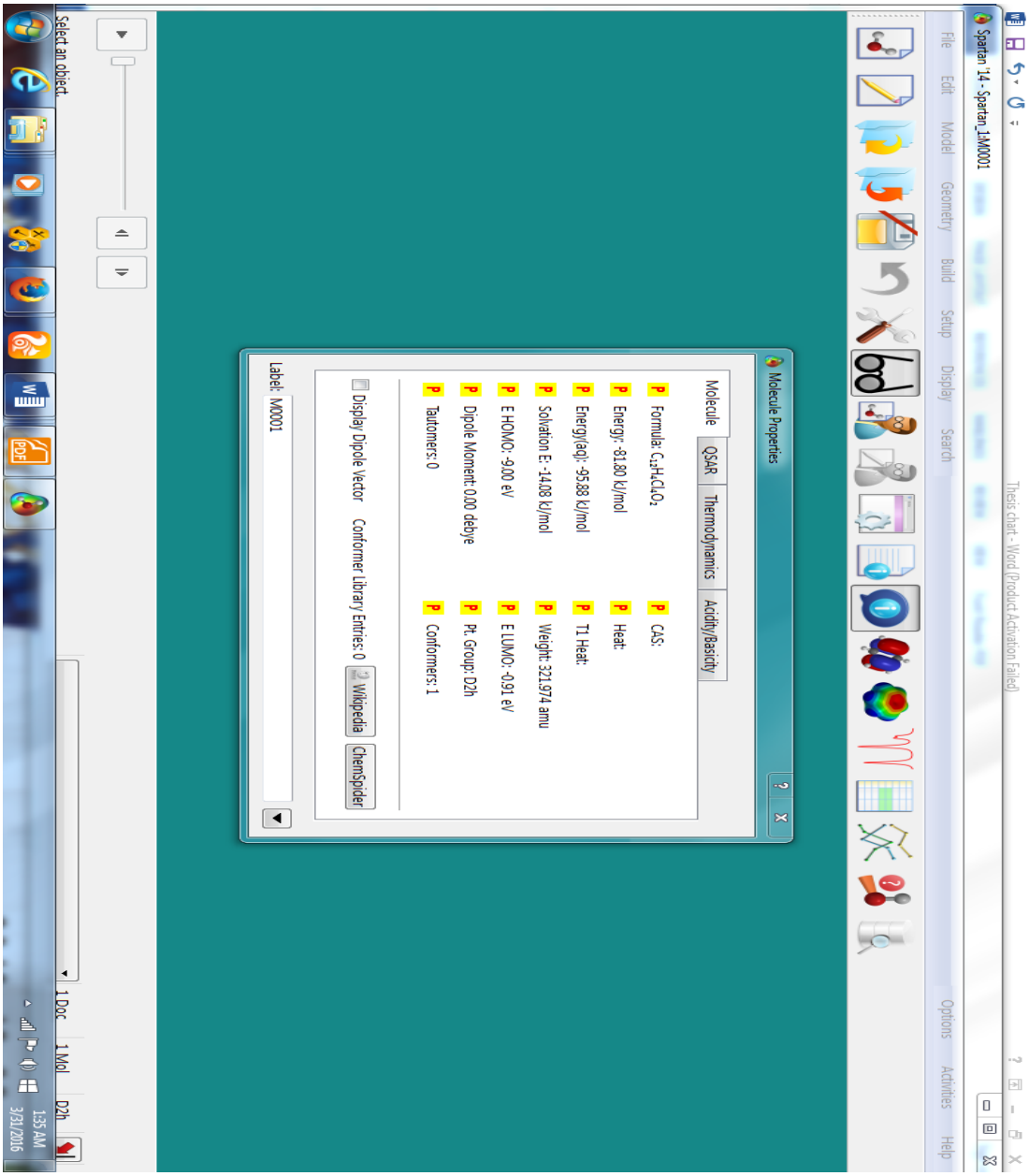


Figure 3.11Dialog box showing the molecular properties (descriptors)

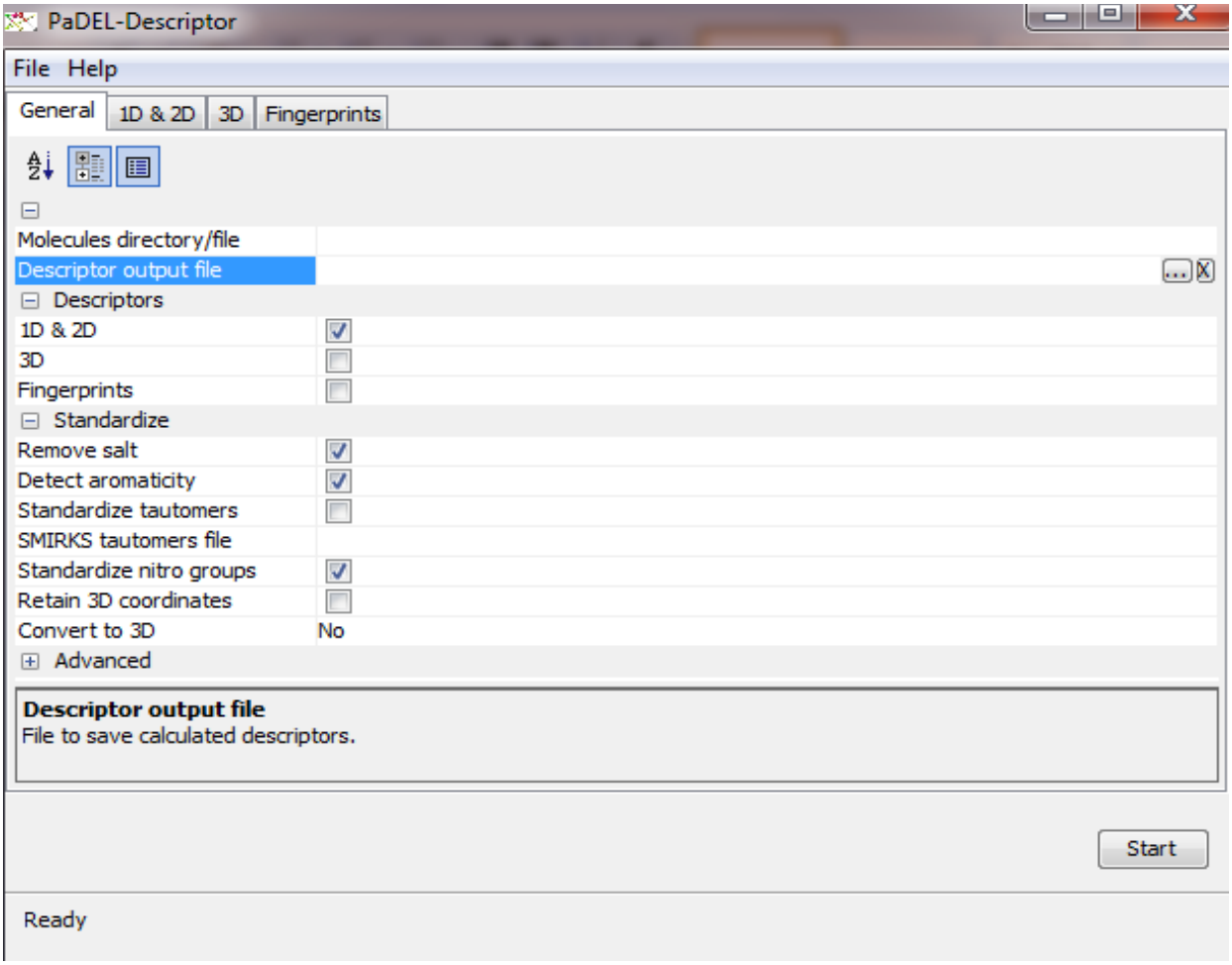


Figure 3.12: Padel-Descriptor interface

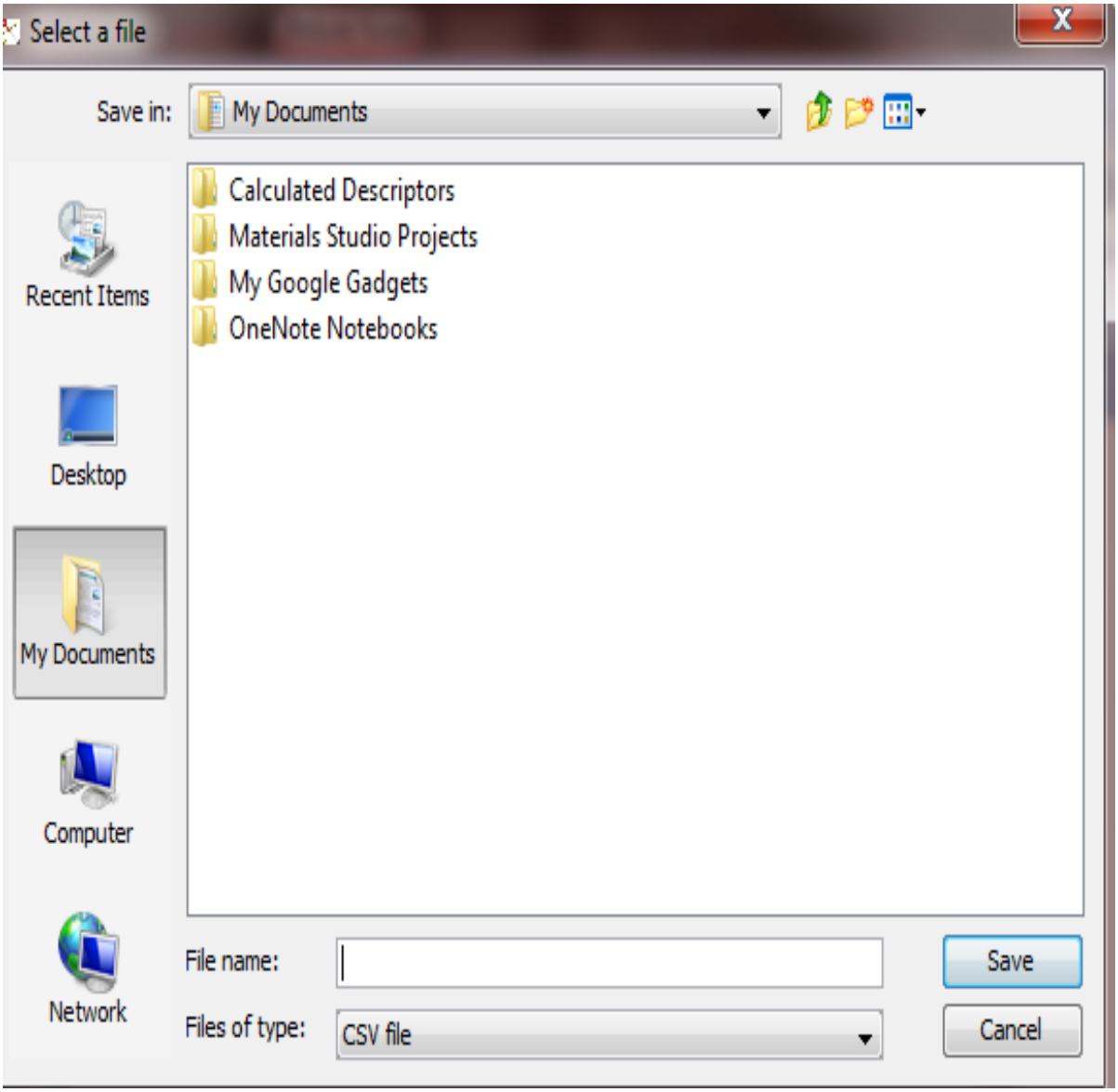


Figure 3.13. Descriptor's file

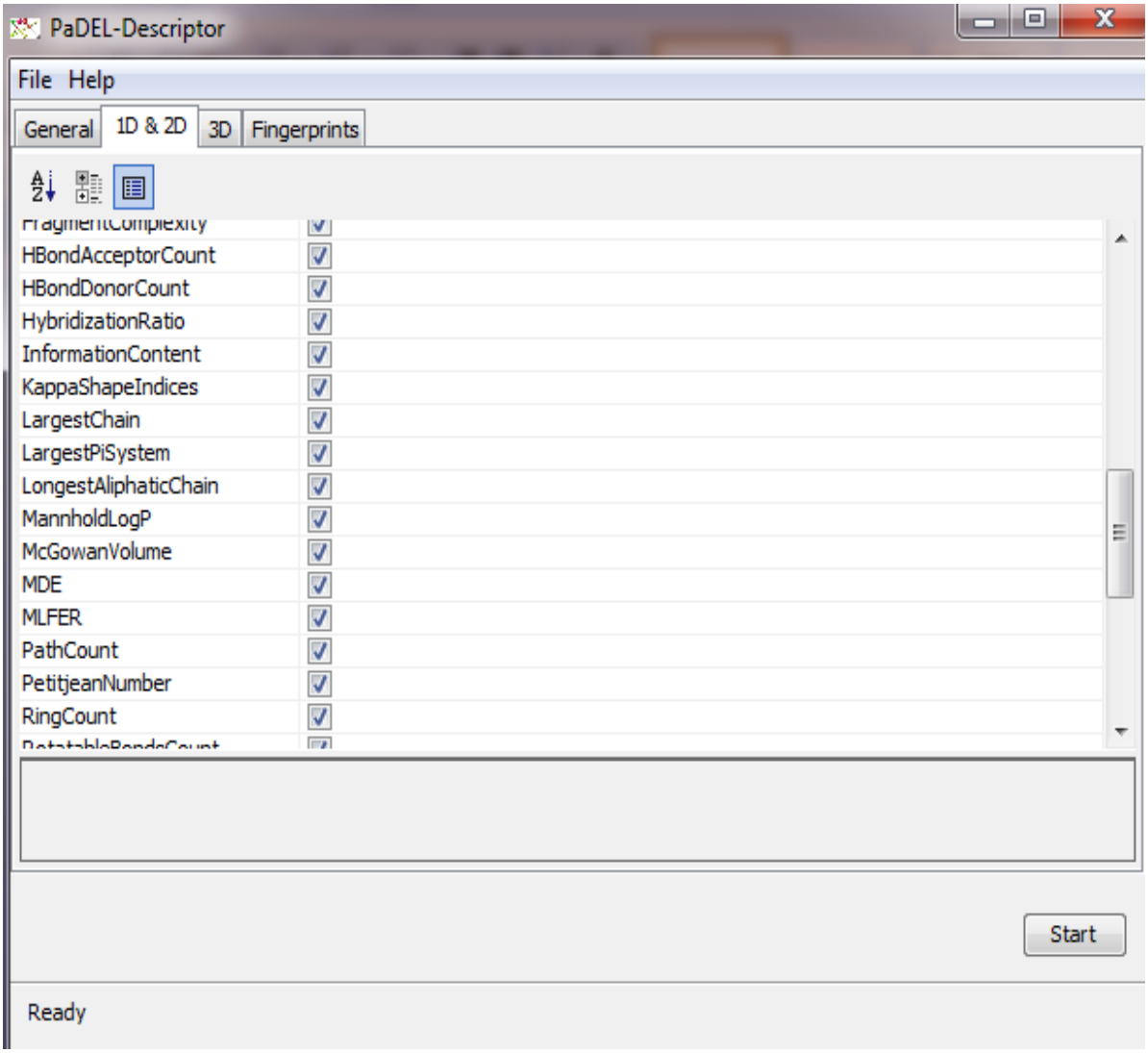


Figure 3.14. Calculation of 1D, 2D and 3D descriptors

3.5 Calculation of molecular descriptors using mathematical formulae

Some of the 2-D descriptors such as electronegativity (χ), absolute hardness (η), and softness (S) of the molecules were calculated using the following formulae:

$$\text{Electronegativity } (\chi) = -\mu = \frac{(IP+EA)}{2} \quad \ldots\ldots\ldots\ldots\ldots 3.2$$

$$\text{Absolute hardness } (\eta) = \frac{(IP-EA)}{2} \quad \ldots\ldots\ldots\ldots\ldots\ldots 3.3$$

where IP and EA are ionization potential and electron affinity respectively of the compounds.

According to the Koopman's theorem(Klopman, 1968), the IP is simply the eigen value of the HOMO with change of sign and the EA is the eigen value of the LUMO with change of sign hence equation (1) and (2) can be written as :

$$\chi = \frac{E_{LUMO}+E_{HOMO}}{2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4$$

$$\eta = \frac{E_{LUMO}-E_{HOMO}}{2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.5$$

The softness (S) which measures the extend of chemical reactivity, is define the reciprocal of hardness.

$$S = \frac{1}{\eta} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.6$$

The energy gap (ΔE) was calculated as:

$$\Delta E = E_{HOMO} - E_{LUMO} \quad \ldots\ldots\ldots\ldots 3.7$$

The following 2-Dimension descriptors were calculated as follows:

$E_{LUMO} - 1,\ E_{LUMO} - 2,\ E_{LUMO} + 1,\ E_{LUMO} + 2$ (Luiz *et al.*, 2011).

Parr et al. proposed an electrophilicity index (ω) as a measure of energy lowering due to maximal electron flow among donor and acceptor (Katrizky *et al.*, 2004). They defined the electrophilic index as follows:

$$\omega = \frac{\mu^2}{2\eta} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.8$$

Molar polarization: the dipole moment induced per unit volume v is called molar polarization Pm and is defined by the Clausius Mossotti equation as follows:

$$\mathrm{Pm} = \frac{\varepsilon - 1}{\varepsilon + 2} \times \frac{MW}{Q} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.9$$

Where MW is the molecular weight, Q is the density, and ε the electric constant which were calculated using Padel descriptor and Spartan.

Sum of square charges on all atoms (SSCA), was calculated as follows:

$$\mathrm{SSCA} = \sum q^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.10$$

Few of the different types of molecular descriptors calculated by Spartan14v.1.1.0 software,Padel descriptor tool kit, and theoretical calculations are presented in Table 3.2.

All the 0D, 1D, 2D, and 3D descriptors calculated by Spartan, Pardel and Theoretical calculations were transferred into the excel sheet in Figure. 3.15,prior to the correlation analysis.

Table 3.2 Various types of descriptors (0D, 1D, 2D and 3D) calculated by the softwares, Spartan"14", Java-jar PaDel descriptor and theoretical calculations

S/N	Descriptor Name	Descriptor Symbol	Descriptor type	Software/ method
1	Total number of atom	nAtom	0-Dimension	PaDel
2	Number of hydrogen atom	nH	0-Dimension	PaDel
3	Number of carbon atom	nC	0-Dimension	PaDel
4	Number of chlorine atom	nCl	0-Dimension	PaDel
5	Number of oxygen atom	nO	0-Dimension	PaDel
6	Number of benzene rings	nBR	0-Dimension	PaDel
7	Molecular weight	Mw	0-Dimension	Spartan "14"
8	Most negative charge in a molecule	Qmin	0-Dimension	PaDel
9	Sum of square of charges on N atom	QNss	0-Dimension	PaDel
10	Total positive charge	TPC	0-Dimension	PaDel
11	Most positive charge on hydrogen atom	QHmax	0-Dimension	PaDel
12	Mean negative charge	MNC	0-Dimension	PaDel
13	Mean positive charge	MPC	0-Dimension	PaDel
14	Most positive charge on oxygen	QOmax	0-Dimension	PaDel
15	Number of halogens	nX	0-Dimension	PaDel
16	Number of single bonds	nBondsS	1-Dimension	PaDel

	including bond of H			
17	Number of single bonds excluding bond to H and aromatic bonds	nBondsS3	1-Dimension	PaDel
18	Number of double bonds	nBondsD	1-Dimension	PaDel
19	Number of aromatic bonds	nAromBond	1-Dimension	PaDel
20	Number of aromatic atoms	naAromAtom	1-Dimension	PaDel
21	Number of alkyl group	nAl	1-Dimension	PaDel
22	Number of heavy atom not halogen	nHeavy	1-Dimension	PaDel
23	Overall solute H-bond basicity	MLFR-B	2-Dimension	PaDel
24	Number of basic group	nBase	2-Dimension	PaDel
25	Overall solute H-bond acidity	MLFER-A	2-Dimension	PaDel
26	Excessive molar refraction	MLFER-E	2-Dimension	PaDel
27	Energy of HOMO	E_{HOMO}	2-Dimension	Spartan "14"
28	Energy of LUMO	E_{LUMO}	2-Dimension	Spartan "14"
29				
30	Dipole moment	DM	2-Dimension	Spartan "14"
31	Electrostatic charge	Eq	2-Dimension	Spartan "14"
32	Mulliken charge	Mq	2-Dimension	Spartan "14"
33	Heat of formation	Hq	2-Dimension	Spartan "14"

34	Standard free energy	G^o	2-Dimension	Spartan "14"
35	Standard enthalpy	H^o	2-Dimension	Spartan "14"
36	Standard entropy	S^o	2-Dimension	Spartan "14"
37	Heat of vaporization	Hv	2-Dimension	Spartan "14"
38	Electronegativity	□	2-Dimension	Theoretical calculation
39	Absolute hardness	□	2-Dimension	Theoretical calculation
40	Softness	S	2-Dimension	Theoretical calculation
41	Energy gap	ΔE	2-Dimension	Theoretical calculation
42	Solute-gas hexadecane partition coefficient	MLFER-L	2-Dimention	PaDel
43	Number of acidic group	nAcid	2-Dimention	PaDel
44	Sum of atomic polarizability	aPol	2-Dimention	PaDel
45	Crippen's logP	ClogP	2-Dimention	PaDel
46	Number of H-bond accetor CDK H-bond acceptor count descriptor algorithm	nHBACC	2-Dimention	PaDel
47	Number of H-bonding acceptor	nHBAC3	2-Dimention	PaDel
48	Fragment complexity	FragC	2-Dimention	PaDel

49	Eccentric connectivity	ECCEN	2-Dimention	PaDel
50	Mannhold log	MlogP	2-Dimention	PaDel
51	Number of atoms in largest pi system	nAtomP	2-Dimention	PaDel
52	Hybridization ratio	HybRatio	2-Dimention	PaDel
64	Sum of valence degree	SVD	2-Dimention	PaDel
65	Partition coefficient	Log P	2-Dimention	PaDel
66	Connolly solvent excluded volume	CSV	2-Dimention	PaDel
67	Crippen logP	Crippen logP	2-Dimention	PaDel
68	Crippen molar refractivity	Crippen MR	2-Dimention	PaDel
69	Molar refractivity	MR	2-Dimention	PaDel
70	Number of rotatable bond	nRotB	2-Dimention	PaDel
71	Fraction of rotatable bond excluding terminal bond	RotBTFrac	2-Dimention	PaDel
72	Fraction of rotatable bond including terminal bond	RotBTfrac	2-Dimention	PaDel
73	Number of rotatable bond including terminal bond	nRotBT	2-Dimention	PaDel
74	Topological radius	Toporadius	2-Dimention	PaDel
75	Topological diameter	Topodiameter	2-Dimention	PaDel
76	Petijean's topological shape index	Topo shape	2-Dimention	PaDel

77	Global topological charge index	GTCI	2-Dimention	PaDel
78	Topological polar surface area	Topo PSA	2-Dimention	PaDel
79	Vertex adjacent information	VAdjMat	2-Dimention	PaDel
80	Molecular ID	WTPT-1	2-Dimention	PaDel
81	Molecular ID/ number of atoms	WTPT-2	2-Dimention	PaDel
82	Sum of path length starting from hero atoms	WTPT-3	2-Dimention	PaDel
83	First kappa shape index	Kier1	2-Dimention	PaDel
84	Second kappa shape index	Kier2	2-Dimention	PaDel
85	Third kappa shape index	Kier3	2-Dimention	PaDel
86	Valence path descriptor	VP-1	2-Dimention	PaDel
87	Wiener path number	WPATH	2-Dimention	PaDel
88	Wiener polarity number	WPOL	2-Dimention	PaDel
89	Sum of squares of atom degree over all heavy atoms	Zagreb	2-Dimention	PaDel
90	Density	ρ	3-Dimension	Spartan “14”
91	Dielectric constant	k	3-Dimension	Spartan “14”
92	Surface tension	T	3-Dimension	Spartan “14”
93	Molar volume	Vm	3-Dimension	Spartan “14”
94	Molar surface area	Am	3-Dimension	Spartan “14”

95	Octanol/waterpartition coefficient	logP	3-Dimension	Spartan “14”
96	Freezing point	FP	3-Dimension	Chm3Dpro 21.0v
97	Boiling point	BP	3-Dimension	Chm3Dpro 21.0v
98	Melting point	MP	3-Dimension	Chm3Dpro 21.0v
99	Critical temperature	CT	3-Dimension	Chm3Dpro 21.0v
100	Critical pressure	CP	3-Dimension	Chm3Dpro 21.0v
101	Critical volume	CV	3-Dimension	Chm3Dpro 21.0v
102	Wiener index	WI	3-Dimension	Chm3Dpro 21.0v
103	Total valence connectivity	TVC	3-Dimension	Chm3Dpro 21.0v
104	Sum of degree radius	SDR	3-Dimension	Chm3Dpro 21.0v
105	Connolly solvent	CS	3-Dimension	Chm3Dpro 21.0v
106	Gravitational index of heavy	GRAV-1	3-Dimension	PaDel

	atom			
107	Gravitational index-Hydrogen included	GRAVH-1	3-Dimension	PaDel
108	Moment of inertia along X-axis	MOMI-X	3-Dimension	PaDel
109	Moment of inertia along Y-axis	MOMI-Y	3-Dimension	PaDel
110	Moment of inertia along Z-axis	MOMI	3-Dimension	PaDel
111	Radius of gyration	MOMI-R	3-Dimension	PaDel
112	Geometrical radius	geomRadius	3-Dimension	PaDel
113	Geometrical diameter	geomDiamter	3-Dimension	PaDel
114	Petitjean geometrical shape	geomShape	3-Dimension	PaDel

Molecule	Pec50	EHOMO	ELUMO	ELUMO-1	ELUMO-2	ELUMO+1	ELUMO+2	DE	n	S	X	DM	E	Eaq	MW	BD	CBD	SE	MA	MV
1	8	-484.41	-173.31	-174.31	-175.31	-172.31	-171.31	-311.1	155.55	0.006429	-328.86	0	-6434144	6434157	341.97	1	1	-13.36	263.12	24
2	7.1	-495.95	-187.14	-188.14	-189.14	-186.14	-185.14	-308.81	154.405	0.006476	-341.545	1.0107	-7640757	-7640768	356.419	1	1	-11.03	277.55	25
3	6.8	-496.12	-186.12	-187.12	-188.12	-185.12	-184.12	-310	155	0.006452	-341.12	1.53	-7640752	-7640764	356.41	1	1	-12.5	277.14	25
4	6.66	-472.85	-155.9	-156.9	-157.9	-154.9	-153.9	-316.95	158.475	0.00631	-314.375	3.13	-5227513	-5227531	287.529	1	1	-17.76	247.62	23
5	6.55	-506.61	-202.87	-203.87	-204.87	-201.87	-200.87	-303.74	151.87	0.006585	-354.74	0.26	-8847369	-8847376	390.864	1	1	-6.87	291.88	27
6	6.1	-488.45	-175.75	-176.75	-177.75	-174.75	-173.75	-312.7	156.35	0.006396	-332.1	1.05	-6434146	-6434158	321.974	1	1	-12.16	263.42	24
7	5.96	-500.01	-190.14	-191.14	-192.14	-189.14	-188.14	-309.87	154.935	0.006454	-345.075	0.85	-7640761	-7640770	356.419	1	1	-9.65	278.13	25
8	5.89	-471.61	-157.58	-158.58	-159.58	-156.58	-155.58	-314.03	157.015	0.006369	-314.595	1	-5227520	-5227536	287.529	1	1	-15.79	247.96	23
9	7.15	-479.12	-182.5	-183.5	-184.5	-181.5	-180.5	-296.62	148.31	0.006743	-330.81	3.49	-6434111	-6434123	321.974	1	1	-12.67	260.88	24
10	5.17	-495	-194	-195	-196	-193	-192	-301	150.5	0.006645	-344.5	1.77	-7640747	-7640756	356.419	1	1	-9.49	276.73	25
11	4.89	-471.01	-155.68	-156.68	-157.68	-154.68	-153.68	-315.33	157.665	0.006343	-313.345	2.58	-5227503	-5227519	287.529	1	1	-15.57	247.08	23
12	5.49	-457.37	-140.5	-141.5	-142.5	-139.5	-138.5	-316.87	158.435	0.006312	-298.935	1.81	-4020896	-4020914	253.084	1	1	-18.47	232.76	21
13	5	-526.82	-232.55	-233.55	-234.55	-231.55	-230.55	-294.27	147.135	0.006796	-379.685	0	-1.1E+07	-1.1E+07	459.754	1	1	-13.01	320.7	2
14	4	-439.9	-115.57	-116.57	-117.57	-114.57	-113.57	-324.33	162.165	0.006167	-277.735	1.71	-2814252	-2814277	218.639	1	1	-24.37	216.5	20
15	8.82	-486.39	-183.59	-184.59	-185.59	-182.59	-181.59	-302.8	151.4	0.006605	-334.99	0	-2.9E+07	-2.9E+07	499.778	1	1	-19.99	280.91	26
16	8.83	-485.42	-179.01	-180.01	-181.01	-178.01	-177.01	-306.41	153.205	0.006527	-332.215	0.14	-1.8E+07	-1.8E+07	410.876	1	1	-16.65	272.01	25
17	9.35	-485.41	-175.74	-176.74	-177.74	-174.74	-173.74	-309.67	154.835	0.006458	-330.575	0.02	-1753205	-1.8E+07	410.876	1	1	16.6	272.02	25
18	7.94	484.91	-174.58	-175.58	-176.58	-173.58	-172.58	659.49	-329.745	-0.00303	155.165	0.07	-1.2E+07	-1.2E+07	366.425	1	1	-14.98	267.57	24
19	7.03	-493.16	-180.85	-181.85	-182.85	-179.85	-178.85	-312.31	156.155	0.006404	-337.005	0.96	-2.9E+07	-2.9E+07	499.778	1	1	19.27	281.91	2
20	8.7	-499.36	186.46	185.46	184.46	187.46	188.46	-685.82	342.91	0.002916	-156.45	0.86	-2.9E+07	-2.9E+07	499.77	1	1	-17.29	279.11	26
21	7.77	-214.64	-500.19	-501.19	-502.19	-499.19	-498.19	285.55	-142.775	-0.007	-357.415	0.67	-3.5E+07	-3.5E+07	578.674	1	1	-17.72	300.92	28
22	8.18	-496.28	-231.5	-232.5	-233.5	-230.5	-229.5	-264.78	132.39	0.007553	-363.89	0.87	-3.5E+07	-3.5E+07	578.674	1	1	-19.08	299.31	2
23	8.93	-473.98	-172.19	-173.19	-174.19	-171.19	-170.19	-301.79	150.895	0.006627	-323.085	1.48	-2.2E+07	-2876136	420.88	1	1	-20.94	261.48	2
24	7.81	-460.69	-144.8	-145.8	-146.8	-143.8	-142.8	-315.89	157.945	0.006331	-302.745	0	-1.5E+07	-1.5E+07	341.98	1	1	-21.79	241.93	2
25	6.53	-444.82	-118.82	-119.82	-120.82	-117.82	-116.82	-326	163	0.006135	-281.82	1.71	-8363813	-8363839	263.09	1	1	-25.46	220.93	20
26	4.53	-503.57	-156.68	-157.68	-158.68	-155.68	-154.68	-346.89	173.445	0.005766	-330.125	1	-2616841	-2616862	202.64	1	1	-21.3	207.14	19
27	3.55	-504.68	-164.24	-165.24	-166.24	-163.24	-162.24	-340.44	170.22	0.005875	-334.46	1.9	-2616840	-2616861	202.64	1	1	-20.25	208.35	19
28	4.38	-500.89	-160.89	-161.89	-162.89	-159.89	-158.89	-340	170	0.005882	-330.89	2.37	-2616840	-2616861	202.64	1	1	-20.64	208.34	19
29	3	-506.92	-158.73	-159.73	-160.73	-157.73	-156.73	-348.19	174.095	0.005744	-332.825	2.42	-2616834	-2616857	202.64	1	1	-22.98	208.8	19

Figure 3.15 Collection of the calculated descriptors using all the methods.

3.6 Selection and Elimination of Molecular Descriptors

After the descriptors have been obtained, it is assumed that many of these descriptors were more or less co-linear or having the same value for all compounds, that in turn cause problems for many statistical analysis method. In order to obtain the best model possible, variable elimination and selection techniques was performed. Variable elimination is the processwhere variables

are physically removed from the dataset. Variables are eliminated for two reasons. First, they are eliminated if they have a small variance, below some threshold value. The second reason for variable removal is the existence of redundancy (exact linear dependencies between subsets of the variables) and multicollinearity (high multiple correlations between subsets of the variables). Thus, variable selection is the process where one searches for the best descriptor combinations (Sofie, 2009). The most spurious variables and poor variables were physically removed and the most informative once were selected. The best descriptor combination were then subjected to regression analysis (MLR).

3.7 Data Set Division

The data was randomly split into training set and test set using *datasetdivisionGUI1.2* software. From the folder containing the software, double clicking on the software file, made the interface of the software appear as shown in Figure. 3.16. By clicking on *browse*option at right side of the interface corresponding to the icon '*select input file'* at the left side of the software, the*choose file* dialog box where the file containing the data set was imported into the software as shown in Figure. 3.17. Then, clicking on the *browse* option corresponding to the icon '*select output directory',* in Figure 3.16,the folder where the result division of data set should be kept appeared as presented in Figure. 3.18. By clicking on the icon *start* at the bottom of Figure 3.16,

Figure 3.16: The interface for dataset division software

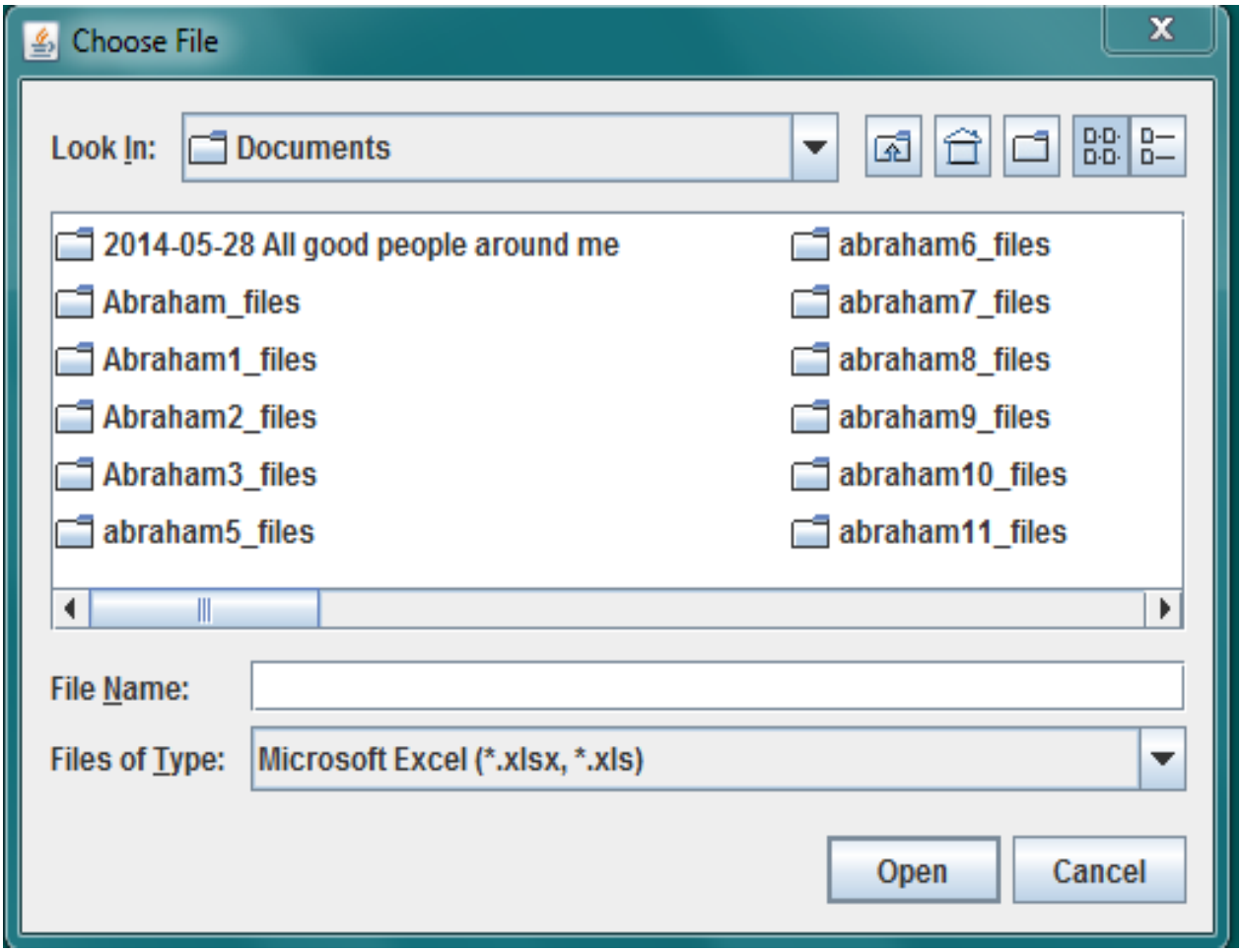


Figure.3.17:“Choose file” dialog box

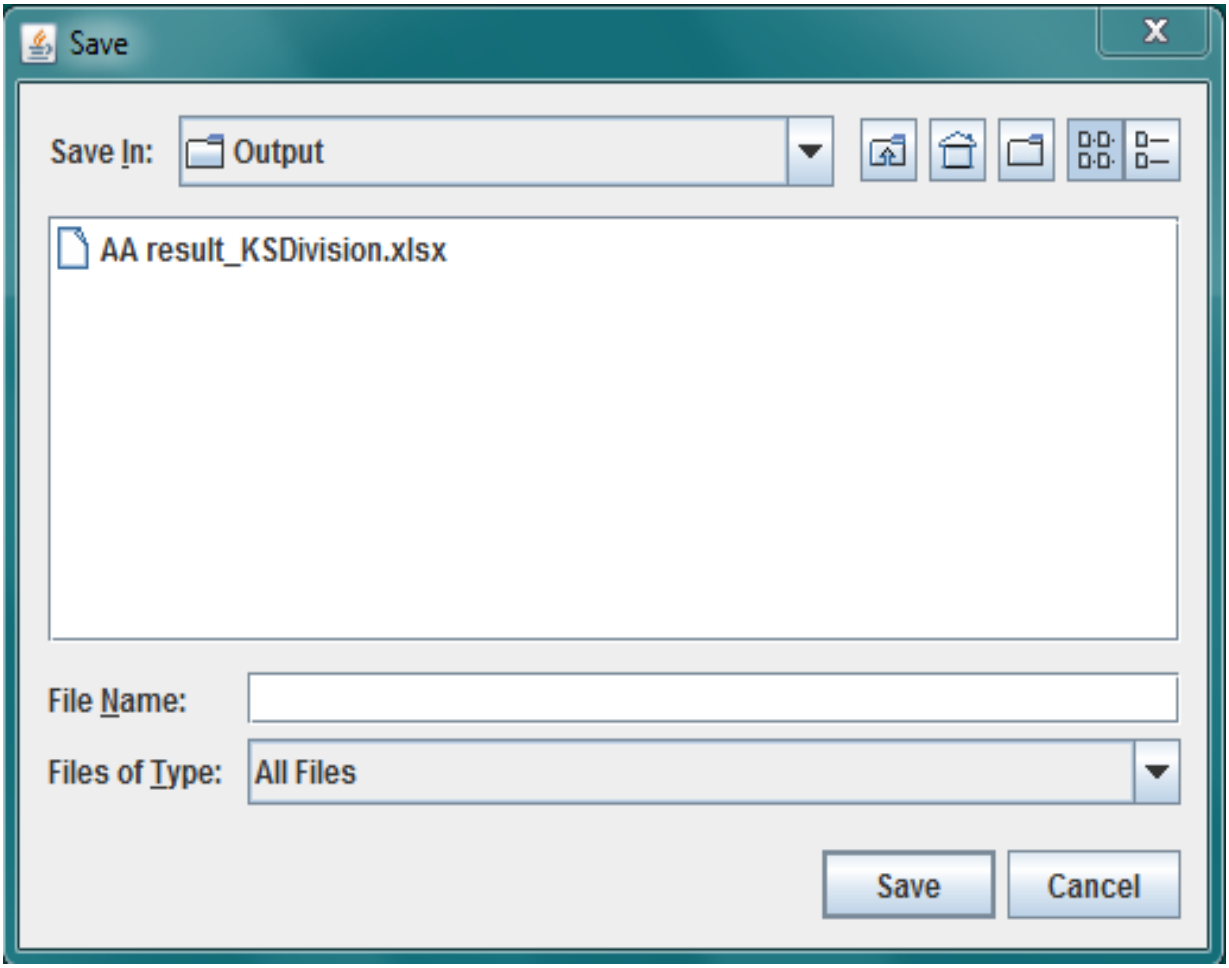


Figure 3.18:“Save file” dialog box

a dialog box appeared for specifying the percent of training set (shown in Figure. 3.19) and finally the Ok option was clicked and result was written in the file chosen. The training set was used in developing the model while the test set was used to evaluate its prediction ability.70% of the data set in each case was used as training set and the remaining 30% as test se (see Figure. 3.19).

3.8 Correlation Analysis using Material Studio 7.0

Double clicking on the *Material Studio 7.0*filelocated on the desktop of the computer, the interface of the software was displayed on the screen of the computer (Figure 3.20). Then clicking on the icon *New,* located at left of the menu bar, a new page in form of an excel sheet was displayed (Figure 3.21). The large number of selected molecular descriptors saved in an excel sheet was imported into the sheet of Material Studio 7.0 with the toxicities of the compounds occupying the first column, then all descriptors including the toxicities were highlighted (Figure 22). By clicking on the icon *statistics,* located on the menu bar, five options (initial analysis, data reduction, cluster analysis, cluster validation and model building) appeared. Upon the selection of the option *initial analysis,* four other options (univariate analysis, standardize data, transform data and correlation matrix) appeared. Clicking on the option *correlation matrix,* a large square matrix of about 450 by 450 showing the correlation of toxicity of the compounds with descriptors was displayed as presented in Figure 3.23. All the correlation coefficients within the range of 0.10 to 0.99 were checked prior to the model building.

3.9 Model Building using Genetic Functional Approximation (GFA)

In this work, all the models were developed using genetic function approximation (GFA) technique. The genetic function approximation algorithm was first anticipated by: (1) Holland's genetic algorithm (Holland, 1975) and (2) Friedman's multivariate adaptive regression splines

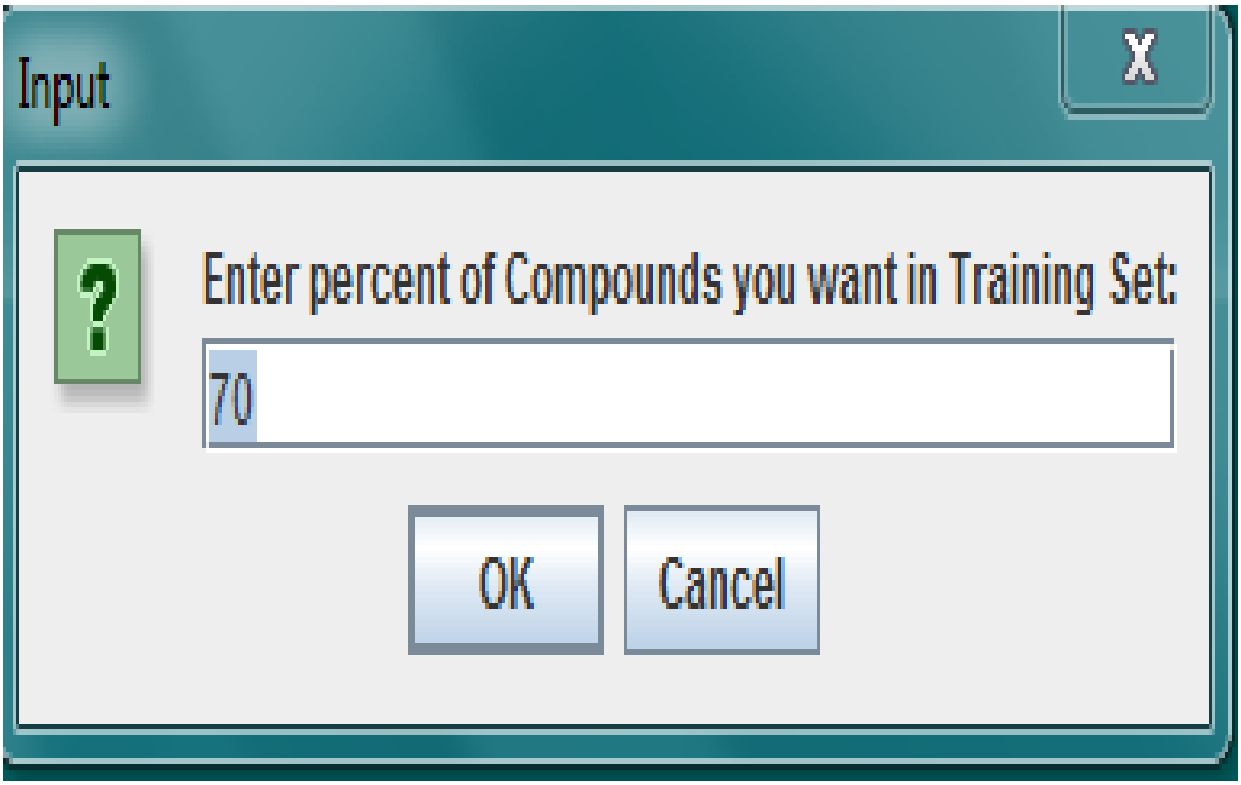


Figure 3.19: Dialog box for imputing the percent of the training set

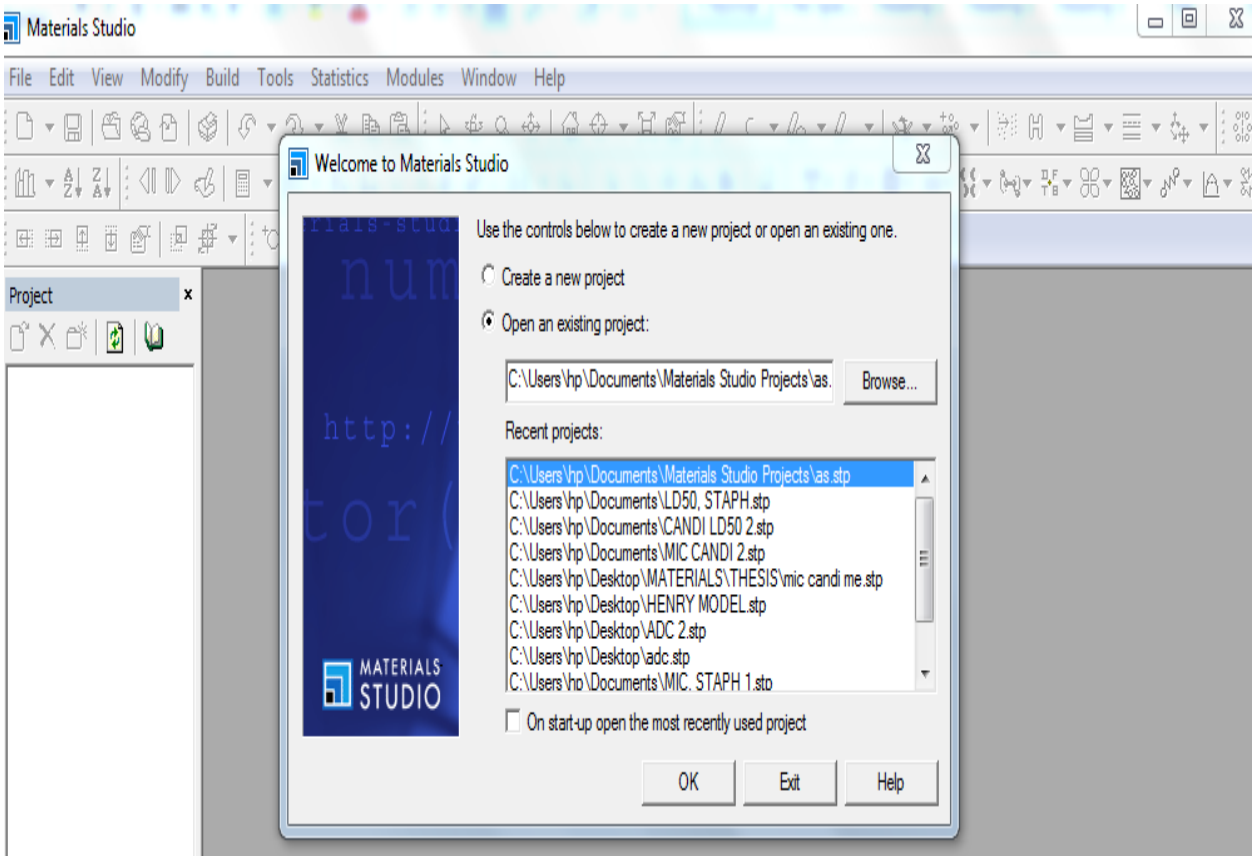


Figure 3.20Material Studio's work space window and project window

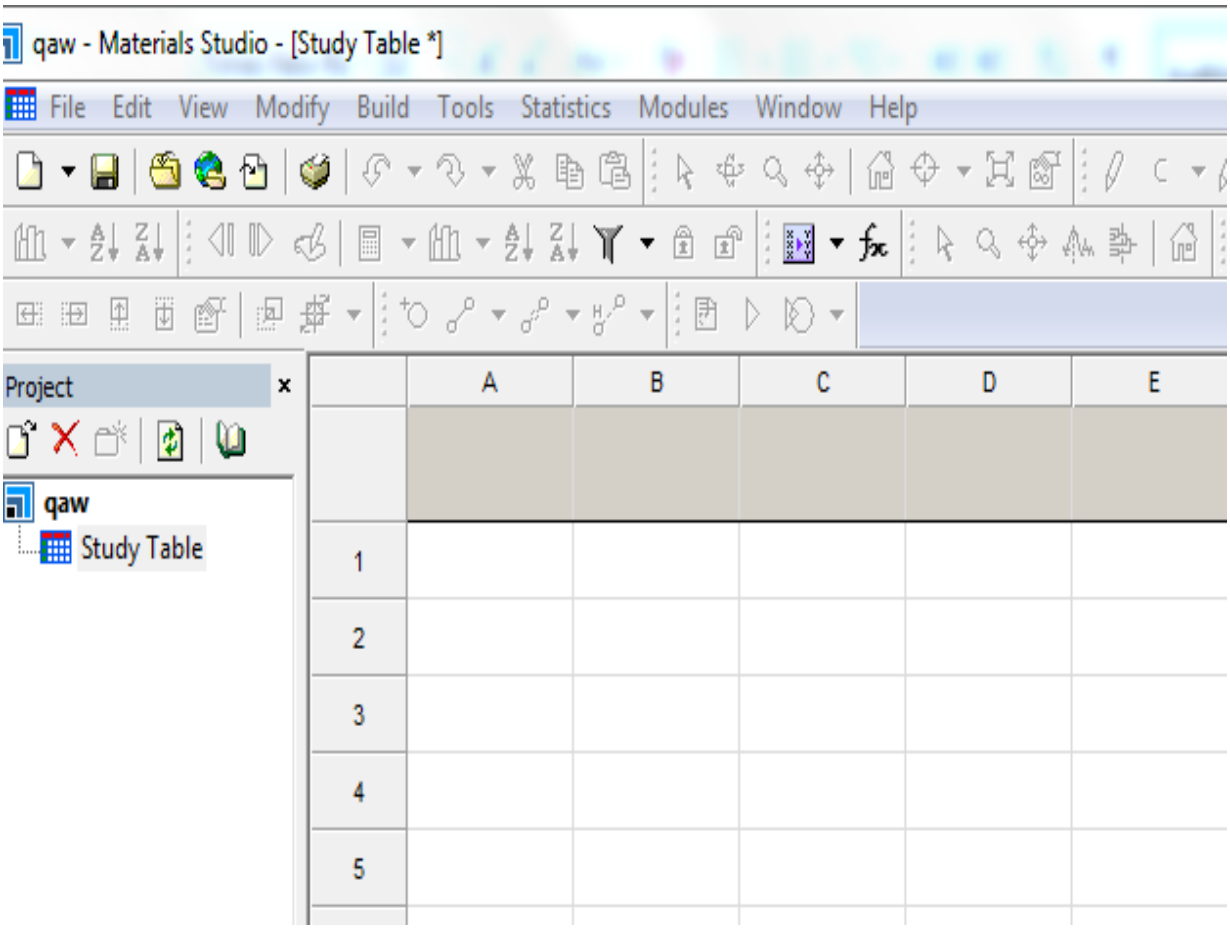


Figure 3.21Material Studio's *Study table* window

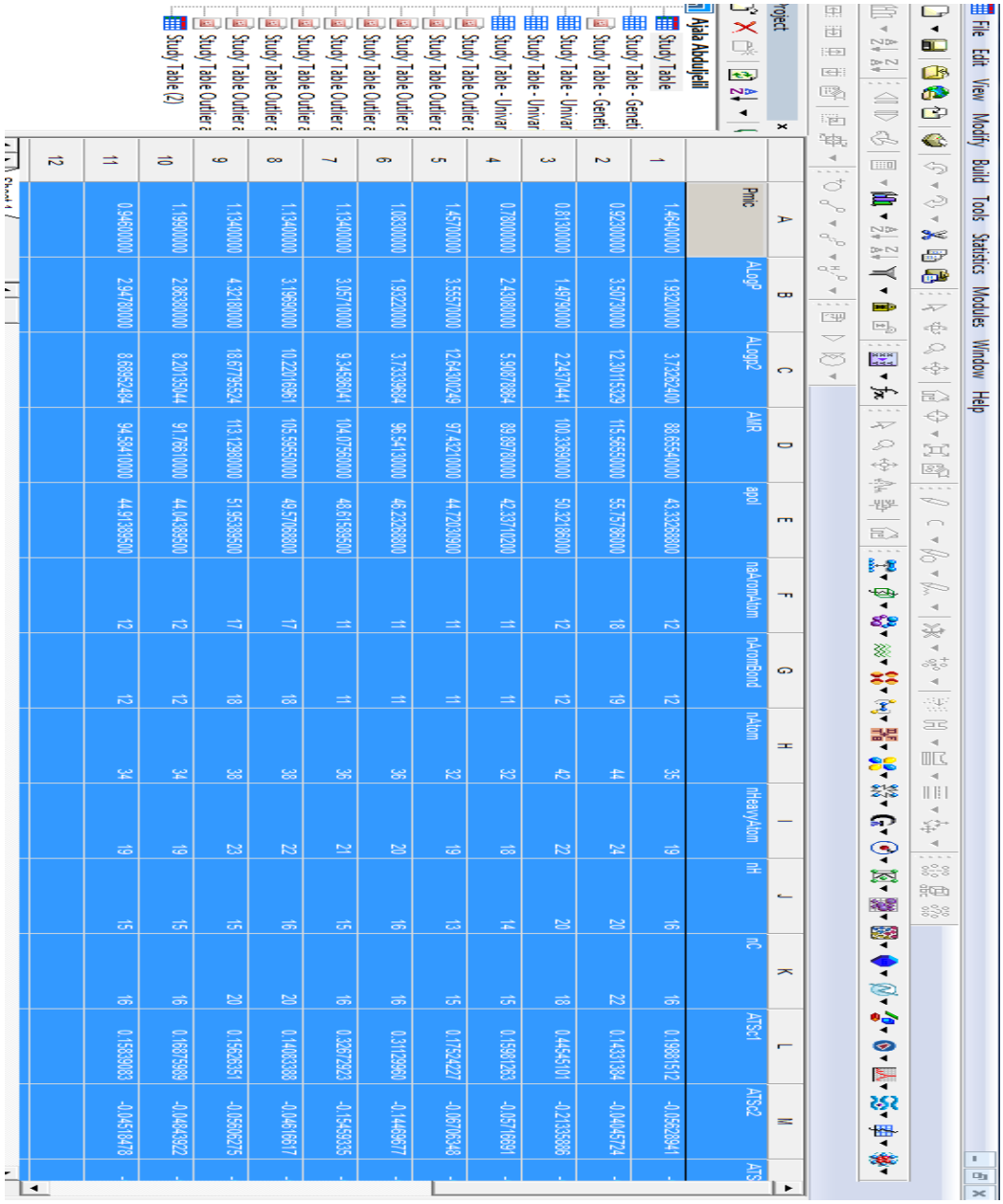
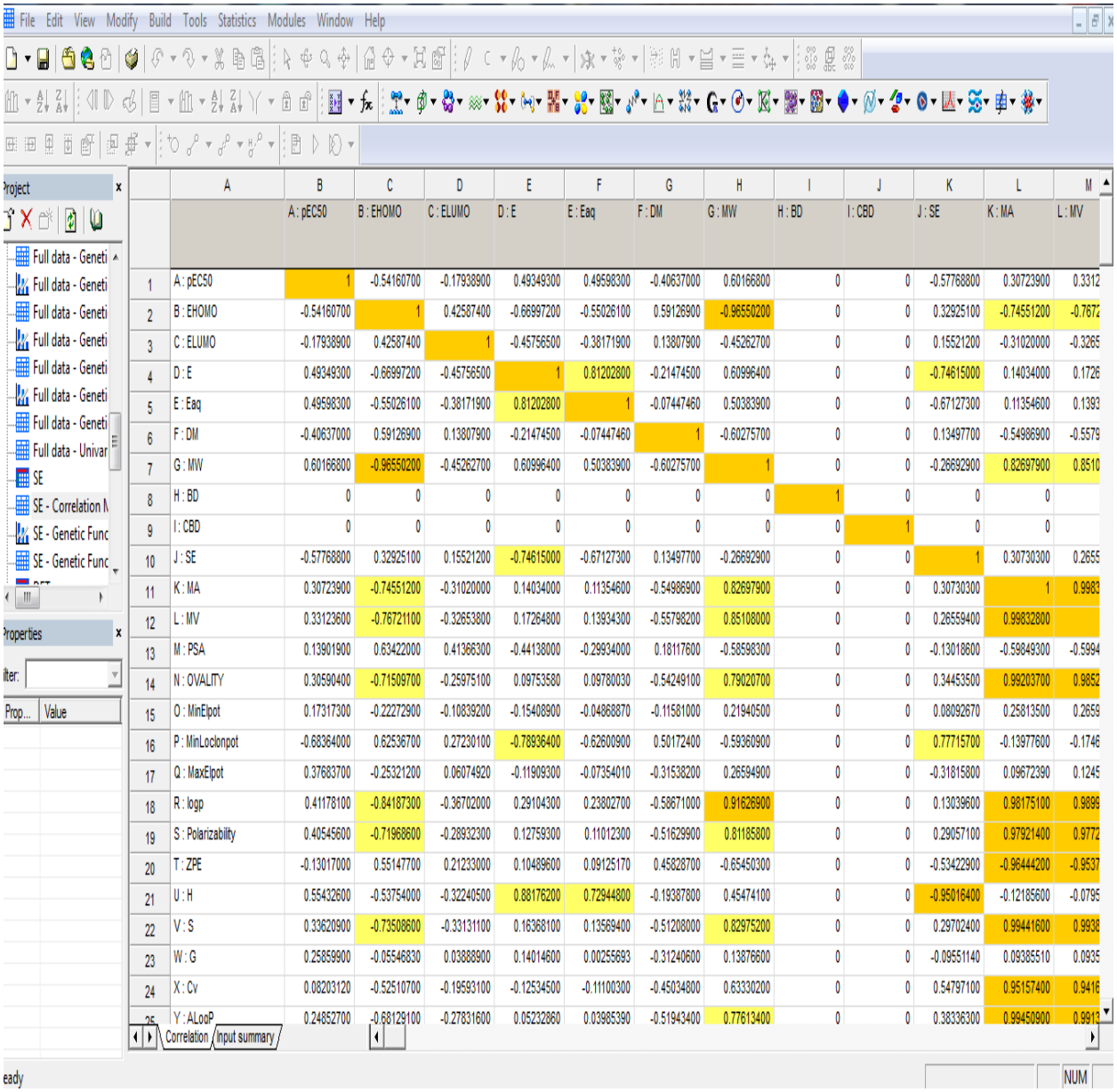


Figure 3.22 Highlighted data for the correlation analysis



	A	B	C	D	E	F	G	H	I	J	K	L	M
		A : pEC50	B : EHOMO	C : ELUMO	D : E	E : Eaq	F : DM	G : MW	H : BD	I : CBD	J : SE	K : MA	L : MV
1	A : pEC50	1	-0.54160700	-0.17938900	0.49349300	0.49598300	-0.40637000	0.60166800	0	0	-0.57768800	0.30723900	0.3312
2	B : EHOMO	-0.54160700	1	0.42587400	-0.66997200	-0.55026100	0.59126900	-0.96550200	0	0	0.32925100	-0.74551200	-0.7672
3	C : ELUMO	-0.17938900	0.42587400	1	-0.45756500	-0.38171900	0.13807900	-0.45262700	0	0	0.15521200	-0.31020000	-0.3265
4	D : E	0.49349300	-0.66997200	-0.45756500	1	0.81202800	-0.21474500	0.60996400	0	0	-0.74615000	0.14034000	0.1726
5	E : Eaq	0.49598300	-0.55026100	-0.38171900	0.81202800	1	-0.07447460	0.50383900	0	0	-0.67127300	0.11354600	0.1393
6	F : DM	-0.40637000	0.59126900	0.13807900	-0.21474500	-0.07447460	1	-0.60275700	0	0	0.13497700	-0.54986900	-0.5579
7	G : MW	0.60166800	-0.96550200	-0.45262700	0.60996400	0.50383900	-0.60275700	1	0	0	-0.26692900	0.82697900	0.8510
8	H : BD	0	0	0	0	0	0	0	1	0	0	0	
9	I : CBD	0	0	0	0	0	0	0	0	1	0	0	
10	J : SE	-0.57768800	0.32925100	0.15521200	-0.74615000	-0.67127300	0.13497700	-0.26692900	0	0	1	0.30730300	0.2655
11	K : MA	0.30723900	-0.74551200	-0.31020000	0.14034000	0.11354600	-0.54986900	0.82697900	0	0	0.30730300	1	0.9983
12	L : MV	0.33123600	-0.76721100	-0.32653800	0.17264800	0.13934300	-0.55798200	0.85108000	0	0	0.26559400	0.99832800	
13	M : PSA	0.13901900	0.63422000	0.41366300	-0.44138000	-0.29934000	0.18117600	-0.58598300	0	0	-0.13018600	-0.59849300	-0.5994
14	N : OVALITY	0.30590400	-0.71509700	-0.25975100	0.09753580	0.09780030	-0.54249100	0.79020700	0	0	0.34453500	0.99203700	0.9852
15	O : MinElpot	0.17317300	-0.22272900	-0.10839200	-0.15408900	-0.04868870	-0.11581000	0.21940500	0	0	0.08092670	0.25813500	0.2659
16	P : MinLocIonpot	-0.68364000	0.62536700	0.27230100	-0.78936400	-0.62600900	0.50172400	-0.59360900	0	0	0.77715700	-0.13977600	-0.1746
17	Q : MaxElpot	0.37683700	-0.25321200	0.06074920	-0.11909300	-0.07354010	-0.31538200	0.26594900	0	0	-0.31815800	0.09672390	0.1245
18	R : logp	0.41178100	-0.84187300	-0.36702000	0.29104300	0.23802700	-0.58671000	0.91626900	0	0	0.13039600	0.98175100	0.9899
19	S : Polarizability	0.40545600	-0.71968600	-0.28932300	0.12759300	0.11012300	-0.51629900	0.81185800	0	0	0.29057100	0.97921400	0.9772
20	T : ZPE	-0.13017000	0.55147700	0.21233000	0.10489600	0.09125170	0.45828700	-0.65450300	0	0	-0.53422900	-0.96444200	-0.9537
21	U : H	0.55432600	-0.53754000	-0.32240500	0.88176200	0.72944800	-0.19387800	0.45474100	0	0	-0.95016400	-0.12185600	-0.0795
22	V : S	0.33620900	-0.73508600	-0.33131100	0.16368100	0.13569400	-0.51208000	0.82975200	0	0	0.29702400	0.99441600	0.9938
23	W : G	0.25859900	-0.05546830	0.03888900	0.14014600	0.00255693	-0.31240600	0.13876600	0	0	-0.09551140	0.09385510	0.0935
24	X : Cv	0.08203120	-0.52510700	-0.19593100	-0.12534500	-0.11100300	-0.45034800	0.63330200	0	0	0.54797100	0.95157400	0.9416
25	Y : ALogP	0.24852700	-0.68129100	-0.27831600	0.05232860	0.03985390	-0.51943400	0.77613400	0	0	0.38336300	0.99450900	0.9913

Figure 3.23 Square correlation matrix between the pEC_{50} and the calculated descriptors

(MARS) algorithm (Rogers, 1992). In Genetic Function algorithm, an individual model is represented as one dimensional string of bits (Raichurkar *et al.,* 2003). A distinctive feature of GFA is that it produces a population of several models instead of generating a single model, as do most other statisticalmethods. Genetic algorithm makes superior models to those developed using stepwiseregression techniques because it selects the basis functions genetically (Niculescu, 2003).

3.9.1 Procedure for Building GFA Model

On the desktop of the computer, the Material Studio software was launched by double clicking on the material studio software. The windows, work space and project windows appeared as presented in Figure3.24. On Figure 3.24interface, "create a new project" was followed by clicking the Ok button. A new document dialog box containing study table appeared. This is presented in Figure3.25. Selecting the study table and then clicking the option *Ok* button, the material studio study table appeared as shown in Figure. 3.26. The collection of all the descriptors along with values of pEC_{50} saved in an excel sheet was imported into the spread sheet of the material studio study table.From the spread sheet containing the toxicities (EC_{50}) of the training set compounds in the first column and descriptors in subsequent columns on the interface of the Material studio 7.0 software, the column containing the toxicities (EC_{50}) was highlighted as presented in Figure. 3.27. By highlighting the *statistic menu* on the menu bar, five options (initial analysis, data reduction, cluster analysis, cluster validation and model building) appeared. By highlighting the option *model building,* five other options (Multiple Linear Regression, Partial Least Square, Genetic Function Approximation and Outlier Analysis) appeared. By clicking on the option *Genetic Function Approximation,* dialog box appeared showing four options at the upper part of the dialog box (variable selection, equation data,

parameters and advanced). By clicking on the option *variable selection*, all the descriptors correlating with EC50 within the range of correlation coefficient of 0.10 to 0.99 (predictor variables) were marked. By clicking on the option *equation data* ,initial and the maximum equation length were selected as 10 for each. By clicking on the option *parameters,* population,

maximum generation and number of top equation were selected as 100, 500 and 10 respectively. Mutation probability was set to 0.1. By clicking on the option *advanced*, the number of variable to plot and plot interval were selected as 10 and 5 respectively. This is presented in Figure 3.28. By clicking *Ok,* calculation began immediately. The results of the calculation were displayed in a tabular form at the bottom of the study table.

QSAR equations (models), generated by GFA were mathematical relationship between the sets of the descriptors (independent variables), x_{ij} and the pEC50 of toxicity values (dependent variables), y_i. for compound i, the linear relationship relating descriptors, $x_{i,}$, x_{i2}, x_{i3} ... to toxicity values, y_i, is given as:

$$y_i = x_{i1}m_i + x_{i2}m_2 + \cdots + x_{in}b_n + e_i \ldots\ldots\ldots\ldots\ldots\ldots \quad 3.11$$

Where m is the linear slope expressing the correlations between descriptors x_{ij} with the toxicity values, y_i of the compound i, and e_i is a constant. Typically, they can usually be found through the regression analysis.

3.10 Model Validation

Internal and external validation parameters were used to evaluate the reliability and predictive ability of the models. The validation parameters were compared with the standard for generally acceptable QSAR model as reported in Table 3.3.

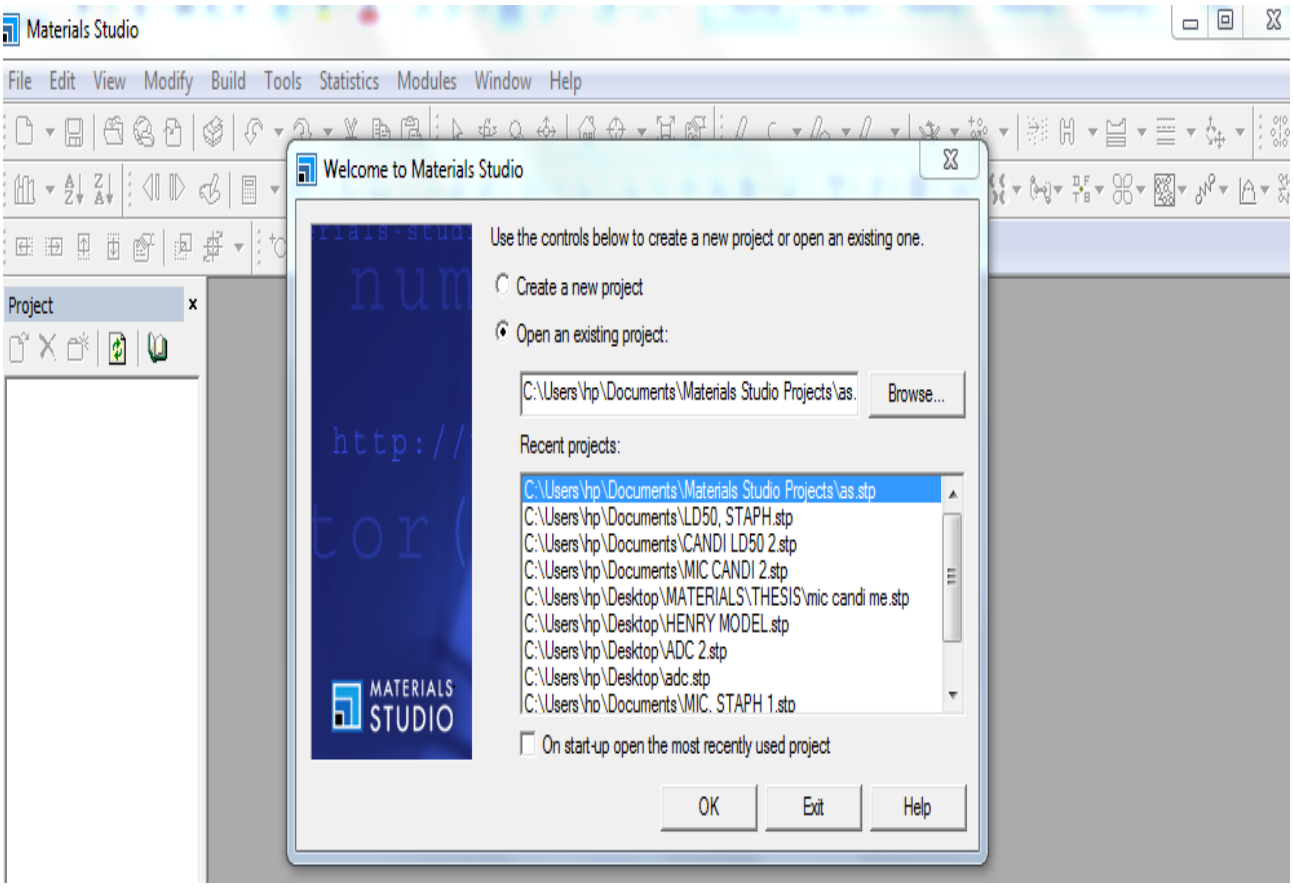


Figure 3.24: Material Studio's work space window and project window

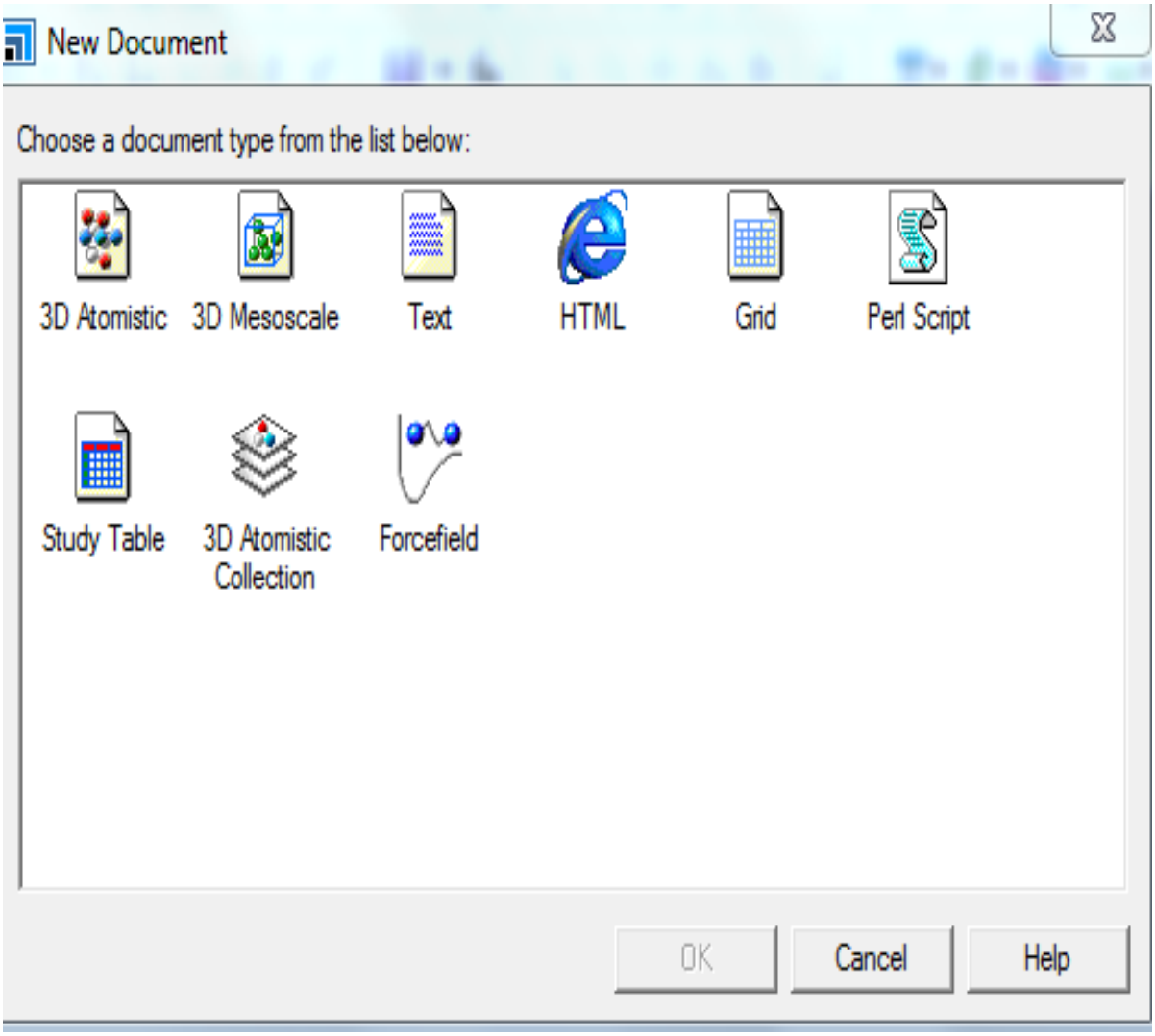


Figure 3.25: Material Studio's *document* window

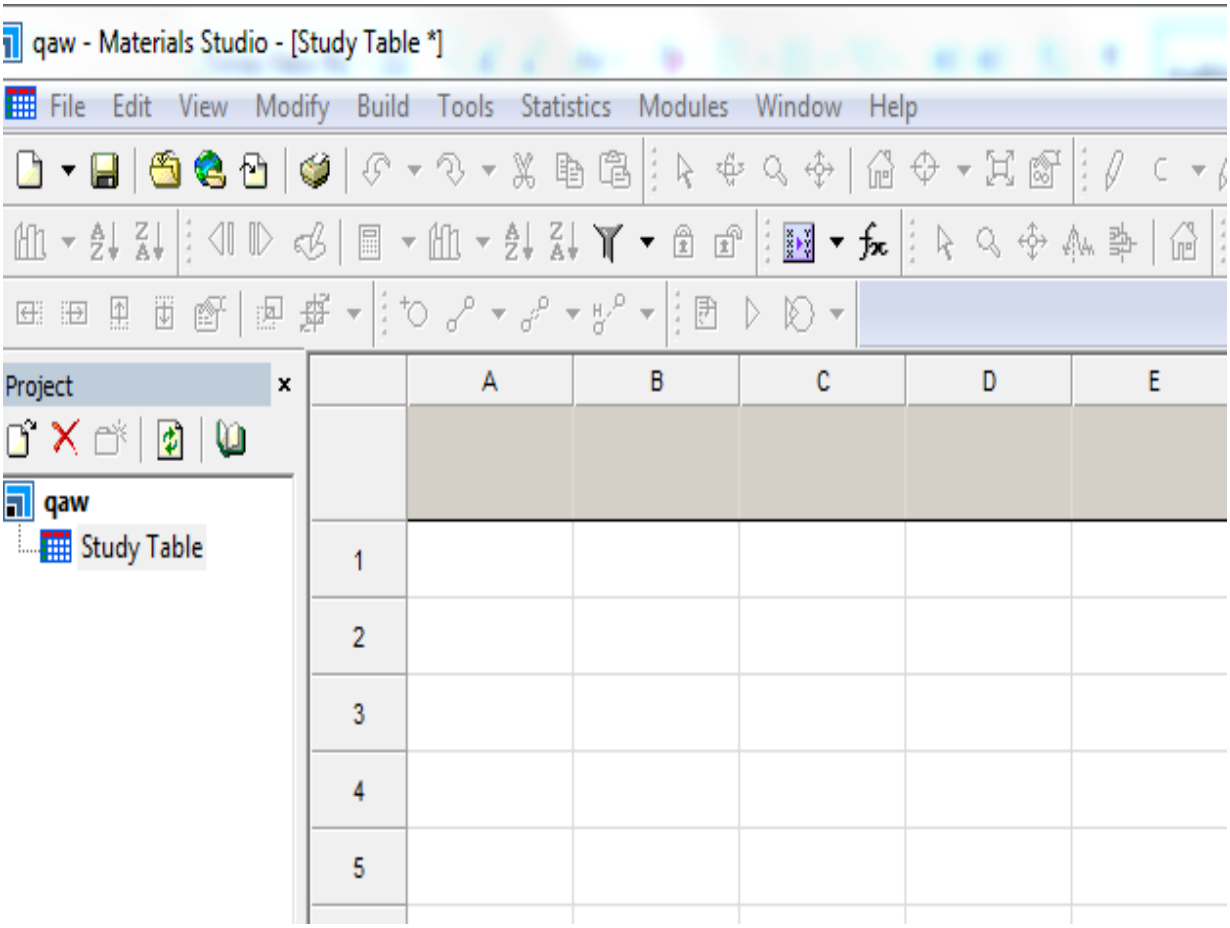
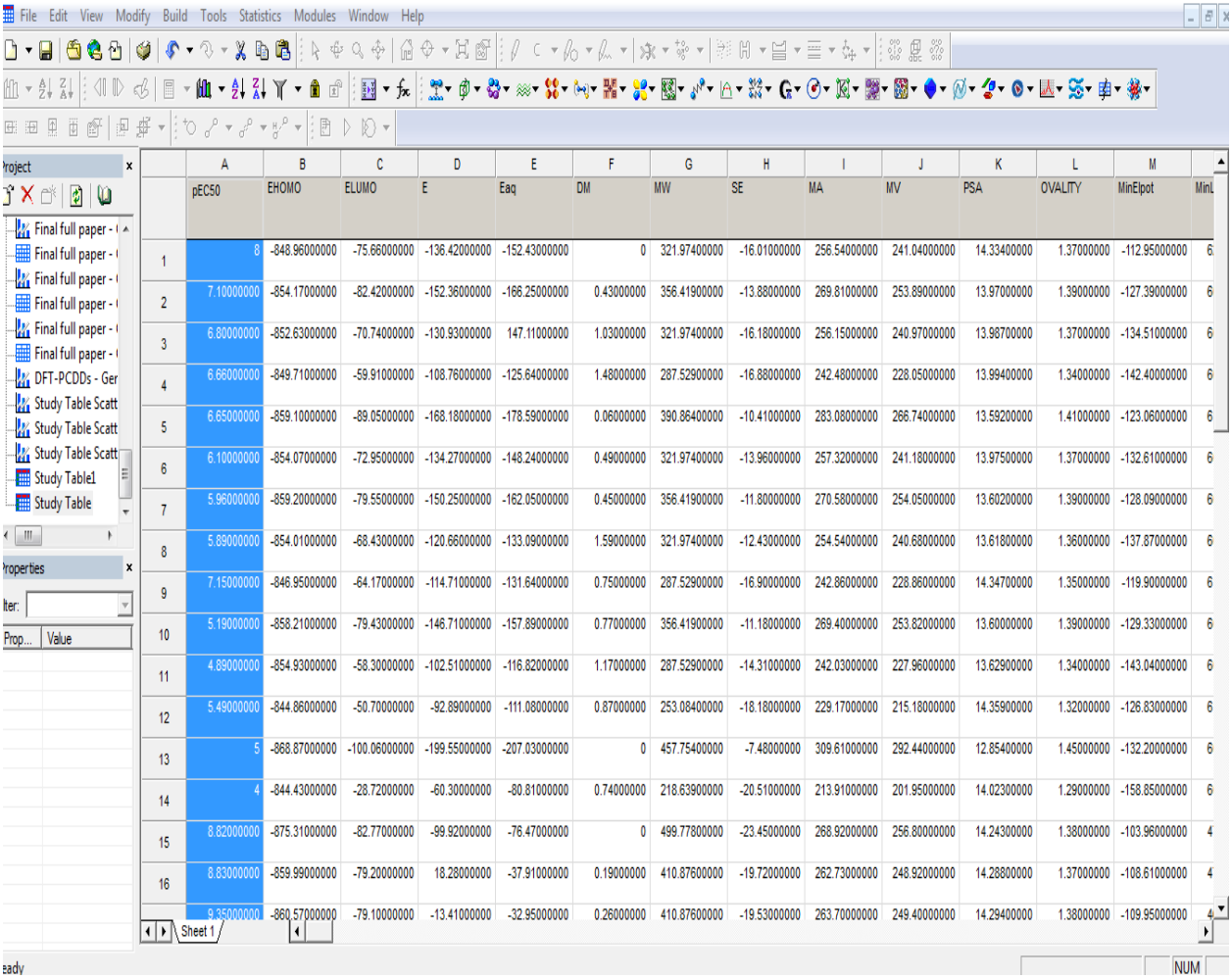


Figure 3.26: Material Studio's *Study table* window



	A	B	C	D	E	F	G	H	I	J	K	L	M
	pEC50	EHOMO	ELUMO	E	Eaq	DM	MW	SE	MA	MV	PSA	OVALITY	MinElpot
1	8	-848.96000000	-75.66000000	-136.42000000	-152.43000000	0	321.97400000	-16.01000000	256.54000000	241.04000000	14.33400000	1.37000000	-112.95000000
2	7.10000000	-854.17000000	-82.42000000	-152.36000000	-166.25000000	0.43000000	356.41900000	-13.88000000	269.81000000	253.89000000	13.97000000	1.39000000	-127.39000000
3	6.80000000	-852.63000000	-70.74000000	-130.93000000	147.11000000	1.03000000	321.97400000	-16.18000000	256.15000000	240.97000000	13.98700000	1.37000000	-134.51000000
4	6.66000000	-849.71000000	-59.91000000	-108.76000000	-125.64000000	1.48000000	287.52900000	-16.88000000	242.48000000	228.05000000	13.99400000	1.34000000	-142.40000000
5	6.65000000	-859.10000000	-89.05000000	-168.18000000	-178.59000000	0.06000000	390.86400000	-10.41000000	283.08000000	266.74000000	13.59200000	1.41000000	-123.06000000
6	6.10000000	-854.07000000	-72.95000000	-134.27000000	-148.24000000	0.49000000	321.97400000	-13.96000000	257.32000000	241.18000000	13.97500000	1.37000000	-132.61000000
7	5.96000000	-859.20000000	-79.55000000	-150.25000000	-162.05000000	0.45000000	356.41900000	-11.80000000	270.58000000	254.05000000	13.60200000	1.39000000	-128.09000000
8	5.89000000	-854.01000000	-68.43000000	-120.66000000	-133.09000000	1.59000000	321.97400000	-12.43000000	254.54000000	240.68000000	13.61800000	1.36000000	-137.87000000
9	7.15000000	-846.95000000	-64.17000000	-114.71000000	-131.64000000	0.75000000	287.52900000	-16.90000000	242.86000000	228.86000000	14.34700000	1.35000000	-119.90000000
10	5.19000000	-858.21000000	-79.43000000	-146.71000000	-157.89000000	0.77000000	356.41900000	-11.18000000	269.40000000	253.82000000	13.60000000	1.39000000	-129.33000000
11	4.89000000	-854.93000000	-58.30000000	-102.51000000	-116.82000000	1.17000000	287.52900000	-14.31000000	242.03000000	227.96000000	13.62900000	1.34000000	-143.04000000
12	5.49000000	-844.86000000	-50.70000000	-92.89000000	-111.08000000	0.87000000	253.08400000	-18.18000000	229.17000000	215.18000000	14.35900000	1.32000000	-126.83000000
13	5	-868.87000000	-100.06000000	-199.55000000	-207.03000000	0	457.75400000	-7.48000000	309.61000000	292.44000000	12.85400000	1.45000000	-132.20000000
14	4	-844.43000000	-28.72000000	-60.30000000	-80.81000000	0.74000000	218.63900000	-20.51000000	213.91000000	201.95000000	14.02300000	1.29000000	-158.85000000
15	8.82000000	-875.31000000	-82.77000000	-99.92000000	-76.47000000	0	499.77800000	-23.45000000	268.92000000	256.80000000	14.24300000	1.38000000	-103.96000000
16	8.83000000	-859.99000000	-79.20000000	18.28000000	-37.91000000	0.19000000	410.87600000	-19.72000000	262.73000000	248.92000000	14.28800000	1.37000000	-108.61000000
	8.35000000	-860.57000000	-79.10000000	-13.41000000	-32.95000000	0.26000000	410.87600000	-19.53000000	263.70000000	249.40000000	14.29400000	1.38000000	-109.95000000

Figure 3.27. Imported descriptors showing the highlighted column of the pEC_{50}

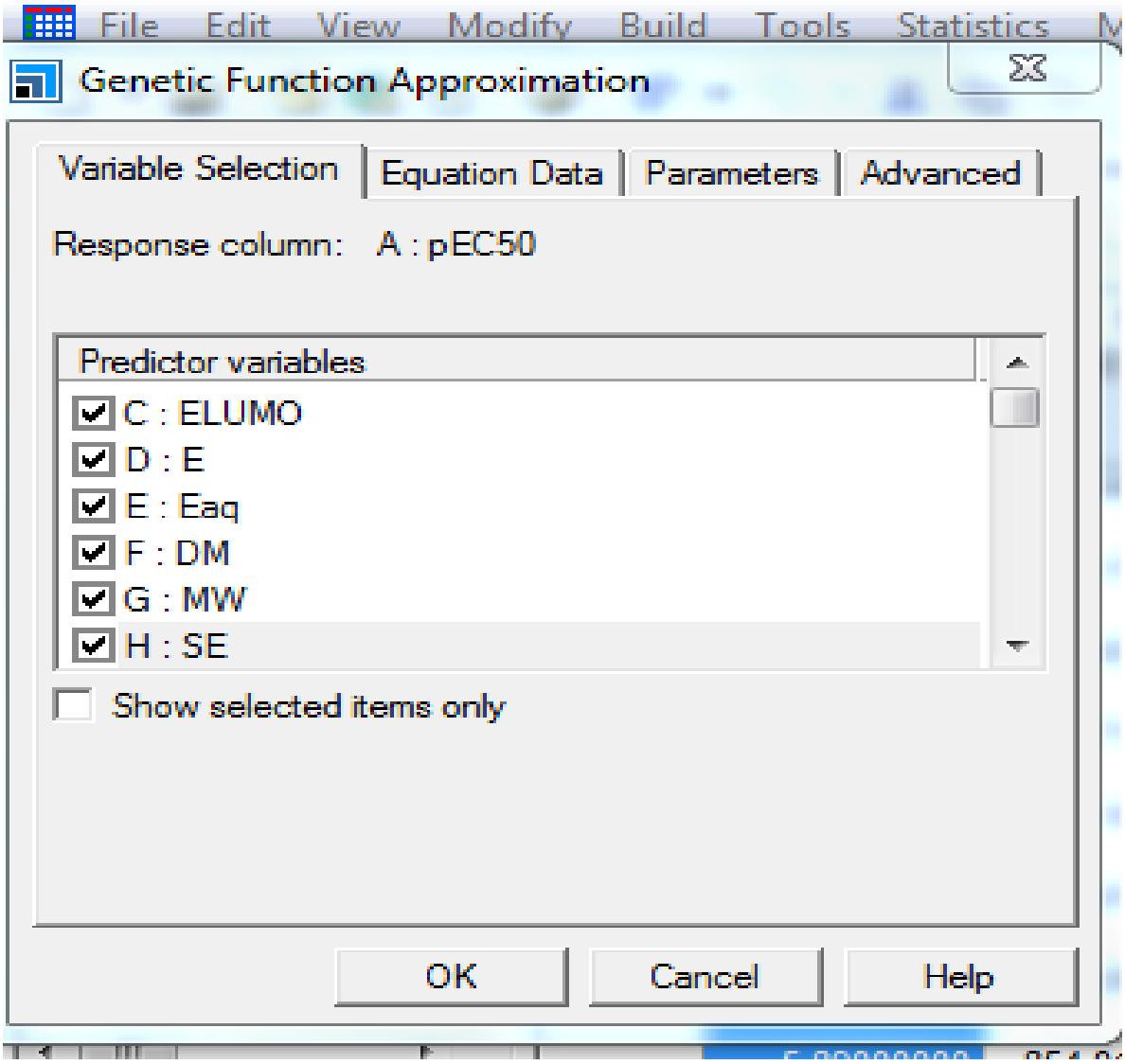


Figure.3.28. Material Studio's Genetic function approximation wind showing the selection of the correlated descriptors.

3.10.1 Internal validation parameters

(a) Friedman LOF: The slightly-modified Friedman lack-of-fit (LOF) for each generated equation is calculated as follows:

$$LOF = \frac{SSE}{M\left[1-\lambda\left(\frac{c+dp}{M}\right)\right]^2} \quad \text{.....3.12}$$

Where, SSE is the sum of squares of errors, c is the number of terms in the model, other than the constant term, d is a scaled smoothing parameter, p is the total number of descriptors contained in all model terms (again ignoring the constant term), M is the number of samples in the training set, λ is a safety factor, with a value of 0.99, to ensure that the denominator of the expression can never become zero, and so the LOF is always well-defined (Kunal et al., 20090.

The scaled smoothing parameter, d, is related to the user-input Scaled LOF smoothnessparameter α by the following expression:

$$d = \alpha\left(\frac{M-C_{max}}{C_{max}}\right) \quad \text{.....3.13}$$

Wherec_{max} is the maximum equation length

(b) Coefficient of multiple determination (R^2)

To assess the goodness-of-fit, the coefficient of multiple determinations is used. R^2 estimates the proportion of the variation in the response that is explained by the predictor.

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{I}(y_i-\bar{y})} \quad \text{... … … …….3.14}$$

Where yi is the observed dependent variable, $\bar{y}$ the mean value of the dependent variable and $\hat{y}$ the calculated dependent variable. If there is no linear relationship between the dependent variable and the descriptors then R^2 = 0.00; if there is a perfect fit then R^2 = 1.00. R^2 values higher than 0.5 indicates that the explained variance by the model is higher than the unexplained one.

(c) Adjusted R^2 (R^2_{adj})

The value of R^2 can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. It follows R^2 should be used with caution. This can be avoided by using another statistical parameter the so-called adjusted R^2 (R^2_{adj}).

$$R^2_{adj} = [(n-1)R^2 - p]/(n-p+1) \ldots\ldots\ldots 3.15$$

R^2_{adj} is interpreted similarly to the R^2 value, except that it takes into consideration the number of degrees of freedom. The value of R^2_{adj} decreases if an added variable to the equation does not reduce the unexplained variable. Where p = number of the independent variables in the model, n is the number of molecule in the data training set. (Brandon and Aline, 2015).

(d)Standard error of estimate (SEE)

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{(I-(K+1))}} \ldots\ldots\ldots\ldots\ldots\ldots 3.16$$

Where, y_i = observed dependent variable, $\hat{y}_i$ = calculated dependent variable, K = number of selected descriptors and I = number of molecules

The smaller the value of SEE is, the higher the reliability of the prediction. However, it is not recommended to have the standard error of estimate smaller than the experimental error of the biological data, because it is an indication of over fitted model.

(e)F-value: High value indicates to show that the regression coefficients are significant. The computer value of p should be significant at p □ 0.005(Roy et al., 2015)

The F-value is determined using equation-6

$$F = \frac{\sum_{i=0}^{I}(y_i-\bar{y})^2/(K-1)}{\sum_{i=1}^{I}(y_i-\hat{y}_i)^2/(I-K)} \ldots\ldots\ldots\ldots 3.17$$

Where I is the number of molecules and K is the number of selected descriptors.y_i = observed dependent variable, $\hat{y}_i$ = calculated dependent variable and $\bar{y}$ = mean value of the calculated dependent variable. The higher the F-value, the greater the probability that the equation is significant (Sofie, 2009).

(f)Cross-validation squared correlation coefficient Q^2 (R^2_{cv})

Cross-validation squared correlation coefficient R^2 (LOO-Q^2) is calculated according to the formula:

$$Q^2 = 1 - \frac{\sum(Y_{pred}-Y)^2}{\sum(Y-\bar{Y})^2} \ldots\ldots\ldots 3.18$$

In Eq. (7), Y_{pred} and Y indicate predicted and observed activity values, respectively, and $\bar{Y}$ indicate mean of the calculated activity value. A model is considered acceptable when the value of Q^2 exceeds 0.5. (Wold *et al.*, 1995).

In the case of this research, external validation techniques (LMO-Leave Many Out) was applied in which the 23 compounds of the test set were used for the external validation and the predicted R^2 for the validation was calculated using equation-7.

3.10.2 External validation parameters

(a) Predicted R^2 (R^2_{pred})

The predictive R^2 was calculated based on only molecules not included in the training set (test set). Models are generated based on training set compounds and predictive capacitywas

was judged based on the predictive R^2 (R^2_{pred}) value which was calculated using eqn-5.

$$R^2_{pred} = 1 - \frac{\sum(Y_{pred(test)} - Y_{test})^2}{\sum(Y_{(test)} - \bar{Y}_{training})^2} \quad \ldots\ldots\ldots\ldots 3.19.$$

In Eq. (8), $Y_{pred(test)}$ and Y_{test} indicate predicted and observed toxicity values respectively of the test set compounds and $\bar{Y}_{training}$ indicates mean toxicity of the training set. For a QSAR model, the value of R^2_{pred} should be more than 0.5. All the statistical parameters calculated, agree with the criteria reported in Table-2

(b) Golbraikh and Tropsha's criteria**:** according to (Golbraikh and Tropsha, 2002)models are considered satisfactory, if all the following conditions are met:

(i) R^2_{test} □ 0.5

(ii) $r^2 - r_0^2 / r^2$ □ 0.1

(ii) $r^2 - r'^2_0 / r^2$ □ 0.1

(iv) $0.85 \leq k \leq 1.15$

(v) $0.85 \leq k^I \leq 1.15$

Parameters r^2 and r_0^2 are the squared correlation coefficients between the observed and predicted toxicity values of the compounds with and without intercept, respectively. The parameter r'^2_0 bears the same meaning with r_0^2 but uses the reversed axes. K is the intercept of the plot of the observed and predicted values of the compounds via the origin and K^I the reversed axes intercept (Golbraikh and Tropsha, 2002).

3.11 Procedure for determining Golbraikh and Tropsha parameters

The Microsoft excel package was launched creating its interface as presented in Figure. 3.26. On the work sheet, the data for predicted and observed activities were entered into two adjacent columns. Through the **Insert → Scatter** menus on the interface, a reduced version of the graph appears in a prop up window (Figure 3.27). By clicking +symbol→ **Chart elements**at the top right of the window, *Trend line, axis title, Chart title* were added to the graph by marking the appropriate boxes. Through **Trend line** box → **more options→ Set intercept box**, the graph was either made to have interceptby leaving the intercept box unchecked or otherwise by setting the intercept to 0.0 (origin). The Parameters r^2 and r_0^2 are the squared correlation coefficients between the observed and predicted values of the compounds with and without intercept, respectively. The parameter r'^2_0 bears the same meaning with r_0^2 but uses the reversed axes.k is the slope of the plot of the observed and predicted values of the compounds via the origin and k' the slope of the reversed axes (Golbraikh and Tropsha, 2002).

3.12 Multicollinearity among Descriptors:Variance Inflation Factor (VIF)

The multi-collinearity between the above five descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = \frac{1}{1 - R^2} \ldots\ldots\ldots\ldots\ldots \quad 3.20$$

Where R^2 is the correlation coefficient of the multiple regression between the variables within the models.R^2 is the fraction of the total variance of the Y variable that is explained by the genetic function approximation equation. R^2 is calculated as:

$$\frac{SSR}{SST} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.21$$

where SSR is the sum of squares of regression and SST is the total sum of squares. The closer the value is to 1.0, the better the genetic function approximation equation explains the Y variable If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary (Shapiro *et al.,* 1998; Jaishwal *et al.,* 2004).

3.13Euclidean Based Applicability Domain (AD)

Applicability domain (AD) is the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain. Euclidean based applicability domain helps to ensure that the compounds of the test set are representative of the training set compounds used in model development. It is based on distance scores calculated by the Euclidean distance norms. At first, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1(0 = least

diverse, 1 = most diverse training set compound). Then normalized mean distance score for test set are calculated and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain. This can also be checked by plotting a 'Scatter plot' (normalized mean distance vs. respective activity) including both training and test set. If the test set compounds are inside the domain/area covered by training set compounds that means these compounds are inside the applicability domain otherwise not (Jaworska *et al.,* 2005, Tropsha *et al.,* 2003).

This distance-based method calculate a distance from the test set compounds to the training set compounds. The Euclidian distance is expressed mathematically using equation below:

$$EDij = [\sum w_k (x_k^i - x_k^j)^2]^{1/2} \quad \text{.....3.22}$$

The difference, $x_k^i - x_k^j$ represents the distance from the test set compounds to the training set compounds and w_k is a weighted vector corresponding to the importance of the k^{th} descriptor in the modelcalculated using auto-scaled descriptors, $x_k^i$and x_k^j represent compound from the test set and training set respectively (Netzeva, 2005). The weighting makes it possible to account for relative contribution of each variable to the similarity and improves detection of the AD of the model.

3.14 Procedure for Computing Euclidean Distance using DTC-Euclidean Program

The interface excel was created by launching it. Columns containing descriptors in the model and their values for each compound present in training and test set were arranged on the worksheet saved in separate folders as *csv* file extension. It was ensured thatthe order of descriptors in training set input file was identical to that of test set input file. The DTC-Euclidean program was launched and its folder was opened. The previously saved input files for test and training set compounds were pasted in its folder as shown in Figure. 3.29. By clicking on

Euclidean 1.0 in the folder, a training/test set dialog box shown in appeared. The name of the training and test set input files as saved in the program's folder were entered in the check box provided and *Ok* option was clicked in succession as presented in Figure 3.27. The result of the calculation was provided as *Train_Test_Euclidean*file in the DTC-Euclidean program's folder on Microsoft excel sheet.

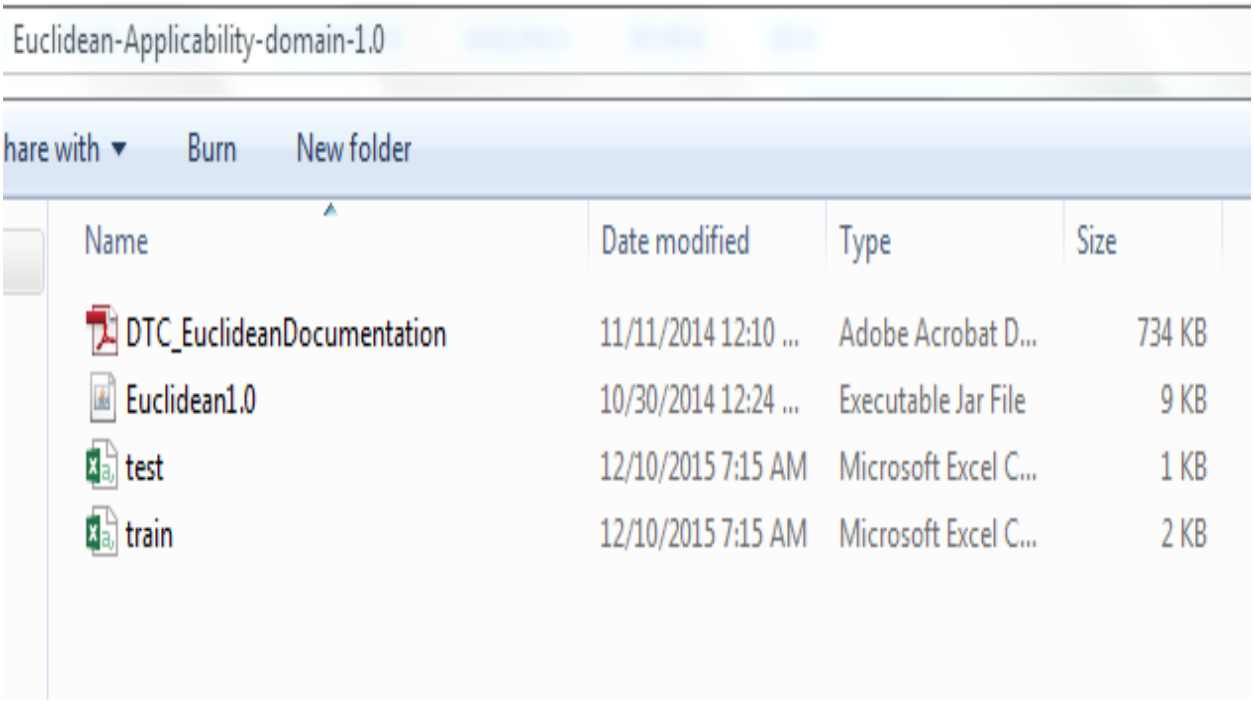


Figure 3.29: The folder of Euclidean-Applicability domain

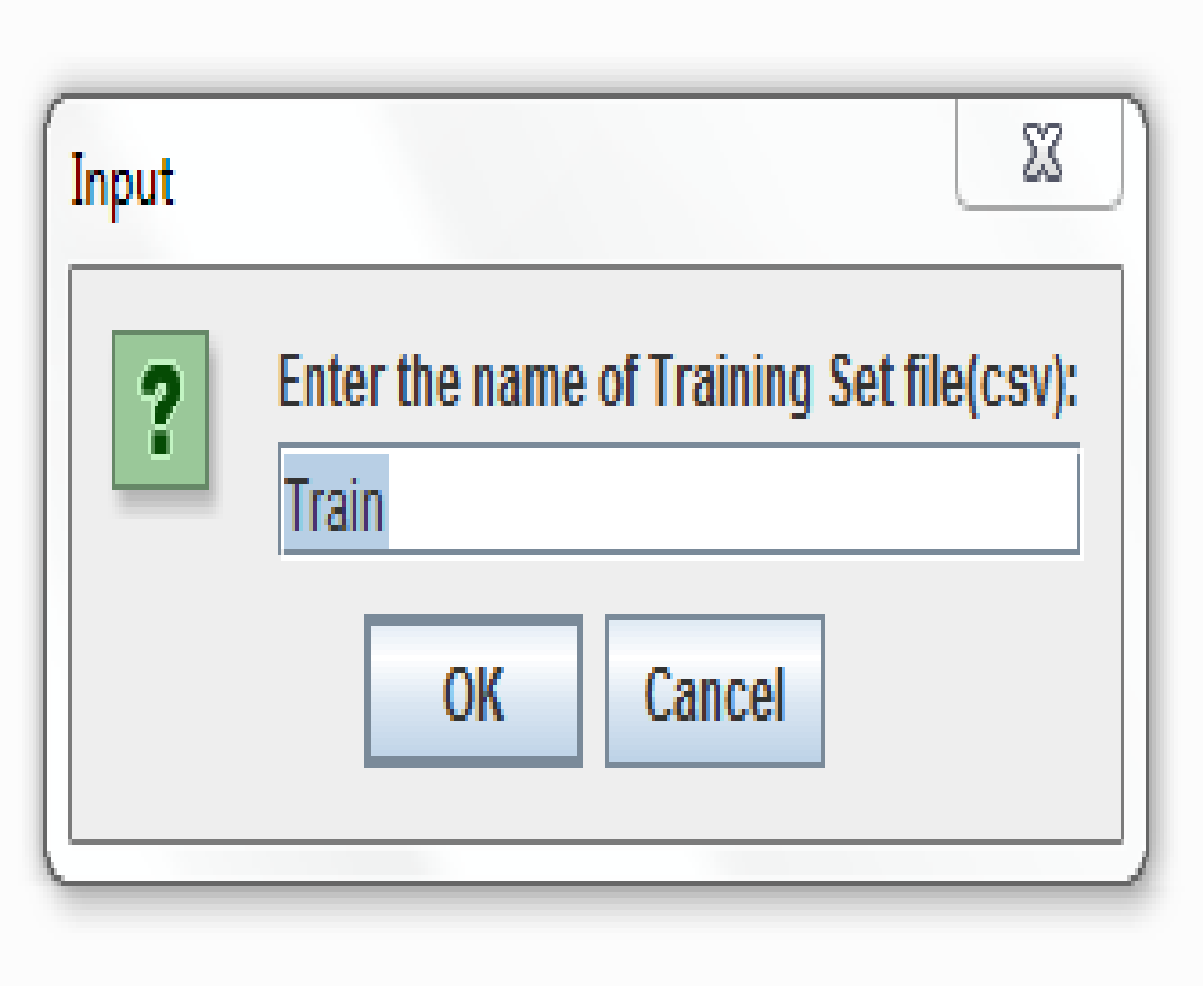


Figure 3.30: Euclidean-Applicability domain training / test set dialog box

Table 3.3: Criteria for selection of acceptable model

S/N	Criteria for selection of model
1	N = number of molecules (> 20 molecules)
2	K= number of descriptors in a model (statistically N/5 descriptor in a model)
3	df = degree of freedom (N-K-1) (higher is better).
4	R^2 =coefficient of determination (> 0.7)
5	R^2cv = cross-validation square correlation (> 0.5)
6	R^2_{adj} = adjusted squared correlation coefficient (> 0.5)
7	R^2_{pred} = predicted coefficient of determination (> 0.5)
8	SEE = standard error of estimate (smaller is better)
9	F-test = F-test for statistical significance of the model (higher is better, for some set of descriptors and compounds)
10	Variance inflation factor (VIF): $1 \leq VIF \leq 10$
11	Intercept: K and K', $0.85 \leq k$ or $k' \leq 1.15$
12	(a) Golbraikh and Tropsha condition: $r^2 - r_0^2 / r^2$ □ 0.1 (b) Golbraikh and Tropsha condition:$r^2 - r'^2_0 / r^2$ □ 0.1

(Pathak*et al.*, 2011; Golbraikh and Tropsha, 2002;Jaiswal *et al.*, 2004)

CHAPTER FOUR

4.0 Results

4.1 Results ofQSAR Modeling of pEC50 forPCDDs, PCDFs and PCBs all together.

QSAR modeling of pEC50 for some of the three classes of polychlorinated aromatic compounds(PCDDs, PDFs and PCBs)was performed and the QSAR results are presented as follows:

4.1.1 GFAgenerated models of pEC50 for PCDDs, PCDDFs and PCBs all together (DFT)

Five (5) QSAR generated GFA models for pEC50 of three classes of some polychlorinated aromatic compounds (PCDDs, PCDFs, and PCBs) are presented as equations (4.1-4.5). Out of the 5 models, model-1 (which is presented as equation 4.1) was selected as the best for predicting the toxicities of polychlorinated aromatic compounds based on the fact that it has the best statistical parameters. The validation parameters of the models agree with the standard reported in Table 3.3.

4.1.2 Statistical/Validation parameters for the generated models

Statistical parameters for internal validation of all the 5 models were calculated andthe results are presented in Table-4.1. There is a good agreement of the validation parameters with the standard reported inTable 3.3.

Table 4.1 Statistical/validation parameters for the generated models.

Statistical parameters	Model 1	Model 2	Model 3	Model 4	Model 5	Standard
Friedman LOF	0.437664	0.439352	0.440601	0.443745	0.443884	Smaller(~zero)
R-squared	0.967342	0.967217	0.967123	0.966889	0.966878	≥0.5
Adjusted R-squared	0.959178	0.959021	0.958904	0.958611	0.958598	≥0.5
Cross validated R-squared	0.940154	0.923084	0.937807	0.933100	0.937382	≥0.5
Significant Regression	Yes	Yes	Yes	Yes	Yes	Yes
Significance-of-regression F-value	118.48328	118.012638	117.666633	116.80482	116.7668	Thelarger value is best
Critical SOR F-value (95%)	2.080618	2.080618	2.080618	2.080618	2.080618	Thelarger value is best
Replicate points	0	0	0	0	0	
Computed experimental error	0.000000	0.000000	0.000000	0.000000	0.000000	~ zero
Lack-of-fit points	40	40	40	40	40	constant
Min expt. error for non-significant LOF (95%)	0.257926	0.258423	0.258790	0.259712	0.259752	~ zero

4.1.3 Comparison of the observed and predicted pEC50

The comparison of the predicted toxicities (pEC50) of the models with the experimental values for the training and test sets are presented in Tables 4.2 and 4.3, respectively. The low residual values shown in the Tables confirm the high predictive power of the models.

4.2.1 Five (5) generated models by GFA based on DFT optimization

Model-1

Y = 0.317123855 * X26 + 2.67319429 - 1.36110203 - 1.453070371 * X165 - 96.631409262 *X167 + 69.872692578 * X170 - 7.877708429 * X228 - 21.439713345 * X277 + 124.942300592 * X281 + 1.835776438 * X293 - 631.436918720……..(4.1)

Descriptor

X26 : Polarizability, X94 : SP-7, X144 : maxHBint5, X165 : ETA_dAlpha_B, X167 : ETA_Epsilon_2, X170 : ETA_Epsilon_5, X228 : n5Ring, X277 : GRAV-2, X281 : GRAVH-3, X293 : MOMI-R

Model-2

Y = 0.337543195 * X26 + 1.986019421 * X94 - 1.841849479 * X144 - 2.591615577 * X165 - 105.290996384 * X167 + 150.560145568 * X175 - 8.806840418 * X228 - 22.985643463 * X277 + 135.442609778 * X281+ 0.491497214 * X345 - 657.102081723…….(4.2)

Descriptors

X26 : Polarizability, X94 : SP-7, X144 : maxHBint5, X165 : ETA_dAlpha_B, X167 :ETA_Epsilon_2, X175 : ETA_Psi_1, X228 : HU : n5Ring, X277 : GRAV-2, X281 : GRAVH-3, X345 : WA.eneg

Model-3

Y = 0.312654553 * X26 + 2.629675114 * X94 - 1.401918491 * X144 - 1.502764637 * X165 - 116.167215005 * X167 + 66.429804381 * X175 - 7.847220091 * X228 - 21.154896743 * X277 + 125.332874970 * X281 + 1.696580396 * X293 - 602.797141503……(4.3)

Descriptors

X26 : Polarizability, X94 : SP-7, X144 : maxHBint5, X165 : ETA_dAlpha_B, X167 : ETA_Epsilon_2, X175 : ETA_Psi_1, X228 : n5Ring, X277 : GRAV-2, X281 : GRAVH-3, X293 : MOMI-R

Model-4

Y = 0.336800494 * X26 + 2.942208255 * X94 - 0.640140511 * X108 - 1.579876967 * X144 - 1.116608578 * X165 + 81.907227030 * X175 - 8.260828699 * X228 - 14.048865479 * X277 + 82.359135734 * X281 + 2.468795662 * X293 - 413.254169614……. (4.4)

Descriptors

X26 : Polarizability, X94 : SP-7, X108 : nHBa, X144 : maxHBint5, X165 : ETA_dAlpha_B, X175 : ETA_Psi_1, X228 : HU : n5Ring, X277 : GRAV-2, X281 : GRAVH-3, X293 : MOMI-R

Model-5

Y = 0.334445535 * X26 - 1.292560192 * X35 + 2.861556067 * X94 - 1.304920784 * X144 - 1.234903575 * X165 - 136.309059489 * X177 - 8.316100485 * X228 - 17.174253798 * X277 + 96.839460363 * X281 + 2.237066146 * X293 - 454.42064568 …… (4.5)

4.1.4 Univariate analysis of the toxicity (pEC50) data

The most important parameters here are skewness and kurtosis. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. Kurtosis which is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution, presented as univariate analysis data in Table-4.4.

Table-4.2. Chemical names along with the observed and calculated toxicity values (pEC_{50}) of the training set compounds**.**

S/N	Chemical name	Actual values	Predicted values	Residual values
1	1,2,3,7,8 pentachlorodibenzo-p-dioxin	7.10000000	6.89480200	0.205198
2	2,3,6,7 Tetrachlorodibenzo-p-dioxin	6.80000000	6.57743100	0.222569
3	2,3,6 Trichlorodibenzo-p-dioxin	6.66000000	6.76089900	-0.100899
4	1,2,3,4,7,8Hexachlorodibenzo-p-dioxin	6.55000000	6.44193700	0.108063
5	1,3,7,8 Tetrachlorodibenzo-p-dioxin	6.10000000	6.11652600	-0.016526
6	1,2,4,7,8 pentachlorodibenzo—dioxin	5.96000000	6.19664600	-0.236646
7	1,2,3,4 Tetrachlorodibenzo-p-dioxin	5.89000000	6.28801600	-0.398016
8	2,3,7 Trichlorodibenzo-p-dioxin	7.15000000	6.86770100	0.282299
9	1,2,4 Trichlorodibenzo-p-dioxin	4.89000000	4.88714900	0.002851
10	2,8-dichlorodibenzo-p-dioxin	5.49000000	5.15069900	0.339301
11	1,2,3,4,6,7,8,9-Octachlorodibenzo-o-dioxin	5.00000000	5.14484700	-0.144847
12	1-chlorodibenzo-p-dioxin	4.00000000	4.24100500	-0.241005
13	2,3,7,8 Tetrabromodibenzo-p-dioxin	8.82000000	9.40408400	-0.584084
14	2,3-Dibromo 7,8-chlorodibenzo-p-dioxin	8.83000000	8.99649900	-0.166499
15	2,8- Dibromo -3,7-dichlorodibenzo-p-dioxin	9.35000000	9.06707700	0.282923
16	2-Bromo-3,7,8-trichlorodibenzo-p-dioxin	7.94000000	8.30822700	-0.368227
17	1,3,7,9-Tetrabromodibenzo-p-dioxin	7.03000000	7.26343800	-0.233438

18	1,3,7,8-Tetrabromodibenzo-p-dioxin	8.70000000	8.47020000	0.229800
19	1,2,4,7,8-Pentabromodibenzo-p-dioxin	7.77000000	7.42604000	0.343960
20	1,2,3,7,8-Pentabromodibenzo-p-dioxin	8.18000000	8.09153100	0.088469
21	2,7-Dibromodibenzo-p-dioxin	7.81000000	7.82283900	-0.012839
22	2-Bromodibenzo-p-dioxin	6.53000000	6.15431200	0.375688
23	1-chlorodibenzofuran	4.53000000	4.36064600	0.169354
24	2-chlorodibenzofuran	3.55000000	4.22833800	-0.678338
25	3-chlorodibenzofuran	4.38000000	4.10933400	0.270666
26	4-chlorodibenzofuran	3.00000000	3.01669200	-0.016692
27	2,3-Dichlorodibenzofuran	5.36000000	5.63232200	-0.272322
28	1,3,6-Trichlorodibenzofuran	5.36000000	5.30147400	0.058526
29	1,3,8-Trichlorodibenzofuran	4.07000000	4.56537300	-0.495373
30	2,3,7-Trichlorodibenzofuran	7.10000000	6.96629000	0.133710
31	2,3,4,8-Tetrachlorodibenzofuran	6.70000000	6.40437800	0.295622
32	2,3,7,8-Tetrachlorodibenzofuran	7.39000000	7.77433500	-0.384335
33	1,2,3,7,8-Pentachlorodibenzofuran	7.13000000	6.97671300	0.153287
34	1,2,4,7,8-Pentachlorodibenzofuran	5.89000000	5.69544900	0.194551
35	2,3,4,7,8-Pentachlorodibenzofuran	7.82000000	7.82693800	-0.006938
36	1,2,3,4,7,8-Hexachlorobenzofuran	6.64000000	6.83519100	-0.195191
37	1,2,3,6,7,8-Hexachlorobenzofuran	6.57000000	6.76563400	-0.195634
38	2,3,4,6,7,8-Hexachlorobenzofuran	7.33000000	7.51370700	-0.183707
39	2,3,6,8-Tetrachlorodibenzofuran	6.66000000	6.24406500	0.415935
40	1,2,3,7-Tetrachlorodibenzofuran	6.96000000	6.86559800	0.094402

41	2,3,4,7,9-Pentachlorodibenzofuran	6.70000000	6.37662100	0.323379
42	1,2,3,7,9-Pentachlorodibenzofuran	6.40000000	6.39655100	0.003449
43	Dibenzofuran	3.00000000	2.89483700	0.105163
44	2,3,4,7-Tetrachlorobiphenyl	7.60000000	7.48510400	0.114896
45	1,2,4,6,8-Pentachlorobiphenyl	5.51000000	5.31940200	0.190598
46	2,3,4,4’-Tetrachlorobiphenyl	4.94000000	4.94644900	-0.006449
47	3,3’,4,4’-Tetrachlorobiphenyl	6.15000000	6.06549900	0.084501
48	3,3’4,4’5-Pentachlorobiphenyl	6.92000000	6.42820400	0.491796
49	2,3,3’,4,4’,5-Hexachlorobiphenyl	5.15000000	5.63058300	-0.480583
50	2,3,3’,4,4’,5’-Hexachlorobiphenyl	5.30000000	5.42050100	-0.120501
51	2,3,4,5-Tetrachlorobiphenyl	3.85000000	3.89186900	-0.041869

Table-4.3: Chemical names along with the observed and the calculated toxicity (pEC_{50}) of the test set compounds.

S/N	Compound names	Actual values	Predicted values	Residual values
1	2,3,7,8-Tetrachlorodibenzo-pdioxin	8.00	7.137742	0.862258
2	1,2,3,4,7pentachlorodibenzo-p-dioxin	5.19	6.416795	-1.2468
3	2,3,7-Tribromodibenzo-p-dioxin	8.93	9.433407	-0.50341
4	2,6-dichlorodibenzofuran	3.61	5.33345	-1.72345
5	2,8-Dichlorodibenzofuran	5.05	5.814313	-0.76431
6	2,3,4-Trichlorodibenzofuran	4.72	6.341931	-1.62193
7	2,3,8-Trichlorodibenzofuran	6.00	6.078967	-0.07897
8	2,6,7 Trichlorodibenzofuran	6.35	5.891172	0.458828
9	2,3,4,6 Tetrachlorodibenzofuran	6.46	6.643481	-0.18348
10	1,2,4,8-Tetrachlorodibenzofuran	5.00	5.764051	-0.76405
11	1,2,4,7,9-Pentachlorodibenzofuran	4.7	4.070704	0.629296
12	1,2,4,6,7,9Hexachlorodibenzofaran	5.08	3.868581	1.211419
13	1,2,3,6-Tetrachlorodibenzofuran	6.46	6.216	0.244
14	1,3,4,7,8-Pentachlorodibenzofuran	6.70	6.636922	0.063078
15	3,4,4’,5-Tetrachlorobiphenyl	4.55	5.973791	-1.42379
16	2’,3,4,4’,5-Pentachlorobiphenyl	4.85	4.9348	-0.0848
17	2,3,3’,4,4’-Pentachlorobiphenyl	5.37	5.668328	-0.29833
18	2,3’,4,4’,5-Pentachlorobiphenyl	5.04	5.392424	-0.35242

19	2,3,4,4’,5-Pentachlorobiphenyl	5.39	5.43461	-0.04461
20	2,2’4,4’,5,5’-Hexachlorobiphenyl	4.26	4.397758	-0.13776
21	2,3’,4,4’,5,5’-Hexachlorobiphenyl	4.80	4.397758	0.402242
22	2,2’,4,4’-Tetrachlorobiphenyl	3.89	4.710143	-0.82014
23	2,3’,4,4’,5’,6-Hexachlorobiphenyl	4.00	3.366028	0.633972

Table-4.4: Univariate analysis of the toxicity data

Statistical parameter	Values
Number of sample points	51
Range	6.35000000
Maximum	9.35000000
Minimum	3
Mean	6.28450980
Median	6.57000000
Variance	2.23024000
Standard deviation	1.50826000
Mean absolute deviation	1.21134000
Skewness	-0.26949300
Kurtosis	-0.48857800

4.1.5 External validation parameters

External validation of the best models were performed by using the models to predict the toxicities (pEC50) of the respective test set compounds and the results which are presented in Table 4.3, revealed a good agreement between the predicted and observed toxicities (pEC50)of the test set molecules.

4.1.5.2 Euclidean base applicability domain for the best model (model-1)

The Euclidean based applicability domain for the training and test set compounds of best model,(models-1) are presented in Tables 4.5 and 4.6, respectively. The normalized mean distance score for all the molecules fall within the range of 0-1 indicating that the models could very well predict the pEC50 (toxicity) of all molecules that falls within their domain.

4.1.5.3 Golbraikh and Tropsha's parameters for the selected model (model-1)

Table 4.7 presents the Golbraikh and Tropsh's parameters for model-1. The results show that the QSTR models meet the Golbraikh and Tropsha criteria for rebust and acceptable QSTR model shown in Table 3.5. Therefore, the robustness and stability of the models are ensured.

4.1.6 Linear relationship of observed and the predicted pEC_{50} values for all the series

The linear relationship between the observed and the predicted (pEC_{50}) for the training and test set for all the three series (PCDDs, PCDFs and PCBs) are presented in Figure. 4.1 and 4.2 respectively.

4.1.7 Plot of residual versus actual toxicity (pEC_{50}) values

The plot of the residual versus actual (pEC_{50}) values for all the three series (PCDDs, PCDFs and PCBs) is the measure of the dispersion of residual EC50values from the actual values is presented in Figure. 4.3.The propagation of the errors on both sides of zero is an indication of the robustness of the QSTR models.

Table-4.5: Euclidean based applicability domain for Model 1 (Training set)

Compound No.	Distance Score	Mean Distance	Normalized Mean Distance
1	1723.088	33.786	0.633
2	1685.574	33.05	0.598
3	1592.952	31.234	0.512
4	1698.473	33.303	0.61
5	1517.455	29.754	0.441
6	1536.902	30.135	0.459
7	1437.604	28.188	0.367
8	1420.358	27.85	0.351
9	1343.221	26.338	0.279
10	1344.468	26.362	0.28
11	1756.132	34.434	0.664
12	1362.401	26.714	0.297
13	1287.707	25.249	0.227
14	1210.207	23.73	0.154
15	1192.205	23.377	0.138
16	1168.586	22.913	0.116
17	1189.654	23.327	0.135
18	1178.727	23.112	0.125
19	1208.791	23.702	0.153
20	1198.849	23.507	0.144
21	1125.535	22.069	0.075
22	1155.417	22.655	0.103
23	1217.06	23.864	0.161
24	1221.275	23.947	0.165
25	1212.898	23.782	0.157
26	1304.822	25.585	0.243
27	1104.281	21.653	0.056

28	1044.77	20.486	0
29	1055.238	20.691	0.01
30	1062.977	20.843	0.017
31	1072.224	21.024	0.026
32	1083.815	21.251	0.036
33	1140.52	22.363	0.089
34	1156.524	22.677	0.104
35	1171.11	22.963	0.118
36	1261.191	24.729	0.202
37	1278.54	25.069	0.218
38	1297.566	25.442	0.236
39	1260.49	24.715	0.201
40	1311.514	25.716	0.249
41	1382.292	27.104	0.315
42	1412.457	27.695	0.343
43	1749.89	34.312	0.658
44	1467.629	28.777	0.395
45	1517.53	29.755	0.441
46	1555.786	30.506	0.477
47	1617.165	31.709	0.534
48	1820.715	35.7	0.724
49	1915.283	37.555	0.813
50	2005.085	39.315	0.897
51	2115.84	41.487	1

Table 4.6: Euclidean based applicability domain for Model 1 (Test set)

Compound No.	Distance Score	Mean Distance	Normalized Mean Distance
1	1729.968	33.921	0.64
2	1452.674	28.484	0.381
3	1130.749	22.172	0.08
4	1111.845	21.801	0.063
5	1108.315	21.732	0.059
6	1056.214	20.71	0.011
7	1069.624	20.973	0.023
8	1077.415	21.126	0.03
9	1065.35	20.889	0.019
10	1100.139	21.571	0.052
11	1141.902	22.39	0.091
12	1337.368	26.223	0.273
13	1284.89	25.194	0.224
14	1354.405	26.557	0.289
15	1658.197	32.514	0.573
16	1659.253	32.534	0.574
17	1696.701	33.269	0.609
18	1738.627	34.091	0.648
19	1777.992	34.863	0.685
20	1876.778	36.8	0.777
21	1962.435	38.479	0.857
22	2056.454	40.323	0.945
23	2154.391	42.243	1.036

Table 4.7: Golbraikh and Tropsha external validation parameters for model-1

S/n	Parameter	Value	Standard
1	r^2	0.657	≥ 0.5
2	$r'_0{}^2$	0.637	≥ 0.5
3	$r_0{}^2$	0.604	≥ 0.5
4	K	1.034	0.85-1.15
5	K'	0.951	0.85-1.15
6	$r^2 - r_0{}^2 / r^2$	0.081	< 0.1
7	$r^2 - r'_0{}^2 / r^2$	0.030	< 0.1

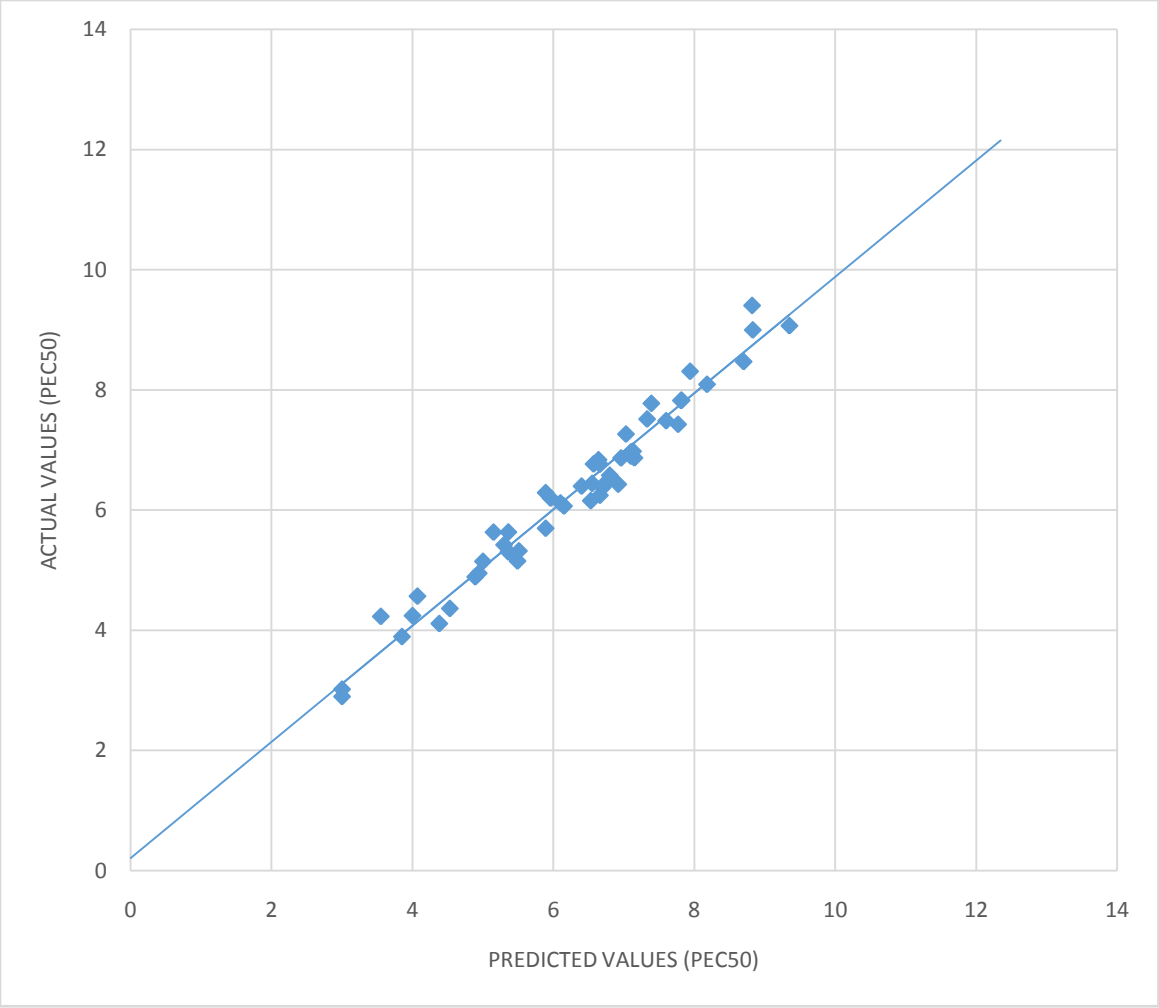


Figure 4.1: Linear relationship of observed and predicted toxicity of data of the training set.

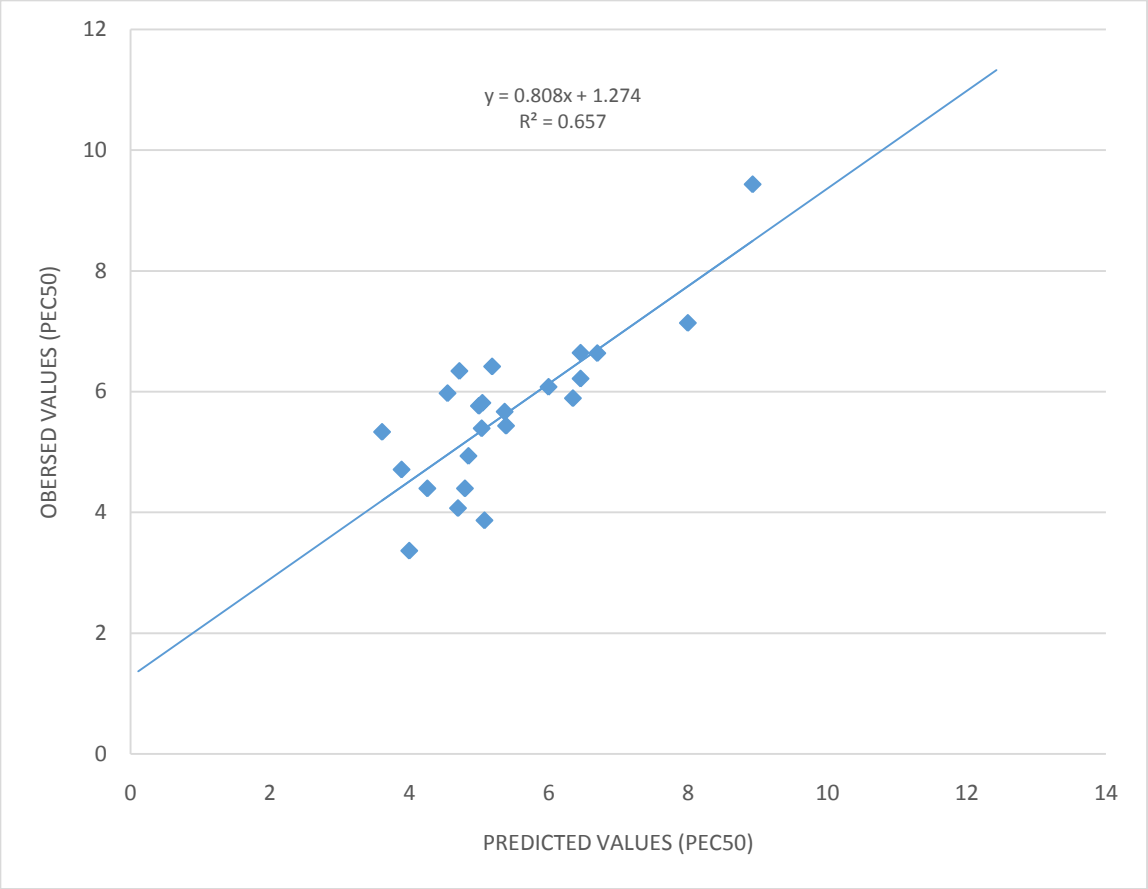


Figure 4.2: Linear relationship of observed and predicted toxicity of data of the test set.

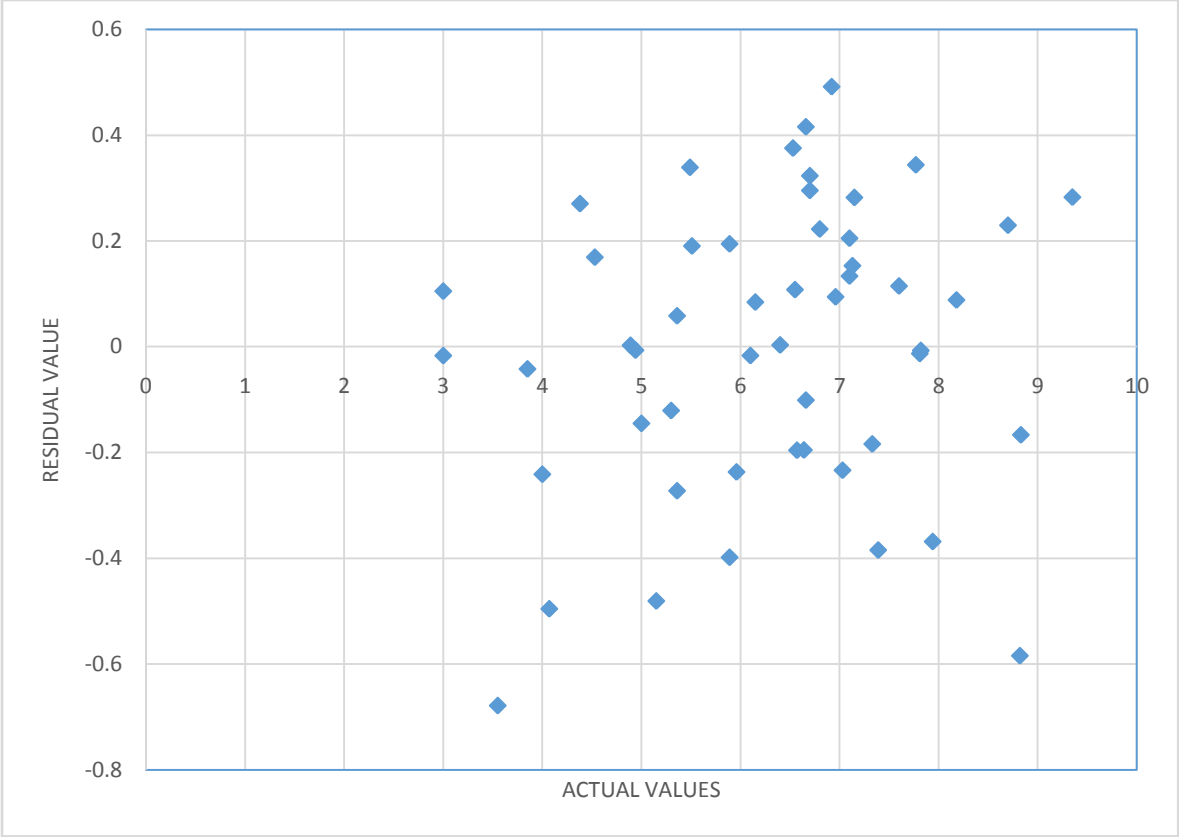


Figure.4.3 A plot of residual values against actual (pEC_{50}) value

4.2 Results of QSTR Modelingof pEC50 for (PCDDs)

QSTR modeling of pEC50 for some polychlorinated aromatic compounds (PCDDs)was performed and the results obtained using semi-empirical and DFT approaches were compared. The QSTR results are presented as follows:

4.2.1 GFA Generated models of pEC50 for PCDDs Using DFT and SE approaches

The five (5) QSTR generated GFA models of pEC50 of some polychlorinated aromatic compounds (PCDDs,) for each of the Density Functional TheoryDFT (BLYP/6-31G*) andthe Semi-empirical SE (AM1) approaches are presented in Table. 4.8. The models presented in Table 4.9were selected as the best (for both DFT and SE) for predicting the toxicities of polychlorinated aromatic (PCDDs) compounds based on the fact that they have the best statistical parameters. The validation parameters of the models agree with the standard reported in Table 3.3.

4.2.2 Correlation matrix for DFT and SE approaches

The descriptors that appeared in the models for both DFT and SE approach are correlated to the pEC50 and the correlation matrix for DFT and SE are presented in Table 4.10 and 4.11 respectively

4.2.3 Statistical/Validation parameters for the models generated for DFT and SE approaches

Statistical parameters for validation of the models for the DFT and SE approaches were calculated and presented in in Table 4.12. There is a good agreement of the validation parameters with the standard reported in Table 3.3.

.

Table-4.8: Four (5) Generated models from DFT and SE techniques for PCDDs

1	DFT	Y = 42.7676 *(BCUTc-1h) + 7.7157 *(VP-3) + 5.9318 *(SssssGe) - 2.22844 *(ETA_dAlpha_B) +32.7353 *(ETA_BetaP) + 1.0527
	SE	Y = 0.18406*(: EHOMO) + 5.08001 *(SP-7) + 10.6744 *(ETA_Shape_P) + 284.0681*(: ETA_EtaP_L) - 0.00225 *(GRAV-4) + 101.5783
2	DFT	Y = 41.0403 *(: BCUTc-1h) + 8.1220 *(nHBd) + 7.31803 *(: SssssGe) - 2.4373*(ETA_dAlpha_B +37.2283 *(ETA_BetaP)+ 15.2040
	SE	Y = 0.1370 *(EHOMO)+ 8.0855 *(SP-7)+ 90.9943 *(: ETA_AlphaP) - 0.0027 *(GRAV-4) - 0.03601 *(Wlambda3.polar) + 66.7416
3	DFT	Y = 6.1456 *(VC-5)+ 6.0276 *(SssssGe) - 2.1371 *(ETA_dAlpha_B) + 38.7000 *(ETA_BetaP)- 6.256765497 *(n5Ring) + 23.3106
	SE	Y = 0.1353 *(EHOMO)+ 8.6647 *(SP-7) + 336.4814 *(ETA_EtaP_L) - 0.00282 *(GRAV-4) - 0.03877 *(Wlambda3.polar) + 46.5057
4	DFT	Y = 6.1653 *(VC-5) + 6.0144 *(SssssGe)- 2.1335 *(ETA_dAlpha_B) + 38.5558 *(ETA_BetaP) - 1.8349*(VAdjMat) + 22.0008

	SE	Y = 0.1931 *(EHOMO) + 6.1332 *(SP-7) - 101.9533 *(ETA_Epsilon_5) - 347.4857 *(WTPT-2) - 0.0028 *(GRAV-4) +975.5696
5	DFT	Y= 6.5646 *(VC-5) + 2.7045 *(VP-3) + 6.2037 *(SssssGe) - 2.2686 *(ETA_dAlpha_B) + 35.3345 *(ETA_BetaP)+ 16.8907
	SE	Y = 0.1904 *(EHOMO) + 6.3303 *(SP-7) - 2.1549*(SsBr) + 114.2770 *(ETA_AlphaP) - 0.00230 *(GRAV-4) + 102.3113

Table 4.9 The best model for DFT and SE for PCDDs

DFT	Y=42.7676*(BCUTc_1h)+7.7158*(VP-3)+5.9318*(SssssGe) -2.2284*(ETA_dApha_B)+32.7353*(ETA_BetaP)+1.0527
SE	Y=0.1841*(EHOMO)+5.0800*(SP-7)+10.6744*(ETA_Shape_P) +284.0681*(ETA_EtaP_L)-0.0023*(GRAV_4)+101.5783

Table-4.10: Correlation matrix for DFT calculated descriptors

	pEC50	BCUTc-1h	VP-3	SssssGe	ETA_dAlpha_B	ETA_BetaP
pEC50	1					
BCUTc-1h	-0.3672	1				
VP-3	0.1166	0.4877	1			
SssssGe	-0.4491	0.6806	0.6530	1		
ETA_dAlpha_B	-0.3274	0.6973	0.8278	0.9602	1	
ETA_BetaP	0.5501	0.0551	0.7339	0.0072	0.2750	1

Table-4.11: Correlation matrix for SE calculated descriptors

	pEC50	*EHOMO*	*SP-7*	*ETA_Shape_P*	*ETA_EtaP_L*	*GRAV-4*
pEC50	1					
EHOMO	-0.4682	1				
SP-7	0.1118	-0.4024	1			
ETA_Shape_P	0.5341	-0.8841	0.7369	1		
ETA_EtaP_L	0.6588	-0.9450	0.2396	0.8140	1	
GRAV-4	0.4395	-0.9255	0.6399	0.9507	0.8773	1

Table-4.12: Statistical/validation parameters for DFT and SE (PCDDs)

Density Functional Theory (DFT)		Semi-Empirical (SE)	
Parameter	Equation1	Parameter	Equation1
Friedman LOF	0.58817700	Friedman LOF	0.38776300
R-squared	0.95165500	R-squared	0.96803200
Adjusted R-squared	0.93893300	Adjusted R-squared	0.95962000
Cross validated R-squared	0.90908100	Cross validated R-squared	0.95180300
Significant Regression	Yes	Significant Regression	Yes
Significance-of-regression F-value	74.80187300	Significance-of-regression F-value	115.07027100
Critical SOR F-value (95%)	2.76172000	Critical SOR F-value (95%)	2.76172000
Replicate points	0	Replicate points	0
Computed experimental error	0.00000000	Computed experimental error	0.00000000
Lack-of-fit points	19	Lack-of-fit points	19
Min expt. error for non-significant LOF (95%)	0.28327500	Min expt. error for non-significant LOF (95%)	0.23000500

4.2.4 Comparison of the observed and predicted pEC50 (PDDs)

The comparison of the predicted experimental and toxicities values (pEC50) of the two modelsfor both DFT and SE approaches are presented in Tables 4.13. The low residual values shown in Table 4.13confirms high predictive power of the models.

4.2.5 Univariate analysis of the of the pEC_{50} data For DFT and SE approaches (PDDs)

Skewness and kurtosis are the most important parameters here. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. Kurtosis which is the fourth moment of the distribution that indicates the profile of the column of data relative to normal distribution, presented as univariate analysis data.The univariate analysis data are presented in Table 4.14

4.2.4Linear relationship of observed pEC_{50} values with the predicted values (PDDs)

The linear relationship of the observed pEC50 with predicted for both DFT and SE approaches are presented in a graphical form as Figure. 4.4 and Figure 4.5 respectively.

4.2.6 Plot of residual versus actual toxicity values for DFT and SE calculated descriptors (PDDs)

Figure 4.6 and 4.7 present the plot of the residual versus actual values for and SE calculated molecule descriptors respectively.

4.3 Results of QSAR modeling of pEC50 for (PCDDs)

The following are the results of the QSTR modeling of PCDDs:

Table-4.13: Actual and predicted toxicity values for DFT and SE (PCDDs)

S/N	Density Functional Theory (DFT)			Semi-Empirical (SE)		
	Actual values pEC50	Predicted values	Residual values	Actual values pEC50	Predicted values	Residual values
1	8.0000	7.8504	0.1496	8.0000	7.6360	0.3640
2	7.1000	7.0987	0.0013	7.1000	7.1386	-0.0386
3	6.8000	6.4319	0.3680	6.8000	6.8130	-0.0130
4	6.6600	6.5102	0.1498	6.6600	6.2991	0.3609
5	6.5500	6.0009	0.5491	6.6500	6.4638	0.1862
6	6.1000	6.4722	-0.3722	6.1000	6.4554	-0.3554
7	5.9600	5.8006	0.1594	5.9600	6.0871	-0.1271
8	5.8900	6.3043	-0.4143	5.8900	5.6780	0.2118
9	7.1500	6.6762	0.4738	7.1500	6.8399	0.3100
10	5.1700	5.6075	-0.4375	5.1900	5.6462	-0.4562
11	4.8900	5.0809	-0.1909	4.8900	4.7512	0.1388
12	5.4900	5.2047	0.2853	5.4900	5.7988	-0.3088
13	5.0000	5.1862	-0.1862	5.0000	4.9205	0.0795
14	4.0000	4.2921	-0.2921	4.0000	4.1177	-0.1177
15	8.8200	8.9274	-0.1074	8.8200	9.1818	-0.3618

16	8.8300	9.1015	-0.2715	8.8300	8.7976	0.0324
17	9.3500	9.1085	0.2415	9.3500	9.1098	0.2402
18	7.9400	8.6603	-0.7203	7.9400	8.5207	-0.5807
19	7.0300	7.1834	-0.1534	7.0300	6.9865	0.0435
20	8.7000	8.6930	0.0070	8.7000	8.6295	0.0705
21	7.7700	7.8034	-0.0334	7.7700	7.5725	0.1975
22	8.1800	8.1236	0.0564	8.1800	8.3289	-0.1490
23	8.9300	8.5154	0.4146	8.9300	8.6621	0.2679
24	7.8100	7.7868	0.0232	7.8100	7.8609	-0.0510
25	6.5300	6.2299	0.3001	6.5300	6.4739	0.0560

Table-14 Univariate analysis for the toxicity data for DFTand Semi-Empirical

Density Functional Theory (DFT)		Semi-Empirical (SE)	
Parameter	Value	Parameter	Value
Number of sample points	25	Number of sample points	25
Range	5.3500	Range	5.3500
Maximum	9.3500	Maximum	9.3500
Minimum	4	Minimum	4
Mean	6.9860	Mean	6.9908
Median	7.0300	Median	7.0300
Variance	2.0054	Variance	1.9994
Standard deviation	1.4453	Standard deviation	1.4432
Mean absolute deviation	1.1834	Mean absolute deviation	1.1784
Skewness	-0.1804	Skewness	-0.1876
Kurtosis	-1.0268	Kurtosis	-1.0170

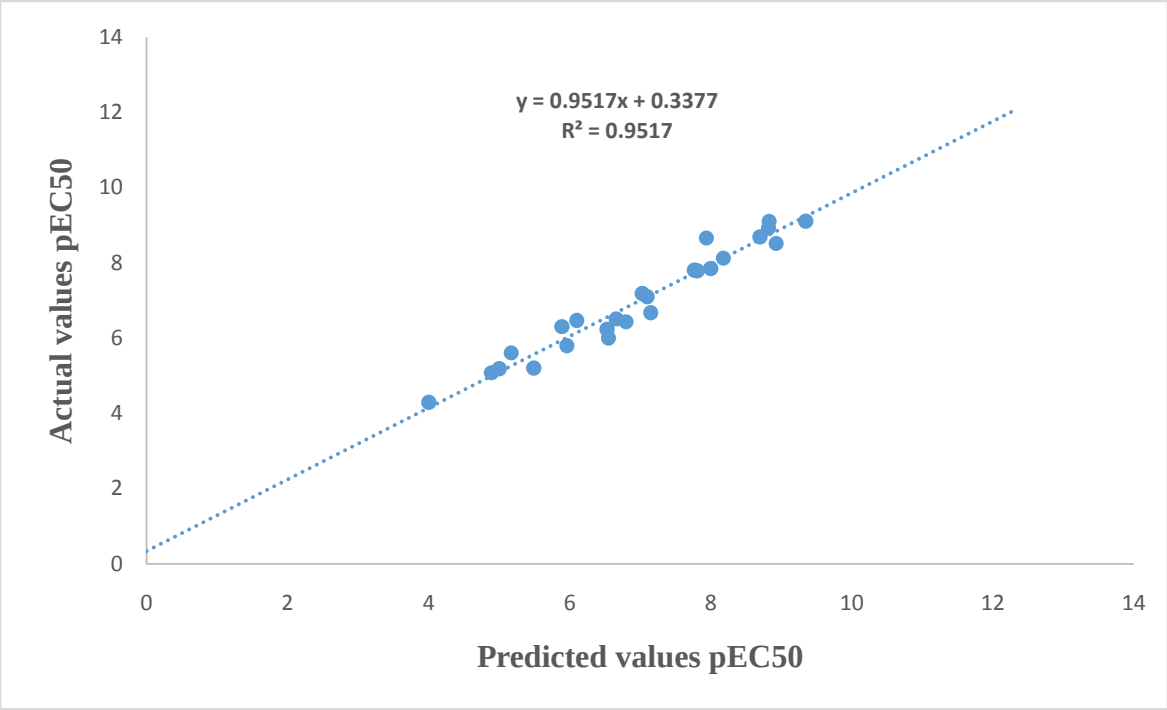


Figure 4.4: Linear relationship between actual toxicity (pEC50) and the predicted for DFT Approach.

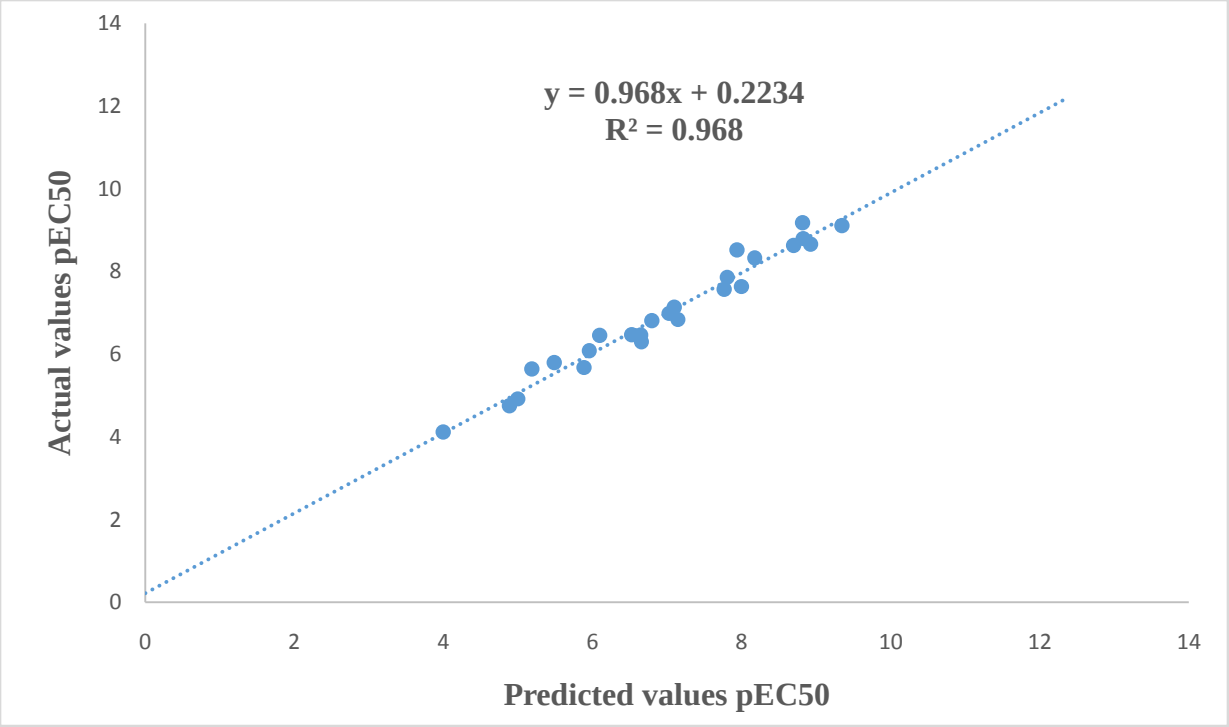


Figure. 4.5: Linear relationship between Actual toxicity (pEC50) and the predicted for SE approach.

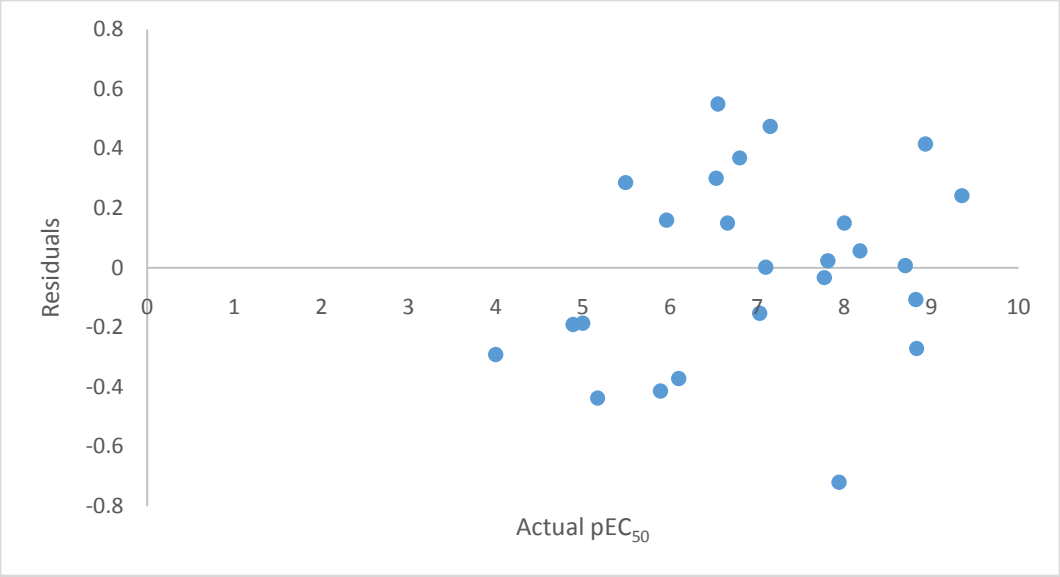


Figure 4.6. Plot of residual versus actual toxicity values for DFT calculated descriptors

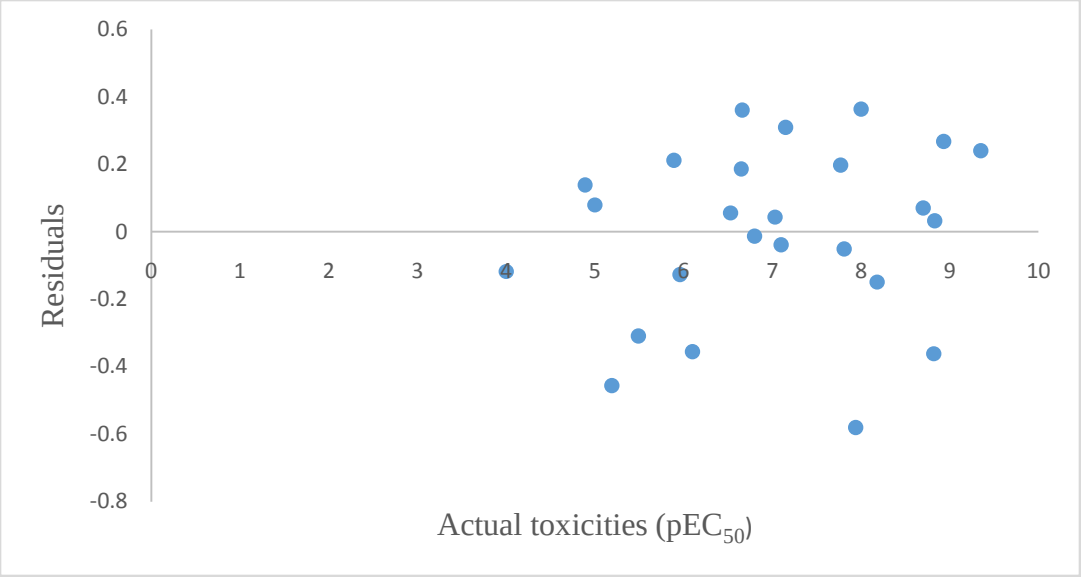


Figure.4.7 Plot of residuals versus actual toxicity values for SE calculated descriptors.

4.3.0 Results of QSTR Modeling of pEC_{50} for PCDFs Using DFT Optimization

QSAR modeling of pEC50 for some of the three classes of polychlorinated aromatic compounds PCDFswas performed DFT(BLYP/6-31G*) and the QSAR results are presented as follows:

4.3.1 GFA Generated models of pEC_{50} for PCDFs

Five (5) QSAR generated GFA models for pEC50 of the class polychlorinated dibenzofuran (PCDFs) were presented as (equations 4.6 to 4.7). Out of the 5 models, model-1 (equation 4.6) was selected as the best for predicting the toxicities of polychlorinated aromatic compounds (PCDFs) based on the fact that it has the best statistical parameters. The validation parameters of the models agree the standard reported in Table 3.3.

4.3.2 Statistical/Validation parameters for the generated models.

Statistical parameters for internal validation of all the five (5) models were calculated and presented in Table4.13. There is a good agreement of the validation parameters with the standard reported in Table 3.3

4.3.3 Comparison of the observed and predicted pEC50

The comparison of the predicted toxicities (pEC50) of the models with the experimental values are presented in Tables 4.14. The low residual values shown in the table confirms the high predictive power of the models.

4.3.4 Univariate analysis of the toxicity (pEC_{50}) data

The univariate analysis data are presented in Table-4.15.

4.3.6 Linear relationship of observed and the predicted pEC_{50} values for PCDFs

The linear relationship between the observed and the predicted (pEC_{50}) for one of the series (PCDFs) is presented in Figure. 4.8

4.4.7 Plot of residual versus actual toxicity (pEC_{50}) values

The plot of the residual versus actual (pEC_{50}) values for one of the series (PCDFs) is presented in Figure. 4.9

Model-1

Y = - 0.022833736 * X7 - 2516.681690760 * X155 - 10.074887633 * X274 - 3.037586747 * X304 - 4.587315200 * X311 + 1148.924708406……..(4.6)

Descriptors

X7: ΔE, X155: ETA_Epsilon_1, X274: MOMI-YZ, X304:WD.mass, X311:Weta2.volume

Model-2

Y = - 0.018210469 * X7 + 21.648748596 * X163 + 118.709554288 * X284 - 4.844391963 * X311 - 38.590840379……(4.7)

Descriptors

X7: DE, X163: ETA_Psi_1, X284: Weta1.unity, X311: Weta2.volume

Model-3

Y = - 714.189206001 * X9 + 21.147237419 * X163 + 117.629297922 * X284 - 4.872513333 * X311- 27.754008563 ……(4.8)

Descriptors

X9: s, X163: ETA_Psi_1, X284: Weta1.unity, X311: Weta2.volume

Model-4

Y = - 892.463510430 * X9 - 2458.279034361 * X155 - 9.974613478 * X274 - 2.971825728 * X304 - 4.623775220 * X311 + 1135.631683428 ……..(4.9)

Descriptors

X9: s, X155: ETA_Epsilon_1, X274: MOMI-YZ, X304: WD.mass, X311: Weta2.volume

Model-5

Y = - 0.022776888 * X7 - 2492.795897707 * X155 - 10.011448000 * X274 - 3.043938772 * X304 - 3.731815324 * X337 + 1137.768184420 …….(4.10)

Descriptors

X7: ΔE, X155: ETA_Epsilon_1, X274: MOMI-YZ, X304: WD.mass, X337: Weta2.polar

4.4.0 Results of QSTR Modeling of pEC_{50} for PCBs obtained using DFT optimization

QSTR modeling of pEC50 for the class of polychlorinated biphenyl (PCBs) was performed using DFT optimization and the QSTR results are presented as follows:

4.4.1 GFA Generated models of pEC_{50} for PCBs

Five (5) QSTR generated GFA models for pEC_{50} of the class, polychlorinated biphenyls (PCBs) are presented as (equation 4.11 to 4.15). Out of the 5 models, model-1 (equation 4.15) was selected as the best for predicting the toxicities of PCBs based on the fact that it has the best statistical parameters. The validation parameters of the models agree with the standard reported in Table 3.3.

4.4.2 Statistical/Validation parameters for the generated models for PCBs

Statistical parameters for the validation of all the 5 models were calculated and presented in Table. 4.16. There is a good agreement of the validation parameters with the standard reported in Table 3.3.

4.4.3 Comparison of the observed and predicted pEC50

The comparison of the predicted toxicity (pEC50) of the models with the experimental values is presented in Tables 4.17. The low residual values shown in Table 4.17 confirms the high predictive power of the models.Table-4.14:

Table 4.15 Statistical/Validation parameters for PCDFs

	Equation 1	Equation 2	Equation 3	Equation 4	Equation 5	Standard
Friedman LOF	0.29556100	0.29567500	0.29716400	0.29841	0.2998	Smaller(~zero)
R-squared	0.92290100	0.90764800	0.90718300	0.922157	0.9217	≥ 0.5
Adjusted R-squared	0.90913400	0.89491000	0.89438100	0.908257	0.9078	≥ 0.5
Cross validated R-squared	0.888597	0.873432	0.849548	0.84712	0.8875	≥ 0.5
Significant Regression	Yes	Yes	Yes	Yes	Yes	Yes
Significance-of-regression F-value	67.034160	71.254393	70.861098	66.3397	66.006	Large value
Critical SOR F-value (95%)	2.582587	2.754787	2.754787	2.5825	2.5825	Small value
Replicate points	0	0	0	0	0	0
Computed experimental error	0.000000	0.000000	0.000000	0.0000	0.0000	0.00
Lack-of-fit points	28	29	29	28	28	---
Minexpt.error fornonsignificant LOF (95%)	0.33486600	0.36121100	0.36211900	0.3364	0.337260	Very small~zero

Table-4.16Comparison of the observed and predicted pEC_{50} for PCDFs

S/N	Actual values (pEC50)	Predicted values (pEC50)	Residual values
1	6.53000000	6.52788400	0.00211600
2	4.53000000	4.14536500	0.38463500
3	3.55000000	3.55076600	-0.0007.660
4	4.38000000	4.46167200	-0.08167200
5	3.00000000	3.12496200	-0.12496200
6	5.36000000	5.49438600	-0.13438600
7	3.61000000	4.42640700	-0.81640700
8	5.05000000	5.04350700	0.00649300
9	5.36000000	5.47666300	-0.11666300
10	4.07000000	4.03831900	0.03168100
11	4.72000000	5.40015600	-0.68015600
12	7.10000000	7.18322700	-0.08322700
13	6.00000000	6.11779600	-0.11779600
14	6.35000000	5.93548900	0.41451100
15	6.46000000	5.88701700	0.57298300
16	6.70000000	6.38054100	0.31945900
17	7.39000000	7.44913600	-0.05913600
18	5.00000000	4.69837800	0.30162200
19	4.70000000	5.24634500	-0.54634500
20	7.13000000	7.49484200	-0.36484200
21	5.89000000	5.64358700	0.24641300
22	7.82000000	7.39179500	0.42820500
23	6.64000000	6.59589500	0.04410500
24	6.57000000	7.41455100	-0.84455100
25	7.33000000	7.26427800	0.06572200
26	5.08000000	5.01867900	0.06132100
27	6.66000000	5.83636700	0.82363300
28	6.46000000	6.21773700	0.24226300

29	6.96000000	6.98408700	-0.02408700
30	6.70000000	6.70221500	-0.00221500
31	6.70000000	6.59181300	0.10818700
32	6.40000000	6.92154200	-0.52154200
33	3.00000000	2.94688200	0.05311800
34	7.60000000	7.18771200	0.41228800

Table-4.17 Results of the univariate analysis for PCDFs

Univariate parameters	A : pEC50
Number of sample points	33
Range	4.82000000
Maximum	7.82000000
Minimum	3
Mean	5.76575758
Median	6.35000000
Variance	1.80759000
Standard deviation	1.36531000
Mean absolute deviation	1.17034000
Skewness	-0.49058000
Kurtosis	-0.94362400

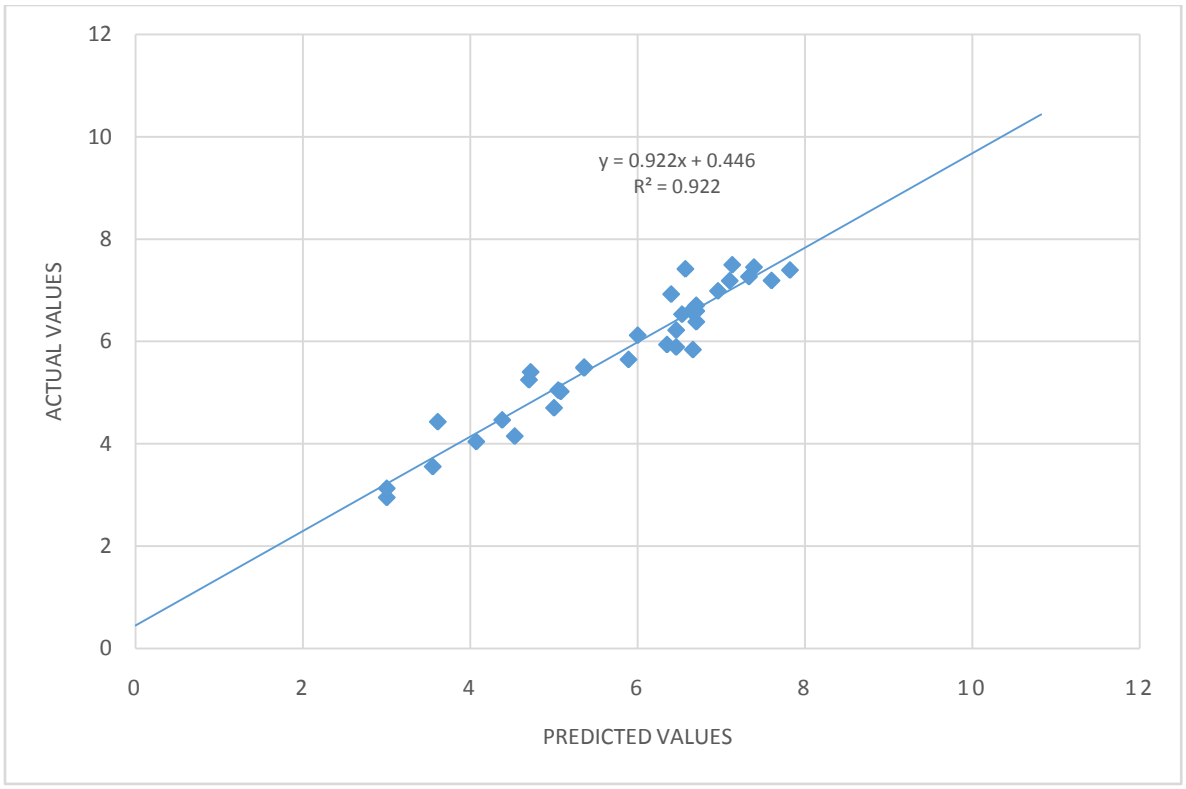


Figure 4.8Linear relationship of observed and the predicted pEC_{50} values for PCDFs

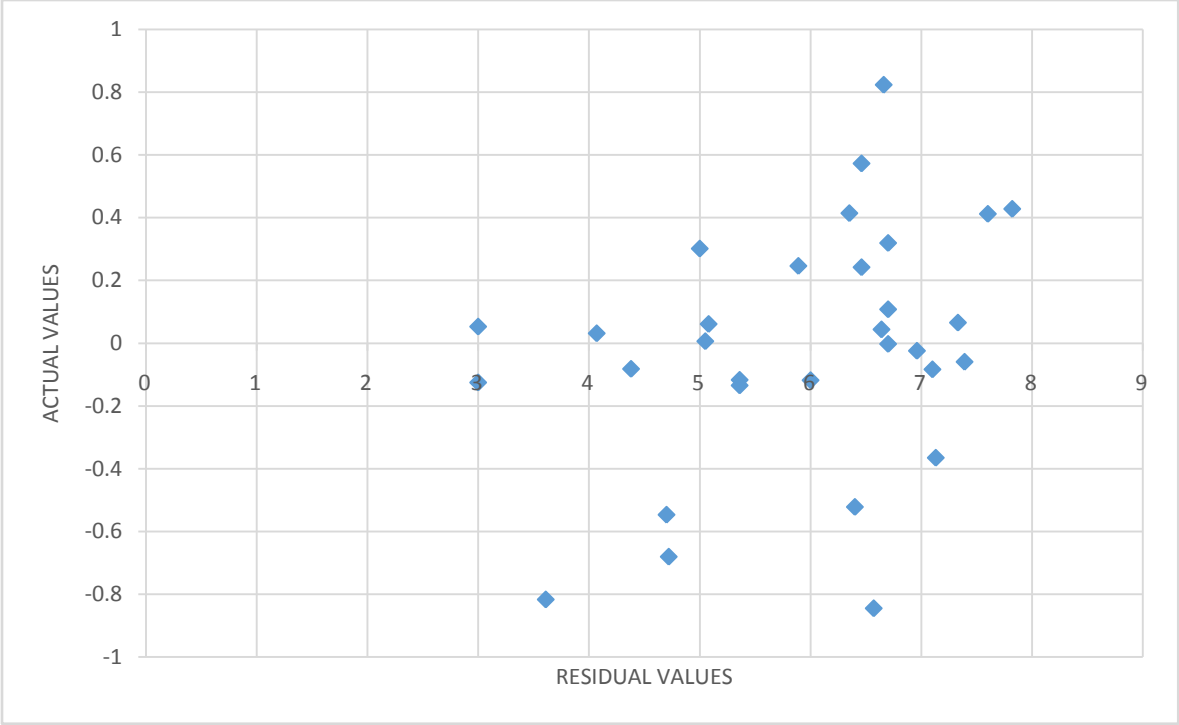


Figure4.9Plot of residual versus actual toxicity (pEC_{50}) values for PCDFs

4.4.4 Univariate analysis of the toxicity (pEC_{50}) data

The most important parameters here are skewness and kurtosis. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. Kurtosis which is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution, presented as univariate analysis data in Table 4.18

4.4.6 Linearrelationship of observed and the predicted pEC_{50} values for PCBs

The linear relationship between the observed and the predicted (pEC_{50}) for one of the series (PCBs) is presented in Figure. 4.10.

4.5.7 Plot of residual versus actual toxicity (pEC_{50}) values

The plot of the residual versus actual (pEC_{50}) values for one of the series (PCBs) is presented in Figure. 4.11.

Model-1

Y = 0.829639808 * X26 + 5.183538692 * X98 - 0.295135065 * X246 + 1.828506774 * X316 + 0.936103312 * X318 - 40.289001916…….. (4.11)

Descriptors

X26: Polarizability, X98: VP-3, X246: Zagreb, X316: Weta2.mass, X318: WT.mass

Model-2

Y = 0.838066521 * X26 + 4.646831237 * X98 + 6.898369161 * X189 + 1.861407164 * X316 + 0.970757393 * X318 - 95.836741280…….(4.12)

Descriptors

X26: Polarizability, X98: VP-3, X189: ETA_Eta, X316: Weta2.mass, X318: WT.mass

Model-3

Y = 0.838066521 * X26 + 4.646831237 * X98 - 2.299456387 * X185 + 1.861407164 * X316 + 0.970757393 * X318 - 42.949244381 …… (4.13)

Descriptors

X26: Polarizability, X98: VP-3, X185: ETA_dBeta, X316: Weta2.mass, X318: WT.mass

Model-4

Y = 0.838066521 * X26 - 1.724592290 * X67 + 4.646831237 * X98 + 1.861407164 * X316 + 0.970757393 * X318 - 35.476011123 …… (4.14)

Descriptors

X26: Polarizability, X67: nBonds, X98: VP-3, X316: Weta2.mass, X318: WT.mass

Model-5

Y = 0.838066521 * X26 + 4.646831237 * X98 - 3.449184580 * X187 + 1.861407164 * X316 + 0.970757393 * X318 - 16.505495932 …….. (4.15)

Descriptors

X26: Polarizability, X98 VP-3, X187 : ETA_Beta_ns_d, X316 : Weta2.mass, X318:WT.mass

Table-4.1 Statistical/Validation parameters for PCBs

	Equation 1	Equation 2	Equation 3	Equation 4	Equation 5	Standard
Friedman LOF	0.45823200	0.46050800	0.46050800	0.46050	0.4605	Small value(~0)
R-squared	0.91861700	0.91821300	0.91821300	0.9182	0.9182	≥ 0.5
Adjusted R-squared	0.87340400	0.87277500	0.87277500	0.8727	0.87277	≥ 0.5
Cross validated R-squared	0.80864800	0.76234200	0.76234200	0.7623	0.76234	≥ 0.5
Significant Regression	Yes	Yes	Yes	Yes	Yes	Yes
Significance-of-regression F-value	20.31757200	20.20827300	20.20827300	20.208	20.2082	—
Critical SOR F-value (95%)	3.50218600	3.50218600	3.50218600	3.5021	3.5025	
Replicate points	0	0	0	0	0	0
Computed experimental error	0.00000000	0.00000000	0.00000000	0.000	0.0000	0.00
Lack-of-fit points	9	9	9	9	9	---
Minexpt. Errorfornon-significant LOF (95%)	0.21617800	0.21671400	0.21671400	0.2167	0.2167	Very small (~0)

Table-4.19 The actual, predicted and residual pEC_{50} values for PCBs

S/N	Actual values for A : pEC50	Equation 1: predicted values	Equation 1: residual values
1	4.94000000	4.93471700	0.00528300
2	6.15000000	5.91385800	0.23614200
3	4.55000000	4.85205200	-0.30205200
4	4.85000000	4.80010200	0.04989800
5	5.37000000	5.18639200	0.18360800
6	5.04000000	5.35222800	-0.31222800
7	5.39000000	5.09155000	0.29845000
8	6.92000000	6.78584200	0.13415800
9	4.26000000	4.55694100	-0.29694100
10	5.15000000	5.05011000	0.09989000
11	4.80000000	4.55694100	0.24305900
12	5.30000000	5.74261600	-0.44261600
13	3.89000000	3.88422200	0.00577800
14	3.85000000	3.95751100	-0.10751100
15	4.00000000	3.79491700	0.20508300

Table-4.20 Results of the univariate analysis for PCBs

Parameters	A : pEC50
Number of sample points	15
Range	3.07000000
Maximum	6.92000000
Minimum	3.85000000
Mean	4.96400000
Median	4.94000000
Variance	0.65089100
Standard deviation	0.83509500
Mean absolute deviation	0.60960000
Skewness	0.61790900
Kurtosis	-0.16243400

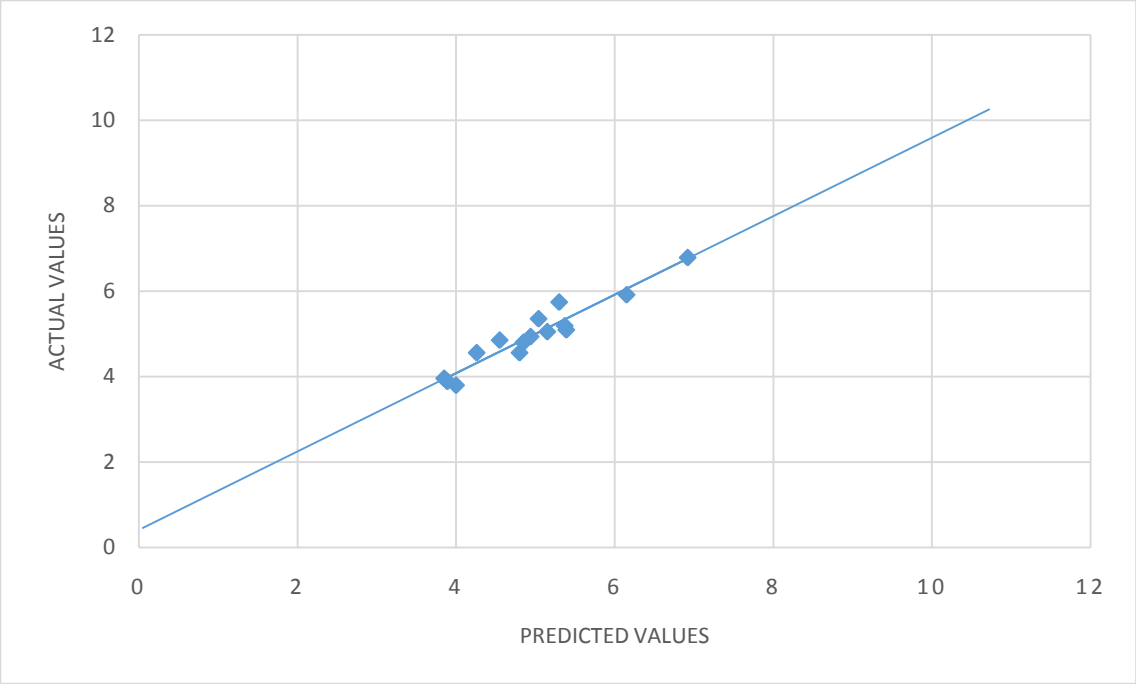


Figure 4.10.Linear relationship of observed and the predicted pEC_{50} values for PCBs

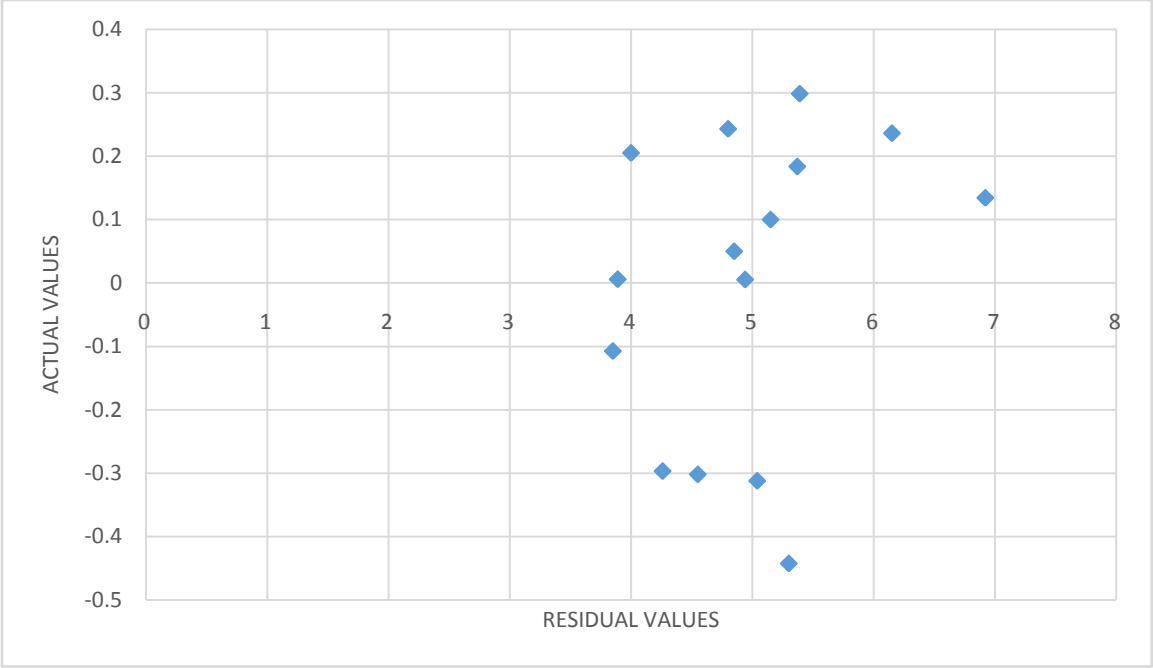


Figure 4.11. Plot of residual versus actual toxicity (pEC_{50}) values for PCBs.

CHAPTER FIVE

5.0 DISCUSSION OF RESULTS

5.1 Geometry Optimization and Calculation of Molecular Descriptors

All the 74 molecules used in this research were completely optimizedusing the level of theory, Density functional theory DFT (B3LY/6-31G*)and Semi-empirical techniques with the basis set AM1. A large number of about 1700 molecular descriptors (0D, 1D, 2D and 3D), which are responsible for encoding important features of the structures, were calculated.The optimization ensures that structures of the optimized molecules were very close to the real structures of the molecules.

5.2 GFA Generated Model for PCDDs, PCDDFs and PCBs all together

Genetic function approximation was used to generate 5 QSTR models on 74 molecules of polychlorinated aromatic compounds using pEC_{50} as dependent variable and the calculated molecular descriptors as independent variables(Polarizability,SP-7,maxHBint5,ETA_dAlpha_B, ETA_Epsilon_2, ETA_Epsilon_5, n5Ring, GRAV-2, GRAVH-3, MOMI-R). Model-1 was selected as the best on the basis of its statistical parameters and was used in predicting the toxicity of polychlorinated aromatic compounds. The statistical parameters of model-1(equation 4.1) were the best set when compared with standard reported in Table 3.1.The statistical parameters of equation-1 include: R^2 =0.967342, R^2_{adj} =0.959178, R^2_{cv}= 0.940154, R^2_{pred}. = 0.7209. This indicates that model-1 is the best among the five models generated.

The predicted toxicities pEC_{50} of both the training set and the test set were in good agreement with theirexperimental toxicities.The predictability of model-1 was evidenced by the low residual values for both the training set and the test sets, respectively.

Table-4.4 shows the statistical parameters of univariate analysis that describe the toxicity data. The most important parameters here are skewness and kurtosis. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. As skewness is positive, the distribution of data value within the column is skewed toward positive values. For a symmetry distribution, the skewness is close to zero. Since the value of the skewness is close to zero, it indicates that there was a symmetry distribution and hence good and reliable model with high predictive power was obtained. Kurtosis is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution

The high linearity (R^2 = 0.9673) for the training set and (R^2=0.6571) for the test set (presented in Figure 4.1 and Figure 4.2 respectively) obtained from the plot of observed pEC_{50} against predicted pEC_{50} also shows the high predictability of the model-1. Furthermore, the propagation of residuals on both sides of zero indicates that no systematic error exists in the development of model-1 (equation 4.1). This is presented in Figure 4.3.

5.2.1 Significance of the descriptors in model 1

The present QSAR model study reveals that polarizability (which determine the dynamical response of a bound system to external fields, and provide insight into a molecule's internal structure), SP-7 (that is the Chi path descriptor with a simple path order 7), ETA_Epsilon_5 (which is themeasure of electronegative atom count 5), GRAVH_3 (cube root of hydrogen-included gravitational index), and MOMI-R (Moment of initia along the radius of gyration) found in model-1 (equation 4.1) are structural features in these polychlorinated aromatic compounds that are significantly responsible for their toxicities.

On the other hand the descriptors like maxHBint5 (Maximum E-state descriptor of strength for potential hydrogen bonds of path length 5), the measure of count of hydrogen bond

acceptor atoms and/or polar surface area(ETA_dApha_B), the measure of electronegative atom count 2(ETA_Epsilon_2, n5Ring), square root of gravitational index of heavy atom(GRAV_2) that also occurred in model-1 (equation 4.1) all have the inverse effects on the toxicities of polychlorinated aromatic compounds. The balance between these descriptors and the ones that have direct effects is what mainly accounts for the net of observable toxicities of the molecules.

The findings in this work agrees with those of Huifeng et al. (2011) and Nandan et al. (2013) about the direct effects of descriptors of polarizability and radius of gyration in explaining toxicities of polychlorinated aromatic compounds.

5.3 GFA GeneratedModels forthe pEC_{50} of PCDDs

Genetic function approximation was used to generate five (5) QSTR models using DFT and Semi-empirical approach on 25 molecules of polychlorinated dioxins (PCDDs) using pEC_{50} as dependent variable and calculated molecular descriptors as independent variables. For the DFT approach, the descriptors used as independent variables are: BCUTc-1h, VP-3, SssssGe, ETA_dAlpha_Shape and ETA_BetaP. For the semi-empirical approach, the following descriptors were used in the model: EHOMO, SP-7, ETA_Shape_P, ETA_Etap_P and GRAV-4.Model-1 was selected as the best for each of the two approaches on the basis of their statistical parameters and were used in predicting the toxicity of polychlorinated dioxins (PCDDs). The models for both DFT and SE and the descriptors used in the models are presented in Table-4.8.

The results of the GFA QSTR model are in conformity with the standard shown in Table 3.3. For DFT:R^2 = 0.95165500, R^2_{adj} = 0.93893300., R^2_{cv} = 0.90908100,LOF= 0.58817700 and for SE: R^2 = 0.96803200, R^2_{adj} = 0.95962000., R^2_{cv} = 0.95180300,LOF= 0.38776300. This confirms the predictive power of the models. These statistical parameters are presented in Table-4.12. The correlation matrix for both DFT and SE which show the correlation of the toxicity with

the molecular descriptors are presented in Table 4.10 and 4.11 respectively.Table-4.13 shows the predicted toxicities (pEC_{50}) of all the 25 molecules which are in very good agreement with the experimental toxicities.The low residual values observed in the Table 4.13 indicates high predictive power of the model.

Figure4.4 and Figure4.5 show plot that describe the linear relationship between the calculated toxicity (pEC_{50}) and the experimental toxicity of the 25 molecules for both the DFT and Semi-empirical calculated descriptors respectively. Most of the compounds in the two figures are along the linear line of the plot. This indicates that, the calculated toxicities (pEC_{50}) are in very good agreement with the experimental values.Figure4.6 and Figure.4.7 for DFT and Semi-empirical calculated descriptors respectively show the plot of the residual values versus experimental values of all the 25 molecules. The propagation of the residual values on both sides of zero indicates that no systematic error exists in the development of the models.

5.3.1 Significance of the descriptors in the models for PCDDs

This study reveals that the following descriptors are found to be responsible for producing toxicities of PCDDs. (DFT: BCUTc-1h, VP-3, SssssGe, ETA_dAlpha_Shape and ETA_BetaP, then for Semi-empirical, we have EHOMO, SP-7, ETA_Shape_P, ETA_Etap_P and GRAV-4). As can be seen in Table-4.9, Extended Topochemical Atom descriptor and Chipath descriptor appear to be found in both of the models generated using DFT and Semi-empirical calculated molecular descriptors. For DFT calculated descriptors, BCUTc-1h, VP-3, SssssGe and ETA_BetaP contribute positively in producing the toxicities of polychlorinated dioxins (PCDDs) because of their positive coefficients in the model. This means that increasing the values of these descriptors will produce high toxicities of these compounds and vice-versa. In the other hand, the descriptor, ETA_dAlpha_Shape which has negative coefficient in the model

contributes negatively in producing the toxicities of these compounds. This indicates that decreasing the values of this descriptor will produce high toxicities of polychlorinated dioxins (PCDDs). For the Semi-empirical calculated descriptors, the descriptors EHOMO, SP-7, ETA_Shape_P and ETA_Etap_P which have positive coefficients in the models contribute positively in producing the toxicities of polychlorinated dioxins (PCDDs). Increasing the values of these descriptors in polychlorinated dioxins will produce higher toxicities (Navin*et al.,* 2012). The descriptor GRAV-4 has negative coefficient in the model and hence contributes negatively in producing the toxicities of these compounds.

5.4 GFA Generated models for the pEC_{50} of PCDFs

Five (5) QSAR models were generated on 34 molecules of polychlorinated aromatic compounds using pEC_{50} as dependent variable and calculated molecular descriptors as independent variables(ΔE, ETA_Epsilon_1,MOMI-YZ, WD.mass, Weta2.volume). Model-1 was selected as the best on the basis of its statistical parameters and was used in predicting the toxicity of polychlorinated aromatic compounds (PCDFs). The results of the GFA QSAR model are in conformity with the standard shown in Table 3.3 as R^2 = 0.92290 , R^2_{adj} = 0.90913, R^2_{cv} = 0.88859, LOF= 0.29556. This confirms the predictive power of the model.

Table-4.14 shows the statistical/validation parameters of all the five (5) model. The statistical quality of the models were determined by the validation parameters like LOF which is the measure of quality of fit, R^2, R^2_{adj}.,R^2_{cv}, F-test. The statistical parameters of model-1(equation 4.6) appears to be statistically better than those of the other four (4) models.Table 4.15 presents the comparison of observed and predicted toxicity (pEC_{50}). The predicted toxicities (pEC_{50}) are in good agreement with their experimental toxicities.The predictability of model-1 is evidenced by the low residual values presented in Table-4.15. The statistical parameters of

univariate analysis that describe the toxicity data is presented in Table-4.16. Skewness and kurtosis are the most important parameters here. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. As skewness is positive, the distribution of data value within the column is skewed toward positive values. For a symmetry distribution, the skewness is close to zero. Kurtosis is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution.

The plot that describes the linear relationship between the calculated toxicity (pEC_{50}) and the experimental toxicity the molecules are presented in Figure. 4.8. Most of the compounds in the two figures are along the linear line of the plot. This indicates that, the calculated toxicities (pEC_{50}) are in very good agreement with the experimental values.Figure 4.9 shows the plot of residuals versus experimental values of data set. The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of GFA model.

5.4.1 Significance of the descriptors in the models for PCDFs

The following descriptors were found to be responsible for producing toxicity of PCDFs:ΔE, ETA_Epsilon_1, MOMI-YZ, WD.mass andWeta2.volume. In this model, all the descriptors have negative coefficient and therefore contribute negatively in producing the toxicity of PCDs. This means that decreasing the values of these descriptors will produce high toxicities of these compounds and vice-versa (Navin*et al.,* 2012).

5.5 GFA Generated models for the pEC_{50} of PCBs

Genetic function approximation was used to generate five (5) QSAR models on the molecules of polychlorinated biphenyls using pEC_{50} as dependent variable and calculated molecular descriptors as independent variables (Polarizability, VP-3, Zagreb, Weta2.mass,WT.mass). Model-1 was selected as the best on the basis of its statistical parameters

and was used in predicting the toxicity of polychlorinated aromatic compounds. The results of the GFA QSAR model are in conformity with the standard reported in Table 3.3 Model-1: R^2 = 0.91862, R^2_{adj} = 0.87340, R^2_{cv} = 0.80864, LOF= 0.45823. This confirms the predictive power of the model.

Table-4.17 shows the statistical/validation parameters of all the five (5) model. The statistical quality of the models were determined by the validation parameters like LOF which is the measure of quality of fit, R^2, R^2_{adj}.,R^2_{cv}, F-test. The statistical parameters of model-1 (equation 4.11) appear to be statistically better than those of the other four (4) models. Table 4.18 presents the comparison of observed and predicted toxicity (pEC_{50}). The predicted toxicities (pEC_{50}) are in good agreement with their experimental toxicities.The predictability of model-1 is evidenced by the low residual values presented in Table-4.11. The statistical parameters of univariate analysis that describe the toxicity data is presented in Table-4.119. Skewness and kurtosis are the most important parameters here. Skewness is the third moment of the distribution, which indicates the symmetry of distribution. As skewness is positive, the distribution of data value within the column is skewed toward positive values. For a symmetry distribution, the skewness is close to zero. Kurtosis is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution.

The plot that describes the linear relationship between the calculated toxicity (pEC_{50}) and the experimental toxicity the molecules are presented in Figure. 4.10 Most of the compounds in the two figure are along the linear line of the plot. This indicates that, the calculated toxicities (pEC_{50}) are in very good agreement with the experimental values.Figure 4.11 shows the plot of residuals versus experimental values of data set. The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of GFA model.

5.5.1Significance of the descriptors in the models for PCBs

The following descriptors were found to be responsible for producing toxicity of PCBs: Polarizability, VP-3, Zagreb, Weta2.mass,WT.mass. In this model, all the descriptors have positive coefficient except the descriptor Zegreb which has negative coefficient of -0.29513. Therefore all of these descriptors except Zegreb contribute positively in producing the toxicity of PCBs. This means that increasing the values of these descriptors will produce high toxicities of these compounds and vice-versa (Navin*et al.,* 2012). But increasing the value of the descriptor Zegreb having negative coefficient in the model will lead to the decrease in toxicity of polychlorinated biphenyls (PCBs).

CHAPTER SIX

6.0 SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1 Summary of the Findings

QSTR analysis on the three series of polychlorinated aromatic compounds (PCDDs, PCDFs and PCBs) with toxicity expressed as $pEC_{50}$was preformed using robust statistical technique GFA, coupled with the use of combination of different classes of descriptors.The generated equations in each QSTR model were analyzed fortheir statistical significance and predictive abilityby using the validation parameters(internal and external validation). Among the models generated, the one with the best statistical significance was selected as the best model for the prediction of the toxicity of polychlorinated aromatic compounds and even other related compounds.

Out of the large number of molecular descriptors generated, only these few of them were found to responsible for producing toxicity of polychlorinated compounds. These descriptors include: polarizability, Chi path descriptor with a simple path length order 7 (SP-7), Extended Topochemical Atomic descriptor (ETA_Epsilon_5) which is the measure of electronegative atom count 5, cubic root of hydrogen-included gravitational index (GRAVH-3) and moment of inertia along the radius of gyration (MOMI-R). All the calculated molecular descriptors were aimed to encode some important information (toxicity) about the structural features of polychlorinated aromatic compounds which could influence the receptor binding affinity. Some of them provided good correlations and statistically reliable models.

Also, Genetic Function Approximation (GFA) technique was used to establish a correlation between (DFT and Semi-empirical calculated molecular descriptors) and

theexperimental toxicities of polychlorinated dioxins. Each of DFT (BLYP/31G*) and Semi-empirical (AM1)calculated descriptors were used to generate 5 QSAR models. The models with the highest statistical significant for both DFT (BLYP/31G*) and Semi-empirical (AM1) calculated descriptors was selected. These models were selected based on their statistical parameters and were used to predict the toxicities of polychlorinated dioxins (PCDDs). The prediction of the toxicity efficiencies for both DFT (BLYP/31G*) and Semi-empirical (AM1) calculated descriptors matched with the experimental measurements. The results of the optimization between DFT and Semi-empirical (AM1) calculated descriptors. This work reveals that semi-empirical (AM1) calculated descriptors gives a little bit better predictions than DFT (BLYP/31G*) calculated molecular descriptors.Our results open very interesting perspectives regarding polychlorinated aromatic compounds with toxicity.

6.2 Conclusion

All the objectives stated in this research were achieved.The generated equations in each QSTR model were analyzed fortheir statistical significance and predictive abilityby using the validation parameters for both internal and external validation for the models generation. Among the models generated, the one with the best statistical significance was selected as the best model for the prediction of the toxicity of polychlorinated aromatic compounds and even dioxin-like compounds. The developed QSTR model had good robustness, predictive ability and mechanism interpretability, which could be applied to predict the toxicity of other related polychlorinated aromatic of compounds.

6.3 Recommendation

Some of the methods and results of the present investigation may be readily extended to other polychlorinated aromatic and dioxin-like compounds and, if so, this investigation will

contribute to understanding the chemistry of the whole class of these dangerous pollutants. Therefore, it is recommended based on the research work that:

ii) Using the molecular descriptors in the models which were aimed to encode some important information about the structural features of these polychlorinated aromatic compounds which could influence the receptor binding affinity, polychlorinated aromatic compounds with low toxicity should designed and even synthesized.

ii) The toxicity test of chemical compounds should be carried out before they are used. The models in this researched should be used to test the toxicity of any polychlorinated compound.

iii) Further QSTR work should be done on the pharmacokinetic properties, docking and 3D-QSAR studies on the Ah receptor binding affinities of these organic pollutants.

iv)Riskassessment which is a science that can be improved upon for many classes of chemicals. Chemical analysis, bioassays, and QSTRs can be used to improve the reliability and reduce the uncertainty in a risk assessment. The combination of these techniques to develop improved methods of toxicity assessment is a promising avenue for those classes of chemicals that are currently lacking sufficient toxicity data.

REFERENCES

A. V. Raichurkar and V. M. Kulkarni (2003). 3D- QSAR of cyclooxygenase-2 Inhibitor by Genetic Function Approximation, Internet Electron. J. mol. Des. 242-261.

Adrian, B., Alexadru T. B. (2004). QSAR for toxicity of polychroodibenzofurans, polychlorodibenzo-1,4-dioxins and polychlorobiphenylsp 163-182.

Ahlborg, U.G., Brouwer, A., Fingerhut, M.A., Jacobson, J.L., Jacobson S.W., Kennedy, S.W…, Van den Berg, M. (1992). Impacts of polychlorinated dibenzo-p-dioxins, dibenzofurans and biphenyls on human and environmental health, with special emphasis on application of the toxic equivalency factor concept. *European Journal of Pharmacology. Environmental Toxicology and Pharmacology Section.* 228:178-199

Alcock, R. E., and Jones, K. C. (1996). Dioxins in the environment: a review of trend data. *Environmental Science & Technology*, 30(11), 3133-3143

Aptula, A.O, Robert, D.W. (2006). Mechanistic application domains for non-animalbased prediction of toxicology endpoints: general principles to reactive toxicity. 19: 1097-1105.

Athanasiadou, M., Cuadra, S.N., Marsh, G., Bergman, A…,Jacobsson, A. (2008). Polybrominated diphenyl ethers (PBDEs) and bioaccumulative hydroxylated PBDE metabolites in young humans from Managua, Nicaragua. *Environmental Health Perspectives.* 116:400-408

Azeddine, A., Rachid, H., Larif, M. (2014). Binding affinies (AhR) of polychlorinated Biphenyls (PCBs), Dibenzo-P-dioxins (PCDDs) and Dibenzofurans (PCDFs) Study Combining DFT and QSAR Result. Vol. 4:4

Azeddine, T.L., Rachid, A., Majdouline, H., and Mohammed, L. (2014). Quantitative structure-toxicity relationship of some polychlorinated aromatic compounds. *IJARCSSE*, 4, 304-305.

Balaban, A. T. (1982). Highly Discriminating Distance-Based Topological Index. *Chemical Physics Letters.* 89, 399–404.

Balaban, A.T., and Beteringhe, A. (2004). QSAR for toxicities of polychlorodibenzofurans, polychlorodibenzo-1,4-dioxins, and polychlorobiphenyls *ARKIVOC*, (i) 163-182 DOI: http://dx.doi.org/10.3998/ark.5550190.0005.116 Paper ID: 04-1081FP

Bandiera, S., Sawyer, T., Romkes, M., Zmudzka, B., Safe, L., Mason, S., Keys, B., Safe, S. (1984). Polychlorinated dibenzofurans (PCDFs): effects of structure on binding to the 2,3,7,8-TCDD cytosolic receptor protein, AHH induction and toxicity.*Toxicology*. 32(2):131-44.

Basak, S. C., Gute, B. D., and Grunwald, G. D. (1997). Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *Journal of Chemical Information and Computer Sciences*. ,37, 651–655.

Basler, A. (1995). Dioxins and related compounds: Status and regulatory aspects in Germany. *Environmental Science and Pollution Research International*2(2):117-21.

Beychok, M.R. (1987). "A data base of dioxin and furan emissions from municipal refuse incinerators". Atmospheric Environment. 21 (1): 29–36.

Bjeldenes, L.F., Kim, J.Y., Grose, K.R., Bartholomew, J.C., and Bradfield, C.A. (1998). Ah-receptor agonists generated from indole-3-carbinolin vitro and in vivo. Comparisons with 2,3,7,8-tetrachlorodibenzo-p-dioxin. Proceedings of the national.

Breiman, L. (2001). Random Forests. Learn. 45, 5–32

Buolamwini, J. K., and Assefa, H. (2002). CoMFA and CoMSIA 3D QSAR and Docking Studies on Conformationally-Restrained Cinnamoyl HIV-1 Integrase Inhibitors: Exploration of a Binding Mode at the Active Site. *Journal of Medicinal Chemistry*, 45, 841–852

Burden, F. R. (1997). A Chemically Intuitive Molecular Index Based on the Eigen values of a Modified Adjacency Matrix. *Quantitative Structure-activity Relationships*,16, 309–314

Caprioara, M., and Diudea, M. (2003). QSAR studies on polychlorinated aromatic compounds.*Indian J Chem*. 42, 1368-1378.

Cerniglia, C. E. (1992). "Biodegradation of polycyclic aromatic hydrocarbons". Biodegradation. **3** (2–3): 351–368.

Chovancova, J., Kocan, A., Jursa, S. (2005). PCDDs, PCDFs and dioxin-like PCBs in food of animal origin (Slovakia). *Chemosphere* 61:1305-1311.

Chuman, H., Karasawa, M., Fujita, T. A. (1998). Novel 3-Dimensional QSAR Procedure – Voronoi Field Analysis. *Quantitative Structure-activity Relationships.* 17, 313–326

Costner, Pat (2005) "Estimating Releases and Prioritizing Sources in the Context of the Stockholm Convention", International POPs Elimination Network, Mexico.

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988). Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins.*Journal of the American Chemical Society*. 110, 5959–5967

Cruciani, G., Pastor, M., and Guba, W. (2000). VolSurf: A New Tool for the Pharmaceutic Optimization of Lead Compounds. *European Journal of Pharmaceutical Science* 2000, 11(Suppl.), S29–S39.

Cruciani, G.; Pastor, M.; Guba, W. VolSurf (2000), A New Tool for the Pharmaceutic Optimization of Lead Compounds.Eur. *Journal of Pharmaceutical Science* .11 (Suppl.), S29–S39

Czuczwa, J. M., and Hites, R.A. (1984). Environmental fate of combustiongenerated polychlorinated dioxins and furans. *Environmental Science & Technology*, 18(6), 444-450

Czuczwa, J. M., McVeety, B. D., and Hites, R. A. (1984). Polychlorinated dibenzo-p-dioxins and dibenzofurans in sediments from Siskiwit Lake, Isle Royale. Science, 226(4674), 568-569

Devillers, Y. and Balaban, A.T. (1999). In eds: Gordan and Breach, Topological indices and related descriptors in QSAR and QSPR, Amsterdam. The Netherland.

Devito, M.T., and Birnbaum, L.S. (1995). Dioxins: model chemicals for assessing receptor-mediated toxicity. *Toxicology* 102(1-2):115-23.

Domingo, J.L., (2007). Levels of PCDD/PCDF PCBs in edible marine species, S., Reifferscheid, G., Hecker, M., Giesy, J. P., ...Hollert, H. (2014). In vitro bioassays for detecting dioxin-like activity—Application potentials and limits of detection, a review. *Science of the Total Environment*, 487, 37-48.

Eisen, H., Hannah, R.R., Allan, C.L., Okey, B., and Nebert D.W. (1993). The Ah-receptor: controlling factor in the induction of drung-metabolizing enzymes by certain chemical carcinogens and other environmental pollutants. Biochemical actions of hormones. Vol. X San Diego, C.A, Academic Press, pp. 227-258.

Espandiari, P., Twaroski, T. P., Festag, M., Lee,E. Y., Glauert, H. P., Robertson, L. W. G. Eds;

Estrada, E., and Uriarte, E., (2001). Recent advances on the role of topological indices in drug discovery research. *Current medicinal chemistry*. 8(13):1573-88.

European Chemicals Agency, 2011, http://www.qsartoolbox.org

Ferguson, A. M., Heritage, T. W., Jonathon, P., Pack, S. E., Phillips, L., Rogan, J., Snaith, P. J. (1997). EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis. *Journal of Computer-Aided Molecular Design*. 11, 143–152.

Fiedler, H.; Bellin, J., and Barnes (1990). Potential for human exposure-toxic dioxins. Chemosphere, 20, 1597.

Fiedler,H. (1996). Sources of PCDD/PCDF and impact on the environment.*Chemosphere*

32, 55-64

Free, S. M., and Wilson, J. W. (1964). A Mathematical Contribution to Structure–Activity Studies. *Journal of Medicinal Chemistry*. 7, 395–399.

Gajewicz, A., Rasulev, B., Dinadayalane, T.C., Urbaszek, P., Puzyn, T., Leszczynska, D., and Leszczynski, J. (2012). Advancing risk assessment of computational approach. *AdvancedDrugDeliveryReviews*.110:1016.

Giesy, J. P., Ludwig, J. P., & Tillitt, D. E. (1994). Deformities in birds of the Great Lakes region. *Environmental science & technology*, 28(3), 128A-135A.

Giesy, J. P., Ludwig, J. P., &Tillitt, D. E. (1994). Deformities in birds of the Great Lakes region. Environmental science & technology, 28(3), 128A-135A.

Goodford, P. J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *Journal of Medicinal Chemistry*., 28, 849–857

Gutman, I. and Trinajstic, N. (1972). Graph Theory and Molecular Orbitals. Total Electron Energy of Alternant Hydrocarbons. *Chemical Physics Letters.* 1972,17, 535–538

Harary, F. (1969). Graph Theory; Addison-Wesley: Reading, MA, 1969

Hawkins, D.M., Basak, S.C., and Mills, D. (2003). Assessing model fit by cross validation. Journal of chemical information and computer sciences, 43(2): 586-587.

Holland, J.H., Adaption in Natural and Artificial Systems, University of MichiganPress, Ann Arbor, MI, 1975.

Hosoya, H. (1971). Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons.Bull. *Chemical Society of Japan*. 44, 2332–2339.

Hou, T., and Xu, X. (2004). Genetic algorithms in chemometrics. *Chem*. 2004; 16: 35-41.

Huisman, M., Koopman-Esseboom, C., Fidler, V., Hadders-Algra, M., van der Paauw, C.G., Tuinstra, L.G., Weisglas-Kuperus, N., Sauer, P.J., Touwen, B.C., Boersma, E.R. (1995). Perinatal exposure to polychlorinated biphenyls and dioxins and its effects on neonatal neurological development. *Early Human Development*. 41:111-127.

Ivanciuc, O., and Balaban, A. T. (1999). The Graph Description of Chemical Structures. In Topological Indices and Related Descriptors in QSAR and QSPR; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, the Netherlands, 1999; pp 59–167

Ivanciuc, O., Balaban, T.S., and Balaban, A. T. (1993). Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices *Journal of mathematical chemistry*. 12, 309–318

Jain, A. N., Koile, K., and Chapman, D. (1994). Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *Journal of Medicinal Chemistry*. 37, 2315–2327

Jaiswal, M., Khadikar, P.V., Scozzafava, A., Supuran, C.T. (2004). carbonic anhydrase inhibitors: the first QSAR on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. *Bioog. Med. Chem. Lett*., 14, 3283-3290.

Janez˘ic, D., Milicevic, A., Nikolic, S., and Trinajstic, N. (2007). Graph Theoretical Matrices in Chemistry; University of Kragujevac: Kragujevac, Serbia, 2007; 205 pp

Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: A review, *Alternative Laboratory Animals*, 33, 445–459

Jouko, T., Terttu, V., Tuomisto, T.J., Mannerheimintie, 166, Helsinki, Finland, 2011. http://creativecommons.org/licenses/by-sa/3.0/deed.fi ISSN 1798-0089 www.thl.fi/dioksiini

Kalson, M., Lobanov V.S. and Katritzky, A.R (1996). Quantum chemical descriptors in QSAR/QSPR studies. Chem. Rew. 96, 1027.

Karelson, M., (2001). Molecular descriptor in (QSAR/QSP)., John Wiley & Sons, New York

Katami, Takeo; Yasuhara, Akio; Okuda, Toshikazu; Shibamoto, Takayuki; et al. (2002). "Formation of PCDDs, PCDFs, and Coplanar PCBs from Polyvinyl Chloride during Combustion in an Incinerator". Environ. Sci. Technol. 36 (6): 1320–1324.

Katritzky, A. R., Mu, L., Lobanov, V. S., Karelson, M. (1996). Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *Journal of Physical Chemistry*. 1996,100, 10400–10407

Katritzky, A.R., Fara, D.C., Yang, H., Karenlson, M., Suzuki, T., Solov'ev V.P. (2009). Varnek. J chem. inf. Comput. Sci. 144, 529.

Katritzky, A.R., Maran, U., Lobanov V.S., krelson M., &Vhem Y (2000). Structurally Diverse Quantitative structure property relationship correlation of TechnologicallynRelevant physical properties. Info. *Computational Science.* 40.1

Katritzky, A.R., Pentrukhin, R., Tatham, D., Basak, S., Benfenatim, E., Karelaon, M., and Maran, U. (2001). Quantitative structure-toxicity relationships for naphathlisoquinoline dericatives as antimalarial agents using molecular descriptors. *J Chem. Info. & Comput. Sci.*,2001; 40: 1–18.

Khaled, K.F. (2011). Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model. *Corrosion sci.* 2011, 53, 3457-3465.

Kier, L. B. (1985). A Shape Index from Molecular Graphs. *Quantitative Structure-activity Relationships*. 4, 109–116.

Kier, L. B., and Hall, L. H. (1997) The Nature of Structure–Activity Relationships and Their Relation to Molecular Connectivity. *European Journal of Medicinal Chemistry*. 1977, 12, 307–312

Kitazawa, A.; Amagai, T.; Ohura, T. (2006). "Temporal Trends and Relationships of Particulate Chlorinated Polycyclic Aromatic Hydrocarbons and Their Parent Compounds in Urban Air". Environmental Science & Technology. 40 (15): 4592

Klebe, G., Abraham, U., Mietzner, T. (1994). Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *Journal of Medicinal Chemistry*.1994, 37, 4130–4146.

Klopman, G. (1968). Chemical reactivity and the concept of charge- and frontier-controlled reactions. *Journal of the American Chemical Society.* 90 (2), pp 223–234

Knight, D.J and Behenu, D. (2002). Alternative to animal testing in the safety evaluation of products, *Alternatives to Laboratory Animals*. 30, 7.

Kohonen, T. (1989). Self-Organization and Associative Memory; Springer: Berlin, Germany.

Kulkarni, P.S., Crespo, J.G., Afonso, C.A. (2008). Dioxins sources and current remediation technologies. A Review. *EnvironmentalInternational*, 34:139-153.

Kunal, R., Roy, P.P., Paul, S., and Mitra,I. (2009). On Two Novel Parameters for Validation Predictive QSAR Models. *Molecules*, 2009, 14, 1660-1701.

Lang, V. J. (1992). Polychlorinated biphenyls in the environment. Journal of *Chromatography. 595*;*1* – 43.

Larsson, M., Hagberg, J., Rotander, A., van Bavel, B., & Engwall, M. (2013). Chemical and bioanalyticalcharacterisation of PAHs in risk assessment of remediated PAH-contaminated soils. *Environmental Science and Pollution Research*, 20(12), 8511-8520.

Livingstone, D. J., and Salt, D. W. (1992). Regression Analysis for QSAR Using Neural Networks. Bioorganic and Medicinal Chemistry Letters. 2, 213–218.

Luiz F. Motto., and Wanda P. A (2011). Quantutative Structure-Activity Relationships (QSAR) of a series of ketone derivative as anti-candida albicans, international Journal of drug discovery. Vol. 3, pp: 100-117.

Mason, G., Farrell, K., Keys, B., Piskorska-Pliszczynska, J., Safe, L., Safe, S. (1986). Polychlorinated dibenzo-p-dioxins: quantitative in vitro and in vivo structure-activity relationships.*Toxicology*. 41(1):21-31.

McKay, G. (2002). Dioxin characterisation, formation and minimisation during municipal solid waste (MSW) incineration: *review*. *Chemical Engineering Journal*, 86(3), 343-368.

Mekenyan, O., Ivanov, J., Veith, G. D., and Bradbury, S. P. (1994). Dynamic QSAR: A New Search for Active Conformations and Significant Stereoelectronic Indices. *Quantitative Structure-activity Relationships.* 13, 302–307.

Mekenyan, O., Karabunarliev, S., and Bonchev, D. (1990). The OASIS Concept for Predicting Biological Activity of Chemical Compounds. *Journal of Mathematical Chemistry*. 4, 207–215

Mekenyan, O., Nikolova, N., and Schmieder, P. (2003). Dynamic 3D QSAR Techniques: Applications in Toxicology. *Journal of Molecular Struct*ure (Theochem) 622, 147–165.

Nandan, K., Ranjan, K., Ahmad, B., and Sah, B. (2013). QSAR studies on polychlorinated aromatic compounds.*IJPSR,* 2013, 4(7), 2691-2695.

Navin, R., Sanmati K. J. (2012). 2D QSAR Study of Substituted 2-Phenyl-Benzimidazole Derivatives as Potent Anti-Allergic Agents. *International Journal of Chemical Technical Research.* Vol.4, No.4, pp 1350-1360.

Netzeva, T.I. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Alternative to Laboratory Animals*,33 (2), 155-173.

Nilsson, U. L.; Oestman, C. E. (1993). "Chlorinated polycyclic aromatic hydrocarbons: Method of analysis and their occurrence in urban air". Environmental Science & Technology. 27 (9): 1826.

Ohura, T. (2007). "Environmental Behavior, Sources, and Effects of Chlorinated Polycyclic Aromatic Hydrocarbons". The Scientific World Journal. **7**: 372–380.

Ohura, T.; Fujima, S.; Amagai, T.; Shinomiya, M. (2008). "Chlorinated Polycyclic Aromatic Hydrocarbons in the Atmosphere: Seasonal Levels, Gas-Particle Partitioning, and Origin". Environmental Science & Technology. **42** (9): 3296.

Pastor, M., Cruciani, G., McLay, I. M., Pickett, S. D., and Clementi, S. (2000). GRid-INdependent Descriptors (GRIND): A Novel Class ofAlignment-Independent Three-Dimensional Molecular Descriptors. *Journal of Medicinal Chemistry*.43, 3233–3243

Patel, V., Chabria, M.T., Brahmkshatriya, P.S., Mahajan, B., and Pael, S. (2013). Quantum mechanical scoring: structure and energetic insights into cyclin-dependent kinase 2-inhibition by pyrazolo [1,5-a] pyrimidines.*Pharmagene*, 2013, 1, 234-282.

Pathak, A.K., Mundada, A.B., and Shrivastava, A. (2011). 2D QSAR study of Quinazoline derivatives as potent antimalarial agent.*Pharmacia,* 2011, 1, P, 57.

Paull, K.D., Friend, S.H., and Weinstein, J.N. (1998). Molecular Descriptors for chemoinformatics*J Chem. Inf. Comput. Sci.* 1998, 38, 189.

Poiger, H.; Pluss, N.; Schlater, C. *Chemosphere* 1989, 18, 265.

Poland, A., and Knutson, J. C. (1982). 2, 3, 7, 8-Tetrachlorodibenzo-thorn-dioxin and related halogenated aromatic hydrocarbons: examination of the mechanism of toxicity. *Annual review of pharmacology and toxicology*, 22(1), 517-554.

Pomarentz, I.; Burke, J.; Firestone, D.; McKinney, J.; Roach, J.; Trotter, W. (1978). QSAR for toxicities of polychlorodibenzofurans, polychlorodibenzo-1,4-dioxinsand polychlorobiphenyls. 24, 133.

Prinz, B. et al. (1990). Polychlorierte dibenzodioxine and dibenzofurane, untersuchungen zurbelastung von gartenboden and nahrurigspflanzen.Staub-Reinhalt Lufi, 50:377-381

Putavy, C., Devillers, J., Domine, D., and Devillers, J. (Ed), Genetic Algorithms in molecular modeling. Academic press London, 1996, p 243.

Puzyn**,** T**.,** Leszczynski**,** J., and Cronin, M. T. (Eds.) Recent Advances in QSAR Studies Methods and Applications ISBN 978-1-4020-9783-6 1 PP03-125.

Raevsky, O. A., Trepalin, S. V., and Razdol'skii, A. N. (2000). New QSAR Descriptors Calculated from Interatomic Interaction Spectra. *Pharmaceutical Chemistry Journal.* 34, 646–649.

Randic, M. (1975). Graph Theoretical Approach to Local and Overall Aromaticity of Benzenoid Hydrocarbons.*Tetrahedron* 1975, 31, 1477–1481

Randic, M. (1996). Molecular Bonding Profiles. *Journal of Mathematical Chemistry*.19, 375–392

Randic, M. (2001). Novel Shape Descriptors for Molecular Graphs. *Journal of Chemical Information and Computer Sciences.* 41, 607–613.

Randic, M.., and Wilkins, C. L., (1979). Graph Theoretical Ordering of Structures as a Basis for Systematic Searches for Regularities in Molecular Data. *Journal of Physical Chemistry*.1979,83, 1525–1540.

Rigo, H. G.; Chandler, A. J.; Lanier, W.S. (1995). The Relationship between Chlorine in Waste Streams and Dioxin Emissions from Waste Combustor Stacks *(PDF).* American Society of Mechanical EngineersReport CRTD. 36. New York, NY: American Society of Mechanical Engineers. *ISBN 0-7918-1222-7.*

Robinson, D. D., Winn, P. J., Lyne, P. D., Richards, W. G. (1999). Self-Organizing Molecular Field Analysis: A Tool for Structure–Activity Studies. *Journal of Medicinal Chemistry*. 42, 573–583

Rogers, D., and Hopfinger, A.J. (1994). Application of Genetic function Approximation to Quantitative structure Activity Relationships and Quantitative property Relationships. *J. Chem. Inf. Comput. Sci.* 1994, 34, 854–866.

Rogers, D., Data analysis using G/SPLINES, in: Advances in Neural Processing Systems 4, Morgan Kaufmann, San Mateo, CA, 1992.

Roggers, D., Devillers, J. (Ed.), "Genetic Algorithm in Molecular Modeling", Academic, London, 1996, pp.87–107

Roy, C.K., Olson, S., Graveley, B.R., Zamore, P.D., Moore, M.J. (2015). Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNAligation. 4(3700).

Safe, S. (1990). Polychlorinated biphenyl (PCBS) dibenzo-p-dioxins (PCDDs), dibenzofurans (CPCDFs), and related compounds: Environmental and mechanistic considerations which

support the development of toxic equivalency factors (Tets). Critical review of toxicology 21:51-88

Safe, S. H. (1994). Polychlorinated biphenyls (PCBs): Environmental impacts, biochemistry, toxicology and mechanism of action. Critical Reviews of Toxicology. 12, 319-395.

Safe, S., Safe, L., and Mullin, M. "Introductory chapter: Polychlorinated biphenyls: "Effect of 4chlorobiphenyls as an initiator of rat liver carcinogenesis". In PCBs – Recent Advances in the Environmental Toxicology and Health Effects; Robertson, L. W.; Hansen, L. Environmental occurrence and analysis". In Environmental Toxin Series 1. Polychlorinated.

Sahu, V.K., and Singh, R.K. (2009). Clean-Soil, air, water, 37,850Walker, J.D., Joworska, J., Comber M.H., I, Schultz T.W and Deardn J.C (2003). Guidelines for developing and using quantitative structure activity relationships. Environmental toxicological and chemistry 22, 165-1665.

Schultz, T.W., Cronin, M.T.D, Walker, J.D., and Aptula, A.O. (2003). Quantitative structure activity relationships (QSARs) in toxicology: a historical perspective. Journal of molecule structure 622, 1-22.

Schuur, J., Selzer, P., and Gasteiger, J. (1996).The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *Journal of Chemical Information and Computer Sciences*. 36, 334–344

Shapiro S, S., Guggenhein, B., Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. *Quantitative) Structure Activity Relationships.* 17, 327-337.

Sofie Van Damme. "Quantum Chemistry in QSAR, Quantum Chemical Descriptors, use, benefits and drawback". Thesis, department of inorganic and physical chemistry Faculty of Science, Universiteit Gent (2009), p 39.

Song, M., Jiang, Q., Xu, Y., Liu, H., Lam, P. K., O'Toole, D. K., & Jiang, G. (2006). AhR-active compounds in sediments of the Haihe and Dagu Rivers, China. *Chemosphere*, 63(7), 1222-1230.

Steiglitz, L., and Vogg, H. (February 1988) "Formation Decomposition of Polychlorodibenzodioxins and Furans in Municipal Waste" Report KFK4379, Laboratorium fur Isotopentechnik, Institut for Heize Chemi, Kerforschungszentrum Karlsruhe.

Sugita, k., Asada, S., Tetsuaki, Y., Mitsuhiro, O., and Sumio, G. (1993). Polychlorinated dibenzo-p-dioxins, dibenzofurans, co-planar PCBs and Mono-Ortho PCBs in urban air. pp.127-130 (organohalogen compounds, volume 12).

Svetnik, V., Liaw, A., Tong, C., Culberson, C., Sheridan, R. P., and Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences.*43, 1947–1958

Tarko, L., and Ivanciuc, O. (2001). QSAR Modeling of the Anticolvulsant Activity of Phylacetanilides with PRECLAV (Property Evaluation by Class Variables).MATCH. *Communications in Mathematical and in Computer Chemistry.* 44, 201–214

Todeschini, R., and Consonni, V. Handbook of Molecular Descriptors; Wiley-VCH: Weinheim, Germany, 2000; 668 pp

Todeschini, R., Consonni, V., and Gramatica P. (2009) Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R (eds.) Comprehensive Chemometrics, volume 4, pp. 129-172 Oxford: Elsevier.

Todeschini, R., Lasagni, M., and Marengo, E. (1994). New Molecular Descriptors for 2D- and 3D-Structures. *Journal of Chemometrics.* 8, 263–273

Todeschini, R.., Moro, G., Boggia, R., Bonati, L., Cosentino, U., Lasagni, M., Pitea, D. (1997). Modeling and Prediction of Molecular Properties. Theory of Grid-Weighted Holistic Invariant Molecular (G-WHIM) Descriptors. *Chemometrics and Intelligent Laboratory* Systems. 36, 65–73

Tropsha, A., Gramatica, P., Gombar, V. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR & Combinatorial Science*, 22, 69–77.

Tudoran M.A Tudoran and M.V Putz (2012). Polycyclic Aromatic hydrocarbon: from cerebro in Silico Eco-toxicity fate. Volume 57(71).

Tuppurainen, K. (1999). EEVA (Electronic Eigenvalue): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies.SAR QSAR *Environ. Res.*1999,10, 39–46.

Tysklind, M.; Lundgren, K.; Rappe, C.; Eriksson, L.; Jonsson, L.; Sjöström, M.; Ahlborg, U.

University Press of Kentucky: Lexington, 2001; 399.

USEPA, Health Assessment Document for 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) and Related Compounds. EPA/600/ Bp-92/001c Estimating Exposure to Dioxin-Like Compounds, EPA/600/6-88/005Cb, Office of Research and Development, Washington, DC, 1994

Van de Berg, M.., De Jogh J., and Poinger, H (1994). Crit. Rev. Toxicology, 24, 1.

Vries, M. D., Kwakkel, R. P., &Kijlstra, A. (2006). Dioxins in organic eggs: a review. NJAS-Wageningen. *Journal of Life Sciences*, 54(2), 207-221.

Wagner, J., Green, A. (1993). "Correlation of chlorinated organic compound emissions from incineration with chlorinated organic input". Chemosphere. 26 (11): 2039–2054.

Walker M.K, &Peterson R.E., (1991). Potencies of plychlorinated dibenzo-p-dioxin, dibenzofuran, and biphenyl congeners, relative to 2,3,7,8-terachlorodibenzo-p-dioxin, for producing early life state mortality in rainbow trout (on corhynchus mykiss). Aq Toxicol 21:219-238.

Walker M.K, Peterson RE (1991). Potencies of plychlorinated dibenzo-p-dioxin, dibenzofuran, and biphenyl congeners, relative to 2,3,7,8-terachlorodibenzo-p-dioxin, for producing early life state mortality in rainbow trout (on corhynchus mykiss). Aq Toxicol 21:219-238.

Walker, J.D Joworska, J. comber M.H., I, Schultz T.W and Deardn J.C (2003). Guidelines for developing and using quantitative structure activity relationships. Environmental toxicological and chemistry 22, 165-1665.

Waller, C. L., and McKinney, J. D. (1992). Application of Genetic function approximation to Quantitative structure Activity Relationships and Quantitative Property Relationships. *Journal of Medicinal Chemistry.*1992, 35, 3660.

Wang and Fang, (2009). Chlorine contribution to quantitative structural-activity relationship models of disinfection by products quantum chemical descriptor and toxicities. Fiu Electronic Thesis and Dissertations. Paper 174:32-39.

Weber et al (2008). Dioxin and Pop contaminated sites contemporary and future relevance and challenges, environmental science and pollution research 15:363-393.

Weber, Kulkarni, Clark (2008). Dioxin and POP-contaminated sites-contemporary future relavance and challenges. Enviromental Science and Pollution Research. 15:363-393

Weber, R., Gaus, C., Tysklind, M., Johnston, P., Forter, M., Hollert H. (2008) Doxin and POP-contaminated sites – contemporary and future relevance and challenges, *Envir. Sci. and Pollution Research*, 2008; 15, 363-393.

Wiener, H. (1947). Structural Determination of Paraffin Boiling Points.J. Am. Chem. Soc.1947,69, 17–20.

Williams D.A Lemke T.L Foye's, Lippin Cott Williams and Wilkins, Baltimore (2002).Principles of medicinal chemistry 5th Ed, 81.

Wold, S., Eriksson, L. (1995). In Chemometrics Methods in Molecular Desisign; van de Waterbeemd, H., Ed VCH, Weinheim, Germany, 1995, pp. 309-318.

Worth, A. P., Bassan, A., Fabjan, E., Gallegos Saliner, A., Netzeva, T. I., Patlewicz, G., Pavan, M., and Tsakovska, I. (2008). The Use of Computational Methods in the Grouping and Assessment of Chemicals – Preliminary Investigations. *Eur. Tech. Rep.*2008,in press

Wu, H., Li, F., Zhao, J., Liu, X., and Zhang, L. (2011). PAPER TITTLE. *Environ. Toxico. pharm.* 2011, 32, 478-485.

Xu, M., Zhang, A., Han, S., and Wang, L.S. (2002). Studies of 3D-Quantitative Structure–Activity Relationships on a Set of Nitroaromatic Compounds: CoMFA, Advanced CoMFA and CoMSIA. *Chemosphere* 2002, 48, 707–715

ZhengG.JLeung A.O., Jiao, L .P., Wong, M .H.,(2008). Polychlorinated dibenzo-p-dioxins and dibenzofurans, pollutions in china. Sources environmental levels and potential human health impacts. Enviromental International, 34:1050-1061.

.

Zupan, J., Novic, M., Gasteiger, J. (1995). Neural Networks with Counter-Propagation Learning Strategy Used for Modelling. Chemom. Intell. Lab. Syst. 27, 175–187

.