

Exploring QSPR modeling for adsorption of organic chemicals by carbon nanotubes (CNTs)

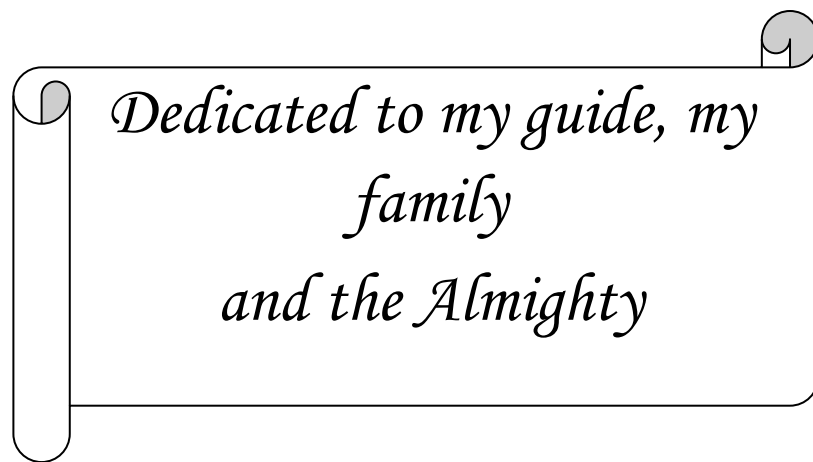


Thesis submitted in partial fulfillment for the requirements of the Degree of
MASTER OF PHARMACY
Faculty of Engineering and Technology

Thesis submitted by
SULEKHA GHOSH
B. PHARM
Registration No. **140843** of **2017-2018**
Examination Roll. No. **M4PHA19001**

Under the Guidance of
DR. KUNAL ROY
Professor

Drug Theoretics & Cheminformatics Laboratory
Division of Medicinal and Pharmaceutical Chemistry
Department of Pharmaceutical Technology
Jadavpur University
Kolkata – 700 032
India
2019

A decorative scroll with a black outline and a light gray shadow. The scroll is unrolled on the left and right sides, with the top and bottom edges curved. The text is centered within the scroll.

*Dedicated to my guide, my
family
and the Almighty*

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research as part of my work on “Exploring QSPR modeling for adsorption of organic chemicals by carbon nanotubes (CNTs)”

All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

NAME: SULEKHA GHOSH

EXAMINATION ROLL NUMBER: M4PHA19001

REGISTRATION NUMBER: 140843 of 2017-2018

THESIS TITLE: “Exploring QSPR modeling for adsorption of organic chemicals by carbon nanotubes (CNTs)”

SIGNATURE WITH DATE:

CERTIFICATE
Department of Pharmaceutical Technology
Jadavpur University
Kolkata - 700 032

This is to certify that Ms. Sulekha Ghosh, B. Pharm. (MAKAUT), has carried out the research work on the subject entitled “Exploring QSPR modeling for adsorption of organic chemicals by carbon nanotubes (CNTs)” under my supervision in Drug Theoretics & Cheminformatics Laboratory in the Department of Pharmaceutical Technology of this university. She has incorporated her findings into this thesis of the same title, being submitted by her, in partial fulfillment of the requirements for the degree of Master of Pharmacy of Jadavpur University. She has carried out this research work independently and with proper care and attention to my entire satisfaction.

Dr. Kunal Roy

Professor,
Drug Theoretics and Cheminformatics
Laboratory,
Department of Pharmaceutical Technology,
Jadavpur University,
Kolkata-700 032

(Prof. Pulok Kumar Mukherjee)
Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

(Prof. Chiranjib Bhattacharjee)
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata

Acknowledgements

*I deem it a pleasure and privilege to work under the guidance of **Dr. Kunal Roy**, Professor, Drug Theoretics & Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata-32. I express my deep gratitude and regards to my revered mentor for suggesting the subject of this thesis and rendering me his thoughtful suggestions and rational approaches to this thesis work. I am greatly indebted to Dr. Kunal Roy for his valuable guidance throughout the work that enabled me to complete the work. With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, perpetual assistance and co-operation from my seniors Probir Kr. Ojha, Khan Kabiruddin, Priyanka De, Mohsin Khan Pathan , Vinay Kumar and Gopala Krishna. Their constant support and helpful suggestions have tended me to accomplish this work in time. I would like to express my special thanks to my friend Reshma Kumari, Joyita Roy , and my juniors Arnab Seth and Sapna Pandey who all have extended their helping hands and friendly cooperation all through my work.*

I am indeed glad to convey cordial thanks to all my friends specially Sourav Roy, Ajeya Samanta Sushmita Basak and Suparna Ghosh. I am thankful to the authority of Jadavpur University and Head of the Department Prof. Dr. Pulok Kumar Mukherjee for providing all the facilities to carry out this work.

I would like to express my special thanks to Bidhan Chandra Ghosh and Chitrlekha Ghosh for their encouragement, helpful criticism and inspiring suggestions throughout the course of this investigation

A word of thanks to all those people associated with this work directly or indirectly whose names I have been unable to mention here. Finally, I would like to thank my parents Mr. Anajit kumar Ghosh , and Mrs. Moni Ghosh and my nephew Prince for all the love and inspirations without which my dissertation work would remain incomplete.

SULEKHA GHOSH

Examination Roll No: M4PHA19001

Department of Pharmaceutical Technology,

Jadavpur University,

Kolkata-700032

Preface

The work presented in this dissertation is spread over a span of two years. The present work has been investigated through *in silico* studies of quantitative structure–property (QSPR) relationship of selected classes of organic pollutants having defined endpoint (K_{SA} and K) towards single and multi-walled carbon nanotubes. *In silico* techniques constitute an integral part of the high throughput screening (HTS) methodology for the screening of new chemical entities with desirable properties. *In silico* methods are capable of providing information about the physicochemical properties of chemicals and the necessary structural fragments influencing the molecular properties. The use of statistical models to predict biological and physicochemical properties started with linear regression models developed by Hansch in 1960s. Since the appearance of computer-aided structure–activity studies, the term “Quantitative structure-activity relationship (QSAR)” has become one of the most popular techniques in medicinal, environmental and synthetic chemistry. QSARs represent predictive models derived from application of statistical tools correlating biological activity/property of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or property.

Nanotechnology has introduced a new generation of adsorbents like carbon nanotubes (CNTs), which have drawn a widespread attention due to their outstanding ability for the removal of various inorganic and organic pollutants. The goal of this study was to develop regression-based quantitative structure–property relationship (QSPR) models for organic pollutants using only easily computable 2D descriptors to explore the key structural features essential for adsorption to multi-walled and single-walled CNTs.

Using the molecular features as independent variables and molecular property as dependent variable, statistically validated models are developed. The statistically significant models are selected for property prediction of untested molecules, which have similar structural features to the compounds used for development of models. If the model shows good predictive power then the compound with those structural features may show efficient response profile and then only the compound may be subjected for subsequent wet laboratory analysis. Thus, the QSPR models help to reduce the number of compounds to be synthesized and tested.

In this present dissertation, we have developed predictive models for adsorption coefficient using easily computable 2D descriptors. The models developed have showed acceptable statistical significance. The models developed were validated rigorously based on internal and external validation strategies. The following analyses have been performed in this dissertation:

Study 1: Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs).

Study 2: Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs.

The work has been presented in this dissertation under the following sections:

Chapter 1 :	Introduction
Chapter 2 :	Present work
Chapter 3 :	Materials and methods
Chapter 4 :	Results and discussions
Chapter 5 :	Conclusion
Chapter 6 :	References
Appendix :	Reprints

Abbreviations

AD	Applicability Domain	MLR	Multiple linear regression
ANN	Artificial neural networks	MWCNTs	Multi-walled carbon nanotubes
C _e	Liquid phase equilibrium concentration	NIPALS	Non-linear iterative partial least squares
CNTs	Carbon nanotubes	OCs	Organic chemicals
DModx	Distance to model x	OECD	Organisation for Economic Co-operation and Development
EVA	Eigen value	PCA	Principal Component Analysis
ETA	Extended Topochemical Atom	PCR	Principal Component Regression
GETAWAY	GEometry, Topology, and Atom-Weights Assembly	PLS	Partial Least Squares
H ₂ S	Hydrogen Sulfide	PRESS	Predicted residual sum of squares
ICP	Intelligent consensus predictor	q _e	Solid phase equilibrium concentration
kNN	k-nearest neighbors	QSAR	Quantitative structure-activity relationship
K _{SA}	Adsorption coefficient related to Specific surface area	QSPR	Quantitative structure-property relationship
K _∞	Adsorption coefficient	rmsep	Root mean square error in prediction
LOO	Leave-one-out	SWCNTs	Single-walled carbon nanotubes
LR	Linear regression	SEE	Standard error of estimate
LV	Latent variables	US-FDA	United States food and drug administration
MAE	Mean absolute error	WHIM	Weighted holistic invariant molecular descriptor
MLOGP2	Moriguchi octanol water partition coefficient	ZnO	Zinc Oxide

TABLE OF CONTENTS

Chapter	Topics	Page nos.
	Acknowledgement	i
	Preface	ii-iii
	Abbreviations	iv
1.	INRODUCTION	1
1.1	Quantitative structure-activity relationship(QSAR)analysis	2
1.1.1	Objective of QSAR	4
1.1.2	Concept of Descriptor	4
1.1.3	Types of Descriptors	5
1.1.4	Classification of QSAR analysis	7
1.1.5	Application of QSAR Studies	9
1.2	Role of Carbon nanotubes as a nanomaterials in pollution management	10
1.2.1	Chemistry of Carbon nanotubes	11
1.2.2	Types of Carbon nanotubes	12
1.2.3	Properties of Carbon nanotubes	13
1.2.4	Application of Carbon nanotubes	14
1.3	Role of predictive QSAR models on the adsorption of CNTs	16
2.	PRESENT WORK	19
2.1	Study 1:Data set 1	20
2.2	Study 2 :Data set 2	21
3.	METHOD AND MATERIALS	22
3.1	Study 1:Data set 1	22
3.2	Study 2 :Data set	40
3.3	General description of methods applied for developing QSPR models	52
3.3.1	Descriptor calculation	52
3.3.2	Data set division	72
3.3.3	Selection of variables using multilayered strategy	74
3.3.4	Model development	74
3.3.5	Computation of different statistical metrics for assessing model quality	76
3.3.6	Software packages employed	87
3.4	Study wise specific description of methodologies utilized in each study	88
3.4.1	Study1: Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs).88	88
4.	RESULTS AND DISCUSSIONS	93
4.1	Study 1: Predictive Quantitative Structure-Property	93

	Relationship (QSPR) Modeling for Adsorption of Organic Pollutants by Carbon Nanotubes (CNTs)	
4.1.1	The descriptors related to hydrophobic interaction	97
4.1.2	The descriptors related to π - π interaction	100
4.1.3	The descriptors related to hydrogen bonding interaction	105
4.1.4	The descriptors related to electrostatic interaction	106
4.2	Study 2: Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs	106
	Descriptors related to hydrophobic interaction	115
4.2.2	Descriptors related to electrostatic interaction	119
4.2.3	Descriptors related to hydrogen bonding interaction	121
4.2.4	Other modeled descriptors essential for adsorption of hazardous SOC's to SWCNTs	122
5.	CONCLUSION	125
5.1	Predictive Quantitative Structure-Property Relationship (QSPR) Modeling for Adsorption of Organic Pollutants by Carbon Nanotubes (CNTs)	125
5.2	Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs	126
6.	REFERENCE	129
	Appendix-Reprints	-

Chapter 1



INTRODUCTION

1. INTRODUCTION

Chemistry is one of the fundamental natural sciences affecting a wide variety of processes. In the other words, we are controlled by a multitude of chemical processes from birth to death. The main objective of scientific discipline lies in exploration of the systematic knowledge of nature and its application aid the human exertion. Different kinds of chemicals influence a large part of human endeavor spanning from laboratory experiments to industrial processes, including household applications. Hence, considering the importance of chemicals from our daily life to complex industrial operations, it will be important to gather suitable knowledge about the chemicals to efficiently modify the behavior of the chemicals (Roy et. al., 2015).

Quantification of chemistry and incorporation of mathematical algorithms in chemical sciences allows the development of a logical basis to define a chemistry-activity/property/toxicity correlation. The aftermath of chemical interactions producing the pharmacological effects of chemicals can be understood from such analysis. Although it sounds simple, such correlation analysis can be amplified in a very broad way to solve complex problems spanning from prediction of drug action in human body to the assessment of environmental hazard produced by chemicals.

Hazardous effect produced by chemicals has been a serious issue of concern since the past. A cherished goal of chemists therefore lies in designing novel methods to control the harmful effect of hazardous chemicals towards the environment. Nanomaterials are specifically used for pollution management because they contain high surface area and possess high adsorption affinity towards the organic contaminants, and they can be modified in several ways to increase their selectivity towards specific target pollutants (Chen et al., 2007). Among them carbon nanotubes (CNT) have been investigated widely as alternative adsorbents for the organic compounds (OCs) removal from the environment.

The chemistry of compounds provides a wide opportunity to scientists for the design and development of purpose specific and harmless novel strategies. Rational strategies are always acceptable in this regard to minimize the amount of biological assessments and thereby aid the

development of potent analogues employing less resource. Quantitative structure–activity relationship (QSAR) studies present such an opportunity in exploring the encoded chemical information of molecules through the development of predictive mathematical models using selected experimental data (Dearden., 2003).

1.1. Quantitative structure-activity relationship (QSAR) analysis

Development of suitable techniques which allow modification of the chemical features of molecules is very useful not only in the field of chemistry but also in other branches of natural sciences. Quantitative structure-activity relationship (QSAR) modeling is one such technique that allows the interdisciplinary exploration of knowledge on compounds covering the aspects of chemistry, physics, biology, and toxicology (Lowis, 1997). Quantitative structure-activity relationship (QSAR) modeling, originally evolved from physical organic chemistry, has seen wide application in the screening of chemicals for their target property thus helping in the prioritization of experimental testing and providing excellent statistical filtering tools of the structure–activity/property relationships (QSAR/QSPR). QSAR has now evolved as a well-recognized tool for application in chemistry when a biological activity or property or toxicity is the end point of the study for a series of chemicals of certain degree of structural similarity.

QSAR modeling can serve as a primary screening technique before different intensive screening methodologies can be performed, such as *in vivo* *in vitro* toxicity determination and molecular docking. QSAR is directly related to the molecular structures of a chemical which correlate with physicochemical, biological or toxicological properties of molecules using various numerical values associated with experimentally derived parameters, which are known as descriptors. It offers an *in silico* tool for the development of predictive models towards various activity and property endpoints of a series of chemicals using the response data that have been determined through experiments and molecular structure information derived computationally or sometimes from experiments. Once developed and validated, such models may be used for prediction of the response/endpoint(s) for new and untested chemicals and also for obtaining a mechanistic interpretation. The naming of QSAR study depends on type of response or the endpoint used for a modeling and is of three classes, namely quantitative structure–property/activity/toxicity relationship (QSPR/QSAR/QSTR) which are composed of physicochemical property, biological activity, and toxicological data, respectively. QSPR, i.e., quantitative structure–property

relationship modeling covers all the area related to biological, toxicological as well as physicochemical behavior. The term QSAR is used to denote all such study. QSAR is mathematically represented as follows:

$$\text{Biological activity} = f(\text{Chemical attributes}) \quad 1.1$$

Chemical attribute is use here to denote the fundamental information of the chemical which can control the response. The main objective is to develop a mathematical correlation, these attributes are precise quantitative chemical information which are derived from experimental analysis or theoretical algorithm that analyze chemistry of the molecule. Physicochemical properties like melting point, boiling point and surface tension are often explain the behavioral manifestation of the chemical species. Hence, the chemical attributes in Eq. (1.1) is often described in terms of the information obtained from the chemical structure and the physicochemical information usually derived using experimental techniques represent the following expression (Katritzky et al., 2002).

$$\text{Response} = f(\text{chemical structure, physicochemical property}) \quad 1.2$$

When we consider a series of chemical information in presence /absence of physicochemical property, response specific QSAR equation can be expressed in following manner:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_nX_n \quad 1.3$$

In this expression, Y is the dependent variable represent the response i.e., activity/property/toxicity being modeled, X_1, X_2, \dots, X_n are the independent variables denoting different structural features or physicochemical properties in the form of numerical quantities or descriptors and a_1, a_2, \dots, a_n are the contributions/coefficients of individual descriptors to the response, and a_0 is a constant.

1.1.1. Objectives of QSAR

The principal objective of any QSAR model is to develop a rational strategy towards the design of new molecule with desired activity. Behavioral manifestation of chemicals is largely depends on structural and physicochemical parameters. Even a minute change in structure can be a cause to a significant change in its pharmacological/toxicological activity. So chemistry is a crucial factor for behavioral determination of a compound.

The principle objectives of QSAR analysis are:

1. Prediction of new analogs of compounds with better property.
2. Better understanding and exploration of the modes of actions.
3. Optimization of the lead compound with decreased toxicity.
4. Reduction of the cost, time and manpower requirement by developing of more effective compounds using a scientifically less exhaustive approach.

To achieve the aforementioned objectives, it is necessary to have a detailed knowledge on the following aspects:

- (i) Detailed knowledge of the mode of action of the molecules.
- (ii) Various factors controlling the experimental condition of the molecules.
- (iii) A thorough examination of molecular structures and their properties.

Quantitative structure-activity relationship is an interdisciplinary study of chemistry, biology, and statistics. By the prediction of the essential structural requirements needed for obtaining a molecule with optimized activity/toxicity/property, QSAR analysis provides a good platform for the synthesis of relatively lesser number of chemicals with improved activity toxicity/property of interest (Tong et al., 2005).

1.1.2. Concept of Descriptors

The predictor variables used in a QSAR analysis are also known as “descriptors” or molecular descriptors. Molecular descriptors are terms that characterize specific information about a studied molecule. They are the “numerical values associated with the chemical constitution for correlation of chemical structure with various physical properties, chemical reactivity, or biological activity (Van de Waterbeemd et al., 1997; Randic, 1997). In other words, the modeled response (activity/property/toxicity) is represented as a function of quantitative values of structural features or properties that are termed as descriptors for a QSAR model.

Cheminformatics methods depend on the generation of chemical reference spaces into which new chemical entities are predictable by the developed QSAR model. The definition of chemical spaces significantly depends on the use of computational descriptors of studied molecular structure, physical or chemical properties, or specific features.

$$\text{Response (activity/property/toxicity)} = f(\text{information in the form of chemical structure or property}) = f(\text{descriptors}) \quad (1.4)$$

The type of descriptors used and the extent to which they can encode the structural features of the molecules that are correlated to the response are critical determinants of the quality of any QSAR model. The descriptors may be physicochemical (hydrophobic, steric, or electronic), structural (based on frequency of occurrence of a substructure), topological, electronic (based on molecular orbital calculations), geometric (based on a molecular surface area calculation), or simple indicator parameters (dummy variables).

An ideal descriptor should possess the following features for the construction of a reliable QSAR model:

1. A descriptor should be relevant to a broad class of compounds.
2. A descriptor must be correlated with the studied biological responses while illustrating insignificant correlation with other descriptors.
3. Calculation of the descriptor should be fast and independent of experimental properties.
4. A descriptor should produce different values for structurally dissimilar molecules, even if the structural differences are little.
5. A descriptor should possess physical interpretability to determine the query features for the studied compounds.

1.1.3. Types of Descriptors

Descriptors can be classified in multiple ways depending on the method of their computation or determination: physicochemical (hydrophobic, steric, or electronic), structural (frequency of occurrence of a substructure), topological, electronic (molecular orbital calculations), geometric (molecular surface area calculation), or simple indicator parameters (dummy variables). In a broader perspective, descriptors (specifically, physicochemical descriptors) can be classified into two major groups: (1) substituent constants and (2) whole molecular descriptors (Todeschini and

Consonni, 2008; Livingstone, 2000). Substituent constants are basically physicochemical descriptors which are designed on the basis of factors, which govern the physicochemical properties of chemical entities. Whole molecular descriptors are expansions of the substituent constant approach, but many of them are also derived from experimental approaches. In QSAR study most commonly used whole molecular descriptors are octanol-water partition coefficient, acidic dissociation constant (pKa), and van der Waals volume (Vw).

The descriptors can be derived experimentally as well as theoretically. The theoretical descriptors are more preferable because chances of error are less. The brief description of theoretical and experimental descriptor is as follows,

Experimental descriptors: Experimental descriptor generally represents various physicochemical properties. During experimental determination care should be taken to decrease the chances of error. These are basically the “whole molecular descriptors” forming the inherent chemical nature of the molecules. Examples of different experimental descriptors are octanol-water partition coefficient i.e., logKo/w, melting point, boiling point, pKa values, rate of reaction, molar refractivity etc. In the field of QSAR studies partition coefficient is one of the most widely used descriptor.

Theoretical descriptors: Theoretical descriptors are computationally determined chemical features which are calculated from mathematical algorithm. Such descriptors are computed for the whole molecule and also for the predefined fragments. In QSAR modeling paradigm, development of different theoretical descriptors introduce a momentum. In early days mathematical graphs are used in solving the problems associated with chemical issues. Graph theory was also used for solving different mathematical problems. In 1736 Euler has first introduced graph theory for solving “Königsberg Bridge problem”. In 1959 William Cullen (Crosland, 1959) first used the concept of chemical graph and prepare affinity diagram for determining chemical forces. Later William Higgins also used such diagram for designating the forces which existing between two atoms. In nineteenth century, Dalton used “ball and stick” models for molecular representation where atoms are represent as ball and bonding forces are resemble as a stick. Cayley is known for denoting the alkane using kenograms (tree graphs) while different conditions required for the development of a chemical graph was proposed by Sylvester. The presentation of structural formula of covalent compounds using graphs is termed as ‘molecular graphs’ or ‘constitutional graphs’. Here, atoms are known as ‘vertices’ and bonds

as ‘edges’. Other forms of graphs in the field of chemistry include reaction graph, synthon graph etc. The graph theory finds its importance during the development of the concept of isomerism which allows the existence of different constitutional isomers. With progress of graphical representation the concept of topology came in consideration. Topology is basically represented as the minimum distance between two connecting objects. Encoding of chemical information in terms of numbers is determined through the concept of matrix. The numbers representing matrix elements were derived from graph theoretical connectivity of chemical structures which upon treatment on a suitable algorithmic operator yielded descriptors. The topological indices, i.e., the descriptors derived using topological information of molecules can encode essential chemical information defining size, shape, branchedness, symmetry, cyclicality etc. In 1947 Wiener and Platt started the journey of theoretical descriptor by introducing Wiener index and Platt number respectively by developing predictive QSAR models on boiling point of hydrocarbons. The Wiener index and Platt number both are topological descriptors which are derived from graph theoretical approach. Randić, Balaban, Kier and Hall, Gutman, Trinajstić, Zagreb contributes a valuable research on the graph theoretical metrics and topological descriptors which gave a momentum to the research of QSAR modeling. In 1970s quantum chemical descriptor were first used to develop predictive QSAR model. The names of Pauling, Coulson, Sanderson, Fukui and Mulliken can be mentioned in this context (Todeschini and Consonni, 2009) for their notable contribution in exploring electronic distribution, charge, chemical bonding etc. In the mid-1980s three dimensional chemical features in the topological formalism has been came in light. This will open the new path for various three dimensional spatial descriptors as well as for 3D-QSAR. Shadow indices, charged partial surface area descriptors, weighted holistic invariant molecular (WHIM) descriptors, gravitational indices, Eigen Value (EVA) descriptors, 3D-MoRSE descriptors, EEVA descriptors, and GEometry, Topology, and Atom-Weights Assembly (GETAWAY) descriptors etc. represent the 3D-descriptors.

1.1.4. Classification of QSAR analysis

QSAR analysis is basically categorized based on their nature of endpoint. QSAR analysis can be a QSAR/QSPR/QSTR model depending on the nature of response variable to be biological activity, physicochemical property, and toxicity respectively. QSAR can be categorized based on dimensionality of the predictor variables such as 0D, 1D, 2D, 3D etc. Classifications of different QSAR techniques based on dimensionality are depicted in **Fig.1**. Apart from that, QSAR

analysis based on the type of chemicals used for modeling as for example, modeling of nonmaterial is named as quantitative nanostructure-activity relationship (QNAR) analysis. Sometimes QSAR methods are also classified into following two categories, such as Linear methods [Linear regression (LR), multiple linear regression (MLR), partial least squares (PLS), and principal component analysis/regression (PCA/ PCR)] and Nonlinear methods [Artificial neural networks (ANN), k-nearest neighbors (kNN), and Bayesian neural nets].

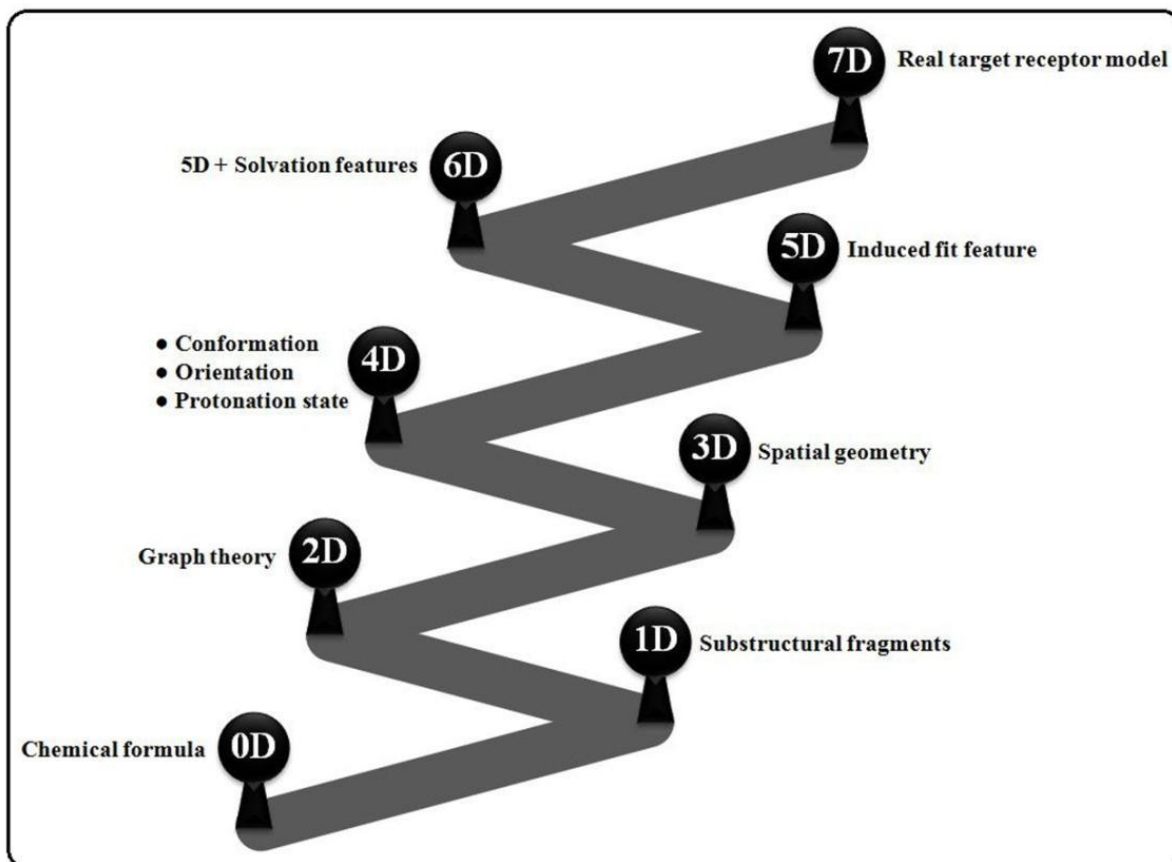


Fig. 1. Classification of QSAR analysis based on dimensionality of the predictor variables (Das, 2016).

1.1.5. Application of QSAR studies

Chemicals are the essential part of human necessity from laboratory to industrial processes as well as household usage. The application of QSAR modeling can be broadly viewed less than three major areas namely biological activity, predictive toxicity, and physicochemical property. Modeling of biological activity includes design and discovery of drugs for recovery in various diseases and disorders like microbial infection, viral infection, cardiovascular disorder, hepatic

damage, cancer, disease of the central nervous system, cholinergic system, adrenergic system etc. The toxicological modeling deals with all type of toxic chemicals including drugs, pharmaceuticals and industrial chemicals. QSAR studies also allows assessing chemical hazards towards the living ecosystem i.e., environmental toxicity. , QSAR has also been found to be beneficial in agricultural sciences where toxicity potential of chemicals is an essential feature, e.g., fungicidal and pesticidal activity. Modeling of property of chemicals encircles a wide field of industrial process chemicals to physicochemical properties of drugs. Hence, we can see that QSAR modeling can be a very good option to predict chemical response using limited resources in any prospective discipline. Needless to say such studies are not only helpful to the users, but also beneficial in making crucial regulatory decisions like biological testing using animal models. In the recent years, QSAR modeling have been observed to be useful for modeling response of novel chemicals like ionic liquids, nanoparticles etc., increasing the area of application manifold. Application of the QSAR technique in combination with other in silico methods has been very fruitful in the drug-discovery paradigm, and some representative examples of such designed drug molecules which were later approved by the US Food and Drug Administration (US-FDA) as drug entities are presented in **Fig. 2**.

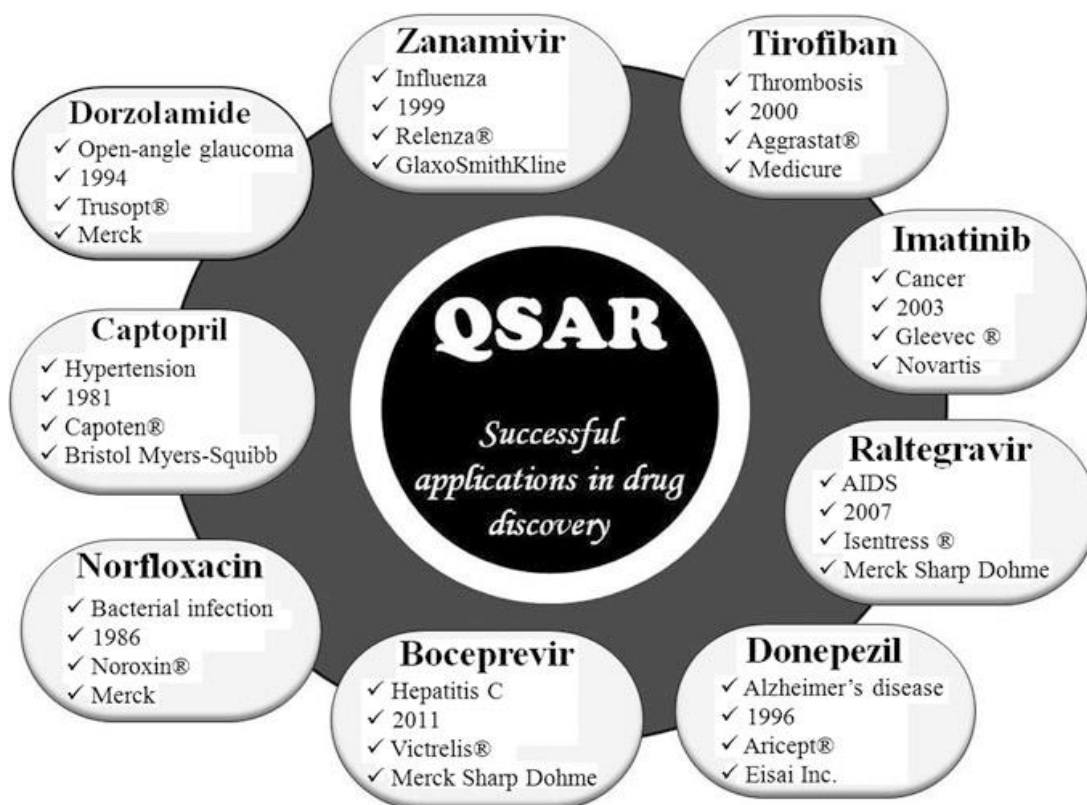


Fig. 2. Examples of drugs designed and developed using different in silico techniques including QSAR modeling analysis and later approved by the US Food and Drug Administration (US-FDA) as drug entities (Roy et al., 2015).

1.2. Role of Carbon nanotubes as a nanomaterial in Pollution management

Fast industrialization and development of agricultural production make the water resources full of heavy metals. Waste water discharge from domestic, industrial or agricultural sources produces a wide range of contaminants and has drawn major concern worldwide since they reduce the quality of water. The presence of heavy metals in water resources is a serious issue for both environment and ecosystem. To make the water as well as environment free from all kind of toxic contaminants several traditional techniques such as reverse osmosis, chemical precipitation, filtration, ion exchange, coagulation and adsorption are used. (Krishnan et al., 1998). Among all of this process adsorption is an efficient technique because it is low cost process and easy to perform. Nanomaterials are one of the most efficient and actively used adsorbent because of their high surface area and ease of synthesis. Carbon nanotubes are such

nanomaterials possess high surface area with light mass density and shows interaction towards toxic environmental contaminants. (Yu et al., 2000; Kang et al., 2006). The contaminants found in waste water are heavy metal ions which are non biodegradable, highly toxic and carcinogenic causing accumulative poisoning, cancer and nervous system damage. CNTs show wider adsorption affinity for such heavy metals as well as environmental pollutants including organic materials and radioactive elements (Ong et al., 2010).

Carbon nanotubes were first discovered by Iijima in 1991 which was a Multi-Walled carbon nanotubes (Iijima, 1991). Benning and co-workers discovered C60 fullerene and single wall carbon nanotubes in 1992. (Benning et al., 1992). CNTs are one of the most studied nanomaterials as they are building blocks of nanotechnology and have gained a great deal of attention in research field (Wang et al., 2017). Due to its extraordinary physical, chemical and electronic properties, a wide variety of applications has been proposed in different fields like nanotechnology, electronic, optics and in the fields of science and technology. CNTs have a great potential for application in various environmental field such as waste water treatment, air pollution monitoring, biotechnologies, renewable energy technologies, super capacitors and green nanocomposites (Ong et al., 2010).

1.2.1. Chemistry of Carbon nanotubes

CNTs are allotropes of carbon, seamless cylinder or tube shaped material having a diameter measuring on the nanometer scale. Carbon nanotubes consist of graphene sheets, the edges of the sheet joint together to form a seamless cylinder. Graphene is known as 2D single layer of graphite. Carbon atoms are hexagonally arranged in lattice structure of graphene as shown in **Fig. 3**. The $2s$ and $2p$ atomic orbital of isolated single carbon atom consist of four valence electrons. In case of graphene, the $2s$, $2p_x$ and $2p_y$ atomic orbitals are hybridized and converted into three sp^2 orbitals. Graphene is stronger than diamond because a sp^2 bond in graphene is stronger than sp^3 bonds in diamond (Kaushik and Majumder, 2015).

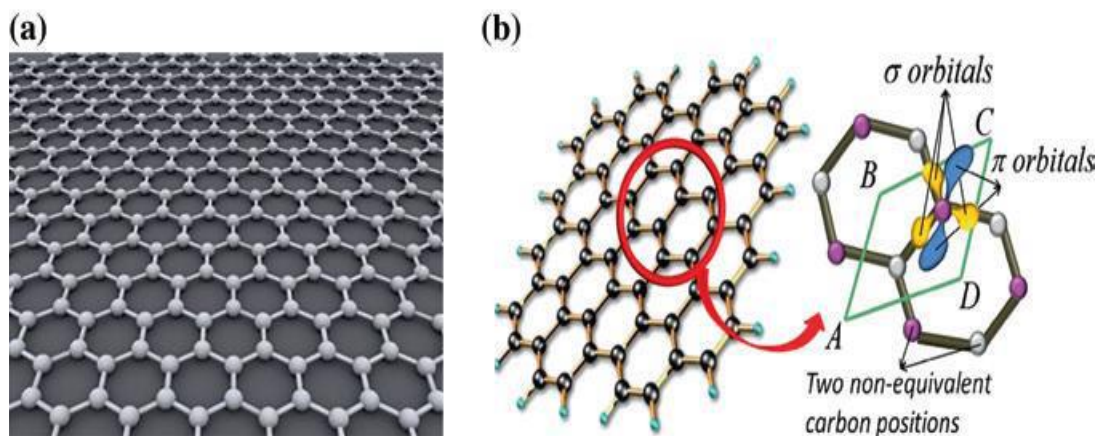


Fig. 3. Basic (a) hexagonal and (b) orbital structure of graphene (Kaushik and Majumder, 2015).

1.2.2. Types of CNTs

CNTs can be divided into three categories on the basis of the number of tubes present in the CNTs. These are described below.

(a) Single-walled CNTs: Single-walled CNTs (SWCNTs) are made of a single graphene sheet rolled upon itself with a diameter of 1–2 nm. The SWNTs were reported to exist in bundles, to have fullerene-like hemispheric caps formed at the end of the tubes (Liew et al., 2016), and to have diameters ranging from 0.7 to 1.6 nm. Compared with carbon nanotubes (CNTs) having multiple walls, SWNTs were predicted to exhibit superior properties due to their strong one-dimensionality and the crystalline perfection of their structure.

(b) Double-walled CNTs: These nanotubes are made of two concentric carbon nanotubes in which the outer tube encloses the inner tube.

(c) Multi-walled CNTs: MWCNTs consist of multiple layers of graphene rolled upon itself with diameters ranging from 2 to 50 nm depending on the number of graphene tubes. These tubes have an approximate inter-layer distance of 0.34 nm. Multi-walled carbon nanotubes (MWCNTs) consist of multiple rolled layers (concentric tubes) of graphene. There are two models that can be used to describe the structures of multi-walled nanotubes. According to *Russian Doll* model, sheets of graphite are arranged in concentric cylinders, e.g., a (0, 8) single-walled carbon nanotube (SWCNT) within a larger (0,17) single-walled nanotube. Apart from that another model known as *Parchment* model stated that, a single sheet of graphite is rolled in around itself, resembling a scroll of parchment or a rolled newspaper. The interlayer distance in

multi-walled nanotubes is close to the distance between graphene layers in graphite, approximately 3.4 Å. The Russian Doll structure is observed more commonly. Its individual shells can be described as SWNTs, which can be metallic or semiconducting. Because of statistical probability and restrictions on the relative diameters of the individual tubes, one of the shells, and thus the whole MWNT, is usually a zero-gap metal (Das, 2013).

1.2.2. Properties of Carbon nanotubes

CNTs reportedly have extremely high surface areas, large aspect ratios, and remarkably high mechanical strength. The tensile strength of CNTs is 100 times greater than that of steel, and the electrical and thermal conductivities approach those of copper (Ebbesen et al., 1996). CNTs are good incorporating agents due to their unique electrical, mechanical and thermal properties (Khalid and Ibrahim, 2013).

(a) Electronic nature of CNTs: CNTs show good electrical properties in chiral forms. If we consider the bonding of CNTs carbon atoms are arranged in a hexagonal lattice, each carbon atom is covalently bonded to three neighbor carbons via sp^2 molecular orbitals. Thus, the fourth valence electron remains free in each unit, and these free electrons are delocalized over all atoms and contribute to the electrical nature of CNTs. Thus, CNTs can be conducting or semi-conducting types depending on the type of chirality (Hahm et al., 2011; Saito et al., 1992). SWNTs, due to the ballistic nature of electron transport, can be described as quantum wires. On the other hand, transport in MWNTs is found to be fairly diffusive or quasi-ballistic. CNTs, due to their electronic nature, can be used in transistors and other switching applications in advanced electronics. The most recent application of nanotubes was as an emitter.

(b) Mechanical properties of CNTs: CNTs have excellent potential as they are the stiffest and toughest structure ever synthesized by scientists. The literature suggests that CNTs are very strong materials, especially in the axial direction. The Young's modulus ranges from 270 to 950 GPa, while the tensile strength is also very high, in the range of 11–63 GPa. Falvo et al. (1997) observed that MWNTs could be bent at sharp angles without undergoing any structural fracturing. Sinnot et al. (1998) also done theoretical work on the mechanical properties of CNTs and found that SWCNTs could exhibit a Young's modulus as high as that of diamond.

(c) *Thermal properties of CNTs:* The incorporation of pristine and functionalized nanotubes to different materials can double the thermal conductivity for a loading of only 1%, showing that nanotubes composite materials may be useful for thermal management applications in industries. Kim et al. measured the thermal conductivity of individual MWNTs and found it to be 3,000 W/K (higher than that of graphite) at room temperature. Beside this they also determined that the value is two orders higher than the magnitude those obtained for bulk MWNTs. A similar study was carried for SWCNTs, with this result being greater than 200 W/m K for SWNTs (Yu et al., 2005). Thermal properties are depends on the atomic arrangement, the diameter and length of the tubes, the number of structural defects and the morphology, as well as the presence of impurities in the CNTs (Kasuya et al., 1996).

(d) *Adsorption property of CNTs:* The outer surface of CNTs provides evenly distributed hydrophobic sites creating strong interaction with organic chemicals. The main adsorption mechanism of CNTs include π - π interactions (between π systems on CNT surfaces and organic molecules with C=C double bonds), hydrogen bonds (because of the functional groups on CNT surfaces), and electrostatic interactions (because of the charged CNT surface). The adsorption phenomena highly depends on the physical properties of carbon nanotubes, types of functional group present on it and the morphology of CNTs. The adsorption mechanism also depends on the molecular morphology and type of functional group present in organic chemicals (OCs). On the other hand CNTs are effective adsorbents for organic chemicals in solid phase extraction and water treatment as compared to C18 (Liu et al., 2004) and activated carbon (Su and Lu, 2007; Wang et al., 2007). Thus, adsorption data of various OCs by CNTs may vary with different factors and requires experimental measurement for proper application (Pan and Xing, 2008).

1.2.3. Application of carbon nanotubes

Nanotechnology is one of the latest and the most developed technologies, presenting many advantages and benefits for new materials with significantly improved properties (Kaushik and Majumder, 2015). Nanotechnology can be used in different applications in various fields, including nano-medicine, energy, the environment, and in sensors (NANOSAFE 2008. Available from: <http://www.nanosafe2008.org>). Although the fields of nanotechnology are vast and new materials come into use regularly, the potential of CNTs is most promising. Since their discovery by Iijima (Iijima, 1991) in 1991, CNTs are the most rapidly growing nanomaterials in the field of nanotechnology due to their various applications. Many investigators and researchers have

dedicated much effort to the creation of novel properties and to expanding the number of novel applications in diverse fields, from materials science, medicine, electronics and energy storage, with many studies focusing on nanotechnology and the use of CNTs as fillers (Helland et al., 2007). More attractive applications of CNTs can be achieved through the use of CNTs for applications that require conductivity and a high absorption capacity and for the creation of high-strength composites, fuel cells, energy conversion devices, field-emission devices, hydrogen storage devices, and semiconductor devices (Baughman et al., 2002; Cao et al., 2001). Wastewater treatment by CNTs is also a rapidly growing field for those who are interested in adsorption studies. The major problem associated with CNTs is their high cost and nonrenewable characteristic. At present, special efforts are in progress to develop certain preparation methods for CNTs which minimize their cost. Some of the very important and promising applications of CNTs are discussed below in detail:

(a) Air pollution filter: CNTs have adsorption capacity and large specific area therefore used as air filter material. CNT membranes can successfully filter carbon dioxide from emissions of different factories and industries.

(b) Water filter: CNT membranes also used for water filtration. It can be used to make the distillation process more convenient and economical. These thin tubes block the large particles and allow the smaller one to pass through CNTs.

(c) Chemical Nanowires: The CNTs finds their applications in nanowire manufacturing using materials such as gold, zinc oxide, gallium, arsenide, etc. The gold based CNT nanowires are used for hydrogen sulfide (H₂S) detection. The zinc oxide (ZnO) based CNT nanowires can be used for light emitting devices.

(d) Sensors: CNT based sensors can be used for detection of temperature, air pressure, chemical gases (such as carbon monoxide, ammonia), molecular pressure, strain, etc. The working principle of these sensors is mainly dependent on the generation of current/voltage. The electric current is generated by the flow of free charged carrier induced on any material.

(e) Loudspeaker: CNTs also finds their applications in loudspeakers manufacturing. Such a loudspeaker is able to produce sound having frequency similar to the sound of lightning producing thunder.

1.3. Role of predictive QSAR models on the adsorption of CNTs

As we can see that CNTs are associated with considerable adsorption affinity, explicit assessment of environmental pollutants (organic materials, heavy metal ions and radioactive elements) is necessary to evaluate the adsorption property of both SWCNTs and MWCNTs. However, considering a sufficient number of such chemicals (pesticides, herbicides and fungicides) synthesized in factories and industries, it will be impracticable to perform an exhaustive testing of chemical hazard. Thus, alternative strategies using limited experimental data can be of much use. The predictive QSAR modeling paradigm investigates the chemical features of the compounds responsible for their high adsorption towards CNTs. Apart from that, this work provides an understanding of the important structural requirements or essential molecular properties and the requisite features of molecules that is important to increase or decrease the adsorption of organic contaminants. The developed models could be useful as preliminary support tools for the identification and prioritization of new potential organic pollutants among already existing chemicals as well as “screening prior to synthesis” procedures to avoid the production, and consequent release into the environment, of new organic pollutants. The models provide an important guidance for the chemist to increase the efficient application of CNTs which may be useful for reducing the environmental pollution. Recent studies have reported predictive QSAR models on various physicochemical properties of organic chemicals towards CNTs. Modeling physicochemical properties enables the design and development of purpose specific efficient analogues, while models property response allows the user to capture specific information on the adsorption coefficient. However, considering the scope of this dissertation we would like to present an account on some of the representative published QSAR models on adsorption of chemicals onto CNTs.

Hassanzadeh et al., (2015) developed a QSPR approach based on whole space GA-RBFN (wsGA-RBFN) and applied to predict the adsorption coefficients ($\log k$), of 40 small molecules on the surface of multi-walled carbon nanotubes (MWCNTs). In their investigation the authors have used, a combination of RBFN and GA for QSPR studies of adsorption of OCs on MWCNTs surface. RBFN is used to construct QSPR model and GA is used to optimize the numerical values of RBFN centers (Hassanzadeh et al., 2015).

Wang et al., (2013) developed 3D-QSPR model for adsorption of aromatic compounds by carbon nanotubes based on physicochemical properties of adsorbed compounds and compared MLR, ANN and support vector machine (SVM) methods (Wang et al., 2013).

Rahimi-Nasrabadi and coworkers (2015) reported a predictive QSPR model using stepwise multiple linear regression (MLR) technique to examine the adsorption property of aromatic chemicals by CNTs. The authors have reported that molar volume and hydrogen bond accepting ability were most influencing factors required for good the adsorption of the aromatic compounds (Rahimi-Nasrabadi et al., 2015).

In another study, Apul et al., used 29 aromatic compounds to develop predictive models with multiple linear regression analysis. They used both QSAR and LSER models for adsorption of organic contaminants by multi-walled carbon nanotubes (MWCNTs). They also stated that, at higher equilibrium concentrations, hydrogen bond donating (A) and hydrogen bond accepting (B) terms play important role on the adsorption coefficient values of the compounds (Apul et al., 2012).

Recently, Ahmadi and Akbari used Monte Carlo method to investigate, quantitative structure–property relationship (QSPR) modelling of adsorption coefficients of 69 aromatic compounds on multi-wall carbon nanotubes (MWCNTs) (Ahmadi and Akbari, 2018). The authors used CORAL software descriptors for the modelling of the surface area normalized adsorption coefficients of small organic compounds on MWCNTs.

Ding et al., used linear solvation energy relationship (LSER) method for prediction of the adsorption coefficient (K) of synthetic organic compounds (SOCs) on single-walled carbon nanotubes (SWCNTs). They used a total of 40 compounds for their study. The results portrayed major contribution Hydrogen bond donating interaction (bB) and cavity formation and dispersion interactions (νV) were controlling the adsorption of SOCs onto SWCNTs (Ding et al., 2015).

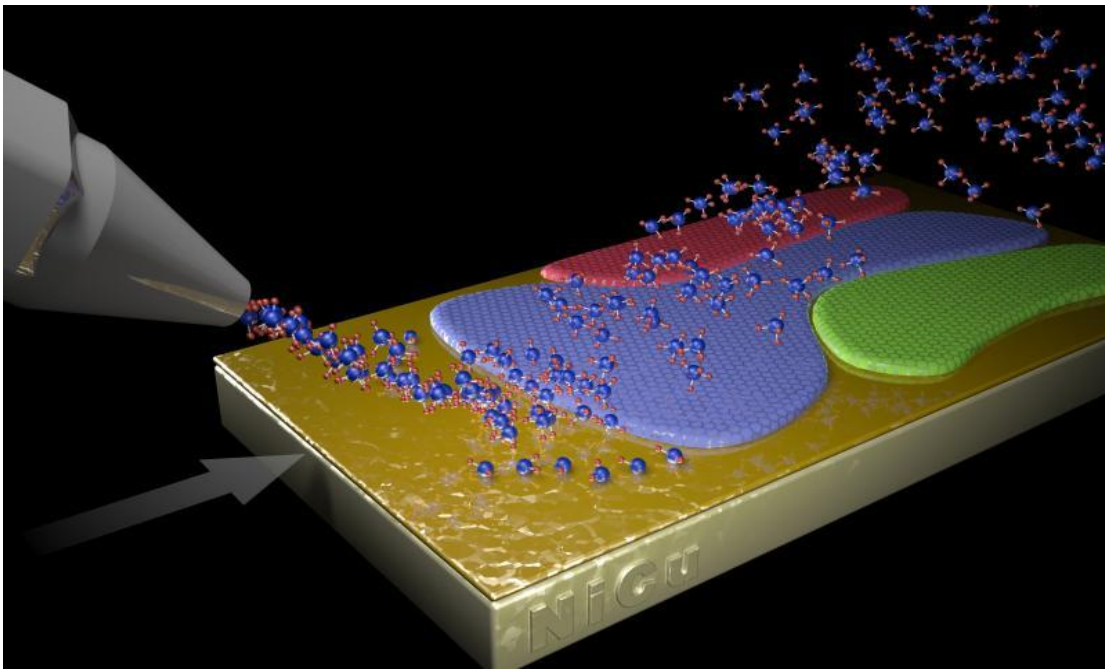
Another study on single-walled carbon nanotubes (SWCNTs) for predicting adsorption equilibrium coefficients ($\log K$) of organic compounds was done by Wang et al., (2019) where the QSAR models were developed by using Multiple linear regression (MLR) and support vector machine (SVM) algorithms. They also stated that, the adsorption of organic compounds toward SWCNTs is mainly determined by van der Waals forces and hydrophobic interactions (Wang et al., 2019).

The Quantitative Ion Character–Activity Relationship (QICAR) method was used by Salahinejada and Zolfonounb for correlating metal ionic characteristics with the maximum adsorption capacity (q_{\max}) of multi-walled carbon for heavy metals. They have used genetic algorithm, and partial least squares (PLS) regression for model development. Furthermore, electronegativity, ionic radius and atomic number of the heavy metal ions plays important role on the adsorption of multi-walled carbon nanotubes (MWCNTs) (Salahinejada and Zolfonounb, 2018).

Liu et al., reported a predictive quantitative structure–activity relationship (QSAR) model to predict the adsorption of 25 simple benzene derivatives on CNTs, and evaluate the interaction between molecule–SWCNTs by density functional theory (DFT) calculations. The authors also stated that, molecules are interacted with SWCNTs through π - π stacking force. The possible mechanism for π - π stacking interaction is generally of two types, i.e., indirectly affecting p–p stacking by altering the electron density of the benzene ring and directly interacting with the nanotube surface (Liu et al., 2014).

In another study, Lata compares the predictive ability of a LSER model for adsorption of OCs by single-walled CNTs (SWCNTs) with a developed QSAR model with using quantum-mechanical descriptors. Further the proposed models were used to predict the adsorption of agrochemicals such as insecticides, pesticides, herbicides, as well as adsorption of endocrine disruptors and biomolecules such as nucleobases and steroid hormones. In addition to that, the mean polarizability of the compounds was reported to be a key quantum-mechanical factor influencing the adsorption property of organic compounds.

Chapter 2



PRESENT WORK

2. PRESENT WORK

The rapid industrial growth leads to an increase in the demand for new inventions and technologies for the benefit of human beings. New chemicals have been introduced for various purposes, which can, however, be a major threat for humans and animals (Latkar and Chakrabarti, 1994). A noticeable amount of organic pollutants is released into the environment via various routes like burning of fossil fuels, wastes from incineration, exhausts from automobiles, agricultural processes and industrial sectors. The disposals of the by-products from the various industries are a challenging job for the environmentalists and for the people of industries. The major problem with pollutants is their effective and safe disposal without affecting the environment further adversely. The organic pollutants (phenols, cresols, alkyl benzene sulfonates, nitro chlorobenzene, chlorinated paraffins, butadiene, synthetic dyes, insecticides, fungicides and pesticides etc.) accumulate in food chain and persist in nature and possess significant threat to the environment (Garg et al., 2007; Randall et al., 1974; Ferner, 2001; Lu et al., 2015).

Recently, nanomaterials are used for pollution management, because they contain high surface area, high adsorption affinity towards the organic contaminants, and they can be modified in several ways to increase their selectivity towards specific target pollutants (Chen et al., 2007). Carbon nanotubes (CNTs) are such type of nanomaterials, which have recently gained special attention from the researchers due to their smaller size, large specific surface area, hollow and layered structure, responsible for their extraordinary adsorption property (Khani and Moradi, 2013; Long and Yang, 2001).

In the present work, diverse organic chemicals with defined adsorption property have been modeled using defined chemometric tool. The compounds modeled in this work are summarized in table 1 and table 2 and derived from the literature (Chayawan, 2016; Ding et al., 2016). We have used only easily predictable 2D descriptors for QSPR model development. For the development of QSAR model, the response values (K_{SA} and K) were expressed in logarithmic scale. The QSAR models developed here provide quantitative insight regarding the essential

structural attributes of the different classes of molecules imparting increased adsorption coefficient value to the molecules. It is known that activity depends solely upon the basic physicochemical properties and structural features of the compounds, hence the use of 2D descriptors helps in finding suitable physicochemical characteristics such as electronegativity, unsaturation, bulk of the compound and hydrogen bonding properties, hydrophobic surface of the molecules, molecular shape and degree of branching etc. Different chemometric tools were employed for determining the correlation of the various types of descriptors and the response. Chemometric tools like stepwise regression, multiple linear regression (MLR), partial least squares (PLS) have been used to establish relation between the various descriptors and the respective activity. The models developed were validated rigorously based on both internal, and external validation strategies. Subsequently, the applicability domain of different models were also performed to check either models are able to predict new set of data of similar class or not.

2.1. Study 1: Dataset 1

Nanotechnology has introduced to the environmentalists a new generation of adsorbents like carbon nanotubes (CNTs) which have drawn a widespread attention due to their outstanding ability for the removal of various inorganic and organic pollutants. The goal of this study was to develop regression-based quantitative structure-property relationship (QSPR) models for organic pollutants and organic solvents using only easily computable 2D descriptors to explore the key structural features essential for adsorption to multi-walled CNTs.

Among the various nanomaterials adsorbents, carbon nanotubes (CNTs) has been investigated deeply as they have a large surface area to volume ratio, inertness towards chemicals, light mass density, porous structure, great physical and chemical properties, small diameter, extraordinary optical and electrical properties, high tensile strength and has efficient affinity towards pollutants. The possibility of the surface modification with different functional groups makes it a good adsorbent (Al-Saidi et al., 2016; Kumar et al., 2014; Mosayebidorcheh and Hatami, 2017) and enhances the reactivity and dispersibility of the carbon nanotubes for environmental protection application.

In this work, QSPR models were developed comprising of a dataset containing 69 organic chemicals with defined end point (adsorption affinity of organic contaminants related to specific

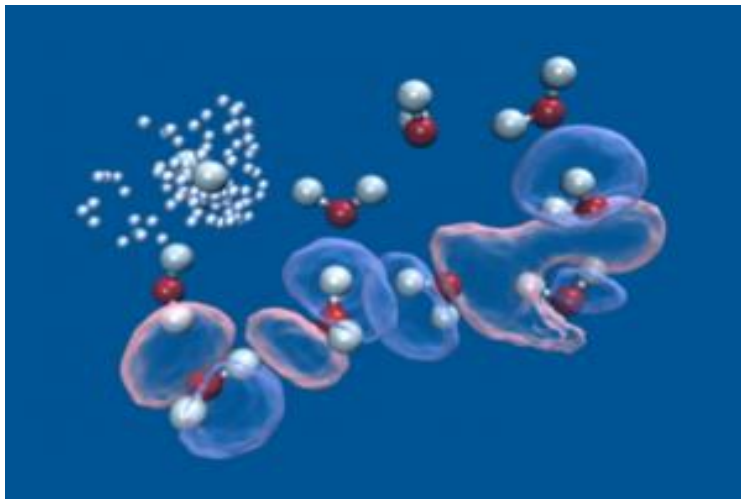
surface area ($\log K_{SA}$) of multi walled carbon nanotubes) to correlate the adsorption affinity ($\log K_{SA}$) in order to determine the structural features which are responsible for adsorption of organic contaminants by multi-walled carbon nanotubes (MWCNTs).

2.2. Study 2: Dataset 2

Introduction of new chemicals for different purposes can be a major threat for humans as well as animals. The use of herbicides, for example, has increased during the last two decades due to the rejuvenation of agriculture. It was reported that 2.5 million ton pesticides were in use worldwide yearly, and the amount is increasing day by day (Pimentel, 1995; Tariq et al., 2007; Carter, 2000). Endocrine disrupting chemicals (EDCs) act like natural hormones and hamper the distribution, as well as metabolic process of natural hormones. EDCs (e.g., ethinyl estradiol) are harmful for the reproductive system of animals and humans (Snyder et al., 2003). Effluents from hospitals or radiological clinics have shown high concentration of antibiotics like sulfamethoxazole and Lincomycin and contrast medium (ipromide), which are responsible for the production of antibiotic resistance bacteria and genes in the aquatic environment (Rand-Weaver et al., 2013; Michael, 2013). Hence, removal of antibiotics as well as pharmaceuticals and contrast medium from water is essential to get purified water.

In this study, we have developed partial least squares (PLS) regression based quantitative structure-property relationship (QSPR) models using adsorption coefficient data of 40 diverse hazardous synthetic organic chemicals (SOCs) onto SWCNTs. The main objectives of our work are: 1) to develop statistically robust and validated QSPR models of hazardous SOC's using 2D descriptors only in order to identify the significant structural features essential for effective adsorption in SWCNTs; 2) to examine the adsorption behavior of diverse synthetic organic chemicals onto SWCNTs; 3) to give a deep insight to understand the mechanisms and factors that are responsible for hazardous SOC's and SWCNTs/functionalized SWCNTs interactions.

Chapter 3



*METHOD AND
MATERIALS*

3. METHOD AND MATERIALS

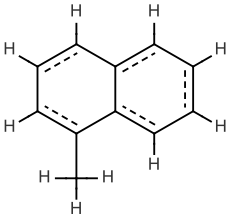
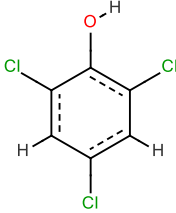
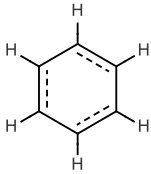
The objective of the present dissertation is to develop or evaluate a clear methodology for the development of a predictive and robust QSPR model by using two dimensional dragon and PaDEL descriptors. In this section, we have described here the details of datasets comprising the structures along with their adsorption coefficient data in logarithmic scale. The section has been divided in the following parts:

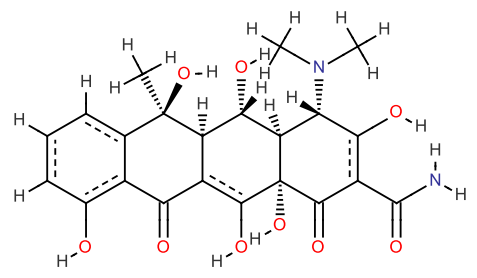
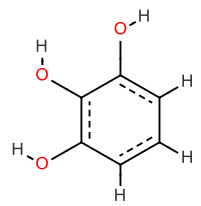
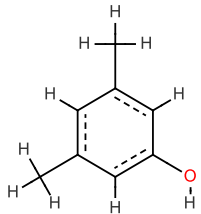
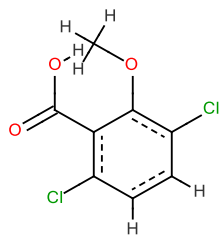
- Details of datasets consisting chemical structures along with their adsorption coefficient data in logarithmic scale.
- General description of methods applied for developing QSPR models.
- Study wise specific description of methodologies utilized in each study.

3.1. Study 1: Dataset 1

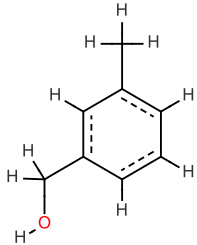
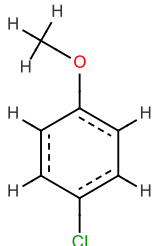
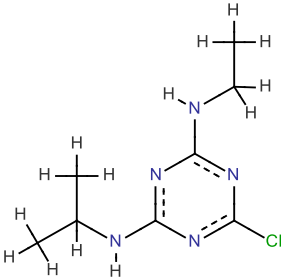
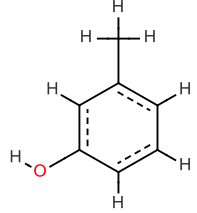
We have developed a model QSPR model, using a data set for diverse organic contaminants with adsorption coefficient ($\log K_{SA}$) of carbon nanotubes reported in the literature (Chayawan, 2016). The dataset involves the adsorption affinity of 69 organic contaminants related to the specific surface area (K_{SA}) of multi-walled carbon nanotubes (MWCNTs). We have not excluded any compound of individual data sets in our modeling analysis. The endpoint values were taken in the logarithmic scale for the modeling purposes. The data set mainly involve adsorption data for synthetic organic compounds like pyrene, naphthalene, phenol, benzene, aniline, benzoate, chloroanisole, alcohol, acetophenone, isophoron, phenanthrene dicamba, atrazine, carbamazepine, pyrimidinone, acetamide, piperidine, propionitrile, acrylic acid, thiodiethanol, ethanolamine, cyclopentanone, acetone and ethylene glycol derivatives. K_{SA} is adsorption coefficients that can be obtained from isotherm data. K_{∞} is the ratio of q_e and C_e (solid and liquid phase equilibrium concentrations, respectively, at infinite dilution conditions with an average of 0.2% aqueous solubility). K_{SA} is the normalized value of K_{∞} and the specific surface area of multi-walled carbon nanotubes (MWCNTs). The data set is given in Tables 3.1.

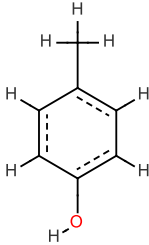
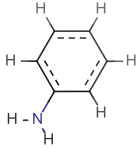
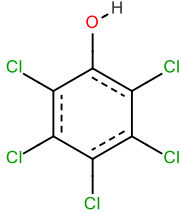
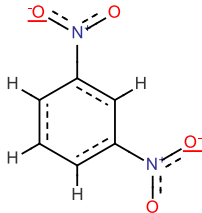
Table 3.1. The chemical name, experimental $\log K_{SA}$ and calculated $\log K_{SA}$ values of the MLR models.

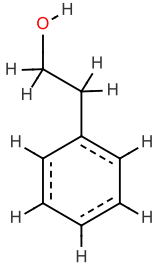
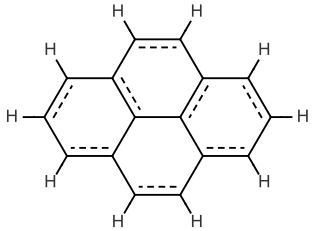
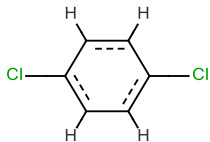
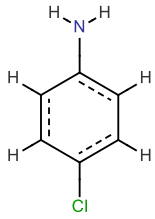
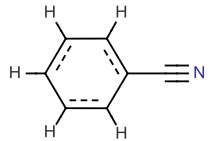
No.	Chemical name	Chemical Structure	$\log K_{SA}$					
			Expt.	Calc.				
				Model N1	Model N2	Model N3	Model N4	Model N5
1*	1-Methylnaphthalene		-0.48	-0.33	-0.41	-0.40	-0.40	-0.36
2	2,4,6-trichlorophenol		-0.81	-0.62	-0.79	-1.28	-1.27	-0.61
3	benzene		-2.47	-2.90	-2.81	-2.52	-2.58	-2.82

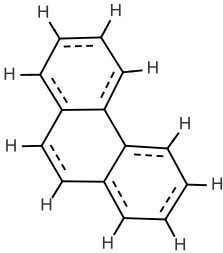
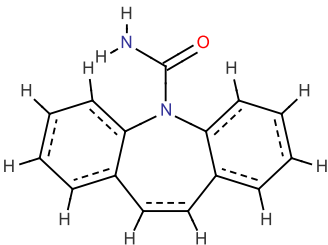
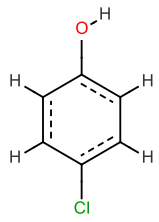
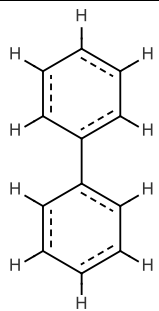
4	oxytetracycline		-0.23	0.12	0.34	0.15	0.15	0.29
5	pyrogallol		-0.98	-1.45	-1.57	-1.37	-1.40	-1.36
6	3,5-dimethylphenol		-1.88	-1.87	-1.88	-2.05	-2.09	-2.13
7	dicamba		-2.64	-2.17	-1.98	-1.72	-1.72	-2.59

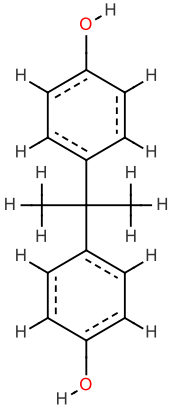
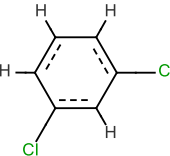
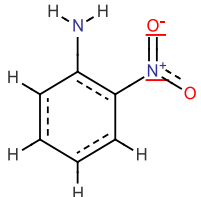
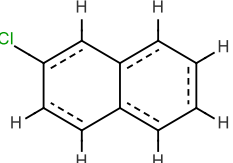
8	2-chlorophenol		-2.16	-1.96	-2.01	-1.99	-2.02	-1.80
9	ortho-dichlorobenzene		-1.81	-1.08	-1.12	-1.40	-1.41	-1.19
10*	1,2,4,5-tetrachlorobenzene		0.2	0.07	-0.02	0.10	0.16	0.48
11	p-nitrophenol		-1.45	-1.14	-1.10	-1.10	-1.11	-1.37
12	4-chlorotoluene		-1.55	-1.39	-1.45	-1.69	-1.71	-1.70

13	3-methylbenzyl alcohol		-2.52	-2.35	-2.75	-2.85	-2.28	-2.08
14	4-chloroanisole		-1.3	-1.92	-1.75	-1.96	-1.99	-1.62
15	atrazine		-0.55	-0.36	-1.34	-0.42	-0.40	-0.42
16	3-methylphenol		-2.29	-2.18	-2.21	-2.17	-2.22	-2.32

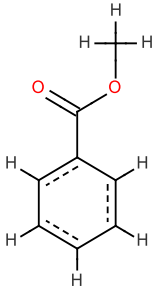
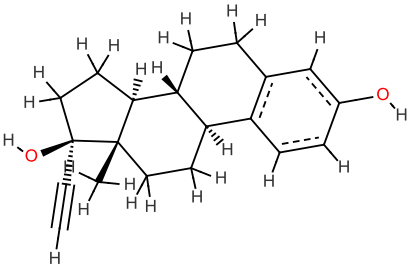
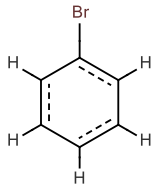
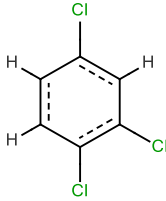
17*	4-methylphenol		-2.18	-2.18	-2.21	-2.17	-2.22	-2.30
18	aniline		-3.01	-3.22	-3.12	-2.97	-3.05	-3.07
19	pentachlorophenol		-0.7	-0.88	-0.97	-0.92	-0.88	-0.89
20*	1,3-dinitrobenzene		-0.75	0.33	-0.13	0.28	0.32	-0.05

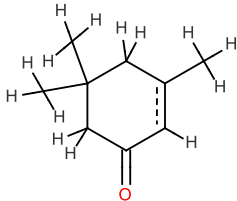
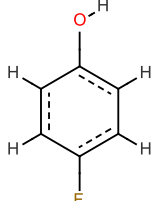
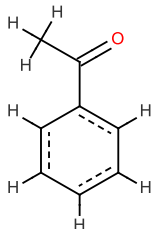
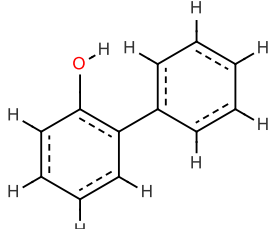
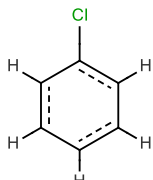
21*	phenylethylalcohol		-2.83	-2.53	-2.75	-2.85	-2.28	-2.20
22	pyrene		1.77	2.02	1.80	2.00	2.06	1.61
23	p-dichlorobenzene		-1.86	-2.04	-2.06	-1.40	-1.41	-1.42
24	4-chloroaniline		-2.9	-2.57	-2.52	-2.61	-2.68	-2.61
25*	benzonitrile		-2.33	-3.15	-2.78	-2.84	-2.91	-2.61

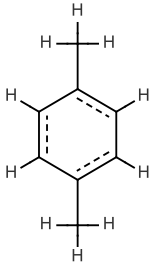
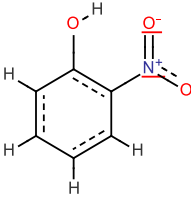
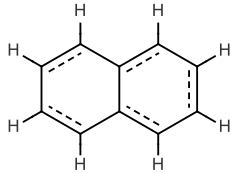
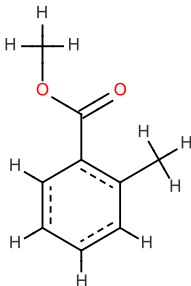
26	phenanthrene		1.13	0.96	0.89	0.92	0.96	0.76
27	carbamazepine		-1.47	-1.25	-0.73	-0.99	-1.03	-1.02
28	4-chlorophenol		-1.5	-1.96	-2.08	-2.07	-2.11	-2.03
29	biphenyl		-0.17	-0.20	-0.02	-0.23	-0.22	-0.32

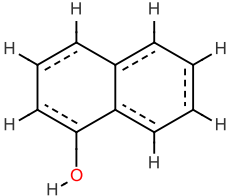
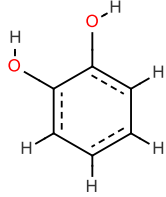
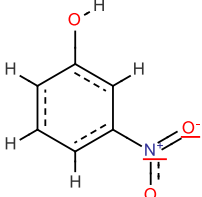
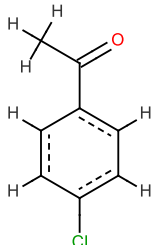
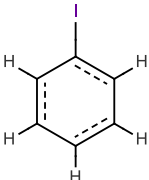
30*	bisphenol A		-0.73	-0.20	0.20	-0.51	-0.50	-0.45
31*	m-dichlorobenzene		-1.72	-1.03	-1.12	-1.40	-1.41	-1.43
32	2-nitroaniline		-0.64	-1.46	-1.71	-1.37	-1.39	-1.38
33*	2-chloronaphthalene		0.36	0.10	-0.04	-0.07	-0.06	-0.14

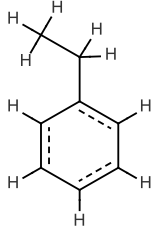
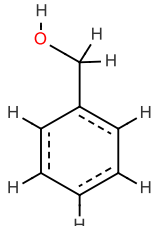
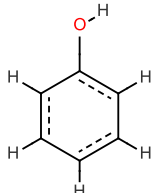
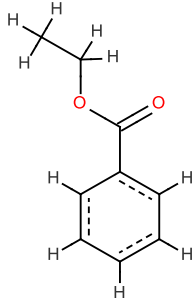
34	azobenzene		0.35	-1.18	-0.63	-1.05	-1.08	-1.20
35	tetracycline		0.21	-0.38	-0.15	-0.25	-0.28	-0.12
36	3-bromophenol		-1.58	-1.67	-1.76	-1.75	-1.78	-1.93
37	4-ethylphenol		-1.75	-2.13	-1.88	-2.05	-2.09	-1.89

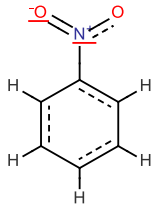
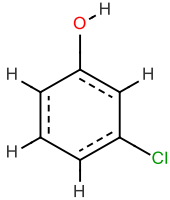
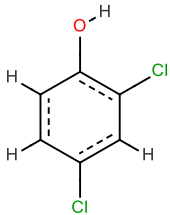
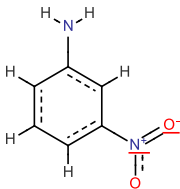
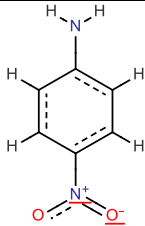
38*	methylbenzoate		-1.67	-1.75	-1.39	-1.60	-1.62	-1.45
39*	17 α -ethynylestradiol		0.99	1.07	1.36	0.84	0.87	0.91
40	bromobenzene		-1.87	-1.80	-1.83	-1.85	-1.88	-1.96
41	1,2,4-trichlorobenzene		-1	-1.02	-1.08	-0.70	-0.67	-0.48

42	isophorone		-2.36	-1.92	-2.01	-2.32	-2.38	-2.37
43	4-fluorophenol		-2.69	-1.97	-2.03	-1.98	-2.02	-2.14
44	acetophenone		-2.11	-2.10	-1.90	-2.06	-2.10	-2.08
45	2-phenylphenol		-1.16	-0.81	-0.53	-0.71	-0.72	-0.75
46	chlorobenzene		-2.35	-2.00	-2.01	-2.01	-2.05	-2.12

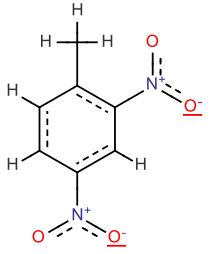
47	p-xylene		-2.11	-1.73	-1.76	-1.95	-1.98	-1.96
48*	2-nitrophenol		-1.69	-1.17	-1.10	-1.10	-1.11	-1.25
49	naphthalene		-0.45	-0.80	-0.93	-0.65	-0.67	-0.74
50	methyl 2-methylbenzoate		-1.25	-1.53	-1.13	-1.53	-1.55	-1.23

51	1-naphthol		-1.24	-1.15	-1.20	-0.94	-0.97	-1.00
52*	catechol		-1.95	-1.99	-2.06	-1.82	-1.86	-1.89
53	3-nitrophenol		-1.32	-1.14	-1.10	-1.10	-1.11	-1.39
54*	4-chloroacetophenone		-1.09	-1.57	-1.45	-1.85	-1.88	-1.58
55	iodobenzene		-1.49	-1.59	-1.64	-1.69	-1.71	-1.80

56	ethylbenzene		-2.18	-2.02	-1.76	-1.95	-1.98	-1.73
57	benzylalcohol		-3.27	-2.66	-3.04	-2.94	-2.38	-2.22
58	phenol		-2.73	-2.48	-2.50	-2.23	-2.28	-2.49
59	ethyl benzoate		-1.23	-1.66	-1.13	-1.53	-1.55	-1.44

60	nitrobenzene		-1.86	-1.46	-1.39	-1.38	-1.40	-1.58
61	3-chlorophenol		-1.75	-1.93	-2.04	-2.03	-2.06	-2.77
62*	2,4-dichlorophenol		-1.28	-1.34	-1.48	-1.72	-1.74	-1.33
63	3-nitroaniline		-1.53	-1.28	-1.71	-1.21	-1.23	-1.37
64*	4-nitroaniline		-1.29	-1.13	-1.71	-1.06	-1.07	-1.21

65	<i>m</i> -nitrotoluene		-1.17	-1.28	-1.17	-1.39	-1.40	-1.50
66	2,4,5-trichlorophenoxyacetic acid		-2.51	-2.60	-2.62	-2.64	-2.64	-1.27
67	4-nitrotoluene		-0.93	-1.28	-1.17	-1.39	-1.40	-1.48
68*	propylbenzene		-1.61	-1.64	-1.23	-1.64	-1.65	-1.51

69	2,4-dinitrotoluene		0.21	0.32	-0.06	0.13	0.17	-0.01
----	--------------------	-----------------------------------------------------------------------------------	------	------	-------	------	------	-------

* denotes the test set compounds

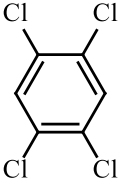
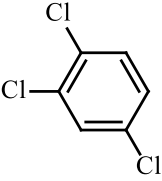
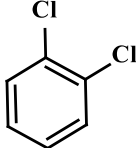
3.2. Study 2: Dataset 2

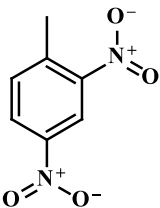
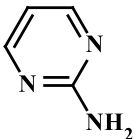
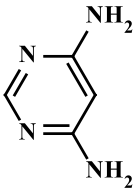
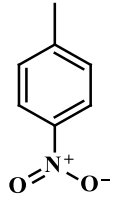
A diverse set of 40 hazardous synthetic organic chemicals (SOC) with defined adsorption coefficient onto SWCNTs reported in the literature (Ding et al., 2016) were used to develop the QSPR models. The whole data set of 40 synthetic organic chemicals was assembled from 14 published articles containing experimental adsorption coefficient (K in, L/kg) values. The adsorption coefficient K was calculated by using following formula:

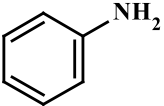

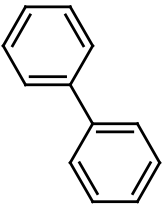
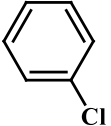

$$K = \frac{q_e}{C_e}$$

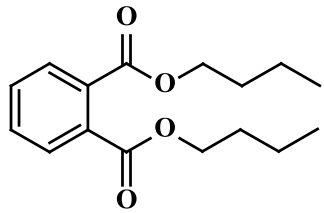
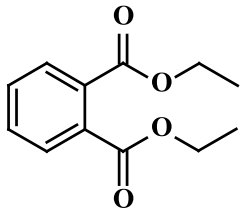
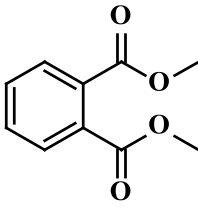
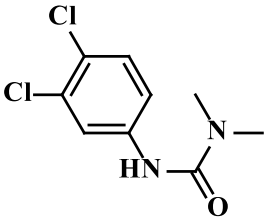
where, q_e (mg/kg) is equilibrium concentration on the surface and C_e (mg/L) is the equilibrium concentration in the aqueous phase of SWCNTs. The adsorption coefficient depends on the equilibrium concentration whenever the adsorption isotherm is nonlinear in nature (Zhao et al., 2014). The effect of concentration on K was investigated. The equilibrium concentration on the surface (q_e) could be obtained from isotherm data at $C_e=0.00002, 0.0002, 0.002, 0.02$ and $0.2 C_s$, (where C_s is the aqueous solubility of the adsorbate). The consequent K values are represented as $K_{0.00002}, K_{0.0002}, K_{0.002}, K_{0.02}$ and $K_{0.2}$ respectively. The endpoint K values were taken in the logarithmic scale for the development of QSPR models. We have used $\log K_{0.002}$ values for the development of QSPR models due to its relatively wide distribution than rest of the $\log K$ values. The data set is given in Table 3.2.

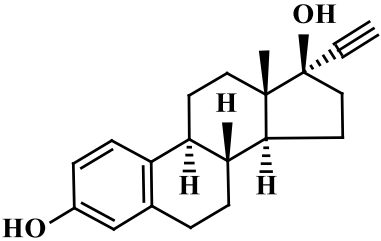
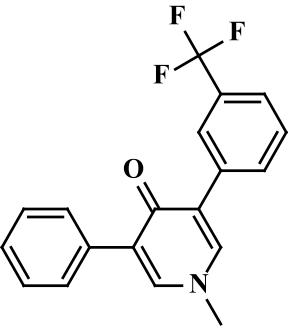
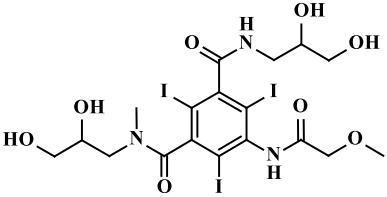
Table 3.2. The chemical name, experimental logK and calculated logK values of the MLR and PLS models.

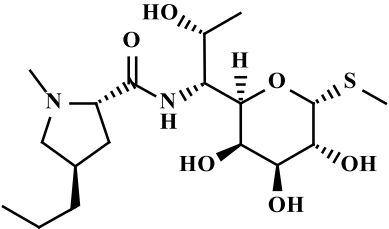
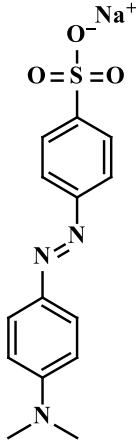
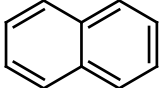
Sl. No.	Chemical name	Chemical Structure	logK					
			Expt.	Predicted				
				Model M1	Model M2	Model M3	Model M4	Model M5
1*	1,2,4,5-Tetrachlorobenzene		3.68	4.11	4.06	4.46	4.36	4.28
2	1,2,4-Trichlorobenzene		2.94	3.03	2.99	3.27	3.21	3.16
3	1,2-Dichlorobenzene		2.33	2.07	1.95	2.11	2.10	2.07

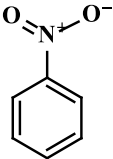
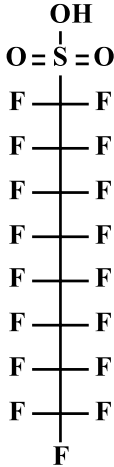
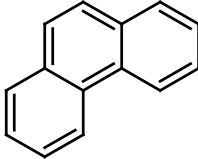
4	2,4-Dinitrotoluene		2.07	1.83	2.05	2.00	2.03	2.09
5*	2-Aminopyrimidine		-0.54	-0.27	-0.41	-0.52	-0.42	-0.40
6	4,6-Diaminopyrimidine		-0.27	0.33	0.19	0.11	0.24	0.19
7	4-Nitrotoluene		1.67	1.28	1.59	1.51	1.49	1.63

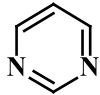
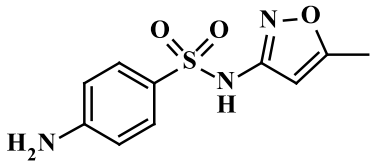

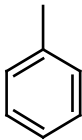
8	Aniline		-0.16	0.03	-0.07	-0.14	-0.08	-0.04
9	Benzene		0.25	-0.28	-0.27	-0.32	-0.35	-0.21
10	Biphenyl		2.87	2.54	2.64	2.38	2.37	2.33
11	Chlorobenzene		1.16	0.90	0.89	0.94	0.92	0.98
12	Cyclohexane		0.44	0.44	0.52	0.56	0.48	0.63

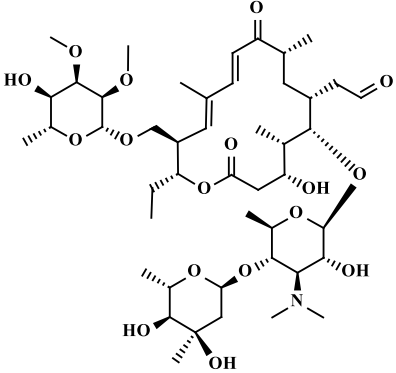
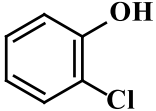
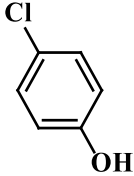
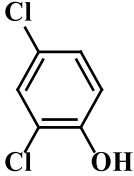
13*	Dibutyl phthalate		2.95	2.28	2.53	2.59	2.47	2.66
14	Diethyl phthalate		1.68	1.62	1.70	1.65	1.65	1.77
15	Dimethyl phthalate		1	1.49	1.46	1.36	1.42	1.50
16	Diuron		2.28	1.80	2.07	2.05	2.02	2.14

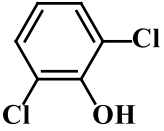
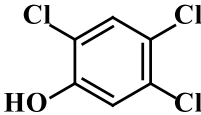
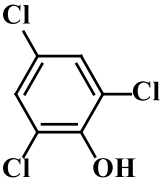
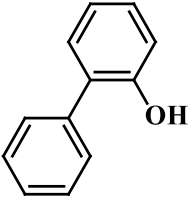
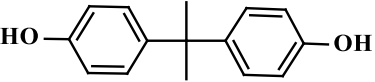
17	Ethinyl estradiol		3.64	3.65	3.33	3.48	3.46	3.49
18	Fluridone		1.81	1.86	2.10	1.68	1.96	2.14
19	Iopromide		0.89	0.94	0.86	0.68	0.80	0.86

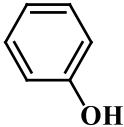
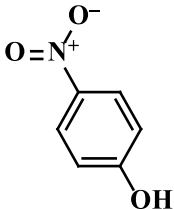
20	Lincomycin	 <p>The structure of Lincomycin is a complex molecule consisting of a 6-membered lactam ring fused to a 5-membered imidazolidine ring, which is further attached to a 6-membered pyranose ring. The pyranose ring has several hydroxyl groups and a methyl group attached to it.</p>	-0.53	-0.49	-0.58	-0.64	-0.12	-0.55
21	Methyl orange	 <p>The structure of Methyl orange is a diazo dye. It consists of a central benzene ring with a sulfonate group (-SO₃⁻Na⁺) at the para position and a dimethylamino group (-N(CH₃)₂) at the other para position. This central ring is connected via an azo group (-N=N-) to another benzene ring, which also has a dimethylamino group (-N(CH₃)₂) at the para position.</p>	0.49	0.57	0.59	0.78	0.90	0.32
22	Naphthalene	 <p>The structure of Naphthalene is a polycyclic aromatic hydrocarbon consisting of two fused benzene rings.</p>	1.8	2.22	2.19	1.86	1.93	1.84

23	Nitrobenzene		0.61	0.67	0.49	0.46	0.55	0.53
24	PFOS		1.29	1.34	1.24	1.92	1.78	1.54
25	Phenanthrene		3.67	3.92	3.81	3.67	3.69	3.54

26	Pyrimidine		-1.56	-1.05	-1.13	-1.29	-1.24	-1.12
27	Sulfamethoxazole		1.43	0.68	0.63	1.15	0.53	0.64
28*	Thiophene		-0.07	-0.84	-0.89	-1.03	-1.00	-0.87
29	Toluene		0.78	0.71	1.14	0.70	0.70	0.75

30	Tylosin		0.43	0.15	0.16	0.63	-0.03	0.05
31*	2-Chlorophenol		0.75	1.00	0.77	0.77	0.87	0.82
32	4-Chlorophenol		0.81	0.86	0.77	0.77	0.84	0.82
33*	2,4-Dichlorophenol		1.76	1.75	1.57	1.66	1.73	1.65

34*	2,6-Dichlorophenol		1.61	1.87	1.57	1.66	1.75	1.65
35*	2,4,5-Trichlorophenol		2.45	2.61	2.40	2.58	2.63	2.51
36*	2,4,6-Trichlorophenol		2.35	2.60	2.40	2.58	2.63	2.51
37	2-Phenylphenol		2.16	2.08	2.07	2.07	2.05	2.17
38	Bisphenol A		1.93	2.27	2.30	2.33	2.29	2.41

39*	Phenol		-0.7	0.04	-0.06	-0.13	-0.06	-0.03
40	p-Nitrophenol		0.44	0.84	0.63	0.60	0.74	0.65

*denotes the test set compounds

3.3. General description of methods applied for developing QSPR models

3.3.1. Descriptor calculation

“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 experiments.” The descriptors were calculated using two software tools namely Dragon software version 6 (http://www.taletе.mi.it/products/dragon_description.htm.) and PaDEL-descriptor (<http://www.yapcsoft.com/dd/padeldescriptor>) software. In this work, we have calculated only 2D descriptors covering constitutional, ring descriptors, connectivity index, functional group counts, atom centered fragments, atom type E-states, 2D atom pairs, molecular properties (using Dragon software version 6) and ETA indices (using PaDEL-Descriptor software). Table 3.3 contains nine types of 2D descriptors which are used during model development.

Table 3.3. Description of 2D descriptors used for the QSAR studies

Dimension of descriptors	Class of descriptors	Representative example
2D	Constitutional indices	Number of atoms, number of non-H atoms, number of binds, number of aromatic bonds, sum of atomic van der Waals volumes (scaled on carbon atom) etc.
2D	Ring descriptors	Number of rings (cyclomatic number), number of circuits ,total ring size, ring perimeter, ring bridge count , molecular cyclized degree, ring fusion density, ring complexity index, number of ring systems, normalized number of ring systems.
2D	Connectivity index	Connectivity index of order 0, connectivity index of order 1, modified Randic index, average connectivity index of order 0.
2D	Functional group	Number of terminal primary C(sp ³), number of total

	count	secondary C(sp ³), number of total tertiary C(sp ³), number of esters (aliphatic), number of esters (aromatic), number of hydrazones, number of hydroxyl groups
2D	Atom centered fragment	CH ₃ R / CH ₄ , CH ₂ R ₂ , CHR ₃ , CR ₄ , CH ₃ X, CH ₂ RX, CH ₂ X ₂ , CHR ₂ X, CHR _X ₂
2D	Atom type E-states	Sum of sCH ₃ , Sum of dCH ₂ E-states, Sum of ssCH ₂ E-states, Sum of sOH E-states, Sum of tN E-states
2D	2D atom pairs	Sum of topological distances between N..N, T(N..O) sum of topological distances between N..O, sum of topological distances between N..S, sum of topological distances between N..P, sum of topological distances between N..F.
2D	Molecular properties	Unsaturation index, hydrophilic factor, Ghose-Crippen molar refractivity, Moriguchi octanol-water partition coeff. (logP), squared Moriguchi octanol-water partition coeff. (logP ²).
2D	ETA indices	First and second generation ETA indices

(a) Constitutional indices

These are the most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry or topology. The most common constitutional descriptors are number of atoms, number of bonds, absolute and relative numbers of specific atom-types, absolute and relative numbers of single, double, triple, and aromatic bonds, number of rings, number of rings divided by the number of atoms or bonds, number of benzene rings, number of benzene rings divided by the number of atoms and information on hybridization (Todeschini and Consonni, 2009). The descriptors used during the model development are constitutional, ring descriptors, connectivity index, functional group counts, atom centered fragments, atom type E-states, 2D atom pairs, molecular properties. The details of all these descriptors are given in **Table 3.4 to Table 3.12** in the following section.

Table 3.4. The details of constitutional indices with their symbols.

Descriptor Symbol	Descriptor
MW	molecular weight
AMW	average molecular weight
Sv	sum of atomic van der Waals volumes (scaled on Carbon atom)
Se	sum of atomic Sanderson electronegativities (scaled on Carbon atom)
Sp	sum of atomic polarizabilities (scaled on Carbon atom)
Si	sum of first ionization potentials (scaled on Carbon atom)
Mv	mean atomic van der Waals volume (scaled on Carbon atom)
Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)
Mp	mean atomic polarizability (scaled on Carbon atom)
Mi	mean first ionization potential (scaled on Carbon atom)
nAT	number of atoms
nSK	number of non-H atoms
nBT	number of bonds
nBO	number of non-H bonds
nBM	number of multiple bonds
SCBO	sum of conventional bond orders (H-depleted)
RBN	number of rotatable bonds
RBF	rotatable bond fraction
nDB	number of double bonds
nTB	number of triple bonds
nAB	number of aromatic bonds
nH	number of Hydrogen atoms
nC	number of Carbon atoms
nN	number of Nitrogen atoms
nO	number of Oxygen atoms
nP	number of Phosphorous atoms
nS	number of Sulfur atoms
nF	number of Fluorine atoms
nCL	number of Chlorine atoms
nBR	number of Bromine atoms
nI	number of Iodine atoms
nB	number of Boron atoms
nHM	number of heavy atoms
nHet	number of heteroatoms
nX	number of halogen atoms
H%	percentage of H atoms
C%	percentage of C atoms
N%	percentage of N atoms
O%	percentage of O atoms
X%	percentage of halogen atoms
nCsp3	number of sp ³ hybridized Carbon atoms
nCsp2	number of sp ² hybridized Carbon atoms
nCsp	number of sp hybridized Carbon atoms

Table 3.5. The details of ring descriptors with their symbols.

Descriptor Symbol	Description
nCIC	number of rings (cyclomatic number)
nCIR	number of circuits
TRS	total ring size
Rperim	ring perimeter
Rbrid	ring bridge count
MCD	molecular cyclized degree
RFD	ring fusion density
RCI	ring complexity index
NRS	number of ring systems
NNRS	normalized number of ring systems
nR03	number of 3-membered rings
nR04	number of 4-membered rings
nR05	number of 5-membered rings
nR06	number of 6-membered rings
nR07	number of 7-membered rings
nR08	number of 8-membered rings
nR09	number of 9-membered rings
nR10	number of 10-membered rings
nR11	number of 11-membered rings
nR12	number of 12-membered rings
nBnz	number of benzene-like rings
ARR	aromatic ratio
D/Dtr03	distance/detour ring index of order 3
D/Dtr04	distance/detour ring index of order 4
D/Dtr05	distance/detour ring index of order 5
D/Dtr06	distance/detour ring index of order 6
D/Dtr07	distance/detour ring index of order 7
D/Dtr08	distance/detour ring index of order 8
D/Dtr09	distance/detour ring index of order 9
D/Dtr10	distance/detour ring index of order 10
D/Dtr11	distance/detour ring index of order 11
D/Dtr12	distance/detour ring index of order 12
nCIC	number of rings (cyclomatic number)
nCIR	number of circuits
TRS	total ring size

Table 3.6. The details of connectivity indices with their symbols.

Descriptor Symbol	Description
X0	connectivity index of order 0
X1	connectivity index of order 1 (Randic connectivity index)
X2	connectivity index of order 2
X3	connectivity index of order 3
X4	connectivity index of order 4
X5	connectivity index of order 5
X0A	average connectivity index of order 0
X1A	average connectivity index of order 1
X2A	average connectivity index of order 2
X3A	average connectivity index of order 3
X4A	average connectivity index of order 4
X5A	average connectivity index of order 5
X0v	valence connectivity index of order 0
X1v	valence connectivity index of order 1
X2v	valence connectivity index of order 2
X3v	valence connectivity index of order 3
X4v	valence connectivity index of order 4
X5v	valence connectivity index of order 5
X0Av	average valence connectivity index of order 0
X1Av	average valence connectivity index of order 1
X2Av	average valence connectivity index of order 2
X3Av	average valence connectivity index of order 3
X4Av	average valence connectivity index of order 4
X5Av	average valence connectivity index of order 5
X0sol	solvation connectivity index of order 0
X1sol	solvation connectivity index of order 1
X2sol	solvation connectivity index of order 2
X3sol	solvation connectivity index of order 3
X4sol	solvation connectivity index of order 4
X5sol	solvation connectivity index of order 5
XMOD	modified Randic index
RDCHI	reciprocal distance sum Randic-like index
RDSQ	reciprocal distance sum inverse Randic-like index
X1Kup	Kupchik connectivity index
X1Mad	connectivity topochemical index
X1Per	perturbation connectivity index

Table 3.7. The details of functional group count descriptors with their symbols.

Descriptor Symbol	Description
nCp	number of terminal primary C(sp3)
nCs	number of total secondary C(sp3)
nCt	number of total tertiary C(sp3)
nCq	number of total quaternary C(sp3)
nCrs	number of ring secondary C(sp3)
nCrt	number of ring tertiary C(sp3)
nCrq	number of ring quaternary C(sp3)
nCar	number of aromatic C(sp2)
nCbH	number of unsubstituted benzene C(sp2)
nCb-	number of substituted benzene C(sp2)
nCconj	number of non-aromatic conjugated C(sp2)
nR=Cp	number of terminal primary C(sp2)
nR=Cs	number of aliphatic secondary C(sp2)
nR=Ct	number of aliphatic tertiary C(sp2)
n=C=	number of allenes groups
nR#CH/X	number of terminal C(sp)
nR#C-	number of non-terminal C(sp)
nROCN	number of cyanates (aliphatic)
nArOCN	number of cyanates (aromatic)
nRNCO	number of isocyanates (aliphatic)
nArNCO	number of isocyanates (aromatic)
nRSCN	number of thiocyanates (aliphatic)
nArSCN	number of thiocyanates (aromatic)
nRNCS	number of isothiocyanates (aliphatic)
nArNCS	number of isothiocyanates (aromatic)
nRCOOH	number of carboxylic acids (aliphatic)
nArCOOH	number of carboxylic acids (aromatic)
nRCOOR	number of esters (aliphatic)
nArCOOR	number of esters (aromatic)
nRCONH2	number of primary amides (aliphatic)
nArCONH2	number of primary amides (aromatic)
nRCONHR	number of secondary amides (aliphatic)
nArCONHR	number of secondary amides (aromatic)
nRCONR2	number of tertiary amides (aliphatic)
nArCONR2	number of tertiary amides (aromatic)
nROCON	number of (thio-) carbamates (aliphatic)
nArOCON	number of (thio-) carbamates (aromatic)
nRCOX	number of acyl halogenides (aliphatic)
nArCOX	number of acyl halogenides (aromatic)
nRCSOH	number of thioacids (aliphatic)
nArCSOH	number of thioacids (aromatic)
nRCSSH	number of dithioacids (aliphatic)
nArCSSH	number of dithioacids (aromatic)
nRCOSR	number of thioesters (aliphatic)
nArCOSR	number of thioesters (aromatic)
nRCSSR	number of dithioesters (aliphatic)
nArCSSR	number of dithioesters (aromatic)
nRCHO	number of aldehydes (aliphatic)
nArCHO	number of aldehydes (aromatic)

nRCO	number of ketones (aliphatic)
nArCO	number of ketones (aromatic)
nCONN	number of urea (-thio) derivatives
nC=O(O)2	number of carbonate (-thio) derivatives
nN=C-N<	number of amidine derivatives
nC(=N)N2	number of guanidine derivatives
nRC=N	number of imines (aliphatic)
nArC=N	number of imines (aromatic)
nRCNO	number of oximes (aliphatic)
nArCNO	number of oximes (aromatic)
nRNH2	number of primary amines (aliphatic)
nArNH2	number of primary amines (aromatic)
nRNHR	number of secondary amines (aliphatic)
nArNHR	number of secondary amines (aromatic)
nRNR2	number of tertiary amines (aliphatic)
nArNR2	number of tertiary amines (aromatic)
nN-N	number of N hydrazines
nN=N	number of N azo-derivatives
nRCN	number of nitriles (aliphatic)
nArCN	number of nitriles (aromatic)
nN+	number of positively charged N
nNq	number of quaternary N
nRNHO	number of hydroxylamines (aliphatic)
nArNHO	number of hydroxylamines (aromatic)
nRNNOx	number of N-nitroso groups (aliphatic)
nArNNOx	number of N-nitroso groups (aromatic)
nRNO	number of nitroso groups (aliphatic)
nArNO	number of nitroso groups (aromatic)
nRNO2	number of nitro groups (aliphatic)
nArNO2	number of nitro groups (aromatic)
nN(CO)2	number of imides (-thio)
nC=N-N<	number of hydrazones
nROH	number of hydroxyl groups
nArOH	number of aromatic hydroxyls
nOHp	number of primary alcohols
nOHs	number of secondary alcohols
nOHt	number of tertiary alcohols
nROR	number of ethers (aliphatic)
nArOR	number of ethers (aromatic)
nROX	number of hypohalogenides (aliphatic)
nArOX	number of hypohalogenides (aromatic)
nO(C=O)2	number of anhydrides (-thio)
nH2O	number of water molecules
nSH	number of thiols
nC=S	number of thioketones
nRSR	number of sulfides
nRSSR	number of disulfides
nSO	number of sulfoxides
nS(=O)2	number of sulfones
nSOH	number of sulfenic (thio-) acids
nSOOH	number of sulfinic (thio-/dithio-) acids
nSO2OH	number of sulfonic (thio-/dithio-) acids
nSO3OH	number of sulfuric (thio-/dithio-) acids

nSO ₂	number of sulfites (thio-/dithio-)
nSO ₃	number of sulfonates (thio-/dithio-)
nSO ₄	number of sulfates (thio-/dithio-)
nSO ₂ N	number of sulfonamides (thio-/dithio-)
nPO ₃	number of phosphites/thiophosphites
nPO ₄	number of phosphates/thiophosphates
nPR ₃	number of phosphanes
nP(=O)O ₂ R	number of phosphonates (thio-)
nP(=O)R ₃ /nPR ₅	number of phosphoranes (thio-)
nCH ₂ RX	number of CH ₂ RX
nCHR ₂ X	number of CHR ₂ X
nCR ₃ X	number of CR ₃ X
nR=CHX	number of R=CHX
nR=CRX	number of R=CRX
nR#CX	number of R#CX
nCHRX ₂	number of CHRX ₂
nCR ₂ X ₂	number of CR ₂ X ₂
nR=CX ₂	number of R=CX ₂
nCRX ₃	number of CRX ₃
nArX	number of X on aromatic ring
nCX _r	number of X on ring C(sp ³)
nCX _r =	number of X on ring C(sp ²)
nCconjX	number of X on exo-conjugated C
nAziridines	number of Aziridines
nOxiranes	number of Oxiranes
nThiranes	number of Thiranes
nAzetidines	number of Azetidines
nOxetanes	number of Oxetanes
nThioethanes	number of Thioethanes
nBeta-Lactams	number of Beta-Lactams
nPyrrolidines	number of Pyrrolidines
nOxolanes	number of Oxolanes
ntH-Thiophenes	number of tetrahydro-thiophenes
nPyrroles	number of Pyrroles
nPyrazoles	number of Pyrazoles
nImidazoles	number of Imidazoles
nFuranes	number of Furanes
nThiophenes	number of Thiophenes
nOxazoles	number of Oxazoles
nIsoxazoles	number of Isoxazoles
nThiazoles	number of Thiazoles
nIsothiazoles	number of Isothiazoles
nTriazoles	number of Triazoles
nPyridines	number of Pyridines
nPyridazines	number of Pyridazines
nPyrimidines	number of Pyrimidines
nPyrazines	number of Pyrazines
n135-Triazines	number of 1-3-5-Triazines
n124-Triazines	number of 1-2-4-Triazines
nHDon	number of aromatic hydroxyls
nHAcc	number of primary alcohols
nHBonds	number of secondary alcohols

Table 3.8. The details of atom centered fragment descriptors with their symbols.

Descriptor Symbol	Description
C-001	CH3R / CH4
C-002	CH2R2
C-003	CHR3
C-004	CR4
C-005	CH3X
C-006	CH2RX
C-007	CH2X2
C-008	CHR2X
C-009	CHRX2
C-010	CHX3
C-011	CR3X
C-012	CR2X2
C-013	CRX3
C-014	CX4
C-015	=CH2
C-016	=CHR
C-017	=CR2
C-018	=CHX
C-019	=CRX
C-020	=CX2
C-021	#CH
C-022	#CR / R=C=R
C-023	#CX
C-024	R--CH--R
C-025	R--CR--R
C-026	R--CX--R
C-027	R--CH--X
C-028	R--CR--X
C-029	R--CX--X
C-030	X--CH--X
C-031	X--CR--X
C-032	X--CX--X
C-033	R--CH..X
C-034	R--CR..X
C-035	R--CX..X
C-036	Al-CH=X
C-037	Ar-CH=X
C-038	Al-C(=X)-Al
C-039	Ar-C(=X)-R
C-040	R-C(=X)-X / R-C#X / X=C=X
C-041	X-C(=X)-X
C-042	X--CH..X
C-043	X--CR..X
C-044	X--CX..X
H-046	H attached to C0(sp3) no X attached to next C

H-047	H attached to C1(sp3)/C0(sp2)
H-048	H attached to C2(sp3)/C1(sp2)/C0(sp)
H-049	H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)
H-050	H attached to heteroatom
H-051	H attached to alpha-C
H-052	H attached to C0(sp3) with 1X attached to next C
H-053	H attached to C0(sp3) with 2X attached to next C
H-054	H attached to C0(sp3) with 3X attached to next C
H-055	H attached to C0(sp3) with 4X attached to next C
O-056	alcohol
O-057	phenol / enol / carboxyl OH
O-058	#NOME?
O-059	Al-O-Al
O-060	Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X
O-061	O--
O-062	O- (negatively charged)
O-063	R-O-O-R
Se-064	Any-Se-Any
Se-065	#NOME?
N-066	Al-NH2
N-067	Al2-NH
N-068	Al3-N
N-069	Ar-NH2 / X-NH2
N-070	Ar-NH-Al
N-071	Ar-NAl2
N-072	RCO-N< / >N-X=X
N-073	Ar2NH / Ar3N / Ar2N-Al / R..N..R
N-074	R#N / R=N-
N-075	R--N--R / R--N--X
N-076	Ar-NO2 / R--N(--R)--O / RO-NO
N-077	Al-NO2
N-078	Ar-N=X / X-N=X
N-079	N+ (positively charged)
F-081	F attached to C1(sp3)
F-082	F attached to C2(sp3)
F-083	F attached to C3(sp3)
F-084	F attached to C1(sp2)
F-085	F attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X
Cl-086	Cl attached to C1(sp3)
Cl-087	Cl attached to C2(sp3)
Cl-088	Cl attached to C3(sp3)
Cl-089	Cl attached to C1(sp2)
Cl-090	Cl attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X
Br-091	Br attached to C1(sp3)
Br-092	Br attached to C2(sp3)
Br-093	Br attached to C3(sp3)
Br-094	Br attached to C1(sp2)
Br-095	Br attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X
I-096	I attached to C1(sp3)
I-097	I attached to C2(sp3)
I-098	I attached to C3(sp3)
I-099	I attached to C1(sp2)
I-100	I attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X

F-101	fluoride ion
Cl-102	chloride ion
Br-103	bromide ion
I-104	iodide ion
S-106	R-SH
S-107	R ₂ S / RS-SR
S-108	R=S
S-109	R-SO-R
S-110	R-SO ₂ -R
Si-111	>Si<
B-112	>B- as in boranes
P-115	P ylids
P-116	R ₃ -P=X
P-117	X ₃ -P=X (phosphate)
P-118	PX ₃ (phosphite)
P-119	PR ₃ (phosphine)
P-120	C-P(X) ₂ =X (phosphonate)

Table 3.9. The details of atom centered fragment descriptors with their symbols.

Descriptor Symbol	Description
T(N..N)	sum of topological distances between N..N
T(N..O)	sum of topological distances between N..O
T(N..S)	sum of topological distances between N..S
T(N..P)	sum of topological distances between N..P
T(N..F)	sum of topological distances between N..F
T(N..Cl)	sum of topological distances between N..Cl
T(N..Br)	sum of topological distances between N..Br
T(N..I)	sum of topological distances between N..I
T(O..O)	sum of topological distances between O..O
T(O..S)	sum of topological distances between O..S
T(O..P)	sum of topological distances between O..P
T(O..F)	sum of topological distances between O..F
T(O..Cl)	sum of topological distances between O..Cl
T(O..Br)	sum of topological distances between O..Br
T(O..I)	sum of topological distances between O..I
T(S..S)	sum of topological distances between S..S
T(S..P)	sum of topological distances between S..P
T(S..F)	sum of topological distances between S..F
T(S..Cl)	sum of topological distances between S..Cl
T(S..Br)	sum of topological distances between S..Br
T(S..I)	sum of topological distances between S..I
T(P..P)	sum of topological distances between P..P
T(P..F)	sum of topological distances between P..F
T(P..Cl)	sum of topological distances between P..Cl
T(P..Br)	sum of topological distances between P..Br
T(P..I)	sum of topological distances between P..I

T(F..F)	sum of topological distances between F..F
T(F..Cl)	sum of topological distances between F..Cl
T(F..Br)	sum of topological distances between F..Br
T(F..I)	sum of topological distances between F..I
T(Cl..Cl)	sum of topological distances between Cl..Cl
T(Cl..Br)	sum of topological distances between Cl..Br
T(Cl..I)	sum of topological distances between Cl..I
T(Br..Br)	sum of topological distances between Br..Br
T(Br..I)	sum of topological distances between Br..I
T(I..I)	sum of topological distances between I..I
B04[C-C]	Presence/absence of C - C at topological distance 4
B04[C-N]	Presence/absence of C - N at topological distance 4
B04[C-O]	Presence/absence of C - O at topological distance 4
B04[C-S]	Presence/absence of C - S at topological distance 4
B04[C-P]	Presence/absence of C - P at topological distance 4
B04[C-F]	Presence/absence of C - F at topological distance 4
B04[C-Cl]	Presence/absence of C - Cl at topological distance 4
B04[C-Br]	Presence/absence of C - Br at topological distance 4
B04[C-I]	Presence/absence of C - I at topological distance 4
B04[C-B]	Presence/absence of C - B at topological distance 4
B04[C-Si]	Presence/absence of C - Si at topological distance 4
B04[C-X]	Presence/absence of C - X at topological distance 4
B04[N-N]	Presence/absence of N - N at topological distance 4
B04[N-O]	Presence/absence of N - O at topological distance 4
B04[N-S]	Presence/absence of N - S at topological distance 4
B04[N-P]	Presence/absence of N - P at topological distance 4
B04[N-F]	Presence/absence of N - F at topological distance 4
B04[N-Cl]	Presence/absence of N - Cl at topological distance 4
B04[N-Br]	Presence/absence of N - Br at topological distance 4
B04[N-I]	Presence/absence of N - I at topological distance 4
B04[N-B]	Presence/absence of N - B at topological distance 4
B04[N-Si]	Presence/absence of N - Si at topological distance 4
B04[N-X]	Presence/absence of N - X at topological distance 4
B04[O-O]	Presence/absence of O - O at topological distance 4
B04[O-S]	Presence/absence of O - S at topological distance 4
B04[O-P]	Presence/absence of O - P at topological distance 4
B04[O-F]	Presence/absence of O - F at topological distance 4
B04[O-Cl]	Presence/absence of O - Cl at topological distance 4
B04[O-Br]	Presence/absence of O - Br at topological distance 4
B04[O-I]	Presence/absence of O - I at topological distance 4
B04[O-B]	Presence/absence of O - B at topological distance 4
B04[O-Si]	Presence/absence of O - Si at topological distance 4
B04[O-X]	Presence/absence of O - X at topological distance 4
B04[S-S]	Presence/absence of S - S at topological distance 4
B04[S-P]	Presence/absence of S - P at topological distance 4
B04[S-F]	Presence/absence of S - F at topological distance 4
B04[S-Cl]	Presence/absence of S - Cl at topological distance 4
B04[S-Br]	Presence/absence of S - Br at topological distance 4
B04[S-I]	Presence/absence of S - I at topological distance 4

B04[S-B]	Presence/absence of S - B at topological distance 4
B04[S-Si]	Presence/absence of S - Si at topological distance 4
B04[S-X]	Presence/absence of S - X at topological distance 4
B04[P-P]	Presence/absence of P - P at topological distance 4
B04[P-F]	Presence/absence of P - F at topological distance 4
B04[P-Cl]	Presence/absence of P - Cl at topological distance 4
B04[P-Br]	Presence/absence of P - Br at topological distance 4
B04[P-I]	Presence/absence of P - I at topological distance 4
B04[P-B]	Presence/absence of P - B at topological distance 4
B04[P-Si]	Presence/absence of P - Si at topological distance 4
B04[P-X]	Presence/absence of P - X at topological distance 4
B04[F-F]	Presence/absence of F - F at topological distance 4
B04[F-Cl]	Presence/absence of F - Cl at topological distance 4
B04[F-Br]	Presence/absence of F - Br at topological distance 4
B04[F-I]	Presence/absence of F - I at topological distance 4
B04[F-B]	Presence/absence of F - B at topological distance 4
B04[F-Si]	Presence/absence of F - Si at topological distance 4
B04[F-X]	Presence/absence of F - X at topological distance 4
B04[Cl-Cl]	Presence/absence of Cl - Cl at topological distance 4
B04[Cl-Br]	Presence/absence of Cl - Br at topological distance 4
B04[Cl-I]	Presence/absence of Cl - I at topological distance 4
B04[Cl-B]	Presence/absence of Cl - B at topological distance 4
B04[Cl-Si]	Presence/absence of Cl - Si at topological distance 4
B04[Cl-X]	Presence/absence of Cl - X at topological distance 4
B04[Br-Br]	Presence/absence of Br - Br at topological distance 4
B04[Br-I]	Presence/absence of Br - I at topological distance 4
B04[Br-B]	Presence/absence of Br - B at topological distance 4
B04[Br-Si]	Presence/absence of Br - Si at topological distance 4
B04[Br-X]	Presence/absence of Br - X at topological distance 4
B04[I-I]	Presence/absence of I - I at topological distance 4
B04[I-B]	Presence/absence of I - B at topological distance 4
B04[I-Si]	Presence/absence of I - Si at topological distance 4
B04[I-X]	Presence/absence of I - X at topological distance 4
B04[B-B]	Presence/absence of B - B at topological distance 4
B04[B-Si]	Presence/absence of B - Si at topological distance 4
B04[B-X]	Presence/absence of B - X at topological distance 4
B04[Si-Si]	Presence/absence of Si - Si at topological distance 4
B04[Si-X]	Presence/absence of Si - X at topological distance 4
B04[X-X]	Presence/absence of X - X at topological distance 4
B05[C-C]	Presence/absence of C - C at topological distance 5
B05[C-N]	Presence/absence of C - N at topological distance 5
B05[C-O]	Presence/absence of C - O at topological distance 5
B05[C-S]	Presence/absence of C - S at topological distance 5
B05[C-P]	Presence/absence of C - P at topological distance 5
B05[C-F]	Presence/absence of C - F at topological distance 5
B05[C-Cl]	Presence/absence of C - Cl at topological distance 5
B05[C-Br]	Presence/absence of C - Br at topological distance 5
B05[C-I]	Presence/absence of C - I at topological distance 5
B05[C-B]	Presence/absence of C - B at topological distance 5

B05[C-Si]	Presence/absence of C - Si at topological distance 5
B05[C-X]	Presence/absence of C - X at topological distance 5
B05[N-N]	Presence/absence of N - N at topological distance 5
B05[N-O]	Presence/absence of N - O at topological distance 5
B05[N-S]	Presence/absence of N - S at topological distance 5
B05[N-P]	Presence/absence of N - P at topological distance 5
B05[N-F]	Presence/absence of N - F at topological distance 5
B05[N-Cl]	Presence/absence of N - Cl at topological distance 5
B05[N-Br]	Presence/absence of N - Br at topological distance 5
B05[N-I]	Presence/absence of N - I at topological distance 5
B05[N-B]	Presence/absence of N - B at topological distance 5
B05[N-Si]	Presence/absence of N - Si at topological distance 5
B05[N-X]	Presence/absence of N - X at topological distance 5
B05[O-O]	Presence/absence of O - O at topological distance 5
B05[O-S]	Presence/absence of O - S at topological distance 5
B05[O-P]	Presence/absence of O - P at topological distance 5
B05[O-F]	Presence/absence of O - F at topological distance 5
B05[O-Cl]	Presence/absence of O - Cl at topological distance 5
B05[O-Br]	Presence/absence of O - Br at topological distance 5
B05[O-I]	Presence/absence of O - I at topological distance 5
B05[O-B]	Presence/absence of O - B at topological distance 5
B05[O-Si]	Presence/absence of O - Si at topological distance 5
B05[O-X]	Presence/absence of O - X at topological distance 5
B05[S-S]	Presence/absence of S - S at topological distance 5
B05[S-P]	Presence/absence of S - P at topological distance 5
B05[S-F]	Presence/absence of S - F at topological distance 5
B05[S-Cl]	Presence/absence of S - Cl at topological distance 5
B05[S-Br]	Presence/absence of S - Br at topological distance 5
B05[S-I]	Presence/absence of S - I at topological distance 5
B05[S-B]	Presence/absence of S - B at topological distance 5
B05[S-Si]	Presence/absence of S - Si at topological distance 5
B05[S-X]	Presence/absence of S - X at topological distance 5
B05[P-P]	Presence/absence of P - P at topological distance 5
B05[P-F]	Presence/absence of P - F at topological distance 5
B05[P-Cl]	Presence/absence of P - Cl at topological distance 5
B05[P-Br]	Presence/absence of P - Br at topological distance 5
B05[P-I]	Presence/absence of P - I at topological distance 5
B05[P-B]	Presence/absence of P - B at topological distance 5
B05[P-Si]	Presence/absence of P - Si at topological distance 5
B05[P-X]	Presence/absence of P - X at topological distance 5
B05[F-F]	Presence/absence of F - F at topological distance 5
B05[F-Cl]	Presence/absence of F - Cl at topological distance 5
B05[F-Br]	Presence/absence of F - Br at topological distance 5
B05[F-I]	Presence/absence of F - I at topological distance 5
B05[F-B]	Presence/absence of F - B at topological distance 5
B05[F-Si]	Presence/absence of F - Si at topological distance 5
B05[F-X]	Presence/absence of F - X at topological distance 5
B05[Cl-Cl]	Presence/absence of Cl - Cl at topological distance 5
B05[Cl-Br]	Presence/absence of Cl - Br at topological distance 5

B05[Cl-I]	Presence/absence of Cl - I at topological distance 5
B05[Cl-B]	Presence/absence of Cl - B at topological distance 5
B05[Cl-Si]	Presence/absence of Cl - Si at topological distance 5
B05[Cl-X]	Presence/absence of Cl - X at topological distance 5
B05[Br-Br]	Presence/absence of Br - Br at topological distance 5
B05[Br-I]	Presence/absence of Br - I at topological distance 5
B05[Br-B]	Presence/absence of Br - B at topological distance 5
B05[Br-Si]	Presence/absence of Br - Si at topological distance 5
B05[Br-X]	Presence/absence of Br - X at topological distance 5
B05[I-I]	Presence/absence of I - I at topological distance 5
B05[I-B]	Presence/absence of I - B at topological distance 5
B05[I-Si]	Presence/absence of I - Si at topological distance 5
B05[I-X]	Presence/absence of I - X at topological distance 5
B05[B-B]	Presence/absence of B - B at topological distance 5
B05[B-Si]	Presence/absence of B - Si at topological distance 5
B05[B-X]	Presence/absence of B - X at topological distance 5
B05[Si-Si]	Presence/absence of Si - Si at topological distance 5
B05[Si-X]	Presence/absence of Si - X at topological distance 5
B05[X-X]	Presence/absence of X - X at topological distance 5
B06[C-C]	Presence/absence of C - C at topological distance 6
B06[C-N]	Presence/absence of C - N at topological distance 6
B06[C-O]	Presence/absence of C - O at topological distance 6
B06[C-S]	Presence/absence of C - S at topological distance 6
B06[C-P]	Presence/absence of C - P at topological distance 6
B06[C-F]	Presence/absence of C - F at topological distance 6
B06[C-Cl]	Presence/absence of C - Cl at topological distance 6
B06[C-Br]	Presence/absence of C - Br at topological distance 6
B06[C-I]	Presence/absence of C - I at topological distance 6
B06[C-B]	Presence/absence of C - B at topological distance 6
B06[C-Si]	Presence/absence of C - Si at topological distance 6
B06[C-X]	Presence/absence of C - X at topological distance 6
B06[N-N]	Presence/absence of N - N at topological distance 6
B06[N-O]	Presence/absence of N - O at topological distance 6
B06[N-S]	Presence/absence of N - S at topological distance 6
B06[N-P]	Presence/absence of N - P at topological distance 6
B06[N-F]	Presence/absence of N - F at topological distance 6
B06[N-Cl]	Presence/absence of N - Cl at topological distance 6
B06[N-Br]	Presence/absence of N - Br at topological distance 6
B06[N-I]	Presence/absence of N - I at topological distance 6
B06[N-B]	Presence/absence of N - B at topological distance 6
B06[N-Si]	Presence/absence of N - Si at topological distance 6
B06[N-X]	Presence/absence of N - X at topological distance 6
B06[O-O]	Presence/absence of O - O at topological distance 6
B06[O-S]	Presence/absence of O - S at topological distance 6
B06[O-P]	Presence/absence of O - P at topological distance 6
B06[O-F]	Presence/absence of O - Cl at topological distance 6
B06[O-Cl]	Presence/absence of O - Br at topological distance 6
B06[O-Br]	Presence/absence of O - I at topological distance 6
B06[O-I]	Presence/absence of O - B at topological distance 6

B06[O-B]	Presence/absence of O - Si at topological distance 6
B06[O-Si]	Presence/absence of O - X at topological distance 6
B06[O-X]	Presence/absence of S - S at topological distance 6
B06[S-S]	Presence/absence of S - P at topological distance 6
B06[S-P]	Presence/absence of S - F at topological distance 6
B06[S-F]	Presence/absence of S - Cl at topological distance 6
B06[S-Cl]	Presence/absence of S - Br at topological distance 6
B06[S-Br]	Presence/absence of S - I at topological distance 6
B06[S-I]	Presence/absence of S - B at topological distance 6
B06[S-B]	Presence/absence of S - Si at topological distance 6
B06[S-Si]	Presence/absence of S - X at topological distance 6
B06[S-X]	Presence/absence of P - P at topological distance 6
B06[P-P]	Presence/absence of P - F at topological distance 6
B06[P-F]	Presence/absence of P - Cl at topological distance 6
B06[P-Cl]	Presence/absence of P - Br at topological distance 6
B06[P-Br]	Presence/absence of P - I at topological distance 6
B06[P-I]	Presence/absence of P - B at topological distance 6
B06[P-B]	Presence/absence of P - Si at topological distance 6
B06[P-Si]	Presence/absence of P - X at topological distance 6
B06[P-X]	Presence/absence of F - F at topological distance 6
B06[F-F]	Presence/absence of F - Cl at topological distance 6
B06[F-Cl]	Presence/absence of F - Br at topological distance 6
B06[F-Br]	Presence/absence of F - I at topological distance 6
B06[F-I]	Presence/absence of F - B at topological distance 6
B06[F-B]	Presence/absence of F - Si at topological distance 6
B06[F-Si]	Presence/absence of F - X at topological distance 6
B06[F-X]	Presence/absence of Cl - Cl at topological distance 6
B06[Cl-Cl]	Presence/absence of Cl - Br at topological distance 6
B06[Cl-Br]	Presence/absence of Cl - I at topological distance 6
B06[Cl-I]	Presence/absence of Cl - B at topological distance 6
B06[Cl-B]	Presence/absence of Cl - Si at topological distance 6
B06[Cl-Si]	Presence/absence of Cl - X at topological distance 6
B06[Cl-X]	Presence/absence of Br - Br at topological distance 6
B06[Br-Br]	Presence/absence of Br - I at topological distance 6
B06[Br-I]	Presence/absence of Br - B at topological distance 6
B06[Br-B]	Presence/absence of Br - Si at topological distance 6
B06[Br-Si]	Presence/absence of Br - X at topological distance 6
B06[Br-X]	Presence/absence of I - I at topological distance 6
B06[I-I]	Presence/absence of I - B at topological distance 6
B06[I-B]	Presence/absence of I - Si at topological distance 6
B06[I-Si]	Presence/absence of I - X at topological distance 6
B06[I-X]	Presence/absence of B - B at topological distance 6
B06[B-B]	Presence/absence of B - Si at topological distance 6
B06[B-Si]	Presence/absence of B - X at topological distance 6
B06[B-X]	Presence/absence of Si - Si at topological distance 6
B06[Si-Si]	Presence/absence of Si - X at topological distance 6
B06[Si-X]	Presence/absence of X - X at topological distance 6
B06[X-X]	Presence/absence of C - C at topological distance 7
B07[C-C]	Presence/absence of C - N at topological distance 7

B07[C-N]	Presence/absence of C - O at topological distance 7
B07[C-O]	Presence/absence of C - S at topological distance 7
B07[C-S]	Presence/absence of C - P at topological distance 7
B07[C-P]	Presence/absence of C - F at topological distance 7
B07[C-F]	Presence/absence of C - Cl at topological distance 7
B07[C-Cl]	Presence/absence of C - Br at topological distance 7
B07[C-Br]	Presence/absence of C - I at topological distance 7
B07[C-I]	Presence/absence of C - B at topological distance 7
B07[C-B]	Presence/absence of C - Si at topological distance 7
B07[C-Si]	Presence/absence of C - X at topological distance 7
B07[C-X]	Presence/absence of N - N at topological distance 7
B07[N-N]	Presence/absence of N - O at topological distance 7
B07[N-O]	Presence/absence of N - S at topological distance 7
B07[N-S]	Presence/absence of N - P at topological distance 7
B07[N-P]	Presence/absence of N - F at topological distance 7
B07[N-F]	Presence/absence of N - Cl at topological distance 7
B07[N-Cl]	Presence/absence of N - Br at topological distance 7
B07[N-Br]	Presence/absence of N - I at topological distance 7
B07[N-I]	Presence/absence of N - B at topological distance 7
B07[N-B]	Presence/absence of N - Si at topological distance 7
B07[N-Si]	Presence/absence of N - X at topological distance 7
B07[N-X]	Presence/absence of O - O at topological distance 7
B07[O-O]	Presence/absence of O - S at topological distance 7
B07[O-S]	Presence/absence of O - P at topological distance 7
B07[O-P]	Presence/absence of O - F at topological distance 7
B07[O-F]	Presence/absence of O - Cl at topological distance 7
B07[O-Cl]	Presence/absence of O - Br at topological distance 7
B07[O-Br]	Presence/absence of O - I at topological distance 7
B07[O-I]	Presence/absence of O - B at topological distance 7
B07[O-B]	Presence/absence of O - Si at topological distance 7
B07[O-Si]	Presence/absence of O - X at topological distance 7
B07[O-X]	Presence/absence of S - S at topological distance 7
B07[S-S]	Presence/absence of S - P at topological distance 7
B07[S-P]	Presence/absence of S - F at topological distance 7
B07[S-F]	Presence/absence of S - Cl at topological distance 7
B07[S-Cl]	Presence/absence of S - Br at topological distance 7
B07[S-Br]	Presence/absence of S - I at topological distance 7
B07[S-I]	Presence/absence of S - B at topological distance 7
B07[S-B]	Presence/absence of S - Si at topological distance 7
B07[S-Si]	Presence/absence of S - X at topological distance 7
B07[S-X]	Presence/absence of P - P at topological distance 7
B07[P-P]	Presence/absence of P - F at topological distance 7
B07[P-F]	Presence/absence of P - Cl at topological distance 7
B07[P-Cl]	Presence/absence of O - Cl at topological distance 6
B07[P-Br]	Presence/absence of O - Br at topological distance 6
B07[P-I]	Presence/absence of O - I at topological distance 6
B07[P-B]	Presence/absence of P - Br at topological distance 7
B07[P-Si]	Presence/absence of P - I at topological distance 7

B07[P-X]	Presence/absence of P - B at topological distance 7
B07[F-F]	Presence/absence of P - Si at topological distance 7
B07[F-Cl]	Presence/absence of P - X at topological distance 7
B07[F-Br]	Presence/absence of F - F at topological distance 7
B07[F-I]	Presence/absence of F - Cl at topological distance 7
B07[F-B]	Presence/absence of F - Br at topological distance 7
B07[F-Si]	Presence/absence of F - I at topological distance 7
B07[F-X]	Presence/absence of F - B at topological distance 7
B07[Cl-Cl]	Presence/absence of F - Si at topological distance 7
B07[Cl-Br]	Presence/absence of F - X at topological distance 7
B07[Cl-I]	Presence/absence of Cl - Cl at topological distance 7
B07[Cl-B]	Presence/absence of Cl - Br at topological distance 7
B07[Cl-Si]	Presence/absence of Cl - I at topological distance 7
B07[Cl-X]	Presence/absence of Cl - B at topological distance 7
B07[Br-Br]	Presence/absence of Cl - Si at topological distance 7
B07[Br-I]	Presence/absence of Cl - X at topological distance 7
B07[Br-B]	Presence/absence of Br - Br at topological distance 7
B07[Br-Si]	Presence/absence of Br - I at topological distance 7
B07[Br-X]	Presence/absence of Br - B at topological distance 7
B07[I-I]	Presence/absence of Br - Si at topological distance 7
B07[I-B]	Presence/absence of Br - X at topological distance 7
B07[I-Si]	Presence/absence of I - I at topological distance 7
B07[I-X]	Presence/absence of I - B at topological distance 7
B07[B-B]	Presence/absence of I - Si at topological distance 7
B07[B-Si]	Presence/absence of I - X at topological distance 7
B07[B-X]	Presence/absence of B - B at topological distance 7
B07[Si-Si]	Presence/absence of B - Si at topological distance 7
B07[Si-X]	Presence/absence of B - X at topological distance 7

Table 3.10 . The details of molecular properties descriptors with their symbols.

Descriptor Symbol	Description
Uc	unsaturation count
Ui	unsaturation index
Hy	hydrophilic factor
AMR	Ghose-Crippen molar refractivity
TPSA(NO)	topological polar surface area using N,O polar contributions
TPSA(Tot)	topological polar surface area using N,O,S,P polar contributions
MLOGP	Moriguchi octanol-water partition coeff. (logP)
MLOGP2	squared Moriguchi octanol-water partition coeff. (logP ²)
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)
ALOGP2	squared Ghose-Crippen octanol-water partition coeff. (logP ²)
SAtot	total surface area from P_VSA-like descriptors
SAacc	surface area of acceptor atoms from P_VSA-like descriptors
SAdon	surface area of donor atoms from P_VSA-like descriptors
Vx	McGowan volume
VvdwMG	van der Waals volume from McGowan volume
VvdwZAZ	van der Waals volume from Zhao-Abraham-Zissimos equation
PDI	packing density index
BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)
BLTD48	Verhaar Daphnia base-line toxicity from MLOGP (mmol/l)
BLTA96	Verhaar Algae base-line toxicity from MLOGP (mmol/l)

Table 3.11. Definitions of different basic ETA parameters

Descriptor Symbol	Description
N_v	Vertex count (excluding hydrogen)
N	Total number of atoms including hydrogens
$\Sigma\alpha$	Sum of α values of all non-hydrogen vertices of a molecule
$[\Sigma\alpha]_P$	Sum of α values of all non-hydrogen vertices each of which is joined to only one other non-hydrogen vertex of the molecule
$[\Sigma\alpha]_X$	Sum of α values of all non-hydrogen vertices each of which is joined to four other non-hydrogen vertex of the molecule
$[\Sigma\alpha]_Y$	Sum of α values of all non-hydrogen vertices each of which is joined to three other non-hydrogen vertex of the molecule
ϵ	Measure of electronegativity
η	The composite ETA index
η_R	The composite index for the reference alkane
$\Sigma\beta'_s$	Sum of β'_s values of all non-hydrogen vertices of a molecule; $\Sigma\beta'_s$ is defined as $[\Sigma\beta'_s]/N_v$.
$\Sigma\beta'_{ns}$	Sum of β'_{ns} values of all non-hydrogen vertices of a molecule; $\Sigma\beta'_{ns}$ is defined as $[\Sigma\beta'_{ns}]/N_v$.
$[\eta'_F]$	Total functionality contribution
$[\eta'_F]_X$	Functionality contribution for the atom/group/fragment X (e.g., -F, -CN, -OH etc.)

Table 3.12. An overview of the newly derived more novel indices under the ETA formalism

Descriptor Symbol	Description
$\Delta\alpha_A = \left\langle \frac{\sum \alpha - [\sum \alpha]_R}{N_V} \right\rangle$	A measure of count of non-hydrogen heteroatoms [N _V stands for total number of atoms excluding hydrogens]
$\Delta\alpha_B = \left\langle \frac{[\sum \alpha]_R - \sum \alpha}{N_V} \right\rangle$	A measure of count of hydrogen bond acceptor atoms and/or polar surface area
$\varepsilon_1 = \frac{\sum \varepsilon}{N}$	A measure of electronegative atom count [N stands for total number of atoms including hydrogens]
$\varepsilon_2 = \frac{\sum \varepsilon_{EH}}{N_V}$	A measure of electronegative atom count [EH stands for <i>excluding hydrogens</i>]
$\varepsilon_3 = \frac{[\sum \varepsilon]_R}{N_R}$	[R stands for reference alkane]
$\varepsilon_4 = \frac{[\sum \varepsilon]_{SS}}{N_{SS}}$	[SS stands for saturated carbon skeleton]
$\varepsilon_5 = \frac{\sum \varepsilon_{EH} + \sum \varepsilon_{XH}}{N_V + N_{XH}}$	[XH stands for those hydrogens which are connected to a heteroatom]
$\Delta\varepsilon_A = \varepsilon_1 - \varepsilon_3$	A measure of contribution of unsaturation and electronegative atom count
$\Delta\varepsilon_B = \varepsilon_1 - \varepsilon_4$	A measure of contribution of unsaturation
$\Delta\varepsilon_C = \varepsilon_3 - \varepsilon_4$	A measure of contribution of electronegativity
$\Delta\varepsilon_D = \varepsilon_2 - \varepsilon_5$	A measure of contribution of hydrogen bond donor atoms
$\psi = \frac{\alpha}{\varepsilon}$	A measure of hydrogen bonding propensity of the atoms
$\psi_1 = \frac{\sum \alpha}{[\sum \varepsilon]_{EH}} = \frac{\sum \alpha / N_V}{\varepsilon_2}$	A measure of hydrogen bonding propensity of the molecules and/or polar surface area
$\Delta\psi_A = \langle 0.714 - \psi_1 \rangle$	A measure of hydrogen bonding propensity of the molecules
$\Delta\psi_B = \langle \psi_1 - 0.714 \rangle$	A measure of hydrogen bonding propensity of the molecules
$\Delta\beta = \sum \beta_{ns} - \sum \beta_s$	A measure of relative unsaturation content
$\Delta\beta' = \frac{\Delta\beta}{N_V}$	A measure of relative unsaturation content
$\sum \beta_{ns(\delta)}$	A measure of lone electrons entering into resonance
$\sum \beta'_{ns(\delta)} = \frac{\sum \beta_{ns(\delta)}}{N_V}$	A measure of lone electrons entering into resonance

3.3.2 Dataset division

Data set division is a very important step for QSPR modeling. The dataset was divided into training and test sets using different data division methods such as Euclidean distance (diversity-based), Kennard Stone, k -means clustering and sorted response. The splitting of the dataset was done in such a way that the training set would capture all the information of the dataset enabling correct predictions for the test set compounds from the corresponding QSPR model. The algorithm of division method is discussed below.

(a) Kennard Stone method (Euclidean distance based): The optimal division of dataset into training and independent test subset is an important and critical step in the QSAR modeling analysis. The steps involved in the selection of training set based on Euclidean-based KS algorithm are as follows (Kennard and Stone, 1969; Snarey et al., 1997)

- The first two compounds of training set are selected by choosing two compounds that are quite farthest apart in terms of Euclidean distance. The normalized mean distance for all the compounds are calculated by following equations:

$$d_{ij} = ||X_i - X_j|| = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (\text{i})$$

$$\bar{d}_i = \sum_{k=1}^m d_{ij} / n - 1 \quad (\text{ii})$$

$$\text{Mean distance}_{(\text{Normalised})} = \frac{(\text{MeanDistance} - \text{Min_MeanDistance})}{(\text{Max_MeanDistance} - \text{Min_MeanDistance})} \quad (\text{iii})$$

where,

d_{ij} = The distance score between two compounds

\bar{d}_i = Mean distance

- Find the compound which has the maximum dissimilarity (maximum minimum distance) from each of the previously selected chemicals and place this chemical in the training set.
- Repeat step 2 until the desired number of chemicals have been added to the training set.

The remaining chemicals were placed in the test set.

(b) k -Means clustering: k -Means is one of the simplest unsupervised learning algorithms (<https://sites.google.com/site/dataclusteringalgorithms>) that solves the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a

certain number of clusters (assume k clusters) fixed *a priori*. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. This algorithm aims at minimizing an objective function known as squared error function given by:

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \quad (3.1)$$

Where, $||x_i - v_j||$ is the Euclidean distance between x_i and v_j

c_i is the number of data points in the i^{th} cluster

' c ' is the number of cluster centers.

The steps involved in the k -means clustering method are as follows:

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

- i) Randomly select ' c ' cluster centers.
- ii) Calculate the distance between each data point and cluster centers.
- iii) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- iv) Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_j \quad (3.2)$$

where, ' c_i ' represents the number of data points in i^{th} cluster.

- v) Recalculate the distance between each data point and new obtained cluster centers.
- vi) If no data point was reassigned then stop, otherwise repeat from step iii).

3.3.3. Selections of variables using multilayered strategy

Variable selection is another important step in QSPR modeling. However, when modelling a particular property or biological activity, it is reasonable to assume that only a small number of descriptors is actually correlated to the experimental response and is, therefore, relevant for building the mathematical model of interest. The subsequent design of a quantitative structure-activity relationship (QSAR) model (regression or discriminant) would lead to poor

performance if little significant features are selected. As a consequence, a key step is the selection of the optimal subset of variables (i.e. molecular descriptors) for the development of the model. In the present thesis work, we have used Stepwise regression technique was used for the selection of optimal descriptors.

3.3.4. Model development

Training set molecules are employed for model development purposes. Various statistical tools required to develop a QSPR model are summarized as follows;

(a) Stepwise regression: Stepwise regression is a type of multiple linear regression equation made step by step which is altered by adding or removing a predictor variable. Forward selection and backward elimination are two parts of stepwise regression method. Forward selection starts with no variable and then ‘statistically significant’ variables are included one by one. In case of backward elimination, initially, all the candidate variables are selected and then deleting statistically insignificant variables one by one. The objective function of the selection in stepwise regression may be “F-for-inclusion”, also known as “steeping criteria”. The F-value is square of t value of incoming variable which signify the regression coefficient. In stepwise regression process, a multiple term linear regression equation is built up using a “steeping criteria” also known as “Fisher criteria”. In this present study, we have fixed the “steeping criteria” or “Fisher criteria” $F=4$ to enter and $F=3.9$ to remove (Das et al., 2017) because at this value of the F-criterion, the descriptors are considered to be significant at the 95% confidence level. In this study, the stepwise regression has been performed using initial pool of descriptors and selected the model descriptors and kept aside. After removing the selected descriptors from the initial pool, stepwise regression was done again and selected the model descriptors and so on. The variable selection approach using stepwise regression technique was applied for the dataset containing 69 organic contaminants. In this way, we have selected 47 descriptors (reduced pool) for the dataset containing 69 organic contaminants.

(b) Best subset selection: The best descriptor combination out of total descriptor sets is selected by best subset selection software developed in our laboratory

([http://teqip.jdvu.ac.in/QSAR Tools/DTCLab](http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) by evaluating the possible combinations of descriptors. Different statistical parameters such as R^2 , Q^2 , Q^2_{F1} , Q^2_{F2} are also calculated by this method. In this work, we have run best subset selection using the reduced pool of descriptors obtained after applying different variable selection approaches and selected best five multiple linear regression models based on MAE based criteria (Roy et al., 2016).

(c) Intelligent consensus predictor (ICP) (Roy et al., 2018): This software was used to judge the performance of consensus predictions and compares them with the prediction quality obtained from the individual (MLR) models based on MAE based criteria (95%). It is obvious that a single model might not be equally useful in prediction for the whole test set compounds which means one QSAR model may be the best model for prediction of a test compound while other model may be the best predictor for another test compounds. For this reason, we have selected five models (M1-M5) and performed consensus prediction using “Intelligent consensus predictor” tool to explore whether the quality of predictions of test set compounds can be enhanced through an “intelligent” selection of multiple MLR models.

(d) Partial least squares (PLS): Although MLR is relatively simpler and widely used in the QSAR literature, MLR has several shortcomings some of them are mentioned here: (a) it cannot handle strongly correlated predictor variables; (b) it assumes that the predictor variables are noise-free, which is not true in most of the cases; (c) it cannot handle missing values in predictor variables; (d) MLR requires that number of data points much greater than number of predictor variables, which may be impractical in many circumstances leading to suboptimal predictions. These shortcomings of MLR can be obviated by using latent variable modeling (LVM) which is more generalized and robust than MLR (Wold et al., 2001; Geladi and Kowalski, 1986; Wold et al., 1987) LVM uses a dimension reduction technique which may be unsupervised (like principal component analysis) or supervised (like partial least squares). Partial least squares (PLS) technique generalize and combines features of MLR and principle component analysis (PCA) that enables analysis with collinear, noisy, and numerous variables. It allows modeling of several response variables (Y), which is not possible in MLR. The PLS technique finds secondary variables also known as latent variables (LVs) which captures the information of the actual variables. However, the final PLS equation is presented using the actual variables. PLS technique also avoids the

overfitting problem observed in MLR, when the number of predictor variables is high. It may be noted that when the number of LVs equals the number of actual predictor variables in the equation, the PLS model becomes an MLR model. Hence, it is necessary to keep the number of LVs at least one less than the total number of descriptors while running a PLS analysis. The PLS regression is based on the Non-linear iterative Partial Least Squares (NIPALS) algorithm as described in the literature (Wold et al., 2001). In the present study, we have performed PLS using MINITAB (version 14.13) software (<http://www.minitab.com/en-US/default.aspx>).

3.3.5. Computation of different statistical metrics for assessing model quality

3.3.5.1. Quality measures in fitting of a QSPR model

(a) *Squared correlation coefficient (R^2)*: This parameter is termed as the determination coefficient or squared correlation coefficient. The squared correlation coefficient of a model can be obtained from the following equation Eq. 3.3:

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{cal})^2}{\sum (Y_{obs} - Y_{train})^2} \quad (3.3)$$

The R^2 statistic represents the ratio of the regression variance to the original variance where the former is determined using the original variance minus the variance around the line of regression. The R^2 bears a value between zero (no correlation) to one (perfect correlation). A model possessing a value of R^2 more than 0.8 can be considered to elicit acceptable correlation while the quality enhancing with the increasing value of R^2 until it reaches a maximum value of unity (which is unusual in real cases). Y_{obs} and Y_{calc} are the respective observed and calculated values of the response variable and is their mean value. R^2 gives a measure of explained variance. Each additional X variable added to a model increases R^2 . The prime drawbacks of the R^2 parameter lies in the facts that it does not provide any information on whether: (i) the independent variables are a true cause of the changes in the dependent variable, (ii) the correct regression was used, (iii) the most appropriate set of independent variables has been chosen, (iv) the model might be improved by using transformed versions of the existing set of independent variables and (v) whether any collinearities exists in the data or not.

(b) **Adjusted R^2 or R_a^2 :**

$$R_a^2 = \frac{(n-1) \times R^2 - P}{n-p-1} \quad (3.4)$$

Adjusted R_a^2 is a modified version of the determination coefficient and is also known as the explained variance. *The R_a^2 parameter* incorporates the information of the number of samples and the independent variables used in model, and can be defined as follows (Snedecor and Cochran, 1967). Here, R^2 is the determination coefficient of a QSAR model comprising p number of predictor variables and n number of samples. Hence, instead of using only the initial observed (i.e., experimental) and final predicted response values, R_a^2 considers information on the model history in terms of the number of descriptors and number of chemicals used to develop the model (i.e., training set chemicals). The R_a^2 penalizes the R^2 value of a model containing too many independent variables compared to the total number of compounds. The R_a^2 improves only if the addition of a new term enhances the model quality avoiding chance. The R_a^2 value usually is less than the corresponding R^2 value.

(c) **Standard error of estimate (s):** The error in the estimation of individual activity values of the compounds under study using the MLR method can be quantified based on their residual data. The standard error of estimate (SEE or s) for the residuals is calculated by taking the root-mean square of the residuals. The standard error of the estimate is a measure of the accuracy of fitting. Lower values of SEE correspond to improved model acceptability.

$$s = \sqrt{\frac{\sum (Y_{obs} - Y_{calc})^2}{n-p-1}} \quad (3.5)$$

In Eq. 3.26, Y_{obs} and Y_{calc} are the actual and estimated scores respectively, while n is the number of scores and p is the number of descriptors.

(d) **F-value:** F-value is called the variance ratio and is defined

$$F = \frac{\frac{\sum (Y_{calc} - \bar{Y})^2}{p}}{\frac{\sum (Y_{obs} - Y_{calc})^2}{n - p - 1}} \quad (3.6)$$

3.3.5.2 Validation strategies

Both internal and external validation statistics constitute the primary methods for validation of the developed QSPR models. Both the methods have been widely used by different groups of researchers for assessing the predictive ability of the developed model. Another method employs fitting of the dependent X matrix to randomized response parameters. Several metrics are used to check the predictivity of the QSPR models. For the validation of QSPR models, three strategies are primarily adopted: (i) internal validation using the training set molecules and (ii) external validation based on the test set compounds.

3.3.5.2.1 Validation metrics for Training set

(a) **Q² or Q²_{int}**: The models developed from the training set by using stepwise regression or genetic methods have been subjected to internal validation by means of calculating leave-one-out *cross-validation* R² (Q²) and *predicted residual sum of squares* (PRESS) (Wold et al., 1995) and the acceptable models have been further processed for the prediction of toxicity and/or property of the test set compounds. Cross-validated correlation coefficient R² (LOO-Q²) is calculated according to the formula:

$$Q^2 = 1 - \frac{\sum (Y_{obs(training)} - Y_{pred(training)})^2}{\sum (Y_{obs(training)} - \bar{Y}_{training})^2} \quad (3.7)$$

Here, $Y_{obs(training)}$, $Y_{pred(training)}$, and $\bar{Y}_{training}$ are the observed, predicted and the average value of the response variable of the training set. In this technique, one compound is omitted from the data set at random in each cycle and then model is built using the rest of the compounds. The model thus formed in this way is used for the prediction of activity of the omitted compound. The process is iterated until all the compounds are eliminated once. On the basis of the predicting ability of the model, the cross-validated R² (Q²) for the model is determined. Acceptable value of Q² is 0.5

with a maximum value of 1.0 and hence more the value is closer to 1, more will be the internal predictivity of the model.

(b) $r_m^2(\text{LOO})$: It was shown that (Roy et al., 2009) squared cross-validated correlation coefficient alone might not indicate the true predictive capability of a model and hence a modified r^2 [$r_m^2(\text{LOO})$] term was used to indicate the leave-one-out prediction capacity of the model for the training set compounds. The parameter $r_m^2(\text{LOO})$ is defined as:

$$r_{m(\text{LOO})}^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (3.8)$$

where r^2 and r_0^2 are the squared correlation coefficients between the observed and LOO predicted values of the training set compounds with and without intercept respectively. The value of $r_m^2(\text{LOO})$ should be greater than 0.5 for an acceptable model.

(c) Root mean square error in prediction for training set ($rmsep_{\text{int}}$): This parameter suggests that it is possible to determine the internal predictive ability of the training set compounds simply by taking the square root of the squared difference between the observed and predicted response value divided by the number of compounds in the training set (Consonni et al., 2010). Mathematically:

$$rmsep_{\text{int}} = \sqrt{\frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{n_{\text{int}}}} \quad (3.9)$$

where, n_{int} is the number of compounds present in the training set and Y_{obs} and Y_{pred} corresponds to the corresponding observed and LOO predicted response value. It should have a minimum value.

(d) Golbraikh and Tropsha criteria

Golbraikh and Tropsha (Tropsha, 2010) defined a set of criteria to be followed in order to ascertain the external predictive potential of a QSAR model. As we can see that the basic objective of model validation is to determine how close the observed i.e., experimental values are to the corresponding predicted ones. Hence, the simple correlation coefficient between the observed (y) and predicted (\hat{y}) response apparently should give a value of 1 in an ideal case. In this situation, if a regression line is drawn all the points will be located on the line which passes through the origin point (0, 0) in a Cartesian plane. However, in real cases, deviation occurs and the best fitted line poses a definite intercept value. It may be here noted that the plots of

experimental versus fitted or fitted versus experimental response are not equivalent (Besalu et al., 2007). Golbraikh and Tropsha (Tropsha, 2010) proposed that regression of observed (y) against predicted (\hat{y}) or predicted (\hat{y}) against observed (y) response through the origin must be determined and the corresponding slopes k or k' of the regression lines should be close to unity. This process is known as regression through origin (RTO) method, where a regression line is forcefully passed through the origin point (0, 0) and the corresponding regression lines can be presented as $y^{r_0} = k\hat{y}$ and $\hat{y}^{r_0} = k'y$. The slopes k and k' can be defined as follows:

$$k = \frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i^2} \quad (3.10) \quad \text{and} \quad k' = \frac{\sum y_i \hat{y}_i}{\sum y_i^2} \quad (3.11)$$

Golbraikh and Tropsha calculated the determination coefficient values r_0^2 and $r_0'^2$ of the regression lines passing through origin (y against \hat{y} or \hat{y} against y) and, argued that these values should be close to the value of the normal R^2 of the model in case of good predictivity. The r_0^2 and $r_0'^2$ represent the squared correlation coefficient between the observed and predicted response values with and without intercept respectively and can be defined as follows:

$$r_0^2 = 1 - \frac{\sum (\hat{y}_i - y_i^{r_0})^2}{\sum (\hat{y}_i - \bar{\hat{y}})^2} \quad (3.12) \quad r_0'^2 = 1 - \frac{\sum (y_i - \hat{y}_i^{r_0'})^2}{\sum (\hat{y}_i - \bar{\hat{y}})^2} \quad (3.13)$$

Here, \bar{y} and $\bar{\hat{y}}$ refers to the respective mean values of the observed and predicted response data. A set of conditions for model acceptability was proposed by Golbraikh and Tropsha and are summarized below.

- a) $Q_{(LOO)}^2 > 0.5$
- b) $R_{test}^2 > 0.6$
- c) $\frac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \leq k \leq 1.15$ or $\frac{r^2 - r_0'^2}{r^2} < 0.1$ and $0.85 \leq k' \leq 1.15$
- d) $|r_0^2 - r_0'^2| < 0.3$

Here, $Q^2_{(LOO)}$ is for the training set only while rest of the parameters correspond to test set chemicals.

(e) The r_m^2 metrics: Using the concept of regression through origin approach, Roy et. al (2009) introduced a new parameter r_m^2 or modified r^2 that penalizes the R^2 value of a model with respect to an ideal condition (Roy and Mitra, 2012). The r_m^2 metric can be defined as follows:

$$r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right) \quad (3.14) \quad r_m'^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0'^2)}\right) \quad (3.15)$$

where, r^2 is the squared correlation coefficient value between observed and predicted response values, and r_0^2 and $r_0'^2$ are the respective squared correlation coefficients when the regression line is passed through the origin by interchanging the axes. Roy and co-workers (Ojha et al., 2011; Roy et al., 2013) further defined average and difference of the two r_m^2 metric values (i.e., r_m^2 and $r_m'^2$) to be used as the acceptable criteria to judge the predictive ability of a model as follows.

$$\frac{r_m^2 + r_m'^2}{2} > 0.5 \quad (3.16) \quad \Delta r_m^2 = |r_m^2 - r_m'^2| < 0.2 \quad (3.17)$$

The r_m^2 metrics can not only be computed for the test set compounds ($r_m^2_{(\text{test})}$) to judge external predictivity, but it can also be used to determine the internal predictivity of the model using the training set. In the latter case, leave-one-out predicted values ($r_m^2_{(\text{LOO})}$) of the training set observations are used against their observed response. Furthermore, Roy et al. (Roy and Mitra, 2012) also reported the use of the r_m^2 metric in characterizing the overall predictive capability of the model by using leave-one-out predicted values for the training set and equation (i.e., model) based predicted values for the test set together against their corresponding observed response ($r_m^2_{(\text{overall})}$). Later, a rank based r_m^2 (Roy and Kabir, 2012) as well as a scaled Roy et al., 2013 version of the r_m^2 metric was introduced by the same authors' group and these have been used in this present study.

(f) MAE based criteria: In a recent study, Roy et al. (Roy et al., 2016) have shown that the conventional Q^2 based external validation metrics ($Q^2_{\text{ext}(F1)}$, $Q^2_{\text{ext}(F2)}$, $Q^2_{\text{ext}(F3)}$) often provide biased judgment of model predictivity since such metrics are influenced by factors such as response range and distribution of data. Here, the authors have defined a set of criteria using simple 'mean absolute error' (MAE) and the corresponding standard deviation (σ) measure of the predicted residuals to judge the external predictivity of the models. Note that,

$MAE = \frac{1}{n} \times \sum |Y_{obs} - Y_{pred}|$, where Y_{obs} and Y_{pred} are the respective observed and predicted response values of the test set comprising n number of compounds. The response range of training set compounds has been employed here to define the threshold values. Furthermore, the authors have proposed application of the ‘MAE based criteria’ on 95% of the test set data by removing 5% data with high predicted residual values precluding the possibility of any outlier prediction. The criteria are described below.

(i) **Good predictions:** The criteria for good predictions are as follows:

$$MAE \leq 0.1 \times \text{training set range AND } (MAE + 3\sigma) \leq 0.2 \times \text{training set range}$$

In simpler terms, an error of 10% of the training set range should be acceptable while an error value more than 20% of the training set range may be considered as high.

(ii) **Bad predictions:** The predictions considered as bad can be defined using the following criteria:

$$MAE > 0.15 \times \text{training set range OR } (MAE + 3\sigma) > 0.25 \times \text{training set range}$$

Here, a value of MAE more than 15% of the training set range is considered high while an error more than 25% of the training set range is judged as very high.

The predictions which do not fall under either of the above two conditions may be considered as of moderate quality. The above criteria should be applied for judging the quality of test set predictions when the number of data points is at least 10 (statistical reliability) and there is no systematic error in model predictions (statistical applicability).

3.3.5.2.2. Validation metric for Test set

(a) R^2_{pred} or $Q^2_{ext(F1)}$

After the prediction of toxicity and/or property of the test set compounds, this parameter was calculated. It can be defined as:

$$R^2_{pred} = Q^2_{ext(F1)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \bar{Y}_{training})^2} \quad (3.18)$$

where, $Y_{obs(test)}$ is the observed activity of the test set compounds, $Y_{pred(test)}$ is the predicted activity of the test set compounds and $\bar{Y}_{training}$ corresponds to the mean of observed activity of the

training set compounds. R^2_{pred} value for an acceptable model should be greater than 0.5 (maximum value 1).

(b) $Q^2_{ext(F2)}$: This function as a metric for external set validation was described in the paper of Hawkins (Hawkins, 2004) and can be calculated as:

$$Q^2_{ext(F2)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \bar{Y}_{test})^2} \quad (3.19)$$

The only notable difference from $Q^2_{ext(F1)}$ is that, in Equation 3.77 the average value of external or test set is used in the denominator instead the internal or training set average value.

Both these functions $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$ were compared and discussed by Schuurmann *et al.* (Schuurmann *et al.*, 2008).

(c) $Q^2_{ext(F3)}$: This function was described by Consonni *et al.* (Consonni *et al.*, 2009) and is defined as:

$$Q^2_{ext(F3)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2 / n_{test}}{\sum (Y_{obs(training)} - \bar{Y}_{training})^2 / n_{training}} \quad (3.20)$$

Since the terms for summation in the numerator deals totally with the test set values and the denominator with training set values, division with n_{test} and $n_{training}$ of the numerator and denominator summation expression respectively makes the two squares comparable. The threshold value of acceptance for all the three parameters $Q^2_{ext(F1)}$, $Q^2_{ext(F2)}$ and $Q^2_{ext(F3)}$ is 0.5.

(d) $r^2_{m(test)}$: For test set compounds, $r^2_{m(test)}$ has been determined which penalizes a model for large differences between observed and predicted values of the test set compounds. The formula is:

$$r^2_{m(test)} = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (3.21)$$

r^2 and r_0^2 are the squared correlation coefficients between the observed and predicted values of the test set compounds with and without intercept respectively.

(e) **Root mean square error in prediction for test set ($rmsep_{ext}$):** We have also calculated the $rmsep_{ext}$ parameter for the evaluation of external predictive ability of a model as follows (Consonni et al., 2010):

$$rmsep_{ext} = \sqrt{\frac{\sum (Y_{obs} - Y_{pred})^2}{n_{ext}}} \quad (3.22)$$

where, n_{ext} represents the number of training set compounds, and Y_{obs} and Y_{pred} corresponds to the observed and predicted activity of the test set compounds respectively. It should have a minimum value. The $rmsep$ value for test and training set depends on the scale of the response activity and therefore comparison makes no sense when a model is compared to another modeling a different activity.

3.3.5.2.3. Validation metric for overall set

For the purpose of determination of an overall validation strength of a model, we have calculated the overall r_m^2 metric between the observed toxicity and/or property value of a dataset and the calculated and predicted value of the training and test set respectively. The formula is:

$$r_{m(overall)}^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (3.23)$$

The r_m^2 metric has been developed by the present authors' group and has been extensively used by them (Mitra et al., 2009; Roy and Roy, 2009; Roy and Popelier, 2009).

3.3.5.3. Y-randomization

The relationships between the response variable and the descriptors can be checked for further statistical significance by randomization test (Y-randomization) of the models. The method can be executed in two ways namely:

- i) Process randomization and
- ii) Model randomization

In process randomization, random scrambling of the dependent response variables is performed accompanied with fresh selection of variables from the whole descriptor matrix and in model randomization scrambling or randomization of the response variable is performed within the descriptors present in an existing model. We have performed process as well as the model

randomization of the genetic models. A parameter was proposed by Roy and Paul (Roy and Paul, 2009) named R_p^2 that penalises the model R^2 for a small difference between squared mean correlation coefficient (R_r^2) of randomized models and squared correlation coefficient (R^2) of the non-randomized model and was defined as:

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \quad (3.24)$$

and the acceptable value of R_p^2 was proposed to be greater than or at least equal to 0.5. Later a correction for this parameter has been suggested by Todeschini and the rebuilt formula is as follows:

$${}^c R_p^2 = R \times \sqrt{R^2 - R_r^2} \quad (3.25)$$

We have used the new parameter ${}^c R_p^2$ which should be above 0.5 for a good model.

3.3.5.4. Applicability Domain

“The applicability domain of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The applicability domain of a (Q)SAR should be described in terms of the most relevant parameters, i.e., usually those that are descriptors of the model. Ideally the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation.” The AD of QSAR model is characterized by the molecular properties of the training set compounds. The AD criteria help to check whether the test/query compound under consideration is inside the AD or not. The predictability of a QSPR model is good if the molecules were present within the domain of chemical space of training set molecules. Here, we have checked the applicability domain of test set compounds of the developed models, employing the standardization approach using the software developed in our laboratory (Roy et al., 2015) and a DModX (distance to model X) approach (Wold et al., 2001) at 99% confidence level using SIMCA-P software (www.umetrics.com). The predictability of a QSPR model is good if the molecules are present within the domain of the chemical space of the training set molecules.

3.3.5.5. Model validation based on OECD guidelines

To authenticate the applicability of the developed QSPR models and to judge the reliability of the predictions made, the models were further analyzed based on the OECD guidelines (Gramatica, 2007). Thus, the QSPR models developed in this work were validated based on these five guidelines laid down by the OECD. The compliance of the developed models with the OECD guidelines for applicability in regulatory purposes was assessed as follows:

(a) Principle 1 (a defined endpoint): The response parameter (CMC) modeled in the present work for the three different datasets were measured under similar conditions of temperature using identical solvent system. Thus the QSPR models were developed in accordance with the 1st OECD principle.

(b) Principle 2 (an unambiguous algorithm): Various chemometric tools based on specific algorithms were employed for the calculation of the different categories of descriptors and subsequent QSPR model development using specific software packages. Thus the model development pathway employed for the present studies follow a definite algorithm.

(c) Principle 3 (a defined domain of applicability): The domain of applicability of all the statistically significant QSPR models was analyzed in case of all the three datasets using the leverage method. The leverage values thus calculated were subsequently plotted against the standardized predicted residuals in William's plot for identification of response outliers and influential chemicals. Thus the selection of the best QSPR model was done in corroboration with this principle.

(d) Principle 4 (appropriate measures of goodness-of-fit, robustness and predictivity): All the developed models were rigorously validated using internal, external and overall validation techniques. The quality of fitness and the predictive potential of the developed models was assessed based on the different validation metrics while the robustness of the models was judged using the randomization approach. The selection of the most significant models based on the acceptable values of the various validation metrics account for the compliance of the models with the 4th guideline.

(e) Principle 5 (a mechanistic interpretation): All the descriptors appearing in the developed QSPR models could aptly define the essential structural attributes of the molecules imparting optimum CMC values to the cationic surfactants thus signifying suitable mechanistic interpretation of the developed models.

3.3.6. Software packages employed

Marvin Sketch version 5.5.0.1 (<http://www.chemaxon.com/>) was used to draw chemical structures. Descriptors were calculated by PADEL Descriptor software (<http://www.yapcwsoft.com/dd/padeldescriptor>) and Dragon software version 6 (http://www.taletе.mi.it/products/dragon_description.htm). Clustering of the data set was done by “Modified K-Medoid” tool version 1.3 (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) for its splitting into training set and test set. Data Pretreatment version 1.2 was used to remove intercorrelated descriptors. Stepwise regression analysis was done by MINITAB software version 13.14 (<http://www.minitab.com/en-US/default.aspx>). Best subset selection (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) and intelligent consensus predictor (Roy et al., 2018) were used to generate the QSAR models.

3.4. Study wise specific description of methodologies utilized in each study

3.4.1. Study1: Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs)

3.4.1.1. Selection of Dataset: The dataset involves the adsorption affinity of 69 organic contaminants related to the specific surface area (K_{SA}) of multi-walled carbon nanotubes (MWCNTs). The endpoint values were taken in the logarithmic scale for the modeling purposes. The data set mainly involve adsorption data for synthetic organic compounds like pyrene, naphthalene, phenol, benzene, aniline, benzoate, chloroanisole, alcohol, acetophenone, isophoron, phenanthrene dicamba, atrazine, carbamazepine, pyrimidinone, acetamide, piperidine, propionitrile, acrylic acid, thiodiethanol, ethanolamine, cyclopentanone, acetone and ethylene glycol derivatives. K_{SA} is adsorption coefficients that can be obtained from isotherm data. K_{SA} is the specific surface area of multi-walled carbon nanotubes. K_{SA} is the normalized value of K_{∞} which is the ratio of solid and liquid phase equilibrium concentrations at infinite dilution conditions with an average of 0.2% aqueous solubility.

3.4.1.2. Molecular descriptors: The descriptors were calculated using two software tools namely Dragon software version 6 (http://www.taletе.mi.it/products/dragon_description.htm.) and PaDEL-descriptor (<http://www.yapcwsoft.com/dd/padeldescriptor>.) software. In this work, we have calculated only 2D descriptors covering constitutional, ring descriptors, connectivity index,

functional group counts, atom centered fragments, atom type E-states, 2D atom pairs, molecular properties (using Dragon software version 6) and ETA indices (using PaDEL-Descriptor software).

3.4.1.3. Dataset division: Division of the dataset is a very important step for QSAR. The present work deals with three datasets containing diverse organic pollutants or solvents. In each case, all the dataset compounds were divided into a training set and a test set using “Modified k-medoid” clustering technique. Six clusters were generated for the dataset. We have selected approximately 25% compounds of the total data set for the test set and remaining 75% compounds selected for the training set.

3.4.1.4. Variable selection and model development: After the step of dataset division, we have performed data pretreatment to remove inter correlated descriptors from all three sets of datasets. Prior to development of final models, we have tried to extract the important descriptors from the large pool of initial descriptors using various variable selection strategies (Das et al., 2017b). We have run stepwise regression and selected some descriptors. After removing the selected descriptors obtained from first stepwise regression run, we have run again stepwise regression using remaining pool of descriptors and we have repeated the same procedure. In this way, we have selected some manageable number of descriptors and made a pool (reduced pool of descriptors).

The validation of the models was done by both internal and external validation tools. Internal validation metrics measured: R^2 , R_a^2 , and Q_{LOO}^2 . For external validation the measured metrics were: R_{pred}^2 (Q_{F1}^2) and Q_{F2}^2 . The MAE (95%) and RMSE values were also calculated.

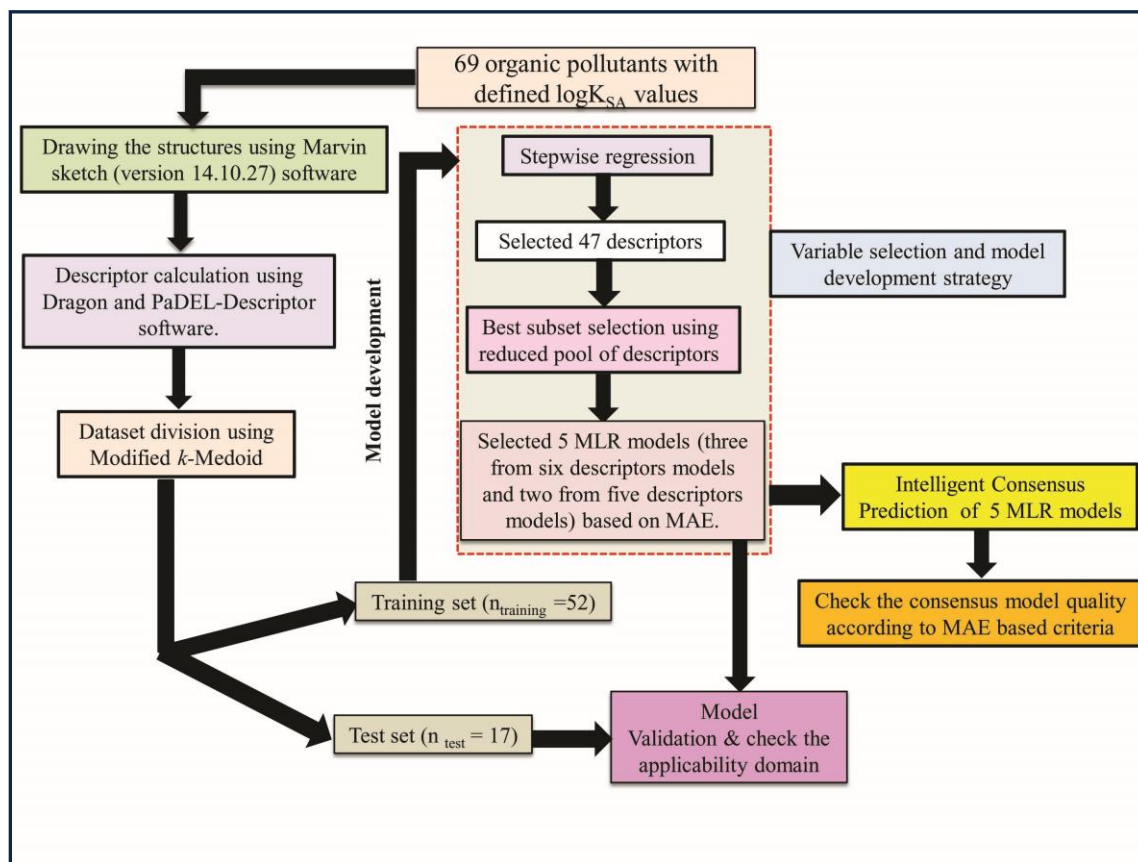


Fig. 4. Schematic representation of the steps involved for the development of QSPR models.

3.3.2. Study 2: Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs

3.3.2.1. Selection of Dataset: A data set of 40 hazardous synthetic organic chemicals (SOC) with defined adsorption coefficient onto SWCNTs reported in the literature (Ding et al., 2016a) were used to develop the QSPR models. The whole data set of 40 synthetic organic chemicals was collected from 14 published articles containing experimental adsorption coefficient (K in, L/kg) values. The endpoint K is the ratio of q_e and C_e . Where q_e is the equilibrium concentration on the surface and C_e is the equilibrium concentration in the aqueous phase of (Zhao et al., 2014).

3.3.2.2. Molecular descriptors: All the structures were drawn by using Marvin sketch software (<http://www.chemaxon.com>). The descriptors were calculated using two software tools, Dragon descriptor version 6 and PaDEL-Descriptor (<http://www.yapcwsoft.com/dd/padeldescriptor>)

software. Constitutional indices, ring descriptors, connectivity indices, functional group count, atom centered fragments, atom type E-state indices, 2D atom pairs and molecular properties were calculated using Dragon software, while extended topochemical atom (ETA) indices were calculated using PaDEL-Descriptor software.

3.3.2.3. Dataset division: In this work, the whole data set was divided by using the “datasetDivisionGUI1.2” (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) software tool. We have employed the Kennard-Stone method for data set division. The selection of objects in the Kennard-Stone algorithm was done in such a manner that they were uniformly distributed throughout the descriptor space of the data set. In this study, 75% compounds were selected for the training set, and the remaining 25% compounds were selected for the test set (i.e., 30 compounds for the training set and 10 compounds for the test set).

3.3.2.4. Variable selection and model development: Prior to the development of the final models, data pretreatment was performed to eliminate intercorrelated descriptors. Various variable selection strategies were employed to prepare the descriptor pool. We have excluded the variables with constant and near constant values (standard deviation less than 0.0001), descriptors with at least one missing value, descriptors with all missing values and descriptors with (absolute) pair correlation larger than or equal to 0.95 from the initial pool of descriptors. Initially, stepwise regression analysis was run and modeled descriptors were selected. The previously selected descriptors were removed from the initial pool of descriptors and stepwise regression analysis was rerun by using remaining pool of descriptors. In this manner, 31 descriptors were selected for the development of final models. Among the best subset equations, five models were selected based on Mean Absolute Error (MAE) criteria (Roy et al., 2016) along with some other parameters, and then carried out partial least squares (PLS) regression (Wold et al., 2001), in each case, using the selected descriptors. Finally, “intelligent consensus prediction” was performed of the test set compounds based on the selected five models using intelligent consensus predictor (ICP) tool (Roy et al., 2018) in order to investigate whether prediction quality of the external set compounds was increased or not through an “intelligent” selection.

We have used various statistical parameters like determination coefficient (R^2), explained variance (R_a^2), variance ratio (F) and standard error of estimate (s) were used for model validation. Apart from that, Q^2 , $r_m^2(LOO)$ and $\Delta r_m^2(LOO)$ were used for internal validation while R^2_{pred} , Q^2_{F2} , CCC, $r_m^2(test)$ and $\Delta r_m^2(test)$ were used for external validation.

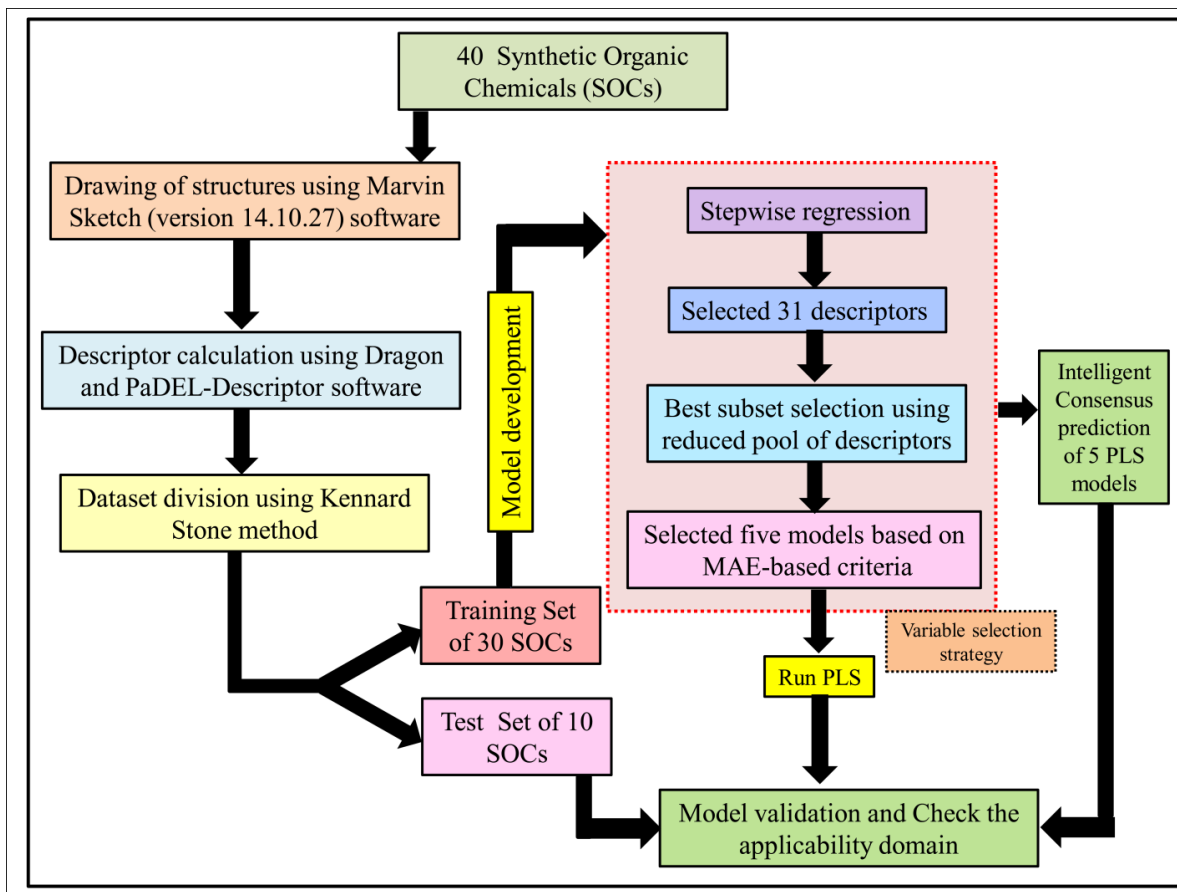


Fig. 5. Schematic representation of the steps involved in the development of final PLS model

Chapter 4



*RESULTS AND
DISCUSSIONS*

4. RESULTS AND DISCUSSIONS

4.1. Study 1: Predictive Quantitative Structure-Property Relationship (QSPR) Modeling for Adsorption of Organic Pollutants by Carbon Nanotubes (CNTs)

In the present work, QSPR models were developed comprising of a dataset containing 69 organic chemicals (Chayawan., 2016) with defined end point (adsorption affinity of organic contaminants related to specific surface area of multi walled carbon nanotubes ($\log K_{SA}$) to correlate the adsorption affinity ($\log K_{SA}$) in order to determine the structural features which are responsible for adsorption of organic contaminants by multi-walled carbon nanotubes (MWCNTs). Thus the aim of this study is:

- (a) Development of QSPR models.
- (b) Validation of models using different validation parameters to judge the statistical quality of the models.
- (c) Interpret the model descriptors and provide some information to the chemist regarding the structural features of organic pollutants which can change the adsorption affinity of organic contaminants by MWCNTs without any experimental work.

The significant descriptors obtained from the five MLR models using the adsorption properties ($\log K_{SA}$) of 69 organic pollutants related to the specific surface area of MWCNTs are Eta_Epsilon_3, X1A, X2A, nOHp, VAdjMat, F04(O-Cl), B05(O-Cl), MLOGP2, T(N..N), O%, and T(O..Cl). We have discussed here all the significant descriptors which are the key properties to alter the adsorption properties of organic pollutants. The definition, contribution and frequency of the modeled descriptors are shown in Tables 4.1. The applicability domain of the developed models using standardization approach showed that one test set compound (compound number **10**) for model **1**, two test set compounds (compound number **10** and **21**) for model **2**, one test set compound (compound number **21**) for model **3** are situated outside the applicability domain while in case of model nos. **4** and **5**, all the test set compounds are situated within the domain of applicability. Statistical quality and different validation parameters of the

models are given in Table 4.2 The scatter plot of observed vs predicted adsorption coefficient related to specific surface area of MWCNTs for all the MLR models are shown in Fig. 6.

Experimental $\log K_{SA}$ vs predicted $\log K_{SA}$ values of 69 organic pollutants

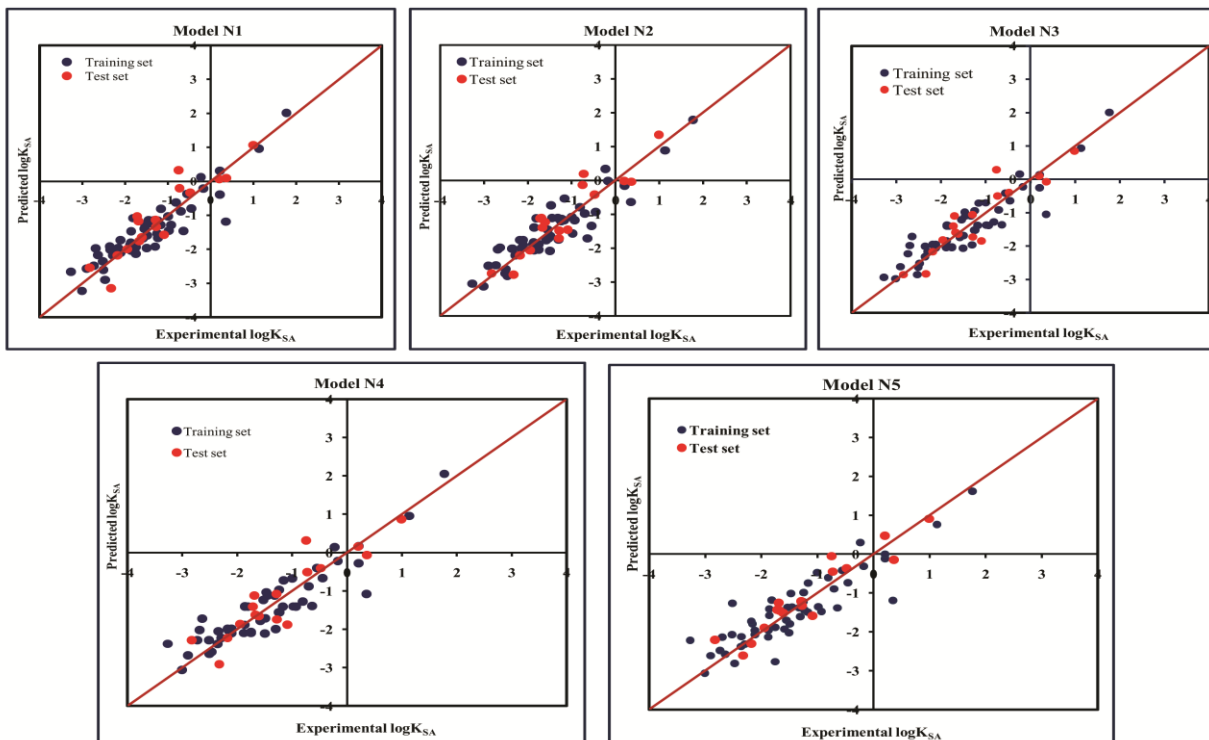


Fig. 6. The scatter plot of the observed and the predicted adsorption coefficient property related to specific surface area of MWCNTs ($\log K_{SA}$) of the developed MLR models (models 1-5).

Table 4.1. Definition and contribution of different descriptors obtained from five PLS models.

Sl. No.	Name of descriptor	Contribution	Discussion	Mechanism	Frequency of descriptors
1	ETA_Epsilon_3	+ve	Summation of electronegativity or ϵ value relative to the total no. of atoms including hydrogen in the connected molecular graph of the reference alkane	Electrostatic interaction	2
2	X1A	-ve	Average connectivity index of order 1	Hydrophobic interaction	1
3	X2A	-ve	Average connectivity index of order 2	Hydrophobic interaction	1
4	nOHp	-ve	Number of primary alcohols	π - π interaction	2
5	VAdjMat	+ve	Vertex adjacency information (magnitude)	Hydrophobic interaction	1
6	F04(O-Cl)	-ve	Number of (O..Cl) fragment at topological distance 4	π - π interaction	1
7	B05[Cl-Cl]	-ve	Presence/absence of Cl - Cl at topological distance 5	π - π interaction	2
8	MLOGP2	+ve	Squared Moriguchi octanol-water partition coefficient ($\log P^2$)	Hydrophobic interaction	5
9	T(N..N)	+ve	Sum of topological distances between N..N	π - π interaction	4

<p>Model N1</p> $\log K_{SA} = 4.29(\pm 2.194) + .0965(\pm 0.014) \times O\% - 16.4(\pm 4.397) \times X1A + 0.145(\pm 0.032) \times T(N..N)$ $- 0.0279(\pm 0.009) \times T(O..CI) - 1.01(\pm 0.094) \times B05(CI - CI) + 0.203(\pm 0.203) \times MLOGP2$ $n_{training} = 52, R^2 = 0.845, R^2_{(adj)} = 0.824, Q^2 = 0.798, S = 0.433, PRESS = 11.003, F = 40.79$ $\overline{r^2}_{m(LOO)} = 0.709, \Delta r^2_{m(LOO)} = 0.087, MAE \text{ based criteria} = \text{Moderate}$ $n_{test} = 17, Q^2_{F1} = 0.809, Q^2_{F2} = 0.795, \overline{r^2}_{m(test)} = 0.783, \Delta r^2_{m(test)} = 0.048,$ $CCC = 0.908, MAE \text{ based criteria} = \text{Moderate}.$ <p>Model N2</p> $\log K_{SA} = -7.19(\pm 0.571) + 0.0805(\pm 0.015) \times O\% - 0.323(\pm 0.662) \times nOHp$ $- 0.0358(\pm 0.009) \times T(O..CL) - 0.943(\pm 0.294) \times B05(CI - CI) + 0.185(\pm 0.019)$ $\times MLOGP2 + 0.958(\pm 0.144) \times VAdjMat$ $n_{training} = 52, R^2 = 0.842, R^2_{(adj)} = 0.821, Q^2 = 0.790, S = 0.437, PRESS = 11.41,$ $F = 39.97, \overline{r^2}_{m(LOO)} = 0.723, \Delta r^2_{m(LOO)} = 0.114, MAE \text{ based criteria} = \text{Moderate}$ $n_{test} = 17, Q^2_{F1} = 0.9830, Q^2_{F2} = 0.818, \overline{r^2}_{m(test)} = 0.805, \Delta r^2_{m(test)} = 0.050,$ $CCC = 0.918, MAE \text{ based criteria} = \text{good}.$ <p>Model N3</p> $\log K_{SA} = -42.3(\pm 7.527) + 0.0973(\pm 0.013) \times O\% - 6.22(\pm 0.323) \times nOHp$ $+ 0.154(\pm 0.031) \times T(N..N) - 0.0407(\pm 0.008) \times T(O..CI) + 0.160(\pm 0.20) \times$ $MLOGP2 + 89.8(\pm 17.51) \times ETA_Epsilon_3$ $n_{training} = 52, R^2 = 0.842, R^2_{(adj)} = 0.821, Q^2 = 0.788, S = 0.436, PRESS = 11.512,$ $F = 40.07, \overline{r^2}_{m(LOO)} = 0.714, \Delta r^2_{m(LOO)} = 0.081, MAE \text{ based criteria} = \text{Good}$ $n_{test} = 17, Q^2_{F1} = 0.783, Q^2_{F2} = 0.768, \overline{r^2}_{m(test)} = 0.712, \Delta r^2_{m(test)} = 0.14,$ $CCC = 0.890, MAE \text{ based criteria} = \text{Good}.$ <p>Model N4</p> $\log K_{SA} = -42.0(\pm 7.743) + 0.101(\pm 0.014) \times O\% + 0.159(\pm 0.032) \times T(N..N)$ $- 0.0411(\pm 0.008) \times T(O..CI) + 0.168(\pm 0.021) \times MLOGP2 + 88.9(\pm 18.01) \times ETA_Epsilon_3$ $n_{training} = 52, R^2 = 0.829, R^2_{(adj)} = 0.811, Q^2 = 0.785, S = 0.449, PRESS = 11.722,$ $F = 44.73, \overline{r^2}_{m(LOO)} = 0.707, \Delta r^2_{m(LOO)} = 0.087, MAE \text{ based criteria} = \text{Good}$ $n_{test} = 17, Q^2_{F1} = 0.812, Q^2_{F2} = 0.799, \overline{r^2}_{m(test)} = 0.748, \Delta r^2_{m(test)} = 0.044,$ $CCC = 0.903, MAE \text{ based criteria} = \text{Good}.$ <p>Model N5</p> $\log K_{SA} = 2.49(\pm 1.36) + 0.0757(\pm 0.016) \times O\% - 17.3(\pm 3.773) \times X2A$ $+ 0.145(\pm 0.036) \times T(N..N) - 0.721(\pm 0.144) \times F04(O - CI)$ $+ 0.15(\pm 0.023) MLOGP2$ $n_{training} = 52, R^2 = 0.793, R^2_{(adj)} = 0.77, Q^2 = 0.743, S = 0.495, PRESS = 13.955,$ $F = 35.17, \overline{r^2}_{m(LOO)} = 0.709, \Delta r^2_{m(LOO)} = 0.087, MAE \text{ based criteria} = \text{Good}$ $n_{test} = 17, Q^2_{F1} = 0.890, Q^2_{F2} = 0.882, \overline{r^2}_{m(test)} = 0.836, \Delta r^2_{m(test)} = 0.090,$ $CCC = 0.940, MAE \text{ based criteria} = \text{Good}$

Table 4.2. Statistical quality and validation parameters obtained from the developed MLR models.

Data set	Type of model		Training set statistics					Test set statistics							
			Model R ²	Model Q ² (LOO)	MAE_train	$\overline{r}_{m(LOO)}^2$	$\Delta r_{m(LOO)}^2$	R ² _{pred} or Q ² F ₁	Q ² F ₂	CCC	$\overline{r}_{m(test)}^2$	$\Delta r_{m(test)}^2$	MAE (100%)	MAE (95%)	MAE
69 organic contaminants	Individual Models	IM1	0.845	0.798	Moderate	0.709	0.087	0.809	0.795	0.908	0.783	0.048	0.319	0.271	Moderate
		IM2	0.842	0.790	Moderate	0.723	0.114	0.830	0.818	0.918	0.805	0.050	0.359	0.323	Good
		IM3	0.842	0.788	Good	0.714	0.081	0.783	0.768	0.890	0.712	0.140	0.340	0.265	Good
		IM4	0.829	0.785	Good	0.709	0.087	0.812	0.799	0.903	0.748	0.044	0.330	0.286	Moderate
		IM5	0.793	0.743	Good	0.709	0.087	0.890	0.882	0.940	0.836	0.090	0.273	0.247	Good
	Consensus Models	CM0	-	-				0.862	0.852	0.929	0.818	0.002	0.284	0.245	Good
		CM1	-	-				0.862	0.852	0.929	0.818	0.002	0.284	0.245	Good
		CM2	-	-				0.865	0.852	0.930	0.820	0.014	0.279	0.241	Good
		CM3	-	-				0.887	0.879	0.941	0.851	0.040	0.263	0.235	Good

CM0=Ordinary consensus predictions.

CM1 = Average of predictions from individual models IM1 through IM5. CM2 = Weighted average predictions from individual models IM1 through IM5.

CM3 = Best selection of predictions (compound-wise) from individual models IM1 through IM5.

*Note that we have run the “Intelligent consensus predictor tool” using the options,

AD: No; Dixon Q-test: No; Euclidean distance: N

4.1.1. The descriptors related to hydrophobic interaction

The descriptor, X1A, indicates average connectivity index of order one, it encodes the ‘chi’ value across one bond which can be calculated on the basis of Kier and Hall’s connectivity index and defined as follows:

$${}^1X = \sum_{b=1}^B (\delta_i \cdot \delta_j)_b^{-0.5}$$

In this equation, b runs over the 1st order sub graphs having n vertices with B edges, δ_i and δ_j are number of other vertices attached to vertex i and j respectively. The negative regression coefficient of this descriptor implies that the higher numerical values of this descriptor are not favorable to enhance the adsorption property of organic pollutants related to the specific surface area of MWCNTs as shown in compound nos. **3 (benzene)**, **56 (ethylbenzene)** and **57 (benzyl alcohol)** (corresponding numerical values of these compounds are 0.5, 0.491, 0.491 respectively showing lower range of adsorption affinity). On the other hand, compounds like **35 (tetracycline)**, **22 (pyrene)** and **26 (phenanthrene)** show better adsorption affinity ($\log K_{SA}$) due to their lower numerical value of this descriptor.

Another significant descriptor, X2A, indicates average connectivity index of order 2, encodes the ‘chi’ value across two bonds which can be calculated on the basis of Kier and Hall’s connectivity index and defined in the following equation:

$${}^2X = \sum_{b=2}^B (\delta_i \cdot \delta_j)_b^{-0.5}$$

Here, b runs over the 2nd order sub graphs having n vertices with B edges, δ_i and δ_j are number of other vertices attached to vertex i and j respectively. This descriptor is also having a negative contribution towards the adsorption profile ($\log K_{SA}$) of organic pollutants by MWCNTs as evidenced by the negative regression coefficient. This indicates that the adsorption property of organic pollutants decreases with an increase in the numerical value of this descriptor as shown in compounds **3 (benzene)**, **18(aniline)** and **40 (bromobenzene)** and vice versa in case of compounds **22 (pyrene)**, **26 (Phenanthrene)** and **35 (tetracycline)**.

The VAdjMat descriptor represents vertex adjacency information, gives information about molecular dimension and hydrophobicity. This descriptor can be calculated by using the following formula:

$$VAdjMat = 1 + \log_2(m)$$

Here, m depicts the number of heavy-heavy bonds. This descriptor contributed positively towards the adsorption property ($\log K_{SA}$) of organic pollutants as indicated by the positive

regression coefficient. Thus, the higher numerical value of this descriptor is influential to the adsorption affinity of organic pollutants. This indicates that hydrophobicity plays a crucial role to alter the adsorption property of organic pollutants by MWCNTs. As for example, compounds **22 (pyrene)**, **26 (Phenanthrene)** and **35 (tetracycline)** show higher range of adsorption property as these compounds contain higher numerical value of this descriptor whereas compounds **3 (benzene)**, **55 (iodobenzene)** and **46 (chlorobenzene)** show lower range of adsorption property as these compounds contain higher numerical value of this descriptor. From this descriptor, it can be suggested that the hydrophobic organic pollutants can easily be adsorbed by MWCNTs through hydrophobic interactions between the pollutants and CNTs.

The next descriptor, MLOGP2, represents squared Moriguchi octanol water partition coefficient, calculated from the regression equation of Moriguchi logP model (Moriguchi et al., 1994; Ojha, and Roy., 2018) consisting of 13 parameters as depicted in the below mentioned equation.

$$\begin{aligned} \log P = & -1.244(CX)^{0.6} - 1.017(NO)^{0.9} + 0.406PRX - 0.145(UB)^{0.8} + 0.511HB \\ & + 0.268POL - 2.215AMP + 0.912ALK - 0.392RNG - 3.684QN + 0.474NO_2 \\ & + 1.582NCS + 0.773BLM - 1.041 \end{aligned}$$

Here, 'CX' depicts summation of weighted number of carbon atoms; 'NO' depicts total number of N and O atoms; 'PRX' represents Proximity effect of N/O; 'UB' represents number of unsaturated bonds including semi polar bonds; 'POL' depicts number of aromatic polar substituent; 'AMP' depicts amphoteric property; 'ALK' represents dummy variable for alkane, alkene; 'RNG' depicts Indicator variable for presence of ring structure except benzene and its condensed ring; 'QN' represents Quaternary nitrogen; 'NO₂' represents number of nitro groups; 'HB' represents a dummy variable for the presence of intermolecular hydrogen bond; 'NCS' depicts 'isothiocyanito; thicyanito; 'BLM' represents a dummy variable for the presence of β-lactam.

The positive regression coefficient of this descriptor indicates that hydrophobicity plays a crucial role to regulate the adsorption property of organic pollutants. The highly hydrophobic organic pollutants can easily get adsorbed by MWCNTs as evidenced by the compounds **22 (pyrene)**, **26 (phenanthrene)** and **34 (azobenzene)** as their corresponding MLOGP2 values are 22.653, 18.762 and 10.539 respectively whereas hydrophilic molecules are poorly adsorbed by MWCNTs as evidenced by the compounds **18 (aniline)**, **57 (benzylalcohol)** and **63 (3-nitroaniline)** as their corresponding MLOGP2 values are 2.268, 2.532 and 1.816 respectively.

Therefore, it can be inferred that the organic pollutants are adsorbed to the CNTs through hydrophobic interactions. Thus, for proper adsorption, organic pollutant should be hydrophobic in nature. MLOGP2 is not strictly a 2D descriptor. Here, a term ‘intramolecular H-bonds’ is used to calculate MLOGP value which is conformation dependent.

The information obtained from the descriptors X1A, X2A, VAdjMat and MLOGP2 suggested that adsorption of organic pollutants related to specific surface area of MWCNTs may occur through hydrophobic interactions. Molecular connectivity index (X1A and X2A) has a direct relationship with count of interacting C-H bonds present in a molecule. The number of C-H bonds in a molecule is equal to the number of H atoms. As the C-H bond increases, the hydrophobicity of the molecule increases. The δ value (depends on the number of H atoms, definition of a δ value for a carbon atom in a molecular graph is: $\delta = 4 - H$) is decreasing with the average connectivity index. Thus, the hydrophobic interaction between the organic contaminants and MWCNTs is reduced and the adsorption of organic pollutants related to the specific surface area of MWCNTs may also reduce (Kier and Hall, 2002)

The descriptors VAdjMat and MLOGP2 give information about hydrophobicity of molecules. It is obvious that the hydrophobic organic pollutants will interact with hydrophobic CNTs through hydrophobic interactions. This is implied that the hydrophobic organic pollutants can easily be adsorbed by MWCNTs through hydrophobic interactions. The descriptors involved for hydrophobic interaction are graphically depicted in Fig.7.

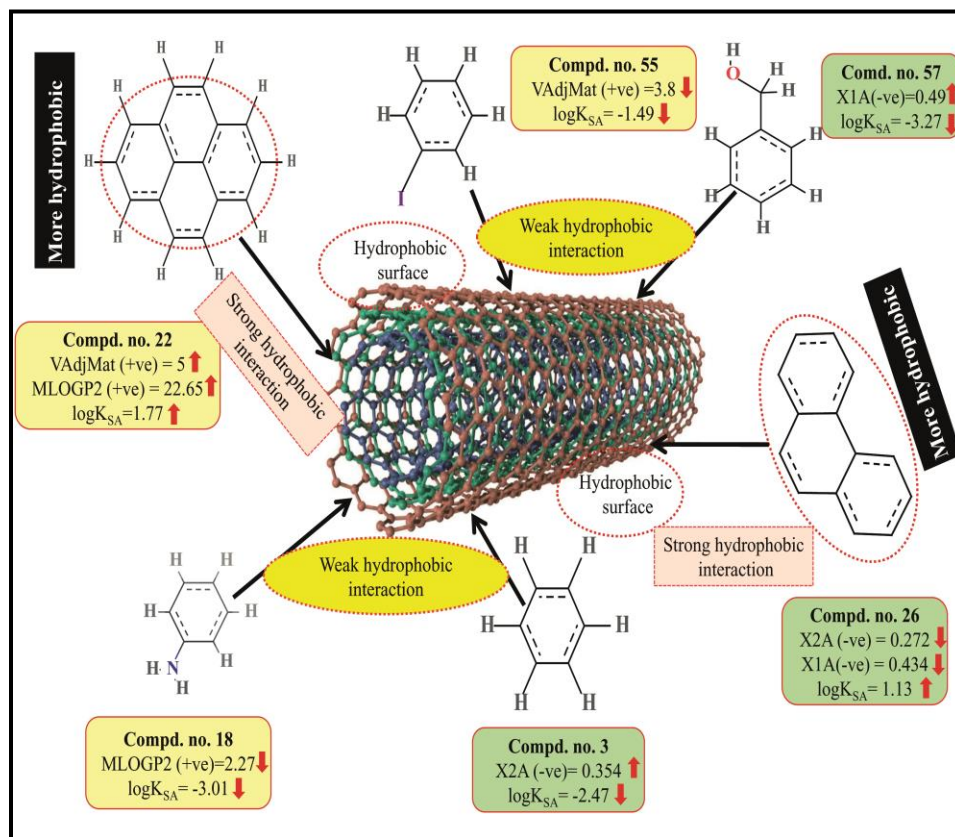


Fig.7. Mechanistic interpretation of the descriptors related to hydrophobic interaction between organic pollutants and MWCNTs.

4.1.2. The descriptors related to π - π interaction

A functional group count descriptor, nOHp, describes the number of primary alcohols. The negative regression coefficient of this descriptor points out that the primary alcoholic group is not favored to enhance the adsorption property ($\log K_{SA}$) of organic pollutants as found in compounds **13** (3-methyl benzyl alcohol) and **57** (benzyl alcohol). On the contrary, organic pollutants do not containing any primary alcoholic group have higher adsorption affinity ($\log K_{SA}$) as shown in compounds **22** (pyrene), **26** (Phenanthrene) and **34** (azobenzene). Thus, the organic pollutants which do not contain any primary alcoholic group may be highly adsorbed by MWCNTs.

F04[O-Cl] is a 2D atom pair descriptor which indicates the number of (O-Cl) fragments at a topological distance 4. The negative regression coefficient of this descriptor indicates that the frequency of O-Cl fragment at the topological distance 4 is inversely proportional to the

adsorption property of organic pollutants. A higher number of this fragment correlates to lower adsorption property of organic pollutants as observed in compounds **7 (dicamba)**, **61 (3-chlorophenol)** and **66 (2,4,5-trichlorophenoxyacetic acid)** (these compounds contain 3, 1 and 1 such fragments respectively at topological distance 4), while a lower numerical value of this descriptor correlates to higher adsorption property of organic pollutants as observed in compounds **22 (pyrene)**, **26 (phenanthrene)**, **34 (azobenzene)** and **69 (2,4-dinitrotoluene)** (these compounds contain no such fragments at topological distance 4). Thus, presence of this fragment at the topological distance 4 may hinder adsorption of the organic pollutants by MWCNTs. Adsorption of organic contaminants to the CNTs decreases when frequency of (O-Cl) fragment at topological distance 4 increases. Compound **2 (2,4,6-trichlorophenol)** also contains a O-Cl fragment but not at topological distance 4. So, the adsorption affinity related to the specific surface area of MWCNTs value of compound 2 is ($\log K_{SA}$ value=-0.81) not low as compared to compounds **7 (dicamba)**, **61 (3-chlorophenol)** and **66 (2,4,5-trichlorophenoxyacetic acid)** (these compounds contain 3, 1 and 1 such fragments respectively at topological distance 4 and the $\log K_{SA}$ value is -2.64, -1.75 and -2.51 respectively).

T(O..Cl), a 2D atom pair descriptor, indicates sum of topological distance between oxygen and chlorine. The negative regression coefficient of this descriptor suggested that higher numerical value of this descriptor is detrimental to enhance the adsorption property of organic pollutants related to specific surface area of MWCNTs as shown in compounds **2 (2,4,6-trichlorophenol)**, **7(dicamba)** and **66 (2,4,6-trichlorophenoxyacetic acid)**. On the other hand, the organic pollutants containing no such fragments have higher adsorption property as shown in compounds **22 (pyrene)**, **26 (phenanthrene)** and **34 (azobenzene)**. From this observation, it can be suggested that the organic pollutants without (O..Cl) fragments may adsorb better to MWCNTs surface.

A 2D atom pair descriptor, B05(Cl-Cl), describes the presence or absence of Cl-Cl fragments at topological distance 5. The negative regression coefficient of this descriptor indicates that the presence of the Cl-Cl fragment at the topological distance 5 may reduce the adsorption property of organic pollutants related to the specific surface area of MWCNTs ($\log K_{SA}$). A higher number of this fragment correlates to lower adsorption property of organic pollutants as observed in compounds **7 (dicamba)**, **41 (1,2,4-trichlorobenzene)** and **66 (2,4,5-trichlorophenoxyacetic acid)** (containing one such fragment each) while absence of this fragment in organic pollutants

correlates to higher adsorption property as evidenced from compounds **22 (pyrene)**, **26 (phenanthrene)** and **34 (azobenzene)**. From this descriptor, it can be suggested that presence of this fragment at topological distance 5 may retard adsorption of the organic pollutants by MWCNTs.

Another 2D atom pair descriptor, T(N..N), indicates the sum of topological distances between two nitrogen atom. A positive contribution towards the adsorption property of organic pollutants related to the specific surface area of MWCNTs ($\log K_{SA}$) indicated that for better adsorption of organic pollutants by MWCNTs, the topological distance between two nitrogen atoms should be more as shown in compounds **4 (oxytetracycline)**, **35 (tetracycline)** and **69 (2,4-dinitrotoluene)** (as their corresponding topological distance between two nitrogen atoms are 5, 5 and 4 respectively) and vice versa in case of compounds **42 (isophorone)**, **43 (4-fluorophenol)** and **44 (acetophenone)**. Thus, it can be inferred that the topological distances between two nitrogen atoms should be more for better adsorption of organic pollutants by MWCNTs.

As discussed earlier in the introduction section that π - π interactions is one of the key mechanisms for the adsorption of organic pollutants to CNTs. The information obtained from these descriptors nOHp, F04[O-Cl], B05[Cl-Cl], T(N..N) and T(O..Cl) strongly support this statement. The descriptor nOHp weakens the π - π interaction that occurs between the organic pollutants and CNTs. In this case, the hydroxyl group is alcoholic in nature (aliphatic hydroxyl group). Thus, it cannot donate the lone pair of electrons to the aromatic ring (not directly bonded with aromatic carbon) and ultimately weakens the π - π interactions of the aromatic ring though it can form hydrogen bonds with the surface modified CNTs. On the other hand, the phenolic hydroxyl group can donate the lone pair of electrons to the aromatic ring (bonded directly to the aromatic carbon atom) as discussed previously (section 3.1) thus strengthening the π - π interactions between organic pollutants and CNTs. In case of phenolic hydroxyl group, it can also act as a π donor but it is not possible in case of alcoholic hydroxyl group. From this observation, it can be suggested that aliphatic hydroxyl (alcoholic) group is not favorable for the adsorption affinity of organic pollutants to the CNTs. In case of the descriptors B05[Cl-Cl], T(O..Cl) and F04[O-Cl], the chlorine atom has an electron inductive effect and decreases the electron density in the benzene ring, which compensates for the electron-donating effect of the oxygen atom (in case of compound nos. **7** and **66**), even after -OH dissociated into -O⁻. The withdrawing inductive character of chlorine substituents decreases the electron density of the

p-chlorophenol ring compared with that of phenol ring. Thus, when O-Cl or Cl-Cl fragment is present in an aromatic molecule, it decreases the electron density of that aromatic ring (compared with that of -OH substituted benzene ring (phenolic) or benzene ring itself) and ultimately, electron donor-acceptor interactions does not occur easily between CNTs and organic contaminants. Hence, the compound could not be easily adsorbed to the MWCNTs. In case of the descriptor T(N..N), the lone pair of electrons of nitrogen atom can be donated to the ring system (when directly attached) and enhance the π - π interaction with the CNTs. The nitrogen can be present as amino form (electron donating) or in nitro form (electron withdrawing). Both the forms strengthen the π - π interaction between the organic pollutants and CNTs by increasing or decreasing the π -electrons density of the aromatic ring system and act as π electron donor or acceptor respectively. If the nitrogen is not attached directly to the aromatic ring system then adsorption happens through electrostatic interaction between nitrogen of the pollutants and hydrogen of CNTs by forming dipoles when they are coming close to each other, the position of placing nitrogen atom hardly matters here. The descriptors influencing the π - π interaction are graphically represented in Fig.8.

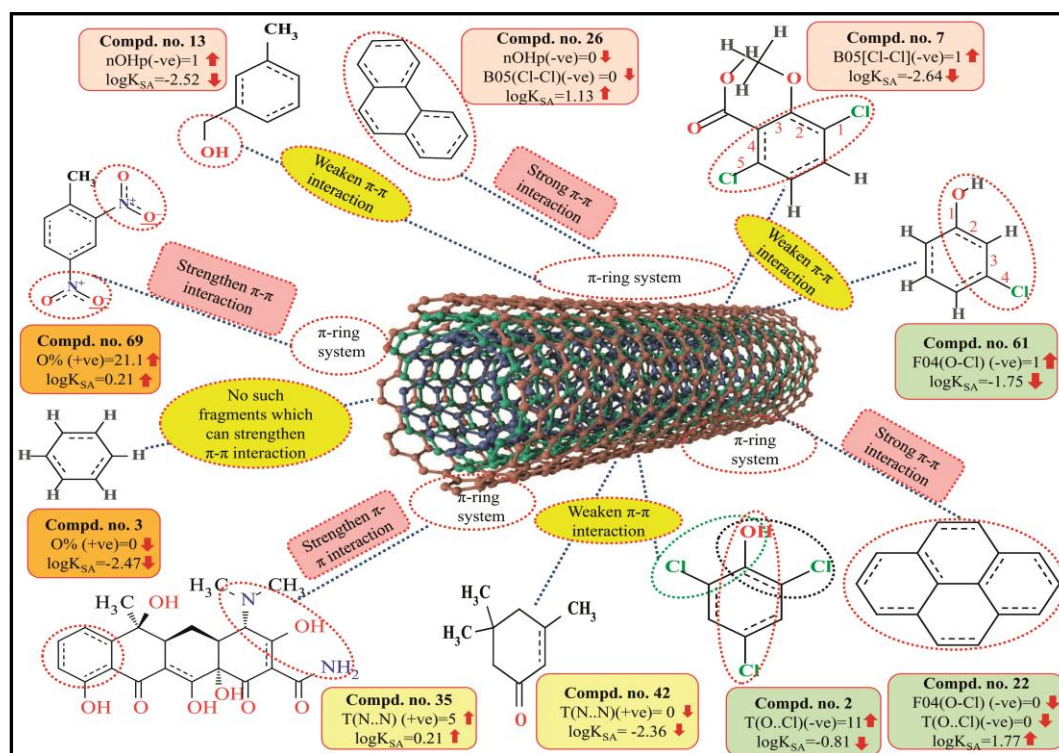


Fig. 8. Mechanistic interpretation of the descriptors related to π - π interaction between organic pollutants and MWCNTs.

4.1.3. The descriptors related to hydrogen bonding interaction

The descriptor, O%, indicates the percentage of oxygen atoms present in a particular molecule. The positive regression coefficient of this descriptor suggested that presence of oxygen atom is highly influential to adsorb the organic pollutants on the surface of MWCNTs. As for example, compounds **4** (oxytetracycline), **35** (tetracycline) and **69** (2,4-dinitrotoluene) show better adsorption affinity as their corresponding percentage of oxygen atom is 15.8, 14.3 and 21.1 respectively. On the contrary, compounds **3** (benzene), **18** (aniline) and **24** (4-chloroaniline) show poor adsorption affinity as these compounds do not contain any oxygen atom. The oxygen atom may be present in different organic pollutants in keto, phenolic (favorable for adsorption) or alcoholic forms (not favorable for adsorption as discussed in this section previously). These different types of oxygen may interact with CNTs in different ways like hydrogen bonding, strengthening the π - π interactions and electrostatic interactions. On the other hand, a high percentage of oxygen atoms may enhance the polarity of the pollutants. As the side wall of the CNTs is also electrically polarized, the polar group of organic pollutants can easily be adhered to the surface of CNTs. The descriptor involved for hydrogen bonding interaction is given in Fig. 9.

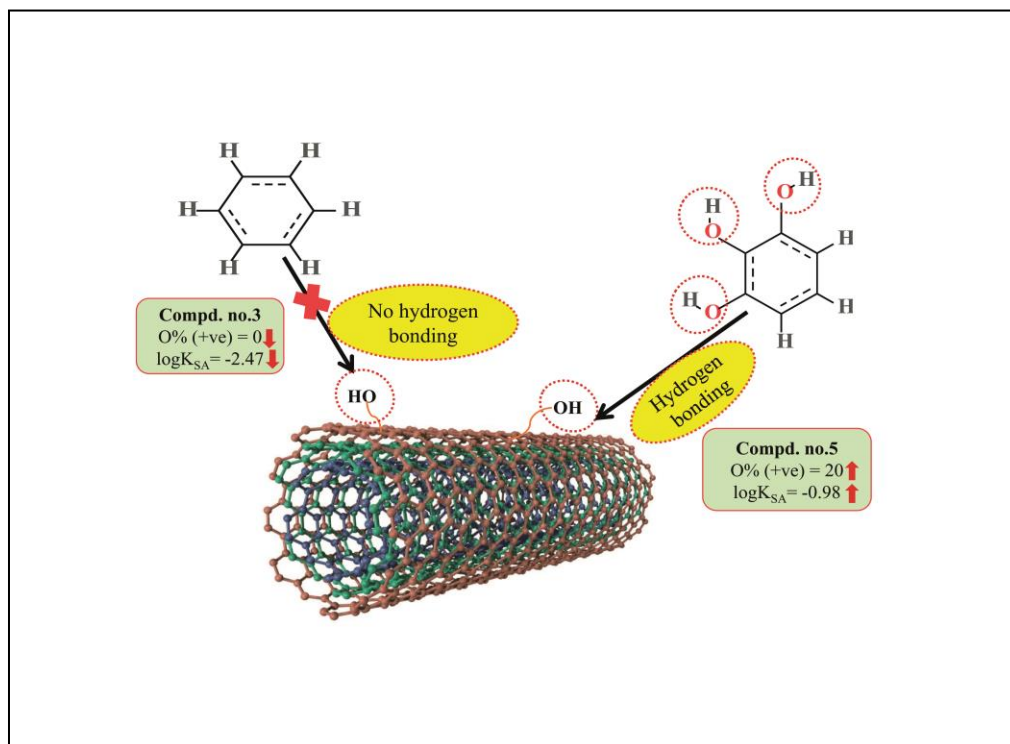


Fig. 9. Mechanistic interpretation of the descriptors related to hydrogen bonding interaction between organic pollutants and MWCNTs.

4.1. 4. The descriptors related to electrostatic interaction

The descriptor, Eta_Epsilon_3, indicates summation of epsilon values relative to the total number of atoms including hydrogen in the connected molecular graph of the reference alkane which can be calculated by the following equation.

$$\epsilon_3 = \epsilon_R / N_R$$

In the above equation, ϵ denotes electronegativity; N_R denotes number of atom present in reference alkane. This descriptor has a positive contribution towards the adsorption property of organic pollutants related to the specific surface area of MWCNTs. This indicated that electron rich organic pollutants will be highly adsorbed by MWCNTs. Thus, the higher numerical value (due to strong electrostatic interactions between organic pollutants and CNTs) of this descriptor is required to increase the adsorption property of organic pollutants by MWCNTs as shown in compound nos. **22 (pyrene)**, **26 (phenanthrene)** and **35 (tetracycline)** and vice versa in case of compound nos. **7 (dicamba)**, **13 (3-methylbenzyl alcohol)** and **18 (aniline)** (due to weak electrostatic interactions between these organic pollutants and CNTs).

The information obtained from the descriptor O% suggested that the organic pollutants can adhere to the surface of MWCNTs by electrostatic interaction. There may be a chance to form electrostatic interactions between organic pollutants (negative charged atom like oxygen atom of hydroxyl group) and MWCNTs (sidewall of the CNTs are electrically polarizable thus polar molecules can easily adhere to their surface). The descriptors involved for electrostatic interaction are shown graphically in Fig. 10.

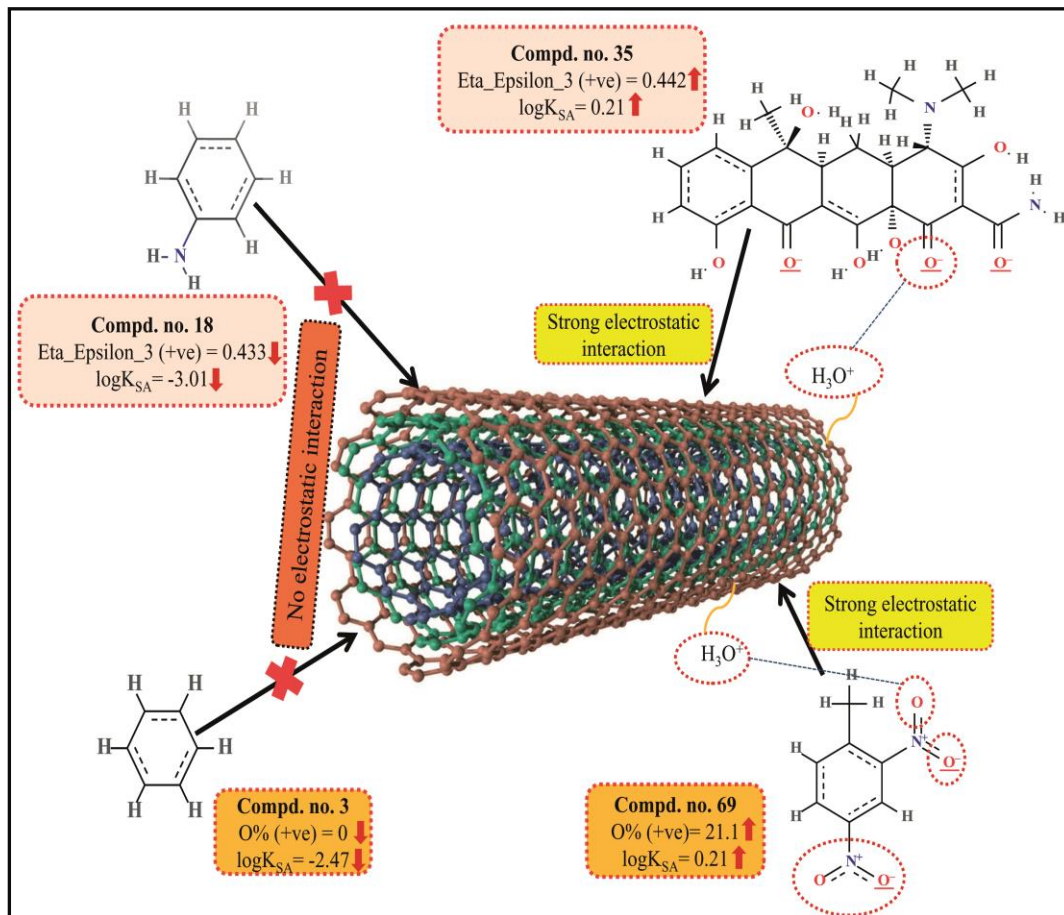


Fig. 10. Mechanistic interpretation of the descriptors related to electrostatic interaction between organic pollutants and MWCNTs.

4.2. Study 2: Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs

In the current study, five PLS models [Box 2] were developed for the dataset containing 40 diverse hazardous SOC having significant adsorption affinity for SWCNTs, using a reduced descriptor pool obtained by stepwise regression method, as discussed in Methods and Materials section. We have validated the models using various internal and external validation parameters, which showed that the models are statistically significant (**Table 4.3**). The MAE based criteria of all the models were passed which indicates that all the models are acceptable. We have also checked the consensus predictivity of all the individual models (1-5) using “Intelligent consensus predictor” tool to see whether the quality of predictions of test set compounds can be enhanced or not. The consensus predictivity of the test set compounds were found to be better than the

individual models based on MAE based criteria as depicted in **Table 4.3** (the winner model is CM3). The descriptors obtained from the individual models are discussed elaborately in this section. To judge the predicting ability of the developed PLS models, we have also validated the model using Golbraikh and Tropsha criteria, and the results are given in **Table 4.4**. The results showed that the models are acceptable based on these criteria. Randomization test was also performed by using the SIMCA-P software to verify whether the model was obtained by any chance or not. The intercept of both R^2 and Q^2 values are below the stipulated values of $R^2_{int} < 0.4$ and $Q^2_{int} < 0.05$ which confirmed that the models are not obtained by any chance (Fig. 11-15). The definitions and contributions of different descriptors obtained from five PLS models are depicted in **Table 4.5**. In equations as depicted below, $n_{training}$ is the number of compounds used to develop the models, and n_{test} is the number of compounds used for external prediction. The leave one out (LOO) cross validated correlation coefficient Q^2 ($Q^2=0.861-0.901$) above the critical value of 0.5 signifies the statistical reliability of the models. The predictive R^2 (R^2_{pred}) or Q^2_{F1} ($Q^2_{F1}=0.898-0.929$) and Q^2_{F2} ($Q^2_{F2}=0.897-0.928$) show good predictive ability of the models. We have also checked the applicability domain of the compounds using the standardization approach. All the compounds are found to be present within the AD. The scatter plot of observed vs. predicted adsorption coefficient for five PLS models are depicted in **Fig. 16**.

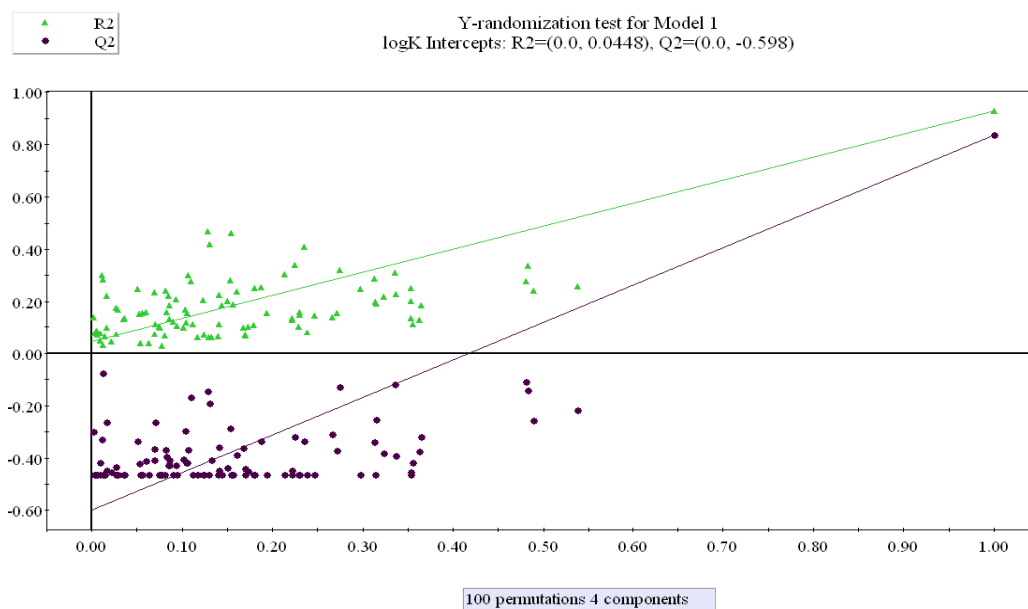


Fig. 11. The randomization plot for the QSPR model 1 derived from PLS analysis.

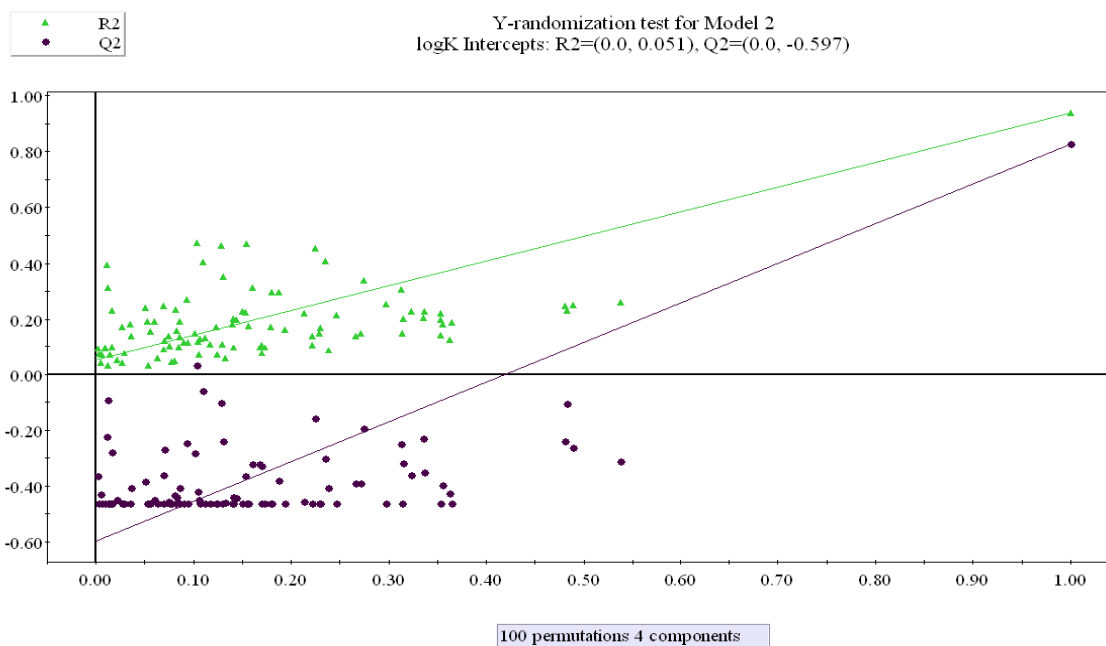


Fig. 12. The randomization plot for the QSPR model 2 derived from PLS analysis

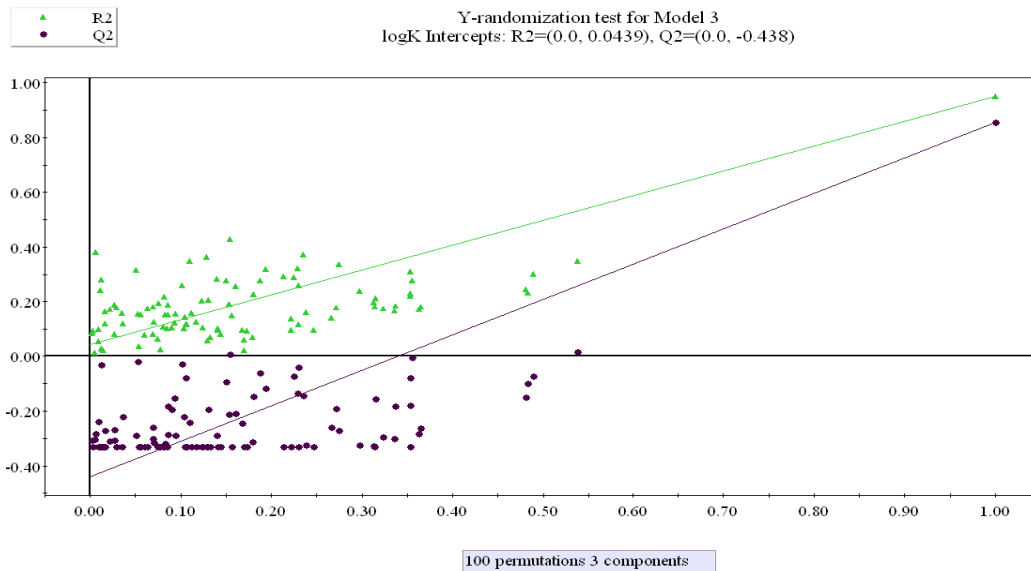


Fig. 13. The randomization plot for the QSPR model 3 derived from PLS analysis.

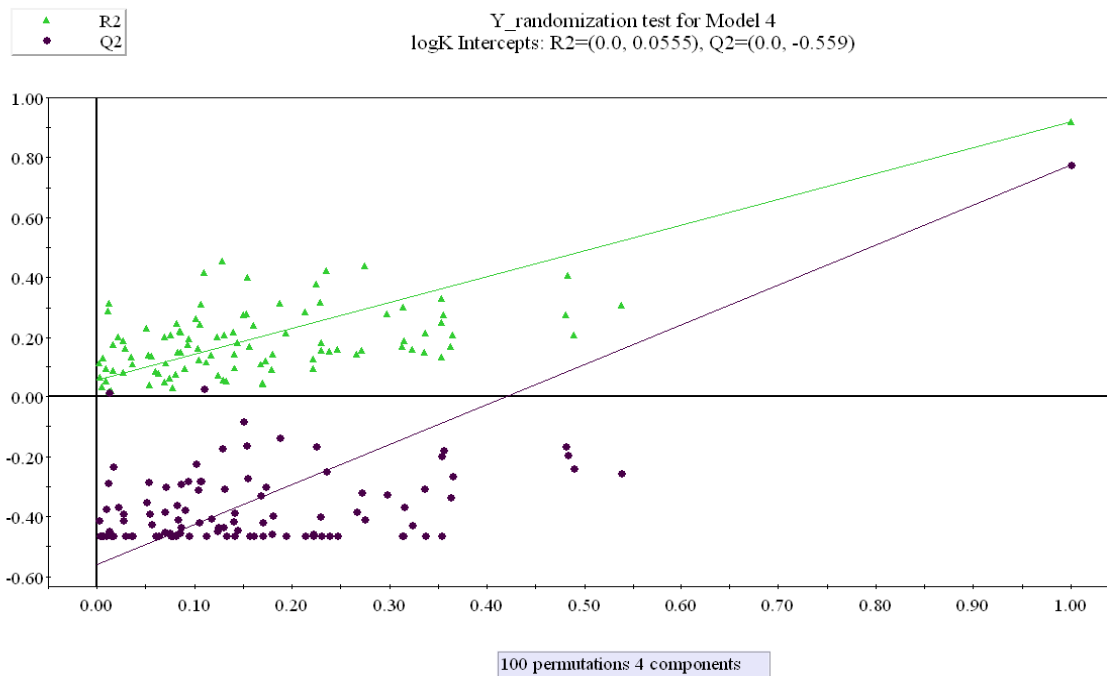


Fig. 14. The randomization plot for the QSPR model 4 derived from PLS analysis.

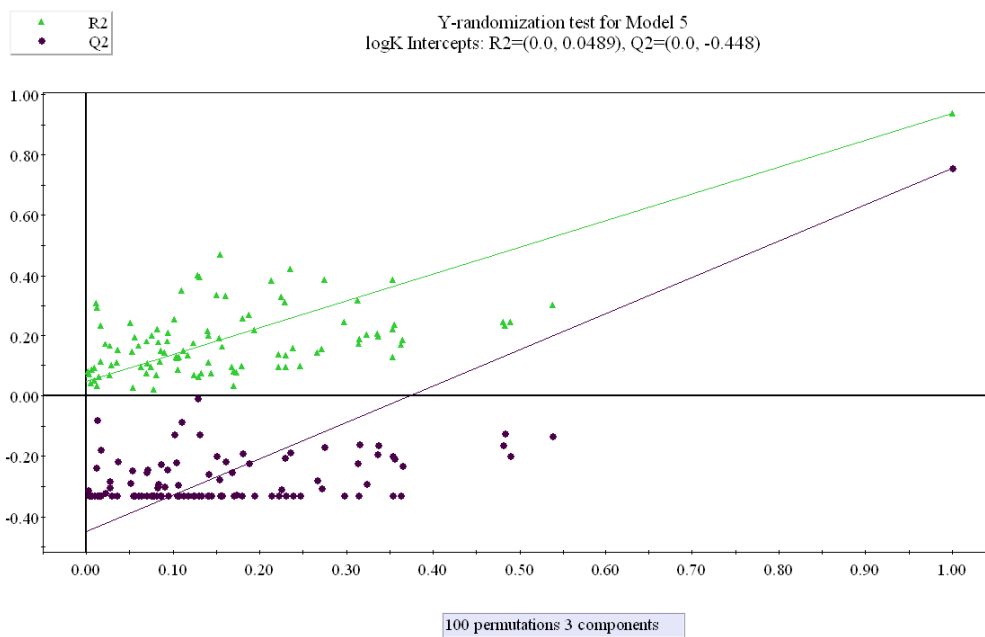


Fig. 15. The randomization plot for the QSPR model 5 derived from PLS analysis.

Box 2

Model1 :

$$\log k = 2.881 + 0.153 \times MLOGP2 + 4.183 \times ETA_Shape_Y$$

$$- 1.472 \times nRNR2 - 11.116 \times X2A - 0.569 \times B07[C - S]$$

$$n_{training} = 30, LV = 4, R^2 = 0.928, R^2_{(adj)} = 0.916, Q^2 = 0.894, S = 0.348, PRESS = 4.452, F = 80.52$$

$$\overline{r^2_{m(LOO)}} = 0.850, \Delta r^2_{m(LOO)} = 0.056, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.900, Q^2_{F2} = 0.899, \overline{r^2_{m(test)}} = 0.869, \Delta r^2_{m(test)} = 0.043, CCC = 0.950$$

MAEbasedcriteria = Moderate

Model2 :

$$\log k = -1.125 + 0.169 \times MLOGP2 + 4.565 \times ETA_Shape_Y - 1.316 \times nRNR2$$

$$- 0.639 \times B07[C - S] + 0.463 \times B04[C - C]$$

$$n_{training} = 30, LV = 4, R^2 = 0.938, R^2_{(adj)} = 0.928, Q^2 = 0.901, S = 0.324, PRESS = 4.175, F = 94.43,$$

$$\overline{r^2_{m(LOO)}} = 0.860, \Delta r^2_{m(LOO)} = 0.036, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.929, Q^2_{F2} = 0.928, \overline{r^2_{m(test)}} = 0.904, \Delta r^2_{m(test)} = 0.054, CCC = 0.965, MAEbasedcriteria = Good$$

Model3 :

$$\log k = -1.291 + 0.190 \times MLOGP2 + 4.871 \times ETA_Shape_Y - 1.887 \times nRNR2$$

$$+ 0.231 \times H - 051 + 0.351 \times B06[C - O]$$

$$n_{training} = 30, LV = 3, R^2 = 0.949, R^2_{(adj)} = 0.943, Q^2 = 0.890, S = 0.287, PRESS = 4.620, F = 161.42$$

$$\overline{r^2_{m(LOO)}} = 0.846, \Delta r^2_{m(LOO)} = 0.063, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.901, Q^2_{F2} = 0.900, \overline{r^2_{m(test)}} = 0.851, \Delta r^2_{m(test)} = 0.061, CCC = 0.956, MAEbasedcriteria = Moderate$$

Model4 :

$$\log k = -0.448 + 0.176 \times MLOGP2 + 5.052 \times ETA_Shape_Y - 1.520 \times nRNR2$$

$$+ 0.240 \times B06[C - O] - 2.245 \times X2A$$

$$n_{training} = 30, LV = 4, R^2 = 0.920, R^2_{(adj)} = 0.907, Q^2 = 0.861, S = 0.366, PERSS = 5.686, F = 72.39$$

$$\overline{r^2_{m(LOO)}} = 0.806, \Delta r^2_{m(LOO)} = 0.059, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.898, Q^2_{F2} = 0.897, \overline{r^2_{m(test)}} = 0.875, \Delta r^2_{m(test)} = 0.065, CCC = 0.953, MAEbasedcriteria = Good$$

Model5 :

$$\log k = -1.117 + 0.179 \times MLOGP2 + 4.564 \times ETA_Shape_Y - 1.437 \times nRNR2$$

$$+ 0.457 \times B06[C - O] - 0.496 \times B07[C - S]$$

$$n_{training} = 30, LV = 3, R^2 = 0.937, R^2_{(adj)} = 0.930, Q^2 = 0.886, S = 0.321, PRESS = 4.803, F = 128.07$$

$$\overline{r^2_{m(LOO)}} = 0.842, \Delta r^2_{m(LOO)} = 0.008, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.923, Q^2_{F2} = 0.922, \overline{r^2_{m(test)}} = 0.905, \Delta r^2_{m(test)} = 0.053, CCC = 0.963, MAEbasedcriteria = Good$$

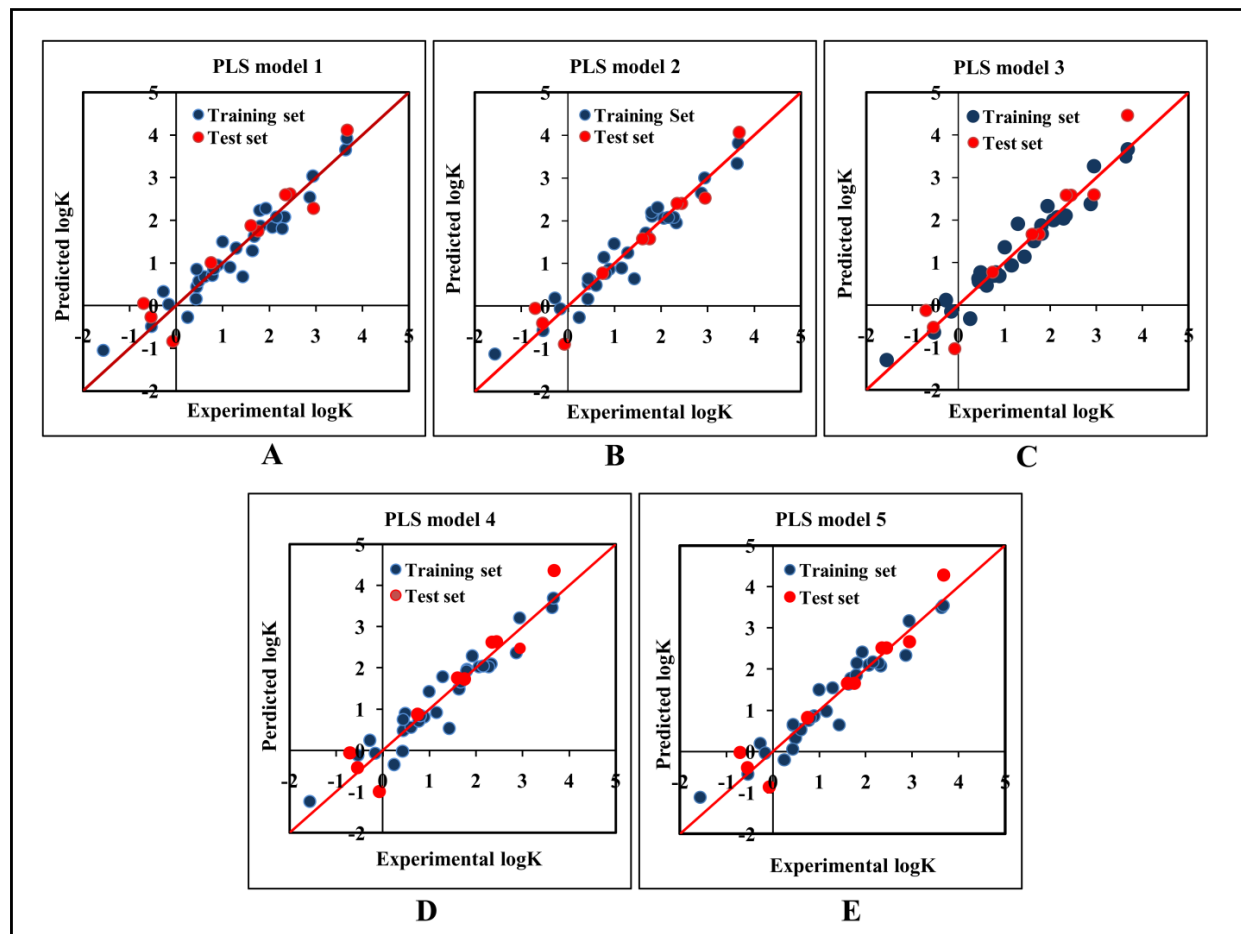


Fig.16. The scatter plot of the observed and the predicted adsorption coefficient ($\log K$) of the developed PLS models (models 1-5).

Table 4.3. Statistical quality and validation parameters obtained from the developed PLS models.

Type of Model	Training set statistics					Test set statistics								
	Model R ²	Model Q ² _(LOO)	MAE _{train}	$\overline{r}_{m(LOO)}^2$	$\Delta r_{m(LOO)}^2$	R ² _{pred} or Q ² _{F1}	Q ² _{F2}	CCC	$\overline{r}_{m(test)}^2$	$\Delta r_{m(test)}^2$	MAE (100%)	MAE (95%)	MAE	
IM1 (LV=4)	0.928	0.894	Good	0.850	0.056	0.900	0.899	0.950	0.867	0.043	0.382	0.338	Moderate	
IM2 (LV=4)	0.938	0.901	Good	0.860	0.036	0.929	0.928	0.965	0.904	0.054	0.275	0.214	Good	
IM3 (LV=3)	0.949	0.890	Good	0.846	0.063	0.901	0.900	0.956	0.851	0.061	0.320	0.249	Moderate	
IM4 (LV=4)	0.920	0.861	Good	0.806	0.059	0.898	0.897	0.953	0.875	0.065	0.359	0.295	Good	
IM5 (LV=3)	0.937	0.886	Good	0.842	0.008	0.923	0.922	0.963	0.905	0.053	0.322	0.263	Good	
CM0						0.914	0.913	0.959	0.894	0.059	0.282	0.219	Good	
CM1						0.912	0.911	0.958	0.890	0.060	0.327	0.268	Good	
CM2						0.913	0.912	0.959	0.889	0.060	0.322	0.262	Good	
CM3						0.938	0.937	0.971	0.903	0.044	0.250	0.189	Good	

CM0 = Ordinary consensus predictions

CM1 = Average of predictions from individual models IM1 through IM5

CM2 = Weighted average predictions from individual models IM1 through IM5

CM3 = Best selection of predictions (compound-wise) from individual models IM1 through IM5

*Note that we have run the “Intelligent consensus predictor tool” using the options, AD: No; Dixon Q-test: No; Euclidean distance cut-off: 0.4

Table 4.4. Validation results for the final PLS models obtained according to Golbraikh and Tropsha's criteria.

Sl. No.	Models	r^2	$[(r^2-r_0^2)/r^2]$	$[(r^2-r_0'^2)/r^2]$	k	k'	Remarks
	Threshold value	>0.6	<0.1	<0.1	$0.85 < k < 1.15$	$0.85 < k' < 1.15$	
1	M1	0.906	-0.094	-0.088	0.940	1.013	Passed
2	M2	0.932	-0.065	-0.068	0.981	0.982	Passed
3	M3	0.927	-0.065	-0.072	0.910	1.054	Passed
4	M4	0.918	-0.079	-0.082	0.912	1.049	Passed
5	M5	0.933	-0.065	-0.067	0.936	1.031	Passed

Table 4.5. Definition and contribution of different descriptors obtained from five PLS models.

Sl. no.	Name of the descriptors	Contribution	Discussion	Mechanism	Frequency
1	MLOGP2	+ve	Squared Moriguchi octanol-water partition coeff. ($\log P^2$)	Hydrophobic interaction	5
2	ETA_Shape_Y	+ve	$ETA_Shape_Y = (\sum \alpha)_Y / \sum \alpha$, $(\sum \alpha)_Y$ stands for summation of α values of the vertices that are joined to three other non-hydrogen vertices in the connected molecular graph. Gives a measure of molecular shape.		5
3	nRNR2	-ve	Number of tertiary amines (aliphatic)	Unable to form hydrogen bond due to the absence of free hydrogen atom.	5
4	B07[C-S]	-ve	Presence/absence of C-S at topological distance 7		3
5	B04[C-C]	+ve	Presence/absence of C-C at topological distance 4	Hydrophobic interaction	1
6	X2A	-ve	average connectivity index of order 2	Hydrophobic interaction	2
7	H-051	+ve	H attached to alpha-C	Electrostatic interaction. H atoms attached	1

				to α carbon atom can easily donate protons and may involve in electrostatic interaction.	
8	B06[C-O]	+ ve	Presence/absence of C-O at topological distance 6	Formation of hydrogen bond	3

4.2.1. Descriptors related to hydrophobic interaction

The descriptor MLOGP2, represents squared Moriguchi octanol water partition coefficient, calculated from the regression equation of Moriguchi logP model (Moriguchi et al., 1994; Ojha and Roy, 2018) consisting of 13 parameters.

The positive regression coefficient of this descriptor indicates that hydrophobicity is directly correlated with the adsorption property of organic pollutants. Thus, the organic pollutants bearing highly hydrophobic property can easily get adsorbed onto SWCNTs as evidenced by the compounds **25 (Phenanthrene)**, **2 (1,2,4-trichlorobenzene)** and **17 (Ethinyl estradiol)** as their corresponding MLOGP2 values are 18.762, 16.507 and 16.033 respectively, whereas less hydrophobic organic pollutants are poorly adsorbed onto SWCNTs as evidenced by the compounds **6 (4,6-Diaminopyrimidine)**, **26 (Pyrimidine)** and **8 (Aniline)**, as their corresponding MLOGP2 values are 0.256, 0 and 2.268 respectively. Therefore, it can be inferred that the hazardous SOCs get adsorbed onto the SWCNTs through hydrophobic interactions. For proper adsorption, synthetic organic chemicals should be hydrophobic in nature. MLOGP2 is not strictly a 2D descriptor as its numerical value depends on intermolecular H-bonds (as it depends on molecular conformation).

The next descriptor B04[C-C] is a 2D binary fingerprint descriptor corresponding to presence/absence of C-C bond at topological distance 4. The positive regression coefficient of this descriptor indicates that presence of C-C bond at the topological distance 4 is important for good adsorption of SWCNTs. The descriptor is related to the size of molecule. If the size of the molecule increases, hydrophobic interaction of the molecule with SWCNTs also increases hence adsorption coefficient also increases. As for example, compounds **25 (phenanthrene)**, **22 (Naphthalene)** and **17 (ethinyl estradiol)** contain single C-C bond at the topological distance 4,

and their corresponding adsorption coefficient values are 3.67, 1.8 and 2.87 respectively (higher adsorption coefficient values). While absence of such fragment decreases the adsorption of organic pollutants to SWCNTs as shown in compounds **26** (pyrimidine), **6** (4,6-diaminopyrimidine) and **8** (aniline) (adsorption coefficient -1.56, -0.27 and -0.16 respectively). Another significant descriptor, X2A, indicates average connectivity index of order 2, it encodes the ‘chi’ value across two bonds, which can be calculated on basis of Kier and Hall’s connectivity index and defined in the following equation:

$${}^2X = \sum_{b=2}^B (\delta_i \cdot \delta_j \cdot \delta_k)_b^{-0.5}$$

Here, b runs over the 2nd order sub graphs having n vertices with B edges, δ_i , δ_j and δ_k are number of other vertices attached to vertex i, j and k respectively. This descriptor has a negative contribution towards the adsorption coefficient (logK) of organic pollutants by SWCNTs as evidenced by the negative regression coefficient. This indicates that the adsorption property of hazardous SOCs decreases with an increase in the numerical value of this descriptor. For example, compounds **26** (Pyrimidine), **8**(Aniline) and **6**(4,6-Diaminopyrimidine) have descriptor values 0.354,0.343 and 0.338 in that order, and their corresponding adsorption coefficient values are -1.56,-0.16 and -0.27 respectively. If we consider compounds **25**(Phenanthrene) and **17**(Ethinyl estradiol), their descriptor values are less (0.272 and 0.257 respectively), thus their corresponding adsorption coefficient value is high (logK value is 3.67 and 3.64 respectively).

4.2.1.1 Mechanistic interpretation of hazardous SOCs containing higher and lower adsorption coefficient based on hydrophobic interaction

Phenanthrene (Compound **25**) (shown in **Fig. 17**) is a poly aromatic hydrocarbon (PAH) and non ionic in nature. Its MLOGP2 value is 18.76. Due to its hydrophobic property, it can strongly interact with hydrophobic surface of SWCNTs. The B04[C-C] value for phenanthrene is 1 (positive contribution) and X2A value is also low (0.272) (negative contribution) which supports that strong interactions occur between phenanthrene and SWCNTs (Chen et al., 2008). Phenanthrene has an intensive electron donor property. Phenanthrene donates its π -electron and can easily get converted to a cationic form, and thus more easily interacts with the surface of functionalized (with -COOH) SWCNTs (Gotovac et al., 2007).

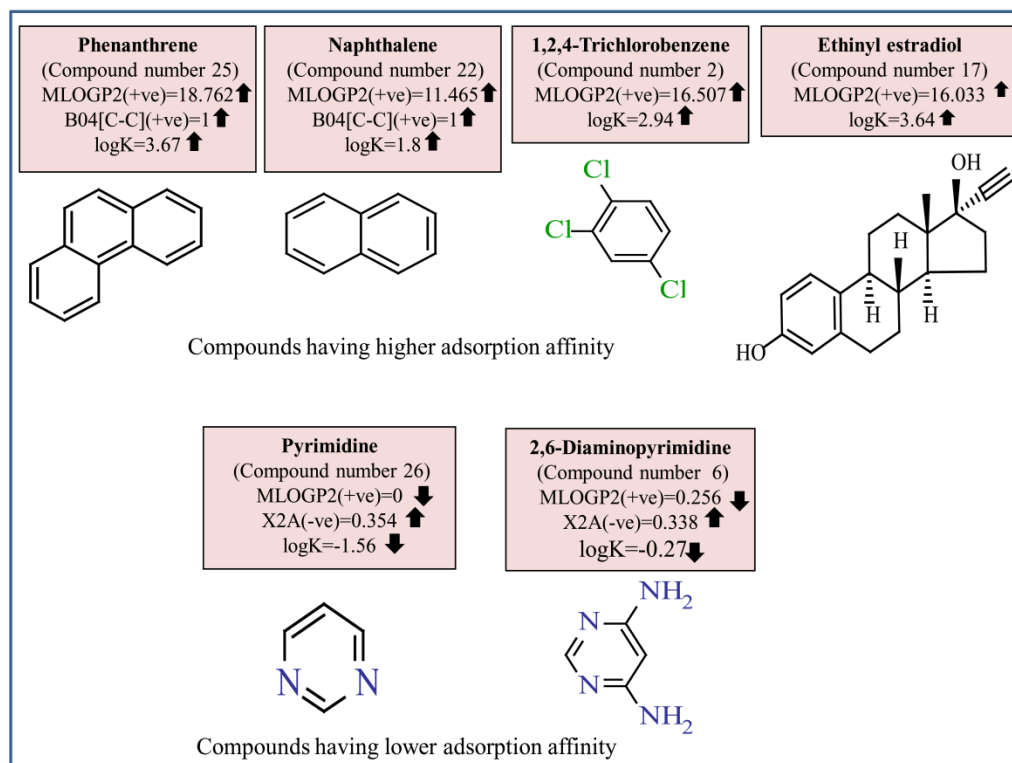


Fig. 17. Contribution of descriptors MLOGP2, B04[C-C] and X2A related to hydrophobic interaction between SWCNTs and synthetic organic chemicals

When we compare between Phenanthrene (compound **25**) and Naphthalene (compound **22**) (**Fig. 17**), both are poly aromatic hydrocarbons (PAHs) and non ionic in nature. Their corresponding MLOGP2 values are 18.76 and 11.46 respectively. Due to their hydrophobic property, they can strongly interact with hydrophobic surface of SWCNTs. The B04[C-C] value for phenanthrene and naphthalene is also 1. Both phenanthrene and naphthalene show strong interactions with SWCNTs. They only differ in polarity and electron donor-acceptor ability (Chen et al., 2008). Thus, the adsorption coefficient of phenanthrene ($\log K=3.67$) is higher than naphthalene ($\log K=1.8$) because naphthalene contains less number of aromatic rings and is therefore less hydrophobic than phenanthrene.

Another example, 1,2,4-Trichlorobenzene (compound number **2**) (**Fig. 17**) present in the data set, shows higher adsorptive property ($\log K=2.94$) due to its bulkiness. As the size of chlorine atom is high, substitution of three chlorine atoms in the benzene ring is responsible for the bulkiness of molecule. SWCNTs shows molecular sieving effect, so based on the unit surface area (Chen et al., 2007), it shows a stronger affinity towards trichlorobenzene. As the volume of molecule

increases, the molecule would preferentially like to be in the non-polar phase. As a result, the partition-coefficient also increases (logP value for benzene and chlorobenzene are 2.13 and 2.84 respectively). The numerical value for MLOGP2 (squared Moriguchi octanol-water partition coefficient) of this compound is also high (21.476). At the same time, the molecule contains a benzene ring, which is responsible for π - π interaction with graphene sheets of carbon nanotubes. Chlorobenzene is also able to participate in hydrogen bonding (though moderately). Thus, the hazardous SOCs containing bulky hydrophobic groups (reflected in the MLOGP2 descriptor) is influential for adsorption of organic pollutants to SWCNTs.

Another compound, Ethinyl estradiol (compound number **17**) (**Fig. 17**) shows good adsorptive property to the SWCNTs due to its hydrophobicity (Borisover and Graber, 2003). The higher MLOGP2 value (16.033), presence of single C-C fragment at topological distance 4 and low X2A value (0.257) also give evidences for its hydrophobicity as well as higher adsorptive property. Another mechanism, π - π electron donor-acceptor interaction, also supports the higher adsorptive property of Ethinyl estradiol onto SWCNTs. Due to the presence of two phenolic groups (charge donor), it can strongly interact with SWCNTs through π - π electron donor-acceptor interaction. (Chen et al., 2007; Zhao et al., 2002). Hydrogen bonding between two -OH groups of Ethinyl estradiol and SWCNTs is also possible, which supports a favorable mechanism for adsorption of this compound with SWCNTs (Pan and Xing., 2008).

We can consider 4,6-diaminopyrimidine(compound number**6**) in comparison to pyrimidine (compound **26**) (**Fig. 17**) (lower range of adsorption coefficient). Pyridine is an electron deficient system in comparison to benzene. Thus, it can weaken the π - π electron donor acceptor interaction with SWCNTs (Wang et al., 2010b). On the other hand, 4,6-diaminopyrimidine contains two amino groups which are strong electron donating groups and increase the electron density of aromatic ring. They may form stronger π - π interaction as compared to pyrimidine. The numerical value of X2A descriptor for 4,6-Diaminopyrimidine and Pyrimidine are 0.338 and 0.358 respectively. For all these reasons, the adsorption coefficient value of 2,6-diaminopyrimidine (logK=-0.27) and pyrimidine (logK=-1.56, lowest active compound present in dataset) are in the lower range.

Thus, the information obtained from the descriptors, MLOGP2, B04[C-C] and X2A suggested that the organic pollutants can adhere to the surface of SWCNTs by strong hydrophobic interaction.

4.2.2. Descriptors related to electrostatic interaction

The descriptor H-051 indicates the number of H atoms attached to α carbon atom. Such H atoms are very active in nature. They can easily donate protons and may involve in electrostatic interaction between SWCNTs and synthetic organic chemicals. The positive regression coefficient of this descriptor indicates that organic pollutants contain higher number of such hydrogen atoms have good adsorption property as shown in compounds **30 (Tylosin)** and **27 (Sulfamethoxazole)**. The numerical values of H-051 for compounds **30** and **27** are 5 and 3, respectively, and their corresponding logK values are 0.43 and 1.43, respectively. On the other hand, in case of compounds **6(4,6-Diaminopyrimidine)** and **8(Aniline)**, the adsorption coefficient values (logK values are -0.27 and -0.16 respectively) decrease due to the absence of α H atoms.

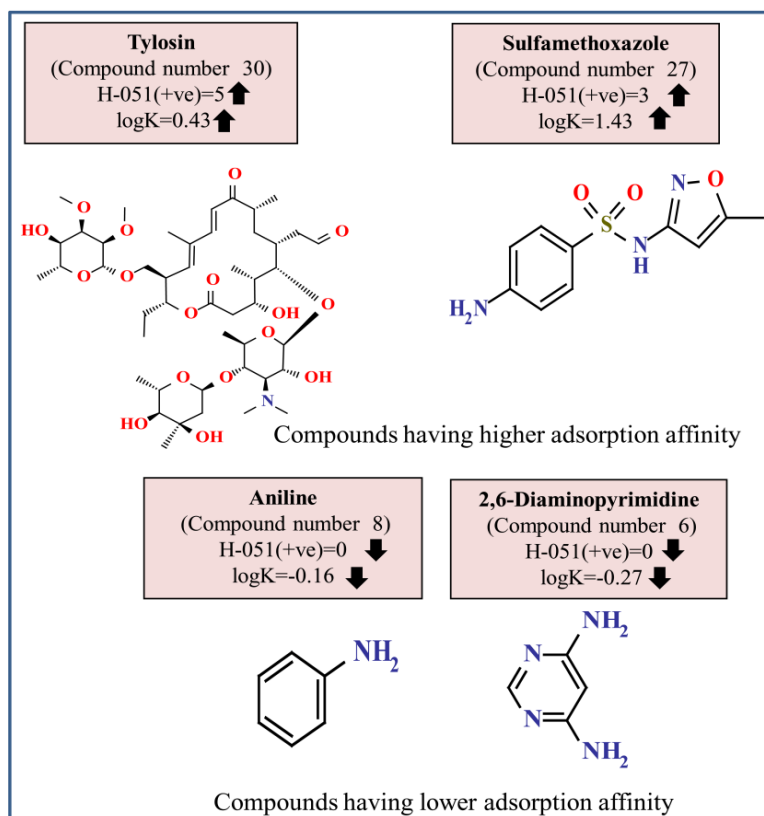


Fig. 18. Contribution of descriptor H-051 related to electrostatic interaction between SWCNTs and synthetic organic chemicals.

4.2.2.1. Mechanistic interpretation of hazardous SOCs containing higher and lower adsorption coefficient based on electrostatic interaction

Molecules like Sulfamethoxazole and Tylosin (**Fig. 18**) are large in size. Larger molecules adopt themselves in such a manner that they can easily fit with the curvature surface and make stable complex with CNTs (Zhou et al., 2001; Richard et al., 2003; Karajanagi et al., 2004; Gurevitch and Srebnik, 2008). The adsorption energy provides the steric energy required for the conformational changes of organic molecules (Pan et al., 2008). Sulfamethoxazole is well adsorbed to the SWCNTs as its corresponding H-051 value is 3. In case of Tylosin, its descriptor value for H-051 is 5 but its adsorption coefficient value is moderate as compared to Sulfamethoxazole, because its MLOGP2 value is very less (1.604). On the other hand, 4,6-diaminopyrimidine (compound **6**) and aniline (compound **8**) show poor adsorption affinity towards the SWCNTs as discussed above.

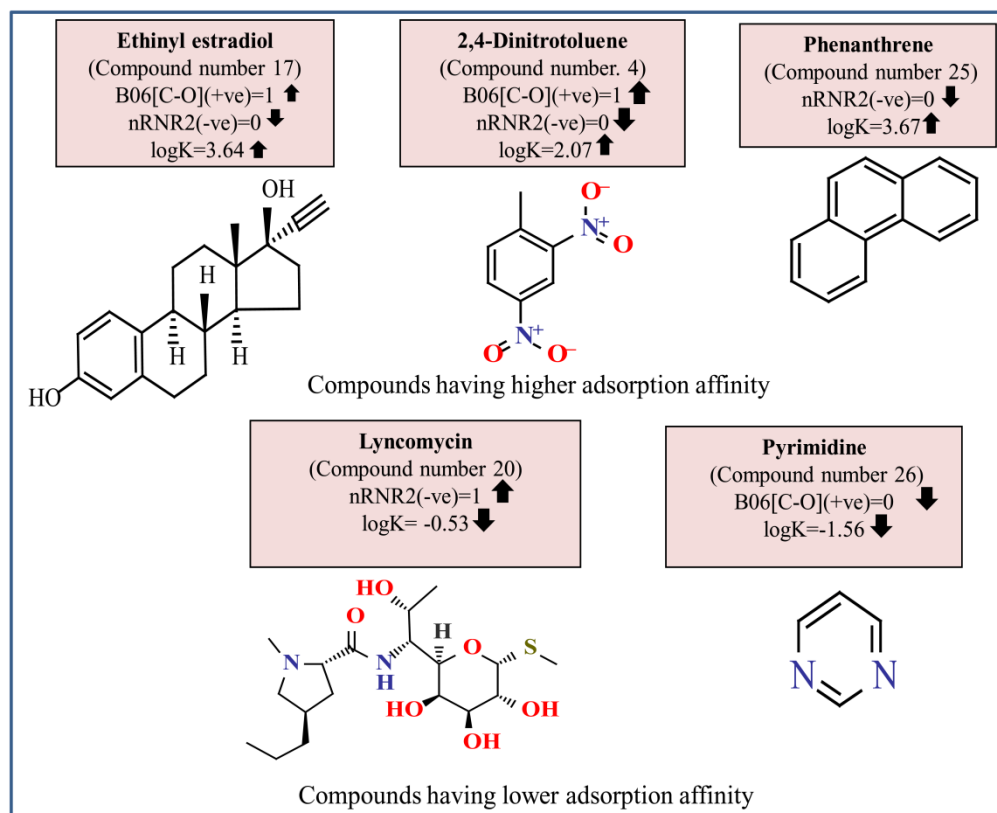


Fig. 19. Contribution of descriptors B06[C-O] and nRNR2 related to hydrogen bonding interaction between SWCNTs and synthetic organic chemicals.

4.2.3. Descriptors related to hydrogen bonding interaction

2D binary fingerprints descriptor, B06[C-O], indicates presence/absence of C-O bond at topological distance 6. B06[C-O] have a positive regression coefficient which implies that presence of C-O fragment at topological distance 6 is beneficial for the adsorption of organic pollutants to SWCNT, as the C-O fragment is capable of forming hydrogen bonds with SWCNTs. For example, each of compounds **17 (Ethinyl estradiol)**, **16(Diuron)** and **4(2,4-Dinitrotoluene)** contains single C-O fragment at the topological distance 6, and their corresponding adsorption coefficient values are 3.64, 2.28 and 2.07, whereas compounds **26(Pyrimidine)**, **6(4,6-Diaminopyrimidine)** and **8(Aniline)** have lower adsorption affinity to the SWCNTs due to the absence of such fragment (**Fig. 19**). Compounds **17 (Ethinyl estradiol)** and **4 (2,4-Dinitrotoluene)** contain C-O fragment at topological distance 6, which indicates that they are capable of forming hydrogen bonds with functionally modified SWCNTs. Therefore, their adsorption coefficient values are in higher (3.64 and 2.07 respectively) range. On the other hand, both compounds **30 (Tylosin)** and **20 (Lyncomycin)** contain one aliphatic tertiary amine, so, they are not capable of forming any hydrogen bond and thus adsorption coefficient value is less (briefly discuss in ETA_Shape_Y).


The descriptor, nRNR2 indicates the number of aliphatic tertiary amines present in a compound. Due to absence of free hydrogen atoms, tertiary amine does not act as a hydrogen bond donor like primary or secondary amine. The negative regression coefficient of this descriptor indicates that higher number of aliphatic tertiary amine weakens the interaction between synthetic organic chemicals and SWCNTs and vice versa. For example, compounds **30 (Tylosin)** and **20 (Lyncomycin)** have descriptor value 1, and their corresponding adsorption coefficient is less (0.43 and -0.53 respectively), while compounds with lower descriptor value (no such group) have higher adsorption coefficient as shown in case of compounds **17(Ethinyl estradiol)** and **25(Phenanthrene)** (logK values 3.64 and 3.67 respectively). If we consider compounds **30(Tylosin)** and **20(Lyncomycin)**, Tylosin is moderately active as compared to Lyncomycin because the latter contains 5 α -H atom (H-051 value is 5) and C-O fragment at topological distance 6 (B06[C-O] value is 1).

4.2.4. Other modeled descriptors essential for adsorption of hazardous SOCs to SWCNTs

The descriptor ETA_Shape_Y is a first generation extended topochemical atom index. ETA_Shape_Y (Roy, 2015) can be calculated by using the following formula:

$$\text{ETA_Shape_Y} = (\sum \alpha)_Y / \sum \alpha$$

$(\sum \alpha)_Y$ stands for summation of α value (a volume measure) of the vertices that are joined to three other non-hydrogen vertices in the connected molecular graph and forming a Y-shaped

structural fragment like tertiary groups (). It gives a measure of molecular shape. The positive regression coefficient of this descriptor indicates that the branching is directly correlated with adsorption of organic pollutants to SWCNTs. The higher degree of branching plays a crucial role to enhance the adsorption affinity of synthetic organic chemicals to SWCNTs as evidenced by the compounds **4** (2, 4-Dinitrotoluene) and **16** (Diuron) (corresponding logK values are 2.07 and 2.28 respectively) (**Fig. 20**) with their corresponding descriptor values are in the higher range (0.408 and 0.340 respectively).

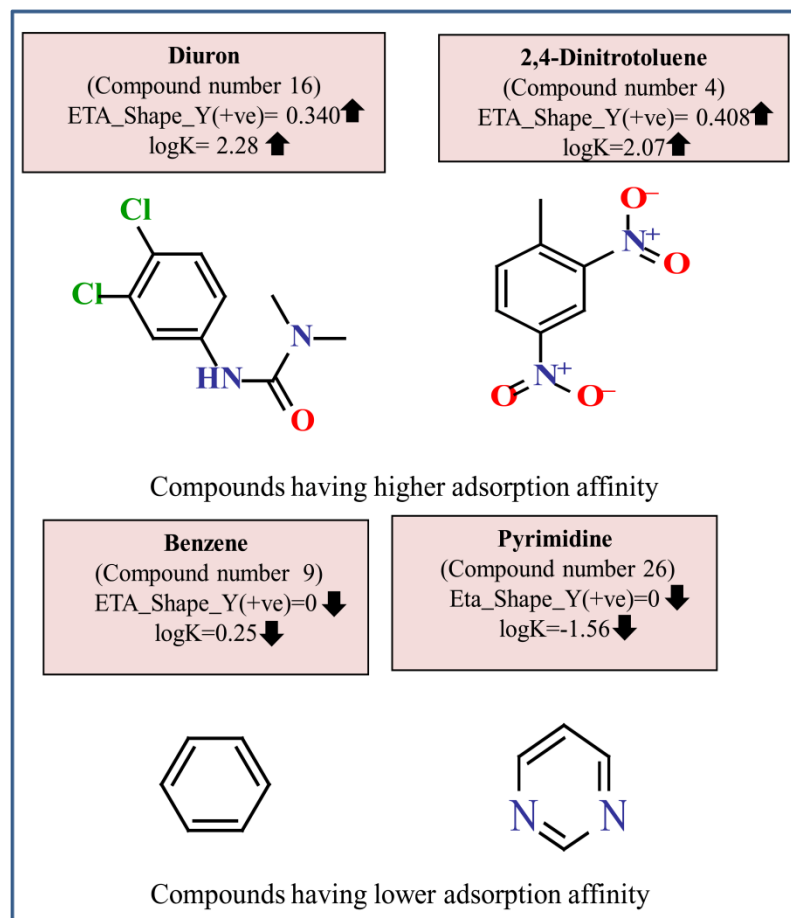


Fig. 20. Contribution of descriptor Eta_Shape_Y for adsorption of synthetic organic chemicals onto SWCNTs

On the other hand, compounds **26** (Pyrimidine) and **9** (Benzene) have the descriptor value of 0 and thus, their corresponding adsorption coefficient is also low (-1.56 and 0.25 respectively). Between 2,4-Dinitrotoluene (compound **4**) and Diuron (compound **16**), the adsorption affinity of Diuron is higher (though its ETA_Shape_Y value is comparatively less) than that of 2,4-Dinitrotoluene due to its hydrophobicity (MLOGP2 values are 7 and 5.02 respectively).

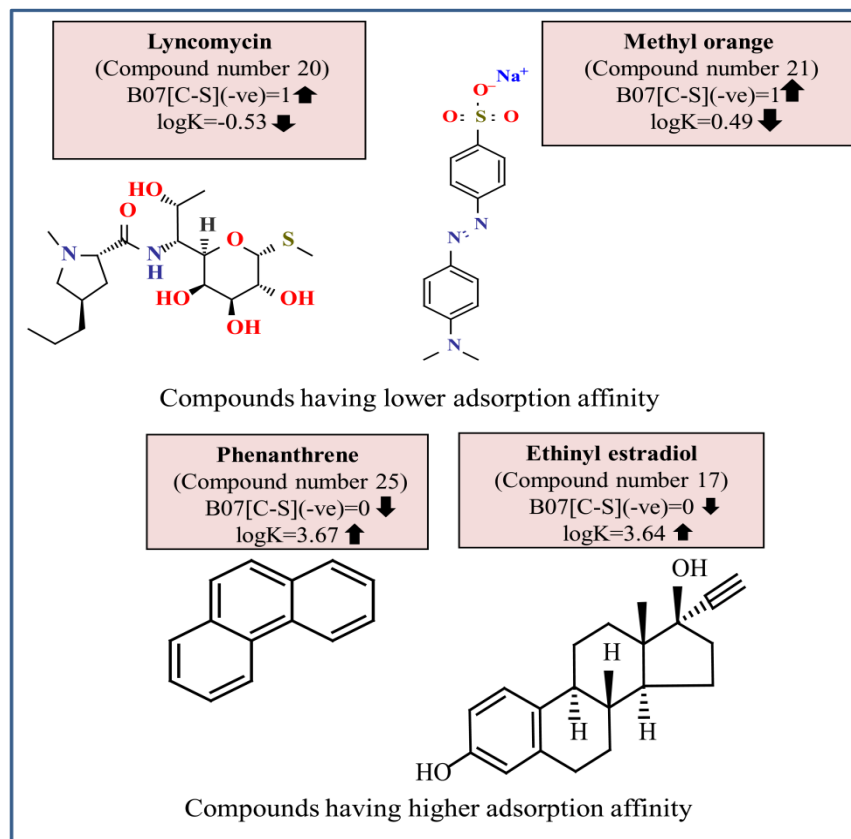


Fig. 21. Contribution of descriptor B07[C-S] for adsorption of synthetic organic chemicals onto SWCNTs

Another descriptor, B07[C-S], indicates the presence/absence of C-S at topological distance 7. The negative regression coefficient of this descriptor indicates that presence of C-S fragment at the topological distance 7 in hazardous SOCs is not beneficial for the adsorption of SWCNTs as evidenced by compounds **20 (Lincomycin)** and **21 (Methyl Orange)** (they contain C-S fragment at the topological distance 7, and their corresponding adsorption coefficient value is -0.53 and 0.49 respectively). On the other hand, compounds **25 (Phenanthrene)** and **22 (naphthalene)** (**Fig. 21**) do not contain any such fragments, so their adsorption coefficient value is higher.

The adsorption coefficient value of Lyncomycin (compound number **20**) is low (in spite of high ETA_Shape_Y) as compared to Methyl orange (-0.53) because of its low hydrophobicity (MLOGP2 is 0.538).

Chapter 5

Conclusion



5. CONCLUSION

Innovative scientific solutions are the key strength for meeting the needs of novel applications. Nanomaterials are used for pollution management, because they contain high surface area, high adsorption affinity towards the organic contaminants, and they can be modified in several ways to increase their selectivity towards specific target pollutants (Chen et al., 2007). Carbon nanotubes (CNTs) are such type of nanomaterials, which have recently gained special attention from the researchers due to their smaller size, large specific surface area, hollow and layered structure, responsible for their extraordinary adsorption property. Hence, it will be important to explore the chemical attributes of the CNTs for effective adsorption in greater extent so that they can be designed for better application.

5. 1. Predictive Quantitative Structure-Property Relationship (QSPR) Modeling for Adsorption of Organic Pollutants by Carbon Nanotubes (CNTs)

The present work deals with a variable selection strategy and development of QSPR model to find out the important structural parameters of organic contaminants which are essential to alter the adsorption property of organic contaminants related to the specific surface area of MWCTs. We have developed a predictive QSPR model using diverse classes of organic contaminants with reported experimental $\log K_{SA}$ values. Various type of descriptors including constitutional indices, ring descriptors, connectivity indices, functional group count, atom centered fragments, atom type E-state indices, 2D atom pairs and extended topochemical atom (ETA) indices descriptors were used to developed the predictive models. The whole dataset was divided into a training set and a test set based on modified *k*-Medoids method. We have developed statistically robust QSPR models using different chemometric tools like stepwise regression, best subset selection and intelligent consensus predictor (ICP). We have checked the statistical quality of the final models using various internal and external parameters like Q^2 , R^2_{pred} , Q^2_{F2} , CCC, r_m^2 metrics and MAE based criteria. The MAE based criteria were employed in case of both internal and external validation.

The insights obtained from developed model suggested that: i) Electronegative atoms like N, Cl, O, F etc. will increase adsorption of organic contaminants by MWCNTs. ii) Average connectivity index of order one and two causes decrease in adsorption affinity of organic contaminants. iii) Presence of higher number of primary alcohol will decrease the adsorption property of organic contaminants. iv) Hydrophobicity is an important structural parameter to enhance the adsorption property of organic contaminants. The highly hydrophobic organic pollutants can adsorb easily by MWCNTs. v) To enhance the adsorption property of organic contaminants by MWCNTs, the number of (O-Cl) fragment at the topological distance 4 should be lower. vi) To enhance the adsorption property of organic contaminants, the number of (Cl-Cl) fragment at the topological distance 5 should be higher. vii) The sum of the topological distances between two nitrogen atoms should be high for increase the adsorption property of organic contaminants. viii) Presence of oxygen atoms in a molecule is important to enhance the adsorption of organic contaminants. ix) Sum of topological distances between oxygen and chlorine atoms should be less to increase the adsorption property of organic contaminants. Thus, this work provides an understanding of the important structural requirements or essential molecular properties and the requisite features of molecules that is important to increase or decrease the adsorption of organic contaminants. The developed models could be useful as preliminary support tools for the identification and prioritization of new potential organic pollutants among already existing chemicals as well as “screening prior to synthesis” procedures to avoid the production, and consequent release into the environment, of new organic pollutants. The models provide an important guidance for the chemist to increase the efficient application of MWCNTs which may be useful for reducing the environmental pollution.

5. 2. Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs

In this present study, we have developed PLS QSPR models for a dataset containing 40 diverse synthetic organic chemicals (herbicides, fungicides, EDCs, PAH, contrasting agent, dyes) having defined adsorption affinity for SWCNTs, by applying different strategies. We have validated the models using various internal and external validation parameters, which showed that the models were statistically significant. We have also checked the consensus predictivity of all the individual models (IM1-IM5) using “Intelligent consensus predictor” tool and found that the

consensus predictivity of the test set compounds was better than the individual models based on MAE based criteria as depicted in Table 4.3 (winner model is CM3)

The present study shows how the chemical and structural features of diverse hazardous SOCs alter the adsorption property to SWCNTs. From the insights obtained from five PLS models, we have concluded that hydrophobic surface of the molecules, molecular shape and degree of branching, presence of two carbon atoms at topological distance 4, number of H atom attached with α -C atom, presence of carbon and oxygen atom at the topological distance 6 can enhance the adsorption of hazardous SOCs to the SWCNTs. On the other hand, number of tertiary aliphatic amine, presence of carbon and sulphur at topological distance 7 may be detrimental for the adsorption of hazardous SOCs to the SWCNTs. The adsorption mechanism as evidenced from different contributed descriptors is depicted in **Fig 22**. Among all the above mentioned descriptors, MLOGP2 has the strongest impact on the adsorption of hazardous SOCs onto SWCNTs. The conclusions drawn in the present study are also supported by several studies published previously. Sun et al. (Sun et al., 2012) and Wang et al. (Wang et al., 2010a) suggested that hydrophobic interaction is very crucial for adsorption of hazardous SOCs to SWCNTs. Ding et al. (Ding et al., 2016b) reported that the potency of adsorption is positively correlated with hydrophobicity and it is the principal reason behind the adsorption capacities of different hazardous SOCs. Thus, the developed models give information about the important structural requirements or essential molecular properties and the requisite features of molecules that are important to increase or decrease the adsorption of the hazardous SOCs onto SWCNTs. The information obtained from the developed models may be useful for the management of environmental pollution.

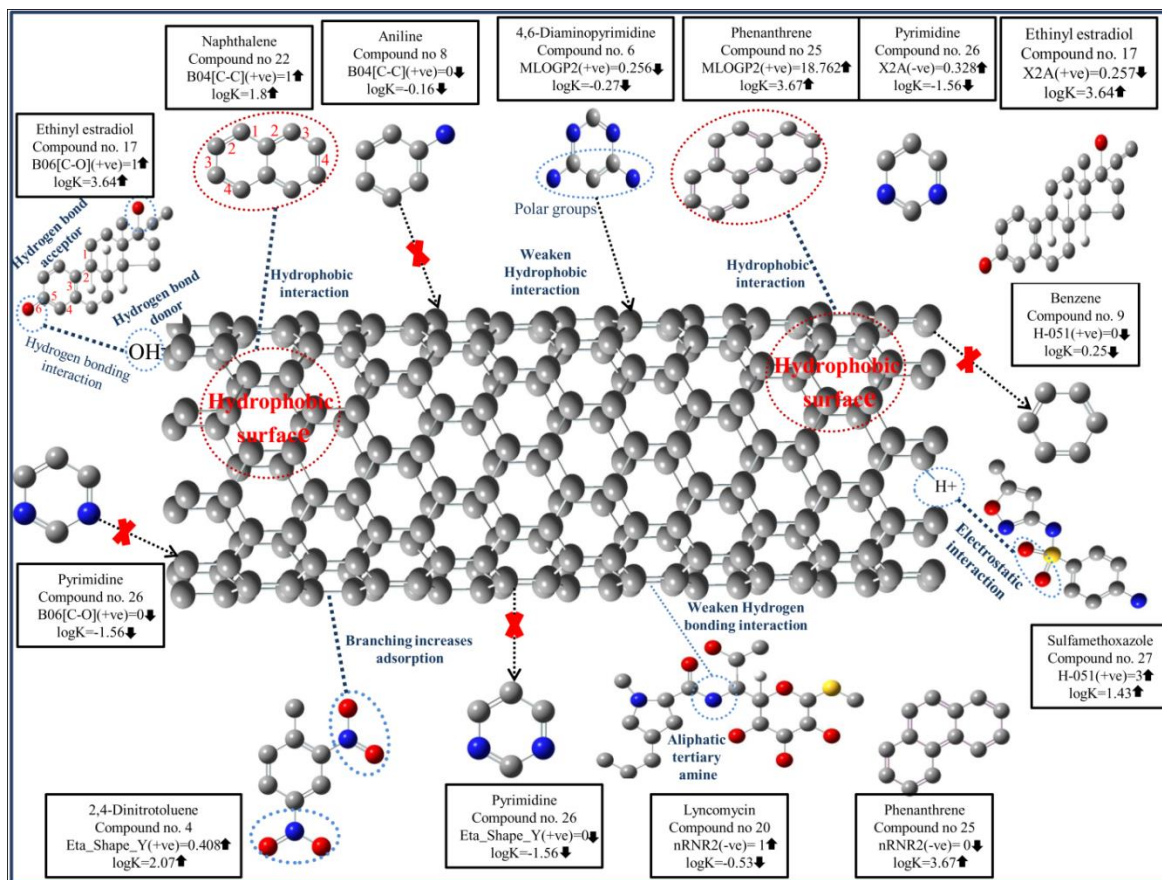
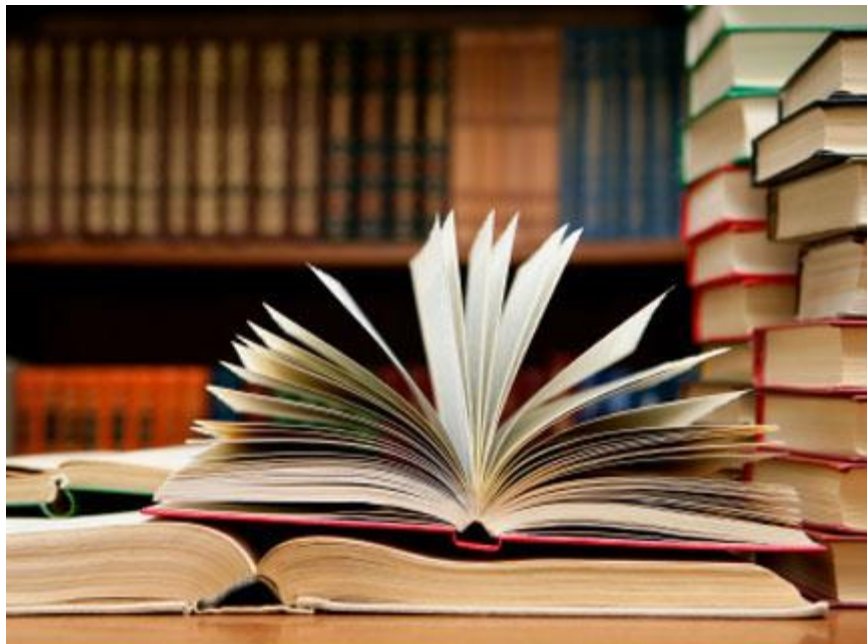


Fig. 22. Adsorption mechanism of contributed descriptors for the adsorption of synthetic organic chemicals onto SWCNTs/functionalized SWCNTs.

To the best of our knowledge, this work presents the QSPR modeling for adsorption coefficient of organic chemicals towards carbon nanotubes which may serve as an efficient tool giving some insight for designing and synthesis of potent molecules to address the increasing threat of environmental pollution throughout the world. Furthermore, the employed chemical descriptors, recognized as the crucial component in QSPR analysis, are based on two dimensional representation of molecular structures allowing prompt application for unknown molecules.

Chapter 6



REFERENCES

6. REFERENCE

- Ahmadi, S. and Akbari, A., 2018. Prediction of the adsorption coefficients of some aromatic compounds on multi-wall carbon nanotubes by the Monte Carlo method. *SAR and QSAR in Environmental Research*, 29(11), pp.895-909.
- Apul, O.G., Wang, Q., Shao, T., Rieck, J.R. and Karanfil, T., 2012. Predictive model development for adsorption of aromatic contaminants by multi-walled carbon nanotubes. *Environmental science & technology*, 47(5), pp.2295-2303.
- Al-Saidi, H.M., Abdel-Fadeel, M.A., El-Sonbati, A.Z. and El-Bindary, A.A., 2016. Multi-walled carbon nanotubes as an adsorbent material for the solid phase extraction of bismuth from aqueous media: Kinetic and thermodynamic studies and analytical applications. *Journal of Molecular Liquids*, 216, pp.693-698.
- Baughman, R.H., Zakhidov, A.A. and De Heer, W.A., 2002. Carbon nanotubes--the route toward applications. *science*, 297(5582), pp.787-792.
- Borisover, M. and Graber, E.R., 2003. Classifying NOM- organic sorbate interactions using compound transfer from an inert solvent to the hydrated sorbent. *Environmental science & technology*, 37(24), pp.5657-5664.
- Benning, P.J., Poirier, D.M., Ohno, T.R., Chen, Y., Jost, M.B., Stepniak, F., Kroll, G.H., Weaver, J.H., Fure, J. and Smalley, R.E., 1992. C 60 and C 70 fullerenes and potassium fullerides. *Physical Review B*, 45(12), p.6899.

- Besalu, E., de Julian-Ortiz, J.V. and Pogliani, L., 2007. Trends and plot methods in MLR studies. *Journal of chemical information and modeling*, 47(3), pp.751-760.
- Cao, A., Zhu, H., Zhang, X., Li, X., Ruan, D., Xu, C., Wei, B., Liang, J. and Wu, D., 2001. Hydrogen storage of dense-aligned carbon nanotubes. *Chemical physics letters*, 342(5-6), pp.510-514.
- Carter, A.D., 2000. Herbicide movement in soils: principles, pathways and processes. *Weed Research (Oxford)*, 40(1), pp.113-122.
- Chen, J., Chen, W. and Zhu, D., 2008. Adsorption of nonionic aromatic compounds to single-walled carbon nanotubes: effects of aqueous solution chemistry. *Environmental science & technology*, 42(19), pp.7225-7230.
- Chen, W., Duan, L. and Zhu, D., 2007. Adsorption of polar and nonpolar organic chemicals to carbon nanotubes. *Environmental science & technology*, 41(24), pp.8295-8300.
- Chayawan, V., 2016. Quantum-mechanical parameters for the risk assessment of multi-walled carbon-nanotubes: A study using adsorption of probe compounds and its application to biomolecules. *Environmental Pollution (1987)*, 218, pp.615-624.
- Consonni, V., Ballabio, D. and Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques. *Journal of chemometrics*, 24(3-4), pp.194-201.
- Consonni, V., Ballabio, D. and Todeschini, R., 2009. Comments on the definition of the Q² parameter for QSAR validation. *Journal of chemical information and modeling*, 49(7), pp.1669-1678.
- Crosland, M.P., 1959. The use of diagrams as chemical 'equations' in the lecture notes of William Cullen and Joseph Black. *Annals of Science*, 15(2), pp.75-90.

- Das, R.N., 2016. Exploring “ETA” Indices for Effective Encoding of Chemical Information in Modeling Different Toxicity Endpoints of Ionic Liquids.
- Das, S., 2013. A review on Carbon nano-tubes-A new era of nanotechnology. *Int J Emer Tech Adv Eng*, 3(3), pp.774-783.
- Das, S., Ojha, P.K. and Roy, K., 2017a. Development of a temperature dependent 2D-QSPR model for viscosity of diverse functional ionic liquids. *Journal of Molecular Liquids*, 240, pp.454-467.
- Das, S., Ojha, P.K. and Roy, K., 2017b. Multilayered variable selection in QSPR: a case study of modeling melting point of bromide ionic liquids. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, 2(1), pp.106-124.
- Dearden, J.C., 2003. In silico prediction of drug toxicity. *Journal of computer-aided molecular design*, 17(2-4), pp.119-127.
- Ding, H., Chen, C. and Zhang, X., 2016a. Linear solvation energy relationship for the adsorption of synthetic organic compounds on single-walled carbon nanotubes in water. *SAR and QSAR in Environmental Research*, 27(1), pp.31-45.
- Ding, H., Li, X., Wang, J., Zhang, X. and Chen, C., 2016b. Adsorption of chlorophenols from aqueous solutions by pristine and surface functionalized single-walled carbon nanotubes. *Journal of Environmental Sciences*, 43, pp.187-198.
- Ebbesen, T.W., Lezec, H.J., Hiura, H., Bennett, J.W., Ghaemi, H.F. and Thio, T., 1996. Electrical conductivity of individual carbon nanotubes. *Nature*, 382(6586), p.54.
- Falvo, M.R., Clary, G.J., Taylor II, R.M., Chi, V., Brooks Jr, F.P., Washburn, S. and Superfine, R., 1997. Bending and buckling of carbon nanotubes under large strain. *Nature*, 389(6651), p.582.

- Ferner, D.J., 2001. Toxicity, heavy metals. *Med. J*, 2(5), p.1.
- Geladi, P. and Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185, pp.1-17.
- Garg, U.K., Kaur, M.P., Garg, V.K. and Sud, D., 2007. Removal of hexavalent chromium from aqueous solution by agricultural waste biomass. *Journal of Hazardous materials*, 140(1-2), pp.60-68.
- Gotovac, S., Yang, C.M., Hattori, Y., Takahashi, K., Kanoh, H. and Kaneko, K., 2007. Adsorption of polyaromatic hydrocarbons on single wall carbon nanotubes of different functionalities and diameters. *Journal of colloid and interface science*, 314(1), pp.18-24.
- Gramatica, P., 2007. Principles of QSAR models validation: internal and external. *QSAR & combinatorial science*, 26(5), pp.694-701.
- Gurevitch, I. and Srebniak, S., 2008. Conformational behavior of polymers adsorbed on nanotubes. *The Journal of chemical physics*, 128(14), p.144901.
- Hahm, M.G., Hashim, D.P., Vajtai, R. and Ajayan, P.M., 2011. A review: controlled synthesis of vertically aligned carbon nanotubes. *Carbon letters*, 12(4), pp.185-193.
- Hansch, C., Leo, A., Hoekman, D.H., 1995. Exploring QSAR: structure activity relations in chemistry and biology. *American Chemical Society*.
- Hassanzadeh, Z., Kompany-Zareh, M., Ghavami, R., Gholami, S. and Malek-Khatabi, A., 2015. Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. *Journal of Molecular Structure*, 1098, pp.191-198.
- Hawkins, D.M., 2004. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1), pp.1-12.

- Helland, A., Wick, P., Koehler, A., Schmid, K. and Som, C., 2007. Reviewing the environmental and human health knowledge base of carbon nanotubes. *Environmental health perspectives*, 115(8), pp.1125-1131.
- Ibrahim, K.S., 2013. Carbon nanotubes-properties and applications: a review. *Carbon letters*, 14(3), pp.131-144.
- Kang, I., Heung, Y.Y., Kim, J.H., Lee, J.W., Gollapudi, R., Subramaniam, S., Narasimhadevara, S., Hurd, D., Kirikera, G.R., Shanov, V. and Schulz, M.J., 2006. Introduction to carbon nanotube and nanofiber smart materials. *Composites Part B: Engineering*, 37(6), pp.382-394.
- Karajanagi, S.S., Vertegel, A.A., Kane, R.S. and Dordick, J.S., 2004. Structure and function of enzymes adsorbed onto single-walled carbon nanotubes. *Langmuir*, 20(26), pp.11594-11599.
- Khani, H. and Moradi, O., 2013. Influence of surface oxidation on the morphological and crystallographic structure of multi-walled carbon nanotubes via different oxidants. *Journal of Nanostructure in Chemistry*, 3(1), p.73.
- Katritzky, A.R., Fara, D.C., Petrukhin, R.O., Tatham, D.B., Maran, U., Lomaka, A. and Karelson, M., 2002. The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. *Current Topics in Medicinal Chemistry*, 2(12), pp.1333-1356
- Kier, L.B. and Hall, L.H., 2002. The meaning of molecular connectivity: A bimolecular accessibility model. *Croatica chemica acta*, 75(2), pp.371-382.
- Lata, S., 2018. Concentration dependent adsorption of aromatic organic compounds by SWCNTs: Quantum-mechanical descriptors for nano-toxicological studies of

- biomolecules and agrochemicals. *Journal of Molecular Graphics and Modelling*, 85, pp.232-241.
- Liew, K.M., Jianwei, Y. and Zhang, L.W., 2016. *Mechanical Behaviors of Carbon Nanotubes: Theoretical and Numerical Approaches*. William Andrew.
- Liu, G., Wang, J., Zhu, Y. and Zhang, X., 2004. Application of multiwalled carbon nanotubes as a solid-phase extraction sorbent for chlorobenzenes. *Analytical letters*, 37(14), pp.3085-3104.
- Liu, Y., Zhang, J., Chen, X., Zheng, J., Wang, G. and Liang, G., 2014. Insights into the adsorption of simple benzene derivatives on carbon nanotubes. *RSC Advances*, 4(101), pp.58036-58046.
- Long, R.Q. and Yang, R.T., 2001. Carbon nanotubes as superior sorbent for dioxin removal. *Journal of the American Chemical Society*, 123(9), pp.2058-2059.
- Lowis, D.R., 1997. HQSAR: a new, highly predictive QSAR technique. *Tripos Technical Notes*, 1(5), p.17.
- Kasuya, A., Saito, Y., Sasaki, Y., Fukushima, M., Maedaa, T., Horie, C. and Nishina, Y., 1996. Size dependent characteristics of single wall carbon nanotubes. *Materials Science and Engineering: A*, 217, pp.46-47
- Kaushik, B.K. and Majumder, M.K., 2015. Carbon nanotube: Properties and applications. In *Carbon Nanotube Based VLSI Interconnects* (pp. 17-37). Springer, New Delhi.
- Kennard, R.W. and Stone, L.A., 1969. Computer aided design of experiments. *Technometrics*, 11(1), pp.137-148.
- Krishnan, A., Dujardin, E., Ebbesen, T.W., Yianilos, P.N. and Treacy, M.M.J., 1998. Young's modulus of single-walled nanotubes. *Physical review B*, 58(20), p.14013.

- Kumar, S., Bhanjana, G., Dilbaghi, N. and Umar, A., 2014. Multi walled carbon nanotubes as sorbent for removal of crystal violet. *Journal of nanoscience and nanotechnology*, 14(9), pp. 7054-7059.
- Iijima, S., 1991. Helical microtubules of graphitic carbon. *Nature*, 354(6348), p.56.
- Latkar, M. and Chakrabarti, T., 1994. Performance of up flow anaerobic sludge blanket reactor carrying out biological hydrolysis of urea. *Water environment research*, 66(1), pp.12-15.
- Livingstone, D.J., 2000. The characterization of chemical structures using molecular properties. A survey. *Journal of chemical information and computer sciences*, 40(2), pp.195-209.
- Lu, Y., Song, S., Wang, R., Liu, Z., Meng, J., Sweetman, A.J., Jenkins, A., Ferrier, R.C., Li, H., Luo, W. and Wang, T., 2015. Impacts of soil and water pollution on food safety and health risks in China. *Environment international*, 77, pp.5-15.
- Michael, I., Rizzo, L., McArdell, C.S., Manaia, C.M., Merlin, C., Schwartz, T., Dagot, C. and Fatta-Kassinos, D., 2013. Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: a review. *Water research*, 47(3), pp.957-995.
- Mitra, I., Saha, A. and Roy, K., 2009. Quantitative structure–activity relationship modeling of antioxidant activities of hydroxyl benzalacetones using quantum chemical, physicochemical and spatial descriptors. *Chemical biology & drug design*, 73(5), pp.526-536.
- Moriguchi, I., Hirono, S., Nakagome, I. and Hirano, H., 1994. Comparison of reliability of log P values for drugs calculated by several methods. *Chemical and pharmaceutical bulletin*, 42(4), pp.976-978.

- Mosayebidorcheh, S. and Hatami, M., 2017. Heat transfer analysis in carbon nanotube-water between rotating disks under thermal radiation conditions. *Journal of Molecular Liquids*, 240, pp.258-267.
- Ojha, P.K., Mitra, I., Das, R.N. and Roy, K., 2011. Further exploring rm2 metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems*, 107(1), pp.194-205.
- Ojha, P.K. and Roy, K., 2018. Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. *Food and Chemical Toxicology*, 112, pp.551-562.
- Ong, Y.T., Ahmad, A.L., Zein, S.H.S. and Tan, S.H., 2010. A review on carbon nanotubes in an environmental protection and green engineering perspective. *Brazilian Journal of Chemical Engineering*, 27(2), pp.227-242.
- Pan, B. and Xing, B., 2008. Adsorption mechanisms of organic chemicals on carbon nanotubes. *Environmental science & technology*, 42(24), pp.9005-9013.
- Pan, B., Lin, D., Mashayekhi, H. and Xing, B., 2008. Adsorption and hysteresis of bisphenol A and 17 α -ethinyl estradiol on carbon nanomaterials. *Environmental science & technology*, 42(15), pp.5480-5485.
- Pimentel, D., 1995. Amounts of pesticides reaching target pests: environmental impacts and ethics. *Journal of Agricultural and environmental Ethics*, 8(1), pp.17-29.
- Pratim Roy, P., Paul, S., Mitra, I. and Roy, K., 2009. On two novel parameters for validation of predictive QSAR models. *Molecules*, 14(5), pp.1660-1701.
- Rahimi-Nasrabadi, M., Akhoondi, R., Pourmortazavi, S.M. and Ahmadi, F., 2015. Predicting adsorption of aromatic compounds by carbon nanotubes based on quantitative structure property relationship principles. *Journal of Molecular Structure*, 1099, pp.510-515.

- Randall, J.M., Hautala, E. and Waiss Jr, A.C., 1974, May. Removal and recycling of heavy metal ions from mining and industrial waste streams with agricultural by-products. In *Proceedings of the Fourth Mineral Waste Utilization Symposium. Chicago, IL* (pp. 329-334).
- Randic, M., 1997. On characterization of chemical structure. *Journal of chemical information and computer sciences*, 37(4), pp.672-687.
- Rand-Weaver, M., Margiotta-Casaluci, L., Patel, A., Panter, G.H., Owen, S.F. and Sumpter, J.P., 2013. The read-across hypothesis and environmental risk assessment of pharmaceuticals. *Environmental science & technology*, 47(20), pp.11384-11395.
- Richard, C., Balavoine, F., Schultz, P., Ebbesen, T.W. and Mioskowski, C., 2003. Supramolecular self-assembly of lipid derivatives on carbon nanotubes. *Science*, 300(5620), pp.775-778.
- Roy, K. ed., 2015. *Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment*. IGI Global.
- Roy, K., Ambure, P., Kar, S. and Ojha, P.K., 2018. Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models?. *Journal of Chemometrics*, 32(4), p.e2992.
- Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S. and Das, R.N., 2013. Some case studies on application of “rm2” metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *Journal of computational chemistry*, 34(12), pp.1071-1082

- Roy, K., Das, R.N., Ambure, P. and Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 152, pp.18-33.
- Roy, K. and Kabir, H., 2012. QSPR with extended topochemical atom (ETA) indices, 3: modeling of critical micelle concentration of cationic surfactants. *Chemical engineering science*, 81, pp.169-178.
- Roy, K., Kar, S. and Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 145, pp.22-29.
- Roy, K., Kar, S. and Das, R.N., 2015. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press.
- Roy, K. and Mitra, I., 2012. On the use of the metric rm^2 as an effective tool for validation of QSAR models in computational drug design and predictive toxicology. *Mini reviews in medicinal chemistry*, 12(6), pp.491-504.
- Roy, K. and Paul, S., 2009. Exploring 2D and 3D QSARs of 2, 4-diphenyl-1, 3-oxazolines for ovicidal activity against *Tetranychus urticae*. *QSAR & Combinatorial Science*, 28(4), pp.406-425.
- Roy, K. and Popelier, P.L., 2009. Predictive QSPR modeling of the acidic dissociation constant (pKa) of phenols in different solvents. *Journal of Physical Organic Chemistry*, 22(3), pp.186-196.
- Roy, K. and Roy, P.P., 2009. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS,

- GFA, G/PLS and ANN techniques. *European journal of medicinal chemistry*, 44(7), pp.2913-2922.
- Saito, R., Fujita, M., Dresselhaus, G. and Dresselhaus, U.M., 1992. Electronic structure of chiral graphene tubules. *Applied physics letters*, 60(18), pp.2204-2206.
- Salahinejad, M. and Zolfonoun, E., 2018. An exploratory study using QICAR models for prediction of adsorption capacity of multi-walled carbon nanotubes for heavy metal ions. *SAR and QSAR in Environmental Research*, 29(12), pp.997-1009.
- Schuuuermann, G., Ebert, R.U., Chen, J., Wang, B. and Kuuhne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient Test set activity mean vs training set activity mean. *Journal of Chemical Information and Modeling*, 48(11), pp.2140-2145.
- Sinnott, S.B., Shenderova, O.A., White, C.T. and Brenner, D.W., 1998. Mechanical properties of nanotubule fibers and composites determined from theoretical calculations and simulations. *Carbon*, 36(1-2), pp.1-9.
- Snarey, M., Terrett, N.K., Willett, P. and Wilton, D.J., 1997. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling*, 15(6), pp.372-385.
- Snyder, S.A., Westerhoff, P., Yoon, Y. and Sedlak, D.L., 2003. Pharmaceuticals, personal care products, and endocrine disruptors in water: implications for the water industry. *Environmental engineering science*, 20(5), pp.449-469.
- Su, F. and Lu, C., 2007. Adsorption kinetics, thermodynamics and desorption of natural dissolved organic matter by multiwalled carbon nanotubes. *Journal of Environmental Science Health, Part A*, 42(11), pp.1543-1552. and

- Sun, K., Zhang, Z., Gao, B., Wang, Z., Xu, D., Jin, J. and Liu, X., 2012. Adsorption of diuron, fluridone and norflurazon on single-walled and multi-walled carbon nanotubes. *Science of the Total Environment*, 439, pp.1-7.
- Tariq, M.I., Afzal, S., Hussain, I. and Sultana, N., 2007. Pesticides exposure in Pakistan: a review. *Environment international*, 33(8), pp.1107-1122.
- Todeschini, R. and Consonni, V., 2008. *Handbook of molecular descriptors* (Vol. 11). John Wiley & Sons.
- Todeschini, R. and Consonni, V., 2009. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons.
- Tong W, Hong H, Xie Q, Shi L, Fang H, Perkins R, *CurrComput Aided Drug Des* 2005,1: 195.
- Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. *Molecular informatics*, 29(6-7), pp.476-488.
- Van de Waterbeemd, H., Carter, R.E., Grassy, G., Kubinyi, H., Martin, Y.C., Tute, M.S. and Willett, P., 1997. Glossary of terms used in computational drug design (IUPAC Recommendations 1997). *Pure and applied chemistry*, 69(5), pp.1137-1152.
- Wang, F., Yao, J., Sun, K. and Xing, B., 2010a. Adsorption of dialkyl phthalate esters on carbon nanotubes. *Environmental science & technology*, 44(18), pp.6985-6991.
- Wang, L., Zhu, D., Duan, L. and Chen, W., 2010b. Adsorption of single-ringed N-and S-heterocyclic aromatics on carbon nanotubes. *Carbon*, 48(13), pp.3906-3915.
- Wang, Q., Apul, O.G., Xuan, P., Luo, F. and Karanfil, T., 2013. Development of a 3D QSPR model for adsorption of aromatic compounds by carbon nanotubes: comparison of multiple linear regression, artificial neural network and support vector machine. *RSC Advances*, 3(46), pp.23924-23934.

- Wang, S.G., Liu, X.W., Gong, W.X., Nie, W., Gao, B.Y. and Yue, Q.Y., 2007. Adsorption of fulvic acids from aqueous solutions by carbon nanotubes. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, 82(8), pp.698-704.
- Wang, Ya, Jingwen Chen, Weihao Tang, Deming Xia, Yuzhen Liang, and Xuehua Li. "Modeling adsorption of organic pollutants onto single-walled carbon nanotubes with theoretical molecular descriptors using MLR and SVM algorithms." *Chemosphere* 214 (2019): 79-84.
- Wang, Y., Yan, F., Jia, Q. and Wang, Q., 2017. Assessment for multi-endpoint values of carbon nanotubes: Quantitative nanostructure-property relationship modeling with norm indexes. *Journal of Molecular Liquids*, 248, pp.399-405.
- Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), pp.37-52.
- Wold, S., Eriksson, L. and Clementi, S., 1995. Statistical validation of QSAR results. *Chemometric methods in molecular design*, pp.309-338.
- Wold, S., Sjostrom, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), pp.109-130.
- Yu, M.F., Files, B.S., Arepalli, S. and Ruoff, R.S., 2000. Tensile loading of ropes of single wall carbon nanotubes and their mechanical properties. *Physical review letters*, 84(24), p.5552.
- Yu, C., Shi, L., Yao, Z., Li, D. and Majumdar, A., 2005. Thermal conductance and thermopower of an individual single-wall carbon nanotube. *Nano letters*, 5(9), pp.1842-1846.

- Zhao, J., Buldum, A., Han, J. and Lu, J.P., 2002. Gas molecule adsorption in carbon nanotubes and nanotube bundles. *Nanotechnology*, 13(2), p.195.
- Zhao, Q., Yang, K., Li, W. and Xing, B., 2014. Concentration-dependent polyparameter linear free energy relationships to predict organic compound sorption on carbon nanotubes. *Scientific reports*, 4, p.3888.
- Zhou, G., Duan, W. and Gu, B., 2001. First-principles study on morphology and mechanical properties of single-walled carbon nanotube. *Chemical Physics Letters*, 333(5), pp.344-349.

Chapter 7



Appendix

PAPER



Cite this: *Environ. Sci.: Nano*, 2019, 6, 224

Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs)†

Joyita Roy,‡ Sulekha Ghosh,‡ Probir Kumar Ojha and Kunal Roy *

Nanotechnology has introduced a new generation of adsorbents like carbon nanotubes (CNTs), which have drawn a widespread attention due to their outstanding ability for the removal of various inorganic and organic pollutants. The goal of this study was to develop regression-based quantitative structure–property relationship (QSPR) models for organic pollutants and organic solvents using only easily computable 2D descriptors to explore the key structural features essential for adsorption to multi-walled CNTs and improve the dispersibility index of single-walled CNTs. The statistical results of the developed models showed good quality and predictivity based on both internal and external validation metrics (dataset 1: R^2 range of 0.893–0.920, $Q_{(LOO)}^2$ range of 0.863–0.895, Q_{F1}^2 range of 0.887–0.919; dataset 2: R^2 range of 0.793–0.845, $Q_{(LOO)}^2$ range of 0.743–0.798, Q_{F1}^2 range of 0.783–0.890; dataset 3: $R^2 = 0.830$, $Q_{(LOO)}^2 = 0.775$, $Q_{F1}^2 = 0.945$). We have also tried to explore whether the quality of the predictions of test set compounds can be enhanced through an “intelligent” selection of multiple models using the “Intelligent consensus predictor” tool. The consensus results suggested that the consensus predictivity of the test set compounds gave better results than those from the individual MLR models based on different criteria (dataset 1: $Q_{F1}^2 = 0.935$, $Q_{F2}^2 = 0.935$, $MAE_{(95\%)} = \text{good}$; dataset 2: $Q_{F1}^2 = 0.887$, $Q_{F2}^2 = 0.879$, $MAE_{(95\%)} = \text{good}$). The contributed descriptors obtained from different models suggested that the organic pollutants may adsorb to the CNTs through hydrogen bonding interactions, π – π interactions, hydrophobic interactions and electrostatic interaction. Based on the observations obtained from the developed models, we have inferred that the adsorption of the organic pollutants onto the CNTs can be enhanced by the following factors: a higher number of aromatic rings, high unsaturation or electron richness of molecules, the presence of polar groups substituted in the aromatic ring, the presence of oxygen and nitrogen atoms, the size of the molecules, and the hydrophobic surface of the molecules. On the other hand, the presence of C–O groups, aliphatic primary alcohols and the presence of chlorine atoms may retard the adsorption of organic pollutants. The results also suggest that the organic solvents bearing the >N- fragment, a higher degree of branching (compactness), polar solvents with low donor number and lower ionization potential may be better solvents for enhancing the dispersibility of single-walled CNTs.

Received 22nd September 2018,
Accepted 16th November 2018

DOI: 10.1039/c8en01059e

rsc.li/es-nano

Environmental significance

Nanotechnology has introduced a new generation of adsorbents such as carbon nanotubes (CNTs), which have drawn widespread attention due to their outstanding ability for the removal of various inorganic and organic pollutants. The goal of this study was to develop quantitative structure–property relationship (QSPR) models to explore the key structural features of organic pollutants, which are essential for adsorption to multi-walled CNTs. We have also developed models to investigate the characteristics that can improve the dispersibility of single-walled CNTs. This information may be helpful in the process of removal of the harmful and toxic contaminants/disposal of the by-products from various industries by increasing the adsorption of pollutants and the dispersibility of CNTs, thus making a pollution-free environment.

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.

E-mail: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in;

Fax: +91 33 2837 1078; Tel: +91 98315 94140

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8en01059e

‡ These authors contributed equally.

1. Introduction

A noticeable amount of organic pollutants is released into the environment *via* various routes like the burning of fossil fuels, wastes from incineration, exhausts from automobiles, agricultural processes and industrial sectors. The disposal of

the by-products from the various industries is a challenging job for environmentalists and industries. The major problem with pollutants is their effective and safe disposal without further affecting the environment adversely. The organic pollutants (phenols, cresols, alkyl benzene sulfonates, nitro chlorobenzene, chlorinated paraffins, butadiene, synthetic dyes, insecticides, fungicides and pesticides, *etc.*) accumulate in the food chain and persist in nature and cause a significant threat to the environment.^{1–4} The United States Environmental Protection Agency (EPA) has set maximum contamination levels (MCLs) and maximum contamination level goals (MCLG) for each pollutant, with no ill health effects. Sometimes the MCL level goes beyond the MCLG level because of the problem in determining small quantities of contaminants and due to lack of availability of treatment technologies and analytical methods.^{5–14} Thus, for the protection of the environment, the use of new and advanced materials is important. In recent years, greater focus has been placed on nanostructures as adsorbents and catalysts for removing the harmful and toxic contaminants from the environment.^{15–17} Among the various nanomaterial adsorbents, carbon nanotubes (CNTs) have been thoroughly investigated because they have a large surface area to volume ratio, inertness towards chemicals, light mass density, porous structure, great physical and chemical properties, small diameter, extraordinary optical and electrical properties, high tensile strength and efficient affinity towards pollutants. The possibility of surface modification with different functional groups makes CNTs good adsorbents^{18–20} and enhances their reactivity and dispersibility for environmental protection applications.

SWCNTs have some unique mechanical, electrical and thermal properties but possess poor solubility as well as poor dispersibility in aqueous and other common organic solvents.²¹ They possess high polarizability along with van der Waals interactions and hydrophobic surface, so they are able to form aggregates with each other and with other biological and chemical systems to produce mixtures of aggregates, specifically in water.^{22,23} This bundling or entangling feature of SWCNTs causes difficulties in the dispersion of CNTs in various solvents or matrices.^{24–26} This also prevents the exploration of the chemistry of CNTs at a molecular level and hinders their applications²⁷ as well as limits the availability of adsorption sites for the adsorption of pollutants on the CNT surface.²⁸ The morphology variation of CNTs may also result in a difference in their aggregation tendencies, which may additionally impact their adsorption capability. The major interactions are van der Waals interactions, π - π stacking, and hydrophobic interactions for dispersibility, as suggested by many researchers.²⁹

Hyung *et al.*³⁰ reported that organic contaminants can interact with carbon nanotubes in aquatic systems and increase their stability and transport and thus, the mobility of the adsorbed organic matters on CNTs can be enhanced. The popularity of CNTs has increased since Long and Yang first reported that they can efficiently remove dioxins as compared

to activated carbon.³¹ The sorption studies performed on CNTs for metal ions³² and organic contaminants, such as butane,³³ trihalomethanes,³⁴ dioxin,³¹ xylenes,³⁵ chlorophenols,³⁶ 1,2-dichlorobenzene,³⁷ resorcinol³⁸ and polycyclic aromatic hydrocarbons (PAHs),^{15,39} suggest that CNTs can remove both organic and inorganic pollutants from water and gases.

Although a large number of pollutants are reported in the literature, adsorption data is available for only around 70 000 pollutants.⁴⁰ The determination of experimental data for a large number of pollutants is time-consuming as well as laborious and costly. The surface properties of CNTs can be modified by treating them with some active chemicals so that the CNTs do not aggregate or form bundles and hence, the dispersion of CNTs can be enhanced. QSPR modeling of organic pollutants/solvents using adsorption properties/dispersibility index by CNTs can, therefore, be of great importance for researchers and practitioners. The quantitative structure–property relationship (QSPR) approach is easier than the thermodynamic model since the input parameters of QSPR can be more easily obtained as compared to the thermodynamic models.⁴¹ QSPR not only reduces the experimental work but also predicts the features based on the chemical structures. Thus, the rationalization ideas obtained from such models provide the researchers with a conceptual framework upon which a firm discussion can be based. Recently, a great deal of work has been done with QSPR and linear surface energy relationship (LSER) modeling to develop predictive models for CNTs, including the adsorption of organic chemicals (OCs) by CNTs,^{41–47} dispersibility of CNTs in organic solvents^{48–51} and other properties similar to CNTs. In the past, some work has been done by researchers, for example, linear LSER models were developed by Xia *et al.*⁴³ using the biological surface index (BSAI) for the prediction and characterization of the intermolecular adsorption of OCs by CNTs. Apul *et al.*⁴⁵ reported a 3D-QSPR modeling applying the same data sets for the adsorption of aromatic compounds by CNTs and compared it with MLR, ANN and SVM methods. Another QSPR model was reported by Yilmaz *et al.*⁴⁸ using additive descriptors and quantum-chemical descriptors for the determination of the dispersibility of CNTs in different organic solvents.

The objective of the present study has been to develop statistically significant QSPR models of organic pollutants with multiple-endpoints using only easily computable 2D descriptors to explore the key structural features that are essential for adsorption to MWCNTs. We have also developed a QSPR model for organic solvents to investigate the characteristics of molecules that can improve the dispersibility of SWCNTs and may overcome the drawbacks of SWCNTs. A variable selection strategy was also employed prior to the development of final models to reduce noise in the input. We have also tried to explore whether the quality of predictions of test set compounds can be enhanced through the “intelligent” selection of multiple MLR models using an “Intelligent consensus predictor” tool.

2. Methods and materials

2.1. Dataset

We have developed QSPR models separately, using three different data sets for diverse organic contaminants with multiple-endpoints of carbon nanotubes reported in the literature.^{41,44,52} The first dataset involves the defined adsorption affinity properties (k_{∞}) of 59 organic contaminants by multi-walled carbon nanotubes (MWCNTs). The second dataset involves the adsorption affinity of 69 organic contaminants related to the specific surface area (k_{SA}) of multi-walled carbon nanotubes (MWCNTs), and the third data set involves 29 organic solvents with defined dispersibility index values (C_{max}) for single-walled carbon nanotubes (SWCNTs). We have not excluded any compound of individual data sets in our modeling analysis. All the endpoint values were taken in the logarithmic scale for the modeling purposes. The first two data sets mainly involve adsorption data for synthetic organic compounds like pyrene, naphthalene, phenol, benzene, aniline, benzoate, chloroanisole, alcohol, acetophenone, isophoron, phenanthrene dicamba, atrazine, carbamazepine, pyrimidinone, acetamide, piperidine, propionitrile, acrylic acid, thiodiethanol, ethanolamine, cyclopentanone, acetone and ethylene glycol derivatives, while the third data set is related to different types of solvents. The dispersibility of single-walled carbon nanotubes (SWCNTs) was measured in different solvent ranges. Here, C_{max} (mg mL^{-1}) represents the maximum dispersibility of single-walled carbon nanotubes, K_{∞} and K_{SA} are both adsorption coefficients that can be obtained from isotherm data. K_{∞} is the ratio of q_e and C_e (solid and liquid phase equilibrium concentrations, respectively, at infinite dilution conditions with an average of 0.2% aqueous solubility). K_{SA} is the normalized value of K_{∞} and the specific surface area of multi-walled carbon nanotubes (MWCNTs). The data sets are given in Tables S1, S2 and S3 in the ESI† section.

2.2. Descriptor calculation

“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 experiments”. All the dataset compounds were drawn using the Marvin Sketch software.⁵³ The descriptors were calculated using two software tools, namely, Dragon software version 6,⁵⁴ and PaDEL-descriptor⁵⁵ software. In this work, we have calculated only 2D descriptors covering constitutional, ring descriptors, connectivity index, functional group counts, atom centered fragments, atom type E-states, 2D atom pairs, molecular properties (using Dragon software version 6) and ETA indices (using PaDEL-Descriptor software).

2.3. Data set division

Division of the dataset is a very important step for QSPR. The present work deals with three datasets containing diverse organic pollutants or solvents. In each case, all the dataset

compounds were divided into a training set and a test set using the “Modified k-medoid” clustering technique. The clustering technique categorizes a set of compounds into clusters so that the compounds present in the same cluster are similar to each other. On the other hand, when two compounds belong to two different clusters, they are said to be dissimilar in nature. The indicative compounds within a cluster are called medoids. This technique tends to select k from most middle objects or compounds as the initial medoid. Three clusters were generated for the dataset containing 59 and 29 compounds, while six clusters were generated for the dataset containing 69 compounds. We have selected approximately 25% of compounds from each data set for the test set and the remaining 75% of compounds were selected for the training set. The purpose of the training set was to develop the model and the test set was used to validate the model for prediction purposes. The same strategy was applied in the case of all three datasets for training and test set division.

2.4. Variable selection and model development

After the dataset division step, we performed data pretreatment to remove intercorrelated descriptors from all three sets of datasets. Prior to the development of final models, we tried to extract the important descriptors from the large pool of initial descriptors using various variable selection strategies.^{56,57} In case of the dataset containing 59 and 69 organic pollutants, we separately ran a stepwise regression and selected some descriptors in each case. After removing the selected descriptors obtained from the first stepwise regression run, we ran the stepwise regression again using the remaining pool of descriptors, and we repeated the same procedure. In this way, we selected some manageable numbers of descriptors and made a reduced pool of descriptors. In the case of the dataset containing 29 compounds, we developed GA equations and made a descriptor pool using the descriptors obtained from the GA (genetic algorithm) equations. After that, we ran the best subset selection for all three datasets using the reduced pools of descriptors. For this, we used a tool developed in our laboratory.⁵⁸ Five (three models were selected) and four (two models were selected) descriptor models were generated in the case of the dataset containing 59 organic pollutants, whereas six (three models were selected) and five (two models were selected) descriptor models were generated for the dataset containing 69 organic pollutants. Among the equations generated from the best subset selection, we selected five models, five models and four models for 59, 69 and 29 compounds, respectively, based on MAE criteria.⁵⁹ Descriptors were selected from the GA and stepwise regression models and a descriptor pool was generated. Finally, the selected models were run using the intelligent consensus predictor (ICP) tool developed in our laboratory⁶⁰ to explore whether the quality of predictions of external compounds could be enhanced through an “intelligent” selection of multiple models (in this report, five models were selected).

The multilayered strategies like data pretreatment,⁵⁸ stepwise regression,⁶¹ genetic method⁶² and best subset

selection⁵⁸ were involved for the selection of variables prior to the development of the final models and different steps are discussed separately in the ESI† section.

2.4.1. Intelligent consensus predictor (ICP).⁶⁰ This software was used to judge the performance of consensus predictions in comparison to their quality obtained from the individual (MLR) models based on the MAE based criteria (95%). It is obvious that a single model might not be equally useful for prediction for the whole test set compounds, which means that one QSPR model may be the best model for prediction of a test compound while the other model may be the best predictor for another test compounds. For this reason, we have selected five models in the case of a dataset containing 59 (M1–M5) and 69 (N1–N5) organic contaminants, and performed consensus prediction using the “Intelligent consensus predictor” tool to explore whether the quality of the predictions of the test set compounds could be enhanced through an “intelligent” selection of multiple models. The steps involved in the development of both MLR and PLS models are represented schematically in Fig. 1.

2.5. Statistical validation metrics

In order to judge the predictivity and reliability of the developed QSPR models, we have examined the statistical quality, applying both internal and external validation metrics. In this work, we have used various statistical parameters like determination coefficient R^2 , explained variance R^2_{adj} , variance ratio (F), and standard error of estimate (s). These parameters are

not sufficient to evaluate the predictive potential of the model, so we have used some other classical parameters for validation of the models. The internal predictivity parameters like the leave-one-out cross-validated correlation coefficient (Q^2_{LOO}), and external predictivity parameters like R^2_{pred} or Q^2_{F1} , Q^2_{F2} and concordance correlation coefficient (CCC), were also calculated. We also calculated some r^2_{m} parameters like $r^2_{\text{m(LOO)}}$ and $\Delta r^2_{\text{m(LOO)}}$ for internal validation and $r^2_{\text{m(test)}}$ and $\Delta r^2_{\text{m(test)}}$ for external validation.⁶³ The basic objective of the predictive performance of QSPR models is to investigate the prediction errors of an external set, which should be within the chemical and response-based domain of the internal set (*i.e.*, training set). The Q^2_{ext} -based metrics (*i.e.*, R^2_{pred} and Q^2_{F2}) are not always able to provide the correct indication of the prediction quality because of the influence of the response range as well as the distribution of the values of response in both the training and test set compounds.⁵⁹ Thus, we have also validated the models using the mean absolute error (MAE) criteria for both external and internal validation.⁵⁹ The error based metrics were used to determine the true indication of the prediction quality in terms of prediction error since they do not evaluate the performance of the model in comparison with the mean response (Roy *et al.*, 2016 (ref. 59)). The threshold values of Q^2 , Q^2_{F2} , R^2_{pred} , $r^2_{\text{m(test)}}$, $r^2_{\text{m(LOO)}}$ are 0.5 and for CCC, it is 0.750.^{64,65} The limit for $\Delta r^2_{\text{m(test)}}$ and $\Delta r^2_{\text{m(LOO)}}$ is 0.2. Recently, Roy *et al.* reported that a single model might not be equally useful in the prediction for the whole test set compounds, *i.e.*, one QSPR model may be the best model for prediction of a test compound while the other

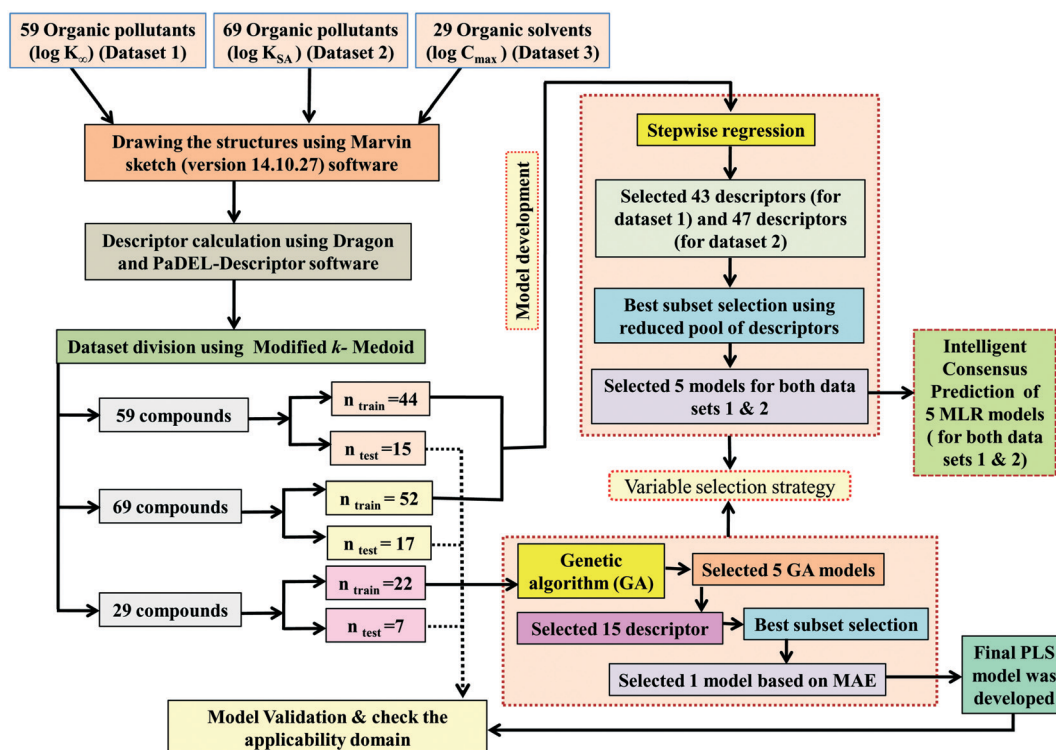


Fig. 1 Schematic representation of the steps involved in the development of QSPR models.

model may be the best predictor for another test compound. For this reason, we have also performed Intelligent consensus prediction (ICP) using multiple QSPR models to determine whether the quality of the predictions of test set compounds can be enhanced through an “intelligent” selection. Here, a simple average of predictions from all the models is not considered; only ‘qualified models’ are taken into account.

2.6. Applicability domain

“The applicability domain of a (Q)SAR is the physicochemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The applicability domain of a (Q)SAR should be described in terms of the most relevant parameters, i.e., usually those that are descriptors of the model. Ideally, the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation”. The AD of the QSAR model is characterized by the molecular properties of the training set compounds. The AD criteria help to check whether the test/query compound under consideration is inside the AD or not. Here, we have checked the applicability domain of test set compounds of the developed models, employing the standardization approach (for first two data sets) using the software developed in our laboratory⁶⁶ and a DModX (distance to model X) approach⁶⁷ at 99% confidence level using SIMCA-P software⁶⁸ (for the third data set). The predictability of a QSPR model is good if the molecules are present within the domain of the chemical space of the training set molecules.

2.7. Software used

Marvin Sketch version 5.5.0.1 (ref. 53) was used to draw chemical structures. Descriptors were calculated by the PADEL-Descriptor software⁵⁵ and Dragon software version 6.⁵⁴ Clustering of each data set was done by the “Modified K-Medoid” tool version 1.3 (ref. 58) for its splitting into a training set and a test set. Data Pretreatment version 1.2 was used to remove intercorrelated descriptors. Stepwise regression analysis was done by the MINITAB software version 13.14.⁶⁹ Genetic Algorithm was done by using the Genetic Algorithm tool version 4.1.⁵⁸ Best subset selection⁵⁸ and intelligent consensus predictor tool⁶⁰ were used to generate the QSPR models.

3. Results and discussion

We have developed QSPR models (five MLR models for each of the datasets containing 59 and 69 organic contaminants, and one PLS model for the dataset containing 29 organic contaminants) for three datasets containing diverse organic pollutants with defined adsorption affinities for MWCNTs (for datasets 1 and 2), and the dispersibility index of SWCNTs (for dataset 3), using reduced descriptors pools obtained by different strategies as discussed in the Materials and methods section. We checked the statistical quality of all the

individual models using both internal and external validation parameters, which showed that the models are statistically significant (Table 1). We also checked the MAE-based criteria for all the models.⁵⁹ All the models passed the MAE-based criteria.⁵⁹ Besides the routinely used validation parameters, we also checked the consensus predictions (for datasets 1 and 2 only) using the developed MLR models employing a newly developed “Intelligent consensus predictor” tool⁶⁰ to check whether the quality of the predictions of the test set compounds can be enhanced through an “intelligent” selection of multiple MLR models. We found that the consensus predictions of multiple MLR models are better (based on MAE based criteria) than the results obtained from the individual models as shown in Table 1 (here, in both cases, the winner model is CM3). It was also found that the consensus predictions of the test set compounds are better as compared to the individual MLR models based on not only the MAE-based criteria but also the other external validation metrics used in this work as shown in Table 1. All the individual models are mentioned below and the descriptors are discussed elaborately. In the equation, n_{training} is the number of compounds used to develop the models and n_{test} is the number of compounds used for the external prediction of the developed models. The values of leave-one-out (LOO) cross-validated correlation coefficient (Q^2) (Q^2 in the range of 0.863–0.895 for dataset 1; 0.743–0.798 for data set 2 and 0.775 for dataset 3) above the critical value of 0.5 signify the statistical reliability of the models. The predictability of the models was judged by means of predictive R^2 (R^2_{pred}) or Q^2_{F1} (Q^2_{F1} range of 0.887–0.919 for dataset 1; 0.783–0.890 for data set 2 and 0.945 for dataset 3) and Q^2_{F2} (Q^2_{F2} range of 0.886–0.919 for dataset 1; 0.768–0.882 for data set 2 and 0.938 for dataset 3), which show the good predictive ability of the models. The statistical results of all the models are summarized in Table 1. The PLS model developed from dataset 3 was also validated using a randomization test through randomly reordering (100 permutations) the dependent variable ($\log C_{\text{max}}$) using the SIMCA-P software.⁶⁸ Here, the intercept values for both R^2 and Q^2 are below the stipulated values ($R^2_{\text{int}} < 0.4$ and $Q^2_{\text{int}} < 0.05$), which confirmed that the developed model was not obtained by chance (Fig. S1 in ESI†). We have also checked the intercorrelation among the modeled descriptors for MLR models based on the Pearson correlation coefficient using the SPSS software.⁷⁰ The results showed that there is no intercorrelation between the modeled descriptors.

From the observations obtained from the modeled descriptors, it has been found that the organic pollutants may interact with the MWCNTs through different mechanisms like hydrogen bonding interactions, hydrophobic interactions, π - π interactions and electrostatic interactions as discussed below.

3.1. Dataset 1 : 59 organic pollutants

The significant descriptors obtained from the five MLR models (see Models M1–M5) for the adsorption properties

Table 1 Statistical quality and validation parameters obtained from the developed MLR and PLS models

Dataset	Type of model	Training set statistics					Test set statistics								
		Model R^2	Model $Q^2_{(LOO)}$	MAE_train	$r^2_{m(LOO)}$	$\Delta r^2_{m(LOO)}$	R^2_{pred} or Q^2_{FI}	Q^2_F	CCC	$r^2_{m(est)}$	$\Delta r^2_{m(test)}$	MAE (100%)	MAE (95%)	MAE	
59 organic contaminants	Individual models (M1–M5)	IM1	0.920	0.895	Good	0.851	0.078	0.887	0.886	0.934	0.745	0.104	0.271	0.240	Good
		IM2	0.912	0.892	Good	0.848	0.079	0.916	0.915	0.952	0.817	0.072	0.221	0.197	Good
		IM3	0.905	0.880	Good	0.832	0.075	0.919	0.919	0.954	0.825	0.069	0.213	0.189	Good
		IM4	0.893	0.872	Good	0.821	0.092	0.918	0.917	0.953	0.806	0.074	0.213	0.187	Good
		IM5	0.893	0.863	Good	0.808	0.086	0.915	0.914	0.950	0.798	0.076	0.222	0.199	Good
Consensus models	CM0	—	—	—	—	—	0.917	0.916	0.952	0.800	0.074	0.227	0.203	Good	
	CM1	—	—	—	—	—	0.917	0.916	0.952	0.800	0.074	0.227	0.203	Good	
	CM2	—	—	—	—	—	0.919	0.919	0.953	0.803	0.073	0.221	0.196	Good	
69 organic contaminants	Individual models (N1–N5)	CM3	—	—	—	—	0.935	0.935	0.962	0.812	0.059	0.187	0.163	Good	
		IM1	0.845	0.798	Moderate	0.709	0.087	0.809	0.795	0.908	0.783	0.048	0.319	0.271	Moderate
		IM2	0.842	0.790	Moderate	0.723	0.114	0.830	0.818	0.918	0.805	0.050	0.359	0.323	Good
	IM3	0.842	0.788	Good	0.714	0.081	0.783	0.768	0.890	0.712	0.140	0.340	0.265	Good	
	IM4	0.829	0.785	Good	0.709	0.087	0.812	0.799	0.903	0.748	0.044	0.330	0.286	Moderate	
Consensus models	IM5	0.793	0.743	Good	0.709	0.087	0.890	0.882	0.940	0.836	0.090	0.273	0.247	Good	
	CM0	—	—	—	—	—	0.862	0.852	0.929	0.818	0.002	0.284	0.245	Good	
	CM1	—	—	—	—	—	0.862	0.852	0.929	0.818	0.002	0.284	0.245	Good	
29 organic contaminants	CM2	—	—	—	—	—	0.865	0.855	0.930	0.820	0.014	0.279	0.241	Good	
	CM3	—	—	—	—	—	0.887	0.879	0.941	0.851	0.040	0.263	0.235	Good	
	P1	0.830	0.775	Good	0.689	0.115	0.945	0.938	0.991	0.909	0.048	0.152	—	Good	

CM0 = Ordinary consensus predictions. CM1 = Average of predictions from individual models IM1 through IM5. CM2 = Weighted average predictions from individual models IM1 through IM5. CM3 = Best selection of predictions (compound-wise) from individual models IM1 through IM5. *Note that we have run the “Intelligent consensus predictor tool” using the options, AD: No; Dixon Q-test: No; Euclidean distance: No.

($\log K_{oc}$) of 59 organic chemicals on MWCNTs are X0v, nArOH, B01[C-O], B06[C-Cl], Ui, F03[O-O], F04[N-O], ETA_BetaP, minsCH₃, B03[O-O] and nHBint4, which regulate the adsorption properties of the organic pollutants. The contribution of the descriptors can be easily identified from the regression coefficient of the independent variables. In this case, all the descriptors contributed positively (positive regression coefficients), except the B01[C-O] descriptor (negative regression coefficient). The definition, contribution and frequency of the contributed descriptors are shown in Table S4 in the ESI.† We have checked the applicability domain of the developed MLR models using the standardization approach to confirm whether there is any compound present outside the applicability domain or not. It was found that one compound (compound number 41) for model M1 is situated outside the applicability domain, while compound number 56 is situated outside the domain of applicability in case of models M2, M3, M4 and M5; however, these compounds showed good predictivity based on the models. The scatter plot of the observed vs. predicted adsorption coefficient for all the MLR models are shown in Fig. 2.

$$\begin{aligned} \text{Model M1. } \log k_{oc} = & -4.62(\pm 0.337) + 0.834(\pm 0.155) \times \text{Ui} \\ & + 0.663(\pm 0.220) \times \text{B06[C-Cl]} \\ & + 0.641(\pm 0.057) \times \text{X0v} \\ & + 0.600(\pm 0.091) \times \text{nArOH} \\ & - 0.611(\pm 0.121) \times \text{B01[C-O]} \end{aligned}$$

$$n_{\text{training}} = 44, R^2 = 0.920, R_{\text{adj}}^2 = 0.908, S = 0.294, F = 85.93,$$

$$\text{PRESS} = 4.267, Q^2 = 0.895, \overline{r_{\text{m(L100)}}^2} = 0.851,$$

$$\Delta r_{\text{m(L100)}}^2 = 0.078, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 15, Q_{F1}^2 = 0.887, Q_{F2}^2 = 0.886, \overline{r_{\text{m(test)}}^2} = 0.745, \Delta r_{\text{m(test)}}^2 = 0.104,$$

$$\text{CCC} = 0.934, \text{MAE} = \text{Good}$$

$$\begin{aligned} \text{Model M2. } \log k_{oc} = & -8.51(\pm 0.722) + 0.803(\pm 0.048) \times \text{X0v} \\ & + 0.681(\pm 0.146) \times \text{F03[O-O]} \\ & + 0.415(\pm 0.144) \times \text{F04[N-O]} \\ & + 3.27(\pm 0.491) \times \text{ETA_BetaP} \\ & + 0.204(\pm 0.067) \times \text{minsCH}_3 \end{aligned}$$

Experimental $\log K_{oc}$ vs predicted $\log K_{oc}$ values of 59 organic pollutants

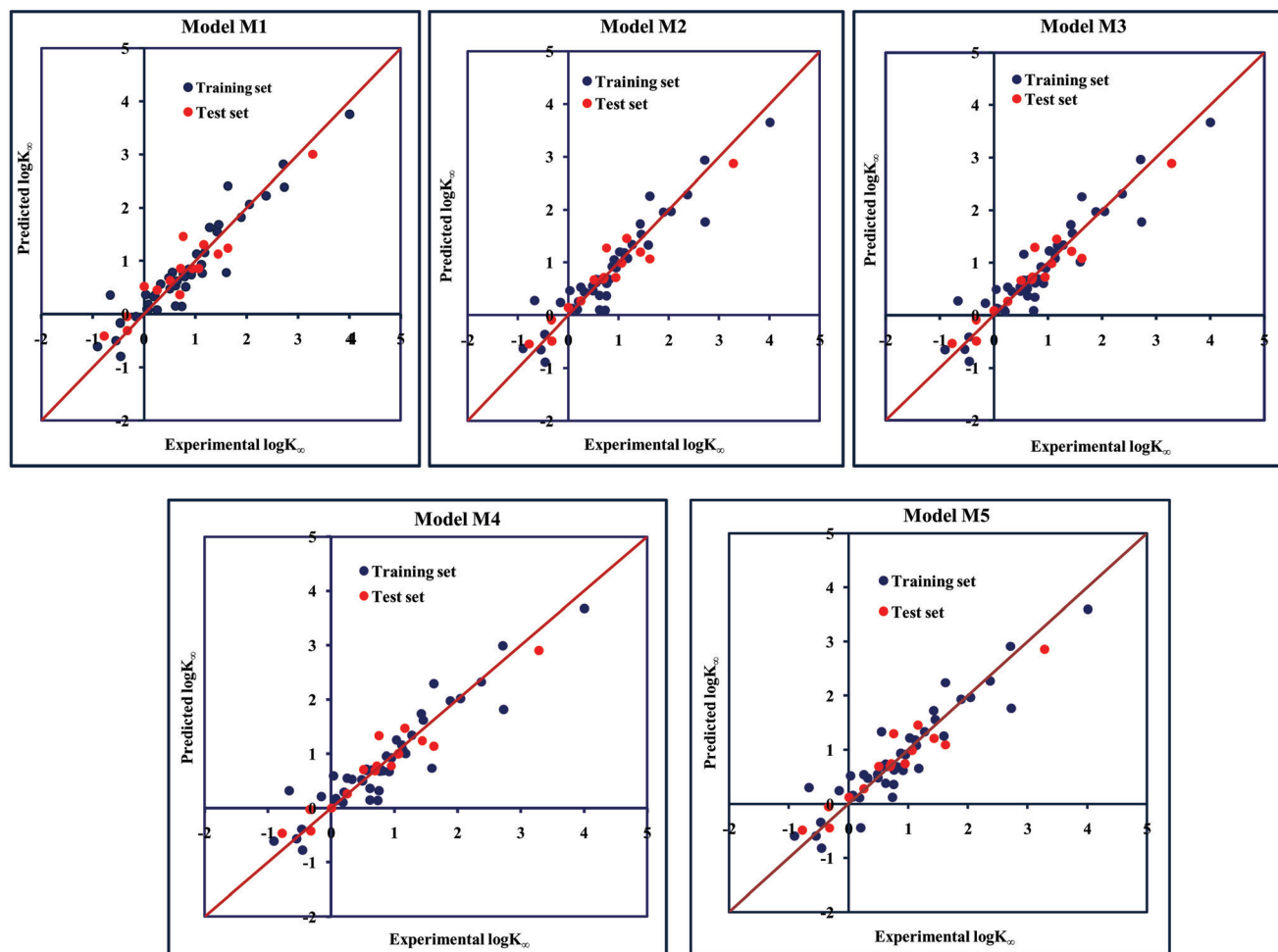


Fig. 2 The scatter plot of the observed and the predicted adsorption coefficient property ($\log K_{oc}$) of the developed MLR models (models M1–M5).

$$n_{\text{training}} = 44, R^2 = 0.912, R_{\text{adj}}^2 = 0.900, S = 0.306, F = 78.66,$$

$$\text{PRESS} = 4.356, Q^2 = 0.892, \overline{r_{\text{m(LOO)}}^2} = 0.848,$$

$$\Delta r_{\text{m(LOO)}}^2 = 0.079, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 15, Q_{\text{F1}}^2 = 0.916, Q_{\text{F2}}^2 = 0.915, \overline{r_{\text{m(test)}}^2} = 0.817, \Delta r_{\text{m(test)}}^2 = 0.072,$$

$$\text{CCC} = 0.952, \text{MAE} = \text{Good}$$

Model M3. $\log k_{\infty} = -8.68(\pm 0.746) + 0.802(\pm 0.050) \times \text{X0v}$
 $+ 0.603(\pm 0.272) \times \text{B03[O-O]}$
 $+ 3.39(\pm 0.503) \times \text{ETA_BetaP}$
 $+ 0.213(\pm 0.069) \times \text{minsCH}_3$
 $+ 0.412(\pm 0.148) \times \text{nHBint}_4$

$$n_{\text{training}} = 44, R^2 = 0.905, R_{\text{adj}}^2 = 0.893, S = 0.318, F = 72.57,$$

$$\text{PRESS} = 4.840, Q^2 = 0.880, \overline{r_{\text{m(LOO)}}^2} = 0.832,$$

$$\Delta r_{\text{m(LOO)}}^2 = 0.075, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 15, Q_{\text{F1}}^2 = 0.919, Q_{\text{F2}}^2 = 0.919, \overline{r_{\text{m(test)}}^2} = 0.825, \Delta r_{\text{m(test)}}^2 = 0.069,$$

$$\text{CCC} = 0.954, \text{MAE} = \text{Good}$$

Model M4. $\log k_{\infty} = -8.72(\pm 0.782) + 0.785(\pm 0.052) \times \text{X0v}$
 $+ 0.650(\pm 0.158) \times \text{F03[O-O]}$
 $+ 3.51(\pm 0.527) \times \text{ETA_BetaP}$
 $+ 0.202(\pm 0.073) \times \text{minsCH}_3$

$$n_{\text{training}} = 44, R^2 = 0.893, R_{\text{adj}}^2 = 0.882, S = 0.334, F = 81.11,$$

$$\text{PRESS} = 5.164, Q^2 = 0.872, \overline{r_{\text{m(LOO)}}^2} = 0.821,$$

$$\Delta r_{\text{m(LOO)}}^2 = 0.092, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 15, Q_{\text{F1}}^2 = 0.918, Q_{\text{F2}}^2 = 0.917, \overline{r_{\text{m(test)}}^2} = 0.806, \Delta r_{\text{m(test)}}^2 = 0.074,$$

$$\text{CCC} = 0.953, \text{MAE} = \text{Good}$$

Model M5. $\log k_{\infty} = -8.42(\pm 0.773) + 0.785(\pm 0.052) \times \text{X0v}$
 $+ 3.29(\pm 0.526) \times \text{ETA_BetaP}$
 $+ 0.199(\pm 0.072) \times \text{minsCH}_3$
 $+ 0.566(\pm 0.137) \times \text{nHBint}_4$

$$n_{\text{training}} = 44, R^2 = 0.893, R_{\text{adj}}^2 = 0.882, S = 0.333, F = 81.33,$$

$$\text{PRESS} = 5.543, Q^2 = 0.863, \overline{r_{\text{m(LOO)}}^2} = 0.808,$$

$$\Delta r_{\text{m(LOO)}}^2 = 0.086, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 15, Q_{\text{F1}}^2 = 0.915, Q_{\text{F2}}^2 = 0.914, \overline{r_{\text{m(test)}}^2} = 0.798, \Delta r_{\text{m(test)}}^2 = 0.076,$$

$$\text{CCC} = 0.950, \text{MAE} = \text{Good}$$

3.1.1. The descriptors related to hydrogen bonding interactions. The functional group count descriptor, $n\text{ArOH}$, repre-

sents the number of aromatic hydroxyl groups present in the compound. This descriptor influences the adsorption properties of organic pollutants by MWCNTs as indicated by its positive regression coefficient. Thus, the compounds containing a large number of aromatic hydroxyl groups may enhance the adsorption properties of organic pollutants by MWCNTs as shown in compounds **13** (pyrogallol) (containing 3-OH groups), **5** (2-phenyl phenol) (containing 1-OH group) and **14** (2,4,6 trichlorophenol) (containing 1-OH group). On the other hand, the compounds containing no aromatic hydroxyl groups are detrimental for the adsorption affinity of organic pollutants by MWCNTs as shown in compounds **18** (4-chloroaniline), **36** (benzyl alcohol) and **42** (phenethyl alcohol) (these compounds contain no aromatic hydroxyl groups). Although some compounds containing no aromatic hydroxyl groups still show high adsorption affinity for the organic pollutants by MWCNTs, it is due to some other dominating descriptors present in the model. Thus, the substitution of electron donating groups like hydroxyl groups in the aromatic ring of organic pollutants could enhance the adsorption on MWCNTs.

A 2D atom pair descriptor, F04[N-O], indicates the frequency of the N-O fragment at topological distance 4. The positive regression coefficient of the descriptor suggests that an increase in N-O fragments at topological distance 4 is directly proportional to the adsorption affinity of organic pollutants. The greater number of fragments correlates to higher adsorption properties as observed in the case of compounds **19** (2-nitroaniline) and **27** (3-nitrophenol), while the absence of such fragments at topological distance 4 has no influence on the adsorption by MWCNTs as shown in compounds **18** (4-chloroaniline), **36** (benzyl alcohol) and **42** (phenethyl-alcohol). This descriptor also indicates that the frequency of two electronegative atoms of organic pollutants (electron donating or electron withdrawing groups) should be situated at topological distance 4 for better adsorption on MWCNTs. In the case of compound number **19**, nitrogen (-NH₂ group) acts as an electron donor and oxygen (-NO₂ group) acts as an electron withdrawing group, whereas in the case of compound number **27**, nitrogen (-NO₂ group) acts as an electron withdrawing group, and oxygen (-OH group) acts as an electron donating group.

The E-state descriptor, nHBint4 indicates the count of potential internal hydrogen bonds separated by four edges. The positive regression coefficient suggests that hydrogen bonds of organic pollutants have the propensity to play a dominant role in enhancing the adsorption properties. Thus, the organic pollutants bearing hydrogen-bonded groups separated by four path lengths are conducive to adsorption as shown in compounds **13** (pyrogallol), **19** (2-nitroaniline) and **48** (3-chlorophenol), whereas the absence of such fragment in organic pollutants are detrimental to the adsorption affinity as shown in compounds **6** (benzene), **11** (phenol) and **42** (phenethyl alcohol).

B03[O-O] is a 2D atom pair descriptor that indicates the presence or absence of the O-O fragment at topological

distance 3. The positive regression coefficient of the descriptor indicates that the higher the frequency of this fragment, the higher is the adsorption affinity. Thus, the presence of the O–O fragment at topological distance 3 favors the adsorption of organic pollutants by MWCNTs as shown in compounds no. 12 (catechol) and 13 (pyrogallol), while compounds no. 6 (benzene), 42 (phenethyl alcohol) and 36 (benzyl alcohol) show low adsorption because these compounds have no such fragments at topological distance 3.

Hydrogen bonding is one of the key mechanisms for the adsorption of organic contaminants on CNTs. The information obtained from the descriptors $n\text{ArOH}$, $\text{F04}[\text{N-O}]$, $n\text{HBint4}$, $\text{F03}[\text{O-O}]$ and $\text{B03}[\text{O-O}]$ suggested that there may be some hydrogen bonding interactions between organic pollutants and MWCNTs, which regulate the adsorption affinity (Fig. 3) of organic pollutants toward MWCNTs. In the case of the descriptor $n\text{ArOH}$, the aromatic hydroxyl group may form hydrogen bonds with the hydroxy/carboxylic groups on the CNTs surface and the hydrogen bonds may also form between the surface-adsorbed aromatic hydroxyl group-containing organic pollutants (phenolics) and dissolved phenolics. Here, the hydroxyl group is always connected to an aromatic ring. Thus, it is obvious that this aromatic ring of organic pollutants themselves can interact with CNTs by π - π interactions. The descriptor, $\text{F04}[\text{N-O}]$, also suggested that besides the hydrogen bonding interactions, there may also be a chance to form electrostatic interactions. The electron-withdrawing groups like NO_2 may also strengthen the π - π interactions formed between the benzene derivatives (acting as π -acceptor) and CNTs (acting as π -donor). In the case of $\text{B03}[\text{O-O}]$, two oxygen atoms (hydroxyl groups) are separated by topological distance 3 and can interact with CNTs by hydrogen bonding interactions. These two electronegative atoms of organic pollutants could also interact electrostatically with CNTs and strengthen the π - π interactions formed between the organic pollutants and MWCNTs.^{39,71} It is worth noting that although the C–O bond is detrimental to the adsorption of organic pollutants on CNTs, the frequency of the O–O fragment at topological distance 3 can suppress the detrimental effect of the C–O group and influence the adsorption affinity of organic pollutants on MWCNTs. The descriptors involved in the hydrogen bonding interactions between the organic pollutants and MWCNTs are depicted in Fig. 3.

3.1.2. The descriptors related to hydrophobic interactions.

A 2D atom pair descriptor, $\text{B06}[\text{C-Cl}]$, represents the presence or absence of the C–Cl bond at topological distance 6. The positive regression coefficient of this parameter suggests that the presence of such a fragment at topological distance 6 enhances the adsorption affinity of organic pollutants towards the MWCNTs as shown in compounds 50 (4-chloroacetophenone) and 57 (2-chloronaphthene). On the other hand, compounds like 11 (phenol), 22 (4-methylphenol) and 43 (3-methylbenzyl alcohol) show poor adsorption affinity for the MWCNTs due to the absence of such a fragment.

The descriptor $X0v$ indicates a valence connectivity index of the order 0, which can be calculated through Kier and

Hall's connectivity index as shown below. This descriptor contributed positively to the adsorption affinity of organic pollutants for the MWCNTs. Thus, the size of the organic pollutants plays a crucial role in regulating the adsorption affinity of organic pollutants to MWCNTs. It has been found that on increasing the numerical value of this descriptor, the adsorption affinity of organic pollutants for MWCNTs also increases, as shown in the case of compounds 1 (pyrene), 58 (azobenzene) and 5 (2-phenyl phenol) (bigger in size), while the adsorption affinity of organic pollutants for MWCNTs decreases in the case of compounds 6 (benzene), 11 (phenol) and 36 (benzyl alcohol) (smaller in size).

The valence connectivity index of the zeroth order can be calculated by the following:

$$X0v = \sum_{i=1}^n (\delta_i^v)^{-0.5}$$

$$\delta_i^v = \frac{Z_i^v - hi}{Z_i - Z_i^v - 1}$$

In the above equation, δ_i^v = the valence vertex degree, Z_i^v = valence electrons in the i th atom, hi = the number of hydrogen atoms connected to the i th atom, Z_i = the number of electrons in the i th atom.

The E-state indices of a particular atom in a certain molecule provide information on its electronic state of that particular atom, which in turn depends on π bonds, the lone pair of electrons and δ bonds that inform the quantitative availability of the valence electrons.⁷² The descriptor minsCH_3 indicates the minimum atom type E-state CH_3 . The positive regression coefficient of this descriptor indicates that the presence of the CH_3 group has an important role in influencing the adsorption properties of organic pollutants. The numerical value of this descriptor is directly proportional to the adsorption property, which suggests that with increasing the numerical value of this descriptor, the adsorption affinity of the organic pollutants also increases as evidenced by compounds 10 (2,4-dinitrotoluene), 50 (4-chloroacetophenone) and 52 (1-methylnaphthalene). On the other hand, the adsorption affinity of organic pollutants decreases with the absence of the CH_3 group as shown in compounds 6 (benzene), 11 (phenol) and 36 (benzyl alcohol).

Hydrophobic interactions between organic pollutants and CNTs are also an important mechanism for better adsorption. The descriptors, $\text{B06}[\text{C-Cl}]$, $X0v$ and minsCH_3 suggest that the organic pollutants may be adsorbed onto the MWCNTs by hydrophobic interactions. In the case of $\text{B06}[\text{C-Cl}]$ and $X0v$, the size of the molecules (for $\text{B06}[\text{C-Cl}]$, the distance between C and Cl atoms is six, which reflects the size of the molecules) plays an important role in the adsorption affinity. The size enhances the surface area of molecules, which can regulate the hydrophobic interactions between organic pollutants and MWCNTs. The methyl group (information obtained from minsCH_3 descriptor) and CNTs are hydrophobic in nature. Thus, an increase in the minsCH_3 value

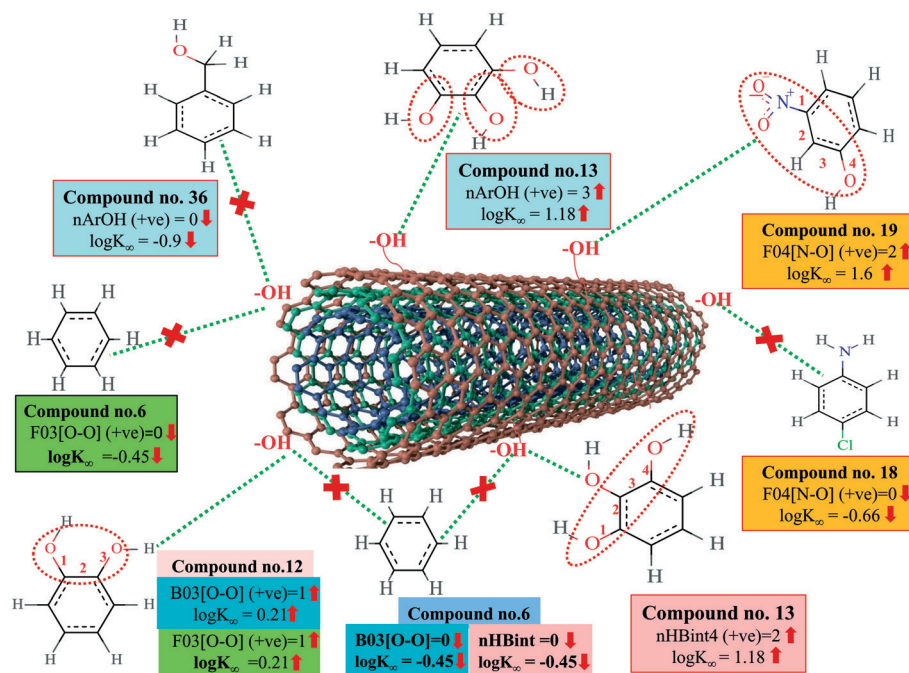


Fig. 3 Mechanistic interpretation of the descriptors related to hydrogen bonding interactions between organic pollutants and MWCNTs (dataset 1).

would indicate a higher degree of unsaturation and would enhance the reactivity. There is, therefore, a chance for hydrophobic interactions between organic pollutants and MWCNTs, which reflects better adsorption. The descriptors involved in hydrophobic interactions between organic pollutants and CNTs are depicted in Fig. 4.

3.1.3. The descriptors related to π - π interactions. The descriptor, U_i , gives information about the unsaturation index, which contributes positively to the adsorption affinity of or-

ganic pollutants by MWCNTs as indicated by the positive regression coefficient. From this descriptor, it has been suggested that the presence of unsaturated inorganic pollutants plays a crucial role in enhancing the adsorption affinity. This was demonstrated in compounds 1 (pyrene), 10 (2,4-dinitrotoluene) and 58 (azobenzene) (the numerical values of this descriptor are 3.392, 3 and 3, respectively), and *vice versa* in the case of compounds 11 (phenol), 36 (benzyl alcohol) and 42 (phenethyl alcohol) (the numerical values of this

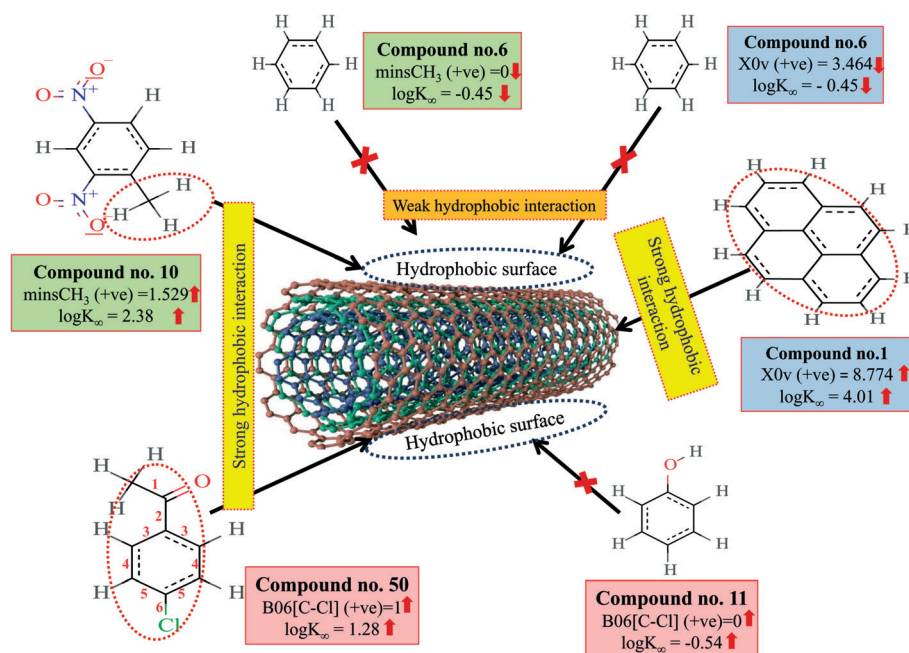


Fig. 4 Mechanistic interpretation of the descriptors related to the hydrophobic interaction between organic pollutants and MWCNTs (dataset 1).

descriptor are 2 in each compound). Here, the compounds, **1** (pyrene), **10** (2,4-dinitrotoluene) and **58** (azobenzene) have a higher range of unsaturation index values due to the presence of a large number of double bonds.

The ETA index, ETA_BetaP, gives a measure of sigma, pi and non-bonded (*i.e.*, lone pairs capable of forming resonance with the aromatic system) electrons relative to the molecular size. Therefore, electron-richness (unsaturation) relative to the molecular size of organic pollutants is an important parameter for regulating the adsorption properties. The positive regression coefficient of this parameter indicates that the electron densities of the molecules should be higher for increasing the adsorption affinity of organic pollutants for MWCNTs, as found in compounds **1** (pyrene), **28** (1,3-dinitrobenzene) and **58** (azobenzene), whereas the compounds with low electron density show a lower range of adsorption affinities as shown in compounds **36** (benzyl alcohol), **42** (phenethyl alcohol) and **43** (3-methylbenzyl alcohol). Thus, it can be concluded that the molecules should be electron-rich for higher adsorption properties of organic pollutants.

The π - π interaction is another important mechanism involved in the adsorption of organic pollutants to CNTs. The information obtained from Ui and ETA_BetaP descriptors suggested that the organic pollutants can adsorb to MWCNTs by strong π - π interactions. The descriptors B03[O-O], F03[O-O] and F04[N-O] suggested that the [O-O] fragments at topological distance 3 and the [N-O] fragments at the topological distance 4 may strengthen the π - π interactions formed between organic pollutants and MWCNTs. The descriptor Ui suggested that unsaturation plays a crucial role for the adsorption of organic pollutants to MWCNTs. CNTs also con-

tain a large number of double bonds (unsaturation), so there is a chance to form strong π - π interactions between organic pollutants and MWCNTs, which reflects the better adsorption of these pollutants to MWCNTs; hence, a higher number of double bonds of organic pollutants enhance the adsorption affinity to MWCNTs. The descriptor, ETA_BetaP suggested that unsaturation (electron-richness) relative to the molecular size of organic pollutants plays a crucial role in regulating the adsorption properties. From this descriptor, it can be inferred that the adsorption affinity of organic pollutants to MWCNTs is increased due to the π - π interactions. The descriptors involved in π - π interactions between organic pollutants and CNTs are described graphically in Fig. 5.

3.1.4. The descriptors related to electrostatic interactions. F03[O-O], a 2D atom pair descriptor, indicates the frequency of the O-O fragment at topological distance 3. The positive regression coefficient of this descriptor suggests that presence of a greater number of O-O bonds at the topological distance 3 might be beneficial for the adsorption affinity of organic pollutants for MWCNTs as shown in compounds **12** (catechol) and **13** (pyrogallol), whereas the opposite happens in the case of compounds **6** (benzene), **42** (phenethyl alcohol) and **43** (3-methylbenzyl alcohol) (where, no O-O fragment is present at topological distance 3). This fragment may also strengthen the π - π interactions formed between organic pollutants and MWCNTs.^{73,74} Like B03[O-O], this descriptor also suppresses the detrimental effect of the C-O group as discussed earlier in this section.

The information obtained from the descriptors, F03[O-O], B03[O-O] and F04[N-O] suggests that the organic pollutants can adhere to the surface of the MWCNTs by strong

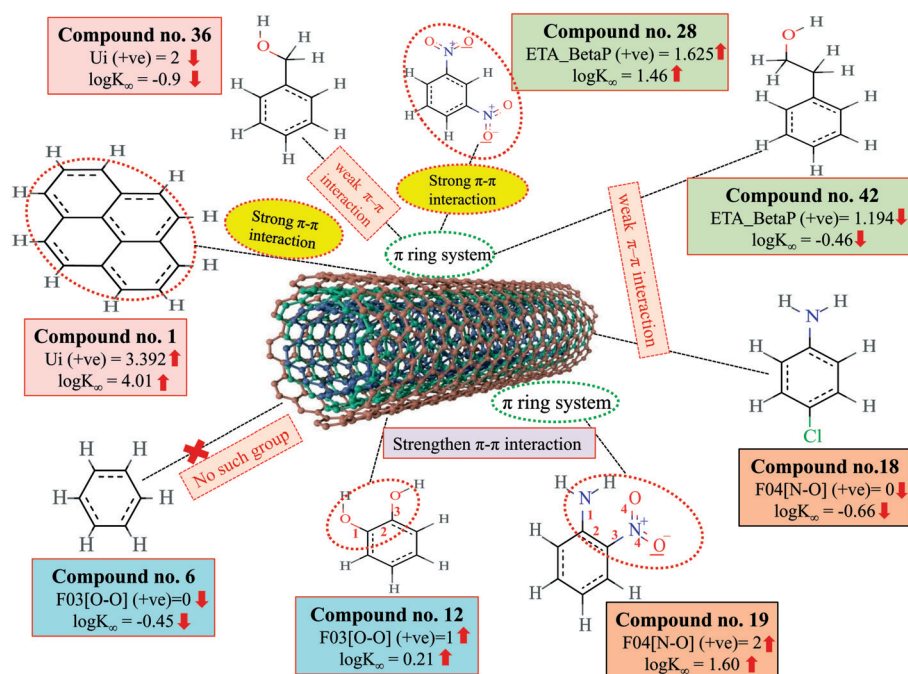


Fig. 5 Mechanistic interpretation of the descriptors related to the π - π interactions between organic pollutants and MWCNTs (dataset 1).

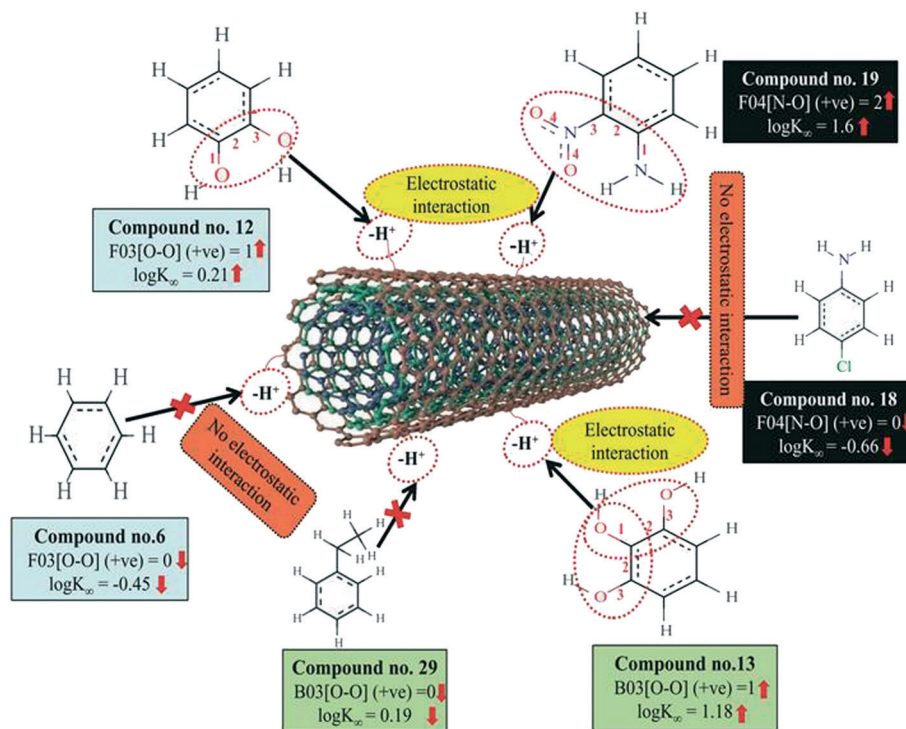


Fig. 6 Mechanistic interpretation of the descriptors related to the electrostatic interactions between organic pollutants and MWCNTs (dataset 1).

electrostatic interactions. The descriptors F03[O-O] and B03[O-O] indicate that the frequency or presence/absence of two electronegative atoms (electron donating group) at the topological distance 3 is essential to enhance the adsorption affinity of organic pollutants to MWCNTs. Thus, there may be a chance to form electrostatic interactions between organic pollutants (negatively charged atom like oxygen atom of the hydroxyl group) and MWCNTs (the sidewall of the CNTs are electrically polarizable and thus polar molecules can easily adhere to their surface). The descriptors involved for electrostatic interactions between organic pollutants and CNTs are represented graphically in Fig. 6.

The 2D atom pair descriptor, B01[C-O], indicates the presence or absence of the C-O bond at topological distance 1. The negative regression coefficient of the descriptor supports that the presence of this fragment at topological distance one is detrimental to the adsorption affinity of organic pollutants by MWCNTs, though it can form hydrogen bonds with MWCNTs. For example, compounds like 1 (pyrene), 57 (2-chloronaphthalene) and 58 (azobenzene) have higher adsorption affinity value due to the absence of such fragments at topological distance 1, whereas compounds like 11 (phenol), 36 (benzyl alcohol) and 42 (phenethyl alcohol) have lower adsorption affinity due to the presence of one C-O bond in each compound.

3.2. Dataset 2: 69 organic pollutants

The significant descriptors obtained from the five MLR models using the adsorption properties ($\log K_{SA}$) of 69 organic pollutants related to the specific surface area of

MWCNTs are Eta_Epsilon_3, X1A, X2A, nOHp, VAdjMat, F04(O-Cl), B05(O-Cl), MLOGP2, T(N...N), O%, and T(O...Cl). We have discussed here all the significant descriptors, which are the key properties for altering the adsorption properties of organic pollutants. The definition, contribution and frequency of the modeled descriptors are shown in Table S5 in the ESI.† The applicability domain of the developed models using the standardization approach showed that one test set compound (compound number 10) for model N1, two test set compounds (compound number 10 and 21) for model N2, one test set compound (compound number 21) for model N3 are situated outside the applicability domain, while in the case of model nos. 4 and 5, all the test set compounds are situated within the domain of applicability. The scatter plot of observed vs. predicted adsorption coefficient related to the specific surface area of MWCNTs for all the MLR models are shown in Fig. 7.

$$\begin{aligned} \text{Model N1. } \log K_{SA} = & 4.29(\pm 2.194) + 0.0965(\pm 0.014) \times \text{O\%} \\ & - 16.4(\pm 4.397) \times \text{X1A} + 0.145(\pm 0.032) \\ & \times \text{T(N}\cdots\text{N)} - 0.0279(\pm 0.009) \\ & \times \text{T(O}\cdots\text{Cl)} - 1.01(\pm 0.294) \\ & \times \text{B05(Cl}\cdots\text{Cl)} + 0.203(\pm 0.022) \\ & \times \text{MLOGP2} \end{aligned}$$

$$n_{\text{training}} = 52, R^2 = 0.845, R^2_{(\text{adj})} = 0.824, Q^2 = 0.798, S = 0.433,$$

$$\text{PRESS} = 11.003, F = 40.79, r^2_{m(\text{LOO})} = 0.709,$$

$$\Delta r^2_{m(\text{LOO})} = 0.087, \text{MAE} = \text{Moderate}$$

Experimental $\log K_{SA}$ vs predicted $\log K_{SA}$ values of 69 organic pollutants

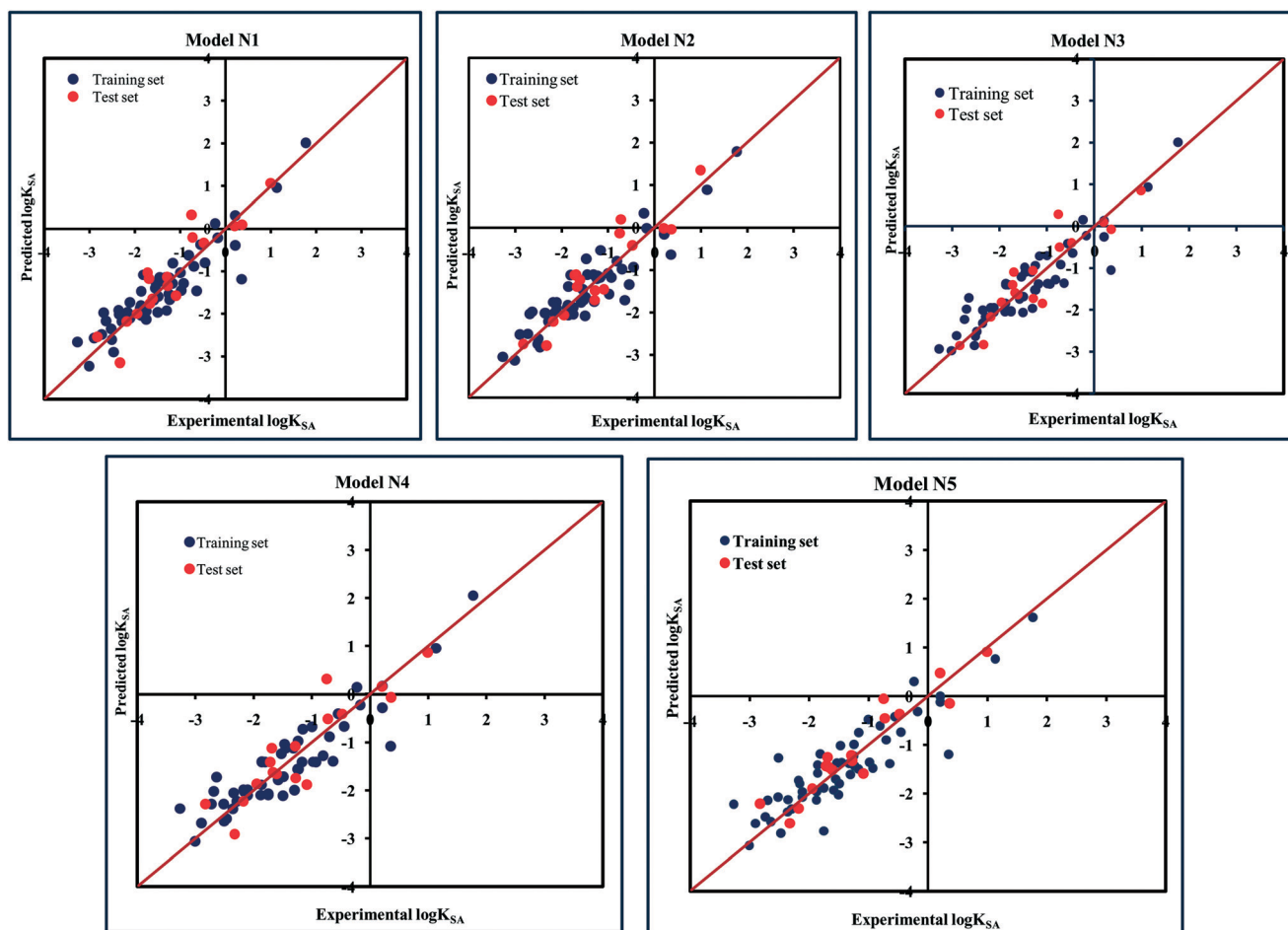


Fig. 7 The scatter plots of the observed and the predicted adsorption coefficient properties related to the specific surface area of MWCNTs ($\log K_{SA}$) of the developed MLR models (models N1–N5).

$$n_{\text{test}} = 17, Q_{F1}^2 = 0.809, Q_{F2}^2 = 0.795, \overline{r_{m(\text{test})}^2} = 0.783, \Delta r_{m(\text{test})}^2 = 0.048, \\ \text{CCC} = 0.908, \text{MAE} = \text{Moderate}$$

$$n_{\text{test}} = 17, Q_{F1}^2 = 0.830, Q_{F2}^2 = 0.818, \overline{r_{m(\text{test})}^2} = 0.805, \Delta r_{m(\text{test})}^2 = 0.050, \\ \text{CCC} = 0.918, \text{MAE} = \text{Good}$$

$$\text{Model N2. } \log K_{SA} = -7.19(\pm 0.571) + 0.0805(\pm 0.015) \times \text{O\%} \\ - 0.662(\pm 0.323) \times \text{nOHp} \\ - 0.0358(\pm 0.009) \times T(\text{O}\cdots\text{Cl}) \\ - 0.943(\pm 0.294) \times \text{B05}(\text{Cl}\cdots\text{Cl}) \\ + 0.185(\pm 0.019) \times \text{MLOGP2} \\ + 0.958(\pm 0.144) \times \text{VAdjMat}$$

$$\text{Model N3. } \log K_{SA} = -42.3(\pm 7.527) + 0.0973(\pm 0.013) \times \text{O\%} \\ - 0.622(\pm 0.323) \times \text{nOHp} \\ + 0.154(\pm 0.031) \times T(\text{N}\cdots\text{N}) \\ - 0.0407(\pm 0.008) \times T(\text{O}\cdots\text{Cl}) \\ + 0.160(\pm 0.20) \times \text{MLOGP2} \\ + 89.8(\pm 17.51) \times \text{ETA_Epsilon_3}$$

$$n_{\text{training}} = 52, R^2 = 0.842, R_{(\text{adj})}^2 = 0.821, Q^2 = 0.790, S = 0.437, \\ \text{PRESS} = 11.41, F = 39.97, \overline{r_{m(\text{LOO})}^2} = 0.723, \\ \Delta r_{m(\text{LOO})}^2 = 0.114, \text{MAE} = \text{Moderate}$$

$$n_{\text{training}} = 52, R^2 = 0.842, R_{(\text{adj})}^2 = 0.821, Q^2 = 0.788, S = 0.436, \\ \text{PRESS} = 11.512, F = 40.07, \overline{r_{m(\text{LOO})}^2} = 0.714, \\ \Delta r_{m(\text{LOO})}^2 = 0.081, \text{MAE} = \text{Good}$$

$$n_{\text{test}} = 17, Q_{F1}^2 = 0.783, Q_{F2}^2 = 0.768, \overline{r_{m(\text{test})}^2} = 0.712, \Delta r_{m(\text{test})}^2 = 0.14, \\ \text{CCC} = 0.890, \text{MAE} = \text{Good}$$

$$\text{Model N4. } \log K_{\text{SA}} = -42.0(\pm 7.743) + 0.101(\pm 0.014) \times \text{O\%} \\ + 0.159(\pm 0.032) \times T(\text{N}\cdots\text{N}) \\ - 0.0411(\pm 0.008) \times T(\text{O}\cdots\text{Cl}) \\ + 0.168(\pm 0.021) \times \text{MLOGP2} \\ + 88.9(\pm 18.01) \times \text{ETA_Epsilon}_3$$

$$n_{\text{training}} = 52, R^2 = 0.829, R_{(\text{adj})}^2 = 0.811, Q^2 = 0.785, S = 0.449, \\ \text{PRESS} = 11.722, F = 44.73, \overline{r_{m(\text{LOO})}^2} = 0.709, \\ \Delta r_{m(\text{LOO})}^2 = 0.087, \text{MAE} = \text{Good}$$

$$n_{\text{test}} = 17, Q_{F1}^2 = 0.812, Q_{F2}^2 = 0.799, \overline{r_{m(\text{test})}^2} = 0.748, \Delta r_{m(\text{test})}^2 = 0.044, \\ \text{CCC} = 0.903, \text{MAE} = \text{Moderate}$$

$$\text{Model N5. } \log K_{\text{SA}} = 2.49(\pm 1.36) + 0.0757(\pm 0.016) \times \text{O\%} \\ - 17.3(\pm 3.773) \times \text{X2A} + 0.145(\pm 0.036) \\ \times T(\text{N}\cdots\text{N}) - 0.721(\pm 0.144) \\ \times \text{F04}(\text{O}\cdots\text{Cl}) + 0.158(\pm 0.023) \\ \times \text{MLOGP2}$$

$$n_{\text{training}} = 52, R^2 = 0.793, R_{(\text{adj})}^2 = 0.77, Q^2 = 0.743, S = 0.495, \\ \text{PRESS} = 13.955, F = 35.17, \overline{r_{m(\text{LOO})}^2} = 0.709, \\ \Delta r_{m(\text{LOO})}^2 = 0.087, \text{MAE} = \text{Good}$$

$$n_{\text{test}} = 17, Q_{F1}^2 = 0.890, Q_{F2}^2 = 0.882, \overline{r_{m(\text{test})}^2} = 0.836, \Delta r_{m(\text{test})}^2 = 0.090, \\ \text{CCC} = 0.940, \text{MAE} = \text{Good}$$

3.2.1. The descriptors related to the hydrophobic interaction. The descriptor, X1A, indicates an average connectivity index of the order one, it encodes the ‘chi’ value across one bond, which can be calculated on the basis of Kier and Hall’s connectivity index and defined as follows:

$${}^1X = \sum_{b=1}^B (\delta_i \cdot \delta_j)_b^{-0.5}$$

In this equation, b runs over the 1st order subgraphs having n vertices with B edges; δ_i and δ_j are the number of other vertices attached to vertex i and j , respectively. The negative regression coefficient of this descriptor implies that the higher numerical values of this descriptor are not favorable to enhance the adsorption properties of organic pollutants related to the specific surface area of MWCNTs as shown in compounds 3 (benzene), 56 (ethylbenzene) and 57 (benzyl al-

cohol) (the corresponding numerical values of these compounds are 0.5, 0.491, 0.491, respectively, showing a lower range of adsorption affinity). On the other hand, compounds like 35 (tetracycline), 22 (pyrene) and 26 (phenanthrene) show better adsorption affinity ($\log K_{\text{SA}}$) due to their lower numerical values of this descriptor.

Another significant descriptor, X2A, indicates an average connectivity index of the order 2, and encodes the ‘chi’ value across two bonds, which can be calculated on the basis of Kier and Hall’s connectivity index, defined in the following equation:

$${}^2X = \sum_{b=2}^B (\delta_i \cdot \delta_j)_b^{-0.5}$$

Here, b runs over the 2nd order subgraphs having n vertices with B edges, δ_i and δ_j are the numbers of other vertices attached to vertex i and j , respectively. This descriptor also has a negative contribution towards the adsorption profile ($\log K_{\text{SA}}$) of organic pollutants by MWCNTs as evidenced by the negative regression coefficient. This indicates that the adsorption properties of organic pollutants decrease with an increase in the numerical value of this descriptor as shown in compounds 3 (benzene), 18 (aniline) and 40 (bromobenzene), and *vice versa* in the case of compounds 22 (pyrene), 26 (phenanthrene) and 35 (tetracycline).

The VAdjMat descriptor represents the vertex adjacency information and gives information about molecular dimension and hydrophobicity. This descriptor can be calculated by using the following formula:

$$\text{VAdjMat} = 1 + \log_2(m)$$

Here, m depicts the number of heavy-heavy bonds. This descriptor contributed positively towards the adsorption properties ($\log K_{\text{SA}}$) of organic pollutants as indicated by the positive regression coefficient. Thus, the higher numerical value of this descriptor is influential toward the adsorption affinity of organic pollutants. This indicates that hydrophobicity plays a crucial role in altering the adsorption properties of organic pollutants by MWCNTs. For example, compounds 22 (pyrene), 26 (phenanthrene) and 35 (tetracycline) show a higher range of adsorption properties as these compounds contain higher numerical values of this descriptor. Compounds 3 (benzene), 55 (iodobenzene) and 46 (chlorobenzene) show a lower range of adsorption properties as these compounds contain higher numerical values of this descriptor. It is therefore suggested that the hydrophobic organic pollutants can easily be adsorbed by MWCNTs through hydrophobic interactions between the pollutants and CNTs.

The next descriptor, MLOGP2, represents the squared Moriguchi octanol–water partition coefficient, calculated from the regression equation of the Moriguchi $\log P$

model^{75,76} consisting of 13 parameters as depicted in the following equation.

$$\begin{aligned} \log P = & -1.244(\text{CX})^{0.6} - 1.017(\text{NO})^{0.9} + 0.406\text{PRX} - 0.145(\text{UB})^{0.8} \\ & + 0.511\text{HB} + 0.268\text{POL} - 2.215\text{AMP} + 0.912\text{ALK} \\ & - 0.392\text{RNG} - 3.684\text{QN} + 0.474\text{NO}_2 + 1.582\text{NCS} \\ & + 0.773\text{BLM} - 1.041 \end{aligned}$$

'CX' depicts the summation of the weighted number of carbon atoms; 'NO' depicts the total number of N and O atoms; 'PRX' represents the proximity effect of N/O; 'UB' represents the number of unsaturated bonds including semi-polar bonds; 'POL' depicts the number of aromatic polar substituents; 'AMP' depicts the amphoteric property; 'ALK' represents the dummy variable for alkanes and alkenes; 'RNG' depicts the indicator variable for the presence of a ring structure, except for benzene and its condensed ring; 'QN' represents quaternary nitrogen; 'NO₂' represents the number of nitro groups; 'HB' represents a dummy variable for the presence of intermolecular hydrogen bonds; 'NCS' depicts isothiocyanato or thiocyanato; 'BLM' represents a dummy variable for the presence of β -lactam.

The positive regression coefficient of this descriptor indicates that hydrophobicity plays a crucial role in regulating the adsorption properties of organic pollutants. The highly hydrophobic organic pollutants can easily be adsorbed by MWCNTs as evidenced by compounds 22 (pyrene), 26 (phenanthrene) and 34 (azobenzene) as their corresponding MLOG2 values are 22.653, 18.762 and 10.539, respectively, whereas hydrophilic molecules are poorly adsorbed by MWCNTs as evidenced by compounds 18 (aniline), 57

(benzylalcohol) and 63 (3-nitroaniline) as their corresponding MLOGP2 values are 2.268, 2.532 and 1.816 respectively. Therefore, it can be inferred that the organic pollutants are adsorbed onto the CNTs through hydrophobic interactions. Thus, for proper adsorption, organic pollutants should be hydrophobic in nature. Note that this was also observed in the case of the VAdjMat descriptor as discussed previously. MLOGP2 is not strictly a 2D descriptor. Here, the term 'intramolecular H-bonds' is used to calculate the MLOGP value, which is conformation dependent.

The information obtained from the descriptors X1A, X2A, VAdjMat and MLOGP2 suggested that the adsorption of organic pollutants related to the specific surface area of MWCNTs may occur through hydrophobic interactions. The molecular connectivity index (X1A and X2A) has a direct relationship with the count of interacting C-H bonds present in a molecule. The number of C-H bonds in a molecule is equal to the number of H atoms. As the C-H bond increases, the hydrophobicity of the molecule increases. The δ value (depends on the number of H atoms, the definition of a δ value for a carbon atom in a molecular graph is: $\delta = 4 - H$) decreases with the average connectivity index. Thus, the hydrophobic interactions between the organic contaminants and MWCNTs are reduced and the adsorption of organic pollutants related to the specific surface area of MWCNTs may also be reduced.⁷⁷

The descriptors VAdjMat and MLOGP2 give information about the hydrophobicity of molecules. It is obvious that the hydrophobic organic pollutants will interact with hydrophobic CNTs through hydrophobic interactions. This implies that the hydrophobic organic pollutants can be easily adsorbed by

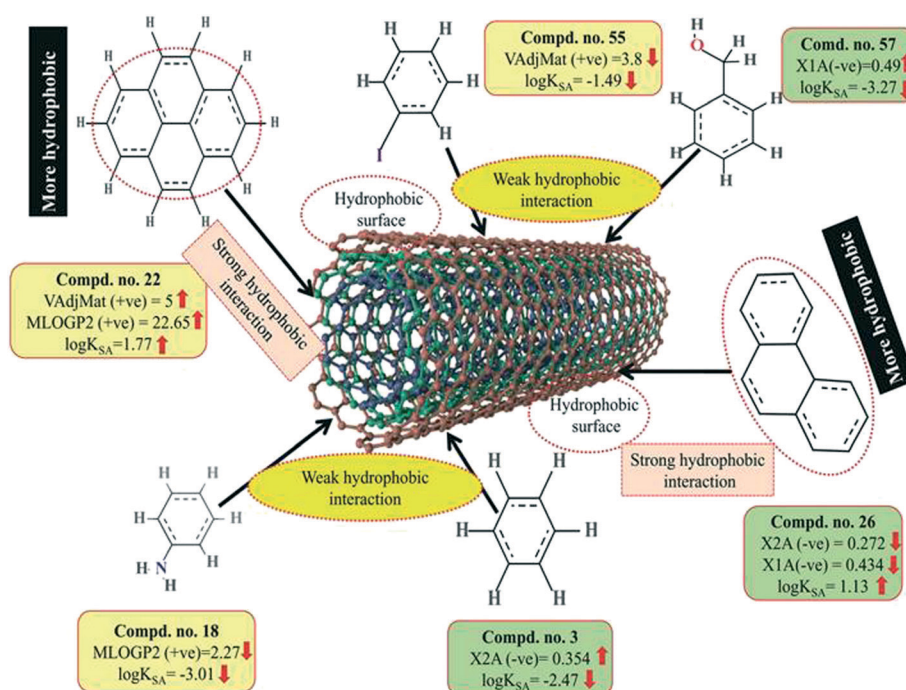


Fig. 8 Mechanistic interpretation of the descriptors related to the hydrophobic interactions between organic pollutants and MWCNTs (dataset 2).

MWCNTs through hydrophobic interactions. The descriptors involved for hydrophobic interaction are graphically depicted in Fig. 8.

3.2.2. The descriptors related to the π - π interactions. A functional group count descriptor, nOHp, describes the number of primary alcohols. The negative regression coefficient of this descriptor points out that the primary alcoholic group is not favored to enhance the adsorption properties ($\log K_{SA}$) of organic pollutants as found in compounds 13 (3-methyl benzyl alcohol) and 57 (benzyl alcohol). On the contrary, organic pollutants that do not contain any primary alcoholic groups have higher adsorption affinities ($\log K_{SA}$) as shown in compounds 22 (pyrene), 26 (phenanthrene) and 34 (azobenzene). Thus, the organic pollutants that do not contain any primary alcoholic groups may be highly adsorbed by MWCNTs.

F04[O-Cl] is a 2D atom pair descriptor that indicates the number of (O-Cl) fragments at a topological distance of 4. The negative regression coefficient of this descriptor indicates that the frequency of the O-Cl fragment at the topological distance 4 is inversely proportional to the adsorption properties of organic pollutants. A higher number for this fragment correlates to lower adsorption properties of organic pollutants, as observed in compounds 7 (dicamba), 61 (3-chlorophenol) and 66 (2,4,5-trichlorophenoxyacetic acid) (these compounds contain 3, 1 and 1 such fragments, respectively, at a topological distance of 4), while a lower numerical value of this descriptor correlates to a higher adsorption property of organic pollutants as observed in compounds 22 (pyrene), 26 (phenanthrene), 34 (azobenzene) and 69 (2,4-dinitrotoluene) (these compounds contain no such fragments at topological distance 4). Thus, the presence of this fragment at the topological distance 4 may hinder the adsorption of the organic pollutants by MWCNTs. The adsorption of organic contaminants to the CNTs decreases when the frequency of the (O-Cl) fragment at topological distance 4 increases. Compound 2 (2,4,6-trichlorophenol) also contains a O-Cl fragment but not at topological distance 4. Therefore, the adsorption affinity related to the specific surface area of the MWCNTs value of compound 2 is ($\log K_{SA}$ value = -0.81) not low as compared to compounds 7 (dicamba), 61 (3-chlorophenol) and 66 (2,4,5-trichlorophenoxyacetic acid) (these compounds contain 3, 1 and 1 such fragments, respectively, at topological distance 4 and the $\log K_{SA}$ values are -2.64, -1.75 and -2.51, respectively).

T(O \cdots Cl), a 2D atom pair descriptor, indicates the sum of the topological distance between oxygen and chlorine. The negative regression coefficient of this descriptor suggests that a higher numerical value of this descriptor is detrimental to enhancing the adsorption properties of organic pollutants related to the specific surface area of MWCNTs as shown in compounds 2 (2,4,6-trichlorophenol), 7 (dicamba) and 66 (2,4,6-trichlorophenoxyacetic acid). On the other hand, the organic pollutants containing no such fragments have higher adsorption properties as shown in compounds 22 (pyrene), 26 (phenanthrene) and 34 (azobenzene). From this observa-

tion, it can be inferred that the organic pollutants without (O \cdots Cl) fragments may be better adsorbed onto the MWCNTs surface.

A 2D atom pair descriptor, B05(Cl-Cl), describes the presence or absence of Cl-Cl fragments at topological distance 5. The negative regression coefficient of this descriptor indicates that the presence of the Cl-Cl fragment at the topological distance 5 may reduce the adsorption property of organic pollutants related to the specific surface area of MWCNTs ($\log K_{SA}$). A higher number of this fragment correlates to lower adsorption property of organic pollutants as observed in compounds 7 (dicamba), 41 (1,2,4-trichlorobenzene) and 66 (2,4,5-trichlorophenoxyacetic acid) (containing one such fragment each) while absence of this fragment in organic pollutants correlates to higher adsorption property as evidenced from compounds 22 (pyrene), 26 (phenanthrene) and 34 (azobenzene). From this descriptor, it can be suggested that the presence of this fragment at topological distance 5 may retard adsorption of the organic pollutants by MWCNTs.

Another 2D atom pair descriptor, T(N \cdots N), indicates the sum of the topological distances between two nitrogen atoms. A positive contribution towards the adsorption properties of organic pollutants related to the specific surface area of MWCNTs ($\log K_{SA}$) indicates that for better adsorption of organic pollutants by MWCNTs, the topological distance between two nitrogen atoms should be greater, as shown in compounds 4 (oxytetracycline), 35 (tetracycline) and 69 (2,4-dinitrotoluene) (as their corresponding topological distances between two nitrogen atoms are 5, 5 and 4, respectively), and *vice versa* in the case of compounds 42 (isophorone), 43 (4-fluorophenol) and 44 (acetophenone). Thus, it can be inferred that the topological distances between two nitrogen atoms should be greater for the better adsorption of organic pollutants by MWCNTs.

As discussed earlier in the introduction section, π - π interactions are one of the key mechanisms for the adsorption of organic pollutants to CNTs. The information obtained from these descriptors, nOHp, F04[O-Cl], B05[Cl-Cl], T(N \cdots N) and T(O \cdots Cl), strongly support this statement. The descriptor nOHp weakens the π - π interaction that occurs between the organic pollutants and CNTs. In this case, the hydroxyl group is alcoholic in nature (aliphatic hydroxyl group) and cannot donate the lone pair of electrons to the aromatic ring (not directly bonded to the aromatic carbon) and ultimately weaken the π - π interactions of the aromatic ring, though it can form hydrogen bonds with the surface modified CNTs. On the other hand, the phenolic hydroxyl group can donate the lone pair of electrons to the aromatic ring (bonded directly to the aromatic carbon atom) as discussed previously (section 3.1), thus strengthening the π - π interactions between organic pollutants and CNTs. In the case of the phenolic hydroxyl group, it can also act as a π donor, but this is not possible in case of the alcoholic hydroxyl group. From this observation, it can be suggested that the aliphatic hydroxyl (alcoholic) group is not favorable for the adsorption affinity of organic pollutants to the CNTs. In case of the descriptors B05[Cl-Cl], T(O \cdots Cl) and

F04[O-Cl], the chlorine atom has an electron inductive effect and decreases the electron density in the benzene ring, which compensates for the electron-donating effect of the oxygen atom (in the case of compounds 7 and 66), even after $-OH$ dissociated into $-O^-$. The withdrawing inductive character of chlorine substituents decreases the electron density of the *p*-chlorophenol ring as compared with that of the phenol ring. Thus, when the O-Cl or Cl-Cl fragment is present in an aromatic molecule, it decreases the electron density of that aromatic ring (as compared with that of the $-OH$ substituted benzene ring (phenolic) or the benzene ring itself) and ultimately, electron donor-acceptor interactions do not occur easily between CNTs and organic contaminants. Hence, the compound could not be easily adsorbed to the MWCNTs. In case of the descriptor T(N \cdots N), the lone pair of electrons of the nitrogen atom can be donated to the ring system (when directly attached) and enhance the π - π interaction with the CNTs. The nitrogen can be present as the amino form (electron donating) or in the nitro form (electron withdrawing). Both forms strengthen the π - π interactions between the organic pollutants and CNTs by increasing or decreasing the π -electron density of the aromatic ring system and act as π electron donor or acceptor, respectively. If the nitrogen is not directly attached to the aromatic ring system, then adsorption happens through electrostatic interactions between the nitrogen of the pollutants and the hydrogen of CNTs by forming dipoles when they are close to each other; the position of the nitrogen atom hardly matters here. The descriptors influencing the π - π interaction are graphically represented in Fig. 9.

3.2.3. The descriptors related to hydrogen bonding interactions. The descriptor, O%, indicates the percentage of oxygen atoms present in a particular molecule. The positive regression coefficient of this descriptor suggests that the presence of oxygen atom is highly influential in the adsorption of the organic pollutants on the surface of MWCNTs. For example, compounds 4 (oxytetracycline), 35 (tetracycline) and 69 (2,4-dinitrotoluene) show better adsorption affinity as their corresponding percentages of oxygen atoms are 15.8, 14.3 and 21.1, respectively. In contrast, compounds 3 (benzene), 18 (aniline) and 24 (4-chloroaniline) show poor adsorption affinity as these compounds do not contain any oxygen atoms. The oxygen atom may be present in different organic pollutants in keto, phenolic (favorable for adsorption) or alcoholic forms (not favorable for adsorption as discussed previously). These different types of oxygen may interact with CNTs in different ways, e.g., hydrogen bonding, strengthening the π - π interactions and electrostatic interactions. On the other hand, a high percentage of oxygen atoms may enhance the polarity of the pollutants. Since the sidewalls of the CNTs are also electrically polarized, the polar group of organic pollutants can easily adhere to the surface of the CNTs. The descriptor involved for hydrogen bonding interactions is given in Fig. 10.

3.2.4. The descriptors related to the electrostatic interactions. The descriptor, Eta_Epsilon_3, indicates the summation of epsilon values relative to the total number of atoms including hydrogen in the connected molecular graph of the reference alkane, which can be calculated by the following equation.

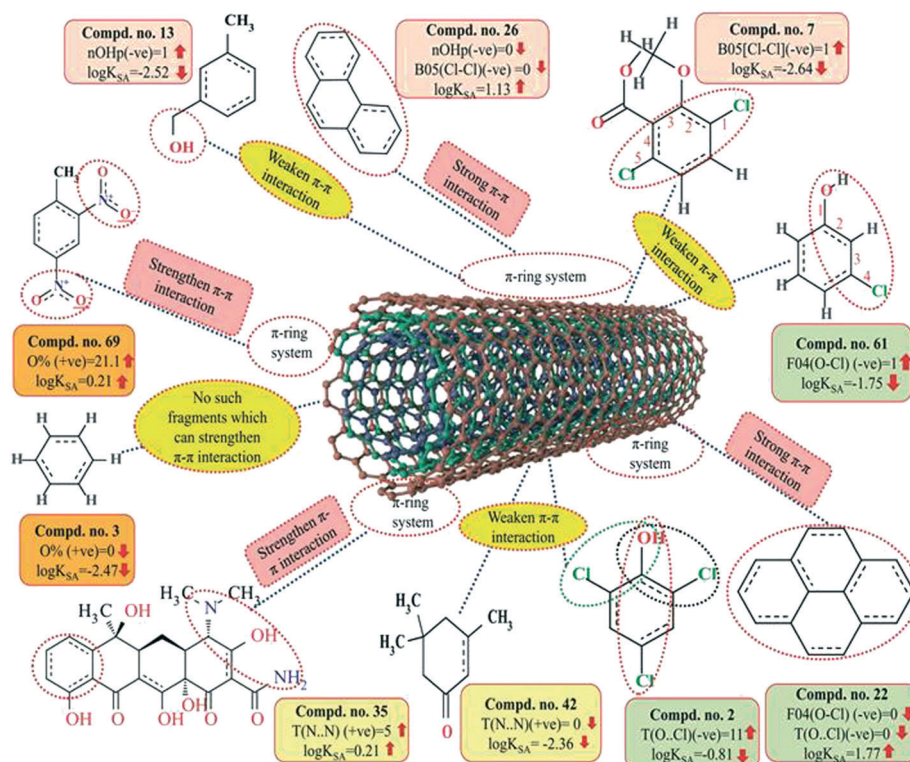


Fig. 9 Mechanistic interpretation of the descriptors related to π - π interactions between organic pollutants and MWCNTs (dataset 2).

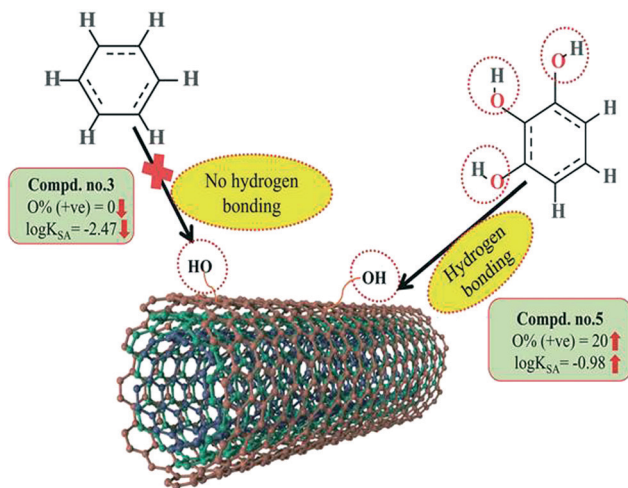


Fig. 10 Mechanistic interpretation of the descriptors related to hydrogen bonding interactions between organic pollutants and MWCNTs (dataset 2).

$$\varepsilon_3 = \varepsilon_R / N_R$$

ε denotes electronegativity, N_R denotes the number of atoms present in the reference alkane. This descriptor has a positive contribution towards the adsorption properties of organic pollutants related to the specific surface area of MWCNTs. This indicates that the electron-rich organic pollutants will be highly adsorbed by MWCNTs. Thus, the higher numerical value (due to strong electrostatic interactions between organic pollutants and CNTs) of this descriptor is required to increase the adsorption properties of organic pollutants by MWCNTs as shown in compounds 22

(pyrene), 26 (phenanthrene) and 35 (tetracycline) and *vice versa* in the case of compounds 7 (dicamba), 13 (3-methylbenzyl alcohol) and 18 (aniline) (due to weak electrostatic interactions between these organic pollutants and CNTs).

The information obtained from the descriptor O% suggests that the organic pollutants can adhere to the surface of MWCNTs by electrostatic interactions. There may be a chance to form electrostatic interactions between organic pollutants (negatively charged atoms like the oxygen atom of the hydroxyl group) and MWCNTs (sidewalls of the CNTs are electrically polarizable, thus polar molecules can easily adhere to their surface). The descriptors involved in electrostatic interactions are shown graphically in Fig. 11.

3.3. Dataset 3 : 29 organic solvents

The significant descriptors obtained from the PLS model using the dispersibility index ($\log C_{\max}$) values of 29 organic solvents to SWCNTs are minsssN, SpMin3_Bhe, VPC-6 and SpMin6_Bhi (arranged according to the variable importance plot, Fig. S2 in ESI†). The modeled descriptors, which are the key properties altering the dispersibility indexes of organic solvents, are discussed below. We have also checked the applicability domain of test set compounds using the DModX approach (99% confidence level) to find out whether any test set compounds lie outside of the AD (D-critical = 4.559). The results suggested that the entire test set compounds lie within the AD, except for compound number 29 (Fig. S3 in ESI†). The scatter plot of the observed *vs.* predicted dispersibility index of SWCNTs in different solvents are presented in Fig. 12.

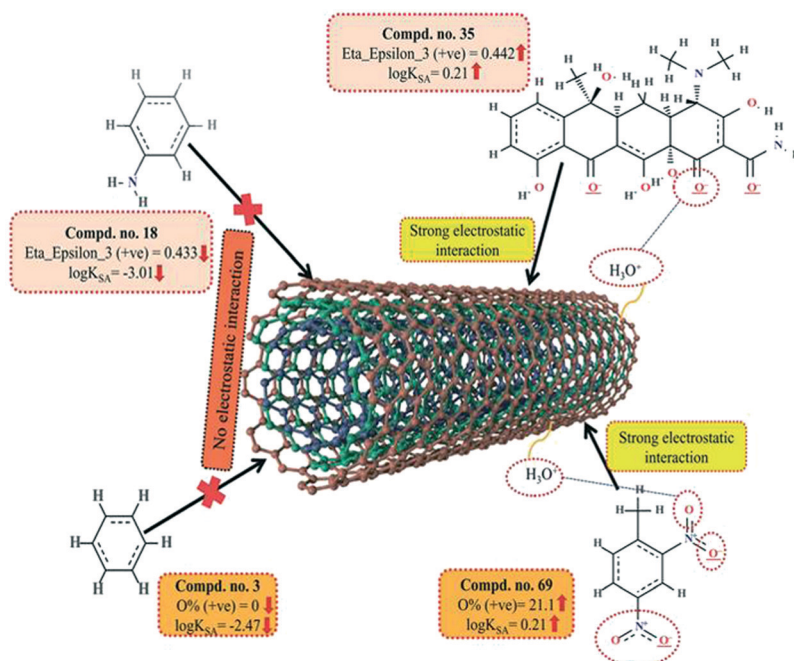


Fig. 11 Mechanistic interpretation of the descriptors related to the electrostatic interactions between organic pollutants and MWCNTs (dataset 2).

$$\text{Model P1. } \log C_{\max} = -1.379 + 1.379 \times \text{VPC-6} - 0.949 \\ \times \text{SpMin3_Bhe} + 0.659 \times \text{minsssN} \\ - 0.375 \times \text{SpMin6_Bhi}$$

$$n_{\text{training}} = 22, R^2 = 0.830, R_{\text{adj}}^2 = 0.810, S = 0.372, F = 29.34,$$

$$\text{PRESS} = 5.164, Q^2 = 0.775, r_{\text{m(LOO)}}^2 = 0.689,$$

$$\Delta r_{\text{m(LOO)}}^2 = 0.115, \text{MAE} = \text{Good},$$

$$n_{\text{test}} = 7, Q_{F1}^2 = 0.945, Q_{F2}^2 = 0.938, r_{\text{m(test)}}^2 = 0.909, \Delta r_{\text{m(test)}}^2 = 0.048,$$

$$\text{CCC} = 0.991, \text{MAE} = \text{Good}$$

The most significant descriptor, minsssN, indicates the minimum atom type E-state >N-. The E-state variable encodes the intrinsic electronic state of each atom present in the molecular graph. The intrinsic electronic state of the atom is changed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. Atoms that possess π and lone pairs of electrons or are terminal atoms possess higher positive values for the E-state index. Atoms that do not have π and lone pairs of electrons and are present at the interior part of a molecule possess lower E-state values. An increase in the minsssN value would indicate the higher electronegativity of the organic solvents, which is beneficial for the dispersibility of SWNTs. The positive regression coefficient of this descriptor indicates that nitrogen atoms connected to other heavy atoms play an important role in influencing the dispersibility of SWNTs in different organic solvents. The numerical values of this descriptor are directly proportional to the dispersibility of SWNTs, suggesting that the dispersibility index of the SWNTs will increase with increasing the number of such fragments as evidenced by the compounds 1 (1,3-dimethyltetrahydro-2(1H)-pyrimidinone), 2 (1-butylpyrrolidin-2-

one) and 5 (3-(2-oxo-1-pyrrolidinyl)propanenitrile). On the other hand, the absence of such fragments in different organic solvents decreases the dispersibility index of SWCNTs as shown in compounds 24 (cyclohexanone), 27 (formamide) and 28 (benzyl alcohol). Thus, from this descriptor, it can be suggested that the dispersibility of CNTs may be enhanced through electrostatic interactions.

The second highest significant descriptor, *SpMin3_Bhe*, is defined as the smallest absolute eigenvalue of Burden modified matrix-n3/weighted by the relative Sanderson electronegativities.⁷⁸ The negative contribution shown by *SpMin3_Bhe* indicates that the dispersibility index of SWCNTs in various solvents can be increased by decreasing the numerical value of *SpMin3_Bhe* as shown in compounds 9 (dimethylimidazolidinone), 10 (dimethyl acetamide) and 16 (acrylic acid). On the other hand, the dispersibility of SWCNTs can be decreased by increasing the numerical value of *SpMin3_Bhe* as shown in compounds 22 (benzyl benzoate) and 26 (triethyleneglycol). The *SpMin3_Bhe* descriptor weighted by the relative Sanderson electronegativity suggests that the electronegativity of the solvents and polar interactions with CNTs play an important role in the dispersibility of the SWCNTs. It can be concluded that polar interactions can have an optimum value. Thus, polar solvents with low donor number are preferred for the dispersibility of the CNTs or it would be better to state that solvents with medium polarity are satisfactory.

The third highest significant descriptor, VPC-6, is a type of topological descriptor, which indicates the chi valance path cluster of order 6. This descriptor differentiates the molecules according to their size, degree of branching, flexibility and overall shape. Chi cluster descriptor (VPC-6) is an indicator of the *n*th degree of branching and thus implicates the effect of substitution in a molecule. The organic solvent molecules that are relatively compact have higher values of this descriptor,⁷⁹ suggesting that a small sized molecule with compactness is most probably a better solvent for SWCNTs. It has a positive contribution toward the dispersibility index of SWCNTs in different organic solvents. This indicates that the degree of branching of organic solvents increases the dispersibility index of SWCNTs as shown in compounds 1 (1,3-dimethyltetrahydro-2(1H)-pyrimidinone), 3 (1-benzylpyrrolidin-2-one), and 9 (dimethyl-imidazolidinone), and *vice versa* in case of compounds 10 (dimethyl acetamide), 16 (acrylic acid) and 17 (2,2'-thiodiethanol).

The least significant descriptor, SpMin6_Bhi indicates the smallest absolute eigenvalue of Burden modified matrix - n6/weighted by the relative first ionization potential.

A modified Burden matrix Q is defined as follows:

$$[Q]_{ij} = Z_i + 0.1\delta_i + 0.01 \times n_i^\pi \text{ and } [Q]_{ij} = 0.4/d_{ij}$$

where, Z_i depicts the atomic number of the *i*th atom, d_i depicts the number of non-hydrogen neighbors of the *i*th atom (*i.e.*, the vertex degree), n_i^π depicts the number of π electrons, and d_{ij} depicts the topological distance between the *i*th and

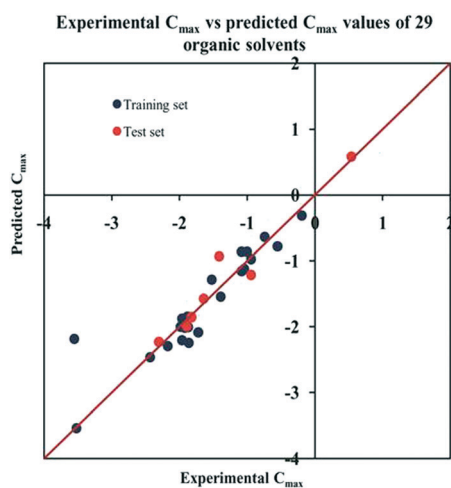


Fig. 12 The scatter plot of the observed and the predicted dispersibility index of SWCNTs ($\log C_{\max}$) of the developed PLS model (model P1).

*j*th atoms.⁷⁸ A larger ionization potential of a molecule suggests that higher energy is required to convert the molecule into cationic form, whereas a smaller ionization potential can easily convert the molecule into cationic form, which helps in the easy interaction of the cationic form of the molecule to the π -system of the carbon nanotube through π -cationic interactions. This descriptor is inversely proportional to the dispersibility of SWNTs, suggesting that with increasing the ionization potential, the dispersibility index of the SWNTs decreases as evidenced by compounds 27 (formamide), 16 (acrylic acid), and 9 (dimethyl-imidazolidinone). On the other hand, the dispersibility index of organic solvents increases in the case of compounds 2 (1-butylpyrrolidin-2-one) and 5 [3-(2-oxo-1-pyrrolidinyl)propanenitrile]. The effects of the contributed descriptors on the dispersibility of SWCNTs in diverse organic solvents are summarized graphically in Fig. 13.

4. Overview and conclusions

MLR and PLS regression-based strategies were employed to develop QSPR models of organic pollutants (datasets 1 & 2) and organic solvents (dataset 3). Multiple endpoints related to CNTs (adsorption coefficient, adsorption coefficient related to specific surface area of MWCNTs and dispersibility index) were used to explore the key structural features that influence the adsorption and dispersibility of the investigated molecules towards MWCNTs and SWCNTs, respectively. The models were developed using 2D descriptors only. Prior to the development of the final models, different strategies for variable selection were performed to extract the most significant descriptors for the generation of the final MLR (5

models for both datasets 1 and 2) and PLS (a single model for dataset 3) models. Extensive validation of the developed models was performed, which showed good predictability and robustness. The QSPR models were developed in compliance with the OCED principles. We also used the “Intelligent consensus predictor” tool to explore whether the quality of the predictions of test set compounds could be enhanced through an “intelligent” selection of multiple MLR models (in the case of datasets 1 and 2). The results showed that based on the MAE-based criteria, the consensus predictions of multiple MLR models are better than the results obtained from the individual models. In both cases, the winning model was CM3. The insights obtained from the developed MLR models for datasets 1 and 2 are as follows: (i) the descriptors like U_i , F03[O–O], F04[N–O], ETA_BetaP, nOHp, O%, T(N··N), T(O··Cl) and F04[O–Cl] influence the adsorption of organic pollutants either by π - π interactions or by strengthening π - π interactions. (ii) n ArOH, F03[O–O], B03[O–O], nHBint, F04[N–O], Eta_Epsilon_3 and O% descriptors favor the adsorption of organic pollutants through electrostatic interactions. (iii) The organic pollutants adsorbed through hydrogen bonding interactions are indicated by n ArOH, F03[O–O], B03[O–O], nHBint, F04[N–O] and O%. (iv) The descriptors minsCH₃, B06[C–Cl], X0v, VAdjMat, MLOGP2, X2A and X1A are essential for the adsorption of organic pollutants through hydrophobic interactions. These observations were further supported by the following discussion: the organic adsorbates of CNTs were mostly aromatic compounds, confirming that aromatic compounds have a better interaction with CNTs than the non-aromatic pollutants, due to their π electron richness and flat conformation. The

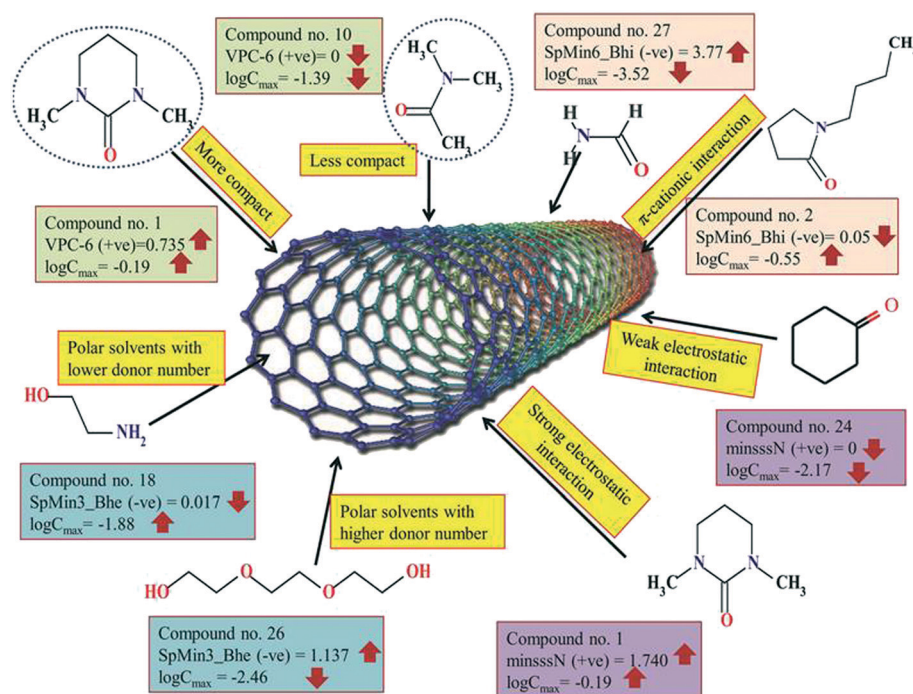


Fig. 13 The effects of the contributed descriptors on the dispersibility of SWCNTs in diverse organic solvents.

systematic understanding of aromatic contaminants is therefore critical since aromaticity plays an important role in adsorption. Several studies have suggested that π - π interactions are crucial for the adsorption of organic compounds to CNTs,^{71,80,81} which in turn depends on the size and shape of the molecules, due to the curvature of the CNTs and its substituents. The π -system of the organic pollutants interacts with the π -system of the CNTs through π - π interactions and the interactions increase with the number of aromatic rings in the adsorbates.^{39,82} Both electron withdrawing groups (e.g. $-\text{NO}_2$ and $-\text{Cl}$) and electron donating groups (e.g. $-\text{NH}_2$, $-\text{OH}$) strengthen the π - π interactions between the pollutants and MWCNTs^{73,74} by acting as π -electron acceptors and π -electron donors, respectively. The hydroxyl group was investigated as an electron donating substituent on adsorptive interactions among pollutants and MWCNTs, since the hydroxyls, by dissociating to $-\text{O}^-$ (which has stronger electron donating ability), strengthen the n - π electron donor-acceptor (EDA) mechanism. Compounds with no aromatic ring (no π electrons) interact through hydrophobic forces. A study also suggested that CNTs act as strong adsorbents for hydrophobic compounds due to hydrophobic interactions.^{15,16,33,83-85} Hydroxyl groups (phenolic form) can interact through various means, such as (i) hydrophobic interactions (ii) electrostatic interactions (both attraction and repulsion) (iii) hydrogen bonding interactions and (iv) enhancing π - π interactions. As the number of hydroxyl groups (phenolics) in the pollutants increases, the hydrophobicity decreases. Thus, it can be considered as a major factor in the adsorption of phenolics to CNTs. Hydrogen bonding can also be a major interaction between hydroxyl-containing pollutants and substituted carbon nanotubes.^{86,87} Hydroxyl and amino group interactions can be related to the electronic features. In one experiment, it was observed that 1-naphthylamine has better adsorption to treated CNTs than the untreated CNTs, and there was an additional observation that although both 2,4-dichlorophenol and 2-naphthol contain an $-\text{OH}$ group, the adsorption of 2-naphthol was more significant with variation in the functionality of CNTs.⁸⁸ This indicates that when the adsorbates possess electronic properties, the functionality of nanotubes helps with the improvement of adsorption.⁸⁸ Chen *et al.*⁸⁹ reported that nitro group containing pollutants show stronger adsorption than non-polar aromatics. This indicates that along with hydrophobic interactions, there is some other essential interaction that controls the adsorption, which is comparable to the π -electron polarizability that is related to aromatic compounds and electron donating as well as accepting properties, similar to compounds having more than two nitro groups. Nitroaromatic compounds, besides being polar in nature, have electron accepting capacity when interacting with adsorbents having high electron polarizability properties and also have high electron conjugation with the π -electrons of CNTs. Thus, the higher affinity of nitro aromatic compounds as compared to other pollutants is due to π - π electron donor-acceptor interactions; since nitrogen is a strong electron-withdrawing atom, it acts as a π -acceptor and

carbon nanotubes act as the π -donor.⁹⁰⁻⁹³ Hydrogen bonding is also possible between nitro groups of the pollutants, which act as H-acceptors and functional group-substituted carbon nanotubes. The presence of two chlorine atoms causes the electron inductive effect, which may cause a reduction in the electron density of the aromatic ring attached to it, as suggested by Sulaymon and Ahmed *et al.*;⁹⁴ the electron donating effect of the hydroxyl atom attached to the aromatic ring compensates for this by dissociating into the stronger electron donor like $-\text{O}^-$ (oxygen). We can, therefore, conclude that the adsorption of the organic pollutants to the CNTs can be enhanced by the following: a greater number of aromatic rings, high unsaturation or electron richness of the molecule, the presence of polar groups substituted on the aromatic ring, the presence of two oxygen atoms at a topological distance of 3, the presence of nitrogen and oxygen atoms at the topological distance of 4, the size of the molecules, and the hydrophobic surface of the molecules. On the other hand, the presence of carbon and oxygen atoms at a topological distance of 1, aliphatic primary alcohols, the presence of two chlorine atoms at topological distance 5 and the presence of oxygen and chlorine atoms at topological distance 4 may be detrimental and can retard the adsorption of organic pollutants. From the insights obtained from the PLS model for dataset 3, we have interpreted that the organic solvents bearing the $>\text{N}$ - fragment, polar solvents with low donor number, compact molecules and lower ionization potential may be better solvents to enhance the dispersibility of SWCNTs. Dispersibility is directly correlated to the adsorption properties of molecules to CNTs. This PLS model and contributed descriptors can help with the understanding of the mechanism of the dispersion process and predict organic solvents that improve the dispersibility of SWCNTs and may overcome the drawbacks of SWCNTs. This work may, therefore, be helpful in the removal of the harmful and toxic contaminants/disposal of the by-products from the various industries, making it possible to achieve a pollution-free environment.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Financial assistance from the AICTE, New Delhi in the form of a fellowship to JR and SG is thankfully acknowledged. PKO thanks the UGC, New Delhi for financial assistance in the form of a fellowship (Letter number and date: F./PDFSS-2015-17-WES-11996; dated: 06/04/2016). KR thanks CSIR, New Delhi for financial assistance under a Major Research project (CSIR Project No. 01(2895)/17/EMR-II).

References

- 1 U. K. Garg, M. P. Kaur, V. K. Garg and D. Sud, Removal of hexavalent Cr from aqueous 19 solutions by agricultural waste biomass, *J. Hazard. Mater.*, 2007, **140**, 60-68.

- 2 J. M. Randall, E. Hautala and A. C. Waiss Jr, Removal and recycling of heavy metal ions from mining and industrial waste streams with agricultural by-products, in *Proceedings of the fourth mineral waste utilization symposium*, Chicago, 1974.
- 3 D. J. Ferner, Toxicity, heavy metals, *Med. J.*, 2001, 2(5), 1.
- 4 Y. Lu, S. Song, R. Wang, Z. Liu, J. Meng, A. J. Sweetman, A. Jenkins, R. C. Ferrier, H. Li, W. Luo and T. Wang, Impacts of soil and water pollution on food safety and health risks in China, *Environ. Int.*, 2015, 77, 5–15.
- 5 L. B. Franklin, *Wastewater engineering: treatment, disposal and reuse*, McGraw Hill, New York, 1991.
- 6 R. L. Droste, *Theory and practice of water and wastewater treatment*, Wiley, New York, 1997.
- 7 R. N. Goyal, V. K. Gupta, A. Sangal and N. Bachheti, Voltammetric determination of uric acid at a fullerene-C60-modified glassy carbon electrode, *Electroanalysis*, 2005, 17(24), 2217–2223.
- 8 R. N. Goyal, V. K. Gupta and N. Bachheti, Voltammetric determination of adenosine and guanosine using fullerene-C60-modified glassy carbon electrode, *Talanta*, 2007, 71(3), 1110–1117.
- 9 R. N. Goyal, V. K. Gupta and N. Bachheti, Fullerene-C60-modified electrode as a sensitive voltammetric sensor for detection of nandrolone, *Anal. Chim. Acta*, 2007, 597, 82–89.
- 10 R. N. Goyal, V. K. Gupta, N. Bachheti and R. A. Sharma, Electrochemical sensor for the determination of dopamine in presence of high concentration of ascorbic acid using a fullerene-C60 coated gold electrode, *Electroanalysis*, 2008, 20, 757–764.
- 11 R. N. Goyal, M. Oyama, V. K. Gupta, S. P. Singh and S. Chatterjee, Sensors for 5-hydroxytryptamine and 5-hydroxyindole acetic acid based on nanomaterial modified electrodes, *Sens. Actuators, B*, 2008, 134, 816–821.
- 12 R. N. Goyal, V. K. Gupta and S. Chatterjee, Fullerene-C60-modified edge plane pyrolytic graphite electrode for the determination of dexamethasone in pharmaceutical formulations and human biological fluids, *Biosens. Bioelectron.*, 2009, 24, 1649–1654.
- 13 D. Z. John, *Handbook of drinking water quality: standards and controls*, Van Nostrand Reinhold, New York, 1990.
- 14 E. A. Laws, *Aquatic pollution: an introductory text*, Wiley, New York, 3rd edn, 2000.
- 15 K. Yang, L. Z. Zhu and B. S. Xing, Adsorption of polycyclic aromatic hydrocarbons by carbon nanomaterials, *Environ. Sci. Technol.*, 2006, 40, 1855–1861.
- 16 K. Yang, X. Wang, L. Zhu and B. Xing, Competitive sorption of pyrene, phenanthrene, and naphthalene on multiwalled carbon nanotubes, *Environ. Sci. Technol.*, 2006, 40, 5804–5810.
- 17 Y. H. Li, Z. Di, J. Ding, D. Wu, Z. Luan and Y. Zhu, Adsorption thermodynamic, kinetic and desorption studies of Pb²⁺ on carbon nanotubes, *Water Res.*, 2005, 39(4), 605–609.
- 18 H. M. Al-Saidi, M. A. Abdel-Fadeel, A. Z. El-Sonbati and A. A. El-Bindary, Multi-walled carbon nanotubes as an adsorbent material for the solid phase extraction of bismuth from aqueous media: kinetic and thermodynamic studies and analytical applications, *J. Mol. Liq.*, 2016, 216, 693–698.
- 19 S. Kumar, G. Bhanjana, N. Dilbaghi and A. Umar, Multi walled carbon nanotubes as sorbent for removal of crystal violet, *J. Nanosci. Nanotechnol.*, 2014, 14, 7054–7059.
- 20 S. Mosayebidorcheh and M. Hatami, Heat transfer analysis in carbon nanotube-water between rotating disks under thermal radiation conditions, *J. Mol. Liq.*, 2017, 240, 258–267.
- 21 N. Nakashima, Soluble carbon nanotubes: Fundamental and applications, *Int. J. Nanosci.*, 2005, 4, 119–137.
- 22 D. A. Britz and A. N. Khlobystov, Noncovalent interactions of molecules with single walled carbon nanotubes, *Chem. Soc. Rev.*, 2006, 35(7), 637–659.
- 23 L. A. Girifalco, M. Hodak and R. S. Lee, Carbon nanotubes, buckyballs, ropes, and a universal graphitic potential, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2000, 62(19), 13104.
- 24 E. Hammel, X. Tang, M. Trampert, T. Schmitt, K. Mauthner, A. Eder and P. Pötschke, Carbon nanofibers for composite applications, *Carbon*, 2004, 42(5–6), 1153–1158.
- 25 T. Liu, I. Y. Phang, L. Shen, S. Y. Chow and W. D. Zhang, Morphology and mechanical properties of multiwalled carbon nanotubes reinforced nylon-6 composites, *Macromolecules*, 2004, 37(19), 7214–7222.
- 26 Y. S. Song and J. R. Youn, Influence of dispersion states of carbon nanotubes on physical properties of epoxy nanocomposites, *Carbon*, 2005, 43(7), 1378–1385.
- 27 K. E. Geckeler and T. Premkumar, Carbon nanotubes: are they dispersed or dissolved in liquids?, *Nanoscale Res. Lett.*, 2011, 6(1), 136.
- 28 A. Abbas, A. M. Al-Amer, T. Laoui, M. J. Al-Marri, M. S. Nasser, M. Khraisheh and M. A. Atieh, Heavy metal removal from aqueous solution by advanced carbon nanotubes: critical review of adsorption applications, *Sep. Purif. Technol.*, 2016, 157, 141–161.
- 29 O. V. Kharissova, B. I. Kharisov and E. G. de Casas Ortiz, Dispersion of carbon nanotubes in water and non-aqueous solvents, *RSC Adv.*, 2013, 3(47), 24812–24852.
- 30 H. Hyung, J. D. Fortner, J. B. Hughes and J. H. Kim, Natural organic matter stabilizes carbon nanotubes in the aqueous phase, *Environ. Sci. Technol.*, 2007, 41, 179–184.
- 31 R. Q. Long and R. T. Yang, Carbon nanotubes as superior sorbent for dioxin removal, *J. Am. Chem. Soc.*, 2001, 123, 2058–2059.
- 32 G. P. Rao, C. Lu and F. Su, Sorption of divalent heavy metal ions from aqueous solution by carbon nanotubes: a review, *Sep. Purif. Technol.*, 2007, 58, 224–231.
- 33 J. Hilding, E. A. Grulke, S. B. Sinnott, D. Qian, R. Andrews and M. Jagtoyen, Sorption of butane on carbon multiwall nanotubes at room temperature, *Langmuir*, 2001, 17, 7540–7544.
- 34 C. S. Lu, Y. L. Chung and K. F. Chang, Adsorption of trihalomethanes from water with carbon nanotubes, *Water Res.*, 2005, 39, 1183–1189.

- 35 C. J. M. Chin, L. C. Shih, H. J. Tsai and T. K. Liu, Adsorption of o-xylene and p-xylene from water by SWCNTs, *Carbon*, 2007, 45, 1254–1260.
- 36 Q. Liao, J. Sun and L. Gao, Adsorption of chlorophenols by multiwalled carbon nanotubes treated with HNO₃ and NH₃, *Carbon*, 2008, 46, 553–555.
- 37 X. J. Peng, Y. H. Li, Z. K. Luan, Z. C. Di, H. Y. Wang, B. H. Tian and Z. P. Jia, Adsorption of 1, 2-dichlorobenzene from water to carbon nanotubes, *Chem. Phys. Lett.*, 2003, 376, 154–158.
- 38 Q. Liao, J. Sun and L. Gao, The adsorption of resorcinol from water using multi-walled carbon nanotubes, *Colloids Surf.*, 2008, 312, 160–165.
- 39 S. Gotovac, H. Honda, Y. Hattori, K. Takahashi, H. Kanoh and K. Kaneko, Effect of nanoscale curvature of single-walled carbon nanotubes on adsorption of polycyclic aromatic hydrocarbons, *Nano Lett.*, 2007, 7, 583–587.
- 40 D. C. Luehrs, J. P. Hickey, P. E. Nilsen, K. A. Godbole and T. N. Rogers, Linear solvation energy relationship of the limiting partition coefficient of organic solutes between water and activated carbon, *Environ. Sci. Technol.*, 1996, 30, 143–152.
- 41 O. G. Apul, Q. Wang, T. Shao, J. R. Rieck and T. Karanfi, Predictive model development for adsorption of aromatic contaminants by multi-walled carbon nanotubes, *Environ. Sci. Technol.*, 2012, 47, 2295–2303.
- 42 M. Rahimi-Nasrabadi, R. Akhoondi, S. M. Pourmortazavi and F. Ahmadi, Predicting adsorption of aromatic compounds by carbon nanotubes based on quantitative structure property relationship principles, *J. Mol. Struct.*, 2015, 1099, 510–515.
- 43 X. R. Xia, N. A. Monteiro-Riviere and J. E. Riviere, An index for characterization of nanomaterials in biological systems, *Nat. Nanotechnol.*, 2010, 5, 671–675.
- 44 V. Chayawan, Quantum-mechanical parameters for the risk assessment of multiwalled carbon-nanotubes: a study using adsorption of probe compounds and its application to biomolecules, *Environ. Pollut.*, 2016, 218, 615–624.
- 45 O. G. Apul, P. Xuan, F. Luo and T. Karanfil, Development of a 3D QSPR model for adsorption of aromatic compounds by carbon nanotubes: comparison of multiple linear regression, artificial neural network and support vector machine, *RSC Adv.*, 2013, 3, 23924–23934.
- 46 O. G. Apul, Y. Zhou and T. Karanfil, Mechanisms and modeling of halogenated aliphatic contaminant adsorption by carbon nanotubes, *J. Hazard. Mater.*, 2015, 295, 138–144.
- 47 Z. Hassanzadeh, M. Kompany-Zareh, R. Ghavami, S. Gholami and A. Malek-Khatabi, Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface, *J. Mol. Struct.*, 2015, 1098, 191–198.
- 48 H. Yilmaz, B. Rasulev and J. Leszczynski, Modeling the dispersibility of single walled carbon nanotubes in organic solvents by quantitative structure-activity relationship approach, *Nanomaterials*, 2015, 5, 778–791.
- 49 M. Salahinejad and E. Zolfonoun, QSAR studies of the dispersion of SWNTs in different organic solvents, *J. Nanopart. Res.*, 2013, 15, 2028.
- 50 M. Rofouei, M. Salahinejad and J. B. Ghasemi, An alignment independent 3D-QSAR modeling of dispersibility of single-walled carbon nanotubes in different organic solvents, *Fullerenes, Nanotubes, Carbon Nanostruct.*, 2014, 22, 605–617.
- 51 A. Heidari and M. H. Fatemi, Hybrid docking-Nano-QSPR: an alternative approach for prediction of chemicals adsorption on nanoparticles, *NANO*, 2016, 11, 1650078.
- 52 S. D. Bergin, Z. Sun, D. Rickard, P. V. Streich, J. P. Hamilton and J. N. Coleman, Multicomponent solubility parameters for single-walled carbon, nanotube–solvent mixtures, *ACS Nano*, 2009, 3, 2340–2350.
- 53 <http://www.chemaxon.com>.
- 54 http://www.taletе.mi.it/products/dragon_description.htm.
- 55 <http://www.yapcwsoft.com/dd/padeldescriptor>.
- 56 S. Das, P. K. Ojha and K. Roy, Multilayered variable selection in QSPR: a case study of modeling melting point of bromide ionic liquids, *Int. J. Quant. Struct.-Prop. Relat.*, 2017, 2(1), 106–124.
- 57 P. K. Ojha and K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, *Chemom. Intell. Lab. Syst.*, 2011, 109(2), 146–161.
- 58 http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab.
- 59 K. Roy, R. N. Das, P. Ambure and R. B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemom. Intell. Lab. Syst.*, 2016, 152, 18–33.
- 60 K. Roy, P. Ambure, S. Kar and P. K. Ojha, Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models?, *J. Chemom.*, 2018, 32(4), 2992.
- 61 R. B. Darlington, in *Regression and linear models*, New York, McGraw-Hill, 1990.
- 62 D. Rogers and A. J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships, *J. Chem. Inf. Comput. Sci.*, 1994, 34, 854–866.
- 63 P. K. Ojha, I. Mira, R. N. Das and K. Roy, Further exploring r_m^2 metrics for validation of QSPR models, *Chemom. Intell. Lab. Syst.*, 2011, 107(1), 194–205.
- 64 I. Lawrence and K. Lin, Assay validation using the concordance correlation coefficient, *Biometrics*, 1992, 599–604.
- 65 N. Chirico and P. Gramatica, Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.*, 2011, 51(9), 2320–2335.
- 66 K. Roy, S. Kar and P. Ambure, On a simple approach for determining applicability domain of QSAR models, *Chemom. Intell. Lab. Syst.*, 2015, 145, 22–29.
- 67 S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, 58, 109–130.
- 68 UMETRICS, *UMETRICS SIMCA-P 10.0*, Umea, Sweden, 2002, info@umetrics.com, www.umetrics.com.
- 69 <http://www.miniTab.com/en-US/default.aspx>.
- 70 SPSS is statistical software of SPSS Inc., USA, 1999.

- 71 Y. Zhang, S. L. Yuan, W. W. Zhou, J. J. Xu and Y. Li, Spectroscopic evidence and molecular simulation investigation of the pi-pi interaction between pyrene molecules and carbon nanotubes, *J. Nanosci. Nanotechnol.*, 2007, 7, 2366–2375.
- 72 L. H. Hall and L. B. Kier, Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information, *J. Chem. Inf. Comput. Sci.*, 1995, 35(6), 1039–1045.
- 73 L. M. Woods, S. C. Bădescu and T. L. Reinecke, Adsorption of simple benzene derivatives on carbon nanotubes, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, 75(1–9), 155415.
- 74 A. Star, T. R. Han, J. C. P. Gabriel, K. Bradley and G. Gruner, Interaction of aromatic compounds with carbon nanotubes: correlation to the Hammett parameter of the substituent and measured carbon nanotube FET response, *Nano Lett.*, 2003, 3, 1421–1423.
- 75 I. Moriguchi, S. Hirono, I. Nakagome and H. Hirano, Comparison of reliability of log P values for drugs calculated by several methods, *Chem. Pharm. Bull.*, 1994, 42(4), 976–978.
- 76 P. K. Ojha and K. Roy, Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules, *Food Chem. Toxicol.*, 2017, 112, 551–562.
- 77 L. B. Kier and L. H. Hall, The meaning of molecular connectivity: A bimolecular accessibility model, *Croat. Chem. Acta*, 2002, 75(2), 371–382.
- 78 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*, John Wiley & Sons, vol. 41, 2009.
- 79 K. P. Singh and S. Gupta, Nano-QSAR modeling for predicting biological activity of diverse nanomaterials, *RSC Adv.*, 2014, 4(26), 13215–13230.
- 80 F. S. Su and C. S. Lu, Adsorption kinetics, thermodynamics and desorption of natural dissolved organic matter by multiwalled carbon nanotubes, *J. Environ. Sci. Health, Part A: Toxic/Hazard. Subst. Environ. Eng.*, 2007, 42, 1543–1552.
- 81 Z. W. Wang, C. L. Liu, Z. G. Liu, H. Xiang, Z. Li and Q. H. Gong, π - π interaction enhancement on the ultrafast third-order optical nonlinearity of carbon nanotubes/polymer composites, *Chem. Phys. Lett.*, 2005, 407, 35–39.
- 82 F. Tournus, S. Latil, M. I. Heggie and J. C. Charlier, π -stacking interaction between carbon nanotubes and organic molecules, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2005, 72(1–5), 75431.
- 83 S. B. Fagan, A. G. S. Filho, J. O. G. Lima, J. M. Filho, O. P. Ferreira, I. O. Mazali, O. L. Alves and M. S. Dresselhaus, 1, 2-Dichlorobenzene interacting with carbon nanotubes, *Nano Lett.*, 2004, 4, 1285–1288.
- 84 S. Gotovac, Y. Hattori, D. Noguchi, J. Miyamoto, M. Kanamaru, S. Utsumi, H. Kanoh and K. Kanek, Phenanthrene adsorption from solution on single wall carbon nanotubes, *J. Phys. Chem. B*, 2006, 110, 16219–16224.
- 85 J. Zhao and J. Lu, Noncovalent functionalization of carbon nanotubes by aromatic organic molecules, *Appl. Phys. Lett.*, 2003, 82, 3746–3748.
- 86 X. J. Li, W. Chen, Q. W. Zhan, L. M. Dai, L. Sowards, M. Pender and R. R. Naik, Direct measurements of interactions between polypeptides and carbon nanotubes, *J. Phys. Chem. B*, 2006, 110, 12621–12625.
- 87 A. M. Li, Q. X. Zhang, H. S. Wu, Z. C. Zhai, F. Q. Liu, Z. H. Fei, C. Long, Z. L. Zhu and J. L. Chen, A new amine-modified hypercrosslinked polymeric adsorbent for removing phenolics compounds from aqueous solutions, *Adsorpt. Sci. Technol.*, 2004, 22, 807–819.
- 88 W. Chen, L. Duan, L. Wang and D. Zhu, Adsorption of hydroxyl-and amino-substituted aromatics to carbon nanotubes, *Environ. Sci. Technol.*, 2008, 42(18), 6862–6868.
- 89 W. Chen, L. Duan and D. Zhu, Adsorption of polar and nonpolar organic chemicals to carbon nanotubes, *Environ. Sci. Technol.*, 2007, 41(24), 8295–8300.
- 90 L. R. Radovic, C. Moreno-Castilla and J. Rivera-Utrilla, Carbon materials as adsorbents in aqueous solutions, *Chem. Phys. Carbon*, 2001, 227–406.
- 91 C. A. Hunter and J. K. M. Sanders, The nature of π - π interactions, *J. Am. Chem. Soc.*, 1990, 112, 5525–5534.
- 92 J. C. Ma and D. A. Dougherty, The cation- π interaction, *Chem. Rev.*, 1997, 97, 1303–1324.
- 93 C. A. Hunter, K. R. Lawson, J. Perkins and C. J. Urch, Aromatic interactions, *J. Chem. Soc., Perkin Trans. 1*, 2001, 651–669.
- 94 A. H. Sulaymon and K. W. Ahmed, Competitive adsorption of furfural and phenolic compounds onto activated carbon in fixed bed column, *Environ. Sci. Technol.*, 2008, 42, 392–397.



Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs

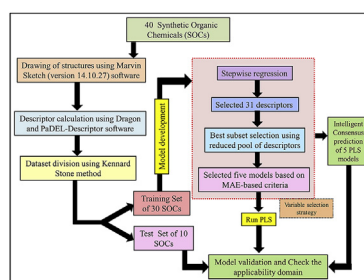
Sulekha Ghosh, Probir Kumar Ojha **, Kunal Roy *

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700 032, India

HIGHLIGHTS

- We develop consensus QSPR models for diverse synthetic organic chemicals having defined adsorption affinity for SWCNTs.
- Only simple 2D descriptors with definite physicochemical meanings have been used.
- The models have been validated extensively with internal and external validation metrics.
- The information obtained from the developed models should be useful for the management of environmental pollution.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 14 February 2019

Received in revised form

12 April 2019

Accepted 15 April 2019

Available online 23 April 2019

Handling Editor: Y Yeomin Yoon

Keywords:

Hazardous

Synthetic organic chemicals

SWCNTs

Adsorption coefficient

QSPR

ABSTRACT

In order to understand the physicochemical properties as well as the mechanisms behind adsorption of hazardous synthetic organic chemicals (SOCs) onto single walled carbon nanotubes (SWCNTs), we have developed partial least squares (PLS)-regression based QSPR models using a diverse set of 40 hazardous SOC having defined adsorption coefficient ($\log K$). The models were extensively validated using different validation parameters in order to assure the robustness and predictivity of the models. We have also checked the consensus predictivity of all the individual models using “Intelligent consensus predictor” tool for possible enhancement of the quality of predictions for test set compounds. The consensus predictivity of the test set compounds were found to be better than the individual models based on not only the MAE based criteria ($MAE_{(95\%)} = \text{Good}$) but also some other validation parameters ($Q^2_{F1} = 0.938$, $Q^2_{F2} = 0.937$). The contributing descriptors obtained from the QSPR models suggested that the hazardous SOC may get adsorbed onto the SWCNTs through hydrophobic interaction as well as hydrogen bonding interactions and electrostatic interaction to the functionally modified SWCNTs. Thus, the developed models may provide knowledge to scientists to increase the efficient application of SWCNTs as a special adsorbent, which may be useful for the management of environmental pollution.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

The rapid industrial growth leads to an increase in the demand for new inventions and technologies for the benefit of human beings. New chemicals have been introduced for various purposes, which can, however, be a major threat for humans and animals

* Corresponding author.

** Corresponding author.

E-mail addresses: probirojha@yahoo.co.in (P.K. Ojha), kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).

(Latkar and Chakrabarti, 1994). The use of herbicides, for example, has increased during the last two decades due to the rejuvenation of agriculture. It was reported that 2.5 million ton pesticides were in use worldwide yearly, and the amount is increasing day by day (Pimentel, 1995; Tariq et al., 2007; Carter, 2000). Endocrine disrupting chemicals (EDCs) act like natural hormones and hamper the distribution, as well as metabolic process of natural hormones. EDCs (e.g., ethinyl estradiol) are harmful for the reproductive system of animals and humans (Snyder et al., 2003). Effluents from hospitals or radiological clinics have shown high concentration of antibiotics like sulfamethoxazole and lincomycin, and contrast medium (ipromide), which are responsible for the production of antibiotic resistance bacteria and genes in the aquatic environment (Rand-Weaver et al., 2013; Michael et al., 2013). Hence, removal of antibiotics as well as pharmaceuticals and contrast medium from water is essential to get purified water. Perfluorinated compounds (PFCs) are synthesized from 1960s, and they have been used as surfactants, fire retardants, paints, adhesives, waxes and polishes (Moody and Field, 2000). Perfluorinated sulfonates (PFOS) are among the most identified perfluorinated compounds because of their high concentration, global sharing, environmental assiduity and bioaccumulation. They are highly soluble in water and, therefore, can easily be transported in water causing water pollution. PFOS are not easily removed from water sources by conventional water purification techniques due to their exceptional stability. Polycyclic aromatic hydrocarbons (PAHs) (naphthalene, phenanthrene, *p*-nitrophenol) are highly toxic substances and are hydrophobic in nature (May et al., 1978; Walters and Luthy, 1984). The principal sources of PAHs are mainly various anthropogenic sources derived from combustion of coal and oil (Nielsen, 1996) exhaust from motor vehicles (Harrison et al., 1996) and effluents from petrochemical plants (Domeno and Nerin, 2003). They are not degraded easily and not efficiently removed from environment by simple physicochemical methods. Chlorophenols have been identified as one of the principle pollutants by the US Environmental Protection Agency (USEPA) (Ahmed Adam and Al-Dujaili, 2013). Chlorophenols are mainly used for the production of pesticides, dyes and biocides. They are carcinogenic and toxic in nature, and thus considered as one of the major sources of environmental pollutants (Okolo et al., 2000). Chlorine is used to disinfect drinking water. Chlorophenols may also be produced during the disinfection of water (Ahmaruzzaman, 2008). The presence of chlorophenol in drinking water causes unpleasant taste and odor even at very low concentration (less than 0.1 mg/L) (Suffet et al., 1999). Huge quantities of *N*- and *S*-heterocyclic (thiophene, 2-aminopyrimidine, 4,6-diaminopyrimidine) aromatics are used by pharmaceutical, dye and pesticide manufacturing industries. Heterocyclic aromatics are common environmental pollutants. They also cannot be removed easily by simple water treatment process, and they do not degrade easily (Song, 2003; Padoley et al., 2008; Bi et al., 2007). Chlorobenzenes (1,2,4,5 tetrachlorobenzene, 1,2,4-trichlorobenzene, 1,2-dichlorobenzene, chlorobenzene) are basically used as solvents, degreasing agents and chemical intermediates. Their short term exposure in animals causes necrosis, restlessness, tremors and muscle spasms, while long term exposure causes numbness, cyanosis, and hyperesthesia in humans. Dialkyl phthalate esters (DPEs) are often used as plasticizers in polyvinyl chloride, polyvinyl acetates, cellulose and polyurethanes. They are also used in many other fields like nanoplasticizer in products such as photographic films, lubricating oils, paints, insect repellents etc. Due to their outstretched use with global production, DPEs have been detected in water, soil and marine ecosystem (Lin et al., 2003; Mackintosh et al., 2004; Zhu et al., 2006). DPEs have been identified as U.S. EPA priority pollutants (<http://www.epa.gov>).

Recently, nanomaterials are used for pollution management, because they contain high surface area, high adsorption affinity towards the organic contaminants, and they can be modified in several ways to increase their selectivity towards specific target pollutants (Chen et al., 2007). Carbon nanotubes (CNTs) are such type of nanomaterials, which have recently gained special attention from the researchers due to their smaller size, large specific surface area, hollow and layered structure, responsible for their extraordinary adsorption property (Khani and Moradi, 2013; Long and Yang, 2001). CNTs were first discovered in 1991; they show interesting physical and chemical properties. They are successfully used in the field of medical and environmental remediation (Kim et al., 2014b; Singh et al., 2014). CNTs are composed of cylindrical graphite sheets, which show high van der Waals index (Lohmann et al., 2005). The graphite sheets of CNTs consist of benzenoid rings, which have sp^2 -hybridized carbon atom and high polarizability. CNTs are hydrophobic in nature and strongly attached with hydrophobic aromatic pollutants by π - π coupling stacking (Lara et al., 2014). CNTs are generally two types, single layered graphitic cylinder with few nanometer diameter nanotubes, also known as single walled carbon nanotubes (SWCNTs), and nanotubes with 2–30 concentric cylinders and 30–50 nm diameter nanotubes, also known as multiwalled carbon nanotubes (MWCNTs) (Petersen et al., 2011). The four adsorption sites present on CNTs are outermost surface, inner cavities, interstitial channels and grooves (Zhao et al., 2002). Recently, more research has been carried out for adsorption of various synthetic organic chemicals like polyaromatic hydrocarbons (PAHs), hydroxyl-, amino-, or chloro-substituted PAHs, herbicides and endocrine disrupting chemicals onto CNTs (Chen et al., 2007; Wang et al., 2010c). Different adsorption mechanisms for carbon nanotubes described in various literature are hydrophobic interaction, π - π interaction and hydrogen bonding interaction (Pan and Xing, 2008; Yang et al., 2006). Kim et al. (2014a, b) reported that SWCNTs strongly adsorb lincomycin, sulfamethoxazole and ipromide as compared to MWCNTs and activated carbon (Kim et al., 2014a). Various LSER models have been developed for prediction of adsorption of hazardous SOCs (synthetic organic chemicals) on MWCNTs. Recently, Mosayebidorcheh and Hatami (Mosayebidorcheh and Hatami, 2017) and Nakashima (2005) reported LSER models for adsorption of aromatic compounds and halogenated aliphatic compounds onto SWCNTs. Chen et al. reported that SWCNTs possess good adsorption property as compared to MWCNTs due to the molecular sieving effect. For this molecular sieving effect, bulky moieties could not access some of the innermost surfaces of the MWNTs (Chen et al., 2007; Roy et al., 2019).

In the present study, we have developed partial least squares (PLS) regression based quantitative structure-property relationship (QSPR) models using adsorption coefficient data of 40 diverse hazardous synthetic organic chemicals (SOCs) onto SWCNTs. The main objectives of our work are: 1) to develop statistically robust and validated QSPR models of hazardous SOCs using 2D descriptors only in order to identify the significant structural features essential for effective adsorption in SWCNTs; 2) to examine the adsorption behavior of diverse synthetic organic chemicals onto SWCNTs; 3) to give a deep insight to understand the mechanisms and factors that are responsible for hazardous SOCs and SWCNTs/functionalized SWCNTs interactions.

2. Method and materials

2.1. The dataset

A diverse set of 40 hazardous synthetic organic chemicals (SOC)

with defined adsorption coefficient onto SWCNTs reported in the literature (Ding et al., 2016a) were used to develop the QSPR models. The whole data set of 40 synthetic organic chemicals were assembled from 14 published articles containing experimental adsorption coefficient (Kin, L/kg) values. The adsorption coefficient K was calculated by using following formula:

$$K = \frac{q_e}{C_e}$$

where, q_e (mg/kg) is equilibrium concentration on the surface and C_e (mg/L) is the equilibrium concentration in the aqueous phase of SWCNTs. The adsorption coefficient depends on the equilibrium concentration whenever the adsorption isotherm is nonlinear in nature (Zhao et al., 2014). The effect of concentration on K was investigated. The equilibrium concentration on the surface (q_e) could be obtained from isotherm data at $C_e = 0.00002, 0.0002, 0.002, 0.02$ and $0.2 C_s$, (where C_s is the aqueous solubility of the adsorbate). The consequent K values are represented as $K_{0.00002}, K_{0.0002}, K_{0.002}, K_{0.02}$ and $K_{0.2}$ respectively. The endpoint K values were taken in the logarithmic scale for the development of QSPR models. We have used $\log K_{0.002}$ values for the development of QSPR models due to its relatively wide distribution than rest of the $\log K$ values. The data set is depicted in Supplementary section (Table S1).

2.2. Descriptor calculation

All the structures were drawn by using Marvin sketch software (<http://www.chemaxon.com>). The descriptors were calculated using two software tools, Dragon descriptor version 6 and PaDEL-Descriptor (<http://www.yapcwsoft.com/dd/padeldescriptor>) software. Constitutional indices, ring descriptors, connectivity indices, functional group count, atom centered fragments, atom type E-state indices, 2D atom pairs and molecular properties were calculated using Dragon software, while extended topochemical atom (ETA) indices were calculated using PaDEL-Descriptor software.

2.3. Dataset division

Data set division is a very important step for model development process. Through dataset division, we can confirm development of statistically robust models which have a potential to predict the activity of new molecules. In this work, the whole data set was divided by using the “datasetDivisionGUI1.2” (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) software tool. We have employed the Kennard-Stone method for data set division. The selection of objects in the Kennard-Stone algorithm was done in such a manner that they were uniformly distributed throughout the descriptor space of the data set. In this study, 75% compounds were selected for the training set, and the remaining 25% compounds were selected for the test set (i.e., 30 compounds for the training set and 10 compounds for the test set). The training set was used for model development, and the test set was used for model validation purposes.

2.4. Variable selection and model development

Prior to the development of the final models, we have performed data pretreatment to eliminate intercorrelated descriptors. Various variable selection strategies were employed to prepare the descriptor pool. We have excluded the variables with constant and near constant values (standard deviation less than 0.0001), descriptors with at least one missing value, descriptors with all missing values and descriptors with (absolute) pair correlation larger than or equal to 0.95 from the initial pool of descriptors.

Initially, we have run stepwise regression analysis and selected the modeled descriptors. Then we have removed the previously selected descriptors from the initial pool of descriptors and rerun stepwise regression using remaining pool of descriptors. In this manner, we have selected 31 descriptors for the development of final models. Among the best subset equations, we have selected five models based on Mean Absolute Error (MAE) criteria (Roy et al., 2016) along with some other parameters, and then carried out partial least squares (PLS) regression (Wold et al., 2001), in each case, using the selected descriptors. Finally, we have performed “intelligent consensus prediction” of the test set compounds based on the selected five models using intelligent consensus predictor (ICP) tool (Roy et al., 2018) in order to investigate whether prediction quality of the external set compounds was increased or not through an “intelligent” selection. The steps involved for development of the final models are depicted in Fig. 1. Some additional details of model development are given in Supplementary Materials.

2.5. Statistical validation metrics

We have examined the statistical quality of the derived models to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. In this work, we have used various statistical parameters like determination coefficient (R^2), explained variance (R^2_a), variance ratio (F) and standard error of estimate (s). But these parameters are not sufficient to judge the actual quality and predictability of the model. So, we have calculated some other classical statistical metrics like leave-one-out-cross-validated correlation coefficient ($Q^2_{(LOO)}$), R^2_{pred} , Q^2_{F2} , concordance correlation coefficient (CCC) and different r^2_m metrics. Among the above said parameters, Q^2 , $r^2_{m(LOO)}$ and $\Delta r^2_{m(LOO)}$ were used for internal validation while R^2_{pred} , Q^2_{F2} , CCC, $r^2_{m(test)}$ and $\Delta r^2_{m(test)}$ were used for external validation. The threshold values of Q^2 , Q^2_{F2} , R^2_{pred} , $r^2_{m(test)}$, $r^2_{m(LOO)}$ are 0.5 and for CCC, it is 0.750. The maximum limit for $\Delta r^2_{m(test)}$ and $\Delta r^2_{m(LOO)}$ is 0.2. Roy et al. (2012) reported that a single model might be not equally helpful to predict all test set compounds, so we have selected five models. Additionally, we have also validated the PLS models using Y-randomization test (Melagraki and Afantitis, 2013) through randomly shuffling (100 permutations) the dependent variable vector ($\log K$) using SIMCA-P software (Umetrics, 2002) to ensure that the model was not obtained by chance. Here, keeping the descriptor matrix intact, the dependent variable vector (Y) is randomly permuted, followed by a PLS run and a new predictive model is developed using the original independent variable matrix. After several repetitions, the new predictive model generates a fresh set of R^2 and Q^2 values. These fresh R^2 and Q^2 values are plotted against the correlation coefficient between the original Y-values and the permuted Y-values. It is expected that the values of fresh R^2 and Q^2 should be low. The PLS model is considered to be valid if the parameter R^2_{int} is less than 0.4 and the parameter Q^2_{int} is less than 0.05. If the opposite happens, then an acceptable model cannot be obtained for the specific modeling method and data. (Zhang et al., 2006; Melagraki and Afantitis, 2015). To judge the predicting ability of the developed PLS models, we have also used an external validation parameters proposed by Golbraikh and Tropsha (2002) (Golbraikh and Tropsha, 2002; Vrontaki et al., 2017). Based on these criteria, the model can be accepted if:

- i) $r^2 > 0.6$
- ii) $(r^2 - r^2_0)/r^2 < 0.1$ or $(r^2 - r^2_0)/r^2 < 0.1$
- iii) $1.15 > k > 0.85$ or $1.15 > k > 0.85$

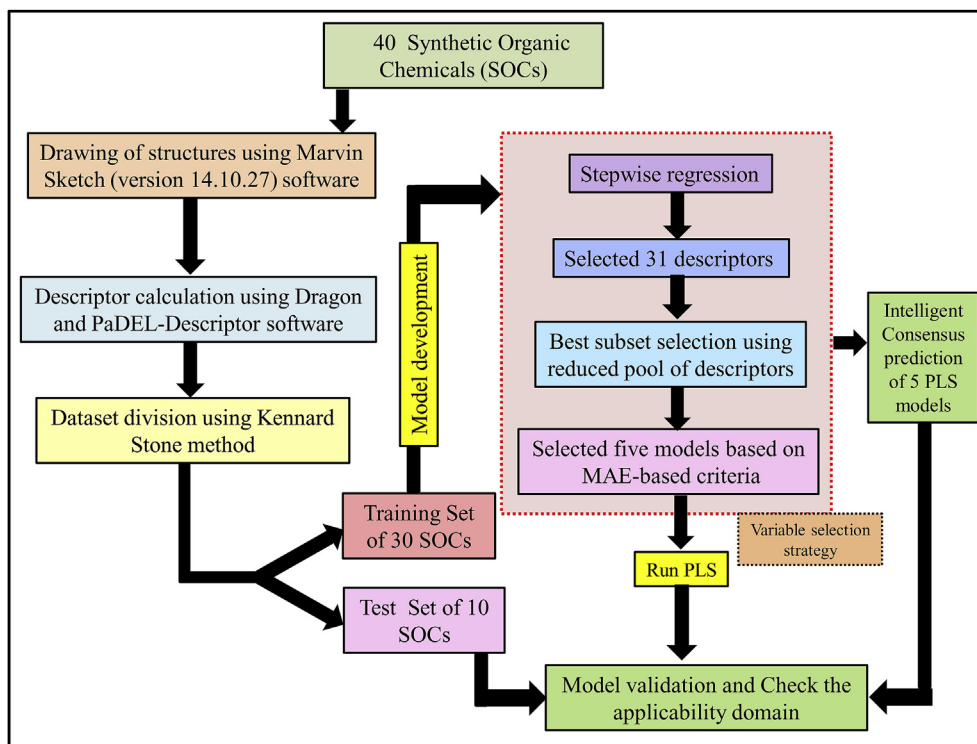


Fig. 1. Schematic representation of the steps involved in the development of final PLS models.

2.6. Applicability domain

The applicability domain is generally defined as the response and chemical structure space within which the training set compounds are occupied. The developed QSPR models are able to make predictions for new compounds properly when the compounds lie within the region of chemical space of the training set molecules. In this present work, we have checked the applicability domain of the developed model using the standardization approach (Roy et al., 2015). All the compounds are found to be present within the AD.

2.7. Software used

We have used Marvin sketch version 5.5.0.1 software to draw chemical structures. PADEL-Descriptor software and Dragon software version 6 were used for descriptor calculation. The data set

division was performed by using “DatasetDivisionGUI1.2” software. Stepwise regression was performed by using MINITAB software version 13.14. The PLS analysis was performed by using Partial Least Squares version 1.0 tool.

3. Results and discussions

In the current study, five PLS models (Box 1) were developed for the dataset containing 40 diverse hazardous SOCs having significant adsorption affinity for SWCNTs, using a reduced descriptor pool obtained by stepwise regression method, as discussed in Methods and Materials section. We have validated the models using various internal and external validation parameters, which showed that the models are statistically significant (Table 1). The MAE based criteria of all the models were passed which indicates that all the models are acceptable. We have also checked the

Table 1
Statistical quality and validation parameters obtained from the developed PLS models.

Type of Model	Training set statistics					Test set statistics							
	Model R ²	Model Q ² _(LOO)	MAE _{train}	$\overline{r^2}_{m(LOO)}$	$\Delta r^2_{m(LOO)}$	R ² _{pred} or Q ² F ₁	Q ² F ₂	CCC	$\overline{r^2}_{m(test)}$	$\Delta r^2_{m(test)}$	MAE _(100%)	MAE _(95%)	MAE
IM1 (LV = 4)	0.928	0.894	Good	0.850	0.056	0.900	0.899	0.950	0.869	0.043	0.382	0.338	Moderate
IM2 (LV = 4)	0.938	0.901	Good	0.860	0.036	0.929	0.928	0.965	0.904	0.054	0.275	0.214	Good
IM3 (LV = 3)	0.949	0.890	Good	0.846	0.063	0.901	0.900	0.956	0.851	0.061	0.320	0.249	Moderate
IM4 (LV = 4)	0.920	0.861	Good	0.806	0.059	0.898	0.897	0.953	0.875	0.065	0.359	0.295	Good
IM5 (LV = 3)	0.937	0.886	Good	0.842	0.008	0.923	0.922	0.963	0.905	0.053	0.322	0.263	Good
CM0						0.914	0.913	0.959	0.894	0.059	0.282	0.219	Good
CM1						0.912	0.911	0.958	0.890	0.060	0.327	0.268	Good
CM2						0.913	0.912	0.959	0.889	0.060	0.322	0.262	Good
CM3						0.938	0.937	0.971	0.903	0.044	0.250	0.189	Good

CM0 = Ordinary consensus predictions.

CM1 = Average of predictions from individual models IM1 through IM5.

CM2 = Weighted average predictions from individual models IM1 through IM5.

CM3 = Best selection of predictions (compound-wise) from individual models IM1 through IM5.

*Note that we have run the “Intelligent consensus predictor tool” using the options, AD: No; Dixon Q-test: No; Euclidean distance cut-off: 0.4.

Table 2

Validation results for the final PLS models obtained according to Golbraikh and Tropsha's criteria.

SI No.	Models	r^2	$[(r^2-r_0^2)/r^2]$	$[(r^2-r_0'^2)/r^2]$	k	k'	Remarks
	Threshold value	>0.6	<0.1	<0.1	$0.85 < k < 1.15$	$0.85 < k' < 1.15$	
1	M1	0.906	-0.094	-0.088	0.940	1.013	Passed
2	M2	0.932	-0.065	-0.068	0.981	0.982	Passed
3	M3	0.927	-0.065	-0.072	0.910	1.054	Passed
4	M4	0.918	-0.079	-0.082	0.912	1.049	Passed
5	M5	0.933	-0.065	-0.067	0.936	1.031	Passed

consensus predictivity of all the individual models (IM1-IM5) using "Intelligent consensus predictor" tool to see whether the quality of predictions of test set compounds can be enhanced or not. The consensus predictivity of the test set compounds were found to be better than the individual models based on MAE based criteria as depicted in Table 1 (the winner model is CM3). The descriptors obtained from the individual models are discussed elaborately in this section. To judge the predicting ability of the developed PLS models, we have also validated the model using Golbraikh and Tropsha criteria, and the results are given in Table 2. The results showed that the models are acceptable based on these criteria. We have performed randomization test using the SIMCA-P software to verify whether the model was obtained by any chance or not. The intercept of both R^2 and Q^2 values are below the stipulated values of $R^2_{int} < 0.4$ and $Q^2_{int} < 0.05$ which confirmed that the models are not obtained by any chance (Figs. S1–S5). The definitions and

contributions of different descriptors obtained from five PLS models are depicted in Table 3. In equations as depicted below, $n_{training}$ is the number of compounds used to develop the models, and n_{test} is the number of compounds used for external prediction. The leave one out (LOO) cross validated correlation coefficient Q^2 ($Q^2 = 0.861–0.901$) above the critical value of 0.5 signifies the statistical reliability of the models. The predictive R^2 (R^2_{pred}) or Q^2_{F1} ($Q^2_{F1} = 0.898–0.929$) and Q^2_{F2} ($Q^2_{F2} = 0.897–0.928$) show good predictive ability of the models. We have also checked the applicability domain of the compounds using the standardization approach. All the compounds are found to be present within the AD. The scatter plot of observed vs. predicted adsorption coefficient for five PLS models are depicted in Fig. 2.

Box 1

Model1 :

$$\log k = 2.881 + 0.153 \times MLOGP2 + 4.183 \times ETA_Shape_Y - 1.472 \times nRNR2 - 11.116 \times X2A - 0.569 \times B07[C - S]$$

$$n_{training} = 30, LV = 4, R^2 = 0.928, R^2_{(adj)} = 0.916, Q^2 = 0.894, S = 0.348, PRESS = 4.452, F = 80.52$$

$$r^2_{m(LOO)} = 0.850, \Delta r^2_{m(LOO)} = 0.056, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.900, Q^2_{F2} = 0.899, r^2_{m(test)} = 0.869, \Delta r^2_{m(test)} = 0.043, CCC = 0.950$$

MAEbasedcriteria = Moderate

Model2 :

$$\log k = -1.125 + 0.169 \times MLOGP2 + 4.565 \times ETA_Shape_Y - 1.316 \times nRNR2 - 0.639 \times B07[C - S] + 0.463 \times B04[C - C]$$

$$n_{training} = 30, LV = 4, R^2 = 0.938, R^2_{(adj)} = 0.928, Q^2 = 0.901, S = 0.324, PRESS = 4.175, F = 94.43,$$

$$r^2_{m(LOO)} = 0.860, \Delta r^2_{m(LOO)} = 0.036, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.929, Q^2_{F2} = 0.928, r^2_{m(test)} = 0.904, \Delta r^2_{m(test)} = 0.054, CCC = 0.965, MAEbasedcriteria = Good$$

Model3 :

$$\log k = -1.291 + 0.190 \times MLOGP2 + 4.871 \times ETA_Shape_Y - 1.887 \times nRNR2 + 0.231 \times H - 051 + 0.351 \times B06[C - O]$$

$$n_{training} = 30, LV = 3, R^2 = 0.949, R^2_{(adj)} = 0.943, Q^2 = 0.890, S = 0.287, PRESS = 4.620, F = 161.42$$

$$r^2_{m(LOO)} = 0.846, \Delta r^2_{m(LOO)} = 0.063, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.901, Q^2_{F2} = 0.900, r^2_{m(test)} = 0.851, \Delta r^2_{m(test)} = 0.061, CCC = 0.956, MAEbasedcriteria = Moderate$$

Model4 :

$$\log k = -0.448 + 0.176 \times MLOGP2 + 5.052 \times ETA_Shape_Y - 1.520 \times nRNR2 + 0.240 \times B06[C - O] - 2.245 \times X2A$$

$$n_{training} = 30, LV = 4, R^2 = 0.920, R^2_{(adj)} = 0.907, Q^2 = 0.861, S = 0.366, PERSS = 5.686, F = 72.39$$

$$r^2_{m(LOO)} = 0.806, \Delta r^2_{m(LOO)} = 0.059, MAEbasedcriteria = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.898, Q^2_{F2} = 0.897, r^2_{m(test)} = 0.875, \Delta r^2_{m(test)} = 0.065, CCC = 0.953, MAEbasedcriteria = Good$$

Model5 :

$$\log k = -1.117 + 0.179 \times MLOGP2 + 4.564 \times ETA_Shape_Y - 1.437 \times nRNR2 + 0.457 \times B06[C - O] - 0.496 \times B07[C - S]$$

$$n_{training} = 30, LV = 3, R^2 = 0.937, R^2_{(adj)} = 0.930, Q^2 = 0.886, S = 0.321, PRESS = 4.803, F = 128.07$$

$$\overline{r^2}_{m(LOO)} = 0.842, \Delta r^2_{m(LOO)} = 0.008, MAE_{basedcriteria} = Good$$

$$n_{test} = 10, Q^2_{F1} = 0.923, Q^2_{F2} = 0.922, \overline{r^2}_{m(test)} = 0.905, \Delta r^2_{m(test)} = 0.053, CCC = 0.963, MAE_{basedcriteria} = Good$$

3.1. Descriptors related to hydrophobic interaction

The descriptor MLOGP2, represents squared Moriguchi octanol water partition coefficient, calculated from the regression equation of Moriguchi logP model (Moriguchi et al., 1994; Ojha and Roy, 2018) consisting of 13 parameters.

The positive regression coefficient of this descriptor indicates that hydrophobicity is directly correlated with the adsorption property of organic pollutants. Thus, the organic pollutants bearing highly hydrophobic property can easily get adsorbed onto SWCNTs as evidenced by the compounds **25 (Phenanthrene)**, **2 (1,2,4-trichlorobenzene)** and **17 (Ethinyl estradiol)** as their corresponding MLOGP2 values are 18.762, 16.507 and 16.033 respectively, whereas less hydrophobic organic pollutants are poorly adsorbed onto SWCNTs as evidenced by the compounds **6 (4,6-Diaminopyrimidine)**, **26 (Pyrimidine)** and **8 (Aniline)**, as their corresponding MLOGP2 values are 0.256, 0, and 2.268, respectively. Therefore, it can be inferred that the hazardous SOCs get adsorbed onto the SWCNTs through hydrophobic interactions. For proper adsorption, synthetic organic chemicals should be hydrophobic in nature. MLOGP2 is not strictly a 2D descriptor as its numerical value depends on intermolecular H-bonds (as it depends on molecular conformation).

The next descriptor B04[C–C] is a 2D binary fingerprint descriptor corresponding to presence/absence of C–C bond at topological distance 4. The positive regression coefficient of this descriptor indicates that presence of C–C bond at the topological distance 4 is important for good adsorption of SWCNTs. The descriptor is related to the size of molecule. If the size of the molecule increases, hydrophobic interaction of the molecule with SWCNTs also increases hence adsorption coefficient also increases. As for example, compounds **25 (phenanthrene)**, **22 (Naphthalene)** and **17 (ethinyl estradiol)** contain single C–C bond at the topological distance 4, and their corresponding adsorption coefficient values are 3.67, 1.8 and 2.87 respectively (higher adsorption

coefficient values). While absence of such fragment decreases the adsorption of organic pollutants to SWCNTs as shown in compounds **26 (pyrimidine)**, **6 (4,6-diaminopyrimidine)** and **8 (aniline)** (adsorption coefficient –1.56, –0.27 and –0.16 respectively).

Another significant descriptor, X2A, indicates average connectivity index of order 2; it encodes the ‘chi’ value across two bonds, which can be calculated on basis of Kier and Hall’s connectivity index and defined in the following equation:

$${}^2X = \sum_{b=2}^B (\delta_i \cdot \delta_j \cdot \delta_k)_b^{-0.5}$$

Here, b runs over the 2nd order sub graphs having n vertices with B edges, δ_i , δ_j and δ_k are number of other vertices attached to vertex i, j and k respectively. This descriptor has a negative contribution towards the adsorption coefficient (logK) of organic pollutants by SWCNTs as evidenced by the negative regression coefficient. This indicates that the adsorption property of hazardous SOCs decreases with an increase in the numerical value of this descriptor. For example, compounds **26 (Pyrimidine)**, **8 (Aniline)** and **6 (4,6-Diaminopyrimidine)** have descriptor values 0.354, 0.343 and 0.338 in that order, and their corresponding adsorption coefficient values are –1.56, –0.16 and –0.27 respectively. If we consider compounds **25 (Phenanthrene)** and **17 (Ethinyl estradiol)**, their descriptor values are less (0.272 and 0.257 respectively), thus their corresponding adsorption coefficient value is high (logK values are 3.67 and 3.64 respectively).

3.1.1. Mechanistic interpretation of hazardous SOCs containing higher and lower adsorption coefficient based on hydrophobic interaction

Phenanthrene (Compound **25**) (shown in Fig. S1 in Supplementary Materials) is a poly aromatic hydrocarbon (PAH) and non ionic in nature. Its MLOGP2 value is 18.76. Due to its hydrophobic property, it can strongly interact with hydrophobic surface of

Table 3
Definition and contribution of different descriptors obtained from five PLS models.

Sl. no.	Name of the descriptors	Contribution	Discussion	Mechanism	Frequency
1	MLOGP2	+ve	Squared Moriguchi octanol-water partition coeff. (logP ²)	Hydrophobic interaction	5
2	ETA_Shape_Y	+ve	ETA_Shape_Y = $(\sum \alpha)_V / \sum \alpha$, $(\sum \alpha)_V$ stands for summation of α values of the vertices that are joined to three other non-hydrogen vertices in the connected molecular graph. Gives a measure of molecular shape.		5
3	nRNR2	-ve	Number of tertiary amines (aliphatic)	Unable to form hydrogen bond due to the absence of 5 free hydrogen atom.	
4	B07[C–S]	-ve	Presence/absence of C–S at topological distance 7		3
5	B04[C–C]	+ve	Presence/absence of C–C at topological distance 4	Hydrophobic interaction	1
6	X2A	-ve	average connectivity index of order 2	Hydrophobic interaction	2
7	H-051	+ve	H attached to alpha-C	Electrostatic interaction. H atoms attached to α carbon 1 atom can easily donate protons and may involve in electrostatic interaction.	1
8	B06[C–O]	+ve	Presence/absence of C–O at topological distance 6	Formation of hydrogen bond	3

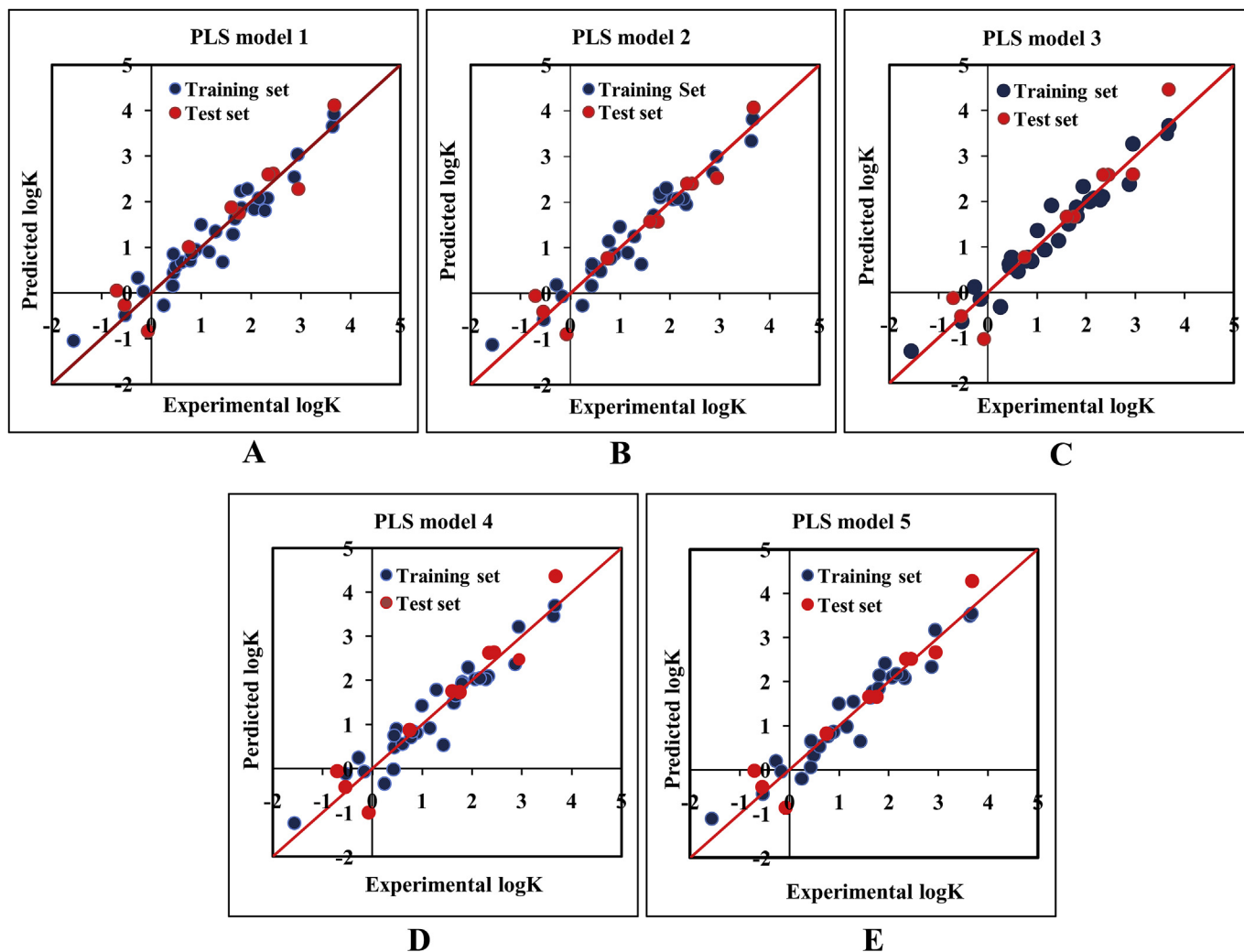


Fig. 2. The scatter plot of the observed and the predicted adsorption coefficient ($\log K$) of the developed PLS models (models M1–M5).

SWCNTs. The B04[C–C] value for phenanthrene is 1 (positive contribution) and X2A value is also low (0.272) (negative contribution) which supports that strong interactions occur between phenanthrene and SWCNTs (Chen et al., 2008). Phenanthrene has an intensive electron donor property. Phenanthrene donates its π -electron and can easily get converted to a cationic form, and thus it more easily interacts with the surface of functionalized (with –COOH) SWCNTs (Gotovac et al., 2007).

When we compare between Phenanthrene (compound **25**) and Naphthalene (compound **22**) (Fig. S6 in Supplementary Materials), both are poly aromatic hydrocarbons (PAHs) and non ionic in nature. Their corresponding MLOGP2 values are 18.76 and 11.46 respectively. Due to their hydrophobic property, they can strongly interact with hydrophobic surface of SWCNTs. The B04[C–C] value for phenanthrene and naphthalene is also 1. Both phenanthrene and naphthalene show strong interactions with SWCNTs. They only differ in polarity and electron donor-acceptor ability (Chen et al., 2008). Thus, the adsorption coefficient of phenanthrene ($\log K = 3.67$) is higher than naphthalene ($\log K = 1.8$) because naphthalene contains less number of aromatic rings and is therefore less hydrophobic than phenanthrene.

Another example, 1,2,4-Trichlorobenzene (compound number **2**) (Fig. S6 in Supplementary Materials) present in the data set, shows higher adsorptive property ($\log K = 2.94$) due to its

bulkiness. As the size of chlorine atom is high, substitution of three chlorine atoms in the benzene ring is responsible for the bulkiness of molecule. SWCNTs shows molecular sieving effect, so based on the unit surface area (Chen et al., 2007), it shows a stronger affinity towards trichlorobenzene. As the volume of molecule increases, the molecule would preferentially like to be in the non-polar phase. As a result, the partition-coefficient also increases ($\log P$ value for benzene and chlorobenzene are 2.13 and 2.84 respectively). The numerical value for MLOGP2 (squared Moriguchi octanol-water partition coefficient) of this compound is also high (21.476). At the same time, the molecule contains a benzene ring, which is responsible for π - π interaction with graphene sheets of carbon nanotubes. Chlorobenzene is also able to participate in hydrogen bonding (though moderately). Thus, the hazardous SOCs containing bulky hydrophobic groups (reflected in the MLOGP2 descriptor) is influential for adsorption of organic pollutants to SWCNTs.

Another compound, Ethinyl estradiol (compound number **17**) (Fig. S6 in Supplementary Materials) shows good adsorptive property to the SWCNTs due to its hydrophobicity (Borisover and Graber, 2003). The higher MLOGP2 value (16.033), presence of single C–C fragment at topological distance 4 and low X2A value (0.257) also give evidences for its hydrophobicity as well as higher adsorptive property. Another mechanism, π - π electron donor-acceptor interaction, also supports the higher adsorptive property

of Ethinyl estradiol onto SWCNTs. Due to the presence of two phenolic groups (charge donor), it can strongly interact with SWCNTs through π - π electron donor-acceptor interaction. (Chen et al., 2007; Zhao et al., 2002). Hydrogen bonding between two -OH groups of Ethinyl estradiol and SWCNTs is also possible, which supports a favorable mechanism for adsorption of this compound with SWCNTs (Pan and Xing, 2008).

We can consider 4,6-diaminopyrimidine (compound number 6) in comparison to pyrimidine (compound 26) (Fig. S6 in Supplementary Materials) (lower range of adsorption coefficient). Pyrimidine is an electron deficient system in comparison to benzene. Thus, it can weaken the π - π electron donor acceptor interaction with SWCNTs (Wang et al., 2010b). On the other hand, 4,6-diaminopyrimidine contains two amino groups which are strong electron donating groups and increase the electron density of aromatic ring. They may form stronger π - π interaction as compared to pyrimidine. The numerical value of X2A descriptor for 4,6-Diaminopyrimidine and pyrimidine are 0.338 and 0.358 respectively. For all these reasons, the adsorption coefficient value of 2,6-diaminopyrimidine ($\log K = -0.27$) and pyrimidine ($\log K = -1.56$, lowest active compound present in dataset) are in the lower range.

Thus, the information obtained from the descriptors, MLOGP2, B04[C-C] and X2A suggested that the organic pollutants can adhere to the surface of SWCNTs by strong hydrophobic interaction.

3.2. Descriptors related to electrostatic interaction

The descriptor H-051, indicates the number of H atoms attached to α carbon atom. Such H atoms are very active in nature. They can easily donate protons and may involve in electrostatic interaction between SWCNTs and synthetic organic chemicals. The positive regression coefficient of this descriptor indicates that organic pollutants contain higher number of such hydrogen atoms have good adsorption property as shown in compounds 30 (Tylosin) and 27 (Sulfamethoxazole). The numerical values of H-051 for compounds 30 and 27 are 5 and 3, respectively, and their corresponding $\log K$ values are 0.43 and 1.43, respectively. On the other hand, in case of compounds 6 (4,6-Diaminopyrimidine) and 8 (Aniline), the adsorption coefficient values ($\log K$ values are -0.27 and -0.16 respectively) decrease due to the absence of α H atoms.

3.2.1. Mechanistic interpretation of hazardous SOCs containing higher and lower adsorption coefficient based on electrostatic interaction

Molecules like Sulfamethoxazole and Tylosin (Fig. S7 in Supplementary Materials) are large in size. Larger molecules adopt themselves in such a manner that they can easily fit with the curvature surface and make stable complex with CNTs (Zhou et al., 2001; Richard et al., 2003; Karajanagi et al., 2004; Gurevitch and Srebnik, 2008). The adsorption energy provides the steric energy required for the conformational changes of organic molecules (Pan et al., 2008). Sulfamethoxazole is well adsorbed to the SWCNTs as its corresponding H-051 value is 3. In case of Tylosin, its descriptor value for H-051 is 5 but its adsorption coefficient value is moderate as compared to Sulfamethoxazole, because its MLOGP2 value is very less (1.604). On the other hand, 4,6-diaminopyrimidine (compound 6) and aniline (compound 8) show poor adsorption affinity towards the SWCNTs as discussed above.

3.3. Descriptors related to hydrogen bonding interaction

2D binary fingerprints descriptor, B06[C-O], indicates presence/absence of C-O bond at topological distance 6. B06[C-O] have a positive regression coefficient which implies that presence of C-O fragment at topological distance 6 is beneficial for the adsorption of


organic pollutants to SWCNT, as the C-O fragment is capable of forming hydrogen bonds with SWCNTs. For example, each of compounds 17 (Ethinyl estradiol), 16 (Diuron) and 4 (2,4-Dinitrotoluene) contains single C-O fragment at the topological distance 6, and their corresponding adsorption coefficient values are 3.64, 2.28 and 2.07, whereas compounds 26 (Pyrimidine), 6 (4,6-Diaminopyrimidine) and 8 (Aniline) have lower adsorption affinity to the SWCNTs due to the absence of such fragment (Fig. S8 in Supplementary Materials). Compounds 17 (Ethinyl estradiol) and 4 (2,4-Dinitrotoluene) contain a C-O fragment at the topological distance 6, which indicates that they are capable of forming hydrogen bonds with functionally modified SWCNTs. Therefore, their adsorption coefficient values are in higher (3.64 and 2.07 respectively) range. On the other hand, both compounds 30 (Tylosin) and 20 (Lyncomycin) contain one aliphatic tertiary amine, so, they are not capable of forming any hydrogen bond, and thus adsorption coefficient value is less (briefly discuss in ETA_Shape_Y).

The descriptor, nRNR2 indicates the number of aliphatic tertiary amines present in a compound. Due to absence of free hydrogen atoms, tertiary amine does not act as a hydrogen bond donor like primary or secondary amine. The negative regression coefficient of this descriptor indicates that higher number of aliphatic tertiary amine weakens the interaction between synthetic organic chemicals and SWCNTs and vice versa. For example, compounds 30 (Tylosin) and 20 (Lyncomycin) have descriptor value 1, and their corresponding adsorption coefficient is less (0.43 and -0.53 respectively), while compounds with lower descriptor value (no such group) have higher adsorption coefficient as shown in case of compounds 17 (Ethinyl estradiol) and 25 (Phenanthrene) ($\log K$ values are 3.64 and 3.67 respectively). If we consider compounds 30 (Tylosin) and 20 (Lyncomycin), Tylosin is moderately active as compared to Lyncomycin because the latter contains 5 α -H atom (H-051 value is 5) and C-O fragment at topological distance 6 (B06 [C-O] value is 1).

3.3.1. Other modeled descriptors essential for adsorption of hazardous SOCs to SWCNTs

The descriptor ETA_Shape_Y, is a first generation extended topochemical atom index. ETA_Shape_Y (Roy, 2015) can be calculated by using the following formula:

$$\text{ETA_Shape_Y} = \frac{(\sum \alpha)_Y}{\sum \alpha}$$

$(\sum \alpha)_Y$ stands for summation of α value (a volume measure) of the vertices that are joined to three other non-hydrogen vertices in the connected molecular graph and forming a Y-shaped structural fragment like tertiary groups (). It gives a measure of molecular shape. The positive regression coefficient of this descriptor indicates that the branching is directly correlated with adsorption of organic pollutants to SWCNTs. The higher degree of branching plays a crucial role to enhance the adsorption affinity of synthetic organic chemicals to SWCNTs as evidenced by the compounds 4 (2,4-Dinitrotoluene) and 16 (Diuron) (corresponding $\log K$ values are 2.07 and 2.28 respectively) (Fig. S9 in Supplementary Materials) with their corresponding descriptor values are in the higher range (0.408 and 0.340 respectively). On the other hand, compounds 26 (Pyrimidine) and 9 (Benzene) have the descriptor value of 0 and thus, their corresponding adsorption coefficient is also low (-1.56 and 0.25 respectively). Between 2,4-Dinitrotoluene (compound 4) and Diuron (compound 16), the adsorption affinity of Diuron is higher (though its ETA_Shape_Y value is comparatively less) than that of 2,4-Dinitrotoluene due to its hydrophobicity (MLOGP2 values are 7 and 5.02 respectively).

Another descriptor, B07[C-S], indicates the presence/absence of

C–S at topological distance 7. The negative regression coefficient of this descriptor indicates that presence of C–S fragment at the topological distance 7 in hazardous SOCs is not beneficial for the adsorption of SWCNTs as evidenced by compounds **20** (Lincosmycin) and **21** (Methyl Orange) (they contain C–S fragment at the topological distance 7, and their corresponding adsorption coefficient values are -0.53 and 0.49 respectively). On the other hand, compounds **25** (Phenanthrene) and **22** (naphthalene) (Fig. S10 in Supplementary Materials) do not contain any such fragments, so their adsorption coefficient value is higher.

The adsorption coefficient value of Lincosmycin (compound number **20**) is low (in spite of high ETA_Shape_Y) as compared to Methyl orange (-0.53) because of its low hydrophobicity (MLOGP2 is 0.538).

4. Overview and conclusion

In this present study, we have developed PLS QSPR models for a dataset containing 40 diverse synthetic organic chemicals (herbicides, fungicides, EDCs, PAH, contrasting agent, dyes) having defined adsorption affinity for SWCNTs, by applying different strategies. We have validated the models using various internal and external validation parameters, which showed that the models were statistically significant. We have also checked the consensus

predictivity of all the individual models (IM1-IM5) using “Intelligent consensus predictor” tool and found that the consensus predictivity of the test set compounds was better than the individual models based on MAE based criteria as depicted in Table 1 (winner model is CM3).

The present study shows how the chemical and structural features of diverse hazardous SOCs alter the adsorption property to SWCNTs. From the insights obtained from five PLS models, we have concluded that hydrophobic surface of the molecules, molecular shape and degree of branching, presence of two carbon atoms at topological distance 4, number of H atom attached with α -C atom, presence of carbon and oxygen atom at the topological distance 6 can enhance the adsorption of hazardous SOCs to the SWCNTs. On the other hand, number of tertiary aliphatic amine, presence of carbon and sulphur at topological distance 7 may be detrimental for the adsorption of hazardous SOCs to the SWCNTs. The adsorption mechanism as evidenced from different contributed descriptors is depicted in Fig. 3. Among all the above mentioned descriptors, MLOGP2 has the strongest impact on the adsorption of hazardous SOCs onto SWCNTs. The conclusions drawn in the present study are also supported by several studies published previously. Sun et al. (2012) and Wang et al. (2010a) suggested that hydrophobic interaction is very crucial for adsorption of hazardous SOCs to SWCNTs. Ding et al. (2016b) reported that the potency of

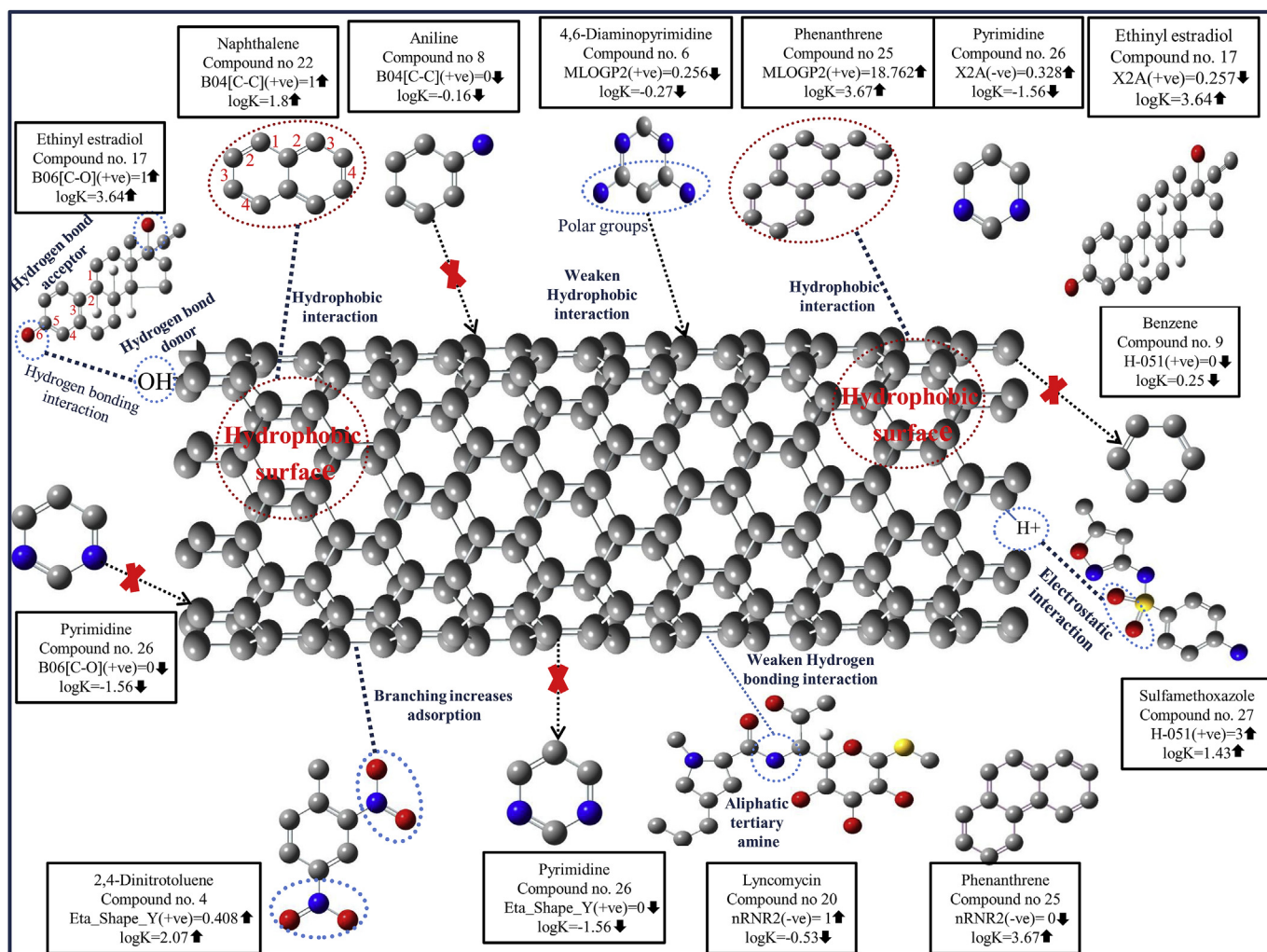


Fig. 3. Adsorption mechanism of contributed descriptors for the adsorption of synthetic organic chemicals onto SWCNTs/functionalized SWCNTs.

adsorption is positively correlated with hydrophobicity, and it is the principal reason behind the adsorption capacities of different hazardous SOCs. Thus, the developed models give information about the important structural requirements or essential molecular properties and the requisite features of molecules that are important to increase or decrease the adsorption of the hazardous SOCs onto SWCNTs. The information obtained from the developed models may be useful for the management of environmental pollution.

Acknowledgement

This work is supported by AICTE, New Delhi in the form of fellowship to SG. Financial assistance from the University Grants Commission, New Delhi, Government of India in the form of a fellowship to PKO (Letter number and date: F/PDFSS-2015-17-WES-11996; Dated: 06/04/2016) is thankfully acknowledged. KR thanks CSIR, New Delhi for financial assistance under a Major Research project (CSIR Project No. 01IJ2895/17/EMR-II).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chemosphere.2019.04.124>.

References

- Domeno, C., Nerin, C., 2003. Fate of polyaromatic hydrocarbons in the pyrolysis of industrial waste oils. *J. Anal. Appl. Pyrol.* 67, 237–246.
- Ahmaruzzaman, M., 2008. Adsorption of phenolic compounds on low-cost adsorbents: a review. *Adv. Colloid Interface Sci.* 143, 48–67.
- Ahmed Adam, O.E.-A., Al-Dujaili, A.H., 2013. The removal of phenol and its derivatives from aqueous solutions by adsorption on petroleum asphaltene. *J. Chem.*, 694029.
- Bi, E., Schmidt, T.C., Haderlein, S.B., 2007. Environmental factors influencing sorption of heterocyclic aromatic compounds to soil. *Environ. Sci. Technol.* 41, 3172–3178.
- Borisover, M., Graber, E.R., 2003. Classifying NOM-organic sorbate interactions using compound transfer from an inert solvent to the hydrated sorbent. *Environ. Sci. Technol.* 37, 5657–5664.
- Carter, A.D., 2000. Herbicide movement in soils: principles, pathways and processes. *Weed. Res.* 40, 113–122.
- Chen, W., Duan, L., Zhu, D., 2007. Adsorption of polar and nonpolar organic chemicals to carbon nanotubes. *Environ. Sci. Technol.* 41, 8295–8300.
- Chen, J., Chen, W., Zhu, D., 2008. Adsorption of nonionic aromatic compounds to single-walled carbon nanotubes: effects of aqueous solution chemistry. *Environ. Sci. Technol.* 42, 7225–7230.
- Ding, H., Chen, C., Zhang, X., 2016a. Linear solvation energy relationship for the adsorption of synthetic organic compounds on single-walled carbon nanotubes in water. *SAR QSAR Environ. Res.* 27, 31–45.
- Ding, H., Li, X., Wang, J., Zhang, X., Chen, C., 2016b. Adsorption of chlorophenols from aqueous solutions by pristine and surface functionalized single-walled carbon nanotubes. *J. Environ. Sci.* 43, 187–198.
- Golbraikh, A., Tropsha, A., 2002. Beware of q^2 ! *J. Mol. Graph. Model.* 20, 269–276.
- Gotovac, S., Yang, C.-M., Hattori, Y., Takahashi, K., Kanoh, H., Kaneko, K., 2007. Adsorption of polyaromatic hydrocarbons on single wall carbon nanotubes of different functionalities and diameters. *J. Colloid Interface Sci.* 314, 18–24.
- Gurevitch, I., Srebnik, S., 2008. Conformational behavior of polymers adsorbed on nanotubes. *J. Phys. Chem.* 128, 144901.
- Harrison, R.M., Smith, D.J.T., Luhana, L., 1996. Source apportionment of atmospheric polycyclic aromatic hydrocarbons collected from an urban location in Birmingham, UK. *Environ. Sci. Technol.* 30, 825–832.
- Karajanagi, S.S., Vertegel, A.A., Kane, R.S., Dordick, J.S., 2004. Structure and function of enzymes adsorbed onto single-walled carbon nanotubes. *Langmuir* 20, 11594–11599.
- Khani, H., Moradi, O., 2013. Influence of surface oxidation on the morphological and crystallographic structure of multi-walled carbon nanotubes via different oxidants. *J. Nanostruct. Chem.* 3, 73.
- Kim, H., Hwang, Y.S., Sharma, V.K., 2014a. Adsorption of antibiotics and iopromide onto single-walled and multi-walled carbon nanotubes. *Chem. Eng. J.* 255, 23–27.
- Kim, H.J., Choi, K., Baek, Y., Kim, D.G., Shim, J., Yoon, J., Lee, J.C., 2014b. High-performance reverse osmosis CNT/polyamide nanocomposite membrane by controlled interfacial interactions. *ACS Appl. Mater. Interfaces* 6, 2819–2829.
- Lara, I.V., Zanella, I., Fagan, S.B., 2014. Functionalization of carbon nanotube by carboxyl group under radial deformation. *Chem. Phys.* 428, 117–120.
- Latkar, M., Chakrabarti, T., 1994. Performance of upflow anaerobic sludge blanket reactor carrying out biological hydrolysis of urea. *Water Environ. Res.* 66, 12–15.
- Lin, Z.-P., Ikonomou, M.G., Jing, H., Mackintosh, C., Gobas, F.A.P.C., 2003. Determination of phthalate ester congeners and mixtures by LC/ESI-MS in sediments and biota of an urbanized marine inlet. *Environ. Sci. Technol.* 37, 2100–2108.
- Lohmann, R., MacFarlane, J.K., Gschwend, P.M., 2005. Importance of black carbon to sorption of native PAHs, PCBs, and PCDDs in Boston and New York harbor sediments. *Environ. Sci. Technol.* 39, 141–148.
- Long, R.Q., Yang, R.T., 2001. Carbon nanotubes as superior sorbent for dioxin removal. *J. Am. Chem. Soc.* 123, 2058–2059.
- Mackintosh, C.E., Maldonado, J., Hongwu, J., Hoover, N., Chong, A., Ikonomou, M.G., Gobas, F.A.P.C., 2004. Distribution of phthalate esters in a marine aquatic food web: comparison to polychlorinated biphenyls. *Environ. Sci. Technol.* 38, 2011–2020.
- May, W.E., Wasik, S.P., Freeman, D.H., 1978. Determination of the solubility behavior of some polycyclic aromatic hydrocarbons in water. *Anal. Chem.* 50, 997–1000.
- Melagraki, G., Afantitis, A., 2013. Enalos KNIME nodes: exploring corrosion inhibition of steel in acidic medium. *Chemometr. Intell. Lab. Syst.* 123, 9–14.
- Melagraki, G., Afantitis, A., 2015. A risk assessment tool for the virtual screening of metal oxide nanoparticles through enalos insiliconano platform. *Curr. Top. Med. Chem.* 15, 1827–1836.
- Michael, I., Rizzo, L., McArdell, C.S., Manaia, C.M., Merlin, C., Schwartz, T., Dagot, C., Fatta-Kassinos, D., 2013. Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: a review. *Water Res.* 47, 957–995.
- Moody, C.A., Field, J.A., 2000. Perfluorinated surfactants and the environmental implications of their use in fire-fighting foams. *Environ. Sci. Technol.* 34, 3864–3870.
- Moriguchi, I., Hirano, S., Nakagome, I., Hirano, H., 1994. Comparison of reliability of log P values for drugs calculated by several methods. *Chem. Pharm. Bull.* 42, 976–978.
- Mosayebidorcheh, S., Hatami, M., 2017. Heat transfer analysis in carbon nanotube-water between rotating disks under thermal radiation conditions. *J. Mol. Liq.* 240, 258–267.
- Nakashima, N., 2005. Soluble carbon nanotubes: fundamentals and applications. *Int. J. Nanosci.* 4, 119–137.
- Nielsen, T., 1996. Traffic contribution of polycyclic aromatic hydrocarbons in the center of a large city. *Atmos. Environ.* 30, 3481–3490.
- Ojha, P.K., Roy, K., 2018. Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. *Food Chem. Toxicol.* 112, 551–562.
- Okolo, B., Park, C., Keane, M.A., 2000. Interaction of phenol and chlorophenols with activated carbon and synthetic zeolites in aqueous media. *J. Colloid Interface Sci.* 226, 308–317.
- Padoley, K.V., Mudliar, S.N., Pandey, R.A., 2008. Heterocyclic nitrogenous pollutants in the environment and their treatment options: an overview. *Bioresour. Technol.* 99, 4029–4043.
- Pan, B., Xing, B., 2008. Adsorption mechanisms of organic chemicals on carbon nanotubes. *Environ. Sci. Technol.* 42, 9005–9013.
- Pan, B., Lin, D., Mashayekhi, H., Xing, B., 2008. Adsorption and hysteresis of bisphenol A and 17 α -ethinyl estradiol on carbon nanomaterials. *Environ. Sci. Technol.* 42, 5480–5485.
- Petersen, E.J., Zhang, L., Mattison, N.T., O'Carroll, D.M., Whelton, A.J., Uddin, N., Nguyen, T., Huang, Q., Henry, T.B., Holbrook, R.D., 2011. Potential release pathways, environmental fate, and ecological risks of carbon nanotubes. *Environ. Sci. Technol.* 45, 9837–9856.
- Pimentel, D., 1995. Amounts of pesticides reaching target pests: environmental impacts and ethics. *J. Agric. Environ. Ethics* 8, 17–29.
- Rand-Weaver, M., Margiotta-Casaluci, L., Patel, A., Panter, G.H., Owen, S.F., Sumpter, J.P., 2013. The read-across hypothesis and environmental risk assessment of pharmaceuticals. *Environ. Sci. Technol.* 47, 11384–11395.
- Richard, C., Balavoine, F., Schultz, P., Ebbesen, T.W., Mioskowski, C., 2003. Supramolecular self-assembly of lipid derivatives on carbon nanotubes. *Science* 300, 775–778.
- Roy, K., 2015. Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment. *IGI Global*, PA.
- Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.* 52, 396–408.
- Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. *Chemometr. Intell. Lab. Syst.* 145, 22–29.
- Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometr. Intell. Lab. Syst.* 152, 18–33.
- Roy, K., Ambure, P., Kar, S., Ojha, P.K., 2018. Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J. Chemom.* 32, e2992.
- Roy, J., Ghosh, S., Ojha, P.K., Roy, K., 2019. Predictive quantitative structure-property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). *Environ. Sci. Nano.* 6, 224–247. <https://doi.org/10.1039/C8EN10059E>.
- Singh, R.K., Patel, K.D., Kim, J.-J., Kim, T.-H., Kim, J.-H., Shin, U.S., Lee, E.-J., Knowles, J.C., Kim, H.-W., 2014. Multifunctional hybrid nanocarrier: magnetic CNTs ensheathed with mesoporous silica for drug delivery and imaging system. *ACS Appl. Mater. Interfaces* 6, 2201–2208.
- Snyder, S.A., Westerhoff, P., Yoon, Y., Sedlak, D.L., 2003. Pharmaceuticals, personal

- care products, and endocrine disruptors in water: implications for the water industry. *Environ. Eng. Sci.* 20, 449–469.
- Song, C., 2003. An overview of new approaches to deep desulfurization for ultra-clean gasoline, diesel fuel and jet fuel. *Catal. Today* 86, 211–263.
- Suffet, I.H.M., Khiari, D., Bruchet, A., 1999. The drinking water taste and odor wheel for the millennium: beyond geosmin and 2-methylisoborneol. *Water Sci. Technol.* 40, 1–13.
- Sun, K., Zhang, Z., Gao, B., Wang, Z., Xu, D., Jin, J., Liu, X., 2012. Adsorption of diuron, fluridone and norflurazon on single-walled and multi-walled carbon nanotubes. *Sci. Total Environ.* 439, 1–7.
- Tariq, M.I., Afzal, S., Hussain, I., Sultana, N., 2007. Pesticides exposure in Pakistan: a review. *Environ. Int.* 33, 1107–1122.
- Umetrics, A.B., 2002. SIMCA-P Version 10.0. Umetrics AB, Umea, Sweden.
- Vrontaki, E., Melagraki, G., Afantitis, A., Mavromoustakos, T., Kollias, G., 2017. Searching for novel Janus kinase-2 inhibitors using a combination of pharmacophore modeling, 3D-QSAR studies and virtual screening. *Mini Rev. Med. Chem.* 17, 268–294.
- Walters, R.W., Luthy, R.G., 1984. Equilibrium adsorption of polycyclic aromatic hydrocarbons from water onto activated carbon. *Environ. Sci. Technol.* 18, 395–403.
- Wang, F., Yao, J., Sun, K., Xing, B., 2010a. Adsorption of dialkyl phthalate esters on carbon nanotubes. *Environ. Sci. Technol.* 44, 6985–6991.
- Wang, L., Zhu, D., Duan, L., Chen, W., 2010b. Adsorption of single-ringed N-and S-heterocyclic aromatics on carbon nanotubes. *Carbon* 48, 3906–3915.
- Wang, X., Liu, Y., Tao, S., Xing, B., 2010c. Relative importance of multiple mechanisms in sorption of organic compounds by multiwalled carbon nanotubes. *Carbon* 48, 3721–3728.
- Wold, S., Sjstrm, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* 58, 109–130.
- Yang, K., Zhu, L., Xing, B., 2006. Adsorption of polycyclic aromatic hydrocarbons by carbon nanomaterials. *Environ. Sci. Technol.* 40, 1855–1861.
- Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., Tropsha, A., 2006. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* 46, 1984–1995.
- Zhao, J., Buldum, A., Han, J., Lu, J.P., 2002. Gas molecule adsorption in carbon nanotubes and nanotube bundles. *Nanotechnology* 13, 195.
- Zhao, Q., Yang, K., Li, W., Xing, B., 2014. Concentration-dependent polyparameter linear free energy relationships to predict organic compound sorption on carbon nanotubes. *Sci. Rep.* 4, 3888.
- Zhou, G., Duan, W., Gu, B., 2001. First-principles study on morphology and mechanical properties of single-walled carbon nanotube. *Chem. Phys. Lett.* 333, 344–349.
- Zhu, J., Phillips, S.P., Feng, Y.-L., Yang, X., 2006. Phthalate esters in human milk: concentration variations over a 6-month postpartum time. *Environ. Sci. Technol.* 40, 5276–5281.