Risk assessment of diverse chemical toxicants towards multiple avian species using QSTR and q-RASTR approaches

Thesis submitted in partial fulfilment of the requirements of the Degree of

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

ABHISEK SAMAL

B. PHARM.

Registration No: 163667 of 2022-2023

Examination Roll No: M4PHG24003

Class roll number: **002211402024**

Under the Guidance of

DR. PROBIR KUMAR OJHA

Associate Professor

Drug Discovery & Development Laboratory

Division of Medicinal and Pharmaceutical Chemistry

Department of Pharmaceutical Technology, Jadavpur University

Kolkata – 700 032

India

2024

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains a literature survey and original research as part of my workon **"Risk assessment of diverse chemical toxicants towards multiple avian species using QSTR and q-RASTR approaches".**

All information in this document has been obtained and presented following academicrules and ethical conduct.

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

NAME: **ABHISEK SAMAL**

EXAMINATION ROLL NUMBER: **M4PHG24003**

REGISTRATION NUMBER: **163667 of 2022-2023**

THESIS TITLE: **"Risk assessment of diverse chemical toxicants towards multiple avian species using QSTR and q-RASTR approaches".**

SIGNATURE: Abhisek Samal

DATE: 28/08/2024

PLACE: Jadavpur university,
Kolkata

CERTIFICATE

Department of Pharmaceutical Technology

Jadavpur University

Kolkata - 700032

This is to certify that **Mr. ABHISEK SAMAL,** B. Pharm. (2018-22), has carried out the research work on the subject entitled ***"Risk assessment of diverse chemical toxicants towards multiple avian species using QSTR and q-RASTR approaches"*** under my supervision in Drug Design & Development Laboratory in the Department of Pharmaceutical Technology of this university. He has incorporated his findings into this thesis of the same title, being submitted by him, in partial fulfillment of the requirements for the degree of Master of Pharmacy of Jadavpur University. He has carried out this research work independently and with proper care and attention to my satisfaction.

Probir Kumar Ojha. 28/08/2024

Dr. Probir Kr. Ojha
Associate Professor,
Dept. of Pharmaceutical Technology
Jadavpur University
Kolkata-700 032, W.B., India

DR. PROBIR KUMAR OJHA

Associate Professor
Drug Discovery & Development Laboratory,
Division of Medicinal and Pharmaceutical
Chemistry,Department of Pharmaceutical
Technology,
Jadavpur
University,Kolkata-
700032

29/8/24.

Head
Dept. of Pharmaceutical Technology
Jadavpur University
Kolkata - 700 032, W.B. India

(Prof. Dr. Amalesh Samanta)

Head, Dept. of Pharmaceutical Technology,

Jadavpur University, Kolkata

Dipak Laha 29. 8. 24

(Prof. Dipak Laha)

Dean, Faculty of Engineering and Technology

Jadavpur University, Kolkata

DEAN
Faculty of Engineering & Technology
JADAVPUR UNIVERSITY
KOLKATA-700 032

Acknowledgements

I deem it a pleasure and privilege to work under the guidance of ***Dr. Probir Kumar Ojha,*** *Associate Professor, Drug Discovery & Development Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata-32. I express my deep gratitude and regard 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. Probir Kumar Ojha*** *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 cooperation from my lab mate* ***Shubha Das,*** *seniors* ***Arnab Bhattacharjee, Ankur Kumar, and Arkaprava Banerjee.*** *Their constant support and helpful suggestions have helped me to accomplish this work in time.*

I offer humble gratitude to ***Prof. Dr. Kunal Roy,*** *Professor, Department of Pharmaceutical Technology, Jadavpur University, for the affection and kindness rendered to me throughout my work.*

I am indeed glad to convey cordial thanks to my friends, ***Minarva Manaswini Behera, Rinki Prasad Bhagat, Sai Satyaprakash Mishra, Totan Das, Himadri Bhusan Sahoo, Simant Kumar Jena,*** *and juniors* ***Pabitra Samanta and Prodipta Bhattacharyya.*** *Lastly, I am thankful to the authority of Jadavpur University and Head of the Department for providing all the facilities to carry out this work.*

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. Ashok Kumar Samal, Ms. Biswajita Samal,*** *and sister* ***Amrita Samal*** *for all the love and inspiration without which my dissertation work would remain incomplete.*

Abhisek Samal

ABHISEK SAMAL
Examination Roll No: **M4PHG24003**
Department of Pharmaceutical Technology,
Jadavpur University,
Kolkata-700032

Preface

This dissertation is presented for the partial fulfilment for the degree of Master of Pharmacy in Pharmaceutical Technology. The work presented in this dissertation is spread over two years, which encompasses the development of Quantitative Structure-Toxicity Relationship (QSTR) and Quantitative Read-Across Structure-Toxicity Relationship (q-RASTR) models using easily interpretable two-dimensional (2D) molecular descriptors for efficient prediction of toxicity of diverse organic compounds towards various avian species. The significance of this research is underscored by its practical application, which extends beyond the realm of theory and into the screening of chemical databases, enabling the identification of substances that may pose risks to both human health and the environment.

The identification and evaluation of toxicity in chemical compounds are of paramount importance in addressing potential health risks, encompassing a spectrum of hazards including carcinogenicity, genotoxicity, immunotoxicology, and developmental and reproductive toxicity. These considerations underscore the integral role of toxicity prediction in the intricate process of drug design and development. While preclinical and clinical trials serve as indispensable means of assessing toxicity before public consumption, they are often characterized by exorbitant costs, extensive labour requirements, prolonged timelines, the potential for inconclusive outcomes, and practical infeasibility in certain scenarios.

In recent years, there has been a significant paradigm shift in the field of toxicology, with *in silico* techniques becoming increasingly prominent as a rational alternative to traditional animal testing for predicting toxicity and chemical properties. Driven by ethical considerations, efficiency gains, and cost-effectiveness, and aligned with the 3Rs (replacement, refinement, and reduction of animals in research), these computational methods offer rapid and versatile solutions for assessing chemical toxicity across various compounds. From predicting diverse toxicity types to aiding in drug discovery and environmental impact assessments, in silico techniques are revolutionizing the way we approach chemical evaluation, aligning with both scientific progress and ethical responsibility in the modern era. The classical approach to QSTR owes much of its foundation to the pioneering research led by Hansch in 1960, utilizing statistical modeling based on linear regression to elucidate the relationships between the structural features of molecules and their activity/toxicity/property. The development of predictive QSTR models represents a significant advancement in our ability to assess the toxicological hazards and properties of chemical toxicants. These models are constructed based on chemical information derived from molecular descriptors, enabling a systematic analysis of how the structural features of chemicals relate to their toxicological behaviour.

QSTR modeling, especially when applied to a large set of toxic compounds, often involves a multitude of descriptors, adding complexity and potentially diminishing reliability and predictiveness. In such cases, the utilization of the Read Across Structure-Toxicity Relationship (RASTR) model becomes a viable alternative. RASTR combines the principles of similarity and error-based estimations, merging elements of both read-across (a non-statistical approach) and traditional QSAR modeling. This approach addresses challenges encountered in QSAR modeling related to external validation and the interpretability of Read Across methods.

Recently, an enhanced iteration of the RASTR model, referred to as q-RASTR (Quantitative Read Across Structure-Toxicity Relationship) modeling, has been introduced. q-RASTR utilizes a blend of similarity and error-based descriptors in its modeling, achieving superior predictive potential compared to both QSTR and read-across predictions. The strength of the q-RASTR method lies in its capacity to incorporate information about similarity and error measures into descriptors, facilitating the development of straightforward, interpretable, transferrable, and reproducible models with enhanced predictive capabilities.

In the present study, predictive QSTR as well as q-RASTR models were developed using different classes of simple 2D descriptors to estimate the toxicity of different organic compounds including pesticides. We attempted to explore the toxicity profile of different diverse chemical compounds and pesticides to make a more realistic move towards risk assessment that could be useful in the development of safer or greener chemicals. The predictive models were constructed strictly catering to OECD guidelines and rigorously validated using various internationally accepted internal and external validation parameters.

The following analyses have been performed in this dissertation:

Study 1: Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across.

Study 2: First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail.

Study 3: Chemometric-based exploration of the toxicological significance of diverse chemical toxicants in wild birds with an application of the q-RASTR approach.

The accomplished 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 discussion

Chapter 5: Conclusion

References

Abbreviation

Abbreviations	**Full forms**	**Abbreviations**	**Full forms**
AD	Applicability domain	OECD	Organization for Economic Co-operation and Development
ANN	Artificial neural network	PPDB	Pesticide property database
SVM	Support Vector Machines	PCA	Principal Component Analysis
SVR	Support vector regression	PCR	Principal Component Regression
SAR	Structure-Activity Relationship	PLS	Partial Least Squares
pLD_{50}	Logarithmic conversion of LD_{50}	pLC_{50}	Logarithmic conversion of LC_{50}
DModX	Distance to Model X	PRESS	Predicted residual sum of squares
R^2	Co-efficient of determination	QAAR	Quantitative activity–activity relationship
R^2adj	Adjusted coefficient of determination	QSAR	Quantitative structure-activity relationships
LV	Latent variable	QSPR	Quantitative structure-property relationship
MAE	Mean absolute error	QSTR	Quantitative structure-toxicity relationship
MLR	Multiple Linear Regression	QTTR	Quantitative toxicity–toxicity relationship
CCC	Concordance correlation coefficient	q-RASAR	Quantitative Read Across Structure-Activity Relationship
CM	Consensus model	q-RASTR	Quantitative Read Across Structure-Toxicity Relationship

IM	Individual model	REACH	Registration, Evaluation, Authorization, and Restriction of Chemicals
Q^2_{LOO}	Cross-validated correlation coefficient	MD	Mallard duck
SVM	Support Vector Machines	RNP	Ring-necked pheasant
SAR	Structure-Activity Relationship	CQ	California quail
SD	Standard deviation	VIP	Variable importance plot
SDEP	The standard deviation of error of prediction	ICP	Intelligent Consensus Predictor

Contents

Chapter				Topic	Page no.
				Acknowledgements	i
				Preface	ii-iv
				Abbreviations	v-vi
1				Introduction	1-30
	1.1			Toxicity and it's various aspects	1
		1.1.1		Chemical toxicity	2
		1.1.2		Environmental toxicity	2
		1.1.3		Pesticide toxicity	2
	1.2			Pesticides, Agriculture and Environment	2
	1.3			Pesticide-related risks to biodiversity	3
	1.4			In silico estimation of pesticide toxicity and risk assessment	4
	1.5			Quantitative structure-activity relationships (QSAR) approach	5
		1.5.1		Application of QSAR/QSTR	7
		1.5.2		Significance of QSAR/QSTR	8
	1.6			Concept of molecular descriptors	8
		1.6.1		Types of descriptors	10
			1.6.1.1	Descriptors commonly used in QSTR studies	10
	1.7			Commonly employed QSAR/QSTR methods for Chemometric model development	12
		1.7.1		Classification of QSAR/QSTR approaches based on the type of Chemometric methods used	13
			1.7.1.1	Linear methods	13
			1.7.1.2	Non-linear methods	15
		1.7.2		Classification of QSAR/QSTR approaches based on dimensionality	18
	1.8			Development of Quantitative models over the period	18
	1.9			The methodology of QSAR/QSTR model generation	20
	1.10			Application of QSAR /QSTR	23
	1.11			Computation of different statistical metrics for assessing model quality	24
			1.11.1	Validation metrics for training set	25
			1.11.2	Validation metric for test set	27
			1.11.3	Y-randomization study	28

			1.11.4	Determination of model applicability domain (AD)	28
		1.12		Literature review	29
2				Present work	32-38
	2.1			Study 1: Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across	32-34
	2.2			Study 2: First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail	34-36
	2.3			Study 3: Chemometric-based exploration of the toxicological significance of diverse chemical toxicants in wild birds with an application of the q-RASTR approach	36-38
3				Materials and methods	40
	3.1			Organization for Economic Cooperation and Development (OECD) guidelines for the QSTR model generation	40
	3.2			Study 1	41
		3.2.1		Dataset preparation	41-57
		3.2.2		Descriptor calculation & data preprocessing	57
		3.2.3		Dataset splitting	57
		3.2.4		Selection of features and model building	58
		3.2.5		Validation metrics of QSTR models	58
		3.2.6		Prediction using read-across algorithm	59
		3.2.7		Applicability domain study of the model	59
		3.2.8		Y-randomization study of the developed models	59
		3.2.9		Analysis of parametric assumptions of the developed models	59
		3.2.10		Application of other machine learning (ML) algorithms	61
		3.2.11		Classification QSTR (LDA-QSTR) model development	61
		3.2.12		Screening of the Pesticide Properties Database	61
	3.3			Study 2	62
		3.3.1		Dataset preparation	62
		3.3.2		Calculation of descriptor & data pretreatment	63
		3.3.3		Splitting of dataset	63
		3.3.4		Feature selection & model generation	64
		3.3.5		Statistical validation of the constructed model	64
		3.3.6		Intelligent Consensus Prediction	65
		3.3.7		Applicability domain study of the developed models	66
		3.3.8		Y-randomization study of the developed models	67

		3.3.9	Application of the constructed model on a prepared external dataset for data gap-bridging	67
	3.4		Study 3	67
		3.4.1	Collection, curation, and preparation of toxicity dataset	67
		3.4.2	Computation of molecular descriptors and data pretreatment	74
		3.4.3	Dataset splitting	75
		3.4.4	Feature selection and QSAR model generation	75
		3.4.5	Read-across and RASAR descriptor calculation	75
		3.4.6	q-RASAR-based model development	76
		3.4.7	Applicability domain study of the generated model	76
		3.4.8	Y-randomization test of the generated models	76
		3.4.9	PPDB database assessment by deploying the developed q-RASTR model and reliability study	77
4			Results and discussion	79-127
	4.1		Study 1	79-93
	4.2		Study 2	94-116
	4.3		Study 3	116-127
5			**Conclusion**	129-131
			References	

CHAPTER - 1

Introduction

1. INTRODUCTION

1. Introduction

1.1 Toxicity and it's various aspects

Toxicity is considered as a multidimensional concept that comprises a variety of dimensions. Toxicity can be defined as the capacity of a chemical to produce detrimental consequences on health and these consequences may affect either one cell, an organ, a group of cells, or the whole body might cause anatomical or functional damage, permanently disturb homeostasis, or increase vulnerability to other chemicals or biological stresses, like infectious illnesses. Toxic effects might be obvious harm to the body or a decline in normal body functions which can only be determined through testing. The growing global population and industrial development have highlighted the significant impact that chemicals, particularly pesticides, have on the planet's ecosystems. Most chemicals have the potential to be poisons since they may harm or even kill people when exposed in excess at certain quantities. Understanding how chemicals interact with the environment is crucial since human society depends on so many different kinds and classes of chemicals. The followings are major chemical classes that have a significant influence on the environment: insecticides, agrochemicals, metals, halogenated hydrocarbons, polycyclic aromatic hydrocarbons, pharmaceuticals for humans and animals, dyes, and synthetic and semi-synthetic substances [1].

The negative or bothersome effects of chemicals on the ecosystem, people, or other living beings are referred to as toxicity [2]. Concerns over the possible effects that new chemicals and environmental pollution may have on human health and the environment are on the rise due to the ongoing synthesis of new chemicals and the pollution of the environment. The impact of dangerous chemicals, medications, food items, pesticides, dyes, and pollutants on the environment is a cause for great concern because, despite the vast majority of compounds being used in commerce, only a small percentage of them have undergone adequate testing to determine their potentially harmful environmental characteristics. Over the past six decades, the amounts of chemicals produced on a big scale have grown from 1 million tons to 400 million tons. It is quite expensive and time-consuming to experimentally determine the environmental parameters such as bioconcentration, biotransformation, and toxic effects of commercial chemicals. As there is a huge quantity of chemicals in regular use today and the rapidity with which new chemicals are synthesized and registered, it is evident that our personnel and resources are inadequate for in-depth testing and focusing on their long-term and chronic effects.

Therefore, the development of quantitative models that can easily and accurately anticipate the environmental behavior of huge sets of chemicals is required. These models, which are supported by strong scientific principles and cutting-edge computational methods, are essential instruments for bridging the gap between the rapidly changing chemical landscape and our ability to thoroughly evaluate its effects on the environment.

1.1.1 Chemical toxicity

Chemicals may have both positive and negative effects on the organisms to which they are exposed, and for thousands of years, people have understood how poisons, medicines, pesticides, and other toxic agents work. Various organisms and the environment are increasingly exposed to a growing number of chemicals as a result of industrialization. People understand that evaluating these compound's effects is necessary due to their potential for harm. Chemical toxicity has become a major worldwide issue in recent times due to the abundance of untested compounds [4].

1.1.2 Environmental toxicity

Human reliance on industrial chemicals including pharmaceuticals, and pesticides is increasing rapidly, mostly in the fields of food production, healthcare, and agriculture. The chemical toxicants in use pose a major risk to the local flora and wildlife due to a lack of necessary eco-toxicological knowledge. Consequently, having a direct or indirect impact on the ecological species that are present in the surroundings. These toxic pollutants may generate metabolic and degradation bi-products that cause unfavorable environmental events that are seen in some organisms.

1.1.3 Pesticide toxicity

Numerous pesticides that are hazardous to animal and human health are dispersed into the environment in large quantities. The use of pesticides carelessly and indiscriminately has annoying effects on biodiversity and the world's ecology. Due to their long-lasting and bioaccumulative nature, pesticides have both acute and long-term negative impacts on both aquatic and non-aquatic habitats. As a result, it's crucial to ascertain the origin, frequency, harmful effects, and ecological destiny of pesticides in addition to conducting an accurate risk assessment.

1.2 Pesticides, Agriculture and Environment

The expanding global population puts an enormous strain on the present agricultural system by increasing the demand for food. Thus, agriculture is essential to the advancement of civilization. A decade of improved agricultural technology and developed fertilizer components have led to a modest improvement in agricultural production. However, several risks, including weeds, fungi, pests, and insects, are having a significant impact on agricultural output. Pesticide is one kind of chemical used in modern agriculture with functions such as preventing pests and insects, controlling different plant diseases, reducing damage from different fungi, minimizing waste, and enhancing crop quality. Over the past several decades, there has been a substantial increase in the usage of agrochemicals in agricultural fields to counteract the detrimental impacts of these threats. These pesticides include nematicides, rodenticides, molluscicides, insecticides, fungicides, herbicides, and other hazardous agrochemicals that are frequently employed for particular goals including disease vector control and crop protection [5].

Pesticide usage on crops is estimated to be 2.5 million tons worldwide annually. Nevertheless, the quantity ingested by pests or comes into contact with them represents a relatively small portion of the overall pesticide application. The majority of research has demonstrated that fewer than 0.3% of pesticides sprayed reach the intended insects [6]. As a consequence, toxic residues of pesticides accumulate in the environment and affect both terrestrial and aquatic food chains. Several researchers reported that currently used pesticides are lack specificity which may responsible for toxicity toward various non-target species including humans and birds. According to research on poisoning and the effects of synthetic pesticides on human health, there have been several instances of farmers and rural laborers becoming intoxicated while applying pesticides [7].

Nowadays, pesticide poisonings are thought to be one of the leading causes of death globally, accounting for 220,000 fatalities and 26 million poisonings annually. The presence of pesticide residues in different ecosystem components worries researchers. Pesticide usage is expanding in response to rising agricultural demand, putting non-target creatures like birds, insects, and aquatic life in jeopardy and upsetting the delicate ecological balance on a worldwide scale. Therefore, from the standpoint of ecosystem safety, it is imperative that a range of endangered species should be protected and restored.

1.3 Pesticide-related risks to biodiversity

1.3.1 Terrestrial biodiversity

Terrestrial biodiversity provides several ecological services, like plant pollination and biodiversity monitoring, making the terrestrial environment indispensable to the ecology. Some reports suggest that thirty-five percent of the food crop yield is attributed to biological pollinators like honey bees and birds. We can't even imagine a world without birds as they are an important part of the environment. Approximately, 10,000 avian species exist on the planet, but as per the report, over the past five centuries, a total of 150 bird species have become extinct and one in eight avian species is at the risk of extinction. Birds are among the most identifiable animal species on the planet. Birds are vital to the world environment because they pollinate plants, spread seeds, maintain ecological circles, and aid in biological conservation [8]. Certain human activities have contributed to the decline of 41% of the 1138 water bird populations, even though birds play a vital role in maintaining ecological balance. Various studies show that the number of common birds and forest birds in Europe reduced by around 10%, while the populations of agricultural birds declined by 48% [9]. Some researchers reported that organophosphate pesticides as well as carbamate pesticides block the AChE enzyme at the post-synaptic membrane of the cholinergic synapse in all the vertebrate species [10] and at large dosages, they can cause convulsion, respiratory collapse, and death. In birds, the rate of binding of organophosphate and carbamate pesticides is faster than in any other vertebrates due to the high activity of AChE in the brain [11]. Numerous literature has reported on the hazardous effects of these pesticides on various birds [12-15].

1.3.2 Aquatic biodiversity

Aquatic organisms are severely affected due to pesticide exposure through the dermal route, breathing route, or oral route. Pesticides have extremely negative impacts on aquatic life, through the skin, respiratory system, or mouth. Herbicides lead to lowering oxygen levels, which causes fish to suffocate and decrease fish breeding. Aquatic plants provide approximately 80% of dissolved oxygen, essential for the survival of the aquatic species [16]. Fishes are susceptible to the range of sub-lethal and lethal effects from pesticides, including behavioral alterations, hematological changes, histopathological changes, genotoxicity, disruption of the endocrine system, and acetylcholine activity alteration. Amphibians are mostly impacted by pesticide-polluted surface waterways. As per reports, carbaryl insecticide has been shown to be harmful to a variety of amphibian species. For instance, the herbicide glyphosate has significantly increased tadpole mortality [17].

1.4 In silico estimation of pesticide toxicity and risk assessment

Numerous chemicals that are hazardous to both animal and human health are dispersed into the environment in large quantities. The use of pesticides carelessly and indiscriminately has resulted in alarming consequences for biodiversity and the preservation and restoration of the world's ecology. Due to their long-lasting and bio-accumulative nature, pesticides have both acute and chronic negative impacts on both aquatic and non-aquatic habitats. As a result, it is crucial to ascertain the origin, frequency, harmful effects, and ecological destiny of pesticides in addition to conducting an accurate risk assessment.

Regretfully, determining the ecotoxicity of these pesticides and their environmental transformation products by in vitro experimental validation is an expensive and time-consuming process. A single pesticide may produce many environmental transformation products of pesticides with various end-points, for which numerous in vivo validations are required that are time-consuming and ethically problematic. Various government organizations such as USEPA (United State Environmental Protection Agency), and EFSA (European Food Safety Agency) have given importance to *in-silico* techniques such as QSAR (Quantitative Structure-Activity Relationships), QSTR (Quantitative Structure-Toxicity Relationships), read-across, RASAR and pharmacophore modeling as suitable alternatives for toxicity assessment. This new scientific trend directed us to develop QSAR-based *in-silico* model. We have developed QSAR models to estimate the environmental toxicity of pesticides in response to this new scientific trend. The qualities that have been identified can also help us to combat the toxicity of pesticides against environmentally friendly insects like butterflies and moths, as well as various birds including avian species and aquatic organisms.

1.5 Quantitative structure-activity relationships (QSARs) approach

Similar molecules can display completely various kinds of biological activities or varying intensities of a single biological activity with just a little structural difference. The QSAR study is focused on this type of relationship between molecular structure and biological activity. QSAR is demonstrated as predictive mathematical models derived from the application of statistical tools correlating biological activity (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or property. Both qualitative (basic SAR) and quantitative (QSAR) correlations are possible. QSAR, or QSPR (quantitative structure-property relationship) approaches link a structure of a molecule to a certain activity or property. The most widely used and well-known in silico methodology for screening novel chemical entities is the QSAR, which has extensive application in the area of drug discovery and chemical toxicity modeling for guiding the experimental design of various chemical compounds.All QSAR research is based on the idea that biological activity is a mathematical function (*f*) of structure or physiochemical properties. Therefore, a basic mathematical equation can be developed and represented as follows in **Eq. 1.1.**

$$\textit{Biological activity} = f\,(\textit{Chemical attributes}) = f\,(\textit{Structural, Properties}) \qquad \mathbf{1.1}$$

The phrase "chemical attributes" describes the characteristics that prescribe how a behavior manifests itself, or, to put it another way, the basic knowledge of the chemicals governing the behavior that is being studied. A behavioral manifestation's physiological characteristics, which reflect its biological roots, provide a clear explanation. The QSAR approach is used to determine the structural characteristics of molecules that are associated with their toxicological profiles. The chemical attributes often characterize information derived directly from the structure, whereas physiological information is obtained through experimental methods that result in the corresponding expression, as shown in **Eq. 1.2.**

$$\textit{Response} = f\,(\textit{Chemical attributes}) = f\,(\textit{Structure, physiological Property}) \qquad \mathbf{1.2}$$

QSAR equation for a particular response can be demonstrated mathematically in terms of chemical information and physiochemical attributes as follows in **Eq. 1.3**

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots\ldots + a_nx_n \qquad \mathbf{1.3}$$

As we are discussing a mathematical correlation, such equations are better stated in terms of variables. Here, Y stands for the response that is being modeled such as activity or toxicity or property, whereas, X_1, X_2, X_3…. Xn represents the independent variables that signify the physiochemical properties in terms of numerical quantities and a_0, a_1, a_2, a_3…...a_n stands for the contribution of individual descriptors with a_0 as a constant term. The primary goal of the QSAR analysis is to quantify chemical characteristics, which is followed by the creation of an appropriate interpretive connection that addresses a specific reaction. Therefore, in this case, mathematics acts as a tool to derive an appropriate connection that is subsequently utilized in accordance with the requirements of the designer. A QSAR investigation includes aspects of biology to account for the biochemical interactions involved, mathematics and statistics for modeling and computation, and chemistry and physics to account for the intrinsic molecular nature. Three easy steps, (a) data preparation, (b) data processing, and (c) data interpretation for a collection of chemicals, can be used to display the QSAR analysis. The response, or endpoint, to be addressed, and the predictor, or independent variables (i.e., X variables) describing the chemical attributes, are the two main sources of the quantitative data. The first step, i.e., the data set preparation includes arrangements and conversion of the data in a suitable form. Typically, two types of endpoints are obtained: response-fixed dose patterns, which show the response induced by the chemical at a fixed dose (concentration), and fixed-dose response patterns, which show the quantity of a chemical required to elicit a given response [18].

1.5.1 Application of QSAR/QSTR

Computational methods have developed into invaluable resources for evaluating the ecological toxicity of chemical toxicants in the environment. QSTR modeling is essential for understanding and predicting the possible risks that chemicals may pose to the environment, along with related approaches.

Data efficiency: QSTR modeling has a special advantage by aiding the prediction of toxicological significance even in situations when there is a lack of available or restricted data on the toxicity of a certain chemical. This is especially helpful for determining the possible ecological impact of recently created or insufficiently researched substances.

Cost-effective and time-efficient: QSTR modeling is time and money-efficient since it eliminates the requirement for in-depth laboratory testing and experiments. Without the resource-intensive procedures usually connected with conventional toxicological investigations, it enables researchers to make well-informed estimates and judgments regarding the possible toxicity of substances.

Ethical considerations: The application of QSTR models in testing and research is aligned with ethical standards. Reducing the utilization of animal testing contributes to the protection of laboratory animals' well-being and is in line with current ethical standards in scientific research.

Predictive ability: QSTR models have the ability to assess particular target endpoints or the toxicological significance of novel compounds. When working with compounds that are within the model's applicability domain, this predictive capacity is especially helpful. These models can be used by researchers to calculate the possible hazards connected to these substances.

Mechanistic insights: Mechanistic insights into the relationships between a chemical's structure and activity or toxicity can be obtained using QSTR modeling. This implies that by examining the molecular characteristics of chemicals, researchers might understand why particular compounds display particular toxicological behaviors.

Regulatory recognition: Various regulatory agencies across the globe, including the US EPA, the Agency for Toxic Substances and Disease Registry (ATSDR), the European Centre for the Validation of Alternative Methods (ECVAM) of the European Union, and the European Union Commission's Scientific Committee on Toxicity, Ecotoxicity, and Environment (CSTEE), recognize the importance of QSTR modeling in assessing chemical toxicity.

1.5.2. Significance of QSAR/QSTR

To create new compounds with more activity and reduced toxicity, ligand-based drug design can make use of the QSAR and QSTR models. Predicting the activity and toxicity of novel chemical entities (NCEs) that fit within the developed models' applicability domain is the primary goal of QSAR/QSTR modeling. Although developing a predictive QSAR/QSTR model may appear straightforward, there are many uses for it in the scientific world. Depending on its chemical makeup, even the same chemical substance might occasionally trigger distinct biological reactions and responses. This makes determining the chemical characteristics causing behavioral changes essential. When it comes to model predictability and making the best use of limited experimental resources with less computational capacity, QSAR/QSTR approaches are helpful.

1.6 Concept of molecular descriptor

Molecular descriptors describe particular details about a molecule under study. They are represented as the numerical value associated with the chemical constitution for correlating chemical structure with various physical attributes, chemical reactivity, and biological activity [19]. In other words, the modeled response (activity/ property/toxicity of query molecules) is represented as a function of quantitative values of structural features or properties that are termed as descriptors for a QSAR model as demonstrated in **Eq.1.4**.

$$\textbf{\textit{Response (toxicity) = f (descriptors)}} \qquad \textbf{1.4}$$

The type of descriptors employed and their capacity to represent the structural characteristics of the molecules have a significant impact on the quality of QSAR models. The descriptors can be topological (hydrophobic, steric, or electronic), physicochemical, geometric (based on a molecular surface area calculation), electronic (based on molecular orbital calculations), structural (based on the frequency of occurrence of a substructure), or simple indicator parameters (dummy variables). The summary of the most ideal characteristics that make a descriptor suitable for the construction of QSTR models is as follows:

- The descriptor should match the structural properties of a particular endpoint with negligible correlation to other descriptors.

- A descriptor should be applicable to a wide range of chemicals.
- The descriptor should produce a unique value for molecules with diverse structures, even when there are little structural variations. This suggested that the descriptor should show low degeneracy and continuity, which means that small structural variation should result in slight changes in descriptor value.
- The descriptor should have a clear mechanical interpretation in order to encode the query characteristics of the molecules.
- Another important aspect is the capacity to map the descriptor values back to the structure for visualization purposes. These visualizations are meaningful only when descriptor values can be linked to structural attributes.

Dimension serves as a constraint in QSTR analysis that controls the character of the study. During predictive model generation, the term dimension refers to the complexity of the modeling technique which describes the degree of the descriptors. Thus, the molecular descriptors can be possibly classified on the basis of the dimension as demonstrated in **Table 1.**

Table 1. Different molecular descriptors on the basis of dimension.

Sl. No.	Dimension of the descriptors	Parameters
1	0D-descriptor	Constitutional indices, molecular property, atom, and bond count.
2	1D-descriptor	Fragment counts fingerprints.
3	2D-descriptor	Topological parameters, structural parameters, and physicochemical parameters including thermodynamic descriptors.
4	3D-descriptor	Electronic parameters, spatial parameters, molecular shape analysis parameters, molecular field analysis parameters, and receptor surface analysis parameters.
5	4D-descriptor	Volsurf, GRID, Raptor, etc. derived descriptors.
6	5D-descriptor	These descriptors consider induced-fit parameters and aim to establish a ligand-based virtual or pseudo-receptor model. These can be explained as 4D-QSAR 1 explicit representations of different induced-fit models. Example: flexible-protein docking.

7	6D-descriptor	These are derived using the representation of various solvation circumstances along with the information obtained from 5D descriptors. They can be explained as 5D-QSAR 1 simultaneous consideration of different solvation models.
8	7D-descriptor	They comprise real receptor or target-based receptor model data.

1.6.1 Types of descriptors

Descriptors can be classified into various types depending on the method of their computation, structural (based on substructure occurrence frequency), topological, electronic (involving molecular orbital calculations), physicochemical (encompassing hydrophobic, steric, or electronic aspects), geometric (utilizing molecular surface area calculations), or simple indicator parameters (represented as dummy variables).

1.6.1.1 Descriptor commonly used in QSTR study

The following descriptors pertain to physiological characteristics and are based on certain findings from scientific experiments. Changes in the physiological qualities will also have an impact on adsorption, distribution, and excretion. Important physicochemical characteristics that impact a drug's chemistry and bioactivity include the substituent that is present in the molecule as well as its electronic, hydrophobic, and steric properties.

Commonly used descriptors in QSAR research are explained in a detailed manner as follows:

- **Physiological descriptors**

The physicochemical descriptors pertain to physicochemical characteristics and are based on certain biological experimentation findings. Changes in physiological properties will also affect adsorption, distribution, and excretion. The chemistry and bioactivity of drugs are influenced by a number of significant physicochemical properties, such as the substituent present in the molecules as well as their electronic, hydrophobic, and steric properties [20].

- **Indicator variables**

Indicator variables are used in the QSAR study due to their simplicity in nature. They can represent the presence or absence of specific substructures in molecules. This method is especially helpful when comparing groups of compounds that are similar except for the coded substructure [21].

- **Topological descriptors**

Topological descriptors are computed using a graphical representation of molecules, therefore they do not need the extensive computations associated with quantum chemical descriptors or the estimation of physicochemical attributes. The 2D graphical topology, which shows the bond connections and atom positions, is necessary for the structural representation. It is based on the graph theory, in which the edges of a molecule represent covalent bonds and the atoms are represented by the vertices [22].

- **Structural descriptors**

A variety of characteristics are included in these descriptors, including the number of chiral centers, molecular weight, rotatable bonds, H-bond donors, and H-bond acceptors. They shed light on the structural characteristics of molecules that may affect how they behave [23].

- **Thermodynamic descriptors**

These descriptors, such as AlogP, AlogP98, Alogp_atypes, Fh2o, Foct, and Hf, are extensively used in QSAR model generation to define thermodynamic properties and characteristics of compounds [24].

- **Electronic descriptor**

Electronic descriptors of molecules are described using electronic characteristics, both at the entire molecule level and within specific sections like atoms, bonds, and molecular fragments. Superdelocalizability(Sr), highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, and the sum of atomic polarizabilities are a few examples [25].

- **Quantum chemical descriptors**

These descriptors include Mulliken atomic charges and Quantum Topological Molecular Similarity (QTMS) descriptors, which focus on bond critical points (BCPs) and their relevance in chemical reactions [26].

- **Spatial descriptor**

These descriptors are calculated based on the spatial arrangements of the molecules and the surface occupied by the molecules. Examples of this class of descriptors include radius of gyration, Jurs descriptors, shadow indices, molecular surface area, density, principal moment of inertia, and molecular volume [27].

- **Information indices**

This method divides molecules according to certain characteristics into subsets of equivalent elements. This category comprises many indices such as atomic composition index, indices based on the A-matrix, D-matrix, E-matrix, and ED-matrix, as well as multigraph information content indices (IC, BIC, CIC, SIC) [28].

- **Molecular shape analysis descriptors**

These descriptors, such as Difference volume (DIFFV), Common overlap steric volume (COSV), Common overlap volume ratio (Fo), Noncommon overlap steric volume (NCOSV), and Root mean square to shape reference (ShapeRMS), are utilized for QSAR model development [29].

- **Molecular field analysis descriptors**

Molecular field analysis (MFA) estimates probe interaction energies on a grid around a bundle of active molecules. Fields are represented using grids, and each energy value at a grid point can be used as a QSAR descriptor [30].

- **Receptor surface analysis descriptors**

Molecular models and receptor surface models interact through interaction energies, which are used as descriptors. These descriptors capture 3D information of interaction energies, considering steric and electrostatic fields at each surface point of the receptor surface [31].

1.7 Commonly employed QSAR/QSTR methods for chemometric model development

The main aim of the QSAR/QSTR research is to develop correlation models that utilize chemical information data and the response of the chemicals (toxicity) within a statistical framework. Regression and classification-based approaches are employed for model generation. In addition to conventional methods, some machine learning techniques are helpful in QSTR/QSAR model development, particularly while working with high dimensional and complex information data that may show nonlinear relationships with response variables [32].

1.7.1 Classification of QSAR/QSTR approaches based on the type of chemometric methods used

1.7.1.1 Linear methods

1.7.1.1.1 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is a commonly used approach in QSAR/QSTR model generation as MLR is a transparent, easy to interpret, simple, and reproducible approach. The generalized form of an MLR equation can be represented as follows in **Eq. 1.3:**

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots + +a_nx_n \qquad \textbf{1.3}$$

In the above equation, Y is the response (dependent variable), and the x_1, x_2, …x_n are descriptors (independent variables) in the model with their corresponding regression coefficient a_1, a_2, …a_n respectively, and a0 is the constant. During the interpretation, the individual descriptors (x_1, x_2,

…x_n) directly depend upon the corresponding value and its algebraic sign. Each regression coefficient should be significant at $p < 0.05$ which can be verified by performing the 't' test. The descriptors present in an MLR model should not be intercorrelated [33].

1.7.1.1.2 Partial least squares (PLS)

When a small dataset contains a large number of noisy and intercorrelated descriptors, partial least squares (PLS) is a better choice as compared to MLR [34], looking for latent variables (LVs) that are functions derived from the original variables. The latent variables aim to capture as much of the underlying factor variation as possible while simultaneously modeling the response.

Linear PLS identifies a set of new variables (LVs) which are linear combinations of the original variables. When the number of latent variables is the same as the number of variables, the PLS essential becomes equivalent to the MLR model. It is important to determine the predictive significance of each PLS component and stop the addition of new components when they are found to be statistically significant. Cross-validation is a frequently used and reliable method for testing the predictive significance. The application of PLS allows the generation of larger QSAR/QSTR models by avoiding overfitting and eliminating most variables.

1.7.1.1.3 Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is an effective method for differentiating between two or more classes of objects, making it a useful tool for classification issues. LDA shares a common goal with MLR when dealing with scenarios where the response variable has categorical values and the molecular descriptors are continuous variables.

LDA mainly aims to model the distinctions between various data classes. The generalized form of the LDA equation is as follows:

$$DF = c_1 \times X_1 + c_2 \times X_2 + \cdots + c_m \times X_m + a \qquad \textbf{1.5}$$

Where, DF represents the discriminant function, which is formed by a linear combination of the discriminating variables. The 'c' represents the discriminant coefficient or weight for that variable, 'X' denotes the score of the respondent on that variable, 'a' denoted as constant, and 'm' indicates the total number of predictor variables. These 'c' coefficients are unstandardized and can be considered as similar to the beta coefficients in a regression equation. They are chosen in order to maximize the separation between the means of criterion (dependent) variables. Normally strong predictors tend to have a large weight. Once the DF is calculated using an existing dataset to classify cases, it is possible to classify new cases (test samples). In a step-wise DF analysis, the model is generated. At every stage, all variables are assessed to find out which one has the biggest impact on the discrimination between groups. The selected variable is then included and the procedure is repeated once more.

1.7.1.1.4 Cluster analysis

Cluster analysis is a tool for exploratory data analysis, used for organizing observed data or cases into two or more categories. In contrast to LDA, cluster analysis does not necessitate any prior knowledge of which elements belong to which clusters. The clusters are defined by an analysis of the data. Cluster analysis maximizes the similarity of cases within each cluster while maximizing the dissimilarity between previously unknown groups. Cluster analysis includes two approaches of analysis;

- **Hierarchical cluster analysis**

Hierarchical cluster analysis detects relatively homogeneous clusters of cases by estimating dissimilarities or distances between objects the most commonly used methods for calculating the distances in a multidimensional space include either Euclidean distances or squared Euclidean distances between objects. Each case is first treated as an individual cluster and then gradually merges these clusters, reducing their count with each step until only one cluster is left. Hierarchical tree diagrams or dendrograms can be generated to show the connection points visually and show how clusters are connected at different dissimilarity levels [34].

- ***k*-Means clustering**

k-Means clustering is a non-hierarchical clustering method used when the number of intended clusters within the objects or cases is known. It functions as a centroid-based, unsupervised clustering technique. Essentially, the *k*-Means algorithm produces exactly k unique clusters. The first step in this process is to create k centroids, one for each cluster, and place them as far apart from one another as possible. The closest centroid is then assigned to each data point in the dataset. As this assignment occurs for all data points, the positions of k centroids are recalculated. The technique is repeated until the centroids no longer move significantly [35].

1.7.1.2 Non-linear methods

1.7.1.2.1 Artificial neural networks (ANN)

It is a computational approach inspired by natural neurons. Artificial neurons are simple tools that are highly interconnected and the connections between neurons transfer the function of the neuron. An artificial neural network creates an empirical relationship between the input variable, also known as independent variables or descriptors (X), and output variables, also known as dependent variables also known as responses(Y), without relying on prior information [36]. The network can be represented by the equation: Y = f(X)+ e. Each neuron, serving as a processing unit, receives stimuli from other neurons via dendrites and transmits stimuli to other neurons through its axon. The strength of the connections between neurons is stored as weight values, and these specific connections are termed synapses. Within a neural network, information is distributed across multiple cells (nodes) and the connections between them, referred to as synapses (weights). The activation signal transforms a function to yield the neuron's output, expressed as Y = f (a). This transformation function can take on various forms, including linearity or non-linearity, such as threshold or sigmoid functions.

1.7.1.2.2 *k*-nearest neighbour method (kNN)

The aim of supervised learning is to establish a classification rule using a set of training objects of known origin. By using this rule, new objects with unknown origins can be categorized into one of the specified classes according to their variable values [37].

The supervised learning process is carried out in several phases. First, a training set is carefully curated, consisting of objects with well-defined classifications and associated features. Consequently, a careful selection of relevant variables for classification takes place, while non-discriminatory or less significant variables are eliminated. Then, a classification rule is formulated using the training set. The efficacy of this classification rule is assessed using an independent test set for validation. There are several clustering techniques that can be used in the process of variable selection. One approach includes organizing the original data in a transposed matrix format, where descriptors occupy rows, and molecules are arranged in columns. From each cluster, one or more representative descriptors are chosen. These methods establish the classifier by evaluating the distances between each object in the training set and approximate functions locally based on neighbouring data points. Typically, Euclidean distance is widely used, although other distance metrics can also be applied. Correlation-based measures are favoured when dealing with strongly correlated variables. For a training set comprising 'n' objects, 'n' distances relative to a test sample are computed, and the closest distance is used to determine class membership. The k-nearest neighbor method (kNN) represents a non-parametric and unbiased approach with versatile applications in both classification and regression tasks.

1.7.1.2.3 Read Across

Read across acts as a non-testing strategy for bridging data gaps by extrapolating toxicological insights from the known toxicity data of compounds exhibiting analogous properties or chemical profiles [38]. It is used in toxicological assessments, where predictions are made within a grouping framework such as the analogue or category approach involving either qualitative or quantitative prediction. In this methodology, the known toxicity data of a chemical, referred to as the "source" chemical, are leveraged to predict the same endpoint or test outcome for another chemical, termed the "target" chemical, which shares scientific similarities. The category approach is based on a group of chemicals with comparable physico-chemical, human health, environmental toxicological, or environmental fate properties, often resulting from structural similarity. Conversely, the analogue approach centres on a smaller subset of closely related substances, typically a target and source substance [39]. Read-across depends on on structural similarity and similar properties or activities between the source and target chemicals. This assessment considers factors such as structure, composition, physical-chemical properties, reactivity, metabolism, and mechanistic similarity. Source analogues are identified based on searches for structurally related compounds, utilizing similarity metrics, or by evaluating structural alerts, potential metabolic precursors, or chemical classes.

1.7.1.2.4 Quantitative read-across structure-toxicity relationship (q-RASTR)

QSTR and read-across approaches have merged to form an emerging method known as Read-across structure–toxicity relationship (RASTR). This approach utilizes the chemical similarity principles of read-across, as an unsupervised step, and later develops into a supervised learning model similar to QSAR [40].

In this approach, a combination of similarity-based and error-based descriptors was employed. This method exhibited superior predictive capability and lower Mean Absolute Error (MAE) as compared to both QSTR and Read across predictions. The effectiveness of the q-RASTR approach relies on its ability to integrate similarity and error measurement information into descriptors. This integration enables the generation of interpretable, transferable, reproducible models with enhanced predictive accuracy [41].

1.7.1.2.5 q-RASTR descriptors

Based on the fundamental principle of read-across, compounds with similar chemical structures are anticipated to exhibit analogous characteristics, commonly known as similarity between the source and target substances. This similarity can result in comparable toxicokinetic and toxicodynamic behaviors. This principle is rooted in a non-statistical methodology and does not depend on mathematically complex models to make predictions of desired chemical compounds. Three distinct techniques are used to estimate compound similarity such as; Gaussian kernel similarity, Euclidean distance, and Laplacian kernel similarity [42].

The RASAR descriptor RA function (LK) is a prediction function produced from read-across by averaging the response values of source compounds, created by averaging the response values of source compounds identified as having structurally analogous properties [43]. The SD activity descriptor represents the weighted standard deviation of activity near n source compounds for a specific target compound. SE is defined as the weighted standard error allied with the activity values of the adjacent n-source compounds for a given target compound. The CVact descriptor characterizes the coefficient of variation of the activity values among the nearest n-source compounds for a specific target compound. MaxPos defines the maximum similarity score between the training set and the target compound. MaxNeg signifies the degree of similarity between a target compound and a close source compound with an activity response value lower than the mean response of the training set.

The absolute difference between MaxPos and MaxNeg for a particular query molecule is demonstrated as Abs Max Pos-Max Neg or Abs Diff. The AvgSim descriptor calculates the similarity mean value among n closely associated compounds for a definite target compound. The gm (Banerjee-Roy coefficient) descriptor estimates the possibility of whether the query compound is active or inactive, with ranging values from -1 to +1. gm*Avg. Sim and gm*SD_Similarity descriptors are found by multiplying gm values with Avg. Sim and SD_Similarity values, respectively. Pos.Avg.Sim defines the average similarity values among the n close source compounds with response values higher than the mean response value of the training set, on the other hand, Neg.Avg.Sim represented as the average similarity values among the n close source compounds with response values lower than the mean response value of the training set [44].

1.7.2 Classification of QSAR/QSTR approaches based on of dimensionality

Table. 2. Classification of the QSAR methodologies on the basis of dimensionality.

Dimension	Methods
0D-QSAR	Models are based on descriptors involving molecular formulas like molecular weight etc.
1D-QSAR	Models are based on the simplex representation of molecular structure (SiRMS) approach.
2D-QSAR	Activity is correlated with physicochemical and structural patterns (connectivity, topology, etc.) of the molecules without consideration of an explicit 3D representation of these properties.
3D-QSAR	Activity is correlated with the three-dimensional structure of the ligands
4D-QSAR	Ligands are represented as an ensemble of configurations
5D-QSAR	As 4D-QSAR + explicit representation of different induced-fit models
6D-QSAR	As 5D-QSAR + simultaneous consideration of different solvation models

1.8 Development of quantitative models over the period

A timeline showing the various approaches that are developed over the period of time to focus on the key molecular structural attributes. Therefore, QSAR methods originated way back in the nineteenth century

1.8.1 De novo design

The De novo QSAR model is a collaborative mathematical model that may encode any molecular information without the need for a descriptor. The models are generated using indicator parameters (binary values 0 or 1) to indicate the presence or absence of groups at specific positions.

i. Hansch's method

In 1962 Hansch et al correlated the Hammett constants and partition coefficients of phenoxyacetic acid with the growth regulatory activity of plants [45]. Two years later they demonstrated that biological activity could be correlated with free energy-related terms linearly. Previously called Linear Free Energy Relationship (LFER), later evolved into an extra thermodynamic approach as expressed by **Eq 1.6.**

$$\log 1/C = a\pi + b\sigma + cEs + \cdots + \text{constant} \qquad \textbf{1.6}$$

Where π = hydrophobic parameter

σ = Hammett electronic descriptor of the substituent

Es = Taft steric constant

a, b, c = appropriate constants

ii. Free Wilson model

Free Wilson approach is genuinely a structure activity-based methodology considering the contributions that every structural component provides to the whole biological process. This model was represented as follows in **Eq 1.7**

$$\llbracket BA \rrbracket\ \_i = \Sigma a\_j\ X\_ij + \mu \qquad \textbf{1.7}$$

Where, μ = overall average biological activity.
BA = biological activity,
a_j = contribution of the j th substituent to biological,
X_j = j th substituent, which carries a value 1 if present, 0 if absent
This de novo approach assumes that the effects of substituents are additive and constant. This approach does not need of physicochemical constant. However, there are certain limitations. The large number of variables is required to describe a smaller number of compounds together with a large number of molecules with varying substituents. Besides, these intra-molecular interactions are not handled well. The constant term (μ) is an overall average of the biological activity of all the compounds used to develop the model.

iii. Fujita Ban model

Fujita Ban modifies the approaches of the Free-Wilson model. In this approach, the biological activity data was expressed in a logarithmic scale. It is a Free-energy-related approach and additive in nature. This model is represented in **Eq 1.8.**

$$log\ A/Ao = \Sigma\ Gixi \qquad \textbf{1.8}$$

Here, A and A_0 are the magnitudes of the activity of substituted and unsubstituted compounds respectively. G_i is the log activity contribution or the log activity enhancement factor of the i^{th} substituent relative to that of H and X_i is the parameter that takes a value 1 or 0 according to the presence or absence of the i^{th} substituent.

1.9 The methodology of QSAR/QSTR model generation

The development of predictive QSAR models consists of various steps such as

1) Dataset preparation
2) Data analysis
3) Data validation
4) Interpretation of data where the "data" relates to the response and predictor variables.

The steps are briefly discussed one by one as follows,

1. **Data preparation**

- The physiological/biological/toxicological response is converted to the respective unit and maintains data consistency.
- Then, drawing of the chemical structures using suitable drawing software like ChemSketch, ChemDraw, Marvin-Sketch, etc. The chemical structures can also be downloaded/collected from public databases such as NIST Chemistry, and PubChem. The configuration should be checked before using the structures.
- Energy minimization operation and conformational analysis should be performed depending on the purpose of modeling.
- A file containing the structure is subjected to software used to calculate descriptors. Initially, data pretreatment was performed to eliminate the intercorrelated descriptors and the constants. Various software can be used for the descriptor calculation.
- There is a single worksheet with different descriptors for each variable and a single column of response (activity/ property/toxicity) that represents all the variables in the QSAR matrix. An additional column representing the name of the chemicals can be added for the quick identification of any compound.

2. **Data analysis**

This phase consists of feature selection, dataset division, and model development.

- The selection of features refers to the identification of the important predictor variables suitable for developing a correlation with the response variable suitable for developing a correlation with the response variables. Usually, various feature selection tools are coupled with one or more model generation methods under the same interface so that the user can select the best predictor variables and simultaneously construct the models using them.

Many applications can generate hundreds or thousands of various molecular descriptors. In chemometric modeling studies, various feature selection tools are performed which include stepwise variable selection, genetic algorithm, best subset selection, variable subset selection, and factor analysis. Typically, only some of them are significantly correlated with the activity. Furthermore, many of the descriptors are inter-correlated. This has negative effects on several aspects of QSAR analysis.

- Some statistical methods require that the number of compounds is significantly greater than the number of descriptors. Using large descriptor sets would require large datasets.
- Selection of the training set chemicals is important in QSTR analysis. According to chemical similarity, the entire dataset is divided into a training set and a test set for the prediction model. The training set (i.e., the equation), while the test set (not used during model development) is used to judge the external predictivity of the model. However, physicochemical descriptors and the chemical similarity principle are the most rational means to select training sets. A higher number of training set chemicals is used in the development of the model. This method is based on the assumption that a molecule structurally very similar to the training set molecules will also be predicted well by the model since the model captures features that are common to the training set molecules and can identify them in the new molecule. It is important to choose the training and test sets in such a way that the test set chemicals fall within the structural domain of the training set chemicals. Otherwise, the model developed using the training set will not be able to make accurate predictions. The methods for the selection of training and test set are as follows;
- *k*-Means clustering and Kennard-Stone selection
- Kohonen's Self-Organizing Map (SOM)
- Principal component analysis (PCA)
- D-optimal design
- Sphere exclusion
- Sorted response

Here, the whole data matrix is first sorted based on the response column followed by a selection of a predefined fraction of chemicals into a training/ test set from different zones maintaining a pattern e.g., every first/second/third/fourth compound, etc. In the random division approach, chemicals are arbitrarily divided into training and test sets following a user-defined fraction.

Sometimes, a combination of response variable-based and predictor variable-based approaches may also be employed e.g., chemicals may be assigned into different structurally similar groups using any of the above-mentioned techniques followed by a selection of chemicals into training/ test set using the sorted response formalism separately from each group.

The model development step dictates that the selected best features are to be combined in a single equation employing an explicit formalism. After the calculation of different features, i.e. descriptors, the construction of the QSTR model is done by using a feature mapping procedure also referred to as the parameter estimation problem. The aim is to build a pure mathematical relationship between the response and the descriptors under investigation. Partial least squares (PLS), multiple linear regression (MLR), etc. are the algorithms used for the development of quantitative regression-based equations while linear discriminant analysis (LDA) generates the classification-based model.

The variable selection tools are accompanied by statistical evaluation of the corresponding model developed from the selected variables as stepwise-MLR, GFA-MLR, G/PLS (genetic PLS), PLS-DA (PLS followed by discriminant analysis), etc.

3. **Model validation**

Determination of statistical reliability becomes the next essential task during the development of predictive models. As the purpose of QSTR analysis isn't simply to develop a model, but also to predict the response of untested/new chemicals, it's important to check for its predictability and stability. Various statistical metrics are calculated to determine the model fitness (R^2, $R^2_{a,}$ etc.), internal stability (Q^2_{LOO}, $r_m{}^2{}_{(LOO)}$) as well as external predictivity ($r_m{}^2{}_{(test)}$, $R^2{}_{(pred)}$), and the values above the threshold limits identify model acceptability. Training set chemicals are used to predict the internal validation (internal stability) only i.e., chemicals used for developing the model, while external predictivity (external validation) refers to the judgment on test set prediction. Some additional validation metrics can also be used to determine the overall predictivity e.g. $r_m{}^2$. For the validation of discriminant model parameters such as specificity, sensitivity, precision, F-value, accuracy, receiver operating characteristic (ROC) analysis, etc. can be employed.

4. **Model interpretation**

Once a QSAR/QSTR model has been developed and considered acceptable from the values of the metrics, the final important part remains with the mechanistic interpretability of the modeled features. Establishing a suitable basis between the chemistry of the chemicals and biological/

toxicological action or physicochemical property helps in understanding the mechanism of action involved. Accordingly, by combining the experimental results and observation from the model, one can explicitly explain each step of the process of behavioral manifestation of chemicals. Such knowledge is useful in designing and developing potent analogues.

1.10 Application of QSAR/QSTR

QSAR presents a suitable option in the rational monitoring of activity/ property/toxicity of chemicals and hence is useful in a wide variety of applications namely biological activity, predictive toxicity, and physicochemical property. Fine-tuning the behavioral nature of chemicals gives fruitful results for a significantly large class of chemicals such as:

- ✓ Pharmaceuticals
- ✓ Agrochemicals
- ✓ Perfumeries
- ✓ Analytical reagents
- ✓ Solvents
- ✓ Surface modifying agents etc.

The chemicals modelled using the QSAR method can be overviewed in three major types, namely:

- ➢ Chemicals of health benefits (drugs, pharmaceuticals, food ingredients, etc.),
- ➢ Chemicals involved in industrial/laboratory processes (solvents, reagents, etc.)
- ➢ The chemicals posing hazardous outcomes are persistent organic pollutants (POPs), toxins, xenobiotics, and volatile organic chemicals (VOCs).

Besides modeling biological activity and toxicity endpoints, it may also be involved in the modeling of ADME which involves in pharmacokinetics profile of drug candidates before its synthesis and hence enhances the efficacy of the designed drug in a biological system. QSTR modeling can be a very good option to predict chemical responses using limited resources in any prospective discipline. Hence, we can see that the simple ideology of QSPR, i.e., the development of a suitable mathematical correlation between the chemical attributes and a response of interest, can be of significant application to serve the human community. QSAR/QSTR plays an encouraging role in achieving this environmental greenness through the design and development of process-specific chemicals with reduced or null hazardous outcomes.

1.11 Computation of different statistical metrics for assessing model quality

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.1.9**.

$$R^2 = 1 - \frac{\Sigma\left(Y_{obs(train)} - Y_{calc(train)}\right)^2}{\Sigma\left(Y_{obs(train)} - \overline{Y}_{train}\right)^2} \quad \textbf{1.9}$$

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 [46]. The R^2 bears a value between zero (no correlations) to one (perfect correlation). A model possessing a value of R^2 more than 0.8 can be considered to elicit an acceptable correlation while the quality enhances with the increasing value of R^2 until it reaches a maximum value of unity (which is unusual in real cases). Yobs and Y_{calc} are the respective observed and calculated values of the response variable. 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 lie in the fact that it does not provide any information on whether:

- The independent variables are a true cause of the changes in the dependent variable,
- The correct regression was used,
- The most appropriate set of independent variables has been chosen,
- The model might be improved by using transformed versions of the existing set of independent variables and
- Whether any collinear ties exist in the data or not.

Adjusted R^2_a **(Eq. 1.10)** 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 the model and can be defined as follows [47]. Here, R^2 is the determination coefficient of a QSTR 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.

$$R_{adj}^2 = \frac{(n-1)\times R^2 - p}{(n-p-1)} \qquad \textbf{1.10}$$

In **Eq.1.10**, 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.

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 the fitting. Lower values of SEE correspond to improved model acceptability as shown in **Eq. 1.11**.

$$S = \sqrt{\frac{\Sigma(Y_{obs}-Y_{calc})^2}{n-p-1}} \qquad \textbf{1.11}$$

Here, 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.

1.11.1 Validation metrics for the training set

1.11.1.1 Q_{LOO}^2

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 $R^2(Q^2)$ and predicted residual sum of squares (PRESS) and the acceptable models have been further processed for the prediction of toxicity and/or property of the test set compounds. The cross-validated correlation coefficient R^2 (LOO$-Q^2$) is calculated according to the formula.

$$Q_{LOO}^2 = 1 - \frac{\Sigma\left(Y_{obs(train)}-Y_{calc(train)}\right)^2}{\Sigma\left(Y_{obs(train)}-\overline{Y}_{train}\right)^2} \qquad \textbf{1.12}$$

Here $Y_{obs(train)}$, $Y_{pred(train),}$ and $\overline{Y}_{train}$ 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 a model is built using the rest of the compounds. The model thus formed in this way is used for the prediction of the 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 R^2 (Q^2) for the model is determined. The acceptable value of Q^2 is 0.5 with a maximum value of 1.0 and hence more the value i.e. closer to 1, the more the internal predictivity of the model.

1.11.1.2 Root mean square error in prediction for the training set (RMSEp)

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 as shown in **Eq. 1.13.**

$$RMSE_p = \sqrt{\frac{\sum\left(Y_{obs(train)} - Y_{calc(train)}\right)^2}{n_{test}}} \qquad 1.13$$

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

1.11.1.3 The r_m^2 metrics

Using the concept of regression through the origin approach [47], introduced a new parameter r^2 or modified that penalizes the R^2 value of a model with respect to an ideal condition [47]. The r^2 metrics can be defined as follows in **Eq. 1.14 and 1.15;**

$$r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right) \qquad 1.14$$

$$r_m'^{\ 2} = r^{\ 2} \times \left(1 - \sqrt{(r^{\ 2} - r_0'^{\ 2})}\right) \qquad 1.15$$

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 [48] further defined the average and difference of the two r^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 in **Eq. 1.16**.

$$\overline{r}_m^{\ 2} = \frac{(r_m^{\ 2} + r_m'^{\ 2})}{2} \qquad 1.16$$

$$\Delta r_m^{\ 2} = |r_m^{\ 2} - r_m'^{\ 2}| \qquad 1.17$$

The r_m^2 metrics can not only be computed for the test set compound ($r^2_{m\,(test)}$) 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 ($r^2_{m\,(LOO)}$) of the training set observations are used against their observed response. Furthermore, Roy et al. [49] 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^2). Later, a rank-based r^2 [48], as well as a scaled [50] version of the r^2 metric, was introduced by the same group of authors and these have been used in this present study.

1.11.2 Validation metrics for the test set

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

For the prediction of toxicity and/or property of the test set compounds, this parameter was calculated. It can be defined as in **Eq. 1.18**.

$$Q_{F1}^2 = 1 - \frac{\Sigma\left(Y_{obs(test)} - Y_{calc(test)}\right)^2}{\Sigma\left(Y_{obs(test)} - \overline{Y}_{train}\right)^2} \qquad \textbf{1.18}$$

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 Ytrain 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).

1.11.2.2 $Q^2_{(F2)}$

This function as a metric for external set validation was described in the paper of Hawkins [51] and can be calculated as in **Eq. 1.19**.

$$Q_{F2}^2 = 1 - \frac{\Sigma\left(Y_{obs(test)} - Y_{calc(test)}\right)^2}{\Sigma\left(Y_{obs(test)} - \overline{Y}_{test}\right)^2} \qquad \textbf{1.19}$$

The only notable difference from $Q^2_{ext\ (F1)}$ is that the average value of the external or test set is used in the denominator instead of the internal or training set average value. Both $Q^2_{(F1)}$ and $Q^2_{\ (F2)}$ were compared and discussed [51]. The threshold value of acceptance for all three parameters $Q^2_{(F1)}$, $Q^2_{(F2)}$, and $Q^2_{(F3)}$ is 0.5.

1.11.3 Y-randomization study

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

- Process randomization

- Model randomization

In process randomization, random scrambling of the dependent response variables is performed accompanied by a 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 the model randomization of the genetic models using SIMCA software. A parameter was proposed by Roy and Paul named R_p^2 that penalizes the model R^2 for a small difference between the squared mean correlation coefficient (R_r^2) of randomized models and the squared correlation coefficient (R^2) of the non-randomized model and was defined as in **Eq. 1.20**.

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \qquad \textbf{1.20}$$

The acceptable value of CR_p^2 was proposed to be greater than or at least equal to 0.5. Later a correction for this parameter has been suggested [52] and the rebuilt formula is as follows in Eq. **1.21.**

$$CR_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \qquad \textbf{1.21}$$

1.11.4 Determination of model applicability domain (AD)

The applicability domain (AD) of a QSTR model can be described as the theoretical region in the chemical space defined by the chemical as well as the response attributes of the model [53]. A definite domain of applicability enables the reliability of the predictive performance of a model. In other words, any QSTR model possesses a defined theoretical domain within which it can provide reliable predictions of other chemicals not used in developing the model. It is not feasible to develop a single model that can contain the chemical information of the whole universe, and accordingly, QSTR models are characterized by different domains. The applicability domain [54] is a theoretical region in chemical space, defined by the model descriptors and modeled response. When a compound is highly dissimilar to all compounds of the modeling set, reliable prediction of its property is unlikely. The concept of AD was used to avoid such an unjustified extrapolation of property predictions. Here, we have applied both the Leverage approach and Distance to model in X-space (DModX) approach for verifying the applicability domain of the best model developed from this study [55].

1.11.4.1 Applicability domain: Standardization approach

The equation to calculate AD is:

$$S_{ki} = \frac{|X_{ki} - X_i\ |}{\sigma X_i} \tag{1.22}$$

Where, k=1, 2, 3 … n_{Comp} (here, n_{Comp} = total number of compounds)

i= 1, 2, 3 … n_{Des} (here, n_{Des} = total number of descriptors)

S_{ki} = Standardized descriptor i for compound k (from the training or test set)

X_{ki} = original descriptor 'i' for compound 'k' (from the training or test set)

X_{ki} = mean value of the descriptor

Xi= for the training set compounds only

σX_i=standard deviation of the descriptor

Xi for the training set compounds only

The standardization approach of the applicability domain **(Eq. 1.22)** is based on the ideal data distribution; 99.7% of the compounds would stay within the range of mean ± 3 standard deviations (SDs). As a result, this range (i.e., mean ± 3SDs) is considered as the area of the majority of the training set compounds. Outside this area, a compound is examined as diverse from the rest of the compounds. So, one should compute the maximum Si(k) value ([Si] max(k)) for the compound k. If the SD value for descriptor i of compound k (Ski) is greater than 3 then the compound is an X-outlier (if it is in the training set) or outside the AD (if it is in the test set) [56].

1.12 Literature review

There are innumerous studies for the prediction of pesticide toxicity (LD_{50}) against different avian species using QSAR approaches have been reported. In 2006, Mazzatorta et al. reported a classification-based QSAR study using the Support Vector Machine (SVM) technique for the estimation of oral toxicity of pesticides against Bobwhite Quail (Colinus virginianus) [57]. In 2015, Basant et al., developed QSAR models using different tree-based modeling approaches like Single Decision Tree (SDT) QSAR, Decision Tree Forest (DTF) QSAR, and Decision Tree Boost (DTB) QSAR to determine the acute oral toxicity of pesticides on multiple avian species, for example, Bobwhite Quail (*Colinus virginianus*), Mallard Duck (*Anas platyrhynchos*), Ring-necked Pheasant (*Phasianus colchicus*), Japanese Quail (*Coturnix japonica*) and House Sparrow (*Passer domesticus*) [58]. In 2017, Halder et al. developed a QSTR model using the Monte Carlo method to predict the acute oral toxicity of some diverse agrochemical pesticides against Bobwhite Quail (*Colinus virginianus*) [59]. In 2020, Kar & Leszczynski developed partial least square (PLS) regression-based individual and intraspecies QSTR models to estimate the acute oral toxicity of certain pesticides in Bobwhite Quail (*Colinus virginianus*), Mallard Duck (*Anas platyrhynchos*) and Japanese Quail (*Coturnix japonica*) [60]. In recent work, Banjare et al. developed classification-based predictive QSTR models for the estimation of acute oral toxicity of pesticides on three different avian species namely Bobwhite Quail (*Colinus virginianus*), Mallard Duck (*Anas platyrhynchos*) and Zebra Finch (*Taeniopygia guttata*) [61]. In 2022, Mukherjee et al. generated a regression-based 2D quantitative structure toxicity relationship (2D QSTR) and quantitative structure toxicity–toxicity relationship (QSTTR) models to predict the toxicological significance of pesticides on five different avian species [62]. Recently podder et al. also developed regression-based QSTR and i-QSTR models for toxicity assessment of pesticides on various avian species, such as mallard duck, bobwhite quail, and zebra finch [63].

CHAPTER - 2

Present work

2. PRESENT WORK

2.1. Study 1: Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across

Pesticides comprise a diverse class of chemicals that are commonly used to control or eliminate pests such as weeds, fungi, insects, and rodents effective crop management. In recent decades, there have been a significant surge in the usage of pesticides, particularly in developing nations that rely on agriculture. [64]. Due to the inherent characteristics, a significant fraction of the applied dose persists as residues on crops and fields [58]. As a result, large concentrations of pesticides have been found in crops, vegetation, and further edible products causing exposure to both animals and humans. According to the reports, prolonged exposure to these substances can cause adverse effects on the neurological, endocrine, reproductive, immunological, cardiovascular, renal, and respiratory systems of an individual [65].

In light of the aforementioned, various regulatory authorities have emphasized the need for the toxicity evaluation of both new and existing pesticides. The avian toxicity tests are essential for regulatory approval and licensing of the active ingredients of pesticides. Aves are significant for ecology and have a huge contribution to biodiversity by performing pollination of plants, rodent control, seed dispersal, and spreading nutrients [62]. According to today’s scenario, one in every eight bird species faces extinction [66]. Therefore, birds are used as a model organism to evaluate toxicity.

Oral toxicity testing is important for determining the toxicological significance of various avian species. Northern bobwhite quail (*Colinus virginianus*) [BQ], Japanese quail (*Coturnix japonica*) [JQ], ring-necked pheasant (*Phasianus colchicus*) [RNP], and mallard duck (*Anas platyrhynchos*) [MD] are the major test species as per OECD norms [67]. The validated wet-lab techniques for the evaluation of compound toxicity towards avians are expensive, unethical, and require a significant amount of time and effort. So the relevant regulatory bodies encourage the employment of potential alternative strategies to achieve the objective. Regulatory agencies like the Environmental Protection Agency (EPA), European Food Safety Authority (EFSA), Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), and European Chemicals Bureau (ECB), have emphasized the potential of computational tools like QSTR, Read-Across, and alternative approaches for investigating the inherent characteristics of chemicals within the realm of toxicokinetics [68].

Some alternative in silico-based approaches were reported previously that offer significant improvements over single-output models for regulatory purposes [69-74]. Speck-Planche et al.

[69] reported the discriminant model based on substructural descriptors for the rational design of new agrochemical fungicides. Speck-Planche et al. [70] also worked on new in-silico methods for the rational design of new insecticidal agents. Speck-Planche et al. [71] also reported the multi-species chemoinformatic methods for assessing the various ecotoxicological profiles in agrochemical fungicides. Speck-Planche et al. [72] also published a work regarding multi-scale QSAR methodology for simultaneous ecotoxicological modeling of pesticides. Jiang et al. [73] worked on boosting tree-assisted multitask deep learning methods for small scientific datasets. a consensus multitask deep learning method was used to model multispecies acute toxic effects by Jain et al. [74]. Even other alternative modeling approaches based on machine learning (ML) tools that have demonstrated significant advancements, particularly in handling nonlinearity aspects and improving predictions were also reported earlier [73-76]. Halder et al. [75] reported the global models employing in-silico methods for Predicting the ecotoxicity of endocrine disruptive chemicals. Samanipour et al. [76] worked on alternative methods for chemical prioritization using molecular descriptors and intrinsic fish toxicity of chemicals

These *in-slico* techniques examine significant structural features that are essential for predicting the biological activity, toxicity, and other characteristics of untested substances. Several research teams published in silico predictions of acute oral toxicity in various species, including rats, mice, and fish [61,77-80]. But in the case of avian oral toxicity, very few in-silico reports are available [58, 61, 62, 63, 66,81].

Herein, we developed QSTR models to interpret the major structural and physicochemical features responsible for their toxicity followed by estimating the toxicity of external datasets in RNP and MD avian species following the OECD guidelines strictly [67]. Alternative tools, such as read-across, are widely used for hazard assessment to fill data gaps. The Read-Across-based predictions assume that a molecule with an unreported experimental endpoint value should have a value similar to molecules that are structurally and/or biologically similar to the query molecule. So, we have conducted the Read-across predictions to improve the test set results. The main motive for choosing the regression-based QSTR approach over others (e.g.: regarding its effectiveness, coping with chemical heterogeneity, and several different species) was to develop a linear relationship between the descriptors and this defined endpoint (pLC_{50}) and to identify the important features responsible for toxicity towards avian species (RNP and MD) as well as data-gap filling. Classification models only focus on the categorical relationship between the input and output variables rather than the exact numerical relationship. On the other hand, regression models can identify the most important features or predictors driving the outcome variable. Additionally, we have also developed classification models as well as employed two different ML algorithms namely SVM, and RF to evaluate their effectiveness in model construction and prediction The present work aimed to design a logical method to assess pesticide toxicity towards avians. Furthermore, screening of the Pesticide Properties DataBase (PPDB) was conducted to evaluate the avian toxicity following the prediction reliability assessment of the QSTR models by the PRI (prediction reliability indicator) tool (http://teqip.jdvu.ac.in/QSAR_Tools/) as a measure of data gaps filling and risk assessment [82]. The robustness, reproducibility, and predictivity of QSTR models were thoroughly validated using globally accepted statistical parameters.

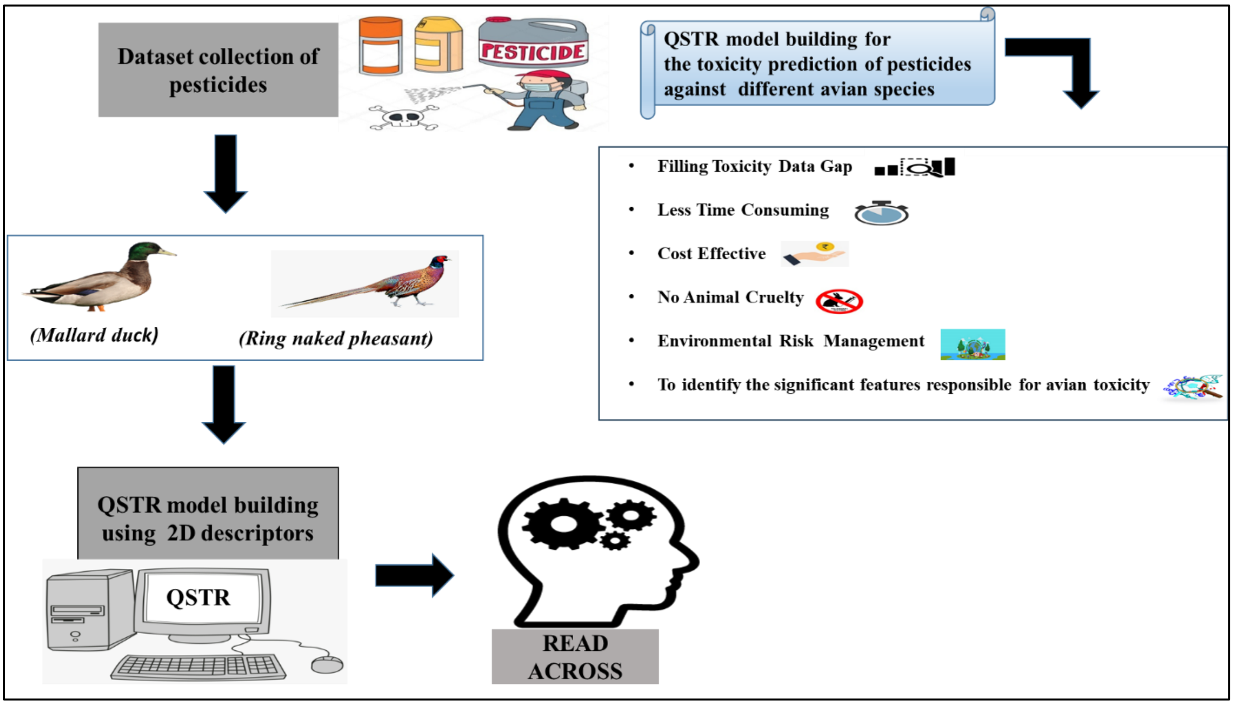

Dataset collection of pesticides
PESTICIDE
QSTR model building for the toxicity prediction of pesticides against different avian species
Filling Toxicity Data Gap
Less Time Consuming
Cost Effective
No Animal Cruelty
Environmental Risk Management
To identify the significant features responsible for avian toxicity
(Mallard duck)
(Ring naked pheasant)
QSTR model building using 2D descriptors
QSTR
READ ACROSS

Fig. 1. The graphical representation of the steps involved in the development of the QSTR model.

2.2. Study 2: First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail

Birds are the essential species for the ecosystem and we can’t even imagine a world without birds. Unfortunately, in today’s world due to increasing the usage of different chemical compounds including pesticides a large number of birds have extinct globally. As per the report, around 150 avian species have been wiped out from the planet since the 1500s and one in eight avian species is at the risk of extinction [66]. Healthy avian populations are a sign of ecological integrity [83] and they also play a significant role in a wide range of functions including pollination, scavenging, seed-dispersing, pest-predator, nutrient cycling, ecosystem engineering, and many more [84]. Therefore, the global decline in bird numbers is a matter of great concern. Human beings ceaselessly manipulate nature to fulfill their demands with increasing population through various activities like deforestation, usage of pesticides, and industrialization [61].

To fulfill the agricultural demand of the rising population and for effective crop management pesticide usage increases rapidly. As a consequence, toxic residues of pesticides accumulate in the environment and affect both terrestrial and aquatic food chains. Several researchers reported that currently used pesticides are lack specificity which may responsible for toxicity toward various non-target species including humans and birds. Birds are susceptible to exposure to pesticides directly through spray treatment and indirectly by feeding, preening, and grooming. Therefore, such effects especially on birds are a threat to the ecosystem and biodiversity, suggested test protocols for oral toxicity in birds.

Considering the aforementioned, a number of regulatory bodies have given priority to the toxicity assessment of pesticides and their derivatives. As the oral route is thought to be the most important for pesticide exposure in avian species, the United States Environmental Protection Agency (USEPA) and the Organization for Economic Co-operation and Development (OECD) proposed test protocols for avian oral toxicity. However, estimating avian toxicity by using animal models is a tough task as well as quite expensive, time-consuming, and unethical. The concept of “3R” (replacement, reduction, and refinement) given by Russell and Burch in 1959, aims to implement other possible approaches for toxicity prediction [54]. Recently, some regulatory agencies like the Environment Protection Agency (EPA), European Chemical Bureau (ECB), and Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) proposed to limit the usage of experimental animals and compel to use of non-animal models or in silico based methods like QSAR or QSTR for toxicity risk assessment, which are considered as a convenient replacement for both in vivo and in vitro approaches, offering advantages in economy and time efficiency.

These techniques can investigate significant structural characteristics and forecast the biological activity or toxicity of the novel compounds [62]. Few researchers reported various individual-species and multi-species QSAR models for toxicity evaluation of various chemicals in fish, rats, and mice [85-87] but in the case of avian species, few *In-silico-based* models are reported. By thorough analysis of various previous works on toxicity assessment, we found that toxicological evaluation for avian species is majorly conducted using either Mallard duck, Bobwhite quail, Japanese quail, Ring-necked pheasant, or zebra finch. However, red-winged blackbirds, house finches, house sparrows, brown-headed cowbirds, and California quail might be utilized optionally or alternatively. Formerly some researchers reported the anti-cholinergic effect of insecticides on California quail [88], which might be considered a threat to their existence. Therefore, we used California quail as the test organism for the toxicological assessment of the chemical pesticides.

The present study deals with avian toxicity assessment of pesticides against California quail

(*Callipepla california*) by constructing QSTR models with partial least square (PLS)-regression by employing 2D descriptors for model development to avoid molecular optimization complexity. We have also attempted to strengthen the prediction quality of the test set compound by intelligent selection of various models using the "Intelligent consensus predictor" tool [89]. All the produced models were built in accordance with the OECD recommendation.

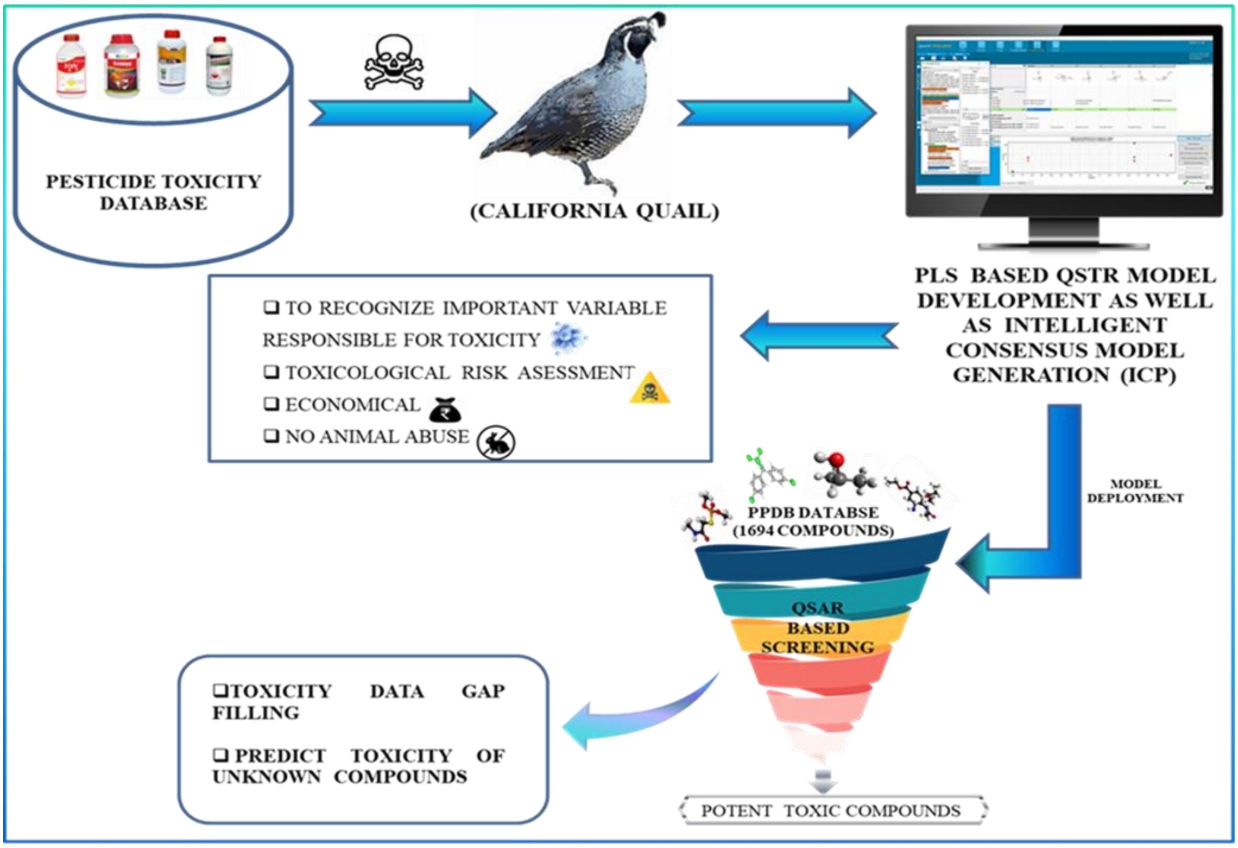

PESTICIDE TOXICITY DATABASE
(CALIFORNIA QUAIL)
PLS BASED QSTR MODEL DEVELOPMENT AS WELL AS INTELLIGENT CONSENSUS MODEL GENERATION (ICP)
❑ TO RECOGNIZE IMPORTANT VARIABLE RESPONSIBLE FOR TOXICITY
❑ TOXICOLOGICAL RISK ASESSMENT
❑ ECONOMICAL
❑ NO ANIMAL ABUSE
MODEL DEPLOYMENT
PPDB DATABSE (1694 COMPOUNDS)
QSAR BASED SCREENING
POTENT TOXIC COMPOUNDS
❑TOXICITY DATA GAP FILLING
❑ PREDICT TOXICITY OF UNKNOWN COMPOUNDS

Fig. 2. The graphical representation of the steps involved in the development of the QSTR model.

2.3.Study 3: Chemometric-based exploration of the toxicological significance of diverse chemical toxicants in wild birds with an application of the q-RASTR approach.

In this modern era, the usage of various chemical compounds has risen enormously to fulfill human requirements. Chemical compounds are mainly designed for different specific purposes, as they tend to air and aquatic transport, these substances have been detected globally. The environment, biodiversity, and wildlife including birds are greatly affected by these chemicals. Birds represent a diverse group of species that play many key ecological roles and offer various services associated with the ecosystem such as nutrient cycling, pollination, seed dispersal, and promoting plant growth and diversity via their herbivory activities [90]. They also function as markers for environmental health. Birds have cultural importance and are also considered symbols of nations and organizations. Regrettably, birds have been recognized as a species that experience the harmful consequences of various chemicals. Various reports suggest that there is a huge drop in the bird population worldwide due to eggshell thinning after exposure to the well-known pesticide dichlorodiphenyltrichloroethane (DDT) [91]. Some researchers reported that a decline in the Gyps vulture population was triggered by diclofenac-induced toxicity [92] and the poisoning of Red kites as a result of exposure to several pesticides and rodenticides [93]. Another research performed by J. W. Macdonald reported that a group of 38 birds which is 19 percent of the total bird population have died as a consequence of adverse environmental factors including chemicals, pesticides, and pharmaceuticals [94].

So, the above reports indicate that there is a significant impact of different chemical compounds on the bird's wildlife ecosystem. Therefore, various regulatory agencies emphasize toxicity assessment of chemical compounds to determine their toxicological effect on various species including both terrestrial and aquatic. For toxicity estimation of a large number of compounds in vitro and in vivo assessment is quite expensive, time-consuming, and needs sacrifice of innocent animals. Therefore, it is essential to attempt various alternative methods of toxicity assessment such as in-silico-based methods which include QSAR, QSTR, and q-RASTR. The in-silico-based approaches associated with chemical toxicology continue to be a better alternative to in-vivo or in-vitro methods of toxicity assessment as they reduce human effort as well as time and cost [95]. Regulatory authorities, chemical industries, and risk evaluators also encourage computational toxicology methods particularly QSTR as suitable technique for early hazard identification and risk assessment. In recent times, few models have been fabricated for toxicity evaluation of various environmental hazards using QSAR or other related in-silico techniques. Various toxicological studies of diverse chemicals against several species such as dogs, fish, and rats have already been performed by using QSAR methods for toxicity prediction [85,96,97]. Recently, few researchers reported avian toxicity studies of diverse chemicals or pesticides using single species [98] as well as multiple avian species. The Read-Across Structure-Activity Relationship is the fusion of both QSAR and the Read-Across approach which improves the reliability of predictions. q-RASAR boosts the predictive ability and diminishes the mean absolute error by using similarity and error-based descriptors [99]. In the recent past, the q-RASAR approach has been used by many researchers for toxicity assessment associated with molecular and environmental contexts. Banerjee et al. developed multiple q-RASAR models for assessing the cytotoxicity of TiO2-based nanoparticles using two sets of toxicity data [100]. Ghosh et al. also use the q-RASAR approach to predict the aquatic toxicity of organic pesticides against fish [101].

In this study, we have generated PLS-based QSTR and q-RASTR models using experimental data with the endpoint LD_{50} of diverse chemical compounds toward wild birds. Wild birds encompass a large group of avian species and represent a major part of our ecosystem. Therefore, it is essential to evaluate the toxicological significance using experimental data on diverse chemicals for wild birds.

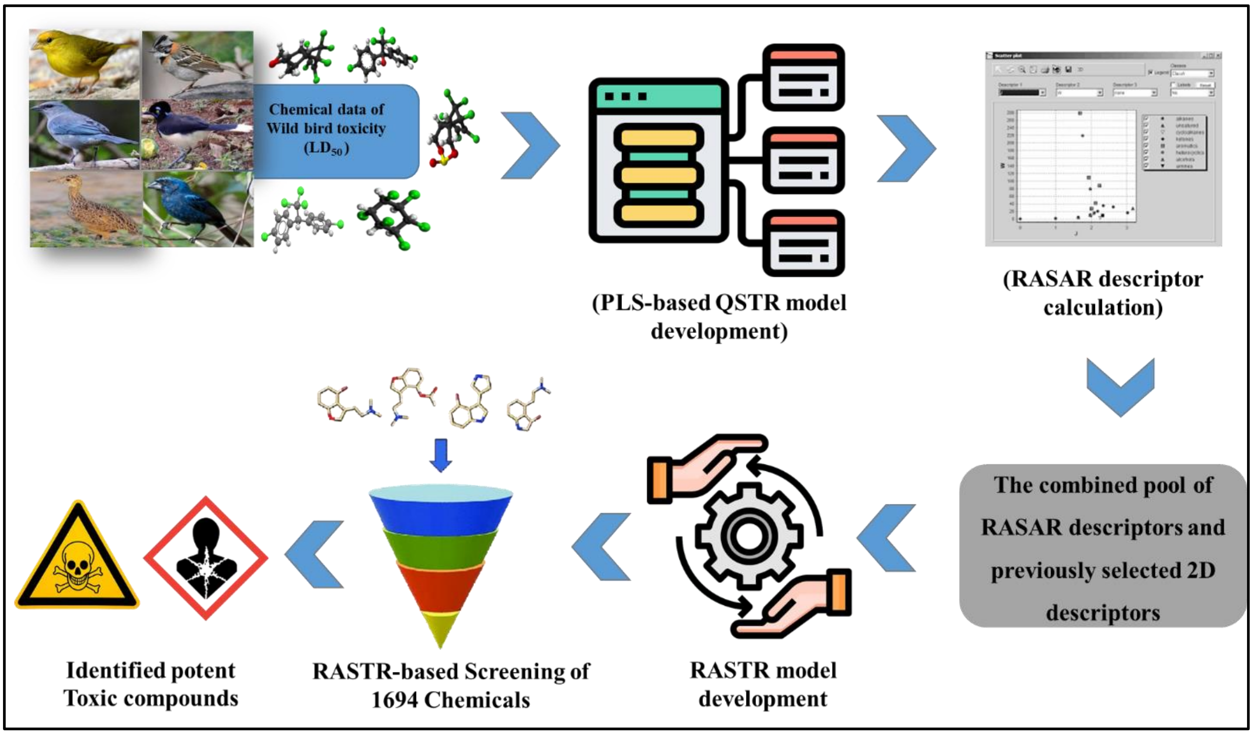

Chemical data of Wild bird toxicity (LD_{50})
(PLS-based QSTR model development)
(RASAR descriptor calculation)
The combined pool of RASAR descriptors and previously selected 2D descriptors
RASTR model development
RASTR-based Screening of 1694 Chemicals
Identified potent Toxic compounds

Fig. 3. The graphical representation of the steps involved in the the q-RASTR model development.

CHAPTER - 3

Materials and Methods

3. MATERIALS AND METHODS

The dissertation presents a transparent methodology for developing QSTR and q-RASTR models using simple 2D descriptors. Our objective has been to ensure clarity and transparency in the process, from the calculation of descriptors to the reduction of the variable matrix, the identification of significant features, and the assessment of the models' reliability and predictive abilities. QSTR models are generally developed in a number of steps, by following the OECD guidelines. This dissertation describes the process we followed in order to complete our studies. In the first part, we provided a general overview of the steps involved in generating a predictive QSTR model; next, we described the methods for each study.

We divided the work into the following parts:

Dataset details: In this section, we provide a comprehensive account of the datasets used in our study. These datasets include information on chemical names and their corresponding activity or toxicity data. This foundational information serves as the bedrock for our research.

Methodological Approach: We represent a general overview of the methodologies and techniques employed in the development of our models. This section outlines the strategies and tools we used to create predictive models for understanding the relationship between chemical structures and toxicity.

3.1 Organization for Economic Cooperation and Development (OECD) guidelines for the QSTR model generation

To develop a QSTR model, certain factors should be considered according to the Organization for Economic Cooperation and Development (OECD). **Table 3** illustrates the five OECD guidelines for the validation of a QSTR model.

Table 3. OECD guidelines for QSTR models.

Serial No.	OECD guidelines	Description
1	A defined end-point	To make sure that all endpoint values in a dataset are the same.
2	An unambiguous algorithm	In order to ensure that the suggested QSAR model is transparent and reproducible
3	In order to ensure that the suggested QSAR model is transparent and reproducible	The necessity to define an AD reflects the fact that QSARs are inherently limited in terms of the sorts of chemical structures, physicochemical characteristics, and mechanisms of action for which they can provide valid predictions.

4	Appropriate measures of goodness of fit, robustness, and predictivity	To make the overall criterion of model validation easier to understand: determination of the internal performance and predictive capability of a model.
5	Mechanistic interpretation, if possible	To ensure that there are assessments of the possibility of a mechanistic interpretation.

3.2 Study 1

3.2.1 Dataset preparation

Here, we developed models using datasets with toxicity endpoint (LC_{50}; defined as the lethal concentration in 50% population) for toxicity prediction in multiple avian species collected from literature [102] which was originally collected from the EPA, Ecotox database (http://cfpub.epa.gov/ecotox/). In this study; 112 pesticides for RNP (Ring-necked pheasant), and 564 pesticides for MD (Mallard duck) were taken for the development of the model. The toxicity endpoint values range from 0.27 to 4.67 in MD and 0.162 to 4.857 in RNP. The two-dimensional structures of the pesticides were sketched using Marvin Sketch 5.5.0.1 (https://chemaxon.com) with the addition of explicit hydrogen atoms as well as proper aromatization. The conversion of structure file formats was carried out using Open Babel v.2.3.2 [103]. Knime workflow (https://www.knime.com/cheminformatics-extensions) was employed for data curation which removes unwanted salts and duplicate compounds. Toxicity in an avian species characterized as an endpoint value (LC_{50}) was converted to millimolar (mM) concentration followed by converting to a negative logarithmic scale, pLC_{50}, for easy interpretation. Some compounds were omitted from the datasets due to high residual values.

Table 4. Canonical smiles with respective experimental pLC_{50} values of the RNP dataset.

Sl No.	Canonical Smiles	pLC50
1	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	3.056
2	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	1.879
3	COP(=S)(OC)Oc1ccc(cc1)S(=O)(=O)N(C)C	3.822
4	COP(=S)(OC)Oc1ccc(SC)c(C)c1	3.139
5*	CCOP(=S)(OCC)Oc1ccc(cc1)[N+](=O)[O-]	2.937
6*	CCOP(=S)(OCC)Oc1ccc2C(=C(Cl)C(=O)Oc2c1)C	3.057
7	C1C2C(C(C1Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	2.979
8	C1(C(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl	2.714
9	CNC(=O)CSP(=S)(OC)OC	2.839
10	C1=NNC(=N1)N	1.225
11	COP(=O)(OC)OC=C(Cl)Cl	2.590
12	CNC(=O)Oc1cccc2ccccc12	1.604
13	CC(=O)C	0.162

14	COc1ccc(cc1)C(c2ccc(OC)cc2)C(Cl)(Cl)Cl	1.839
15	ClC(Cl)C(c1ccc(Cl)cc1)c2ccc(Cl)cc2	2.856
16	ClC(=C(c1ccc(Cl)cc1)c2ccc(Cl)cc2)Cl	2.583
17	CCc1ccc(cc1)C(C(Cl)Cl)c2ccc(CC)cc2	1.788
18	CC(Cl)(Cl)C(=O)O	1.456
19*	ClC1C=CC2C1C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	3.221
20	CCOP(=S)(OCC)SC1OCCOC1SP(=S)(OCC)OCC	2.050
21*	CC(=C)C1CC2=C(O1)C=CC3=C2OC4COC5=CC(=C(C=C5C4C3=O)OC)OC	2.389
22*	OC(=O)Cc1c(Cl)ccc(Cl)c1Cl	1.680
23	COP(=S)(OC)SCN1N=Nc2ccccc2C1=O	2.241
24*	Oc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.788
25	CCC(C)c1cc(cc(c1O)[N+](=O)[O-])[N+](=O)[O-]	2.668
26	CC(Oc1cc(Cl)c(Cl)cc1Cl)C(=O)O	1.777
27	Cc1cc(Cl)ccc1OCC(=O)O	2.001
28	Cc1cc(Cl)ccc1OCCCC(=O)O	1.660
29	OC(=O)CCCOc1ccc(Cl)cc1Cl	1.697
30*	CN(C)C(=O)Nc1ccccc1	1.516
31*	CNC(=O)Oc1ccccc1OC(C)C	2.077
32	ClC1=C(Cl)C2(Cl)C3COS(=O)OCC3C1(Cl)C2(Cl)Cl	2.503
33	OC(c1ccc(Cl)cc1)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.241
34	CCOP(=S)(OCC)Oc1ccc(cc1)S(=O)C	3.318
35	CNC(=O)O\N=C\C(C)(C)SC	2.802
36	Clc1ccc(cc1)S(=O)(=O)c2cc(Cl)c(Cl)cc2Cl	1.852
37	ClC1=C(Cl)C(=O)c2ccccc2C1=O	1.657
38	CCCCC(CC)COC(=O)c1ccccc1C(=O)OCC(CC)CCCC	1.892
39*	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	2.664
40	CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC	2.097
41	COP(=S)(OC)Oc1ccc(c(C)c1)[N+](=O)[O-]	2.786
42	CCNc1nc(Cl)nc(NCC)n1	1.605
43	CC(Cl)(Cl)C(=O)[O-]	1.453
44	ClC(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	1.779
45	CN(C)C(=S)SSC(=S)N(C)C	1.682
46	CNC(=S)[S-]	1.327
47*	CC(COc1ccc(cc1)C(C)(C)C)OS(=O)OCCCl	1.825
48*	COP(=O)(OC)OC(=CC(=O)N(C)C)C	3.731
49	[S-]C(=S)NCCNC(=S)[S-]	1.624
50	CN(C)C(=O)Nc1ccc(Cl)cc1	1.627
51	CCOP(=S)(OCC)Oc1cnccn1	3.537
52*	COP(=S)(OC)Oc1ccc(cc1)[N+](=O)[O-]	3.461
53*	CCOP(=S)(OCC)SCSCC	2.771
54	CCOP(=S)(OCC)SCCSCC	2.636
55	COP(=O)(OC)OC(Br)C(Cl)(Cl)Br	2.176
56	CCS(=O)CCSP(=O)(OC)OC	2.216
57*	C1C2C=CC1C3C2C4(C(=C(C3(C4(Cl)Cl)Cl)Cl)Cl)Cl	3.806
58	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	1.416
59*	CNC(=O)Oc1cc(C)c(N(C)C)c(C)c1	2.419

60	CN(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.66
61*	CON(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.860
62	CCOP(=S)(OCC)Oc1cc(C)nc(n1)C(C)C	3.096
63	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	1.885
64	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	2.003
65*	CCOP(=S)(CC)Sc1ccccc1	2.960
66*	COP(=S)(OC)SCSc1ccc(Cl)cc1	2.297
67*	C[N+](C)(C)CCCl	1.389
68*	Clc1cccc(Cl)c1C#N	2.059
69	CNC(=O)Oc1cccc2CC(C)(C)Oc12	2.586
70*	CCCCCCCC(=O)Oc1c(Br)cc(cc1Br)C#N	1.961
71	CCNc1nc(Cl)nc(NC(C)C)n1	1.634
72*	Nc1c(Cl)c(Cl)nc(C(=O)O)c1Cl	1.683
73	CCCCOCCOC(=O)COc1ccc(Cl)cc1Cl	1.807
74	CNC(=O)Oc1ccc(N(C)C)c(C)c1	2.017
75	CNC(=O)Oc1cc(C)c(SC)c(C)c1	1.767
76	CCOP(=S)(Oc1ccc(cc1)[N+](=O)[O-])c2ccccc2	2.478
77*	CN(C)C(=O)Nc1cccc(c1)C(F)(F)F	1.867
78	CCCC(C)c1cccc(OC(=O)NC)c1	1.646
79	ClC1(Cl)C2(Cl)C3(Cl)C4(Cl)C(Cl)(Cl)C5(Cl)C(Cl)(C1(Cl)C35Cl)C24Cl	2.549
80	CCCCOCCOC(=O)COc1cc(Cl)c(Cl)cc1Cl	1.954
81	CCOC(=O)C(SP(=S)(OC)OC)c1ccccc1	2.062
82*	Cc1cc(cc(c1C)C)OC(=O)NC	1.632
83	CCOP(=S)(OCC)Oc1nc(Cl)c(Cl)cc1Cl	2.802
84	CCCOP(=S)(OCCC)OP(=S)(OCCC)OCCC	1.879
85	COC(=O)NS(=O)(=O)c1ccc(N)cc1	0.487
86	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)cc2)cc1	3.459
87	COP(=S)(OC)Oc1c(cc(c(n1)Cl)Cl)Cl	1.888
88*	CC1=C(Cl)C(=O)N(C(=O)N1)C(C)(C)C	0.913
89	CN(C)C=Nc1ccc(Cl)cc1C	1.877
90	CNC(=O)C=C(C)OP(=O)(OC)OC	4.857
91	COC(=O)C=C(C)OP(=O)(OC)OC	2.959
92*	ClCC1(CCl)C(=C)C2(Cl)C(Cl)C(Cl)C1(Cl)C2(Cl)Cl	2.880
93	CC1(C(=C)C2(C(C1(C(C2(Cl)Cl)(Cl)Cl)Cl)Cl)Cl)C	2.673
94	CC1C(OC(=O)C2C(C=C(C)C)C2(C)C)C=C(CC=CC=C)C1=O	1.817
95	CCOP(=O)(OCC)SCCSCC	2.589
96*	Clc1ccc(c(Cl)c1Cl)c2ccc(Cl)c(Cl)c2Cl	2.456
97	Clc1ccc(c(Cl)c1Cl)c2cccc(Cl)c2Cl	2.475
98*	CC1=CC(=C(C(=C1)OC(=O)NC)C)C	1.632
99	Clc1cc(Cl)cc(c1)c2cc(Cl)cc(Cl)c2	2.347
100	CCN(CC)C(=O)\C(=C(/C)\OP(=O)(OC)OC)\Cl	3.590
101	CCCSP(=O)(OCC)SCCC	3.312
102*	CNC(=O)ON=C(C)SC	1.914
103	CCCCOCCOC(=O)C(C)Oc1cc(Cl)c(Cl)cc1Cl	2.245
104	CC(C)C(=O)[O-]	1.241
105	COP(=O)(OC)OC(=CCl)c1cc(Cl)c(Cl)cc1Cl	1.864

106	CCOP(=S)(OCC)SCCl	1.750
107	CCC(C)c1cccc(OC(=O)N(C)Sc2ccccc2)c1	2.040
108	COC(CN(C)C(=O)Nc1nnc(s1)C(C)(C)C)OC	2.444
109*	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc2cccc(Oc3ccccc3)c2	1.230
110	Clc1ccc(c(Cl)c1)c2cccc(Cl)c2Cl	2.147
111	CCCCCCCCCCCCCC[N+](C)(C)Cc1ccccc1	1.587
112	Cc1cc(c(cc1C)Cl)OC(=O)NC	1.319

***Test set compounds**

Table. 5. Canonical smiles with respective experimental pLC_{50} values of MD dataset.

Sl No.	Canonical_Smiles	pLC_{50}
1	CC(=CC(=O)NC)OP(=O)(OC)OC	4.366
2	COP(=S)(OC)Oc1nc(Cl)n(n1)C(C)C	4.455
3	COP(=S)(OC)Oc1ccc(cc1)S(=O)(=O)N(C)C	3.968
4	CCOP(=S)(OCC)Oc1ccc(cc1)S(=O)C	3.876
5	COP(=S)(OC)Oc1ccc(SC)c(C)c1	3.745
6	CCOP(=S)(OCC)Oc1ccc(cc1)[N+](=O)[O-]	3.583
7	CNC(=O)Oc1cccc2CC(C)(C)Oc12	3.447
8	CCCSP(=O)(OCC)SCCC	3.449
9	CC(=CC(=O)N(C)C)OP(=O)(OC)OC	3.402
10	CN1SC(=CC1=O)Cl	3.174
11	CCCSP(=O)(OCC)Oc1ccc(Br)cc1Cl	3.396
12	CNC(=O)Oc1cc(C)c(N(C)C)c(C)c1	3.170
13	CCOP(=S)(OCC)SCSC(C)(C)C	3.295
14	CCOP(=S)(OCC)OC(Cl)C(Cl)(Cl)Cl	3.152
15	C1CN2CC3=CCOC4CC(=O)N5C6C4C3CC2C61C7=CC=CC=C75	3.197
16	CCOP(=S)(OCC)SCSCC	3.021
17	CCOP(=O)(NC(C)C)Oc1ccc(SC)c(C)c1	2.982
18	CCOP(=S)(OCC)Oc1ccc2C(=C(Cl)C(=O)Oc2c1)C	2.956
19	CCOP(=S)(OCC)Oc1cnccn1	2.771
20	CC1(C(=C)C2(C(C1(C(C2(Cl)Cl)(Cl)Cl)Cl)Cl)Cl)C	2.922
21	CCOP(=S)(OCC)SC(CCl)N1C(=O)c2ccccc2C1=O	2.928
22	ClC1C=CC2C1C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	2.890
23	CCOP(=S)(OCC)SCCSCC	2.730
24	ClCC1(CCl)C(=C)C2(Cl)C(Cl)C(Cl)C1(Cl)C2(Cl)Cl	2.883
25	CCC(C)c1cc(cc(c1O)[N+](=O)[O-])[N+](=O)[O-]	2.648
26	COC1=NN(CSP(=S)(OC)OC)C(=O)S1	2.745
27	CCOP(=S)(OC(C)C)Oc1cnc(nc1)C(C)(C)C	2.741
28	CN(c1c(Br)cc(Br)cc1Br)c2c(cc(cc2C(F)(F)F)[N+](=O)[O-])[N+](=O)[O-]	2.969
29	COc1ccc(cc1NNC(=O)OC(C)C)c2ccccc2	2.660
30	CCN(CC)C(=O)\C(=C(/C)\OP(=O)(OC)OC)\Cl	2.624
31	CN1SC=CC1=O	2.205
32	Nc1ccncc1	2.115
33	CN(C)C(=O)Oc1nc(nc(C)c1C)N(C)C	2.507
34	CNC(=O)ON=C(C(=O)N(C)C)SC	2.456
35	CCCCCCC(C)C1=C(C(=CC(=C1)[N+](=O)[O-])[N+](=O)[O-])OC(=O)C=CC	2.617

36	CNC(=O)Oc1cc(C)c(SC)c(C)c1	2.384
37	CCCSP(=S)(OCC)Oc1ccc(SC)cc1	2.515
38	Cc1cccc2sc3nncn3c12	2.277
39	CCOP(=S)(NC(C)C)Oc1ccccc1C(=O)OC(C)C	2.538
40	CNC(=O)Oc1ccccc1OC(C)C	2.320
41	OC(CC(C1=C(O)c2ccccc2OC1=O)c3ccccc3)c4ccc(cc4)c5ccc(Br)cc5	2.722
42	ClC1=C(Cl)C2(Cl)C3COS(=O)OCC3C1(Cl)C2(Cl)Cl	2.587
43	COP(=S)(Oc1cc(Cl)c(Br)cc1Cl)c2ccccc2	2.583
44	CC(C1CC1)C(O)(Cn2cncn2)c3ccc(Cl)cc3	2.387
45	CCCCCCCCN1SC=CC1=O	2.244
46	CCOP(=S)(CC)Sc1ccccc1	2.303
47	CC(=O)CC(C1=C(O)c2ccccc2OC1=O)c3ccccc3	2.400
48	Cc1c(COC(=O)C2C(\C=C(/Cl)\C(F)(F)F)C2(C)C)cccc1c3ccccc3	2.519
49	COP(=O)(OC)OC=C(Cl)Cl	2.224
50	Oc1c(Br)cc(cc1Br)C#N	2.302
51	CCOc1nc(ns1)C(Cl)(Cl)Cl	2.176
52	OC(c1ccc(Cl)cc1)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.351
53	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.277
54	CNC(=O)ON=C(C)SC	1.933
55	CCC1O[C@]2(CCC1C)C[C@@H]3C[C@@H](CC=C(C)CC(C)C=CC=C4CO[C@@H]5[C@H](O)C(=C[C@@H](C(=O)O3)[C@]45O)C)O2	2.440
56	Clc1ccc(c(Cl)c1Cl)c2ccc(Cl)c(Cl)c2Cl	2.261
57	CC(=CC(=O)OC)OP(=O)(OC)OC	2.051
58	Cc1cc(Cl)ccc1OCC(=O)O	2.001
59	Clc1ccc(CCC(Cn2cncn2)(C#N)c3ccccc3)cc1	2.223
60	CCCCCCCC(=O)Oc1c(Br)cc(cc1Br)C#N	2.273
61	CC(C)(C)C(=O)C1C(=O)c2ccccc2C1=O	2.010
62	CCC(C)c1cccc(OC(=O)N(C)Sc2ccccc2)c1	2.1373
63	CNC(=O)Oc1cc(C)c(C)c(C)c1	1.924
64	Cc1c(F)c(F)c(COC(=O)C2C(\C=C(/Cl)\C(F)(F)F)C2(C)C)c(F)c1F	2.257
65	COP(=S)(OC)Oc1ccc(c(C)c1)[N+](=O)[O-]	1.875
66	CC(C)Oc1cc(N2N=C(OC2=O)C(C)(C)C)c(Cl)cc1Cl	2.140
67	CNC(=O)N(C)c1nnc(s1)C(C)(C)C	1.960
68	CNC(=O)Oc1ccc(N(C)C)c(C)c1	1.911
69	COc1cc2OC[C@H]3Oc4c5C[C@@H](Oc5ccc4C(=O)[C@H]3c2cc1OC)C(=C)C	2.181
70*	CC1(C)C(C=C(Cl)Cl)C1C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2	2.198
71*	CC1OC(C)OC(C)OC(C)O1	1.819
72	Clc1ccc(c(Cl)c1Cl)c2cccc(Cl)c2Cl	2.082
73	COP(=O)(OC)OC(Br)C(Cl)(Cl)Br	2.145
74	COP(=S)(OC)SCSc1ccc(Cl)cc1	2.020
75	CON(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.907
76	COC(=O)c1cc(Cl)cc(N)c1Cl	1.842
77	Clc1ccc(c(Cl)c1)c2cccc(Cl)c2Cl	1.962
78	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	1.996
79	NC(=S)Nc1cccc2ccccc12	1.774

80*	CCC12COCN1COC2	1.621
81	CC(C)(C)[C@H](O)C(=Cc1ccc(Cl)cc1)n1cncn1	1.928
82*	ClC(=C(c1ccc(Cl)cc1)c2ccc(Cl)cc2)Cl	1.949
83	CCOP(=S)(OCC)SC1OCCOC1SP(=S)(OCC)OCC	2.103
84	CCCCCCCCCCCCCC[P+](CCCC)(CCCC)CCCC	2.074
85	CCC(CN1CCOCC1)[N+](=O)[O-]	1.707
86	COC(=O)c1csc(C)c1S(=O)(=O)NC(=O)N2N=C(OC)N(C)C2=O	2.006
87*	CC1(C)[C@@H](\C=C(/Cl)\C(F)(F)F)[C@H]1C(=O)O[C@H](C#N)c2cccc(Oc3ccccc3)c2	2.056
88	ClC1=C(Cl)C(Cl)(C(=C1Cl)Cl)C2(Cl)C(=C(Cl)C(=C2Cl)Cl)Cl	2.077
89*	CCSC(C)CC1CC(=C(C(=NOC\C=C\Cl)CC)C(=O)C1)O	1.956
90	CC1=C(SCCO1)C(=O)Nc2ccccc2	1.757
91	COc1cnc(OC)n2nc(NS(=O)(=O)c3c(OCC(F)F)cccc3C(F)(F)F)nc12	2.049
92*	COC(=O)Nc1nc2ccccc2[nH]1	1.631
93*	CCOc1nc(F)cc2nc(nn12)S(=O)(=O)Nc3c(Cl)cccc3Cl	1.955
94	N#CSCSc1nc2ccccc2s1	1.724
95	CCCCCCCCCCCCCC[N+](C)(C)Cc1ccccc1	1.868
96*	CN(C)C(=O)Nc1cccc(c1)C(F)(F)F	1.712
97*	COC(=O)c1ccc(I)cc1S(=O)(=O)[N-]C(=O)Nc2nc(C)nc(OC)n2	2.050
98*	CN1C=C(C(=O)C(=C1)c2cccc(c2)C(F)(F)F)c3ccccc3	1.860
99*	[O-][N+](=O)\C=C/1\NCCCS1	1.544
100	COC(=O)NC(=S)Nc1ccccc1NC(=S)NC(=O)OC	1.873
101	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(C)cc(C)n2	1.898
102	O=C1NSc2ccccc12	1.514
103	CCOC(=O)[C@@H](C)Oc1ccc(Oc2oc3cc(Cl)ccc3n2)cc1	1.892
104*	CC(C)(C)c1ccc(OC2CCCCC2OS(=O)OCC#C)cc1	1.878
105	CC(C)CCCC(C)C\C=C\C(=C\C(=O)OCC#C)\C	1.775
106	CC1(C)[C@@H](C=C(Br)Br)[C@H]1C(=O)O[C@H](C#N)c2cccc(Oc3ccccc3)c2	2.0369
107	CCCCCCCCCCCCCC=CCCCCCCCC	1.842
108	CCNc1nc(NC(C)(C)C)nc(SC)n1	1.716
109	ClC(Cl)(Cl)C(NC=O)N1CCN(CC1)C(NC=O)C(Cl)(Cl)Cl	1.971
110	CN[C@H]1[C@H](O)[C@@H](O)[C@H](CO)O[C@H]1O[C@H]2[C@H](O[C@H]3[C@H](O)[C@@H](O)[C@H](NC(=N)N)[C@@H](O)[C@@H]3NC(=N)N)O[C@@H](C)[C@]2(O)C=O	2.098
111*	COc1c(Cl)ccc(Cl)c1C(=O)[O-]	1.675
112	Fc1cccc(F)c1C(=O)NC(=O)Nc2ccc(Cl)cc2	1.825
113*	COC1=NN(C(=O)[N-]S(=O)(=O)c2ccccc2OC(F)(F)F)C(=O)N1C	1.952
114	CC(C)(C)N1N=CC(=C(Cl)C1=O)SCc2ccc(cc2)C(C)(C)C	1.891
115	CN(C)C(=S)[S-]	1.409
116	COC(=O)c1ccc(CNS(=O)(=O)C)cc1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	2.025
117*	COc1cc(OC)nc(NC(=O)NS(=O)(=O)Nc2ccccc2C(=O)N(C)C)n1	1.947
118	[O-][N+](=O)N=C1NCCN1Cc2ccc(Cl)nc2	1.726
119	[O-][N+](=O)NC1=NCCN1Cc1ccc(Cl)nc1	1.726
120	CCN(Cc1c(F)cccc1Cl)c2c(cc(cc2[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.943
121*	ClC(Cl)C(c1ccc(Cl)cc1)c2ccc(Cl)cc2	1.822

122	CC(C)(C)C(O)(CCc1ccc(Cl)cc1)Cn2cncn2	1.805
123	CCC1=C(C(=O)[O-])C(=O)C=NN1c2ccc(Cl)cc2	1.760
124*	CCCCOC(=O)[C@@H](C)Oc1ccc(Oc2ccc(cn2)C(F)(F)F)cc1	1.897
125	CC(C)[C@H](C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1)c3ccc(OC(F)F)cc3	1.965
126*	CC(C)[C@H](C(=O)O[C@H](C#N)c1cccc(Oc2ccccc2)c1)c3ccc(Cl)cc3	1.933
127*	[O-]C(=O)C1C2CCC(O2)C1C(=O)[O-]	1.566
128	BrCC(=O)OC\C=C\COC(=O)CBr	1.819
129*	C[C@@H](Oc1ccc(Cl)cc1C)C(=O)O	1.632
130	C[N+](C)(C)CCCl	1.389
131*	CC(=CC1C(C(=O)OCc2coc(Cc3ccccc3)c2)C1(C)C)C	1.830
132*	CC(=CCC\C(=C\CC\C(=C\CO)\C)\C)C	1.648
133*	CC(=CCC\C(=C\CCC(C)(O)C=C)\C)C	1.648
134*	CC(=O)Nc1cc(NS(=O)(=O)C(F)(F)F)c(C)cc1C	1.792
135*	CC(C)(C)C(O)C(Oc1ccc(cc1)c2ccccc2)n3cncn3	1.829
136	CC(C)(C)N(NC(=O)c1ccc(Cl)cc1)C(=O)c2ccccc2	1.820
137*	CC(C)[C@@]1(C)N=C(NC1=O)c2ncc(C)cc2C(=O)[O-]	1.739
138*	CC(C)C(C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1)c3ccc(Cl)cc3	1.924
139*	CC(C)C(O)(c1ccc(OC(F)(F)F)cc1)c2cncnc2	1.795
140*	CC(C)C1(C)N=C(NC1=O)c2nc3ccccc3cc2C(=O)O	1.794
141	CC(C)C1(C)N=C(NC1=O)c2ncccc2C(=O)O	1.718
142	CC(C)CCCCCCCCCCCCCCOCCO	1.798
143*	CC(C)N(C(=O)CCl)c1ccccc1	1.626
144	CC(Cl)(Cl)C(=O)[O-]	1.453
145	CC(Cl)(Cl)C(=O)O	1.456
146	CC(O)CSS(=O)(=O)C	1.532
147*	CC1(C)[C@H](C=C(Cl)Cl)[C@H]1C(=O)O[C@H](C#N)c2ccc(F)c(Oc3ccccc3)c2	1.938
148	CC1(C)CCC(=Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	1.803
149*	CC1(C)N(Cl)C(=O)N(Br)C1=O	1.683
150*	CC1(C)N(CO)C(=O)N(CO)C1=O	1.575
151	CC1=C(Cl)C(=O)N(C(=O)N1)C(C)(C)C	1.636
152	Cc1cc(Cl)ccc1OCCCC(=O)[O-]	1.658
153*	Cc1cc(Cl)ccc1OCCCC(=O)O	1.660
154*	Cc1ncc([N+](=O)[O-])n1CCO	1.534
155	CCC(=O)Nc1ccc(Cl)c(Cl)c1	1.639
156	CCC(C)(C)C(=O)OC1=C(C(=O)OC12CCCCC2)c3cc(Cl)cc(Cl)c3	1.915
157*	CCc1ccc(cc1)C(=O)NN(C(=O)c2cc(C)cc(C)c2)C(C)(C)C	1.8487
158*	CCc1ccc(cc1)C(C(Cl)Cl)c2ccc(CC)cc2	1.788
159	CCc1cccc(CC)c1N(COC)C(=O)CCl	1.732
160*	CCc1nn(C)c(C(=O)NCc2ccc(cc2)C(C)(C)C)c1Cl	1.824
161*	CCCCC(CC)COC(=O)c1ccccc1C(=O)OCC(CC)CCCC	1.892
162	CCCCC(Cn1cncn1)(C#N)c2ccc(Cl)cc2	1.761
163	CCCCCCCC[N+](C)(C)CCCCCCCC	1.733
164*	CCCCCCCCCC[C@H]1O[C@H]1CCCCC(C)C	1.752
165*	CCCCCCCCCCCC(=O)[O-]	1.600
166	CCCCCCCCSC(=O)Oc1cc(Cl)nnc1c2ccccc2	1.879

167*	CCCCN(CC)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.826
168*	CCCCOCCOC(=O)C(C)Oc1cc(Cl)c(Cl)cc1Cl	1.868
169	CCCCOCCOC(=O)COc1cc(Cl)c(Cl)cc1Cl	1.852
170*	CCCCOCCOC(=O)COc1ccc(Cl)cc1Cl	1.807
171*	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])C(C)C)[N+](=O)[O-]	1.791
172	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.826
173*	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])S(=O)(=O)N)[N+](=O)[O-]	1.840
174	CCCOP(=S)(OCCC)OP(=S)(OCCC)OCCC	1.879
175*	CCN(CC(=C)C)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.823
176	CCN(CC)C(=O)c1cccc(C)c1	1.582
177	CCN(CC)C(=O)SCc1ccc(Cl)cc1	1.712
178	CCN(CC)C(=S)SCC(=C)Cl	1.650
179	CCNc1nc(Cl)nc(NC(C)C)n1	1.634
180	CCNc1nc(Cl)nc(NCC)n1	1.605
181	CCOC(=O)C(C)Oc1ccc(Oc2cnc3cc(Cl)ccc3n2)cc1	1.872
182	CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC	1.820
183	CCOC(=O)COc1cc(c(F)cc1Cl)c2nn(C)c(OC(F)F)c2Cl	1.917
184	CCOC(=O)Nc1cccc(OC(=O)Nc2ccccc2)c1	1.778
185*	CCOc1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.859
186	CCOc1ccc(cc1)C(C)(C)COCc2cccc(Oc3ccccc3)c2	1.876
187	CCOP(=S)(OCC)SCCl	1.671
188	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	1.885
189	CCS(=O)CCSP(=O)(OC)OC	1.692
190*	Cl[C@@H]1[C@H](Cl)[C@@H](Cl)[C@H](Cl)[C@H](Cl)[C@H]1Cl	1.764
191	ClC(Cl)(Cl)S(=O)(=O)C(Cl)(Cl)Cl	1.779
192	ClC(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	1.677
193	ClC(Cl)(Cl)SN1C(=O)c2ccccc2C1=O	1.773
194	ClC1(Cl)C2(Cl)C3(Cl)C4(Cl)C(Cl)(Cl)C5(Cl)C(Cl)(C1(Cl)C35Cl)C24Cl	2.037
195	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.755
196	Clc1ccc(c(Cl)c1)C2(Cn3cncn3)CC(Br)CO2	1.877
197	Clc1ccc(cc1)C(=O)c2ccc(Cl)cc2	1.700
198*	Clc1ccc(cc1)c2ccccc2NC(=O)c3cccnc3Cl	1.836
199	Clc1ccc(cc1)S(=O)(=O)c2cc(Cl)c(Cl)cc2Cl	1.852
200	Clc1ccccc1Nc2nc(Cl)nc(Cl)n2	1.741
201*	ClN1C(=O)[N-]C(=O)N(Cl)C1=O	1.595
202*	ClN1C(=O)N(Cl)C(=O)N(Cl)C1=O	1.667
203	CN(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.668
204*	CN(C)C(=O)Nc1ccc(Cl)cc1	1.599
205*	CN(C)C(=O)Nc1ccccc1	1.516
206	CN(C)C(=S)SSC(=S)N(C)C	1.682
207	CN(C)C=Nc1ccc(Cl)cc1C	1.594
208	CN(C)C1=NC(=O)N(C2CCCCC2)C(=O)N1C	1.703
209*	CN(Cc1ccc(Cl)nc1)C(=NC#N)C	1.648
210*	CN\C(=N\[N+](=O)[O-])\NCC1CCOC1	1.606
211	CN1CSC(=S)N(C)C1	1.511
212	CNC(=O)Oc1cccc2ccccc12	1.604

213	COC(=O)C(C)Oc1ccc(Oc2ccc(Cl)cc2Cl)cc1	1.834
214	COC(=O)c1cc(Oc2ccc(Cl)cc2Cl)ccc1[N+](=O)[O-]	1.835
215	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(OC(F)F)cc(OC(F)F)n2	1.9718
216*	COc1cc(OC)n2nc(NS(=O)(=O)c3c(OC)nccc3C(F)(F)F)nc2n1	1.938
217*	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2ncccc2C(=O)N(C)C)n1	1.914
218	COc1ccc(cc1)C(c2ccc(OC)cc2)C(Cl)(Cl)Cl	1.839
219	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2Cl)n1	1.854
220	CON(C(=O)OC)c1ccccc1COc2ccn(n2)c3ccc(Cl)cc3	1.889
221	CON=C(C(=O)OC)c1ccccc1COc2ccccc2C	1.797
222	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	1.711
223*	COP(=S)(OC)Oc1nc(Cl)c(Cl)cc1Cl	1.809
224	CP(=O)(O)CCC(N)C(=O)[O-]	1.556
225	CSC(=O)c1cccc2nnsc12	1.623
226	N#CSCSC#N	1.415
227	Nc1c(Cl)c(Cl)nc(C(=O)O)c1Cl	1.683
228*	Nc1nc(NCl)nc(n1)N(Cl)Cl	1.661
229	Nc1nc[nH]n1	1.225
230*	O=C(Nc1ccccc1)Nc2cnns2	1.644
231	OC(=O)c1c(Cl)ccc2cc(Cl)cnc12	1.684
232	OC(=O)Cc1c(Cl)ccc(Cl)c1Cl	1.680
233	OC(=O)CN(CP(=O)(O)O)CP(=O)(O)O	1.721
234	OCCN(CC[O-])CC[O-]	1.468
235	OCCN1CN(CCO)CN(CCO)C1	1.642
236	OP(=O)(O)CCCl	1.460
237*	CC(=C)C1CCC(C)=CC1	1.435
238	CC(C)C1(C)NC(=NC1=O)c1ncc(C)cc1C([O-])=O	1.739
239*	CC1(C)CCCC(C1)=CC=O	1.483
240*	CC1=C(C)S(=O)(=O)CCS1(=O)=O	1.623
241*	CCCCCCCC1CCC(=O)O1	1.5665
242*	CCCCCCCCCCCCNC(N)=N	1.657
243	COC(=O)c1ccc(C)cc1C1=NC(=O)C(C)(N1)C(C)C	1.760
244*	CO\N=C(\C1=NOCCO1)/c2ccccc2Oc3ncnc(Oc4ccccc4Cl)c3F	1.962
245*	CC1=NNC(=O)N(C1)\N=C\c2cccnc2	1.637
246*	FC(F)(F)c1ccncc1C(=O)NCC#N	1.658
247*	CNC(=N[N+](=O)[O-])NCc1cnc(Cl)s1	1.695
248	CSC1=N[C@](C)(C(=O)N1Nc2ccccc2)c3ccccc3	1.790
249	CNC(NCc1cnc(Cl)s1)=N[N+]([O-])=O	1.695
250*	CON=C(C(=O)OC)c1ccccc1CON=C(C)c2cccc(c2)C(F)(F)F	1.907
251	Cc1c(ccc(c1C2=NOCC2)S(=O)(=O)C)C(=O)c3cnn(C)c3O	1.854
252	Cc1nn(C)c(O)c1C(=O)c2ccc(cc2S(=O)(=O)C)C(F)(F)F	1.852
253*	CC(C)N1\C(=N\C(C)(C)C)\SCN(C1=O)c2ccccc2	1.778
254	O=CCCCC=O	1.290
255	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(c1)[N+](=O)[O-]	1.820
256*	CCCCN1Sc2ccccc2C1=O	1.606
257*	Cc1cc(C)nc(Nc2ccccc2)n1	1.589
258	CC(COc1ccc(Oc2ccccc2)cc1)Oc3ccccn3	1.791

259*	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc2cccc(Oc3ccccc3)c2	1.876
260	CCCOC(=O)NCCCN(C)C	1.558
261*	CCOC(=O)CN(C(=O)CCl)c1c(CC)cccc1CC	1.777
262	CCOC1Oc2ccc(OS(=O)(=O)C)cc2C1(C)C	1.740
263*	Clc1cccc(Cl)c1C#N	1.519
264	CN1COCN(Cc2cnc(Cl)s2)C1=N[N+](=O)[O-]	1.749
265*	CO\C=C(\C(=O)OC)/c1ccccc1Oc2cc(Oc3ccccc3C#N)ncn2	1.889
266	OC(=O)CNCP(=O)(O)O	1.512
267	N(c1ccccc1)c2ccccc2	1.512
268	CC(C)OC(=O)COc1ccc(Cl)cc1Cl	1.702
269	C[C@@H](Oc1ccc(Oc2ncc(Cl)cc2F)cc1)C(=O)OCC#C	1.825
270	CC1(C)CCC(Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	1.786
271*	CC(Oc1ccc(Oc2ncc(Cl)cc2F)cc1)C(=O)OCC#C	1.825
272*	CCOc1cc(ccc1C2COC(=N2)c3c(F)cccc3F)C(C)(C)C	1.836
273*	CC(C)Oc1cccc(NC(=O)c2ccccc2C(F)(F)F)c1	1.790
274	CCC(C)(NC(=O)c1cc(Cl)c(C)c(Cl)c1)C(=O)CCl	1.807
275*	Cc1cc(C)c(C2=C(OC(=O)C(C)(C)C)C3(CCCC3)OC2=O)c(C)c1	1.829
276	CCCCCCCCC=CCCCCCCCC(=O)[O-]	1.726
277	Nc1nc(nc(C(=O)O)c1Cl)C2CC2	1.606
278	CC1(C)[C@H](C=C(Cl)Cl)[C@H]1C(=O)O[C@@H](C#N)c1cccc(Oc2ccccc2)c1	1.895
279	FC(OC(F)(F)F)C(F)(F)Oc1ccc(NC(=O)NC(=O)c2c(F)cccc2F)cc1Cl	1.967
280*	CCS(=O)(=O)c1nc2ccccn2c1S(=O)(=O)NC(=O)Nc3nc(OC)cc(OC)n3	1.946
281	CCSC(=O)N(CC)C1CCCCC1	1.601
282	CCCCOCCOC(=O)COc1nc(Cl)c(Cl)cc1Cl	1.819
283	Nc1cc(Cl)nc(C(=O)O)c1Cl	1.575
284	FC(F)(F)c1ccc(OCCCOc2c(Cl)cc(OCC=C(Cl)Cl)cc2Cl)nc1	1.950
285*	CC(CN1C[C@@H](C)O[C@@H](C)C1)Cc2ccc(cc2)C(C)(C)C	1.735
286	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2ncccc2C(F)(F)F)n1	1.862
287	CC1=CC(=O)NO1	1.247
288*	[O-][N+](=O)\C(=C\c1ccccc1)\Br	1.608
289	C[C@H](O)C(=O)O	1.204
290	C[C@H]1[C@@H](SC(=O)N1C(=O)NC2CCCCC2)c3ccc(Cl)cc3	1.797
291	c1ccc2[nH]c(nc2c1)c3cscn3	1.554
292*	CC(=C[C@@H]1[C@@H](C(=O)OCN2C(=O)C3=C(CCCC3)C2=O)C1(C)C)C	1.770
293	CC(=C[C@@H]1[C@@H](C(=O)OCN2C(=O)CN(CC#C)C2=O)C1(C)C)C	1.753
294*	CC(=CC1[C@@H](C(=O)OCc2cccc(Oc3ccccc3)c2)C1(C)C)C	1.794
295	CC(C)[C@@]1(O)[C@@H](OC(=O)c2ccc[nH]2)[C@@]3(O)[C@@]4(C)C[C@]5(O)O[C@@]6([C@H](O)[C@@H](C)CC[C@]46O)[C@@]3(O)[C@@]15C	1.943
296	CC(C)C(Nc1ccc(cc1Cl)C(F)(F)F)C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2	1.951
297	CC(C)N(C(C)C)C(=O)SCC(=C(Cl)Cl)Cl	1.734
298	CC(C)NC(=O)N1CC(=O)N(C1=O)c2cc(Cl)cc(Cl)c2	1.769
299	CC(C)OP(=S)(OC(C)C)SCCNS(=O)(=O)c1ccccc1	1.849
300	CC(Cl)(Cl)Cl	1.375

301	CC(Oc1ccc(Oc2ncc(cc2Cl)C(F)(F)F)cc1)C(=O)O	1.808
302	CC1(C)CON(Cc2ccccc2Cl)C1=O	1.629
303*	CC1(C)N(Br)C(=O)N(Br)C1=O	1.706
304	CC1(C)N(Br)C(=O)N(Cl)C1=O	1.633
305*	CC1(C)N(Cl)C(=O)N(Cl)C1=O	1.544
306	CC1(OC(=O)N(Nc2ccccc2)C1=O)c3ccc(Oc4ccccc4)cc3	1.823
307	Cc1cc(C)n(CO)n1	1.351
308*	Cc1cc(O)cc(C)c1Cl	1.445
309	CC1CC(OC(=O)C)OC(C)O1	1.491
310	Cc1ccc(cc1)S(=O)(=O)C(I)I	1.875
311*	CC1CCCCC1NC(=O)Nc2ccccc2	1.616
312	Cc1ccn2nc(nc2n1)S(=O)(=O)Nc3c(F)cccc3F	1.762
313	CCCC(=NOCC)C1=C(O)CC(CC(C)SCC)CC1=O	1.765
314*	CCCCC(CC)CN1C(=O)C2C3CC(C=C3)C2C1=O	1.690
315	CCCCC(CC)COC(=O)[C@@H](C)Oc1ccc(Cl)cc1Cl	1.790
316	CCCCC(CC)COC(=O)COc1ccc(Cl)cc1Cl	1.773
317	CCCCCCCCCC[N+](C)(C)CCCCCCC(C)C	1.745
318*	CCCCCCCCCc1ccc(OCCO)cc1	1.672
319*	CCCCCCNC(=N)NC(=N)N	1.518
320*	CCCCOC(=O)[C@@H](C)Oc1ccc(Oc2ccc(cc2F)C#N)cc1	1.803
321*	CCCCOCC(C)O	1.371
322*	CCCCOCCOCCOCCc1cc2OCOc2cc1CCC	1.797
323	CCCN(CCC)C(=O)SCC	1.527
324	CCCOC(=O)c1ccc(nc1)C(=O)OCCC	1.650
325*	CCCOCC(=Nc1ccc(Cl)cc1C(F)(F)F)n2ccnc2	1.789
326	CCCSC(=O)N(CCC)CCC	1.558
327	CCN(CC)C(=O)C(C)Oc1cccc2ccccc12	1.683
328	CCNc1nc(Cl)nc(NC(C)(C)C)n1	1.611
329*	CCNc1nc(NC(C)C)nc(SC)n1	1.607
330	CCOC(=O)C(C)OC(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.914
331	CCOC(=O)C(C)Oc1ccc(Oc2oc3cc(Cl)ccc3n2)cc1	1.808
332*	CCOC(=O)C(Cl)Cc1cc(N2N=C(C)N(C(F)F)C2=O)c(F)cc1Cl	1.865
333*	CCOc1nc(NC)nc(NC(=O)NS(=O)(=O)c2ccccc2C(=O)OC)n1	1.863
334	CCOCCOCCOC(=O)Nc1nc2ccccc2[nH]1	1.717
335	CCOP(=O)([O-])C(=O)N	1.432
336	CCSC(=O)N(CC(C)C)CC(C)C	1.587
337*	ClC(Cl)C(Cl)(Cl)SN1C(=O)[C@@H]2CC=CC[C@@H]2C1=O	1.793
338	Clc1cc(NC(=O)Nc2ccccc2)ccn1	1.644
339	Clc1ccc(C(Cn2ccnc2)OCC=C)c(Cl)c1	1.723
340	CN1C(=O)ON(C1=O)c2ccc(Cl)c(Cl)c2	1.667
341*	CN1SC2=C(CCC2)C1=O	1.441
342	COC(=O)c1c(CC(C)C)c(C2=NCCS2)c(nc1C(F)F)C(F)(F)F	1.848
343*	COC(=O)c1c(Cl)c(Cl)c(C(=O)OC)c(Cl)c1Cl	1.771
344*	COC(=O)c1c(Cl)nn(C)c1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	1.888
345	COC(=O)c1cccc(C)c1S(=O)(=O)NC(=O)Nc2nc(OCC(F)(F)F)nc(n2)N(C)C	1.942
346*	COC(=O)c1ccccc1N	1.429

347*	COC(=O)c1ccccc1S(=O)(=O)NC(=O)N(C)c2nc(C)nc(OC)n2	1.847
348*	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(C)nc(OC)n2	1.831
349	COC(=O)N(C(=O)N1CO[C@]2(Cc3cc(Cl)ccc3C2=N1)C(=O)OC)c4ccc(OC(F)(F)F)cc4	1.972
350*	COC[C@H](C)N(C(=O)CCl)c1c(C)csc1C	1.690
351*	COc1cc(OC)nc(Oc2cccc(Oc3nc(OC)cc(OC)n3)c2C(=O)[O-])n1	1.883
352	COc1cc(OC)nc(Sc2cccc(Cl)c2C(=O)[O-])n1	1.763
353	COc1nc(NC(C)C)nc(NC(C)C)n1	1.603
354	COCC(C)N(C(=O)CCl)c1c(C)csc1C	1.690
355*	CS\C(=N\OC(=O)N(C)SN(C)C(=O)ON=C(C)SC)\C	1.799
356	CSC(=O)c1c(CC(C)C)c(C(=O)SC)c(nc1C(F)F)C(F)(F)F	1.853
357	CSc1nc(NC2CC2)nc(NC(C)(C)C)n1	1.654
358*	Fc1cc2OCC(=O)N(CC#C)c2cc1N3C(=O)C4=C(CCCC4)C3=O	1.799
359	Fc1ccc(Oc2ccnc3cc(Cl)cc(Cl)c23)cc1	1.739
360	Nc1c(Cl)c(F)nc(OCC(=O)O)c1Cl	1.656
361	Nc1nc(N)nc(NC2CC2)n1	1.470
362*	O=C\C=C\c1ccccc1	1.371
363*	OC(=O)c1ccccc1	1.337
364	OC(=O)COc1ccc(Cl)cc1Cl	1.594
365	OC(=O)COc1nc(Cl)c(Cl)cc1Cl	1.659
366*	Oc1ccc(cc1)[N+](=O)[O-]	1.393
367*	Oc1ccc(Cl)cc1Cc2ccccc2	1.590
368	Oc1ccccc1c2ccccc2	1.481
369	OCNC(=O)N(CO)C1N(CO)C(=O)N(CO)C1=O	1.694
370	OCNCC(=O)[O-]	1.267
371	C\C(=N/NC(=O)Nc1cc(F)cc(F)c1)c1ncccc1C([O-])=O	1.773
372	CC(=NNC(=O)Nc1cc(F)cc(F)c1)c1ncccc1C(O)=O	1.774
373*	CC(C)=C[C@@H]1[C@@H](C(=O)O[C@H]2CC(=O)C(CC=C)=C2C)C1(C)C	1.730
374	CC(C)CCCCCOC(=O)C(C)Oc1ccc(Cl)cc1Cl	1.790
375*	CCCCCCCCCCCC(=O)O	1.552
376	CCCCCOC(=O)COc1cc(N2C(=O)C3=C(CCCC3)C2=O)c(F)cc1Cl	1.877
377*	CCCCOCCOCCOCc1cc2OCOc2cc1CCC	1.779
378	ClC(Cl)C(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	1.793
379*	COC(=O)c1ccccc1CS(=O)(=O)NC(=O)Nc1nc(OC)cc(OC)n1	1.863
380	Cl\C=C\C[N+]12CN3CN(CN(C3)C1)C2	1.583
381	CCC(C)(CCC(C)C)C(=O)NC	1.513
382*	CC(C)C1CCC(Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	1.766
383*	COCc1c(F)c(F)c(COC(=O)[C@@H]2[C@@H](C=CC)C2(C)C)c(F)c1F	1.796
384*	CC(C)CC(C)c1sccc1NC(=O)c2cn(C)nc2C(F)(F)F	1.792
385	COC(=O)Nc1cccc(OC(=O)Nc2cccc(C)c2)c1	1.714
386	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(Cl)c1COCC(F)(F)F	1.881
387*	FC(F)(F)c1cnc(CNC(=O)c2c(Cl)cccc2Cl)c(Cl)c1	1.821
388*	CC(=O)Nc1ccc(O)cc1	1.408
389*	CCc1cc(C)cc(CC)c1C2=C(OC(=O)C(C)(C)C)N3CCOCCN3C2=O	1.826
390	CCC(=NOC\C=C\Cl)C1=C(O)CC(CC1=O)C2CCOCC2	1.755

391	COc1cc(C)c(C(=O)c2c(C)c(Br)ccc2OC)c(OC)c1OC	1.828
392*	COc1cc(OCC#C)ccc1CCNC(=O)C(OCC#C)c2ccc(Cl)cc2	1.830
393	Nc1c(Cl)cc(cc1Cl)[N+](=O)[O-]	1.524
394*	OC(c1ccc(Cl)cc1)(c2cncnc2)c3ccccc3Cl	1.724
395	CCCCCCCCSCCO	1.463
396	CN(\C=N\c1ccc(C)cc1C)\C=N\c2ccc(C)cc2C	1.622
397	COC(=O)C[C@@H]1[C@@]2(C)[C@H](O[C@@H]3C[C@H](C(C)=C23)c2ccoc2)[C@@H]2OC[C@@]3(C)[C@H]2[C@]1(C)[C@H](C[C@H]3OC(C)=O)OC(=O)C(C)=CC	1.930
398	CCON=C(CC)C1=C(O)CC(CC1=O)c2c(C)cc(C)cc2C	1.648
399	CC1(C)[C@@H]([C@@H](Br)C(Br)(Br)Br)[C@H]1C(=O)O[C@H](C#N)c2cccc(Oc3ccccc3)c2	1.934
400	CN1CCCC1	0.999
401	CC1(C)C(C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2)C1(C)C	1.587
402	[O-]N1C=CC=CC1=S	1.141
403	BrCC(Br)(CCC#N)C#N	1.424
404	C[N+]1(C)CCCCC1	1.057
405	CC(C)(NC(=O)c1cc(Cl)cc(Cl)c1)C#C	1.408
406	CC(Oc1ccc(Cl)cc1Cl)C(=O)O	1.371
407	CC(Oc1cccc(Cl)c1)C(=O)[O-]	1.300
408	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	1.416
409	CCC(C)Nc1c(cc(cc1[N+](=O)[O-])C(C)(C)C)[N+](=O)[O-]	1.470
410	CCc1cccc(C)c1N(C(C)COC)C(=O)CCl	1.453
411*	CCCCCCCCN1SC(=C(Cl)C1=O)Cl	1.450
412	CCCCOC(=O)COc1ccc(Cl)cc1Cl	1.442
413	CCCCOCCOC(=O)C(C)Oc1ccc(Cl)cc1Cl	1.525
414	CCCN(CCC)c1c(cc(c(N)c1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.544
415	CCN(CC)c1c(cc(c(N)c1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.508
416	CCOC(=O)Cc1cccc2ccccc12	1.330
417	ClC\C=C\Cl	1.045
418*	Clc1c(Cl)c(C#N)c(Cl)c(C#N)c1Cl	1.424
419	CN(C)NC(=O)CCC(=O)O	1.204
420	CNC(=O)O\N=C\C(C)(C)S(=O)(=O)C	1.346
421	CNC1=C(Cl)C(=O)N(N=C1)c2cccc(c2)C(F)(F)F	1.482
422*	COC(C)(C)CCCC(C)C\C=C\C(=C\C(=O)OC(C)C)\C	1.492
423	COc1c(Cl)ccc(Cl)c1C(=O)O	1.344
424	COCC(=O)N(C(C)C(=O)OC)c1c(C)cccc1C	1.446
425	O=C(C(c1ccccc1)c2ccccc2)C3C(=O)c4ccccc4C3=O	1.531
426	O=C1NNC(=O)C=C1	1.049
427	OC(=O)c1ccccc1C(=O)Nc2cccc3ccccc23	1.464
428*	OCC(Br)(CO)[N+](=O)[O-]	1.301
429	CCC(O)=O	0.869
430	CCCCCCCCCCO	1.199
431*	OCOCC12COCN1COC2	1.243
432	CC(C)C1=NN(C(=O)NC(C)(C)C)C(=O)N1N	1.359
433	CCCCC(O)(Cn1cncn1)c2ccc(Cl)cc2Cl	1.470

434	CCSC(=O)N1CCCCCC1	1.158
435	CCOC(=O)NCCOc1ccc(Oc2ccccc2)cc1	1.178
436	CS(=O)(=O)NC(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+]([O-])=O	1.341
437	Clc1ccccc1c2nnc(nn2)c3ccccc3Cl	1.166
438	CCCCOC(=O)C(C)Oc1ccc(Oc2ccc(cn2)C(F)(F)F)cc1	1.185
439	CN(C)C(=O)C(c1ccccc1)c2ccccc2	0.901
440	CSc1nc(NC(C)C)nc(NC(C)C)n1	0.751
441	COC(=O)NS(=O)(=O)c1ccc(N)cc1	0.487
442	CCOCn1c(c2ccc(Cl)cc2)c(C#N)c(Br)c1C(F)(F)F	4.675
443	CCOP(=S)(OCC)Oc1cc(C)nc(n1)C(C)C	3.978
444	CCOP(=O)(SC(C)CC)N1CCSC1=O	3.355
445	CCNS(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	3.504
446	CCOP(=S)(Oc1ccc(cc1)[N+](=O)[O-])c2ccccc2	3.284
447	CN(C)c1ccc(cc1)N=NS(=O)(=O)[O-]	2.967
448	[O-]S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	3.187
449	Clc1ccc(cc1)C(C(=O)C2C(=O)c3ccccc3C2=O)c4ccccc4	2.944
450	CNC(=O)Oc1cccc2OC(C)(C)Oc12	2.670
451	CNC(=O)O\N=C\C(C)(C)SC	2.505
452	CCN(CC)c1nc(C)cc(OP(=S)(OC)OC)n1	2.683
453	CCNc1nc(Cl)nc(NC(C)(C)C#N)n1	2.559
454	COP(=O)(N)SC	2.221
455	Cl[C@@H]1C[C@H]2[C@@H]([C@H]1Cl)[C@]3(Cl)C(=C(Cl)[C@]2(Cl)C3(Cl)Cl)Cl	2.679
456	CC(=O)CC(C1=C([O-])c2ccccc2OC1=O)c1ccccc1	2.538
457	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)cc2)cc1	2.717
458	OC(=O)CCCOc1ccc(Cl)cc1Cl	2.396
459	CNC(=O)CSP(=S)(OC)OC	2.355
460	CCS(=O)(=O)c1cccnc1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	2.495
461	CNC(=O)Oc1cccc(c1)\N=C\N(C)C	2.194
462	Clc1cccc(n1)C(Cl)(Cl)Cl	2.197
463	COC(CN(C)C(=O)Nc1nnc(s1)C(C)(C)C)OC	2.261
464	CCOP(=S)(OCC)SCN1C(=O)Oc2cc(Cl)ccc12	2.345
465	OC(=O)C1(CC1)C(=O)Nc2ccc(Cl)cc2Cl	2.356
466	CCNP(=S)(OC)O\C(=C\C(=O)OC(C)C)\C	2.198
467	COP(=S)(OC)SCN1N=Nc2ccccc2C1=O	2.213
468*	CCCC(=O)Oc1c(Br)cc(cc1Br)C#N	2.103
469*	Clc1cc(Cl)cc(c1)c2cc(Cl)cc(Cl)c2	2.018
470	[S-]C#N	1.309
471*	Oc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.849
472	CSC1=NN=C(C(=O)N1N)C(C)(C)C	1.729
473	CC1(C)COCN1	1.402
474*	CCCN(CC)CC1COC2(CCC(CC2)C(C)(C)C)O1	1.853
475	CCOCN(C(=O)CCl)c1c(C)cccc1CC	1.810
476*	CS(=O)(=O)c1cc(ccc1C(=O)c2cnoc2C3CC3)C(F)(F)F	1.926
477	CC1(C)CNC(=NN=C(\C=C\c2ccc(cc2)C(F)(F)F)\C=C\c3ccc(cc3)C(F)(F)F)NC1	2.055

478*	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2cc(NC=O)ccc2C(=O)N(C)C)n1	2.008
479	CCOC(=O)C(SP(=S)(OC)OC)c1ccccc1	1.852
480	Cc1cc(ccc1NC(=O)c2cccc(I)c2C(=O)NC(C)(C)CS(=O)(=O)C)C(F)(C(F)(F)F)C(F)(F)F	2.177
481*	CCC(CC)Nc1c(cc(C)c(C)c1[N+](=O)[O-])[N+](=O)[O-]	1.782
482	CCCCCCCCO	1.994
483*	[O-][N+](=O)c1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.771
484	CC(C)(C)C(O)C(Oc1ccc(Cl)cc1)n2cncn2	1.772
485	CC1(CCCCC1)C(=O)Nc2ccc(O)c(Cl)c2Cl	1.781
486	CC1C(OC(=O)C2C(C=C(C)C)C2(C)C)C=C(CC=CC=C)C1=O	1.817
487	Cc1ccc2nc3SC(=O)Sc3nc2c1	1.670
488*	CCC(C)(CC)c1cc(NC(=O)c2c(OC)cccc2OC)on1	1.822
489	CCc1cnc(C2=NC(C)(C(C)C)C(=O)N2)c(c1)C(=O)O	1.762
490*	CCCC(C)c1cccc(OC(=O)NC)c1	1.646
491	CCCCNC(=O)OCC#CI	1.749
492	CCCCOC(=O)c1ccccc1C(=O)OCCCC	1.745
493	CCCCSP(=O)(SCCCC)SCCCC	1.798
494	CCN1C(=CC(=O)C(=C1c2ccc(Cl)cc2)C(=O)[O-])C	1.764
495	CCOC(=O)C1=NN(c2ccc(Cl)cc2Cl)C(C)(C1)C(=O)OCC	1.873
496	ClC1=C(Cl)C(=O)c2ccccc2C1=O	1.657
497	ClC1=C(Cl)C(=O)SS1	1.573
498	CN1CCCC1=O	1.297
499*	COc1cc(Cl)c(OC)cc1Cl	1.617
500*	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2CCC(F)(F)F)n1	1.923
501	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2OCCCl)n1	1.905
502	COP(=O)(OC)OC(=CCl)c1cc(Cl)c(Cl)cc1Cl	1.864
503	NC#N	0.924
504	OC(=O)C1C2CCC(O2)C1C(=O)O	1.570
505*	Oc1ccc(c(c1)C(F)(F)F)[N+](=O)[O-]	1.617
506	CC(Oc1ccc(Cl)cc1C)C(O)=O	1.632
507	ClNc1nc(NCl)nc(NCl)n1	1.661
508	FC(C(F)(F)F)C(F)(F)Oc1cc(Cl)c(NC(=O)NC(=O)c2c(F)cccc2F)cc1Cl	2.00786164
509*	CC(C)(C)c1ccc(CCOc2ncnc3ccccc23)cc1	1.784
510	COC1CC(NC(NC(=O)NS(=O)(=O)c2c(Cl)nc3ccccn23)N1)OC	1.9131
511	Clc1ccc(CN2CCSC2=NC#N)cn1	1.695
512*	CC(C)Nc1nc(Cl)nc(NC(C)C)n1	1.650
513	CC[C@H]1CCCC(O[C@H]2CC[C@@H]([C@@H](C)O2)N(C)C)[C@@H](C)C(=O)C2=C[C@H]3[C@@H]4C[C@@H](C[C@H]4C=C[C@H]3[C@@H]2CC(=O)O1)O[C@@H]1O[C@@H](C)[C@H](OC)[C@@H](OC)[C@H]1OC	2.152
514	CCOC(=O)C1CC(=O)C(=C(O)C2CC2)C(=O)C1	1.685
515	FC(F)C(F)(F)Oc1c(Cl)cc(NC(=O)NC(=O)c2c(F)cccc2F)cc1Cl	1.947
516	CCCCN(CC)C(=O)SCCC	1.591
517*	CC(F)c1nc(N)nc(N[C@@H]2[C@@H](C)Cc3ccc(C)cc23)n1	1.761

518	CC(C)N(C)S(=O)(=O)NC(=O)c1cc(N2C(=O)C=C(N(C)C2=O)C(F)(F)F)c(F)cc1Cl	1.977
519*	COc1ccc(cc1OC)\C(=C\C(=O)N2CCOCC2)\c3ccc(Cl)cc3	1.864
520	FC(C(F)(F)F)C(F)(F)Oc1c(Cl)cc(NC(=O)NC(=O)c2c(F)cccc2F)c(F)c1Cl	1.999
521	OC(CN1NC=NC1=S)(Cc2ccccc2Cl)C3(Cl)CC3	1.791
522	COCc1cnc(C2=NC(C)(C(C)C)C(=O)N2)c(c1)C(=O)O	1.738
523*	CCN(CC)CCOCc1ccc(C)cc1	1.596
524*	[O-]C(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.807
525*	[O-]c1ccccc1c2ccccc2	1.478
526	CC(=C[C@H]1[C@H](C(=O)O[C@@H]2CC(=O)C(=C2C)CC#C)C1(C)C)C	1.728
527	CC(Oc1cccc(Cl)c1)C(=O)O	1.552
528	CC1=NN(C(=O)N1C(F)F)c2cc(NS(=O)(=O)C)c(Cl)cc2Cl	1.838
529	CCc1cccc(C)c1N([C@@H](C)COC)C(=O)CCl	1.703
530	CCCC1COC(Cn2cncn2)(O1)c3ccc(Cl)cc3Cl	1.784
531	CCCCCCC(C)OC(=O)COc1nc(F)c(Cl)c(N)c1Cl	1.815
532	CCCCCCCCCC(=O)C	1.481
533*	CCCCCCCCCC[N+](C)(C)CCCCCCCCCC	1.764
534*	CCNC(=O)NC(=O)\C(=N\OC)\C#N	1.547
535	CCOC(=O)C(O)(c1ccc(Cl)cc1)c2ccc(Cl)cc2	1.762
536	CCOC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(Cl)cc(OC)n2	1.868
537	Clc1cc(Cl)cc(c1)C2(CC(Cl)(Cl)Cl)CO2	1.755
538	CN1C(=O)N(C(=O)C=C1C(F)(F)F)c2ccc(Cl)c(c2)C(=O)OC(C)(C)C(=O)OCC=C	1.926
539*	COC(=O)c1sccc1S(=O)(=O)NC(=O)Nc2nc(C)nc(OC)n2	1.838
540	COCC(=O)N(N1CCOC1=O)c2c(C)cccc2C	1.694
541	NC(=O)C(Br)(Br)C#N	1.633
542	NC(=O)N	1.028
543*	Oc1ccc(cc1)C(=O)CBr	1.582
544*	CC(C)=C[C@H]1[C@H](C(=O)O[C@@H]2CC(=O)C(CC=C)=C2C)C1(C)C	1.730
545	CCCCCCCCC(O)=O	1.449
546	CCN(Cc1cccc(c1)S([O-])(=O)=O)c1ccc(cc1)C(=C1C=CC(C=C1)=[N+](CC)Cc1cccc(c1)S([O-])(=O)=O)c1ccccc1S([O-])(=O)=O	2.123
547	CN(C)[C@H]1[C@@H]2[C@@H](O)[C@H]3C(=C(O)[C@]2(O)C(=O)C(C(N)=O)=C1O)C(=O)c1c(O)cccc1[C@@]3(C)O	1.913
548	CC1(OC(=O)N(C1=O)c2cc(Cl)cc(Cl)c2)C=C	1.706
549*	Cn1cc(C(=O)Nc2ccccc2C3CC3C4CC4)c(n1)C(F)F	1.765
550	COc1cccc(C(=O)NN(C(=O)c2cc(C)cc(C)c2)C(C)(C)C)c1C	1.797
551	CCOC(=O)OC1=C(C(=O)N[C@@]12CC[C@@H](CC2)OC)c3cc(C)ccc3C	1.790
552*	CCCC(C)C1(CC=C)C(=O)NC(=NC1=O)[O-]	1.495
553	[S-]C(=NC#N)[S-]	1.065
554	BrCC(=O)OCc1ccccc1	1.359
555	CC(C)(C)C(=O)C(Oc1ccc(Cl)cc1)n2cncn2	1.468
556*	CCCC\C=C\CCC=CCCCCCCOC(=O)C	1.447
557	CCCCNC(=O)n1c(NC(=O)OC)nc2ccccc12	1.462
558	Oc1nc(O)nc(O)n1	1.110

559	[O-][N+](=O)c1cc(c(Cl)c(c1Nc2ncc(cc2Cl)C(F)(F)F)[N+](=O)[O-])C(F)(F)F	1.642
560	CC(C)N1C(=C2C=CC=CC2=NS1(=O)=O)[O-]	1.318
561	CC1(C)[C@H](\C=C(/Cl)\C(F)(F)F)[C@@H]1C(=O)O[C@H](C#N)c2cccc(Oc3ccccc3)c2	1.556
562	CC(C)(C)[C@H](O)[C@H](Cc1ccc(Cl)cc1)n2cncn2	1.167
563	CNC(=O)Oc1cc(C)c(C)cc1Cl	0.727
564	OCC(CO)(CO)[N+](=O)[O-]	0.276

***Test set compounds**

3.2.2 Descriptor calculation & data preprocessing

Descriptors are the numerical representation in which we correlate the chemical structure with any physiochemical property/biological activity/ toxicity. In this study, a total of 9 classes of descriptors were calculated utilizing AlvaDesc 2.02 (https://www.alvascience.com/alvadesc/) [104]. In each dataset, the defective and inter-correlated chemical descriptors were eliminated by V-WSP1.2 (http://teqip.jdvu.ac.in/QSAR_Tools/) software with a standard deviation less than 0.0001 or correlation coefficient greater than 0.95.

3.2.3 Dataset splitting

Dataset division is crucial for QSTR model development. Normally, training set compounds are used to develop the model and test compounds for validation. The validation set is used to assess the model performance and fine-tune the parameters of the model. It tells us how well the model is learning and adapting, allowing for adjustments and optimizations to be made to the model's parameters and hyperparameters (the latter in the case of machine learning-based models) before it is finally tested. The test data set mirrors real-world data the model has never seen before, i.e.: a separate sample of unseen data. Its primary purpose is to offer a fair and final assessment of how the model would perform when it encounters new data in a live, operational environment. This is especially critical to evaluate models effectively along with preventing overfitting [105]. We performed dataset division of four datasets by using rational methods such as the Kennard stone, activity property-based, and Euclidean distance method using Dataset Division GUI 1.2 software as well as using random division method [106]. We also employed modified k-medoid clustering by using Modified k-Medoid 1.3 (http://teqip.jdvu.ac.in/QSAR_Tools/) [107]. After that, the final selection of data-set division methods was done based on the statistical results. The best results come in the Kennard stone method for the MD and JQ data set, the activity property-based method for the BQ dataset, and the random division method for the RNP dataset. In this process of dataset division, the datasets are divided into 75:25 ratios of training and test sets compounds [108].

3.2.4 Selection of features and model building

In the case of model building, feature selection is one of the important phases by which we can find significant descriptors to boost the interpretability and predictive ability of the model [109]. Primarily, we performed a step-wise regression method and genetic algorithm (GA) approaches for feature selection [110] and then we employed the regression-based partial least square (PLS) [111] method through Partial least squares v1.0 tool (http://teqip.jdvu.ac.in/QSAR_Tools/) for model generation.

3.2.5 Validation metrics of QSTR models

A significant step in the formation of a QSTR model is statistical validation, which establishes it's reliability and predictivity [54]. Various internal validation parameters were calculated which involve the determination coefficient (R^2), and leave-one-out (LOO) cross-validated correlation coefficient (Q^2_{LOO}) to judge the reliability and importance of the model. External validation parameters demonstrate the predictivity of QSTR models. The model's external validation is determined using parameters such as $Q^2_{F1}$and Q^2_{F2} [112]. For both internal (${Q^2}_{LOO}$) and external predictive parameters (Q^2_{F1},Q^2_{F2}), the approved threshold value is 0.5.

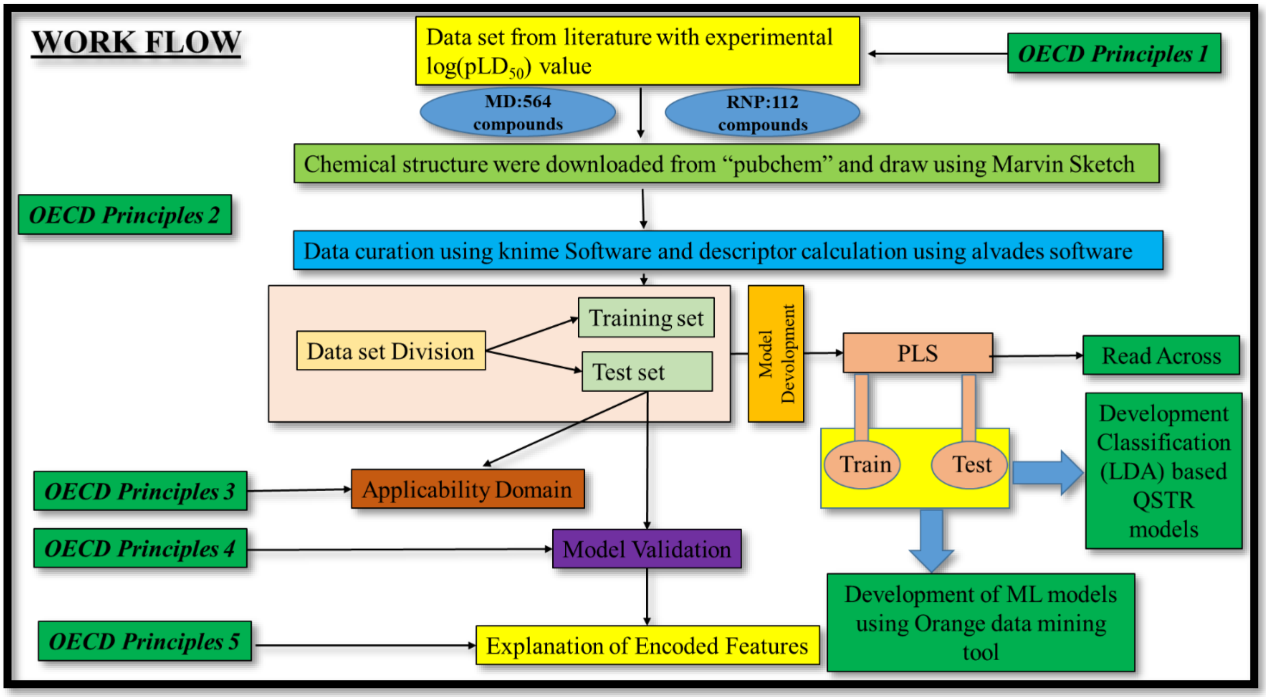

WORK FLOW
Data set from literature with experimental log(pLD$_{50}$) value
OECD Principles 1
MD:564 compounds
RNP:112 compounds
Chemical structure were downloaded from "pubchem" and draw using Marvin Sketch
OECD Principles 2
Data curation using knime Software and descriptor calculation using alvades software
Data set Division
Training set
Test set
Model Development
PLS
Read Across
Train
Test
Development Classification (LDA) based QSTR models
OECD Principles 3
Applicability Domain
OECD Principles 4
Model Validation
Development of ML models using Orange data mining tool
OECD Principles 5
Explanation of Encoded Features

Fig. 4. Workflow of QSTR model development.

3.2.6. Prediction using read-across algorithm

According to the fundamental tenet of read-across, substances with similar chemical structures will also have comparable attributes and it is not utilized in the model development process [113]. Read-across prediction is a similarity-based non-testing technique that is widely used in eco-toxicological data-gap filling. Initially, the training set of the best model was split into sub-training and sub-test sets. These sets were again used to optimize the hyperparameters through Read-Across-v3.1 (http://teqip.jdvu.ac.in/QSAR_Tools/). After similarity-based sorting, similarity threshold values (0 to 1), various distance threshold values (1 to 0), and the numbers of most similar training compounds (2 to 10) were applied. The best setting of hyperparameters obtained from sub-training and sub-test was applied to the original training and test set for the final prediction [114].

3.2.7. Applicability domain study of the model

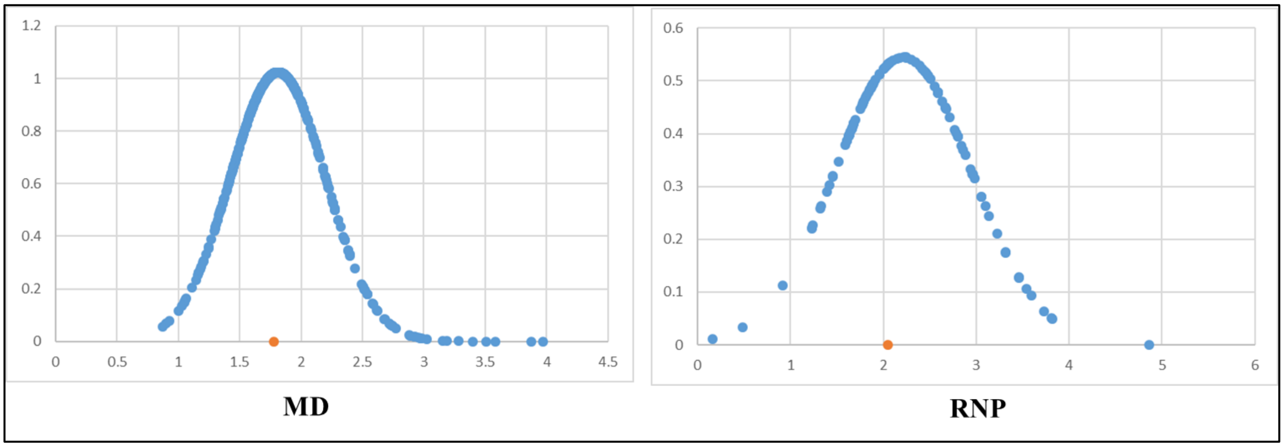
The applicability domain (AD) of a QSAR model has been defined as the chemical structure and response space, considered by the properties of the molecules in the training set [54]. The AD expresses the fact that QSARs are undeniably associated with restrictions in the categories of physicochemical properties, chemical structures, and mechanisms of action for which the models can generate reliable predictions. In the current study, distance to the model in X-space (DModx) has been utilized for AD estimation of constructed PLS models which rely on residuals of response and predictive variables [115].

3.2.8 Y-randomization study

Y-randomization study was carried out to check the chance correlation of the QSTR models with the help of SIMCA-P software [116]. In the Y-randomization test, the descriptor matrix X is kept constant but only the vector Y is scrambled randomly, and a new model is developed using the same set of descriptors. The original model is considered as robust if its validation metrics are better than the random models [117]. The values of the R^2y_{rand} intercept and Q^2y_{rand} intercept should not be more than 0.3 and 0.05 respectively.

3.2.9 Analysis of parametric assumptions of the developed models

To ensure that our model is reliable we carried out some diagnostic tests to check for the existence of multi-collinearity, normal distribution, and homoscedasticity [118-119]. Multicollinearity is when the predictor variables within a regression model are highly correlated with each other, leading to inaccurate results in regression analysis. To identify multicollinearity, we used the variation inflation factor which is a widely used metric. If the VIF is higher than 5, multicollinearity is considered to be present [120]. In statistical regression models, exhibiting multicollinearity can lead to misleading results. For each modeled descriptor, we found that the VIF values were very close to 1. So, it can be concluded that all the independent variables are not collinear with the dependent variable. The function values follow a multidimensional normal distribution with a mean and covariance matrix that depends on the descriptor vectors. We have plotted the normal distribution curve for each (MD and RNP) avian species and provided in **Figs. 5-6**. Homoscedasticity refers to the equal variance of an error in a regression model assessed using the Breusch-Pagan test in our study. A p-value of more than 0.05 indicates the homoscedasticity of the model. In our study, the calculated p-values were not less than 0.05 (0.093-0.209) for all the developed models. Therefore, we fail to reject the null hypothesis, and the model can be considered homoscedastic in nature. All the statistical results of Homoscedasticity and multicollinearity (VIF) for each model are provided in **Tables 6-7**.



MD

RNP

Fig. 5. Normal distribution curve of MD and RNP.

Table 6. Variance Inflation Factor (VIF) results for each model.

MD		**RNP**	
Variables	VIF	Variables	VIF
MW	1.1	X2A	1.0
C-012	1.1	nRCONHR	1.0
B07[O-P]	1.2	nN(CO)2	1.0
Br-094	1.0	B04[C-P]	1.2
B05[C-P]	1.2	B05[P-Cl]	1.2
F04[C-Cl]	1.1	F03[O-S]	1.0

Table 7. Homoscedasticity test results for each model.

Metrics	**MD**	**RNP**
P-value	0.158	0.093
Test statistics	9.29	7.96

3.2.10 Application of other machine learning (ML) algorithms

To estimate the prediction performance of other algorithms, we have employed two different state–of–the–art ML algorithms namely support vector machine (SVM) and Random forest (RF) using the orange data mining tool [120]. The hyperparameters were adjusted to tune the model for optimal performance. The prediction qualities of the ML models were evaluated in terms of R^2, Q^2 $_{\text{Loo,}}$ and MAE values.

3.2.11 Classification QSTR (LDA-QSTR) model development

In the present work, we have developed a classification-based linear discriminant analysis (LDA) QSTR model from the selected set of features and evaluated its performance for its predictive ability. The model development is done using Classification-Based QSAR_v1.0.0 tools (available at http://teqip.jdvu.ac.in/QSAR_Tools/). The model was extensively validated based on different internal and external classification metrics (area under the ROC curve (AUC), accuracy, precision, sensitivity, F-measure, and Matthews correlation coefficient (MCC) [121-122].

3.2.12. Screening of the Pesticide Properties Database

We have collected 1903 chemical data from the Pesticide Properties Database (PPDB) available at (http://sitem.herts.ac.uk/aeru/ppdb/). Knime curation was done to remove duplicates, inorganic salts, and mixtures using the KNIME workflow. Due to the knime curation, some compounds were removed. After the curation, the remaining 1694 compounds were used for the screening process to check the developed model's reliability. The descriptors for these molecules were calculated using the same procedure as in the QSAR modeling process. The predictions were made through the use of individual QSTR models with the help of the PRI (Prediction Reliability Indicator) tool (http://teqip.jdvu.ac.in/QSAR_Tools/). PRI tool categorizes the predictions into three distinct groups: good (composite score 3), moderate (composite score 2), and bad (composite score 1). Additionally, the tool determines the localization of compounds inside the AD. The screened compounds were ranked on the basis of their predicted toxicity and the twenty highest and least toxic compounds which exhibited toxicity towards all four avian species were analysed. The results were further validated extensively based on experimental data reported previously, to establish the real-world applicability of the developed QSTR model.

3.3 Study 2

3.3.1 Dataset preparation

The pesticide toxicity data for California quail were extracted from the EPA ECOTOX database (https://ecotox.ipmcenters.org). The toxicity end-point values range from -0.99 to 2.50. The collected data were curated carefully to eliminate the inorganic salts and organ-metallic compounds from the initial dataset to maintain homogeneity [123]. We have used the remaining 35 compounds with the definite endpoint (LD_{50}) for the model development. For ease of interpretation, the toxicity endpoint values (LD_{50}) were transformed to a negative logarithmic scale (pLD_{50}). The molecular structures of the compounds were drawn by Marvin sketch software with the addition of explicit hydrogen atoms and proper aromatization.

Table 8. Compounds smiles name with respective experimental pLD_{50} values.

Sl. No.	**Smiles**	**pLD_{50}**
1	COP(=O)(OC)OC(=CC(=O)N(C)C)C	2.098
2*	CCOP(=S)(OCC)OC1=NC(Cl)=C(Cl)C=C1Cl	0.710
3	COP(=S)(OC)OC1=CC=C(C=C1)SC2=CC=C(C=C2)OP(=S)(OC)OC	1.392
4*	ClC1=CC=C(C=C1)C(C1=CC=C(Cl)C=C1)C(Cl)(Cl)Cl	-0.224
5	CCC(=O)OC(C(Cl)(Cl)Cl)P(=O)(OC)OC	0.723
6	CC1=C(C=CC(=C1)OP(=S)(OC)OC)SC	1.268
7*	CCOP(=O)(OCC)SC1=CC=C(C=C1)[N+](=O)[O-]	1.236
8	C1CN2CC3=CCOC4CC(=O)N5C6C4C3CC2C61C7=CC=CC=C75	0.475
9	C1C2C(C(C1Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	1.463
10*	C(C(=O)[O-])F	1.221
11	CNC(=O)OC1=CC=CC2=CC=CC=C21	-0.997
12	C1C2C3C(C1C4C2O4)C5(C(=C(C3(C5(Cl)Cl)Cl)Cl)Cl)Cl	2.505
13	CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC.COC1=CC=C(C=C1)C(C2=CC=C(C=C2)OC)C(Cl)(Cl)Cl	-0.471
14	C1=CC(=CC=C1C(C2=CC=C(C=C2)Cl)C(Cl)Cl)Cl	-0.375
15	C1=CC(=C(C=C1O)C(F)(F)F)[N+](=O)[O-]	-0.420
16	CC(C)OC1=CC=CC=C1OC(=O)NC	0.907
17*	CCOP(=S)(OCC)OC1=CC=C(C=C1)S(=O)C	2.413
18	CNC(=O)O/N=C/C(C)(C)SC	1.610
19	P(SCCS(CC)=O)(OC)(OC)=O	0.713
20	CC1=CC(=CC(=C1N(C)C)C)OC(=O)NC	1.493
21	CNC(=O)OC1=C2C=CSC2=CC=C1	-0.349
22	CNC(=O)OC1=C2C=CSC2=CC=C1	-0.732
23*	CCOP(=S)(C1=CC=CC=C1)OC2=CC=C(C=C2)[N+](=O)[O-]	0.949
24	COP(=S)(OC)OC1=NC(=C(C=C1Cl)Cl)Cl	-0.015
25	CC1(C2C(C(C1(C(C2Cl)Cl)C(Cl)Cl)Cl)Cl)C(Cl)Cl	1.242
26*	CCOP(=O)(OCC)SCCSCC	1.386

27	CC(=CC1C(C1(C)C)C(=O)OCC2=COC(=C2)CC3=CC=CC=C3)C	-0.771
28	CC1=CC(=C(C(=C1)OC(=O)NC)C)C	-0.003
29	F[As-](F)(F)(F)(F)F	-0.083
30	COP(=S)(C1=CC=CC=C1)OC2=CC(=C(C=C2Cl)Br)Cl	1.149
31	CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)SC)C	2.219
32*	CC1(C(C1C(=O)OCC2=COC(=C2)CC3=CC=CC=C3)C=C4CCCC4)C	0.418
33	CNC(=O)OC1=CC=CC=C1CCCSC	0.366
34	CC(=NOC(=O)NC)SCCC#N	0.888

***Test set compounds**

3.3.2 Calculation of descriptor & data pretreatment

Molecular descriptors are the numerical representation of chemically comprised values that correlate molecular structure with physicochemical or biological properties [124]. In this current work, we have computed various 2D descriptors such as constitutional indices, ring descriptors, topological indices, connectivity index, functional group counts, atom-centered fragments, atom type E-states, 2D atom pairs molecular properties, and ETA indices using AlvaDesc software version 2.02 (https://www.alvascience.com/alvadesc/) [104]. The unnecessary descriptors (descriptors with a fixed value, highly inter-correlated descriptors, low diverse descriptors, etc.) were removed by employing the data pretreatment tool V-WSP v1.2 (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab) [106] with inter-correlated coefficient |r| > 0.95 and variance < 0.0001. were eliminated using the data preprocessing technique.

3.3.3 Splitting of dataset

Dataset splitting into training and test sets ensures the predictivity of the model during model development. In our present study, the data-set splitting was performed using various dataset division methods, such as Kennard stone, Euclidean distance, activity property based [106] and modified k-medoid clustering technique [107] by using “Dataset Division GUI” version 1.2 and “Modified k-Medoid” version 1.3 software tool correspondingly (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab). However, the best result was obtained from the Euclidean distance division.

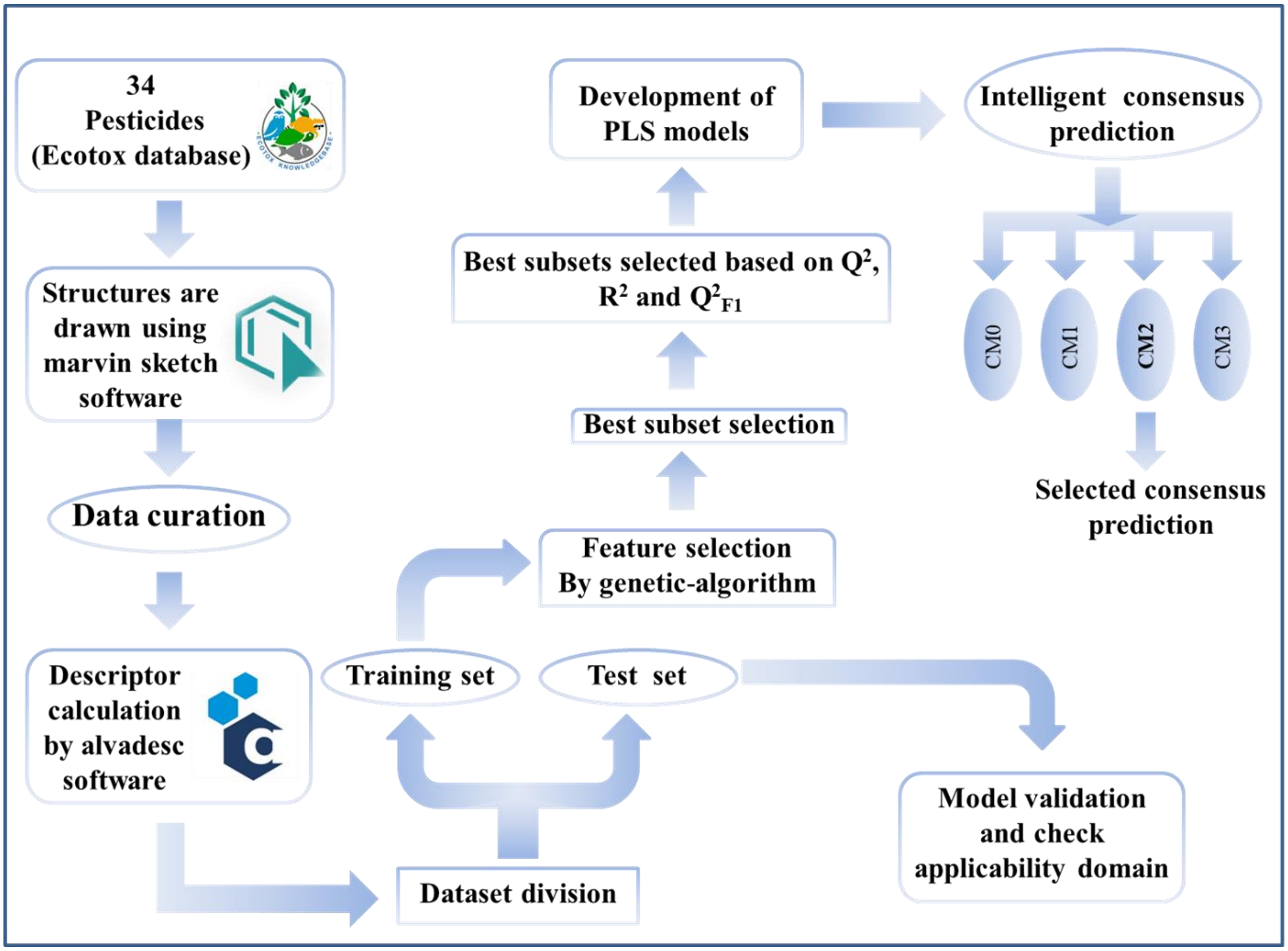

34
Pesticides
(Ecotox database)
Structures are drawn using marvin sketch software
Data curation
Descriptor calculation by alvadesc software
Dataset division
Training set
Test set
Feature selection
By genetic-algorithm
Best subset selection
Best subsets selected based on Q2, R2 and Q2F1
Development of PLS models
Intelligent consensus prediction
CM0
CM1
CM2
CM3
Selected consensus prediction
Model validation and check applicability domain

Fig. 6. Schematic depiction of QSTR model generation.

3.3.4 Feature selection & model generation

Selection of appropriate descriptors constitutes a crucial step before model generation feature selection shrinks the original variable set to obtain a variable sub-set by removing redundant and irrelevant variables, which improves the model's interpretability and predictivity [125]. In the present study, the variable selection was performed using stepwise regression (using Minitab 14 software) [126] and Genetic algorithm (GA) (employing Genetic algorithm 4.1 tools) [106]. The obtained reduced pool of descriptors was subjected to Best-Subset selection2.1 [127] to identify the most significant descriptors for model building. In the current work, the PLS regression approach was adopted to construct the final QSTR models.

3.3.5 Statistical validation of the constructed model

In this work, various statistical validation approaches are employed for measuring robustness and prediction accuracy to establish the significance and reliability of the constructed model using both internal and external validation metrics. For statistical quality assessment as well as internal validation, we calculated metrics such as the determination coefficient (R^2) and leave-one-out cross-validated correlation coefficient ($Q^2_{(LOO)}$) [127]. Internal validation metrics are not enough to assess the performance of the developed model in terms of robustness and predictive ability; therefore, we also validated the predictions for test set compounds using various external validation parameters such as $Q^2_{F1,}$ Q^2_{F2} and CCC (Concordance Correlation Coefficient) to estimate the significance of the developed model. For a better understanding of the prediction quality, we also calculated mean absolute error (MAE) [128]. The approved threshold value for $Q^2_{(LOO)}$ and external validation parameters (Q^2_{F1}, Q^2_{F2}) is 0.5 [128].

3.3.6 Intelligent Consensus Prediction (ICP)

Different classes of descriptors were employed to develop a well-validated QSAR model, which represents various structural and molecular features. An individual QSAR model may either exaggerate some of the descriptors underrate a few descriptors or may completely disregard a few significant features [129]. Thus, consensus models should be generated utilizing individual models. In ICP, we evaluate the consensus model’s performance and correlate it with the individual MLR model-derived prediction quality (95%) based on the MAE criteria. Therefore, in our present study, we execute consensus modeling of selected five PLS-based QSTR models(M1-M5) using the “Intelligent Consensus Predictor (ICP) PLS version 1.2” tool (available at http://dtclab.webs.com/software-tools) to investigate the prediction quality of test set compounds which may be improved by an “intelligent” selection. Four distinct methods of consensus prediction were employed as outlined below:

Model 0 (CM0): This involves calculating the simple average of predictions obtained from each individual model.

Model 1 (CM1): Here, predictions from all eligible individual models are averaged using arithmetic mean.

Model 2 (CM2): Predictions from qualified individual models are averaged with weights assigned to each, creating a weighted average.

Model 3 (CM3): The selection of the best prediction for each compound is determined compound-wise from all eligible individual models.

Consensus predictions are precise enough and depend on numerous models rather than a single model [130]. Here prediction was performed by considering the Dixon Q-test, AD criteria, and Euclidean distance [130].

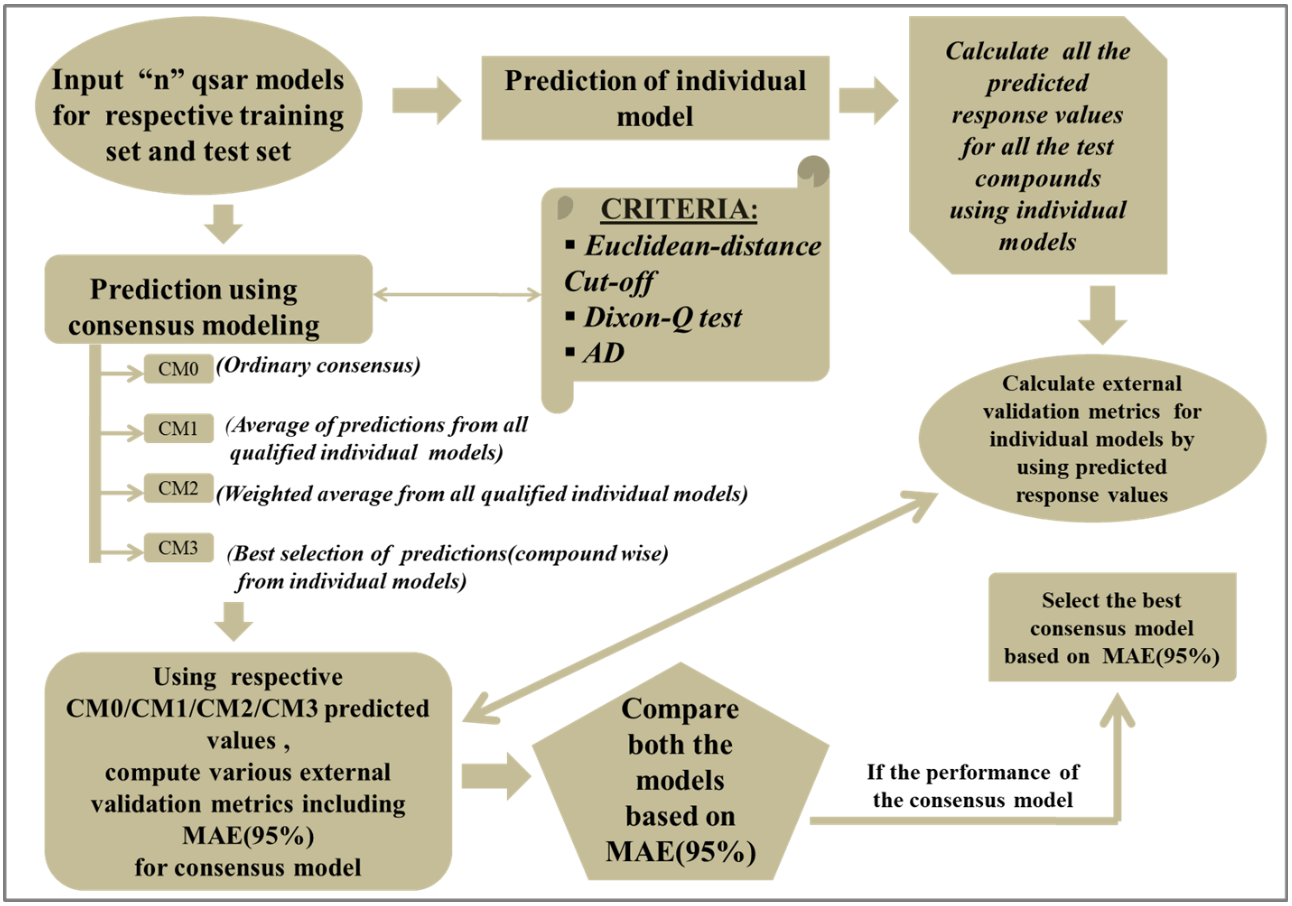

Input "n" qsar models for respective training set and test set
Prediction of individual model
Calculate all the predicted response values for all the test compounds using individual models
CRITERIA:
▪ Euclidean-distance Cut-off
▪ Dixon-Q test
▪ AD
Prediction using consensus modeling
CM0 (Ordinary consensus)
CM1 (Average of predictions from all qualified individual models)
CM2 (Weighted average from all qualified individual models)
CM3 (Best selection of predictions(compound wise) from individual models)
Calculate external validation metrics for individual models by using predicted response values
Select the best consensus model based on MAE(95%)
Using respective CM0/CM1/CM2/CM3 predicted values , compute various external validation metrics including MAE(95%) for consensus model
Compare both the models based on MAE(95%)
If the performance of the consensus model

Fig. 7. Workflow of Intelligent Consensus Prediction.

3.3.7 Applicability domain (AD) study of the developed models

The applicability domain (AD) of a QSAR is the physicochemical, structural, or biological space, enclosed by model descriptors and response on which basis the training set of the model has been developed, and which may be useful for the prediction of novel compounds. Whether the assumptions of the model are satisfied or not is determined by AD. Usually, in this regard, interpolation occurs rather than extrapolation [131]. AD of the developed models was assessed using the DmodX approach using SIMCA-P10.0 software. The basic idea of the DModx approach is based on the residuals of the Y matrix (response variable) and residual of the X matrix (predictor variables) which are of diagnostic value for the quality of the model. The residual standard deviation of the X-residuals of the corresponding row of the residual matrix E offers a summary for each observation as there are large numbers of X-residuals. The standard deviation (SD) is proportional to the distance between the data point and the model plane in X-space, usually called DModX (distance to model in X-space). Here, E is the (N×K) matrix of X-residual, N is the number of observations, and 'k' is the index of X-variables (K=1, 2, 3, 4…, K). If the DModX value is larger than around 2.5 times the overall SD of X-residuals, then the observation is considered to be outside the AD [111].

3.3.8 Y-randomization study

Y-randomization study has been performed to analyze and confirm whether the developed models are produced by any chance [132]. Here, Y-randomization plots are generated for final PLS-based models through the SIMCA-P software [116]. In randomization, the dependent variables were scrambled randomly while keeping the descriptor matrix constant, and by using the same set of variables from the original set, new models were built. The validation metrics obtained from the randomized model should be poorer than the original model otherwise that model should be considered to be developed by any chance [133].

3.3.9 Application of the constructed model on a prepared external dataset for data gap bridging

The capacity of a generated model is defined by its ability to determine the response value of unknown compounds. The developed models were employed to screen the pesticide property database (PPDB) [134] for potential toxicants in *California quail* by using the prediction reliability indicator (PRI) tool [135]. The PRI tool includes AD estimation and enables the categorization of the prediction quality of an external set [136].

3.4 Study 3

3.4.1. Collection, curation, and preparation of toxicity dataset

The toxicity data against wild birds with LD$_{50}$ endpoint was collected from the TOXRIC database [137] (available from https://toxric.bioinforai.tech/) and used for toxicity modeling. Before model generation, we prepare the dataset and curate the data to eliminate duplicate compounds, salts, and impurities. The endpoint values (LD$_{50}$) were converted to a negative logarithmic scale to modify the wide range of LD$_{50}$ into a narrow range and for easy interpretation. Marvin sketch 5.11.5 software (available from https://chemaxon.com/) was executed for drawing structures followed by their structure optimization by adding explicit hydrogen and transforming to aromatic form.

Table 9. Compounds smiles name with respective experimental pLD$_{50}$ values.

Sl. No.	Canonical SMILES	pLD$_{50}$
1*	S=P(N1CC1)(N1CC1)N1CC1	4.527
2	ClC1C(Cl)C(Cl)C(Cl)C(Cl)C1Cl	3.715
3	O=C(O)C=Cc1ccccc1	3.170
4	CCC1C2CC3C4N(C)c5ccccc5C45CC(C2C5O)N3C1O	3.263
5	O=C1c2ccccc2C(=O)c2c1ccc(O)c2O	2.880
6	CC=Cc1cc(OC)c(OC)cc1OC	2.443
7	CNC(C)C(O)c1ccccc1	2.468
8	Oc1cccc(O)c1O	3.225
9	O=[N+]([O-])c1cccc([N+](=O)[O-])c1	3.602

10	O=[N+]([O-])c1ccc(O)c([N+](=O)[O-])c1	4.151
11	ClC1=C(Cl)C2(Cl)C3C4C=CC(C4)C3C1(Cl)C2(Cl)Cl	4.704
12	Nc1ncn(C2OC(CO)C(O)C2O)c(=O)n1	3.387
13	C#CC(O)(C=CCl)CC	3.536
14	NC(Cc1ccc(O)c(O)c1)C(=O)O	3.294
15*	CNC(=O)ON=C(C)SC	4.210
16	COc1ccc2c(c1)N(CC(C)CN(C)C)c1ccccc1S2	3.516
17	Nc1ccc(O)cc1	3.288
18	N#Cc1ccc2c(c1)N(CCCN1CCC(O)CC1)c1ccccc1S2	3.562
19	COC(=O)C1C2CC3c4[nH]c5cc(OC)ccc5c4CCN3CC2CC(OC(=O)c2cc(OC)c(OC)c(OC)c2)C1OC	3.784
20	COP(=S)(OC)Oc1cc(Cl)c(Cl)cc1Cl	3.604
21*	Sc1ccccc1	3.661
22	Nc1ccncc1N	3.162
23	CCOC(=O)c1ccc(N)cc1	3.469
24	CS(=O)(=O)OCCCCOS(C)(=O)=O	3.641
25	Nc1ccncc1	4.599
26	CCCCOc1cc(C(=O)NCCN(CC)CC)c2ccccc2n1	3.912
27	O=[N+]([O-])c1cc(Cl)cc(-c2cc(Cl)cc([N+](=O)[O-])c2O)c1O	4.424
28	OCCN1CCN(CCCN2c3ccccc3Sc3ccc(Cl)cc32)CC1	4.101
29	CN(C)C(=S)SSC(=S)N(C)C	2.904
30	CC=Cc1ccc(OC)cc1	2.671
31	COC12C(COC(N)=O)C3=C(C(=O)C(C)=C(N)C3=O)N1CC1NC12	4.649
32	CN1CCCC1c1cccnc1	3.959
33	CNC(=O)Oc1cccc2ccccc12	3.555
34	O=C(O)O	2.792
35	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	3.842
36	COc1cc2c(c(OC)c1OC)-c1ccc(OC)c(=O)cc1C(NC(C)=O)CC2	4.101
37	CCN(CC)C(=O)C1C=C2c3cccc4[nH]cc(c34)CC2N(C)C1	5.254
38*	C1CN1c1nc(N2CC2)nc(N2CC2)n1	4.852
39*	C(#CCN1CCCC1)CN1CCCC1	3.284
40	C1CN1P1(N2CC2)=NP(N2CC2)(N2CC2)=NP(N2CC2)(N2CC2)=N1	3.197
41	CCOP(=S)(OCC)Oc1cc(C)c(SC)c(C)c1	4.882
42	COP(=S)(OC)Oc1ccc(S(=O)(=O)N(C)C)cc1	5.257
43	COP(=S)(OC)Oc1cc(C)c(SC)c(C)c1	4.465
44*	COP(=S)(OC)Oc1ccc(SC)c(C)c1	5.330
45	CCOP(=S)(OCC)Oc1ccc([N+](=O)[O-])cc1	5.340
46	CCOP(=S)(OCC)Oc1ccc2c(C)c(Cl)c(=O)oc2c1	5.309
47	CNC(=O)CSP(=S)(OC)OC	4.540
48*	COP(=O)(OC)OC=C(Cl)Cl	4.265
49	O=C(O)CF	4.517
50	CNC(=O)Oc1cccc(C(C)C)c1	4.781
51*	COS(C)(=O)=O	3.292
52	O=C1CC2OCC=C3CN4CCC56c7ccccc7N1C5C2C3CC46	4.320
53	Cc1ccc(S(N)(=O)=O)cc1	3.358
54	CCC(=O)c1ccc(N)cc1	3.049

55	CCCCO	1.472
56	CCc1ccc(C(c2ccc(CC)cc2)C(Cl)Cl)cc1	1.533
57	CCC1(C(C)C)C(=O)NC(=O)NC1=O	3.917
58	NNC(N)=S	4.000
59	CC(C)(O)C(C)(O)c1ccc(Cl)cc1	3.826
60*	C[n+]1c2cc(N)ccc2cc2ccc(N)cc21	3.601
61	COP(=S)(OC)SCn1nnc2ccccc2c1=O	4.572
62*	COc1cc(C=O)cc(OC)c1OC	2.667
63*	Cc1cc(OC(=O)N(C)C)n(-c2ccccc2)n1	3.798
64	Cc1c(N)cccc1Cl	2.776
65	NC(=O)c1ccccc1N	2.134
66	Nc1ccccc1[N+](=O)[O-]	2.265
67*	CCC(C)c1cc([N+](=O)[O-])cc([N+](=O)[O-])c1O	4.529
68*	O=[N+]([O-])c1cc(Cl)c(Cl)cc1Cl	3.355
69	COc1ccccc1N	2.465
70	COc1cc(C=CC(=O)O)cc(OC)c1OC	2.751
71*	CN(C)c1ccc(C(=O)c2ccc(N(C)C)cc2)cc1	3.428
72	Cc1c(N=C=O)cccc1N=C=O	3.240
73	O=C(O)c1ccc2ccccc2n1	3.238
74*	Cc1ccccc1F	3.041
75	Nc1ccccc1N	2.910
76	Cc1ccc(N)cc1C	4.335
77	Cc1cc(Cl)ccc1N	3.276
78	Nc1ccc(Cl)c(Cl)c1	2.834
79*	OCC(O)CCl	3.668
80	O=C(O)c1ccc(Cl)c([N+](=O)[O-])c1	3.429
81	CCOP(=S)(OCC)Oc1ccc(Cl)cc1Cl	4.352
82*	O=S(=O)(O)c1ccccc1	3.324
83	O=C(O)c1ccccn1	2.839
84	Nc1cccc(C(=O)O)c1	2.262
85	CC(=O)c1ccc(N)cc1	3.007
86	Nc1ccc([N+](=O)[O-])cc1	3.265
87	CN(C)c1ccc(N(C)C)cc1	3.840
88	C=Cc1ccncc1	3.021
89	COc1ccc(OC)c(N)c1	3.185
90	Nc1ccc(Cl)cc1	3.105
91	Cc1ccc(N)cc1	3.406
92*	CCOP(=O)(OCC)OP(=O)(OCC)OCC	5.348
93	CCOP(=O)(OCC)Oc1cc(C)[nH]n1	3.767
94	Cc1cccc(N)c1	2.646
95*	Nc1cccc(N)c1	2.260
96	Cc1ccncc1	2.343
97	Cc1cccnc1	1.969
98	Oc1cccnc1	2.103
99	Clc1ccccn1	2.055
100	OCC#CCO	3.059

101	CCN(CC)C(=O)C1CN2CCc3cc(OC)c(OC)cc3C2CC1OC(C)=O	3.606
102	CNC(=O)Oc1ccccc1OC(C)C	4.740
103	O=S1OCC2C(CO1)C1(Cl)C(Cl)=C(Cl)C2(Cl)C1(Cl)Cl	4.065
104	CCOP(=S)(OCC)Oc1ccc(S(C)=O)cc1	6.108
105*	CNC(=O)ON=CC(C)(C)SC	5.404
106	Nc1ccccc1C(=O)O	2.262
107*	COc1cc(OC)c([N+](=O)[O-])cc1Cl	3.337
108	Cc1ccc(N)cc1[N+](=O)[O-]	4.677
109	Cc1cc(OC(=O)N(C)C)n(C(C)C)n1	4.390
110*	CNC(=S)C(=S)NC	4.295
111*	COC(=O)C=C(CC(=O)OC)OP(=O)(OC)OC	5.472
112*	COP(=S)(OC)Oc1ccc([N+](=O)[O-])c(C)c1	4.401
113	O=C1C=CC(=O)c2ccccc21	3.075
114*	CN(C)c1ccc(N=NS(=O)(=O)O)cc1	4.105
115	COP(=O)(OC)OC(C)=CC(=O)N(C)C	5.375
116	N#C[Na]	4.088
117	CN(C)P(=O)(OP(=O)(N(C)C)N(C)C)N(C)C	4.415
118*	c1ncncn1	2.908
119	ClC1=C(Cl)C2(Cl)C3C(Cl)OC(Cl)C3C1(Cl)C2(Cl)Cl	5.614
120	CCOP(=S)(OCC)Oc1cnccn1	5.011
121*	COP(=S)(OC)Oc1ccc([N+](=O)[O-])cc1	4.721
122	CCOP(=S)(OCC)SCSCC	5.415
123	CCOP(=S)(OCC)SCCSCC	5.058
124*	CNP(=O)(OC)Oc1ccc(C(C)(C)C)cc1Cl	3.465
125	CCS(=O)CCSP(=O)(OC)OC	3.768
126	CN(C)CCCN(C)C1CCC2C3CC=C4CC(O)CCC4(C)C3CCC21C	2.964
127	CNC(=O)Oc1cc(C)c(N(C)C)c(C)c1	5.346
128*	CCC1CN2CCc3cc(OC)c(OC)cc3C2CC1CC1NCCc2cc(OC)c(OC)cc21	3.932
129	CCOP(=S)(CC)Oc1cc(Cl)c(Cl)cc1Cl	5.319
130*	CNC(=O)Oc1cc(C(C)C)cc(C(C)C)c1	4.371
131	CCOP(=S)(N=C1SCCS1)OCC	5.178
132	CCOP(=S)(OCC)Oc1cc(C)nc(C(C)C)n1	5.182
133	CC(=O)c1cccnc1	2.832
134	Nc1ccc(F)cc1	3.045
135	Nc1cccc(F)c1	3.297
136	Cc1ccc(N)cc1F	4.983
137*	Cc1ccc(N)c(F)c1	3.097
138	Nc1cccnc1	3.849
139	CCOP(=O)(OCC)OC(=CCl)c1ccc(Cl)cc1Cl	4.441
140	O=C1CC(c2ccccc2)Oc2ccccc21	3.475
141	COP(=S)(OC)Oc1ccc([N+](=O)[O-])c(Cl)c1	4.172
142*	Nc1ccccn1	3.474
143	O=C1C=Cc2ccccc2C1=O	3.324
144*	COc1cccc(N)c1	2.340
145*	Nc1ccc(I)cc1	3.340
146*	ClCCN(CCCl)CCCl	3.811

147	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	3.931
148	CCc1ccccc1N	2.208
149*	Cc1ccc(N=C=O)cc1N=C=O	3.240
150	OCc1ccncc1	2.412
151	OCc1ccccn1	2.162
152	CCc1cccc(N)c1	2.583
153	CCc1ccc(N)cc1	3.208
154*	Nc1cccc(O)c1	2.663
155	Cc1ccc(Cl)c(O)c1	2.404
156*	Nc1ccc(N)cc1	3.034
157	Clc1cccnc1	2.180
158*	Nc1ccc(Cl)c([N+](=O)[O-])c1	3.236
159	Cn1cc(NC(=O)c2cc(NC(=O)c3cc(NC=O)cn3C)cn2C)cc1C(=O)NCCC(=N)N	3.825
160	CCCC1CCCCN1	3.354
161*	NC(=O)CF	4.137
162	Cc1cc(OC(=O)N(C)C)nn1C(=O)N(C)C	4.426
163*	CNC(=O)Oc1cc(C)c(Cl)c(C)c1Cl	3.394
164	CNC(=O)Oc1cc(C)c(C)cc1Cl	5.079
165*	CCc1cc(OC(=O)NC)c(Cl)c(C)c1Cl	4.304
166*	CNC(=O)Oc1cc(C)c(Cl)c(C)c1	3.329
167	CCC(C)c1cccc(OC(=O)NC)c1	4.653
168	O[n+]1ccccc1	1.982
169	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	4.246
170	CCCN(CCC)C(=O)SCC	3.277
171	CCOP(=S)(OCC)SCSc1ccc(Cl)cc1	4.786
172*	CCN(CC)CCCl	3.509
173	N#Cc1ccc(N)cc1	3.697
174*	CC(=O)O[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	3.592
175	CC(C)OS(C)(=O)=O	2.682
176*	CCOP(=S)(CC)Sc1ccccc1	4.391
177*	CCOP(=O)(N=C1SCCS1)OCC	5.032
178	COP(=S)(OC)SCSc1ccc(Cl)cc1	4.242
179*	c1ccc(C2(N3CCCCC3)CCCCC2)cc1	4.638
180	Cc1cc[n+](O)cc1	2.042
181*	CC(=O)NN	3.244
182*	CNC(=O)Oc1cccc2sccc12	4.066
183*	On1ccccc1=S	3.104
184*	CC(=O)c1ccccn1	2.083
185	O=[N+]([O-])c1cc[n+]([O-])cc1	4.243
186	CNC(=O)Oc1cccc(C)c1	3.218
187	Nc1ccc(C(=O)c2ccccc2)cc1	2.545
188	COP(=S)(OC)Oc1ccc(S(=O)(=O)c2ccc(OP(=S)(OC)OC)cc2)cc1	4.074
189*	COC(=O)C1C2CC3c4[nH]c5cc(OC)ccc5c4CCN3CC2CC(OC)C1OC	3.632
190	CNC(=O)Oc1cccc2c1OC(C)(C)C2	5.721
191	CCCC(=O)c1ccc(N)cc1	3.587
192*	N#Cc1cc(I)c(O)c(I)c1	3.859

193*	N#Cc1cc(Br)c(O)c(Br)c1	3.743
194*	CCCCCCCC(=O)Oc1c(Br)cc(C#N)cc1Br	3.362
195*	CNP(=O)(NC)Oc1ccccc1	4.187
196	CCOP(=S)(OCC)OC(=CCl)c1cc(Cl)ccc1Cl	3.699
197	CSc1cccc(N)c1	2.268
198*	CC(=O)Nc1ccc(N=NN(C)C)cc1	3.566
199	CNC(=O)Oc1ccccc1	3.179
200	Cc1cnccc1N	4.653
201	CNC(=O)Oc1ccc(N(C)C)c(C)c1	3.619
202*	CNC(=O)Oc1cc(C)c(SC)c(C)c1	4.972
203	CCOP(=S)(Oc1ccc([N+](=O)[O-])cc1)c1ccccc1	5.134
204*	CNC(=O)Oc1cc(C)c(S(C)(=O)=O)c(C)c1	5.155
205	CCNP(=O)(OC)Oc1cc(Cl)c(Cl)cc1Cl	3.754
206	N#Cc1cccc(N)c1	2.322
207	CNC(=O)C(C)SCCSP(=O)(OC)OC	3.914
208	CNC(=O)Oc1ccc(Cl)c(C)c1	3.300
209	CC1OCC2C(CO1)C1(Cl)C(Cl)=C(Cl)C2(Cl)C1(Cl)Cl	3.888
210	CNC(=O)Oc1ccc(Cl)cc1	3.268
211	CNC(=O)Oc1cc(C)cc(C(C)C)c1	4.617
212	CNC(=O)Oc1ccccc1C(C)C	3.536
213*	CNC(=O)Oc1cc(C)c(S(C)=O)c(C)c1	3.759
214*	COP(=S)(OC)Oc1ccc(C#N)cc1	4.908
215	CNC(=O)Oc1cc(C)cc(C)c1	3.378
216	CCOP(=S)(OCC)ON1C(=O)c2cccc3cccc(c23)C1=O	4.312
217	COc1cc(Cl)c(OC)cc1Cl	1.617
218	CNC(=O)Oc1cc(C)c(C)c(C)c1	4.286
219	CCNP(=O)(OCC)Oc1cc(Cl)c(Cl)cc1Cl	4.141
220*	CCC(C)c1ccc(Cl)c(OC(=O)NC)c1	5.003
221	CCOP(=S)(OCC)Oc1nc(Cl)c(Cl)cc1Cl	4.845
222*	CCP(=S)(OC)Sc1ccc(C)cc1	4.643
223*	S=P(NC1CCCCC1)(N1CC1)N1CC1	4.389
224*	COP(=S)(OC)Oc1ccc(SC)cc1	5.723
225	COP(=S)(OC)Oc1ccc(SSc2ccc(OP(=S)(OC)OC)cc2)cc1	5.521
226	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)cc2)cc1	4.163
227	Nc1cc[n+]([O-])cc1	3.112
228	NC(=O)c1cccc(N)c1	2.134
229*	CNC(=O)Oc1ccc(SC)c(C)c1	3.722
230	CCCCNCC1COc2cccc(OCC)c2O1	3.423
231*	CN(C)P(=S)(N1CC1)N1CC1	4.157
232	CCCCCCCC(=O)Oc1c(I)cc(C#N)cc1I	2.696
233*	CNC(=O)Oc1ccc(C)c(C)c1C	3.662
234	COP(=O)(OC)Oc1cc(Cl)c(Cl)cc1Cl	4.229
235	CC(N)=NP(=S)(Oc1ccc(Cl)cc1)Oc1ccc(Cl)cc1	4.951
236*	CCOC(=O)NN	3.642
237	COc1cc(OC)c(OC)cc1C=O	2.667
238*	CC(N)Cc1ccc(Cl)c(Cl)c1	3.434

239	CCOP(=S)(OCC)Oc1cc(Cl)c(Br)cc1Cl	4.283
240	CC(=O)Nc1ccncc1	4.020
241	CC1=C(C(=O)Nc2ccccc2)SCCO1	3.746
242	CN(C)C1CC(c2ccccc2)c2ccccc21	3.500
243	COP(=S)(OC)Oc1nc(Cl)c(Cl)cc1Cl	4.394
244*	CC(C)OC(=O)C(O)(c1ccc(Cl)cc1)c1ccc(Cl)cc1	2.132
245*	COc1cc(Cl)c(OC)cc1N	3.273
246*	C=CCN(CC=C)c1c(C)cc(OC(=O)NC)cc1C	4.324
247	CCOP(=S)(OCC)SC1CCC2C=CCC1S2	3.636
248	CNC(=O)C=C(C)OP(=O)(OC)OC	5.445
249*	CNC(=O)Oc1ccccc1C1OCCO1	4.238
250	CC(=O)Nc1ccc(C)c(Cl)c1	5.150
251	CCNP(=S)(OC)Oc1ccc(C(C)(C)C)cc1Cl	2.729
252*	COC(=O)C=C(C)OP(=O)(OC)OC	5.204
253*	CC(C)=CC1C(C(=O)OCc2coc(Cc3ccccc3)c2)C1(C)C	3.654
254*	OCC(CO)(CCl)CCl	4.857
255*	CCN(CC)C(=O)C(Cl)=C(C)OP(=O)(OC)OC	5.223
256*	CCCSP(=O)(OCC)SCCC	4.760
257	Clc1ccc(C2NCCNCc3ccccc32)cc1	3.435
258	CNC(=O)Oc1ccc(SC)c(C(C)C)c1	5.123
259	O[N+]1=NC2CC1C1C2C2(Cl)C(Cl)=C(Cl)C1(Cl)C2(Cl)Cl	4.186
260	CCOP(=S)(OCC)ON=C(C#N)c1ccccc1	4.724
261	CC1CCC2(O)C3(C)CC4(O)OC2(C1O)C1(O)C3(O)C(OC(=O)c2ccc[nH]2)C(O)(C(C)C)C41C	5.443
262	OCC(O)C1OC2OC(C(Cl)(Cl)Cl)OC2C1O	4.536
263*	CCC(C)(C)c1ccc(Cl)c(OC(=O)NC)c1	4.453
264*	Nc1nc(-c2ccccc2)ns1	3.498
265	CCNC(=O)Oc1ccc([N+](=O)[O-])cc1	3.447
266*	CCCCNC(=O)n1c(NC(=O)OC)nc2ccccc21	3.462
267	COP(=O)(OC)OC(=CCl)c1cc(Cl)c(Cl)cc1Cl	3.563
268*	Cc1nc(N(C)C)nc(OC(=O)N(C)C)c1C	3.900
269	CCOP(=S)(OCC)Oc1cc(C)nc(N(CC)CC)n1	5.045
270	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)c(C)c2)cc1C	4.995
271	CNC(=O)ON=C(C)SCCC#N	4.678
272*	COP(=O)(NC(C)=O)SC	3.116
273	CCC(C(=O)N1CCCC1C)(c1ccccc1)c1ccccc1	3.612
274	CCP(N)(=S)Oc1ccc(SC)c(C)c1	4.912
275	Cc1ccc(N)cc1I	4.987
276	O=S(=O)(F)C1CCC(S(=O)(=O)F)C1	5.255
277*	COC(=O)c1cncn1C(C)c1ccccc1	3.857
278	CCOP(=O)(NC(C)CC)Oc1cc(Cl)c(Cl)cc1Cl	3.681
279	CCNP(=O)(O)Oc1cc(Cl)c(Cl)cc1Cl	4.735
280	CC(C)NP(=O)(S)Oc1cc(Cl)c(Cl)cc1Cl	4.679
281	CNP(=O)(S)OC(C)C	3.228
282*	O=P(O)(O)OC(=CCl)c1ccc(Cl)cc1Cl	4.976
283	CCOP(=O)(OCC)OC(=CSCC)c1ccc(Cl)cc1	4.545

284	CCOP(=O)(OCC)SC(C#N)=NOc1ccccc1	4.992
285*	CCNP(=O)(O)Oc1ccc(Cl)cc1Cl	4.556
286	FC(F)(F)N=C1SC(=Nc2ccccc2)N(c2ccccc2)C1=NC(F)(F)F	2.619
287	CCN(CC)Cc1cc(Cl)cc(Cl)c1OP(O)(=S)CC	4.928
288	CCCCSC(=Nc1cccnc1)SCc1ccc(C(C)(C)C)cc1	1.571
289	CCCCN(CCCC)SN(C)C(=O)Oc1cccc2c1OC(C)(C)C2	4.165
290	CCCCCCCCSC(=O)Oc1cc(Cl)nnc1-c1ccccc1	1.578
291	OC(c1ccc(Cl)cc1)(c1cncnc1)c1ccccc1Cl	3.219
292*	CN(c1c(Br)cc(Br)cc1Br)c1c([N+](=O)[O-])cc([N+](=O)[O-])cc1C(F)(F)F	5.099
293	COP(=O)(OC)ON1C(=O)c2cccc3cccc(c23)C1=O	5.126
294	CCCN(CCOc1c(Cl)cc(Cl)cc1Cl)C(=O)n1ccnc1	2.805
295	O=C1OCCC1N(C(=O)C1CC1)c1cccc(Cl)c1	2.145
296	CC(C)(C)C(O)C(Cc1ccc(Cl)cc1Cl)n1cncn1	1.540
297	O=C1C(N(CO)C(=O)NCO)N(CO)C(=O)N1CO	2.092
298	COC(=O)c1c(Cl)nn(C)c1S(=O)(=O)NC(=O)Nc1nc(OC)cc(OC)n1	2.286
299	Nc1ccc(S)cc1	3.472
300	OCC(Br)(Br)Br	2.951
301	CCCCCCC1C(=O)OC(C)C(NC(=O)c2cccc(NC=O)c2O)C(=O)OC(C)C1OC(=O)CC(C)C	5.040
302*	COC1=CC(=O)c2ccccc2C1=O	2.774
303	CNC(=O)Oc1ccc(C)c(C)c1	3.901
304	COP(=S)(OC)Oc1ccc([N+](=O)[O-])cc1Cl	4.297
305	COP(=O)(OC)Oc1ccc(SC)cc1	5.646
306	C#CCOc1ccccc1OC(=O)NC	3.659
307	c1ccc(-c2nc(-c3ccccn3)nc(-c3ccccn3)n2)nc1	4.744
308	CCOP(=S)(OCC)OP(=S)(OCC)OCC	3.508
309	C#CCOc1cccc(OC(=O)NC)c1	4.136
310*	NCc1ccccn1	2.284
311*	S=c1cc[nH]cc1	2.170
312	CCOP(=S)(OCC)Oc1nn2c(C)cc(C)nc2c1Br	5.215
313*	COP(=O)(OC)OC(C)=CC(=O)OC(C)c1ccccc1	3.747
314	Cc1ccc(N)cc1Cl	4.770
315	Cc1ccc(N)cc1Br	4.990
316*	CCCOC(=O)NCCCN(C)C	1.790
317	CCOP(=S)(OCC)Oc1nn2c(C)cc(C)nc2c1Cl	5.038
318*	CCOP(=O)(O)Sc1ccc(C)cc1	5.161
319	OCc1ccccc1	3.034
320*	Nc1ccccc1	2.219

***Test set compounds**

3.4.2. Computation of molecular descriptors and data pretreatment

The physicochemical and structural descriptors were estimated using Alvadesc software [104]. 3D-based descriptors were omitted and only considered a pool of 2D descriptors which involved atom-based E-state indices, constitutional indices, ring descriptors, 2D atom pairs, molecular properties, connectivity index, functional group numbers, atom-centered fragments, and ETA indices. The identical and non-redundant descriptors were excluded by performing data pre-treatment using the Data pretreatment GUI 1.2 tool (http://teqip.jdvu.ac.in/QSAR_Tools/ DTCLab) with a standard deviation less than 0.001 and a coefficient of correlation value greater than 0.95.

3.4.3. Dataset splitting

The dataset was divided into the training and test sets with the ratio 70:30 by "Dataset Division GUI" version 1.2 (available from http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab). The dataset splitting was carried out by implementing various methods of dataset division such as Kennard stone, activity-property, Euclidean distance based, and modified *k*-Medoid clustering techniques but the best statistics were obtained by using the Euclidean distance-based method.

3.4.4. Feature selection and QSAR model generation

Feature selection is a significant process whose objective is to reduce the redundant, noisy, and irrelevant descriptors towards the model generation without loss of important feature information [138]. In the present study, suitable features are selected from the initial pool of descriptors by genetic algorithm using a java-based tool Genetic algorithm_v4.1 (available from https://teqip.jdvu.ac.in/QSAR_Tools/) [106]. The optimal combination of features was selected by employing the best subset selection method [106] available from (http://teqip.jdvu.ac.in/QSAR_Tools/). Afterward, the selected descriptors were subjected to develop an initial partial least square model (PLS)[22] by the PLS_SingleY_version_1.0 tool (available from https://teqip.jdvu.ac.in/QSAR_Tools/) to diminish the intercorrelated descriptors with the optimum number of latent variables.

3.4.5. Read-across and RASAR descriptor calculation

Read-across is a data gap-filling method based on structural similarity between a target and source compounds in which the toxicity/activity/property of novel compounds are estimated from their structural analouges. The read-across prediction was performed by utilizing 3 similarity-based techniques such as Gaussian kernel (GK) based similarity, Euclidean distance (ED) based similarity and Laplacian Kernel (LK) based similarity methods with hyperparameter optimization which includes Sigma(σ): Gaussian kernel similarity assessment, gamma(γ): Laplacian kernel similarity assessment and number of closed source compound numbers essential for quality prediction [139]. In this current study, we have used various sigma values (0.25-2 with an interval of 0.25), various gamma values (0.25-2 with an interval of 0.25), and the number of close source compounds within the range of 2-10 for hyperparameters optimization. Before optimization, the initial training set is split randomly into sub-train and sub-test sets with the proportion of 3:1. The sub-train and sub-test were deployed to a java-based tool Read-Across-v4.1 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) with different suggested values of σ and γ. On the other hand, other parameters such as the number of close source compounds, the distance threshold, and the threshold of similarity remained constant. The optimized setup has been chosen based on Q^2_{F1}, $Q^2_{F2,}$ MAE, and RMSE. Finally, the optimized set-up was applied to the original training and test sets for final prediction.

The RASAR technique integrates the concept of Read-Across and QSAR [140]. Here, the RASAR descriptor calculation was associated with Read-across prediction followed by hyperparameter optimization. This technique computes novel descriptors like RA function, SD Activity, SE, CVact, CVsim, MaxPos, MaxNeg, Abs Maxpos-MaxNeg, AvgSim, SD Similarity, gm (Banerjee-Roy Coefficient), gm*Avg.Sim, gm*SD Similarity, Pos.Avg.Sim, and Neg.Avg.Sim. by using similarity and error-based measures. The above 15 q-RASAR descriptors were computed by the RASAR-Desc-Calc-v2.0 tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home).

3.4.6. q-RASAR-based model development

After calculating RASAR descriptors, these descriptors were combined with formerly selected structural and physicochemical descriptors for the final q-RASAR model generation. A combination of ten descriptors was identified by performing the best subset selection. The q-RASAR model was generated using a pool of combined descriptors (RASAR, structural, and physiochemical descriptors) by employing PLS regression. This model has been validated vigorously using both internally and externally.

3.4.7. Applicability domain study of the generated model

OECD principle 3 suggests the estimation of applicability domain study of the developed model. The applicability domain is the theoretical chemical space encompassing the model descriptors and response. It is practically impossible to predict the toxicity of every compound by one statistical-based model; there should be some structural similarity between query compounds and training compounds for a reliable prediction [141]. Therefore, estimating the applicability domain of any developed model is required. Here, we perform AD analysis by using the DmodX (distance to model X) approach for the developed PLS-based q-RASTR model with SIMCA-P software [142].

3.4.8. Y-randomization test of the generated model

Y-randomization test is employed to ensure whether the obtained model is generated by any chance or not. In this test, the descriptor variables (X variables) are kept constant and the vector Y is shuffled randomly multiple times and generates a new model using the same sets of descriptors. The model is considered to be robust if the estimates of statistical parameters of the randomized model are poorer than the originally developed model. The R^2y_{rand} intercept and Q^2y_{rand} intercept are not more than 0.3 and 0.05 respectively. In this study, the chance correlation between the descriptor and response variable of the developed model was checked by SIMCA-P software [142].

3.4.9. PPDB database assessment by deploying the developed q-RASTR model and reliability study

For the preparation of an external dataset, PPDB (Pesticides Property Database) consisting of 1902 pesticides was downloaded. The database was curated to eliminate mixtures, salts, and impurities by KNIME workflow. Ultimately, 1694 compounds are remained after curation and used as the external database for screening. Descriptors for these compounds were calculated as done for the modelled dataset by Alvadesc software version 2.02 (https://www.alvascience.com/alvadesc/). This Prepared PPDB database was screened by deploying the generated model using the Prediction Reliability Indicator (PRI) tool [135], which provides assessment and categorization of prediction quality in terms of AD as well as in terms of Good, Moderate, and Bad.

CHAPTER - 4

Results and Discussion

4. Results and discussion

4.1 Study 1

In this study, we have developed PLS models utilizing the toxicity of pesticides ($LogLC_{50}$) on two different avians (MD and RNP) employing a reduced pool of chemical descriptors. The created model's quality is measured by using different internal (R^2, Q^2_{LOO},) and external (Q^2_{F1}, Q^2_{F2},) statistical parameters. The results obtained from PLS models indicated the model's robustness, reliability, and predictivity. All the metrics obtained from QSTR models are depicted in **Table 10**. The read-across algorithm was employed to improve the model's external predictivity External predictivity was improved for the dataset RNP but for MD the external predictivity is slightly diminished in read-across prediction and results are provided in **Table 11**. The obtained results from the Y-randomization test were found to be $R^2 =$ -0.008, $Q^2 =$ -0.0377 (for MD) and $R^2 =$ 0.028, $Q^2 =$ -0.213 (for RNP) which demonstrated that the models were not formed by any chance. AD study depicted that compound **468** in MD, and compound **88** in RNP from the test set are outside the AD as depicted in **Fig. 15** and **Fig. 16** respectively. The tentative reasons or characteristics that designate certain compounds as outliers in each model (above the D-critical line) are due to structural dissimilarity for example, In the case of the MD model C-012, [O-P] fragment at topological distance 7, [C-P] fragment at topological distance 5 and [C-Cl] fragment at topological distance 4 are absent and lastly, for RNP model nRCONHR, [C-P] fragment at topological distance 4, [P-Cl] fragment at topological distance 5, and [O-S] fragment at topological distance 3 is absent. We have developed new QSTR models without the identified outliers and checked the statistical metrics. A visual representation of the correlation between observed and predicted toxicity values has been depicted in the scatter plot (provided in **Fig 8**). Additionally, we used two different ML algorithms namely SVM, and RF to evaluate their effectiveness in model construction and prediction, and the obtained statistical results are demonstrated in **Table 12**.

Table 10. Statistical results of developed PLS models.

Avian Species	**Training set**				**Test set**			
	N_{train}/N_{test}	**LVs**	R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	$MAE_{(test)}$	**Quality(test)**
RNP	82/30	2	0.63	0.53	0.60	0.60	0.34	Moderate
MD	377/162	1	0.60	0.58	0.75	0.63	0.06	Good

Table 11. Read-across-based predictions for four species.

Ring-necked pheasant		
Optimized settings	METRICS	Ylk (Test)
σ =0.5 γ =0.5 No. of similar compounds =10	Q^2_{F1}	0.714
	Q^2_{F2}	0.714
	$RMSE_P$	0.392
	MAE	0.290
Mallard duck		
Optimized settings	METRICS	Yeuc (Test)
σ =0.75 γ =0.75 No. of similar compounds =10	Q^2_{F1}	0.686
	Q^2_{F2}	0.540
	$RMSE_P$	0.114
	MAE	0.081

Table 12. ML model's statistical quality for MD and RNP.

Validation Metrics	ML model's statistical quality			
Model	SVM(MD)	RF(MD)	SVM(RNP)	RF(RNP)
R^2_{LOO}	0.666	0.667	0.641	0.577
Q^2_{LOO}	0.663	0.666	0.505	0.566
$RMSE_C$	0.098	0.098	0.438	0.476
MAE	0.061	0.065	0.340	0.349
Optimum hyperparameter	(*v*-SVM) Regression Cost - 0.50 Complexity bound - 0.55 Kernel-RBF Numerical tolerance-0.0011 Iteration limit-150	No of trees-48 Limit depth of individual trees-5 No of attributes-7	Cost-1.90 Regression loss epsilon-0.30 Kernel-RBF Numerical tolerance-0.0011 Iteration limit-150	No of trees-27 Limit depth of individual trees-5 No of attributes-4

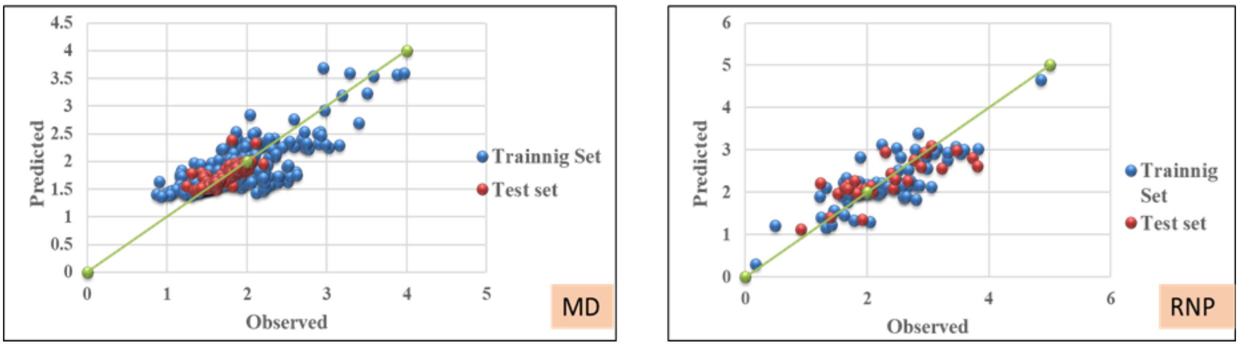


Fig. 8. Scatter plots of developed models.

Several classification-based metrics have been computed with the PLS-based QSTR-read across models for all (MD and RNP) the avian species and reported in the following **Table 13**. Good sensitivity, specificity, and accuracy values indicate the good classification ability of the model. The computed values of the Matthews correlation coefficient [143] indicate an acceptable prediction and an agreement between observed and predicted classification for all the developed models against avian species.

Table 13. Statistical results of the classification-based QSTR models.

Sl no.	LDA-QSTR models	AUC-ROC	Sensitivity	Accuracy	Precision	F-measure	MCC
1	MD (train)	0.88	75.00	83.59	82.60	78.62	0.65
	MD (test)	0.86	75.71	85.71	89.83	82.17	0.71
2	RNP (train)	0.83	63.88	79.74	88.46	74.19	0.60
	RNP (test)	0.87	76.92	84.84	83.33	80.00	0.67

4.1.1. Regression coefficient plot

The descriptor’s positive and negative contribution towards toxicity is provided via a regression coefficient plot. In this investigation, for MD, the descriptors MW, C-012, B07[O-P], Br-094, B05[C-P], and F04[C-Cl] contributed positively towards toxicity on the other hand, in case of RNP, the descriptors nRCONHR and B04[C-P] contributed positively whereas the descriptors X2A, nN(CO)2, B05[P-Cl], and F03[O-S] contributed negatively towards the toxicity. All the relevant plots have been provided in **Fig. 9.** and **Fig. 10**.

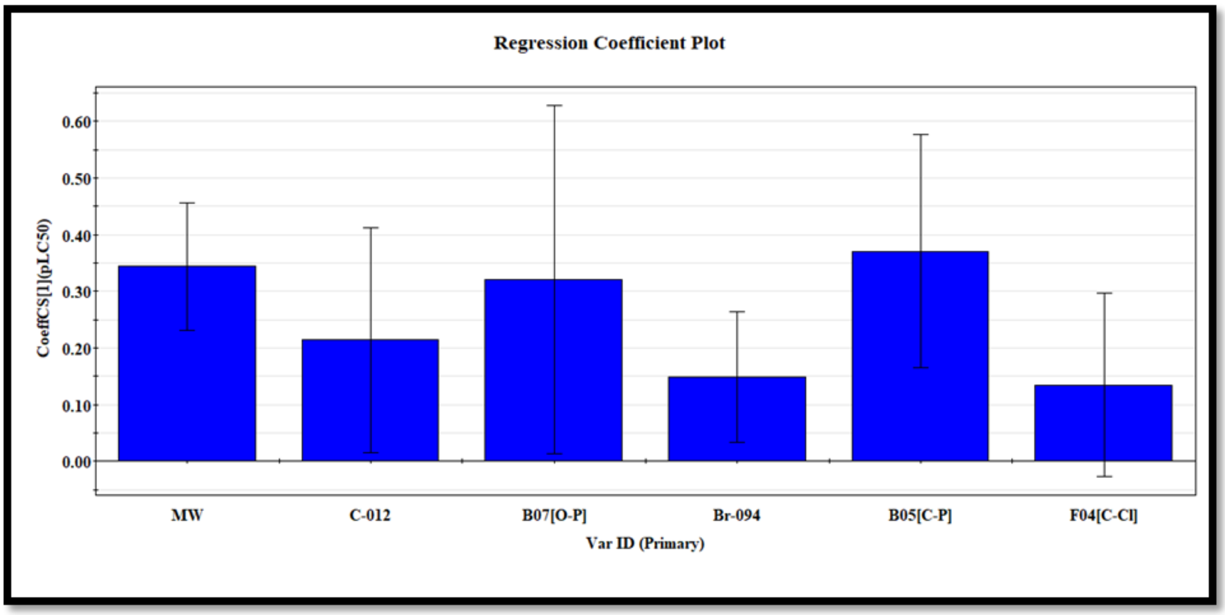


Fig. 9. Regression coefficient plot for MD.

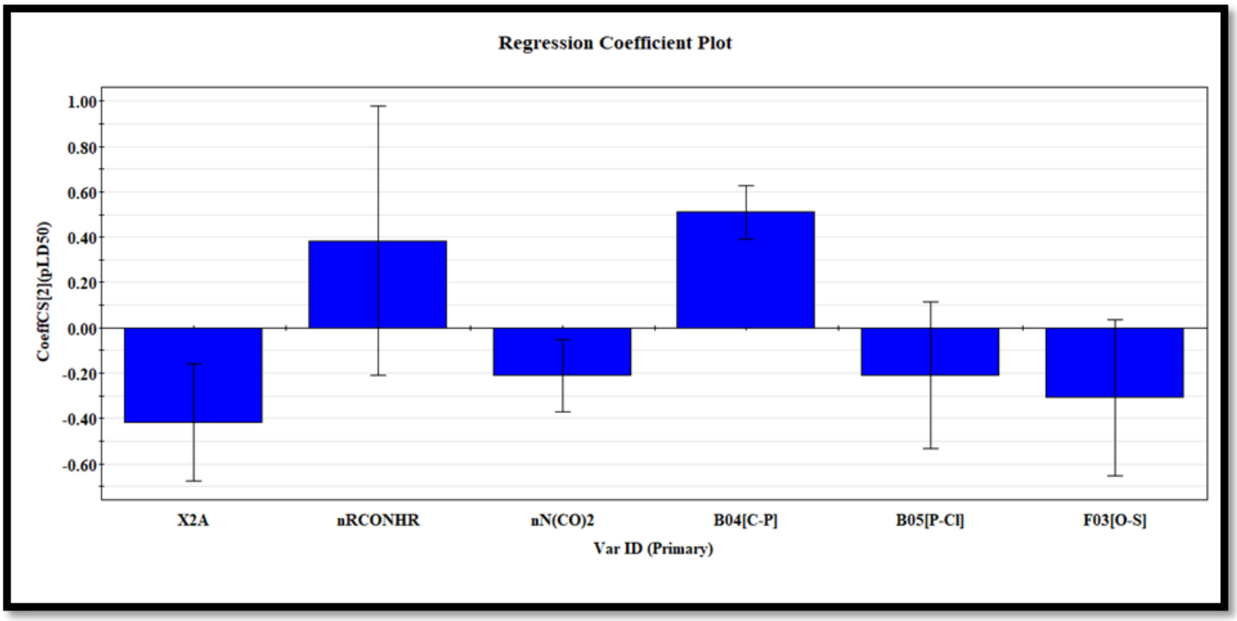


Fig. 10. Regression coefficient plot for RNP.

4.1.2. Variable importance plot (VIP)

The relative importance of model descriptors is illustrated with VIP [144]. Descriptors having the highest and lowest impact on avian species can be recognized from these plots. The significance of the variable is higher whose VIP score is greater than 1. In VIP plot, the descriptors are presented with respect to their significance (higher contribution to lower contribution) and their importance which is in the following order: B05[C-P], MW, B07[O-P], C-012, Br-094, F04[C-Cl)] (in case of MD) and B04[C-P], X2A, nRCONHR, F03[O-S], B05[P-Cl], Nn(CO)2 (in case of RNP) as depicted in **Fig. 11** and **Fig. 12**.

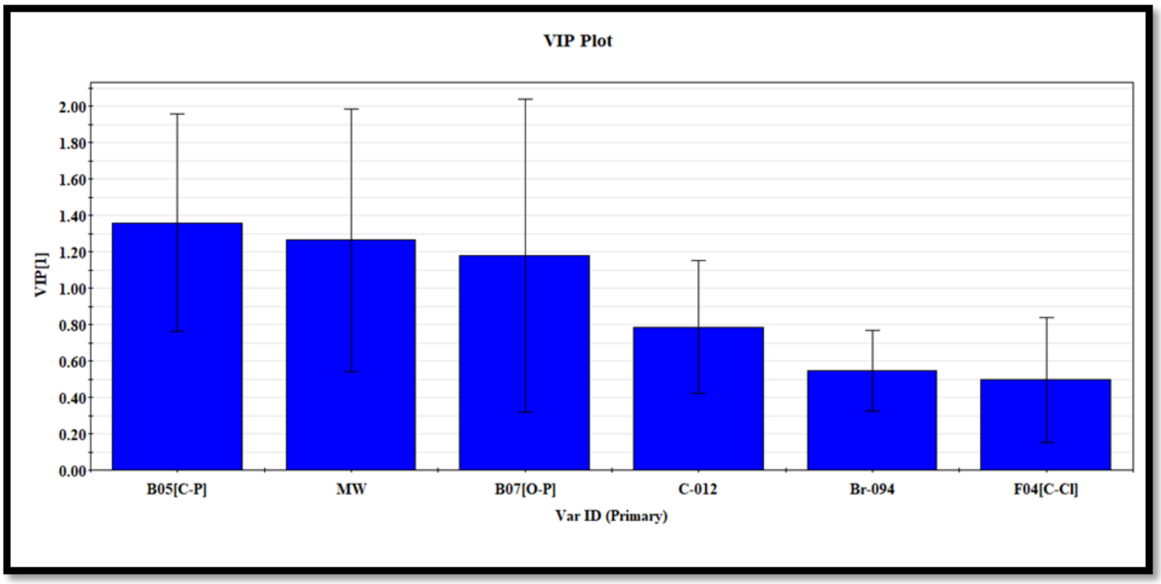


Fig. 11. Variable importance plot of MD.

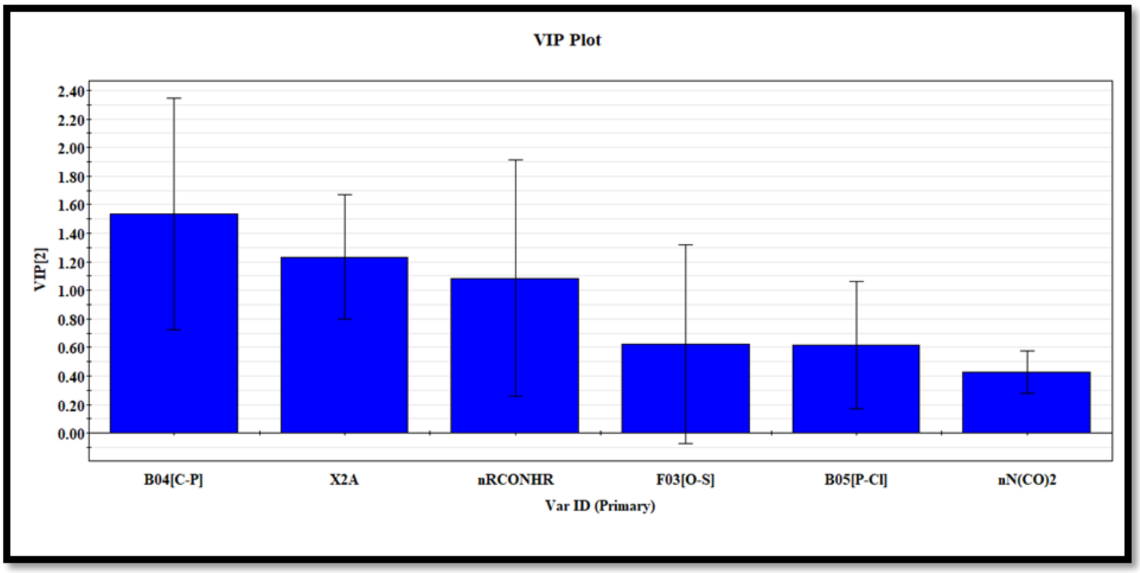


Fig. 12. Variable importance plot of RNP.

4.1.3. Loading plot

The loading plot shows how the independent variables (descriptors) are related to the response variable. The first two components were used to create the loading plot. A descriptor is assumed to have a stronger effect on response value if it is located far from the origin of the plot. On the basis of the loading plot as shown in **Fig.13** and **Fig. 14**, it is interpreted that the X-variables B05[C-P], and B04[C-P] are the most influential descriptors in the case of MD, and RNP respectively.

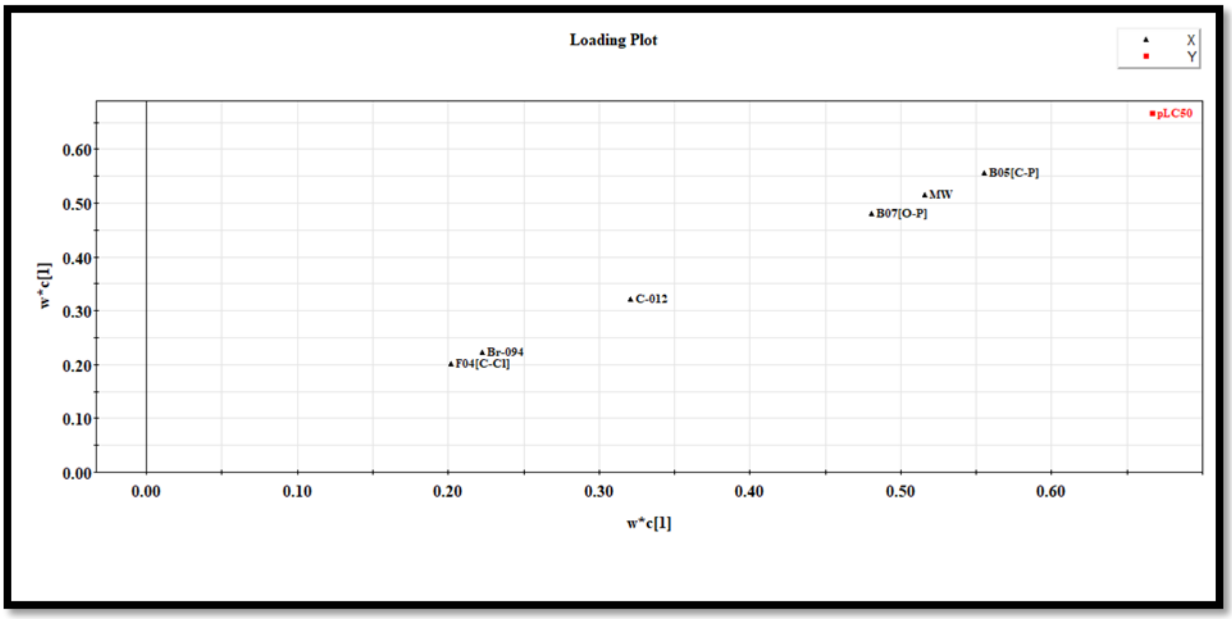


Fig. 13. Loading plot of MD.

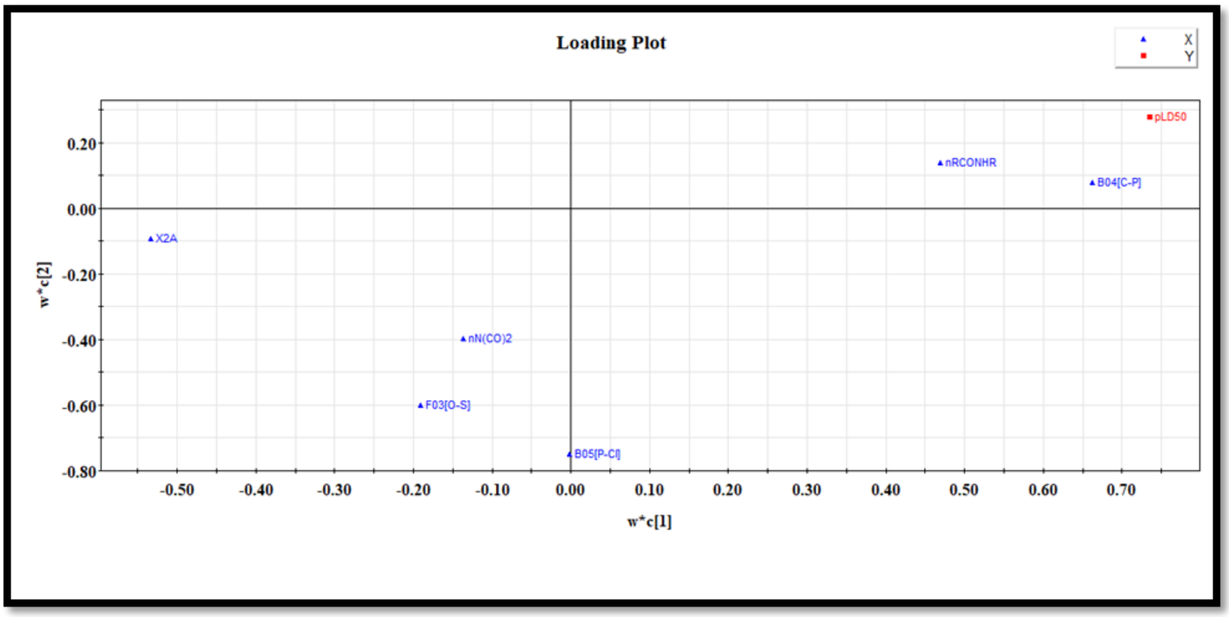


Fig. 14. Loading plot of RNP.

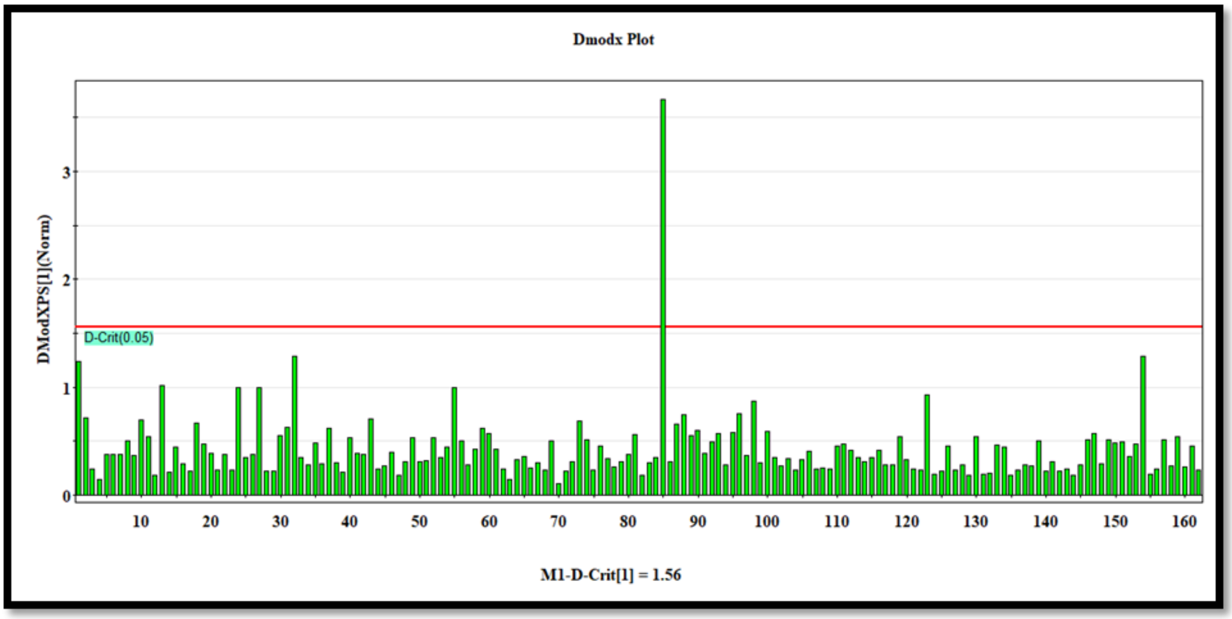


Fig. 15. DmodX plot for MD.

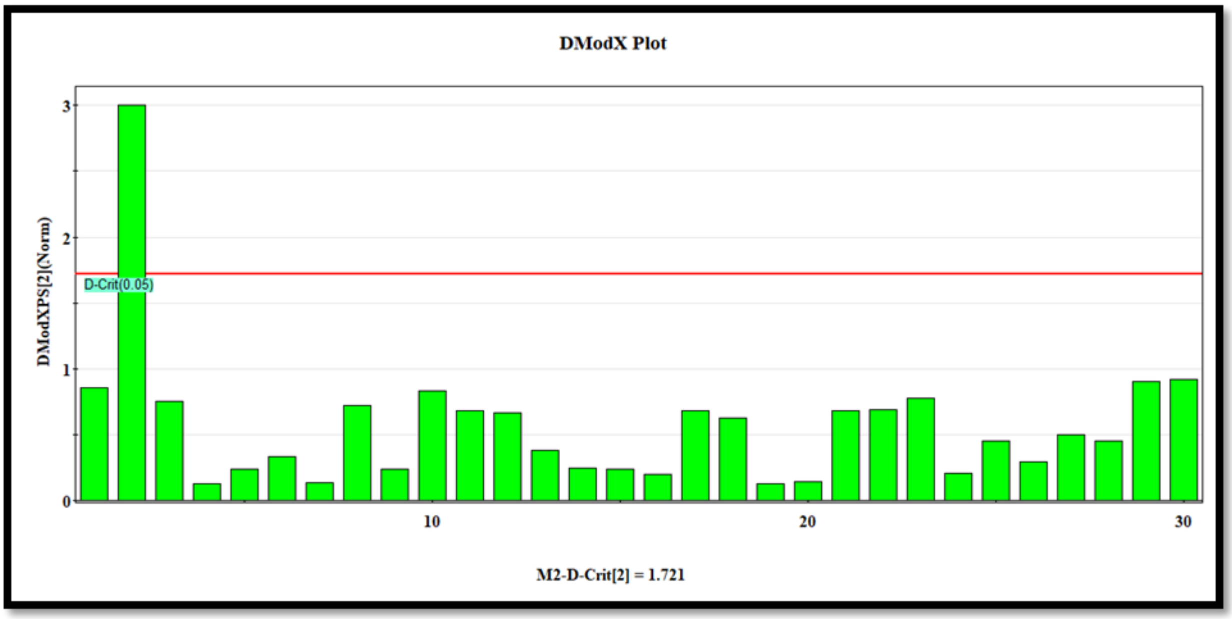


Fig. 16. DmodX plot for RNP.

4.1.4. Mechanistic interpretation of PLS models

Table. 14 and **Figs. 17-18** provide a detailed account of the model descriptors followed by mechanistic interpretations important to identify major structural and physicochemical features.

Table 14. Mechanistic interpretation of descriptors employed in models.

Sl. no	Descriptor	Type	Function	Contribution
MD oral pLC$_{50}$				
1	MW	Constitutional descriptor	Molecular weight	+ve

	Mechanistic interpretation This descriptor is directly related to molecular bulkiness and lipophilicity [145-146]. Usually, lipophilic compounds easily cross the lipophilic membrane of the reference species which ultimately leads to enhancement in toxicity as demonstrated in compound **546** and oppositely occurs in compound **503 (**depicted in **Fig. 17)**.			
2	C-012	Atom-centered fragments	CR2X2 (X is a hetero atom (O, N, S, P, Se, or halogens) and R is a carbon-linked group)	+ve
	Mechanistic interpretation This descriptor enhances the molecular size as well as the electronegativity of the compound due to the presence of heteroatom, which ultimately leads to enhancement in toxicity of diverse pesticides against avian species by incorporating oxidative stress [147] as demonstrated in compound **445**, and vice-versa occurs in compound **144** (given in **Fig. 17**) .			
3	B07[O-P]	2D Atom Pair	Presence of O–P at topological distance 7	+ve
	Mechanistic interpretation Oxygen and phosphorus are highly electronegative atoms and their presence makes the compound more toxic (due to increment in oxidative stress in reference species) [148]. The presence of a long carbon chain (lipophilicity) also contributes to toxicity. This phenomenon is demonstrated in compound **3** and vice versa occurs in the case of compound **145** (illustrated in **Fig. 17**) .			
4	Br-094	Atom-centered fragments	Br attached to C1(sp2)	+ve
	Mechanistic interpretation The Br-094 descriptor refers to the presence of the halogen group (bromine). Thus, the presence of more electronegative/halogen atoms (bromine) makes the compound more toxic as demonstrated in compound **28**. Conversely, the absence			

	of this atom/fragment tends to decrease the toxicity as shown in compound **408** (depicted in **Fig. 17**).			
5	B05[C-P]	2D Atom pair	C–P situated at a topological distance of 5	+ve
	Mechanistic interpretation The presence of the phosphate group enhances the toxicity of the compound [149]. This is evidenced in compound **4**. In opposition, the absence of this fragment tends to decrease the toxicity as shown in compound **530**.			
6	F04[C-Cl]	2D Atom pair	C – Cl situated at topological distance 4	+ve
	Mechanistic interpretation This descriptor refers to the existence of a large electronegative atom such as chlorine, which has a high atomic refractivity and electronegativity [150]. Thus, the presence of more number of this fragment results in high toxicity toward avian species as shown in compound **24** and vice versa occurs in compound **562** (represented in **Fig. 17**).			
RNP oral pLC$_{50}$				
1	X2A	Connectivity indices descriptor	Average connectivity index of order 2	-ve
	Mechanistic interpretation The negative regression coefficient of this descriptor indicates that a higher numerical value of this descriptor leads to a decrease in toxicity as shown in compound **13** and vice versa in the case of compound **51** (given in **Fig. 18**). X2A is inversely correlated with hydrophobic interaction as well as toxicity.			
2	nRCONHR	Functional group count	Presence of secondary aliphatic amides	+ve
	Mechanistic interpretation Aliphatic amides are considered to be toxic as well as reactive [151]. The positive regression coefficient of this descriptor indicates that the presence of this			

	fragment may increase the toxicity as demonstrated in compound **90** and toxicity value may be decreased if the compounds have no such fragment as represented in compound **104** (depicted in **Fig. 18**).			
3	nN(CO)2	Functional group count	Number of imides (-thio)	-ve
	Mechanistic interpretation Generally, this feature helps to facilitate hydrolysis of the compounds which facilitates quick excretion from the body of the reference organism resulting in a reduction of their toxic effects [152] as demonstrated in compound **58** and the absence of this fragment tends to increase the toxicity as shown in compound **101** (illustrated in **Fig. 18)**.			
4	B04[C-P]	2D Atom pair	C – P situated at topological distance 4	+ve
	Mechanistic interpretation The presence of an electronegative atom (like phosphorous) enhances the toxicity of the diverse pesticides by incorporating oxidative stress in avian species [62] as evidenced by compound **3**. On the other hand, the absence of this fragment leads to a decrease the toxicity as shown in compound **10** (represented in **Fig. 18**).			
5	B05[P-Cl]	2D Atom pair	Presence of P-Cl at topological distance 5	-ve
	Mechanistic interpretation The negative regression coefficient of this descriptor indicates that the presence of more number of this fragment reduces the toxicity as demonstrated in compound **105** and oppositely occurs in the case of compound **62** (shown in **Fig. 18)**.			
6	F03[O-S]	2D Atom pair	Frequency of oxygen and sulfur which are situated at topological distance 3.	-ve

	Mechanistic interpretation This descriptor is directly related to the polarity (presence of polar bond) [62] of the compound, as a result, the hydrophilicity of the compound increases, and thus toxicity will decrease which is evidenced by compound **85** and vice versa in the case of compound **9** (represented in **Fig. 18)**.

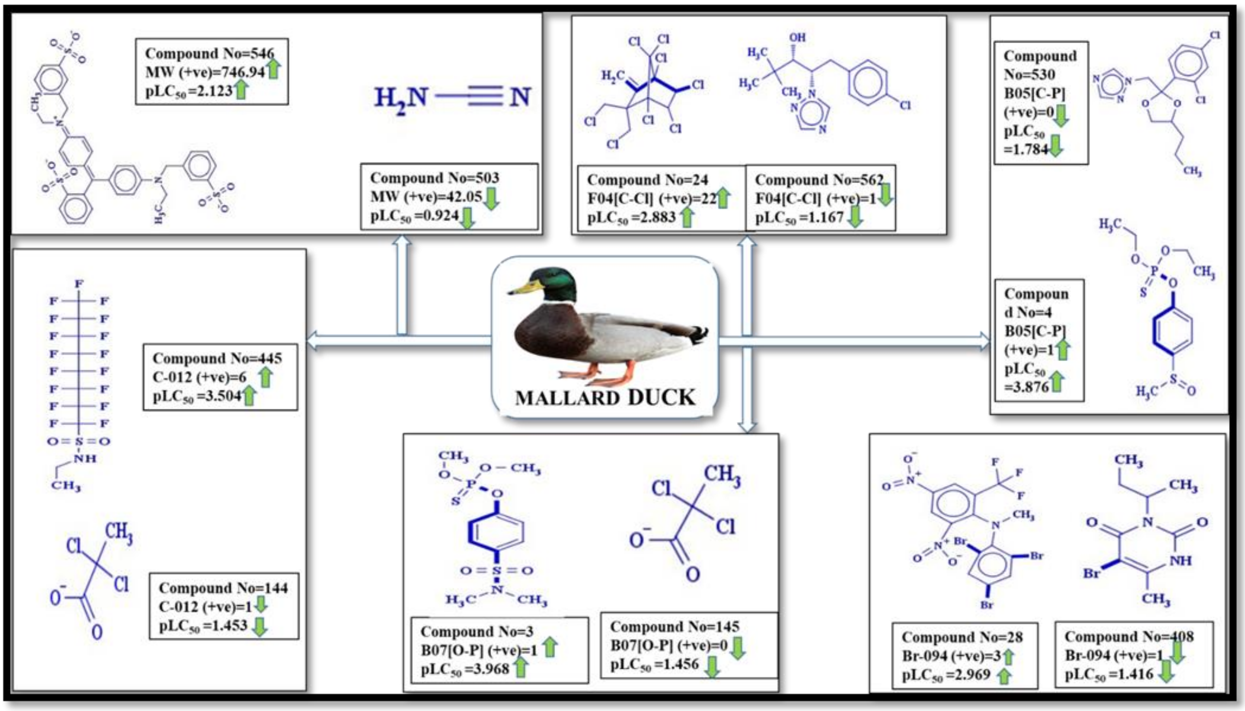

Compound No=546 MW (+ve)=746.94 pLC$_{50}$=2.123
H$_2$N—≡N
Compound No=503 MW (+ve)=42.05 pLC$_{50}$=0.924
Compound No=24 F04[C-Cl] (+ve)=22 pLC$_{50}$=2.883
Compound No=562 F04[C-Cl] (+ve)=1 pLC$_{50}$=1.167
Compound No=530 B05[C-P] (+ve)=0 pLC$_{50}$ =1.784
Compound No=4 B05[C-P] (+ve)=1 pLC$_{50}$ =3.876
MALLARD DUCK
Compound No=445 C-012 (+ve)=6 pLC$_{50}$ =3.504
Compound No=144 C-012 (+ve)=1 pLC$_{50}$ =1.453
Compound No=3 B07[O-P] (+ve)=1 pLC$_{50}$ =3.968
Compound No=145 B07[O-P] (+ve)=0 pLC$_{50}$ =1.456
Compound No=28 Br-094 (+ve)=3 pLC$_{50}$ =2.969
Compound No=408 Br-094 (+ve)=1 pLC$_{50}$ =1.416

Fig. 17 Positive and negative contribution of model descriptors towards MD.

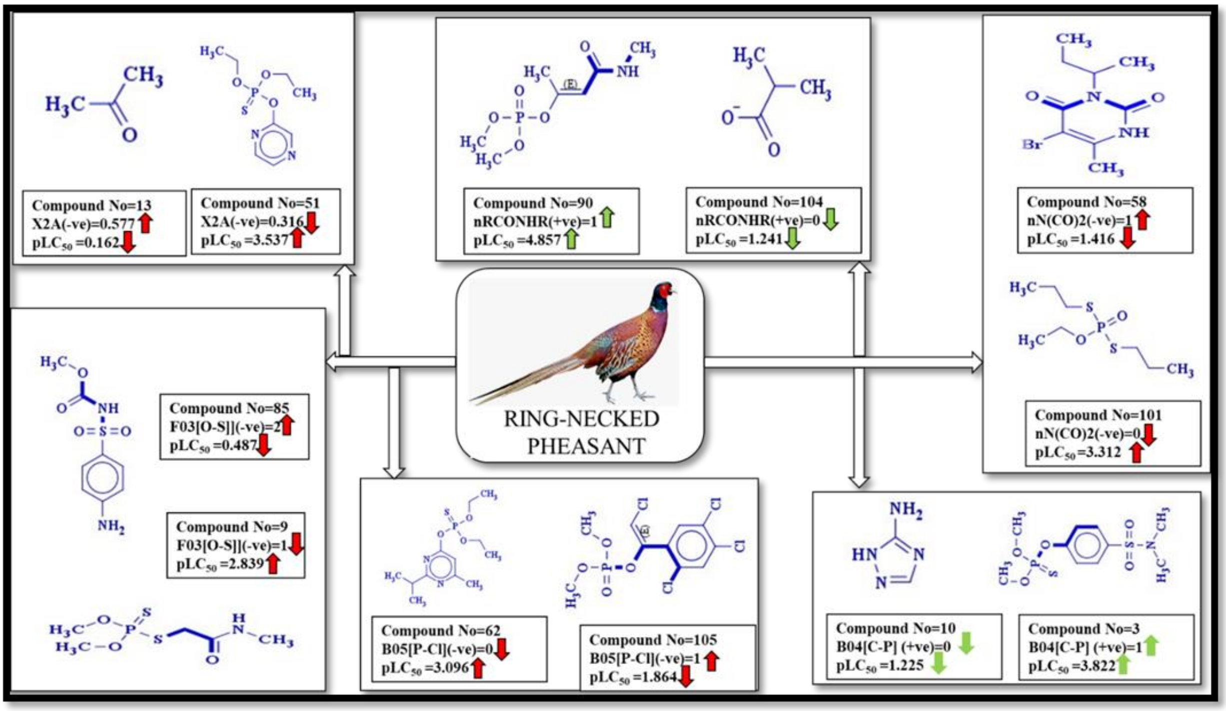


Fig. 18. Positive and negative contribution of model descriptors toward RNP.

4.1.5 PPDB DataBase screening

The pesticide Properties DataBase was screened through the developed models with the help of the software "PRI Tool_PLSversion" (available from http://teqip.jdvu.ac.in/QSAR Tools/) using the developed PLS models. The categorization threshold (mean value of the training set compound) for avian toxicity against MD; RNP ≥ 1.845; 2.191 was applied for prioritization purposes. From the prediction, it was seen that maximum compounds are within the domain of applicability and show prediction quality as "good". The screened chemicals from the Pesticide Properties DataBase with their respective predicted toxicity against MD, and RNP. The compounds were ranked in decreasing order of predicted toxicity for each avian species. The top 20 and least 20 toxic pesticides for all four avian species from the PPDB database are provided in **Table 15.** Further validation of the predicted toxicity of the selected pesticides revealed that apart from fluoroacetamide and sodium-monofluoroacetate all the predicted toxicity corroborated with the previous experimental findings, indicating the practical applicability of the developed models as shown in **Table 15**.

Table 15. Top 20 highly & least toxic pesticides screened from Pesticide Properties Database (PPDB).

Sl. No.	Names of pesticides	Safety and hazards
Top 20 highly toxic pesticides screened from Pesticide Properties Database (PPDB).		
1	Imicyafos	Acute toxic, Irritant.
2	Pirimiphos-ethyl	Acute toxic, Environmental Hazard.
3	Quinothion	Acute toxic
4	Pirimiphos-methyl	Irritant, Health hazard, Environmental hazard
5	Etrimfos	Irritant, Environmental Hazard
6	Buminafos	Acute toxic
7	Diazinon	Irritant, Environmental hazard
8	Quintiofos	Acute toxic
9	Phoxim	Irritant, Health hazard, and Environmental hazard
10	Inezin	Acute toxic
11	Dufulin	Oxidative stress inducer
12	Chlorphoxim	Acute toxic
13	Pyridaphenthion	Irritant
14	Triazophos	Acute toxic, Environmental hazard
15	Isoxathion	Acute toxic, Environmental hazard
16	Naftalofos	Acute toxic
17	Quinalphos	Acute toxic, Environmental hazard
18	Butamifos	Irritant, Environmental hazard
19	Sulprofos	Acute toxic, Environmental hazard
20	Edifenphos	Acute toxic, Environmental hazard
Top 20 least toxic pesticides screened from the Pesticide Properties Database (PPDB)		
1	Ferbam	non-toxic
2	Hexylene glycol	less toxic
3	Bisthiosemi	moderate toxic

4	Choline chloride	less toxic
5	Glutaraldehyde	less toxic
6	Fumaric acid	less toxic
7	Lime sulphur	less toxic
8	Methyl isobutyl ketone	less toxic
9	Sodium tetrathiocarbonate	moderate toxic
10	1,2-dichloropropane	less toxic
11	Metam	less toxic
12	Methylene bisthiocyanate	less toxic
13	Bentonite	Nontoxic
14	Butanethiol	moderate toxic
15	Sodium monochloroacetate	moderate toxic
16	Fluoroacetamide	high toxic
17	Sodium monofluoroacetate	high toxic
18	Propylene glycol	less toxic
19	Peroxyacetic acid	moderate toxic
20	2-hydrazinoethanol	moderate toxic

4.1.6 Comparison with previous study

As the composition of the training and test sets, endpoints used, as well as the algorithms used for model development are not the same, we can’t perform a rigorous comparison, so we have attempted to represent some simple comparative studies between the current work and previously reported literature. Mukherjee et al. [62] developed the models using small data sets in comparison with current work. Basanta et al. (2015) [58] used tree-based approaches to build QSTR and i-QSTR models for various avian species. Banjare et al. (2021) [61] presented QSTR and i-QSTR models for three avian species using a classification approach. Podder et al. [63] developed a regression-based QSTR and i-QSTR model against multiple avian species (MD, BQ, and ZF). Leszczynski et al. [60] reported ecotoxicity QSTR and i-QSTR modeling of chemicals to avian species. While regression models provide explicit quantitative predictions, classification approaches can be useful for data filtering at the outset of research. The current models are built using a regression-based method and a limited number of Simple, 2D, and easily interpretable descriptors. In this work, we have tried to develop the first PLS-based QSTR model considering LC_{50} as an endpoint to assess the toxicity of diverse pesticides against multiple avian species.

Regression-based technique is an assertive and effective approach that can confidently tackle challenges such as descriptor inter-correlation, high levels of noise, collinearity, and a large number of descriptors. In the present work, we have developed the models using large datasets of different avian species. So, it has a wide domain of applicability compared to previous studies. Additionally, we used a read-across algorithm to enhance the external predictivity and it is widely used for data-gap filing as well as widely accepted and recommended by regulatory bodies Apart from the previous studies, consequently, read-across prediction shows a better result than the previous model except for MD. Apart from the previous studies, we get additionally some new findings (specifically observation) which are related to pesticide toxicity towards avian species such as presence of C-012 (CR2X2), B07[O-P] (Presence/absence of O–P at topological distance 7), Br-094 (Br attached to C1(sp2)), B05[C-P] (Presence/absence of C – P at topological distance 5), F04[C-Cl] (Frequency of C–Cl at topological distance 4) and nRCONHR (number of secondary amides (aliphatic)) enhances the pesticides toxicity towards avian species; on the other hands, presence of $nN(CO)_2$ (number of imides (-thio)) and B05[P-Cl] (Presence/absence of P–Cl at topological distance 5) reduces the pesticides toxicity towards avian species. Furthermore, our work highlighted some extra features not mentioned in the previous studies, which are useful for pesticide toxicity assessment viz. molecular weight, presence of heteroatom, presence of bridgehead atoms, secondary aliphatic amide, and molecular refractivity. On the other hand, features like molecular branching and the presence of thio-imides contribute negatively towards the toxicity. The PPDB database was screened using developed models to show the predictivity and application in real-world data of the developed models. The current study's comparison to previously published studies is depicted in **Table 16**.

Table 16. Comparison table with previous works.

Source	Organism used	End point	Model	LV	Features	Training set		Test set	
						R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}
Present study	RNP	LC_{50}	PLS-Read across	2	6	0.63	0.53	0.60-0.71	0.60-0.71
	MD			1	6	0.60	0.58	0.71-0.75	0.63-0.68
Mukherjee et al. 2021 [62]	BQ	LD_{50}	PLS	3	10	0.65	0.58	0.64	0.64
	JQ			2	3	0.73	0.59	--	--
	RNP			2	4	0.76	0.60	0.64	0.64

	MD			2	7	0.65	0.56	0.65	0.57
	HS			1	2	0.91	0.86	0.94	0.88
Mazzatorta et al. 2006 [57]	BQ	LD_{50}	GA-SVM						
Podder et al. 2023 [63]	BQ	LD_{50}	MLR	-	7	0.715–0.719	0.694–0.700	0.722–0.732	0.722–0.732
	MD			-	8	0.689–0.708	0.626-0.695	0.620–0.639	0.620–0.638
	ZF			-	5	0.754–0.758	0.697–0.722	0.787–0.830	0.786-0.829
Banjare et.al. 2021 [61]	BQ	LD_{50}	GA-LDA along with interspecies correlation	-	-	-	-	-	-
	MD			-	-	-	-	-	-
	ZF			-	-	-	-	-	-
Basant et al. 2015 [58]	BQ	LD_{50}	Tree-based QSAR approaches	-	-	-	-	-	-
Leszczynski et al. (2020) [60]	BQ	LD_{50}	GFA-PLS	3	5	0.67	0.63	0.70	0.68
	MD			2	5	0.75	0.67	0.88	0.87
	RNH			3	4	0.89	0.80	0.87	0.87

4.2 Study 2

PLS-based QSTR models were developed using a curated dataset with pLD_{50} endpoint against *California quail*. The external and internal validation metrics of the models have been provided in Table. We used the generated models and performed intelligent consensus prediction to verify whether the prediction quality of test set compounds was improved or not (by an intelligent selection of various PLS models using the ICP tool). It was found that the consensus model 2 (CM2 model) with the best statistical metrics was selected as the winner model as shown in **Table 17**. Y-randomization was carried out to investigate the chance of occurrence of the developed model. R^2y_{rand} and Q^2y_{rand} were found to be less than the standard threshold [117], which assures that the generated models were not obtained by any chance as depicted in **Figs. 40-41**. DModX plots showed that compound 1 for Model 1, and compounds 2, 25, and 31 for Model 5 are outside the AD. On the other hand, there is no outlier compound in Model 2, Model 3, and Model 4. Scatter plot representations of the experimental against predicted toxicity are depicted in **Fig. 19.**

Table. 17. Statistical parameters of the developed PLS and consensus models.

Statistics for Training set				**Statistics for Test set**					
Model	**LV**	**R^2**	**Q^2_{LOO}**	**$Q^2_{(F1)}$**	**$Q^2_{(F2)}$**	**CCC**	**$r^2_{m(test)}$**	**$\Delta r^2_{m(test)}$**	**MAE**
IM1	2	0.701	0.601	0.762	0.690	0.880	0.65	0.15	0.33
IM2	2	0.711	0.612	0.701	0.621	0.821	0.61	0.18	0.40
IM3	2	0.671	0.562	0.753	0.694	0.864	0.68	0.10	0.34
IM4	2	0.674	0.564	0.684	0.597	0.772	0.54	0.001	0.31
IM5	3	0.645	0.531	0.707	0.615	0.815	0.55	0.03	0.31
CM0				0.815	0.751	0.883	0.70	0.10	0.30
CM1				0.812	0.754	0.881	0.70	0.10	0.30
CM2				**0.822**	**0.761**	**0.881**	**0.72**	**0.12**	**0.28**
CM3				0.743	0.672	0.842	0.65	0.14	0.34

(IM: Individual Model, CM: Consensus Model)

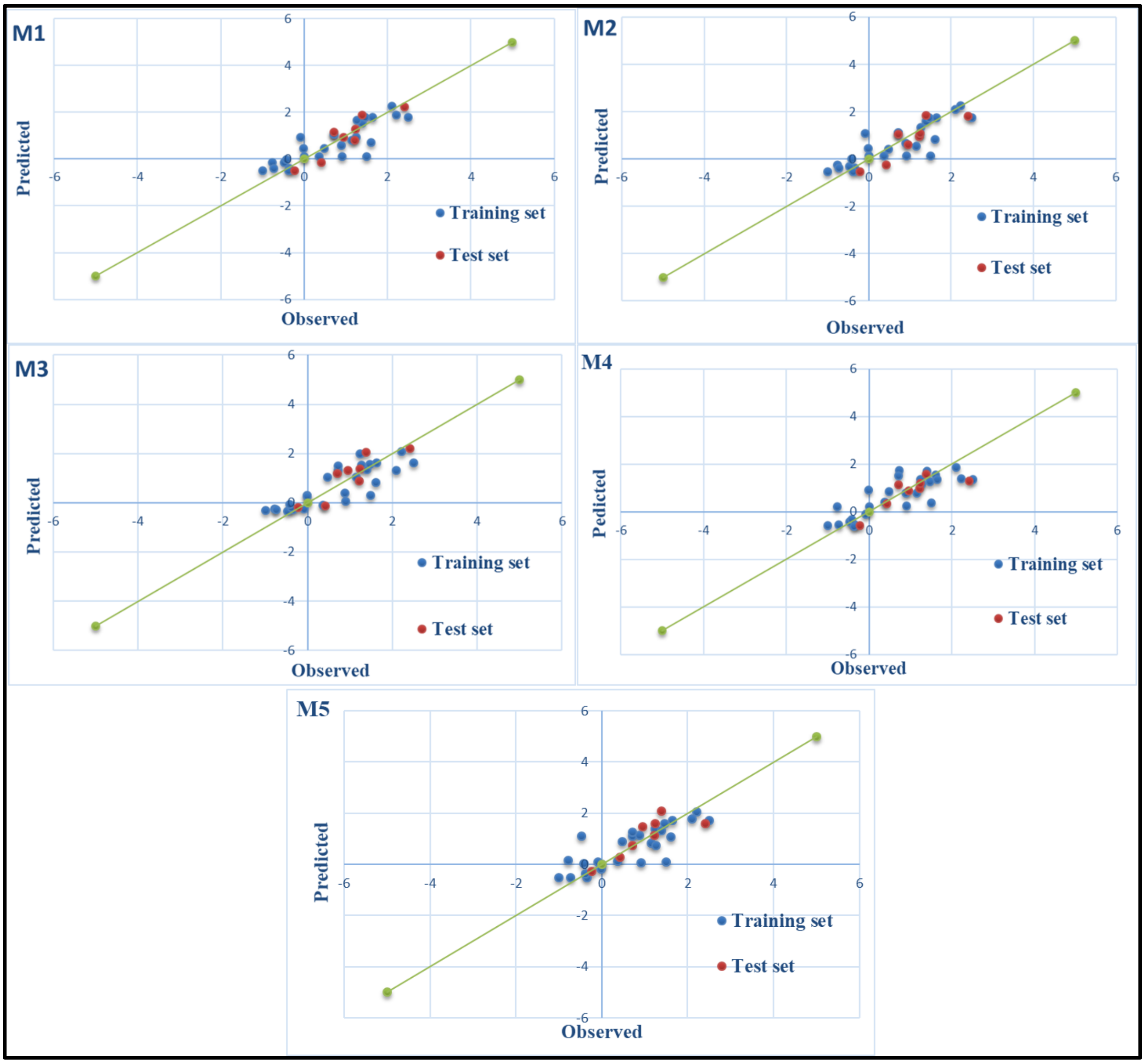


Fig. 19. Scatter plots of the constructed models.

4.2.1. Regression coefficient plot

The positive and negative contribution of the descriptors towards the modeled response value (pLD_{50}) can be categorized from the regression coefficient plot. The descriptors such as F03[C-P], B07[S-S], nR_Cs, F06[S-P], F04[C-P], Fsp3, and C% contribute positively towards toxicity which indicates that the toxicity enhanced with increasing descriptor values while the descriptors like nBM, RBN, AP, Me, F03[O-S] and T(P..Cl) show negative contribution towards toxicity which indicate that the toxicity reduced with increasing the descriptor number. The regression coefficient plot is depicted in **Figs.20 –24.**

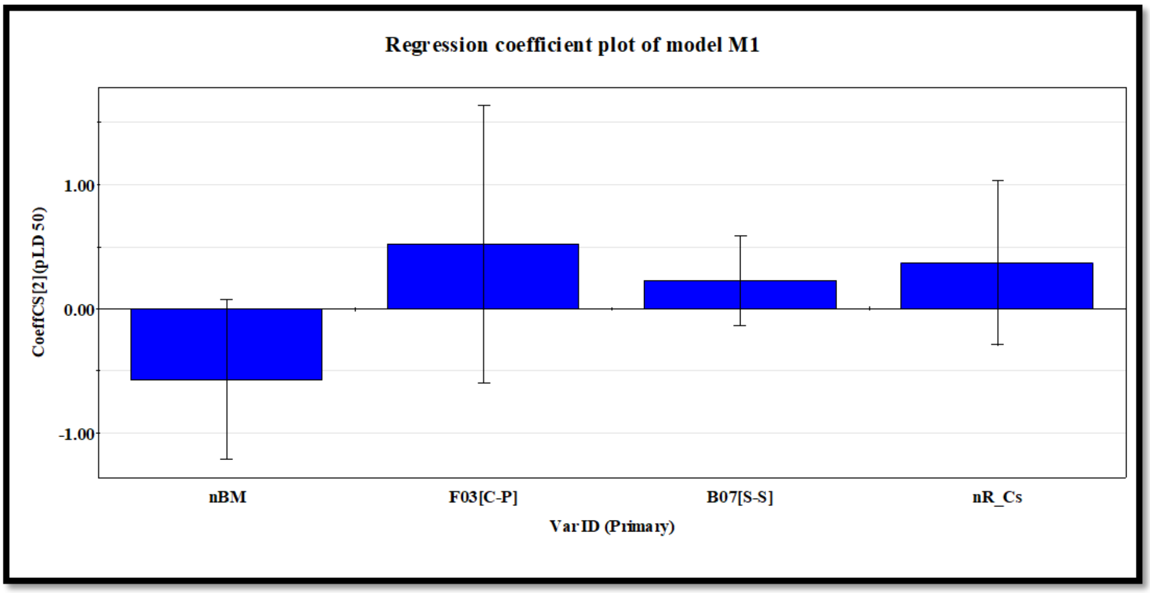


Fig. 20. Regression coefficient plot of model M1

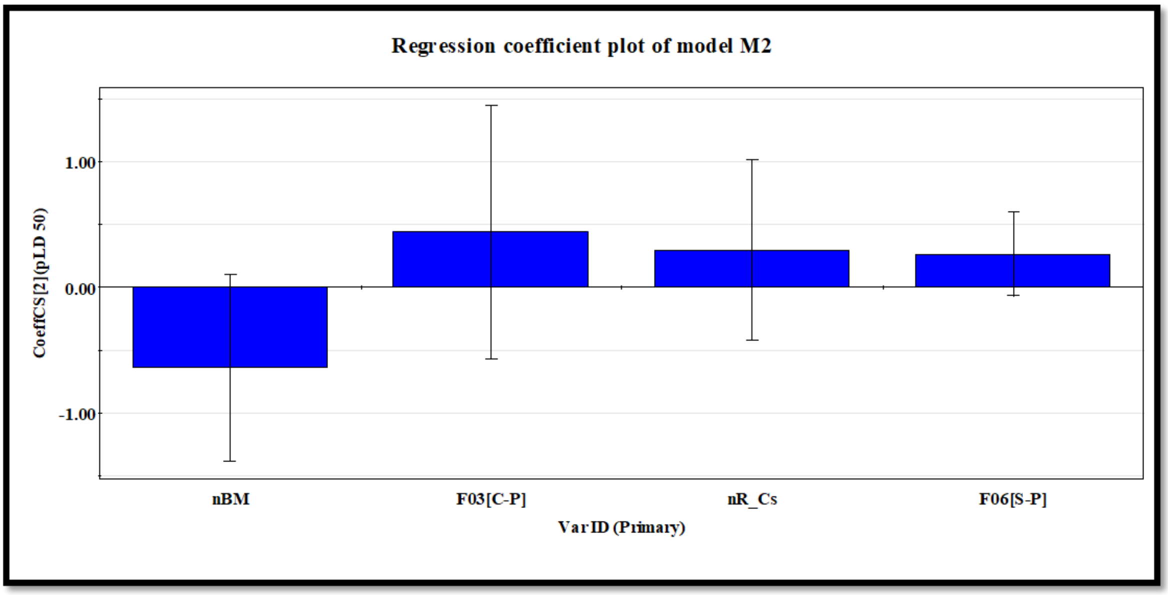


Fig. 21. Regression coefficient plot of model M2

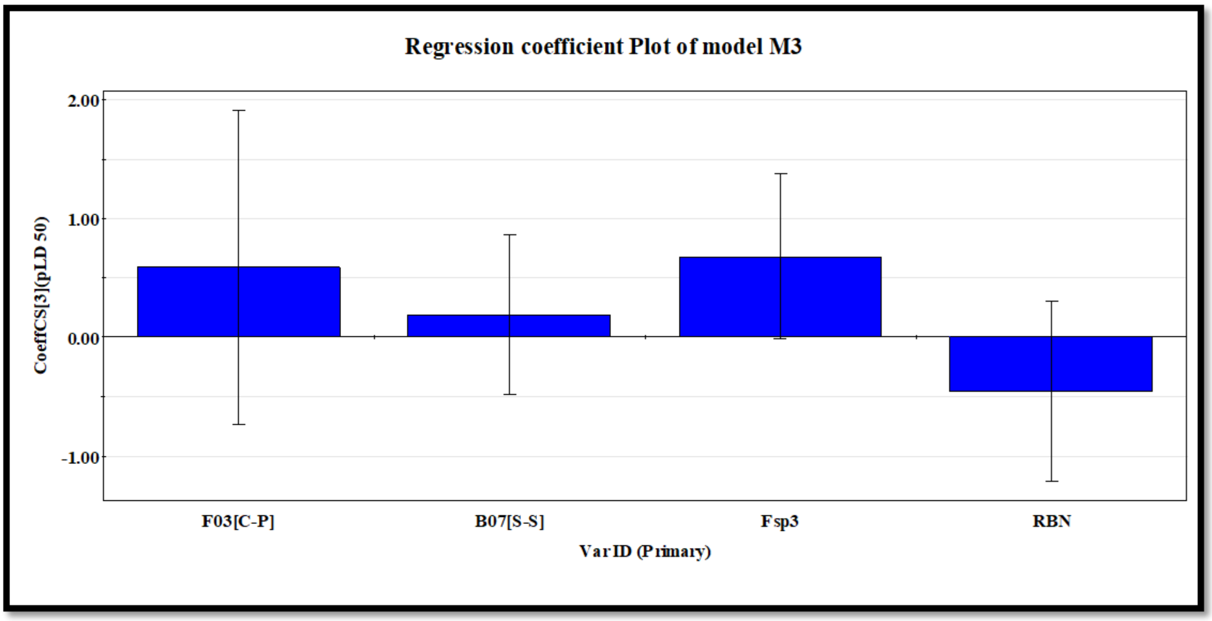


Fig. 22. Regression coefficient plot of model M3

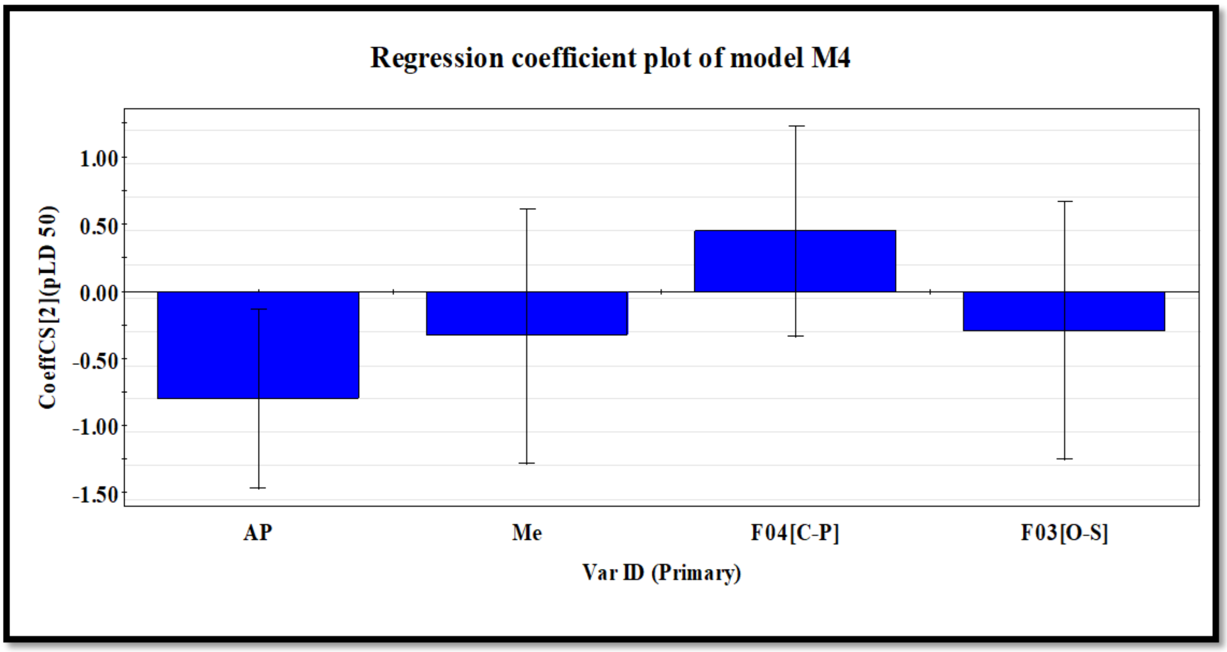


Fig. 23. Regression coefficient plot of model M4.

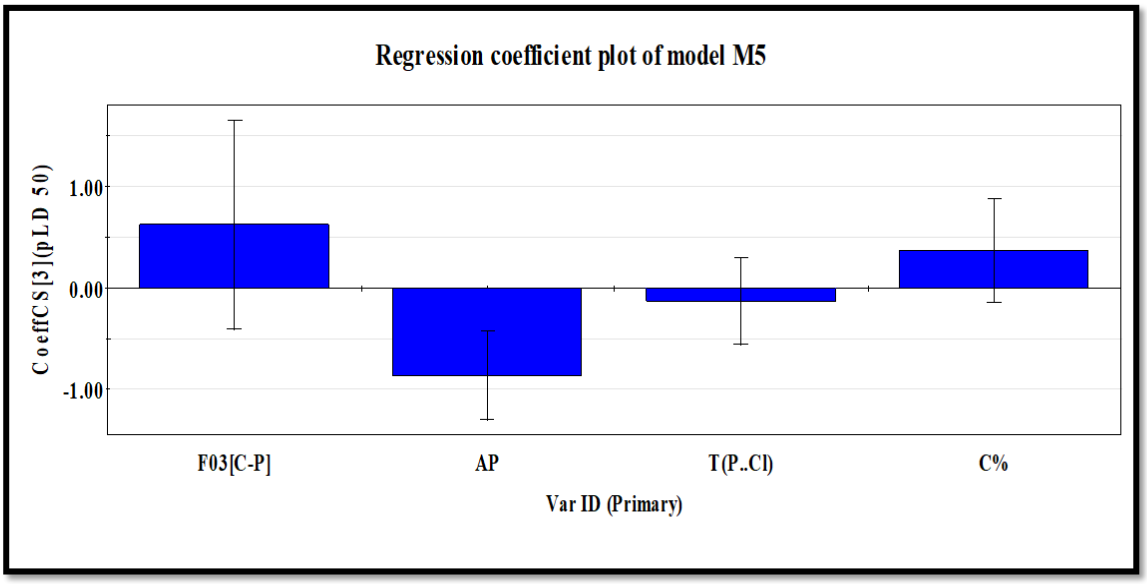


Fig. 24. Regression coefficient plot of model M5**.**

4.2.2. Variable importance plot (VIP)

The significance of the individual descriptors towards toxicity can be described for their importance toward the toxicity from the Variable Importance plot (VIP). The most significant and least significant descriptor contributing to the toxicity can be recognized by this plot. A variable is considered to have high statistical significance if it has a VIP score > 1 as opposed to one with a low VIP value [144]. According to the VIP plot depicted in **Figs. 25-29**, the influential descriptors toward toxicity in the developed model are nBM, nR_Cs, F03[C-P] and B07[S-S] in model M1; nBM, nR_Cs, F03[C-P] and F06[S-P] in model M2; Fsp3, F03[C-P], RBN and B07[S-S] in model M3; AP, F04[C-P], Me and F03[O-S] in model M4 & AP, C%, F03[C-P] and T(P..Cl) in model M5 arranged in higher to lower order as per their VIP score.

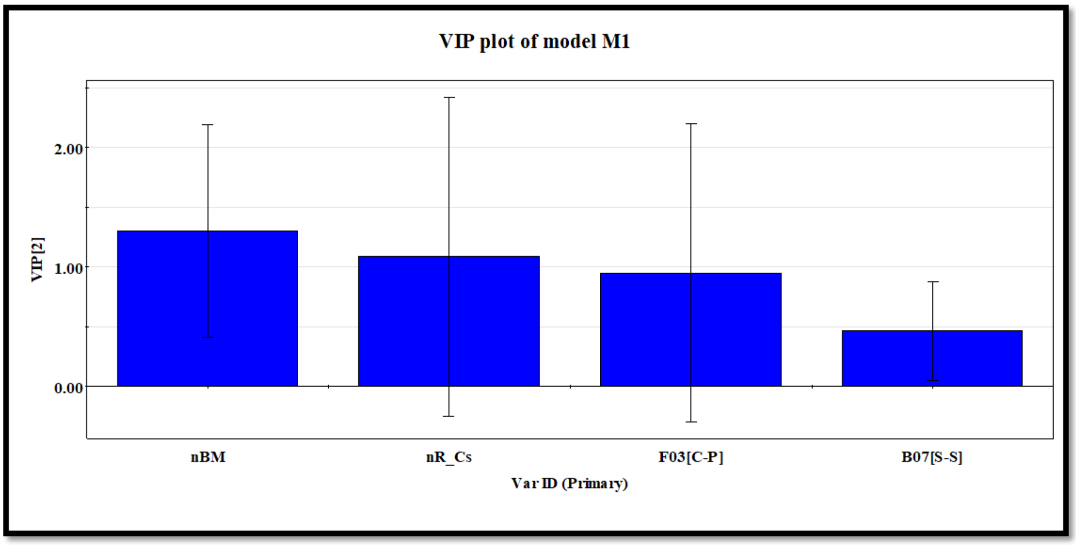


Fig. 25. VIP plot of model M1.

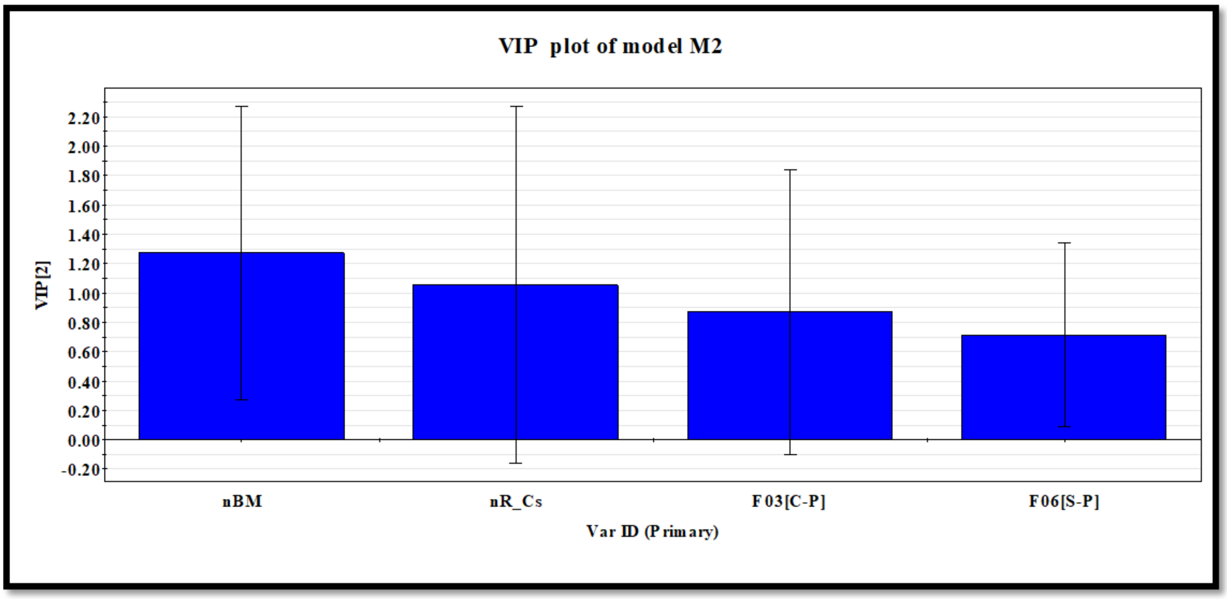


Fig. 26. VIP plot of model M2.

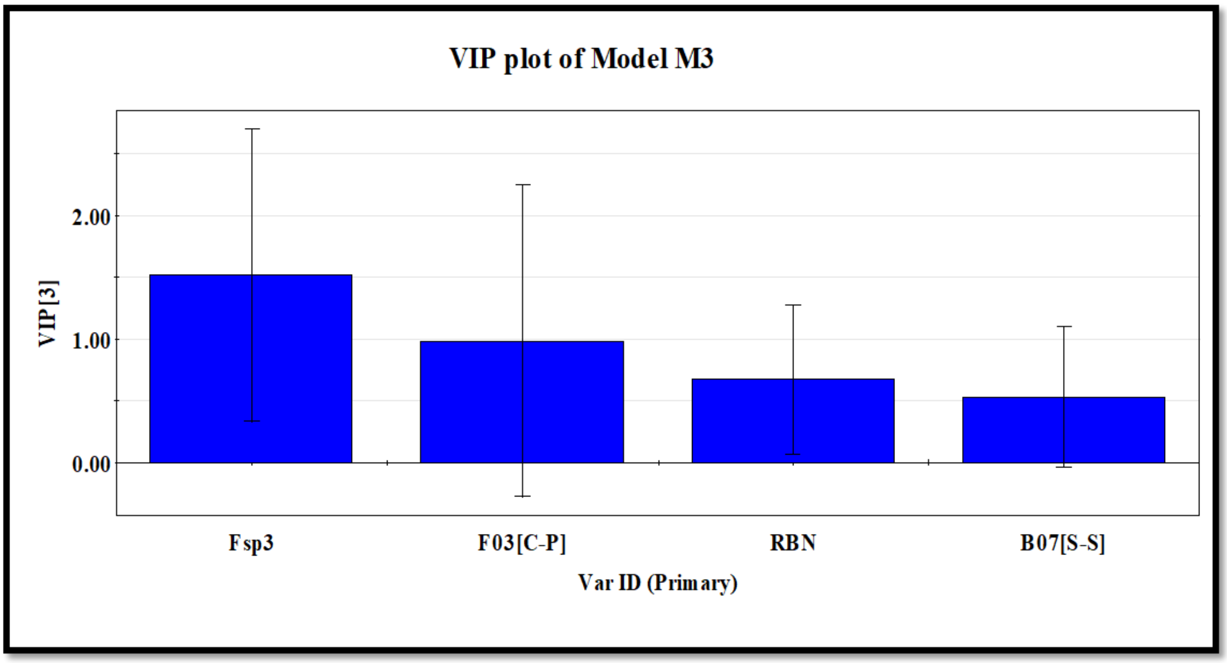


Fig. 27. VIP plot of model M3**.**

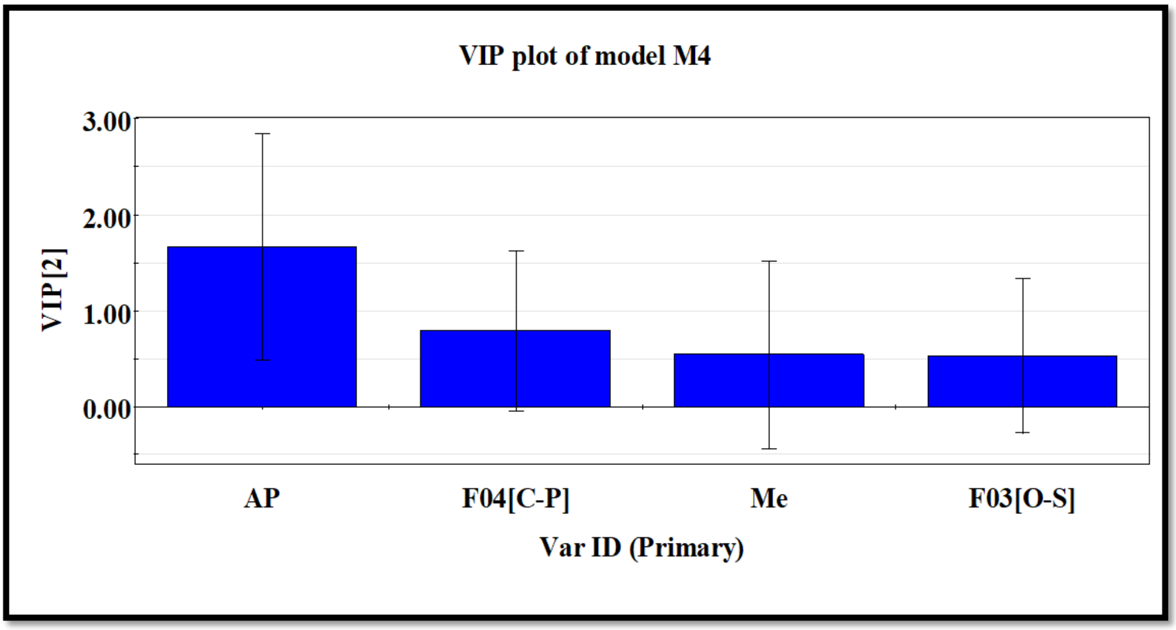


Fig.28. VIP plot of model M4.

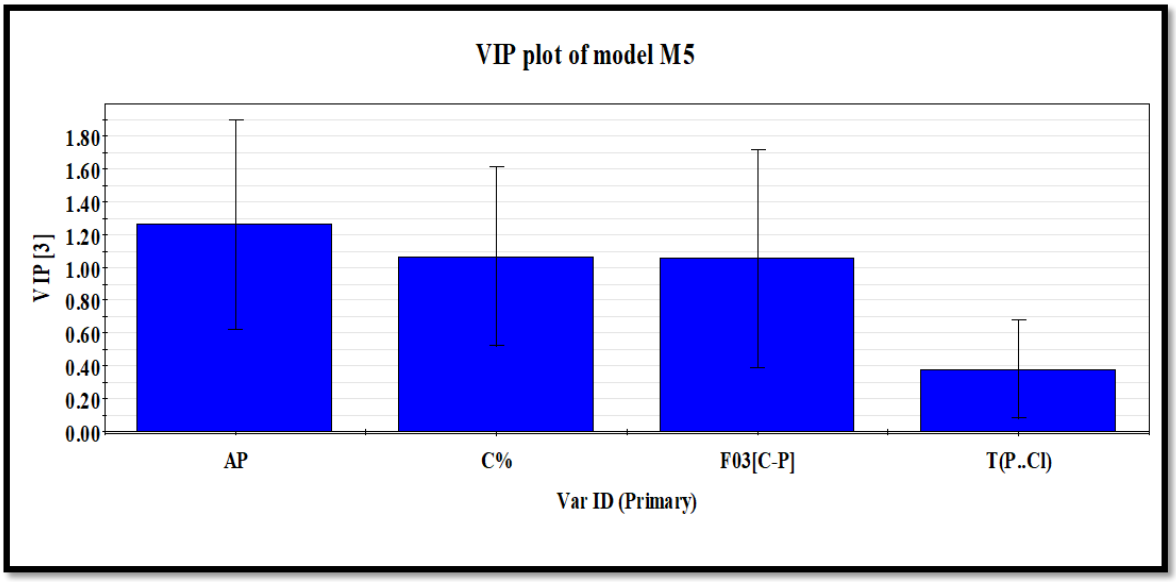


Fig. 29. VIP plot of model M5.

4.2.3. Loading plot of the generated models

The loading plot, portrayed in (**Fig. 30 - Fig. 34)** the relationship between the model's X-variables (independent variables) and Y-variables (dependent variables). The first two components of the developed models were used to generate the loading plot. This plot clarifies how various variables impact the models. The descriptors with maximum distance from the origin are thought to have a higher influence on response value as well as on models. According to the loading plot nBM descriptor in the case of model M2, Fsp3 descriptor in the case of model M3, AP descriptor in the case of model M4 and model M5 were the most impactful variables for the respective models as they were present farthest from the origin.

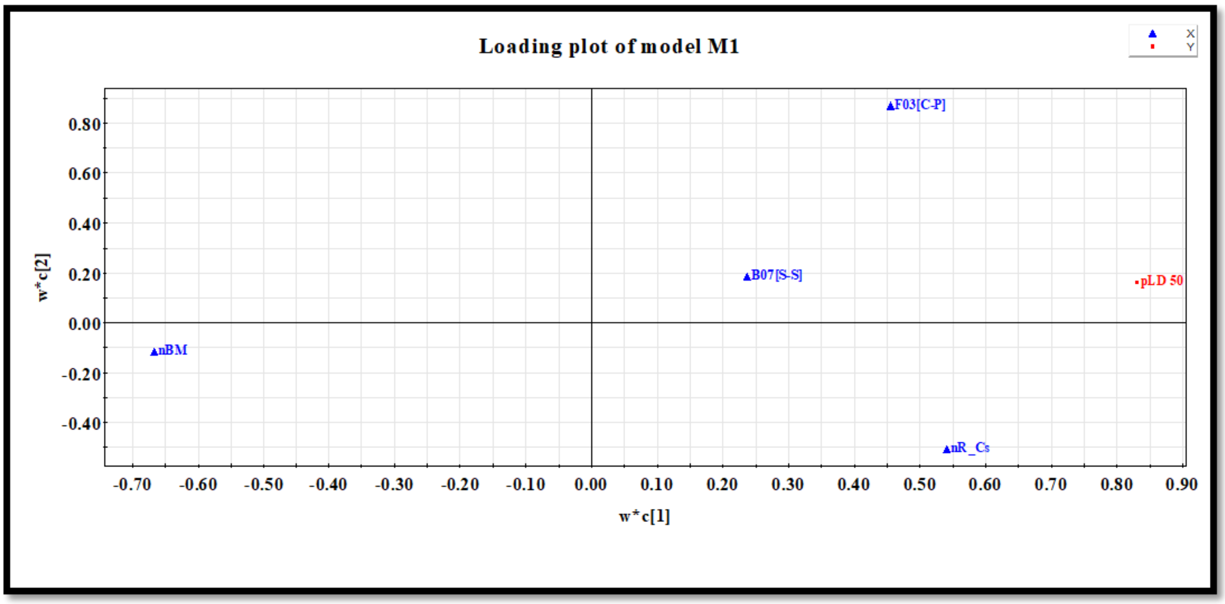


Fig. 30. Loading plot of model M1

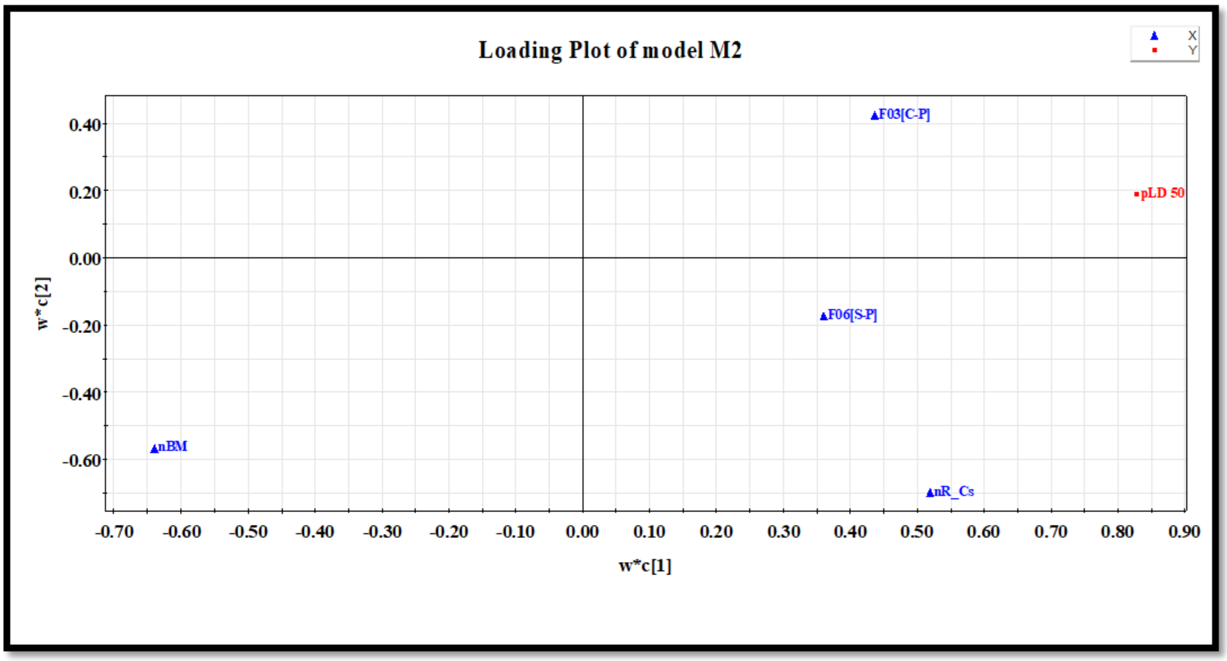


Fig. 31. Loading plot of model M2.

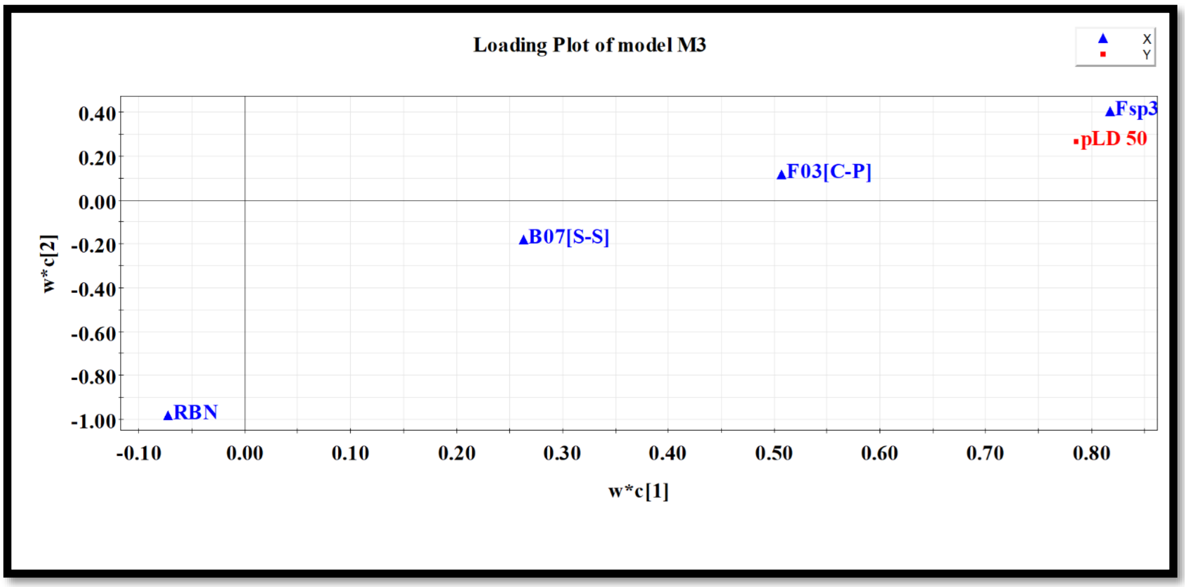


Fig. 32. Loading plot of model M3.

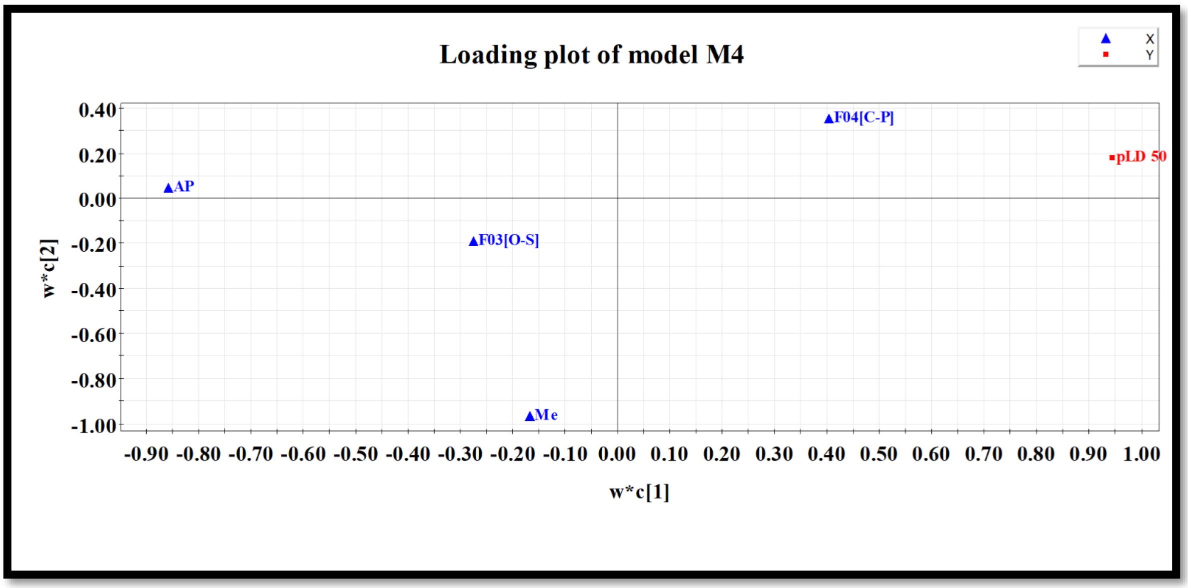


Fig. 33. Loading plot of model M4.

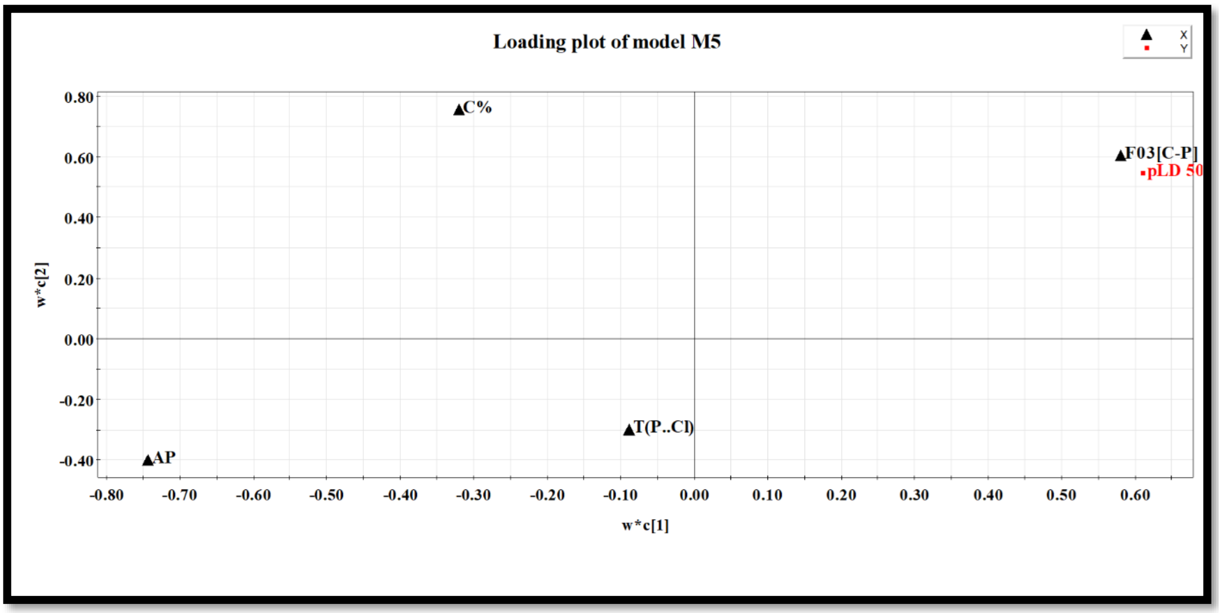


Fig. 34. Loading plot of model M5.

4.2.4 Score plot

A score plot [153] illustrates the disposition of compounds within the hypothetical ellipse representing the latent variable space for reliable prediction. The affirmation of a compound within AD involves confirming its presence within or outside the ellipse on the plot. No compound was found outside the ellipse in the case of models M1, M2, and M3. On the other hand, one compound is situated outside the ellipse in models M4 and M5. The score plots of the developed models are illustrated in **Figs. 35-39.**

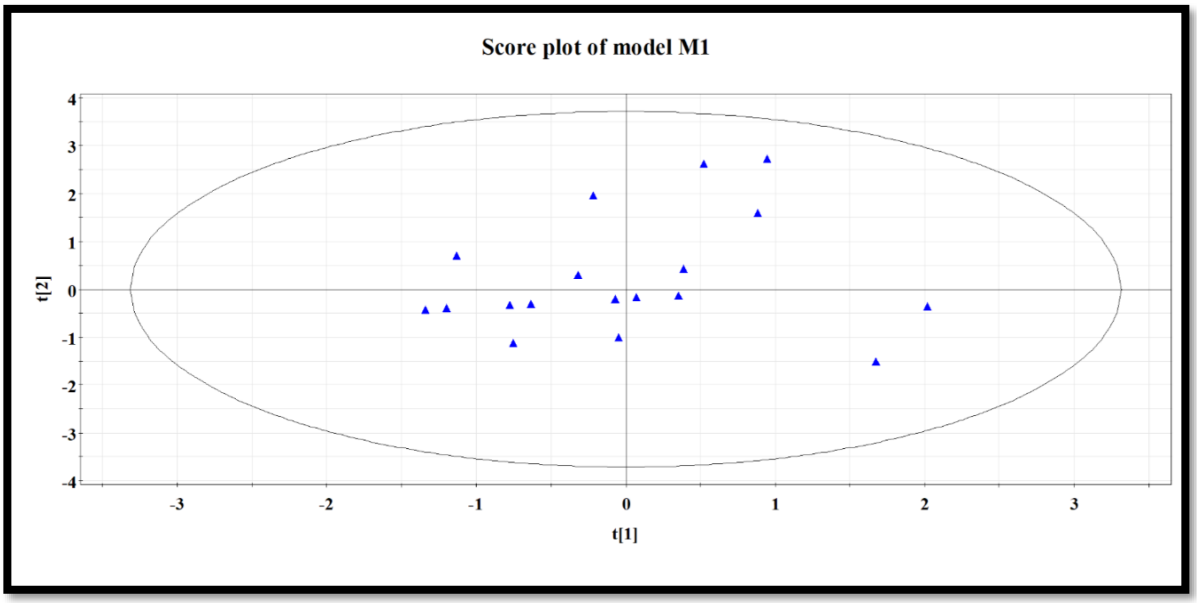


Fig. 35. Score plot of model M1.

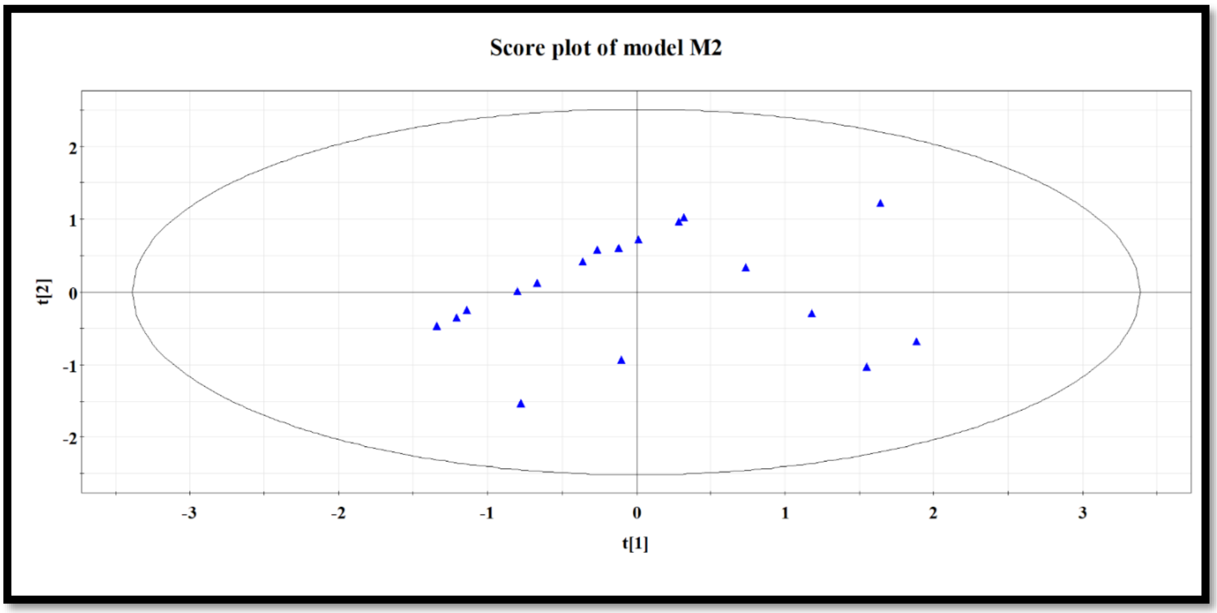


Fig. 36. Score plot of model M2.

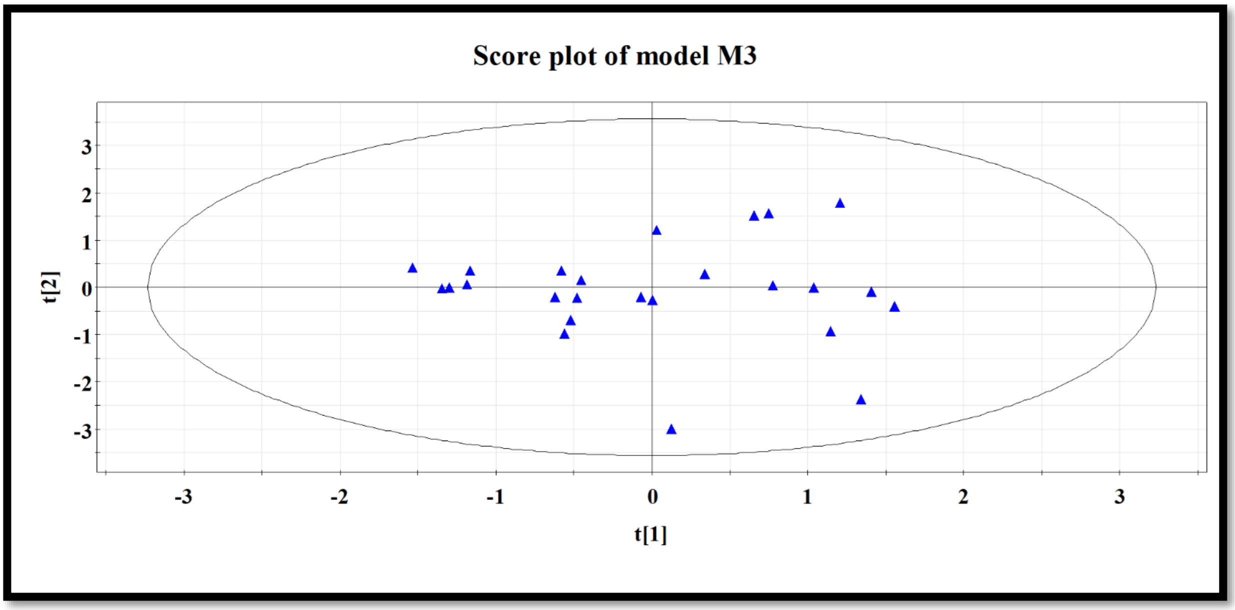


Fig. 37. Score plot of model M3.

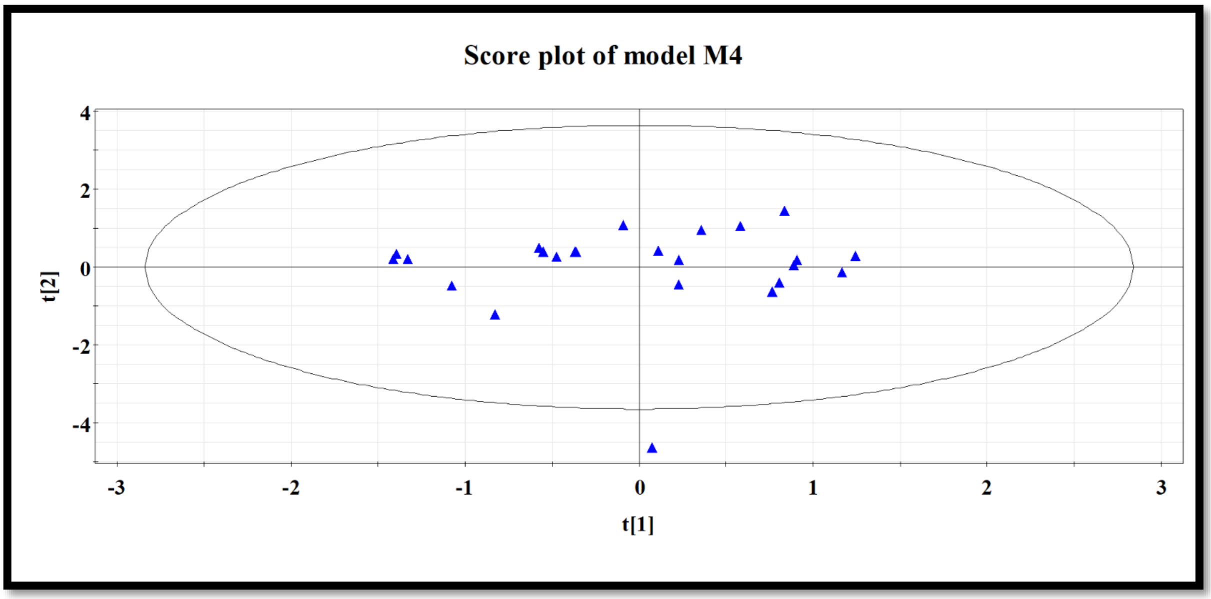


Fig. 38. Score plot of model M4.

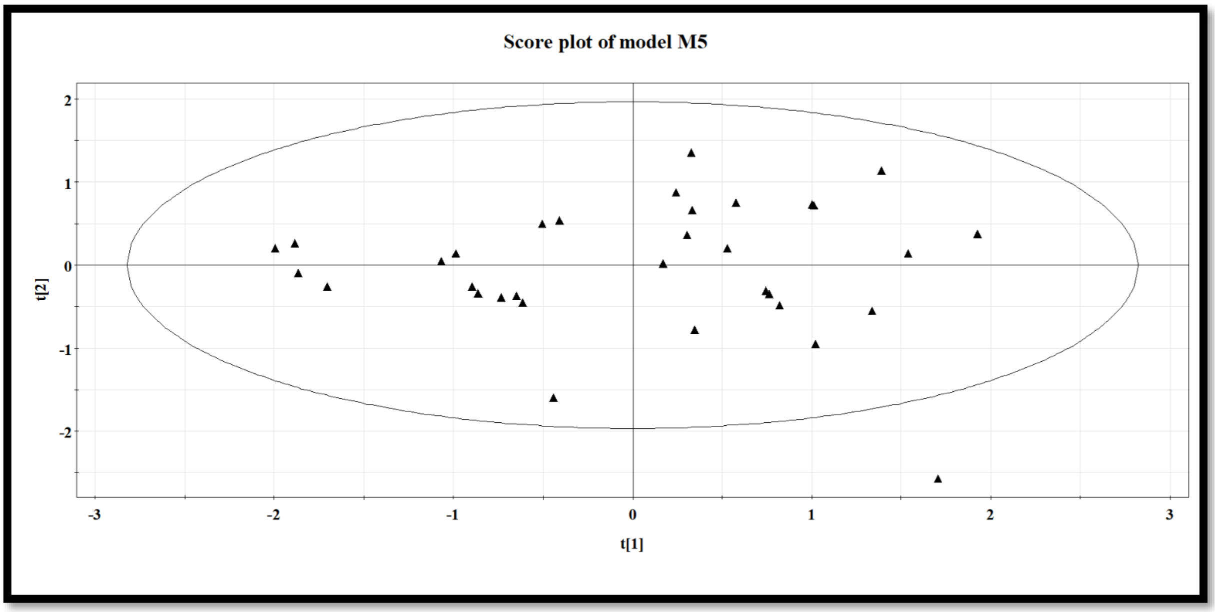


Fig. 39. Score plot of model M5.

4.2.5 Y-randomization study

In our study, we used Y-randomization, where for the training set, the X data (descriptors) remained fixed and Y data (response) were scrambled randomly, and the model was fitted to the permuted data and compared with the best fit. The number of permutations varies; here it is 100 permutations. The horizontal axis contains the correlation coefficient values for those 100 different combinations and the vertical axis contains their respective determination coefficient values (R^2 and Q^2). The basic statistics of randomization models (Q^2 and R^2) should be poor and not within the range of those for acceptable regression models. Otherwise, each resulting model may be considered as a chance correlation. The randomization results (R_Y^2 < 0.3 and Q_Y^2 <0.05) suggested that the models were not obtained by any chance as shown in **Figs. 40-44**.

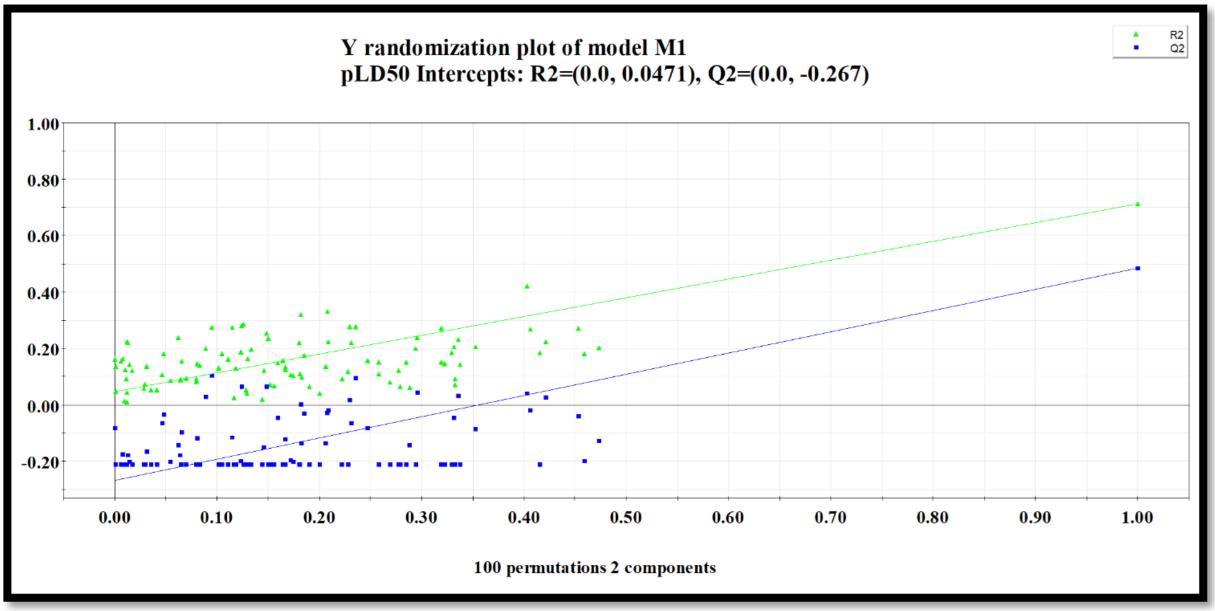


Fig. 40. Y-randomization plot of model M1.

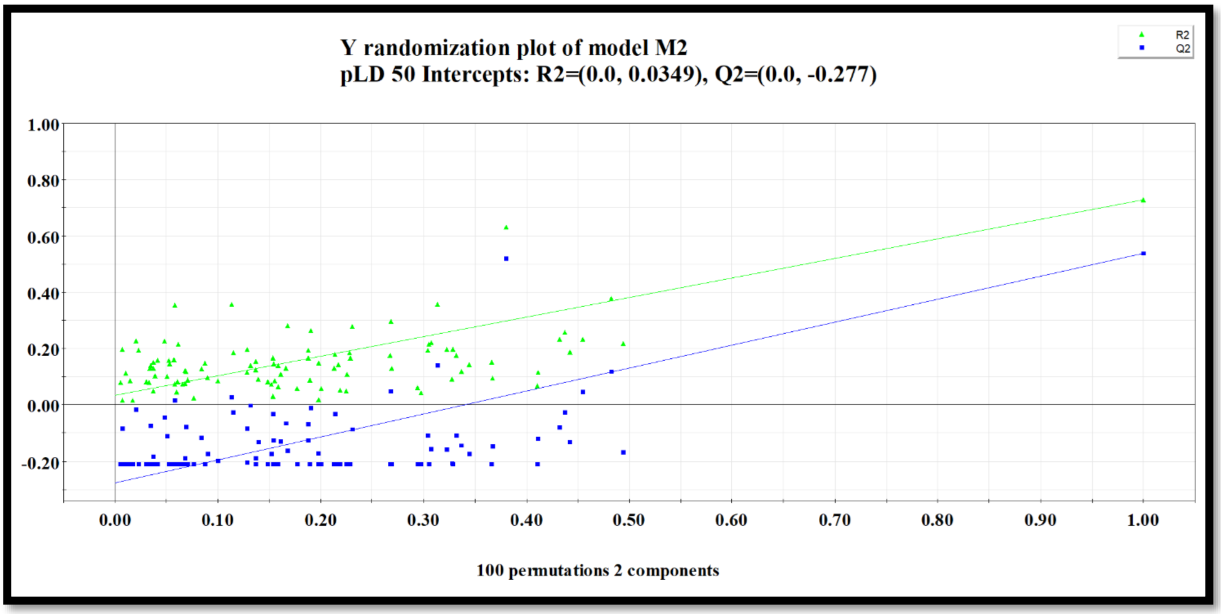


Fig. 41. Y-randomization plot of model M2.

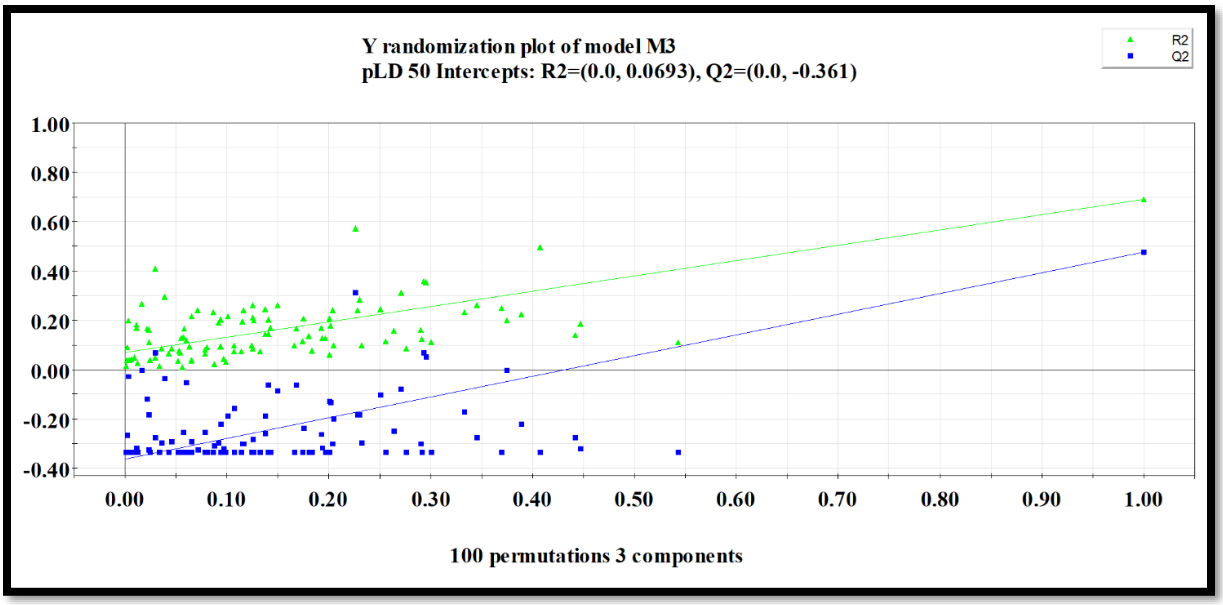


Fig. 41. Y-randomization plot of model M3.

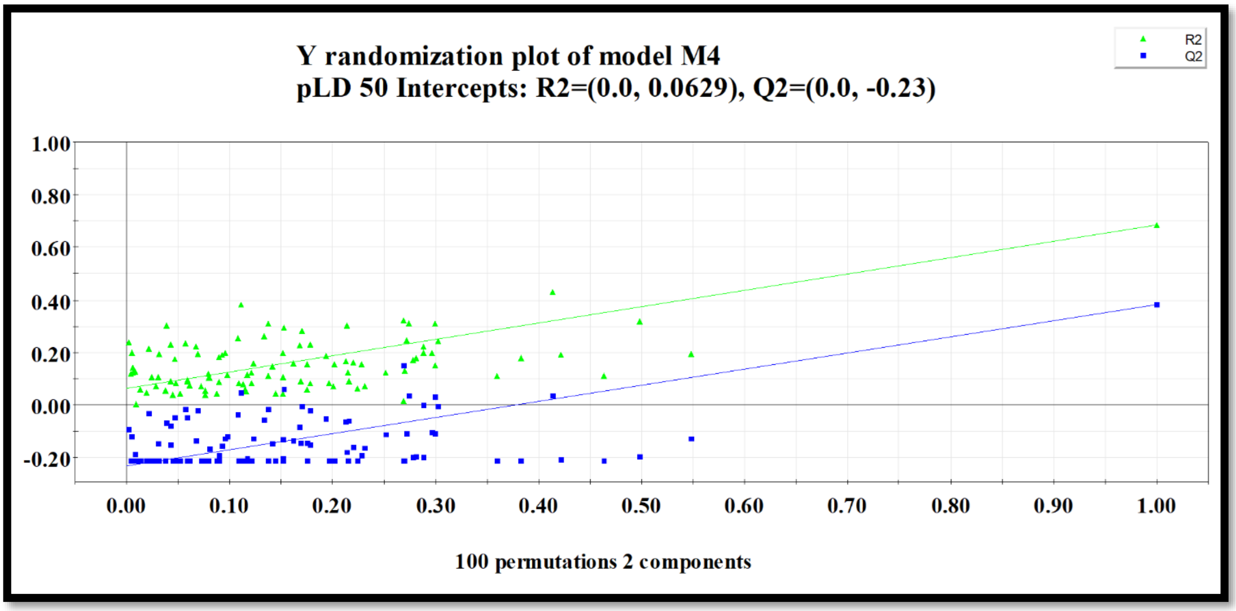


Fig. 42. Y-randomization plot of model M4.

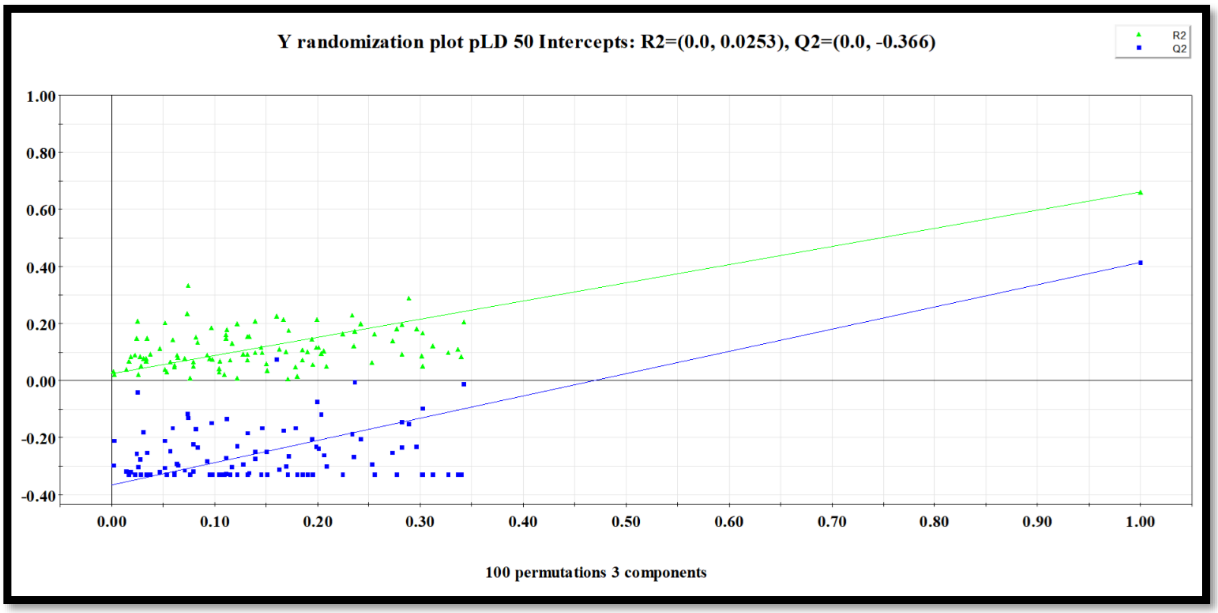


Fig. 43. Y-randomization plot of model M5.

4.2.6 DModX plot

The Dmodx plot **(Figs. 45-49)** shows that 1 compound for Model 2 and Model 4, 2 compounds for Model 5 are outside the AD. On the other hand, there is no outlier in Model 1 and Model 3. Such a low number of outliers signifies that the developed model is reliable and demonstrates the suitability of the same for toxicity prediction.

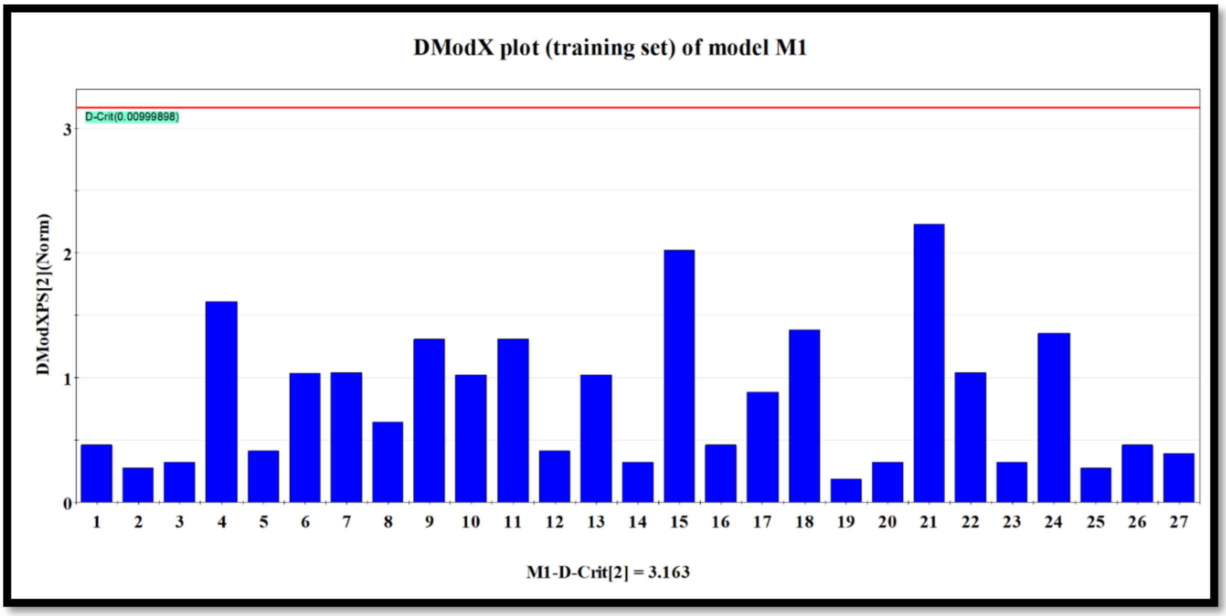


Fig. 45. DmodX plot of model M1.

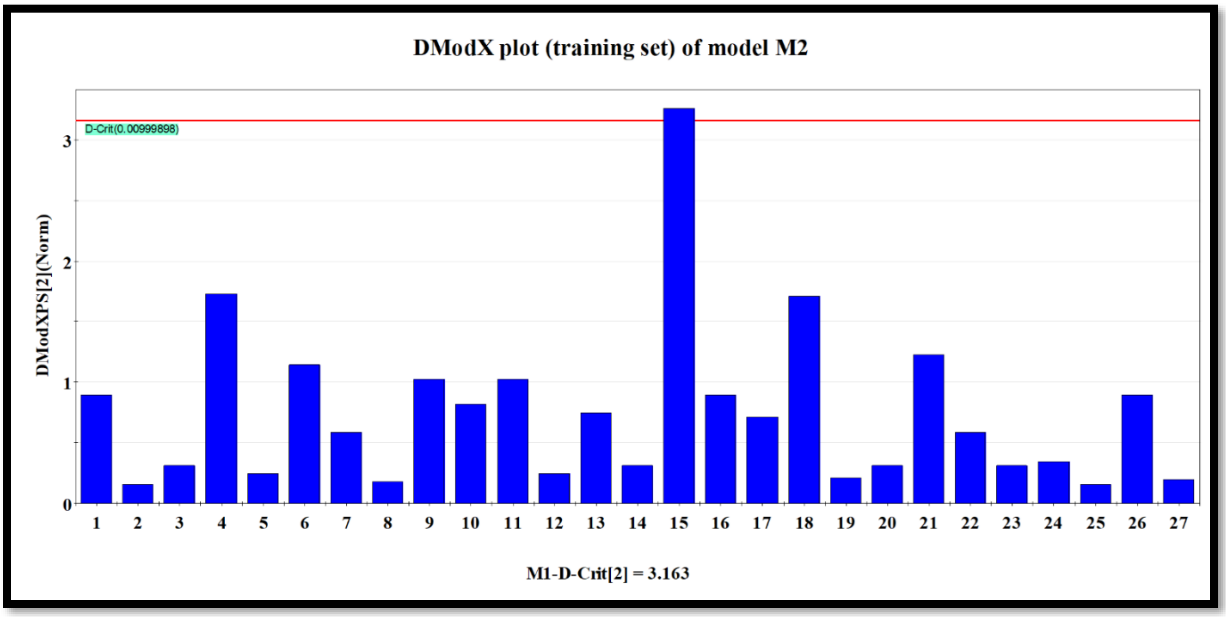


Fig. 46. DmodX plot of model M2.

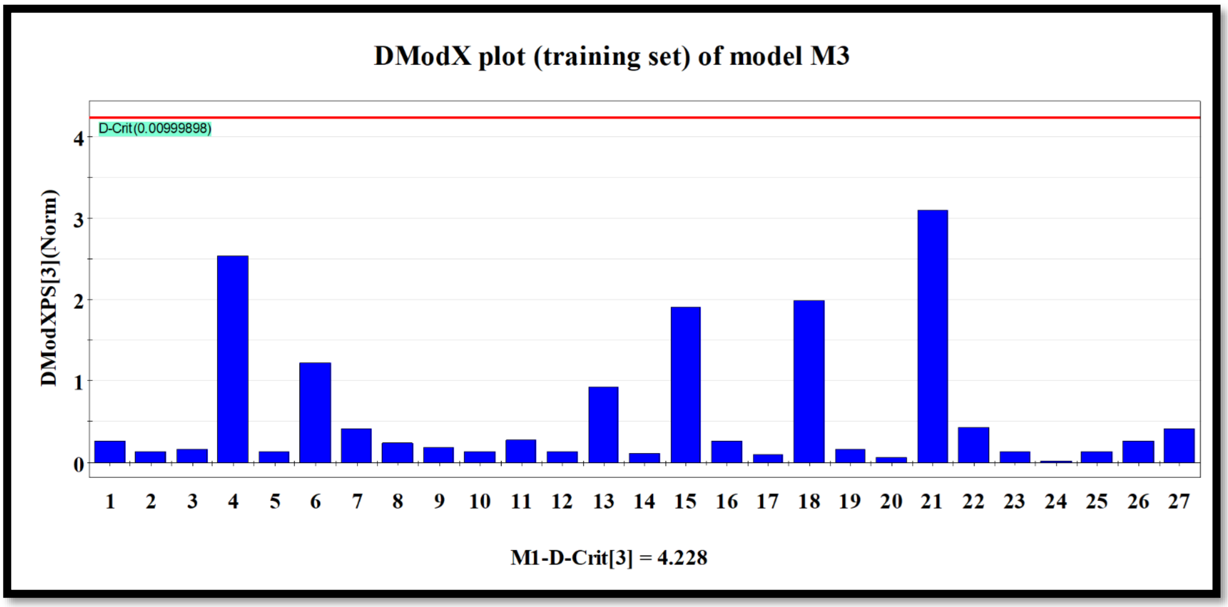


Fig. 47. DmodX plot of model M3.

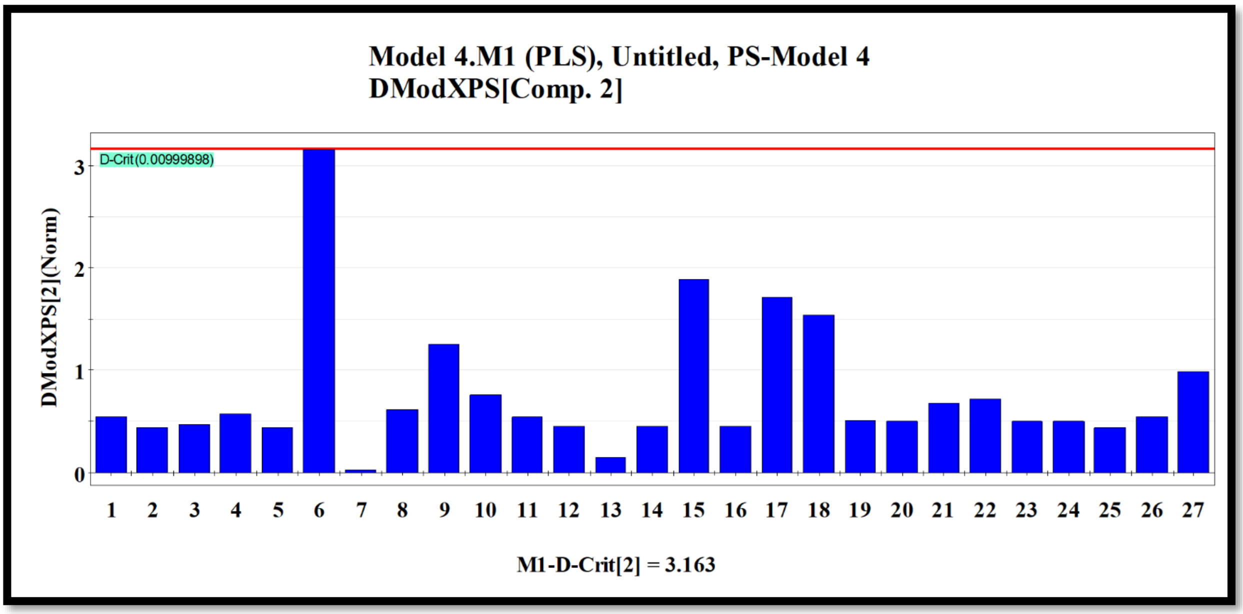


Fig. 48. DmodX plot of model M4.

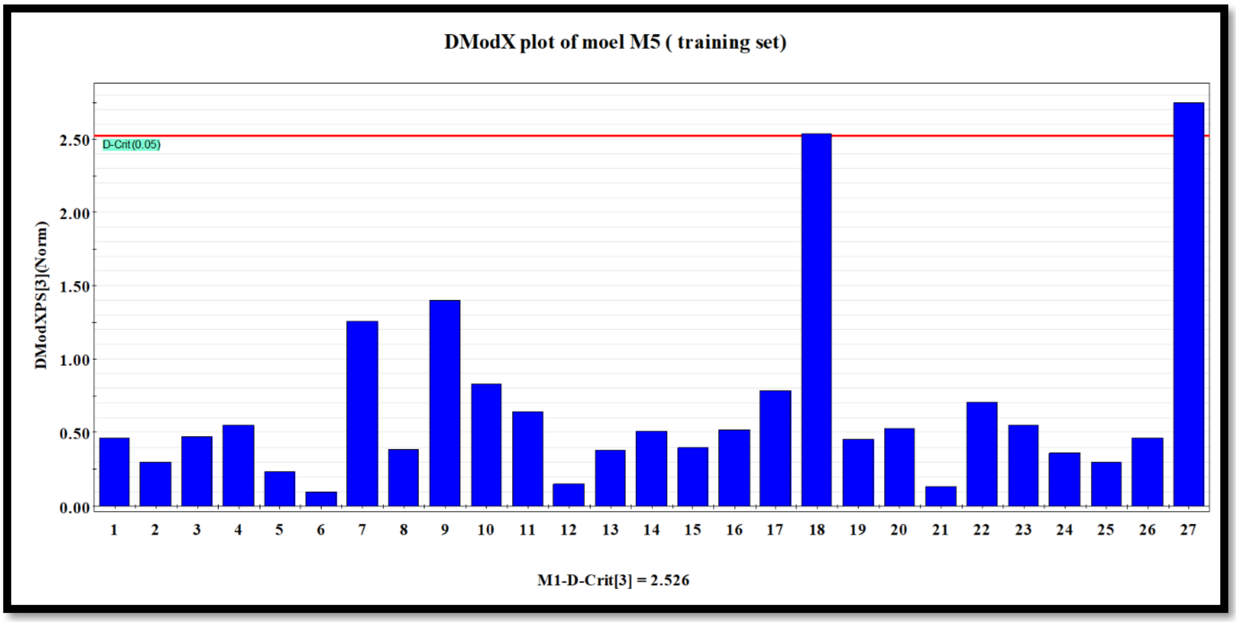


Fig. 49. DmodX plot of model M5.

4.2.7 Mechanistic interpretation of descriptors used in the QSTR model

Table 18. Mechanistic interpretation of descriptors employed in Models

Sl. no	Descriptor	Type	Description	Contribution
1	nBM	Constitutional index	Number of multiple bonds	-ve
	Mechanistic interpretation			

	Increasing the number of multiple bonds will diminish the compound’s toxicity (inversely related to the toxicity as indicated by the negative regression coefficient). We observed that the multiple bonds in the compounds are either situated adjacent to atoms such as nitrogen, sulfur, chlorine, and oxygen, which imparts hydrogen bonding with water or form polar bond which imparts polarity to the compounds. As a result, the hydrophilicity of the respective compounds increases which lowers the toxicity value as demonstrated in compound **8** and the opposite occurs in compound **9** given in **Fig. 51**.			
2	F03[C-P]	2D Atom Pairs	Frequency of C-P at topological distance 3	+ve
	Mechanistic interpretation This fragment represents the presence of a phosphorus atom which is toxic [154]. The toxicity in the respective species is enhanced with the increase of this fragment as depicted in compound **31** and vice versa as demonstrated in compound **5** in **Fig. 50**.			
3	B07[S-S]	2D Atom Pairs	Existence/non-existence of S-S at topological distance 7	+ve
	Mechanistic interpretation This feature characterizes the existence of two sulfur atoms that enhanced the overall electronegativity of the compound. The increase of electronegativity may result in the generation of reactive oxygen species (ROS) [130], which may be responsible for toxicity enhancement toward the respective species as demonstrated in compound **3**, while the opposite occurs in the case of compound **33** depicted in **Fig. 50**.			
4	nR=Cs	Functional group counts	Number of aliphatic secondary C(sp^2)	+ve
	Mechanistic interpretation This descriptor characterizes the number of sp^2 hybridized carbon atoms which means the degree of unsaturation [63]. Generally, unsaturated compounds are more reactive and toxic in nature [63] as demonstrated in compound **12** and vice versa occurs in compound **8** (given in **Fig. 50**).			

<table>
<tr><td rowspan="2">5</td><td>F06[S-P]</td><td><b>2D Atom Pairs</b></td><td>Frequency of S-P at topological distance 6</td><td>+ve</td></tr>
<tr><td colspan="4"><b>Mechanistic interpretation</b>
Phosphorous itself is toxic, so its presence makes the compound more toxic. Sulfur atoms are electronegative. Sulfur atoms may create hydrogen bonds with the DNA of the reference species and make the DNA unstable which leads to the death of the reference organism [155]. Thus, toxicity value may be enhanced if the molecule contains more S and P atoms as shown in compound <b>3</b>, and vice versa as traced in compound <b>6</b> depicted in <b>Fig. 50.</b></td></tr>
<tr><td colspan="5"></td></tr>
<tr><td rowspan="2">6</td><td>Fsp3</td><td>Constitutional indices</td><td>Number of sp3 hybridized carbons/total carbon count</td><td>+ve</td></tr>
<tr><td colspan="4"><b>Mechanistic interpretation</b>
The presence of this descriptor leads to the enhancement of the alkyl chain length of the compound (enhancement in the size of the compound) which will cause toxicity by raising the lipophilicity of that compound [156]. This feature has a positive contribution towards the response which shows the toxicity increases with an increase in the numerical value of the descriptor as depicted in compound <b>25</b> and oppositely occurs as per compound <b>16</b> highlighted in <b>Fig. 50</b>.</td></tr>
<tr><td rowspan="2">7</td><td>RBN</td><td>Constitutional indices</td><td>Number of rotatable bonds</td><td>-ve</td></tr>
<tr><td colspan="4"><b>Mechanistic interpretation</b>
RBN descriptor represents the number of rotatable bonds that contribute negatively towards the modeled response. A molecule with more rotatable bonds has a lesser effect on oral bioavailability as a result chances of inducing toxicity of that compound are also reduced [63] as shown in compound <b>19</b> and vice versa as illustrated in compound <b>1</b> in <b>Fig 51</b>.</td></tr>
<tr><td rowspan="2">8</td><td>AP</td><td>Ring descriptor</td><td>Aromatic proportion</td><td>-ve</td></tr>
<tr><td colspan="4"><b>Mechanistic interpretation</b>
This feature represents the presence of aromatic rings in the compound’s structure. Since aromatic rings are stable and less reactive[153], they loosely</td></tr>
</table>

	interact with any receptor protein. Thereby, making the compound less toxic as demonstrated in compounds **16** and vice versa as shown in compound **18** (Given in **Fig. 51**.)			
9	Me	Constitutional indices	Mean atomic Sanderson electronegativity (scaled carbon atom)	-ve
	Mechanistic interpretation Mean atomic Sanderson electronegativity illustrates the molecular polarity [157] which is responsible for the hydrophilicity of the compound. Hydrophilicity reduces the penetration ability of the compound into the lipophilic cell membrane, which leads to diminishing the toxicity of the compound towards the reference species as depicted in compound **26** and vice versa in compound **31** provided in **Fig. 51**.			
10	F04[C-P]	2D Atom Pairs	Rate of occurrence of C-P at topological distance 4	+ve
	Mechanical interpretation Generally, phosphorus atoms are toxic [158]. The presence of carbon and phosphorous at topological distance 4 increases the size of molecules, making them more lipophilic [82]. Lipophilic compound easily crosses the membrane (more accumulative) of reference organism, making it more toxic as highlighted in compound **31** whereas the absence of this feature lowers the toxicity as shown in compound **19**, provided in **Fig. 50.**			
11	F03[O-S]	2D Atom Pairs	Rate of occurrence of O-S at topological distance 4	-ve
	Mechanistic interpretation The presence of these two polar atoms (oxygen and sulfur) enhanced the overall polarity of the compound, which leads to an increase the hydrophilicity, and hydrophilic compound has a low penetration ability into the lipophilic cell membrane (easily excreted out from the body of reference organism). Thereby, reducing the toxicity of pesticides as shown in compound **34** and inversely occurs in compound **18** displayed in **Fig. 51**.			
12	C%	Constitutional indices	percentage of C atoms	+ve

<table>
<tr><td></td><td colspan="4"><b>Mechanistic interpretation</b>
The alkyl chain length of the corresponding chemical enhanced when the percentage of carbon atoms increased which resulted in a rise in lipophilicity [158] and ultimately leads to enhancement of compound’s toxicity as demonstrated in compound <b>9</b> and oppositely occurs in case of compound <b>5</b> as illustrated in <b>Fig. 50</b>.</td></tr>
<tr><td>13</td><td>T(P..Cl)</td><td>2D Atom Pairs</td><td>The sum of topological distance between P..Cl</td><td>-ve</td></tr>
<tr><td></td><td colspan="4"><b>Mechanistic interpretation</b>
The presence of chlorine atom may responsible to form hydrogen bond with water, which make the compound hydrophilic and reduce the toxicity[62] as evidenced in compound <b>5</b> and inversly occurred in compound <b>20</b> as depicted in <b>Fig. 51.</b></td></tr>
</table>

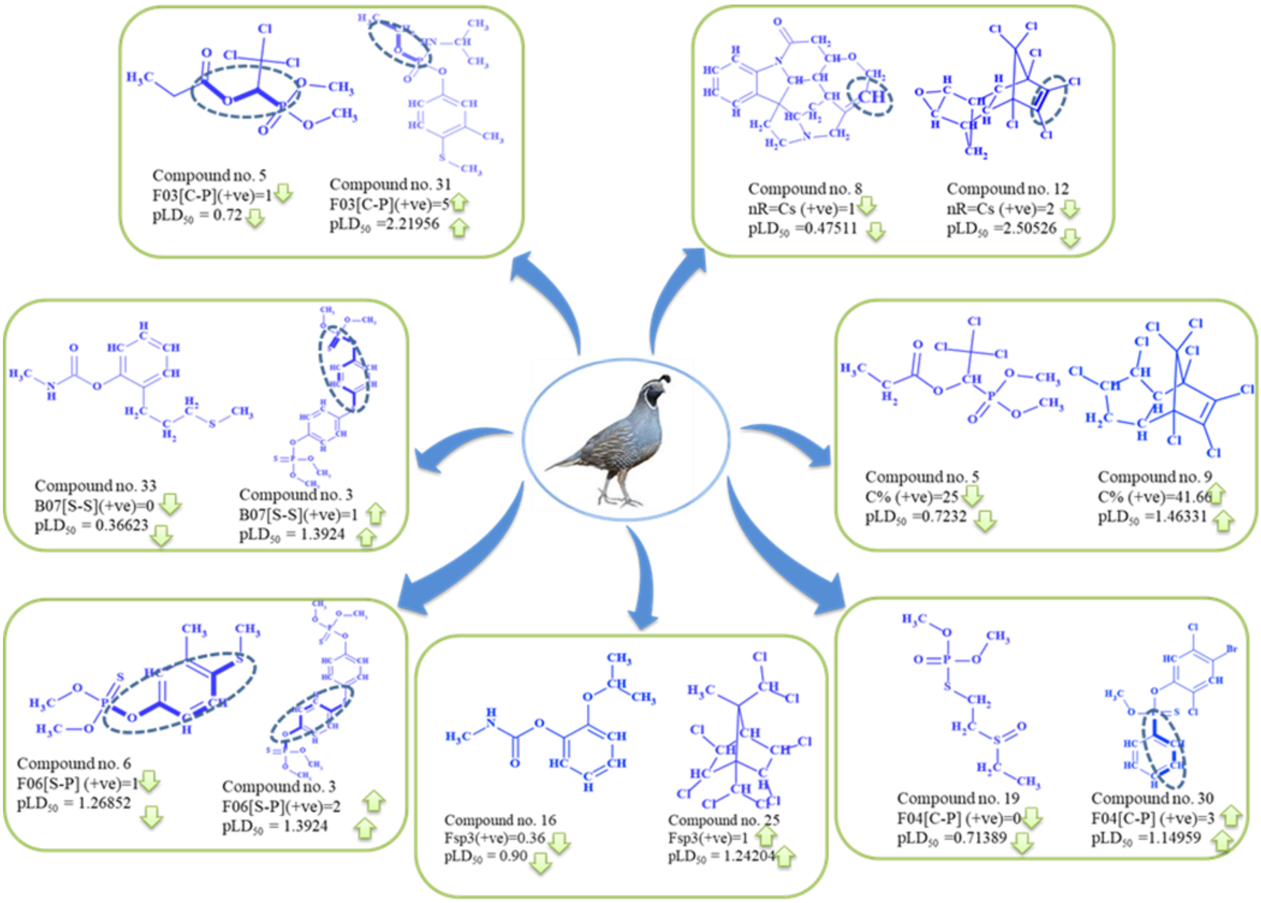

Compound no. 5 F03[C-P](+ve)=1 pLD$_{50}$ = 0.72
Compound no. 31 F03[C-P](+ve)=5 pLD$_{50}$ =2.21956
Compound no. 8 nR=Cs (+ve)=1 pLD$_{50}$ =0.47511
Compound no. 12 nR=Cs (+ve)=2 pLD$_{50}$ =2.50526
Compound no. 33 B07[S-S](+ve)=0 pLD$_{50}$ = 0.36623
Compound no. 3 B07[S-S](+ve)=1 pLD$_{50}$ = 1.3924
Compound no. 5 C% (+ve)=25 pLD$_{50}$ =0.7232
Compound no. 9 C% (+ve)=41.66 pLD$_{50}$ =1.46331
Compound no. 6 F06[S-P] (+ve)=1 pLD$_{50}$ = 1.26852
Compound no. 3 F06[S-P](+ve)=2 pLD$_{50}$ = 1.3924
Compound no. 16 Fsp3(+ve)=0.36 pLD$_{50}$ = 0.90
Compound no. 25 Fsp3(+ve)=1 pLD$_{50}$ = 1.24204
Compound no. 19 F04[C-P] (+ve)=0 pLD$_{50}$ =0.71389
Compound no. 30 F04[C-P] (+ve)=3 pLD$_{50}$ =1.14959

Fig. 50. Depiction of negatively contributed descriptors toward toxicity against California quail.

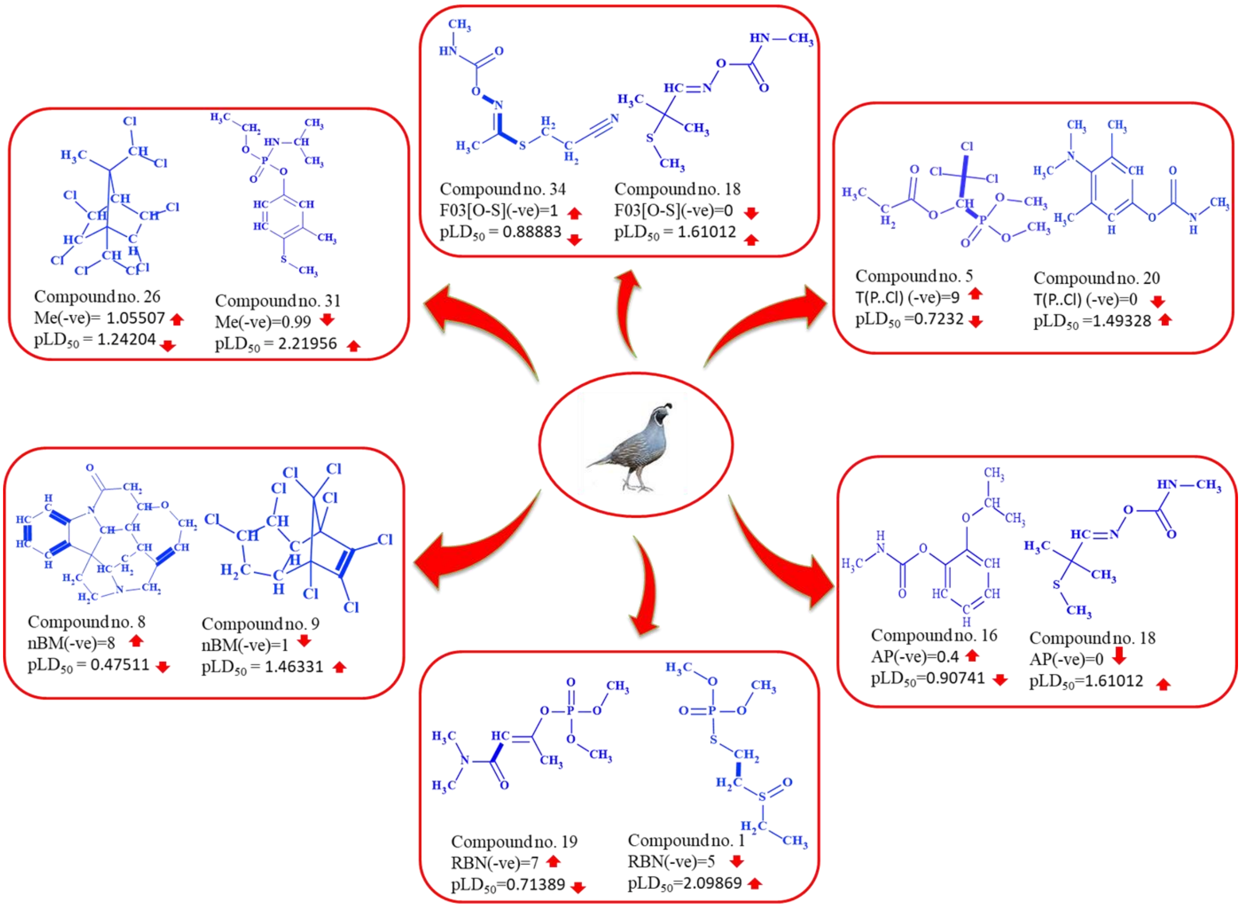


Fig. 51. Depiction of negatively contributed descriptors toward toxicity against California quail.

4.2.8 Screening of prepared external dataset

The PPDB database was screened through the developed models using the PRI tool [136]. After screening with the developed models, it was found that major pesticides are within the applicability domain and with good prediction quality. The screened pesticides are enlisted and categorized in decreasing order of their respective predicted toxicity value. The top 10 and least 10 pesticides according to their predicted values are listed in **Table 19**. Further assessment of the chosen pesticides indicated that all of the expected toxicity coincides with prior experimental values except Fentin chloride, which assures the model's applicability as well as reliability.

Table 19. Top 10 and least 10 toxic screened pesticides from Pesticide Properties DataBase (PPDB).

Top 10 highly toxic pesticides screened from Pesticide Properties Database (PPDB).	
Names of pesticides	Safety and hazards
Buminafos	Acute toxic
Cadusafos	Highly toxic
Hexylthiofos	High toxic (Cramer class-iii)

Sulfotep	Acute toxic
Tetradifon	Acute toxic, Environmental Hazard
Tetraethyl pyrophosphate	Acute toxic, Environmental Hazard
Mipafox	Highly toxic organophosphate
Fosthiazate	Acute toxic, Environmental Hazard
Merphos	Highly toxic
IPSP	Acute toxic, Environmental Hazard
Top 10 least toxic pesticides screened from the Pesticide Properties Database (PPDB)	
Tioxazafen	Low acute toxicity
Fentin hydroxide	Moderately toxic
Clofentezine	Low acute toxicity
Thiabendazole	Low acute toxicity
Fentin chloride	Highly toxic
Diflovidazin	Low toxicity
Fuberidazole	Moderately toxic
Sulcofuron	Non-toxic
Sulcofuron-sodium	Non-toxic
Sulphaquinoxaline	No ecotoxicity data

4.3 Study 3

4.3.1 Assessment of PLS-based QSTR model

In this current work, a PLS-based QSTR model has been constructed with 4 latent variables. Rigorous validation was performed for the assessment of the generated model's performance using various statistical parameters such as determination coefficient (R^2), Leave-one-out cross-validated correlation coefficient (Q^2_{LOO}), and external correlation coefficient (Q^2_{F1}, Q^2_{F2}). The calculated R^2 value of the generated model for the studied dataset crosses the threshold value (0.6) and the cross-validated correlation coefficient (Q^2_{LOO}), Q^2_{F1}, and Q^2_{F2} crosses the acceptable threshold value of 0.5. These validation parameters suggest pretty good predictability of the developed QSAR model.

$$\boldsymbol{pLD_{50}} = 2.99045 + 0.16995 \times \mathrm{nDB} + 0.16876 \times \mathrm{nCt} + 0.66741 \times \mathrm{nArOCON} + 0.16741 \times \text{C-006} + 0.82808 \times \text{H-054} + 1.00087 + \mathrm{B01[O\text{-}P]} + 0.84222 \times \mathrm{B06[S\text{-}P]} + 1.63402 \times \mathrm{B06[O\text{-}F]} + 0.31054 \times \mathrm{B07[C\text{-}N]} - 0.57278 \times \mathrm{B09[C\text{-}N]}$$

$n_{training}$ =210, R^2 =0.636, Q^2_{LOO} =0.601, $r^2m_{(LOO)}$ =0.467, LV=4, $MAE_{(LOO)}$ =0.432,
n_{test} =70, Q^2_{F1}=0.603, Q^2_{F2}=0.558, $r^2_{m(test)}$ =0.389.

4.3.2. Assessment of generated q-RASTR model

We generated q-RASAR models for raising the external predictivity of the developed PLS-based model. Thus, we calculated similarity and error-based read-across derived descriptors and clubbed them with structural and physicochemical features before model development. This combined pool of descriptors encompasses both RA-based similarity and chemical structure attribute-related information. The ultimate feature selection for q-RASAR model construction has been performed from the combined pool of descriptors followed by the selection of the best combination using the best subset selection based on the MAE, cross-validated correlation coefficient (Q^2_{LOO}), and R^2. Finally, PLS-based regression has been used to develop the q-RASAR model with four latent variables, which are depicted as follows;

$$pLD_{50} = 0.60329 + 1.74735 \times \text{Eta\_betaS\_A} + 0.86077 \times \text{H-054} + 0.44739 \times \text{B01[O-P]} + 0.19007 \times \text{B07[C-N]} - 0.13052 \times \text{F04[N-O]} + 0.046 \times \text{F04[O-O]} + 0.49346 \times \text{RA function(LK)} + 0.8103 \times \text{SE(LK)} + 1.50415 \times \text{gm*SD Similarity} - 0.02972 \times \text{sm2(LK)}$$

$n_{training}$=210, R^2 =0.657, Q^2_{LOO} =0.630, $r^2_{m(LOO)}$ =0.501, LV=4, $MAE_{(LOO)}$ =0.421, n_{test} =70, Q^2_{F1}=0.678, Q^2_{F2}=0.642, $r^2_{m(test)}$ =0.520

The generated q-RASAR model was verified by various internal and external validation for it's reliability, robustness, and predictability. A visual depiction of the correlation between observed and predicted toxicity values is represented in the scatter plot. A graphical representation of the relationship between observed and estimated toxicity values is depicted in the scattered plot (**Fig. 52**).

Table 20. Statistical parameters of developed PLS-based QSTR and q-RASTR models.

Model	Internal validation metrics						External validation metrics				
	LV	R^2	Q^2_{Loo}	$\overline{r^2_{m(Loo)}}$	$\Delta rm^2_{(Loo)}$	MAE (train)	Q^2_{F1}	Q^2_{F2}	$\overline{rm^2(test)}$	Δrm^2	MAE (test)
PLS based QSTR	4	0.636	0.601	0.467	0.250	0.432	0.603	0.558	0.389	0.304	0.460
PLS based q-RASTR	4	0.657	0.630	0.501	0.238	0.402	0.678	0.642	0.520	0.232	0.410

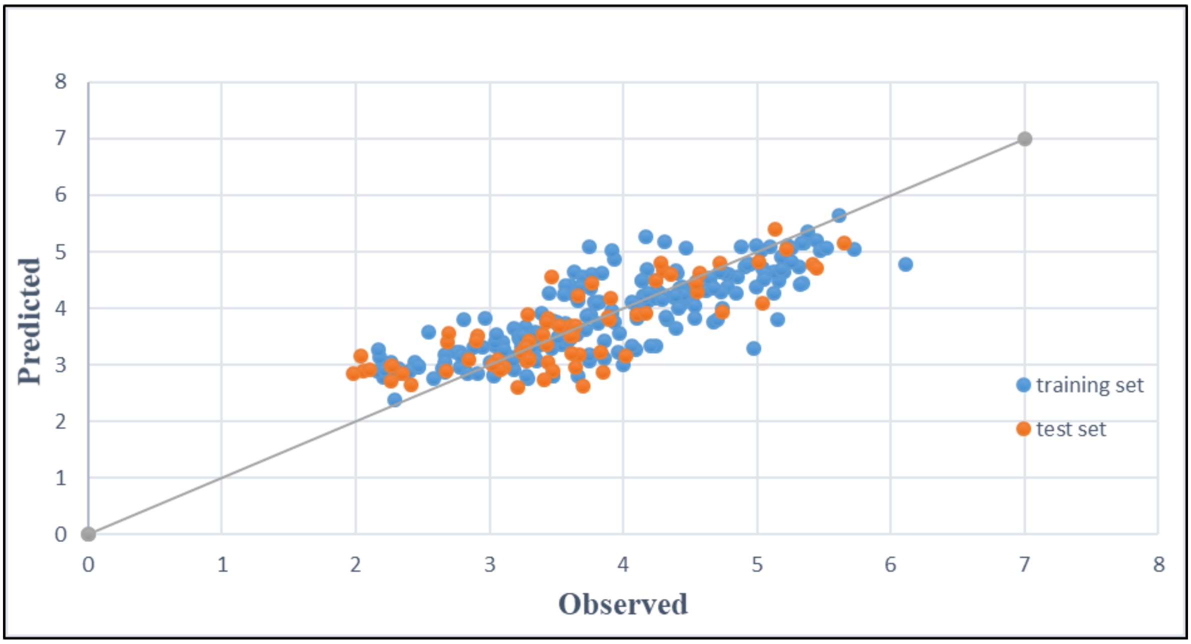


Fig. 52. Scatter plot of the constructed models.

4.3.2.1. Regression coefficient plot

The regression coefficient plot illustrates whether the descriptor of the generated model has a positive or negative impact on the modelled toxicity [28]. Here, molecular descriptors such as Eta_betaS_A, H-054, B01[O-P], B07[C-N], F04[O-O] and RASAR descriptors such as RA function(LK), SE(LK) and gm*SD Similarity has positive contribution towards the model. On the other hand, a single 2D descriptor namely F04[N-O], and a RASAR descriptor i.e. sm2(LK) has a negative contribution towards toxicity.

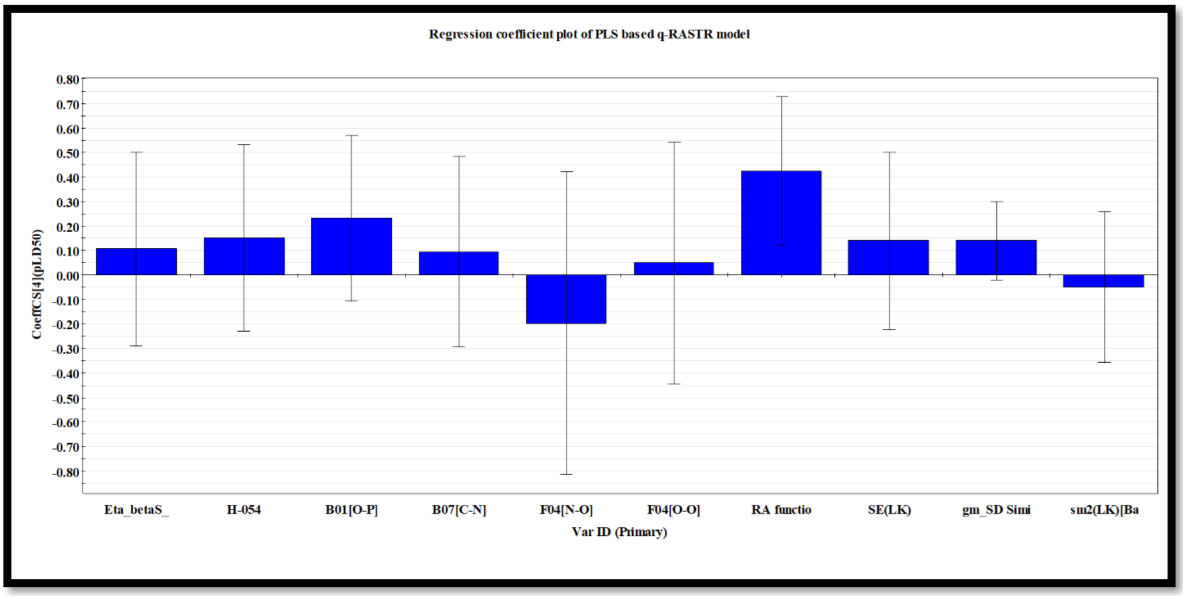


Fig. 53. Regression coefficient plot of the constructed models.

4.3.2.2 Variable importance plot (VIP)

The relative importance of the modeled descriptors about the toxicity is demonstrated by the variable importance plot [111]. Generally, VIP is a column plot where modeled descriptors are represented on the X-axis from left to right descendingly along with the VIP score of individual descriptors plotted on the Y-axis. A descriptor having a VIP value of more than 1 is considered statistically significant. According to the VIP plot, the contributing descriptor's relative importance is arranged in the following order: RA function > B01[O-P] > gm*SD Similarity > Eta_betaS_A > sm2(LK) > SE(LK) > F04[N-O] > H-054 > B07[C-N] > F04[O-O].

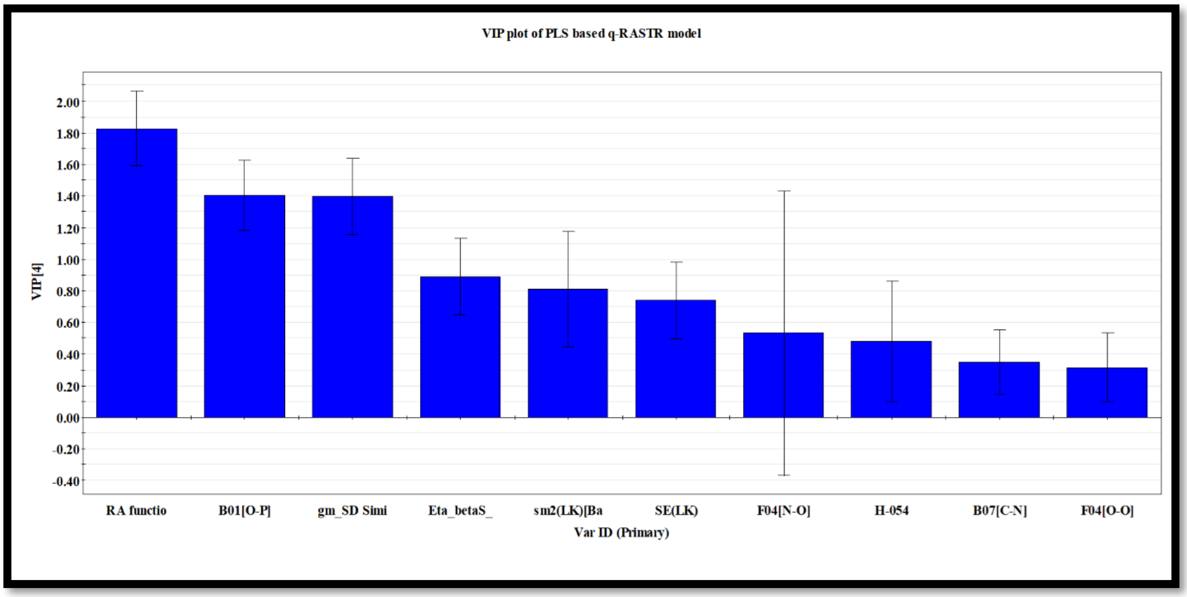


Fig. 54. Variable importance plot (VIP) of the constructed models.

4.3.2.3 Loading plot

The loading plot of the constructed q-RASTR model has been represented by SIMCA-P software. This plot demonstrates the correspondence between the descriptor variable and the response variable with the contribution of the descriptors to the toxicity. The distance of the X-variable from the origin demonstrates the importance of the descriptor. According to the generated loading plot, it was found that the RA function which is a Read-across derived similarity-based RASTR descriptor was present far from the origin and considered as the most influential descriptor for the model toxicity.

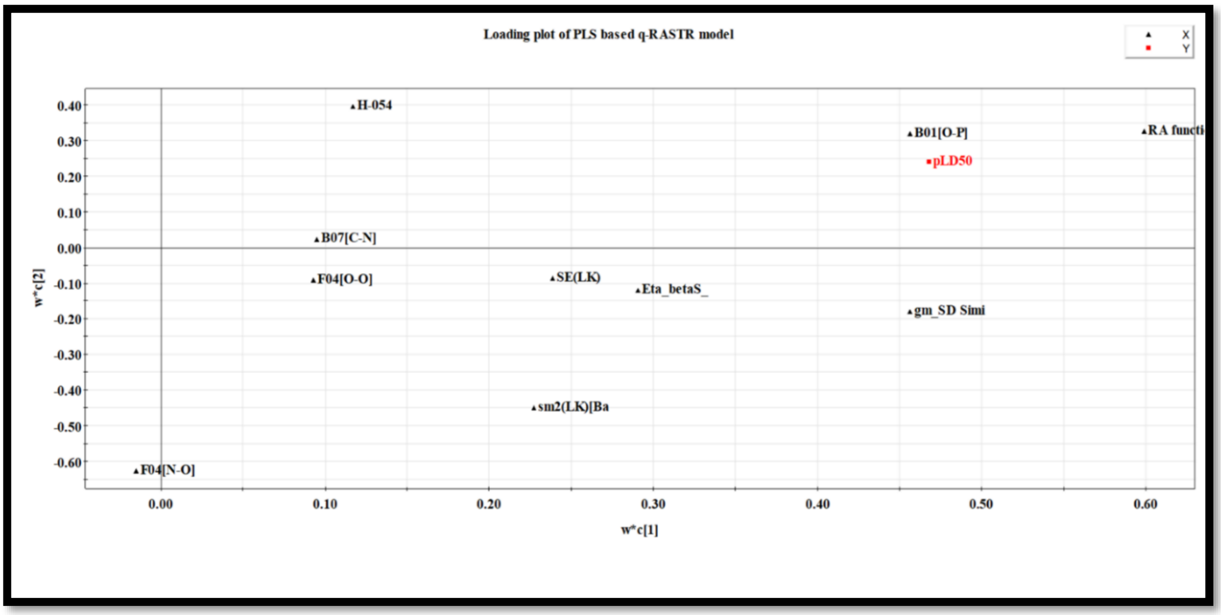


Fig. 55. Loading plot of the constructed models**.**

4.3.2.4 Score plot

The distribution of compounds in the space of latent variable is defined by the obtained scores [160]. Based on a scoring function, a score plot shows where the chemical compounds are located in theoretical chemical space. We found that only 3 compounds are situated outside the ellipse, which demonstrates the model's robustness.

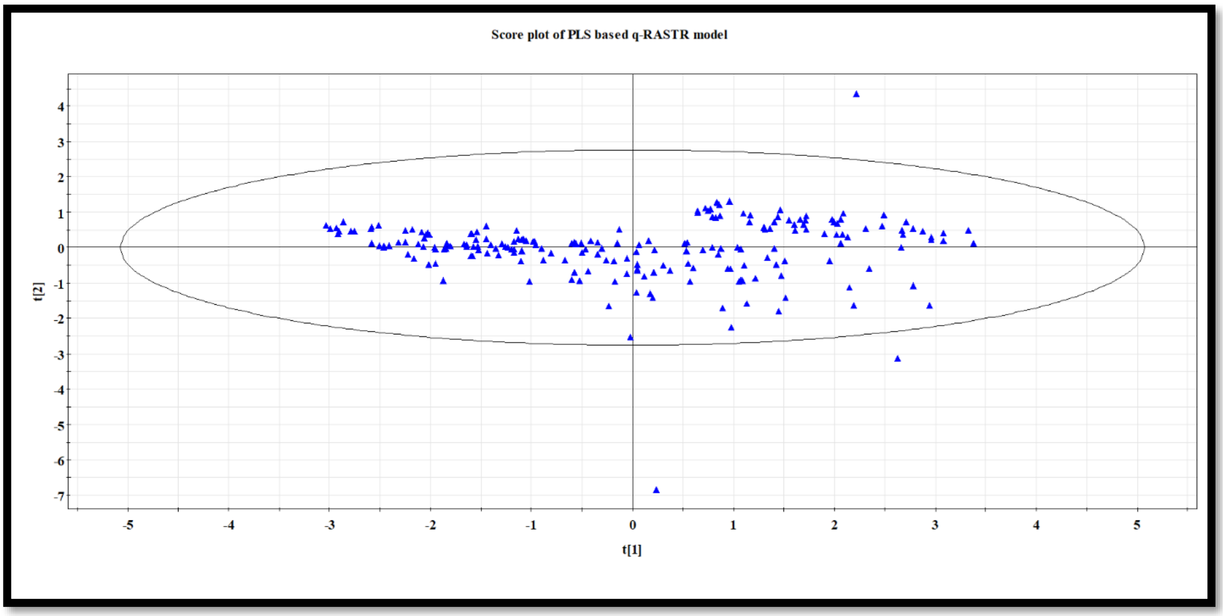


Fig. 56. Score plot of the constructed models.

Table 21. Descriptors and their contribution to the generated q-RASTR model.

Descriptors	Definition	Type	Type of contribution toward toxicity
Eta_betaS_A	Eta Sigma average VEM count	Electro topochemical atom index descriptors	Positive (+ve)
H-054	H attached to C0(sp3) with 3X attached to next C	Atom-centred fragments	Positive (+ve)
B01[O-P]	Presence/absence of O – P at topological distance 1	2D Atom Pairs descriptor	Positive (+ve)
B07[C-N]	Presence/absence of C – N at topological distance 7	2D Atom Pairs descriptor	Positive (+ve)
F04[N-O]	Frequency of N – O at topological distance 4	2D atom pairs descriptor	Negative (-ve)
F04[O-O]	Frequency of O – O at topological distance 4	2D atom pairs descriptor	Positive (+ve)
RA function(LK)	A read-across-obtained prediction function utilizing the Laplacian kernel function similarity-based algorithm	RASAR descriptor	Positive (+ve)
SE(LK)	The weighted standard error relates to the response values of neighboring source compounds	RASAR descriptor	Positive (+ve)
gm*SD Similarity	gm × standard deviation of close source compounds where gm is a concordance measure	RASAR descriptor	Positive (+ve)
sm2(LK)	Similarity coefficient	RASAR descriptor	Negative (-ve)

3.3 Possible mechanical interpretation of the modeled descriptor

Eta_betaS_A

Eta_betaS_A is a descriptor that belongs to the group of extended topochemical atom indices (ETA) descriptors. This descriptor is defined as the summation of β values for all the sigma bonds relative to the vertices number, where β denoted as valence electron mobile (VEM) counts. This can be represented as; Σβ's = Σβs/Nv, where Σβs is the total VEM count and Nv is denoted as the number of vertices [161]. This descriptor represents the electron richness relative to the molecular bulkiness. The electron richness (higher electronegative atoms) in any compound leads to increase the overall electronegativity and increases the production of reactive oxygen species which ultimately cause the death of respective animals [162]. Thus, we can assume that the presence of this feature makes the compound more toxic as evidenced by compound no. 38 (Triethylenemelamine) (pLD_{50}= 4.852, Eta_betaS_A = 0.85) and vice versa as demonstrated by compound no. 294 (Prochloraz) (pLD_{50}= 2.805, Eta_betaS_A = 0.673).

H-054

This is an atom-centered fragment descriptor that denotes the number of hydrogen attached to sp^3 hybridized carbon bound to three electronegative atoms. It is a simple molecular descriptor that expresses the electronegative characteristics of the respective compounds. As we discussed earlier enhancing the electronegativity of any chemical may cause the production of reactive oxygen species (ROS) which is very fatal to the species. Therefore, the existence of this descriptor makes the compound more toxic as demonstrated in compound no. 119 (Isobenzan) (H-054=2, pLD_{50} = 5.614), and the absence of this descriptor makes the compound relatively safer by decreasing toxicity as depicted in compound no. 31 (H-054=1, pLD_{50} = 4.649).

B01[O-P]

B01[O-P] is a 2D atom pair descriptor that characterizes as existence or absence of O and P atoms at topological distance 1. This descriptor represents the presence of two electronegative atoms (oxygen and phosphorous) which raises the overall electronegativity of the compound and resulting oxidative stress leads to the demise of the reference species [62]. Therefore, high electronegativity in a molecule makes the compound more toxic as evidenced in compound no. **177** (B01[O-P]= 1, pLD_{50} = 5.032) and vice versa in case of compound no. **206** (B01[O-P] = 0, pLD_{50} = 2.322).

B07[C-N]

B07[C–N] characterizes the presence of carbon-nitrogen fragments with topological distance 7 in the carbon skeleton of the compound. The presence of this fragment is responsible for higher toxicity due to the presence of electronegative heteroatom such as nitrogen may form hydrogen bonds and electron donor-acceptor (EDA) complexes with the DNA of respective species. Consequently, the stability of the double helix DNA structure may hindered and make the compound more toxic [155]. For example, compound no. **132** has more toxicity value (pLD_{50} = 5.182) as the numerical value of this descriptor is high (B07[C-N] = 1) and on the other hand compound no. **75** has low toxicity value (pLD_{50} = 2.910) as it has low descriptor value (B07[C-N] = 0).

F04[O-O]

F04[O-O] is a 2Datom pair descriptor representing the frequency of 2 oxygen atoms with topological distance 4. This descriptor contributed positively to the toxicity and increased the toxicity by increasing the numerical value of the respective descriptor as it contains highly electronegative atoms such as oxygen. Therefore, the overall electronegativity of the compound has increased which leads to a rise in the toxicity of the compound by forming ROS (reactive oxygen species) [130] as discussed earlier. For example, compound no. **262** has a high toxicity value (pLD_{50} = 4.536,F04[O-O] = 3). On the contrary, compound no. **6** has relatively lower toxicity value (pLD_{50} = 2.443) with descriptor value (F04[O-O] = 1).

F04[N-O]

F04[N-O] is a 2D atom pair descriptor which is defined as the frequency of nitrogen and oxygen atoms at topological distance 4. This descriptor contributed negatively towards the model response, which suggests that the presence of higher number of this fragment will reduce the toxicity as evidenced by compound 298 (F04[N-O] = 14, pLD_{50} = 2.286) and vice bersa in case of compound 284 (F04[N-O] = 4, pLD_{50} = 4.992). The presence of nitrogen and oxygen makes the compound hydrophilic by making hydrogen bonds. Hydrophilic compounds are less toxic.

RA function(LK)

The descriptor RA function(LK), a RASAR descriptor, that acts like latent variables, represent various molecular features by providing a comprehensive understanding of the compound's properties [163]. It can be observed that this descriptor shows a positive contribution towards the model response, which means the toxicity value increases with an increase in descriptor value. For example, compound **224** has a high toxicity value (pLD_{50} = 5.723) with a high numerical value of descriptor (RA function (LK) = 5.061), while compound **148** has a low toxicity value (pLD_{50} = 2.208) as the corresponding low numerical value of descriptor (RA function (LK) = 2.75).

SE(LK)

This is a read-across derived RASAR descriptor encoded as a weighted standard error of the nearby source compound's response value. This descriptor has positive impacts towards the response value. Thus, the toxicity of compounds elevates with elevating this read-across derived RASAR descriptor as shown in compound **276** (SE(LK) = 1.209, pLD_{50} = 5.255) and vice versa in case of compound **207** (SE(LK) = 0.604, pLD_{50} = 3.914).

Sm2(LK)

S_m^2(LK) is a novel similarity coefficient introduced by Banerjee and Roy [164] which found chemical compounds showing abnormal prediction (Activity cliffs). This is directly related to the difference between positive average similarity and negative average similarity [165] as shown in the following formula:

$$S_m^2 = \frac{PosAvgSim - NegAvgSim}{Avg.Sim} \qquad \textbf{3.1}$$

Here, the S_m^2(LK) coefficients are negatively correlated to the response value for the generated model, which indicates that the higher the coefficient value lower the toxicity as shown in compound 18 (sm2(LK) = 1.478, pLD_{50} = 3.562) and vice versa in case of compound 131 (sm2(LK) = 0.521) has high response value (pLD_{50} = 5.178). .

gm*SD Similarity

gm*SD Similarity is an another RASAR descriptor that shows a positive contribution to the model toxicity. It is the product of the g_m (concordance measure) with the nearest compound's standard deviation of similarity values. It's positive contribution can be demonstrated by compound no. **44** with descriptor value (gm*SD Similarity = 0.187) and response value (pLD_{50} = 5.330) and vice versa in case of compound no. **174** having descriptor value (gm*SD Similarity = 0.016), response value (pLD_{50} = 3.592).

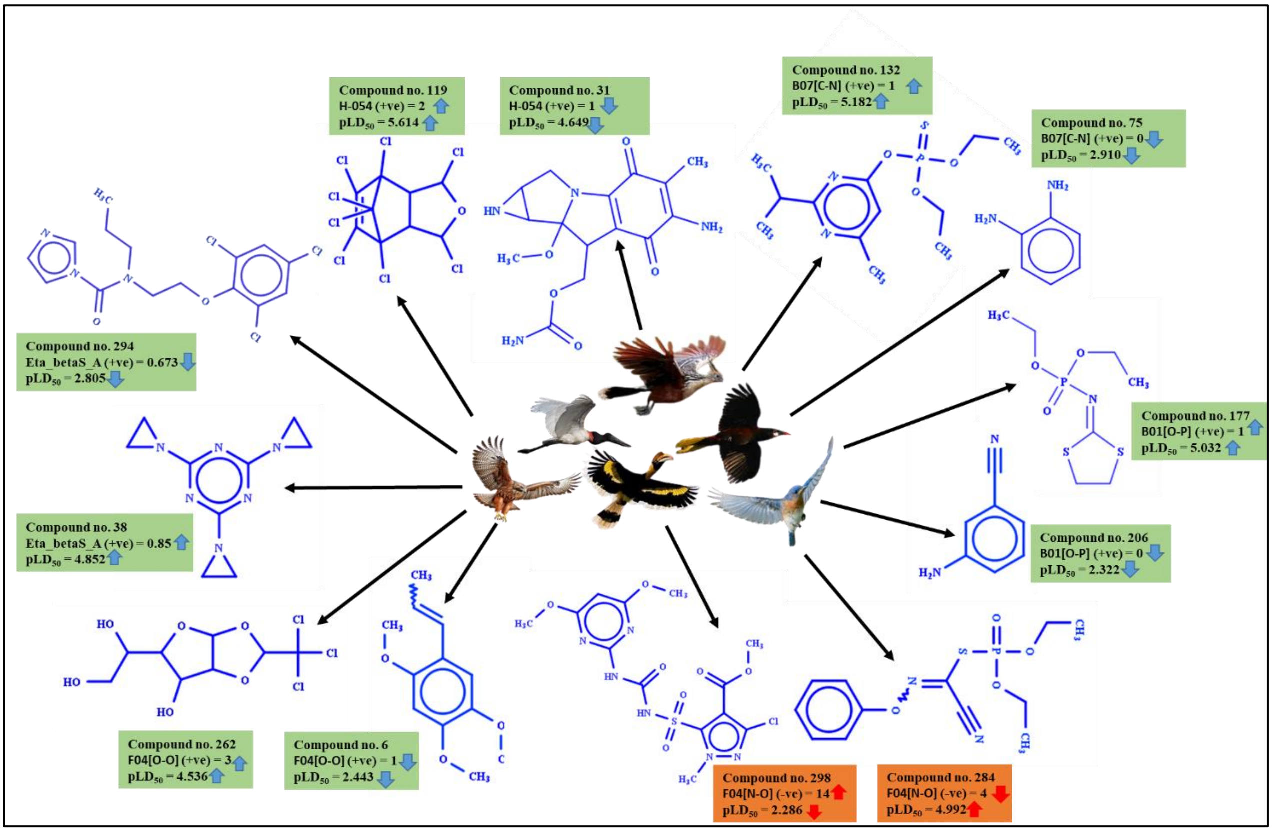


Fig. 57. Mechanistic interpretation of 2D descriptors.

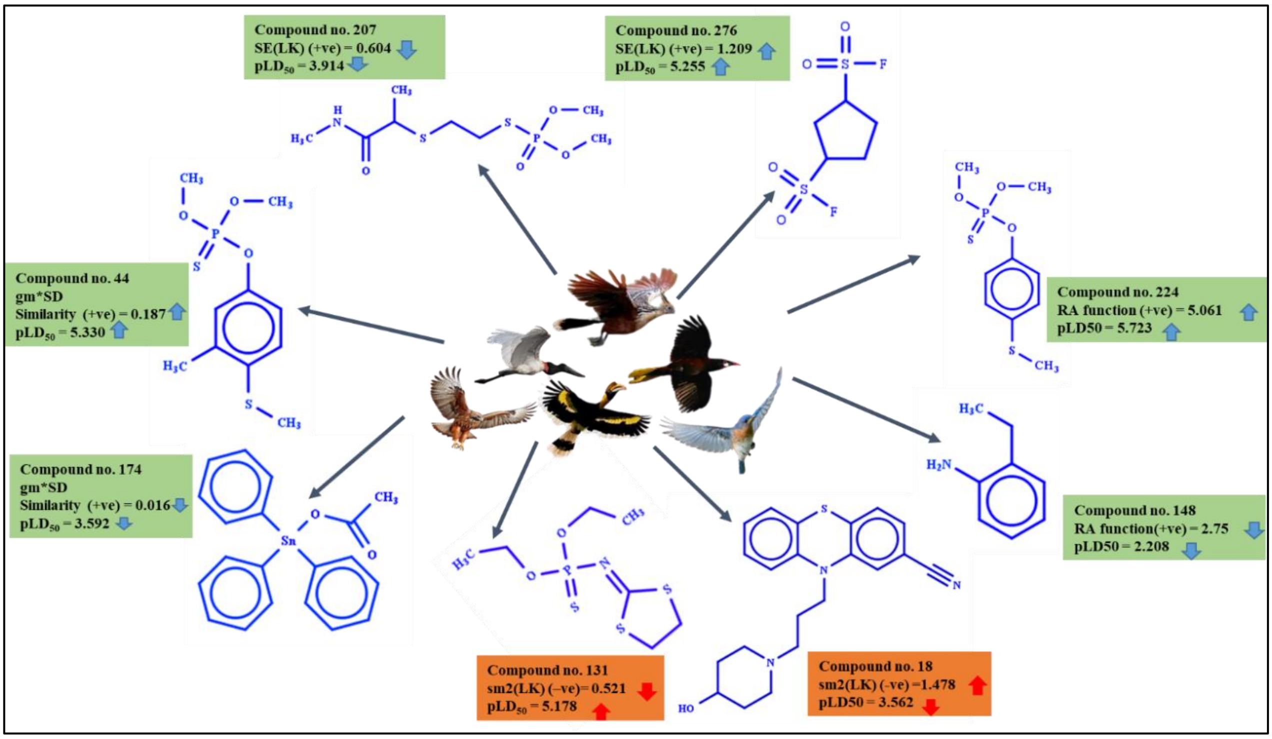


Fig. 58. Mechanistic interpretation of RASAR descriptors.

3.4 Screening of external dataset through the developed model

The developed model has been deployed for the screening of the prepared PPDB dataset by using the PRI tool. The majority of the compounds are within the applicability domain and have good prediction accuracy. The screened compounds are arranged according to their predicted toxicity values in decreasing order. Based on their estimated toxicity value, the top 10 and least 10 toxic compounds are enlisted in **Table 22**.

Table 22.Top 10 highly & least toxic pesticides screened from Pesticide Properties Database (PPDB).

Names of pesticides	Safety and hazards
Top 10 highly toxic pesticides screened from the Pesticide Properties Database (PPDB)	
Amiprofos-methyl	Highly toxic
Methocrotophos	Acutely toxic
Dicrotophos	Highly toxic
Monocrotophos	Very high acutely toxic
Fensulfothion	Highly toxic
Phosnichlor	Moderately toxic

Butathiofos	Highly toxic
Pyraclofos	Highly toxic
Dioxathion	Highly toxic
Dufulin	Low toxic
Top 10 least toxic pesticides screened from the Pesticide Properties Database (PPDB)	
Copper oxychloride	Low to Moderate toxicity
Dimethyl disulfide	Moderately toxic to Birds
Aluminium phosphide	Moderately toxicity
Chlorine dioxide	Moderate toxic
Lime sulphur	Low toxic to earthworms and honeybees
Ammonium thiocyanate	Non-toxic to aquatic invertebrates
1,1,1-acetonitrile	Low to aquatic organisms
Formaldehyde	Low toxic
Metosulam	Low toxic to birds
Methyl isobutyl ketone	Low toxic to rat

CHAPTER - 5

Conclusion

5. Conclusions

The success of any research work is determined by the results and conclusions obtained from the studies, which may reveal previously unknown or undiscovered scientific explanations. These findings may further lead to the development of better understanding and deep knowledge in the specific area in which the studies were performed. Computational chemistry including computer-aided drug design, molecular modeling, and virtual screening techniques are emerging as cost-effective and time-saving methods for introducing new chemicals into the market.

In response to the ethical concerns and the need for more sustainable practices, *in-silico* modeling emerges as a viable alternative to traditional in vivo and in vitro experimentation on living organisms. Through predictive modeling and computational simulations, we can effectively evaluate the behavior of organic chemicals, shedding light on their potential risks and impacts. This approach not only enhances our understanding of chemical interactions but also contributes to the development of ethical and environmentally conscious practices in the chemical industry.

However, there are a lot of limitations to the conventional methods of toxicity assessment. These limitations include ethical concerns related to animal experimentation, significant time and financial investments and the inherent scarcity of comprehensive experimental data. In this regards, the developed QSTR and q-RASTR models emerged to be an effective and adaptable tool for the efficient prediction of toxicity. We can overcome these constraints and deliver precise and quick evaluations of drug toxicity by utilizing data-driven insights and computational modeling.

Our work includes several approaches and combines a wide range of concepts, which come together to produce the results and explanations we offer. Due to their extensive uses in a variety of industries, including food, medicine, cosmetics, and agriculture, organic chemicals highlight the significance of thorough risk assessment. One notable challenge is the substantial data gap that exists concerning the toxic effects of certain chemicals and their largely unidentified environmental consequences.

5.1 Study 1: Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across

In summary, this study employs a range of chemometric tools to predict pesticide toxicity for four different avian species. The research focuses on creating robust and easily interpretable QSTR models based on OECD principles. The study's statistical validation parameters consistently demonstrate the strength and reliability of the constructed PLS models. External validation metrics, employing the read-across algorithm, show slightly superior performance in predicting toxicity, except for the mallard duck dataset. Furthermore, this research develops regression-based models, surpassing previous studies in terms of the dataset's size and the variety of avian species examined. The findings highlight the significance of electronegativity, molecular weight, imide count, lipophilicity, and steric effects in avian toxicity. Notably, the presence of C-P fragments at topological distance 4 and electronegative groups intensifies toxicity, while features like branching and hydrogen bond acceptor characteristics reduce the toxicity.

The validation of the predicted toxicity of the screened compounds by experimental data demonstrated the reliability and feasibility of applying the developed models for screening pesticides, offering valuable support to researchers striving to design eco-friendly and safe chemical pesticides. They effectively bridge gaps in toxicity data and simplify the evaluation of novel pesticides for various bird species. Moreover, these models significantly reduce the time, resources, costs, and the need for animal testing, aligning with the principles of reduction, refinement, and replacement (RRR) in research practices.

5.2. Study 2: First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail

The current work has proposed PLS-based QSTR models against a new avian species (California quail). These models were validated by using various statistical metrics to establish the model’s reliability and robustness. External validation parameters were intensified by using intelligent consensus prediction. Possible mechanistic interpretations of the associated descriptors were demonstrated and we found that the presence of phosphate groups, electronegativity, a high percentage of carbon, unsaturation, mean Sanderson electronegativity, lipophilicity, aromatic proportion, and flexibility have significant effects on toxicity. Particularly, the presence of C-P fragments at exact topological distances, electronegativity, carbon chain length, and degree of unsaturation elevate the toxicity. At the same time, features like the number of rotatable bonds, and aromatic proportion diminish the toxicity. The developed models were employed on a prepared external database which was originally collected from the pesticide properties database (PPDB) and predicted their toxicity to demonstrate the reliability and feasibility of the developed models. After the screening, it can be concluded that the reported models can efficiently fill the gaps in toxicity data and may enlighten researchers and synthetic chemists to design novel, safer, and eco-friendly compounds to reduce the possibility of toxicity specifically toward avian species.

5.3. Study 3: Chemometric-based exploration of the toxicological significance of diverse chemical toxicants in wild birds with an application of the q-RASTR approach

In the current study, a predictive q-RASTR model for the toxicity assessment against the wild birds of diverse chemical toxicants has been generated. Various readily interpreted 2D descriptors were employed for the construction of the final q-RASTR model. The constructed model has been validated externally and internally to verify quantitatively as well as qualitatively by using different statistical parameters to establish reliability, robustness, and predictability. Possible mechanistic interpretation of the modeled descriptors demonstrated that features such as H-054, B01[O-P], B07[C-N], and F04[O-O], majorly indicate the presence of electronegative atoms/hetero atoms along with the presence of extended topochemical atom indices (ETA) descriptors contribute positively towards the toxicity of the chemical toxicants. Furthermore, a few similarity-based RASAR descriptors like RA function, SE(LK), gm*SD Similarity, and Sm2(LK) also contributed positively towards the toxicity of wild birds. Here in this study, the undetermined toxicity values of the PPDB (Pesticide Properties Database) were predicted by deploying the generated q-RASTR model which shows the reliability and feasibility of the developed model. The generated model can be useful for assessing the toxicity of any unknown compound by overcoming limitations such as animal testing, time-consuming, and cost. The generated model and obtained structural information might enlighten the researchers to synthesize safer and environmentally safe chemicals as well as bridge the data gaps in the toxicity database.

REFERENCE

REFERENCES

1. Dearden, J.C., 2002. Prediction of environmental toxicity and fate using quantitative structure-activity relationships (QSARs). Journal of the Brazilian Chemical Society, 13, pp.754-762.

2. Combs, A.B. and Acosta Jr, D., 2007. An introduction to toxicology and its methodologies. Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals, pp.1-20.

3. Mackay, D., Hubbarde, J. and Webster, E., 2003. The role of QSARs and fate models in chemical hazard and risk assessment: Paper prepared for Quantitative Structure-Activity Relationships (QSAR) Proceedings of the QSAR 2002 Conference, Ottawa May 2002. Qsar & Combinatorial Science, 22(1), pp.106-112.

4. Cronin, M.T., 2004. Predicting chemical toxicity and fate. CRC press.

5. Aktar, W., Sengupta, D. and Chowdhury, A., 2009. Impact of pesticides use in agriculture: their benefits and hazards. Interdisciplinary toxicology, 2(1), pp.1-12.

6. Van Der Werf, H.M., 1996. Assessing the impact of pesticides on the environment. Agriculture, Ecosystems & Environment, 60(2-3), pp.81-96.

7. Carvalho, F.P., 2017. Pesticides, environment, and food safety. Food and energy security, 6(2), pp.48-60.

8. Tabur, M.A. and Ayvaz, Y., 2010, June. Ecological importance of birds. In Second International Symposium on Sustainable Development Conference (pp. 560-565).

9. Mitra, A., Chatterjee, C. and Mandal, F.B., 2011. Synthetic chemical pesticides and their effects on birds. Res J Environ Toxicol, 5(2), pp.81-96.

10. Bishop, C.A., 1998. Health of tree swallows (Tachycineta bicolor) nesting in pesticide-sprayed apple orchards in Ontario, Canada. I. Immunological parameters. Journal of Toxicology and Environmental Health Part A, 55(8), pp.531-559.

11. Hill, E.F., 1992. Avian toxicology of anticholinesterases. Clinical and experimental toxicology of organophosphates and carbamates, pp.272-294.

12. Dieter, C.D., Flake, L.D. and Duffy, W.G., 1995. Effects of phorate on ducklings in northern prairie wetlands. The Journal of wildlife management, pp.498-505.

13. Grove, R.A., Buhler, D.R., Henny, C.J. and Drew, A.D., 1998. Declining ring-necked pheasants in the Klamath Basin, California: I. Insecticide exposure. Ecotoxicology, 7, pp.305-312.

14. Hawkes, A.W., Brewer, L.W., Hooper, M.J., Kendall, R.J. and Hobson, J.F., 1996. Survival and cover-seeking response of northern bobwhites and mourning doves dosed with aldicarb. Environmental Toxicology and Chemistry: An International Journal, 15(9), pp.1538-1543.

15. Kupper, J., Baumgartner, M., Bacciarini, L.N., Hoop, R., Kupferschmidt, H. and Naegeli, H., 2007. Carbofuran poisoning in mallard ducks. Schweizer archiv fur tierheilkunde, 149(11), pp.517-520.

16. Helfrich, L.A., Weigmann, D.L., Hipkins, P.A. and Stinson, E.R., 2009. Pesticides and aquatic animals: a guide to reducing impacts on aquatic systems.

17. Relyea, R.A. and Hoverman, J.T., 2008. Interactive effects of predators and a pesticide on aquatic communities. Oikos, 117(11), pp.1647-1658.

18. Gopi Mohan, C., Gandhi, T., Garg, D. and Shinde, R., 2007. Computer-assisted methods in chemical toxicity prediction. Mini reviews in medicinal chemistry, 7(5), pp.499-507.

19. Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. and Aloy, P., 2022. Connecting chemistry and biology through molecular descriptors. Current Opinion in Chemical Biology, 66, p.102090.

20. Todeschini, R., Consonni, V. and Gramatica, P., 2009. Chemometrics in QSAR. In Comprehensive chemometrics (Vol. 4, pp. 129-172). Elsevier.

21. King, R.D. and Srinivasan, A., 1997. The discovery of indicator variables for QSAR using inductive logic programming. Journal of computer-aided molecular design, 11, pp.571-580.

22. Roy, K., 2004. Topological descriptors in drug design and modeling studies. Molecular Diversity, 8, pp.321-323.

23. Stalke, D., 2011. Meaningful structural descriptors from charge density. Chemistry–A European Journal, 17(34), pp.9264-9278.

24. De Benedetti, P.G. and Fanelli, F., 2014. Multiscale quantum chemical approaches to QSAR modeling and drug design. Drug Discovery Today, 19(12), pp.1921-1927.

25. Katritzky, A. R., & Gordeeva, E. V. (1993). Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. Journal of chemical information and computer sciences, 33(6), 835-857.

26. Karelson, M., Lobanov, V.S. and Katritzky, A.R., 1996. Quantum-chemical descriptors in QSAR/QSPR studies. Chemical reviews, 96(3), pp.1027-1044.

27. Lü, J. X., Shen, Q., Jiang, J. H., Shen, G. L., & Yu, R. Q. (2004). QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. Journal of pharmaceutical and biomedical analysis, 35(4), 679-687.

28. Melagraki, G., Afantitis, A., Sarimveis, H., Igglessi-Markopoulou, O., & Supuran, C. T. (2006). QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors.

29. Rhyu, K. B., Patel, H. C., & Hopfinger, A. J. (1995). A 3D-QSAR study of anticoccidial triazines using molecular shape analysis. Journal of chemical information and computer sciences, 35(4), 771-778

30. Collantes, E.R., Tong, W., Welsh, W.J. and Zielinski, W.L., 1996. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. Analytical chemistry, 68(13), pp.2038-2043.

31. Hahn, M., & Rogers, D. (1995). Receptor surface models. 2. Application to quantitative structure-activity relationships studies. Journal of Medicinal Chemistry, 38(12), 2091-2102.

32. Chen, X., Dang, L., Yang, H., Huang, X. and Yu, X., 2020. Machine learning-based prediction of toxicity of organic compounds towards fathead minnow. RSC advances, 10(59), pp.36174-36180.

33. Roy, K., Kar, S., Das, R. N., Roy, K., Kar, S., & Das, R. N. (2015). Statistical methods in QSAR/QSPR. A Primer on QSAR/QSPR Modeling: Fundamental Concepts, 37-59.

34. Bridges Jr, C.C., 1966. Hierarchical cluster analysis. Psychological reports, 18(3), pp.851-854.

35. Burkardt, J., 2009. K-means clustering. Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics.

36. Despagne, F. and Massart, D.L., 1998. Neural networks in multivariate calibration. Analyst, 123(11), pp.157R-178R.

37. Zheng, W. and Tropsha, A., 2000. Novel variable selection quantitative structure− property relationship approach based on the k-nearest-neighbor principle. Journal of chemical information and computer sciences, 40(1), pp.185-194.

38. Berggren, E., Amcoff, P., Benigni, R., Blackburn, K., Carney, E., Cronin, M., Deluyker, H., Gautier, F., Judson, R.S., Kass, G.E. and Keller, D., 2015. Chemical safety assessment using read-across: assessing the use of novel testing methods to strengthen the evidence base for decision making. Environmental health perspectives, 123(12), pp.1232-1240.

39. Patlewicz, G. and Fitzpatrick, J.M., 2016. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. Chemical research in toxicology, 29(4), pp.438-451.

40. Banerjee, A. and Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Molecular Diversity, 26(5), pp.2847-2862.

41. Banerjee, A., Kar, S., Gajewicz-Skretna, A. and Roy, K., 2022. q-RASAR Modeling of Cytotoxicity of TiO2-based Multi-component Nanomaterials.

42. Banerjee, A., Chatterjee, M., De, P. and Roy, K., 2022. Quantitative predictions from chemical

read-across and their confidence measures. Chemometrics and Intelligent Laboratory Systems, 227, p.104613.

43. Chatterjee, M. and Roy, K., 2023. “Data fusion” quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary antibiotic mixtures toward three bacterial species. Journal of Hazardous Materials, p.132129.

44. Roy, K. and Banerjee, A., 2024. Tools, Applications, and Case Studies (q-RA and q-RASAR). In q-RASAR: A Path to Predictive Cheminformatics (pp. 51-88). Cham: Springer Nature Switzerland.

45. Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 194(4824), pp.178-180.

46. Keshavarz, M.H., Shirazi, Z. and Sheikhabadi, P.K., 2021. Risk assessment of organic aromatic compounds to Tetrahymena pyriformis in environmental protection by a simple QSAR model. Process Safety and Environmental Protection, 150, pp.137-147.

47. Roy, K. and Roy, P.P., 2009. QSAR of cytochrome inhibitors. Expert Opinion on Drug Metabolism & Toxicology, 5(10), pp.1245-1266.

48. 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.

49. Roy, K., Ambure, P. and Kar, S., 2018. How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? ACS omega, 3(9), pp.11392-11406.

50. Roy, K., Mitra, I., Ojha, P.K., Kar, S., Das, R.N. and Kabir, H., 2012. Introduction of rm2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. Chemometrics and Intelligent Laboratory Systems, 118, pp.200-210.

51. Hawkins, D.M., 2004. The problem of overfitting. Journal of chemical information and computer sciences, 44(1), pp.1-12.

52. Todeschini, R., Consonni, V. and Gramatica, P., 2009. Chemometrics in QSAR. In Comprehensive chemometrics (Vol. 4, pp. 129-172). Elsevier.

53. Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR & combinatorial science, 26(5), pp.694-701

54. Roy, K., Kar, S. and Das, R.N., 2015. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.

55. Mitra, I., Saha, A. and Roy, K., 2010. Exploring quantitative structure–activity relationship

studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. Molecular Simulation, 36(13), pp.1067-1079.

56. Yun, Y.H., Wu, D.M., Li, G.Y., Zhang, Q.Y., Yang, X., Li, Q.F., Cao, D.S. and Xu, Q.S., 2017. A strategy on the definition of applicability domain of model based on population analysis. Chemometrics and Intelligent Laboratory Systems, 170, pp.77-83.

57. Mazzatorta, P., Cronin, M.T. and Benfenati, E., 2006. A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. QSAR & Combinatorial Science, 25(7), pp.616-628.

58. Basant, N., Gupta, S. and Singh, K.P., 2015. Predicting toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. Journal of Chemical Information and Modeling, 55(7), pp.1337-1348.

59. Halder, A.K., Saha, A. and Jha, T., 2017. Predictive quantitative structure toxicity relationship study on avian toxicity of some diverse agrochemical pesticides by Monte Carlo method: QSTR on pesticides. International Journal of Quantitative Structure-Property Relationships (IJQSPR), 2(1), pp.19-34.

60. Kar, S. and Leszczynski, J., 2020. Is intraspecies QSTR model answer to toxicity data gap filling: Ecotoxicity modeling of chemicals to avian species. Science of The Total Environment, 738, p.139858.

61. Banjare, P., Singh, J. and Roy, P.P., 2021. Predictive classification-based QSTR models for toxicity study of diverse pesticides on multiple avian species. Environmental Science and Pollution Research, 28(14), pp.17992-18003.

62. Mukherjee, R.K., Kumar, V. and Roy, K., 2021. Ecotoxicological QSTR and QSTTR modeling for the prediction of acute oral toxicity of pesticides against multiple avian species. Environmental Science & Technology, 56(1), pp.335-348.

63. Podder, T., Kumar, A., Bhattacharjee, A. and Ojha, P.K., 2023. Exploring regression-based QSTR and i-QSTR modeling for ecotoxicity prediction of diverse pesticides on multiple avian species. Environmental Science: Advances, 2(10), pp.1399-1422.

64. Singh, K.P., Gupta, S., Basant, N. and Mohan, D., 2014. QSTR modeling for qualitative and quantitative toxicity predictions of diverse chemical pesticides in honey bee for regulatory purposes. Chemical Research in Toxicology, 27(9), pp.1504-1515.

65. Mostafalou, S. and Abdollahi, M., 2013. Pesticides and human chronic diseases: evidences, mechanisms, and perspectives. Toxicology and applied pharmacology, 268(2), pp.157-177.

66. Saxena, A.K., Devillers, J., Bhunia, S.S. and Bro, E., 2015. Modelling inhibition of avian aromatase by azole pesticides. SAR and QSAR in Environmental Research, 26(7-9), pp.757-782.

67. OECD; Test No. 223: Avian Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, Effects on Biotic Systems; OECD Publishing: Paris, France, 2010.

68. Nicolotti, O.; Benfenati, E.; Carotti, A.; Gadaleta, D.; Gissi, A.; Mangiatordi, G. F.; Novellino, E. REACH and in silico methods: an attractive opportunity for medicinal chemists. Drug Discovery Today 2014, 19, 1757−1768.

69. Speck-Planche, A., Guilarte-Montero, L., Yera-Bueno, R., Rojas-Vargas, J.A., García-López, A., Uriarte, E. and Molina-Pérez, E., 2011. Rational design of new agrochemical fungicides using substructural descriptors. Pest management science, 67(4), pp.438-445. https://doi.org/10.1002/ps.2082

70. Speck-Planche, A., Natalia Dias Soeiro Cordeiro, M., Guilarte-Montero, L. and Yera-Bueno, R., 2011. Current computational approaches towards the rational design of new insecticidal agents. Current Computer-Aided Drug Design, 7(4), pp.304-314.

71. Speck-Planche, A., Kleandrova, V.V., Luan, F. and Cordeiro, M.N.D., 2012. Predicting multiple ecotoxicological profiles in agrochemical fungicides: a multi-species chemoinformatic approach. Ecotoxicology and environmental safety, 80, pp.308-313.

72. Speck-Planche, A., 2020. Multi-scale QSAR approach for simultaneous modeling of ecotoxic effects of pesticides. Ecotoxicological QSARs, pp.639-660.

73. Jiang, J., Wang, R., Wang, M., Gao, K., Nguyen, D.D., and Wei, G.W., 2020. Boosting tree-assisted multitask deep learning for small scientific datasets. Journal of chemical information and modeling, 60(3), pp.1235-1244.

74. Jain, S., Siramshetty, V.B., Alves, V.M., Muratov, E.N., Kleinstreuer, N., Tropsha, A., Nicklaus, M.C., Simeonov, A. and Zakharov, A.V., 2021. Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. Journal of chemical information and modeling, 61(2), pp.653-663

75. Halder, A.K., Moura, A.S. and Cordeiro, M.N.D., 2023. Predicting the ecotoxicity of endocrine disruptive chemicals: Multitasking in silico approaches towards global models. Science of The Total Environment, 889, p.164337.

76. Samanipour, S., O’Brien, J.W., Reid, M.J., Thomas, K.V. and Praetorius, A., 2022. From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization. Environmental Science & Technology, 57(46), pp.17950-17958.

77. Tugcu, G., Saçan, M.T., Vracko, M., Novic, M. and Minovski, N., 2012. QSTR modelling of the acute toxicity of pharmaceuticals to fish. SAR and QSAR in Environmental Research, 23(3-4), pp.297-310.

78. Song, I.S., Cha, J.Y. and Lee, S.K., 2011. Prediction and analysis of acute fish toxicity of pesticides to the rainbow trout using 2D-QSAR. Analytical Science and Technology, 24(6), pp.544-555.

79. Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L. and Moussa, C.S., 2016. A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. Journal of hazardous materials, 303, pp.28-40.

80. Wang, L.L., Ding, J.J., Pan, L., Fu, L., Tian, J.H., Cao, D.S., Jiang, H. and Ding, X.Q., 2021. Quantitative structure-toxicity relationship model for acute toxicity of organophosphates via multiple administration routes in rats and mice. Journal of Hazardous Materials, 401, p.123724.

81. Zhang, C., Cheng, F., Sun, L., Zhuang, S., Li, W., Liu, G., Lee, P.W. and Tang, Y., 2015. In silico prediction of chemical toxicity on avian species using chemical category approaches. Chemosphere, 122,pp.280-287.

82. Kumar, A., Ojha, P.K. and Roy, K., 2023. QSAR modeling of chronic rat toxicity of diverse organic chemicals. Computational Toxicology, 26, p.100270.

83. Mitra, A., Chatterjee, C. and Mandal, F.B., 2011. Synthetic chemical pesticides and their effects on birds. Res J Environ Toxicol, 5(2), pp.81-96.

84. Mariyappan, M., Rajendran, M., Velu, S., Johnson, A.D., Dinesh, G.K., Solaimuthu, K., Kaliyappan, M. and Sankar, M., 2023. Ecological Role and Ecosystem Services of Birds: A Review. International Journal of Environment and Climate Change, 13(6), pp.76-87

85. Pandey, S.K., Ojha, P.K. and Roy, K., 2020. Exploring QSAR models for assessment of acute fish toxicity of environmental transformation products of pesticides (ETPPs). Chemosphere, 252, p.126508.

86. Kumar, A., Ojha, P.K. and Roy, K., 2024. Safer and greener chemicals for the aquatic ecosystem: Chemometric modeling of the prolonged and chronic aquatic toxicity of chemicals on Oryzias latipes. Aquatic Toxicology, p.106985.

87. Bhhatarai, B. and Gramatica, P., 2010. Per-and polyfluoro toxicity (LC_{50} inhalation) study in rat and mouse using QSAR modeling. Chemical research in toxicology, 23(3), pp.528-539.

88. Wilson, L., Martin, P.A., Elliott, J.E., Mineau, P. and Cheng, K.M., 2001. Exposure of California quail to organophosphorus insecticides in apple orchards in the Okanagan Valley, British Columbia.

Ecotoxicology, 10, pp.79-90.

89. Roy, J., Ghosh, S., Ojha, P.K. and Roy, K., 2019. Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). Environmental Science: Nano, 6(1), pp.224-247.

90. Kuo, D.T., Rattner, B.A., Marteinson, S.C., Letcher, R., Fernie, K.J., Treu, G., Deutsch, M., Johnson, M.S., Deglin, S. and Embry, M., 2022. A critical review of bioaccumulation and biotransformation of organic chemicals in birds. Reviews of Environmental Contamination and Toxicology, 260(1), p.6

91. Newton, I. and Bogan, J., 1974. Organochlorine residues, eggshell thinning and hatching success in British sparrowhawks. Nature, 249(5457), pp.582-583.

92. Movalli, P., Krone, O., Osborn, D. and Pain, D., 2018. Monitoring contaminants, emerging infectious diseases, and environmental change with raptors, and links to human health. Bird Study, 65(sup1), pp.S96-S109.

93. Coeurdassier, M., Poirson, C., PAUL, J.P., Rieffel, D., Michelat, D., Reymond, D., Legay, P., Giraudoux, P. and Scheifler, R., 2012. The diet of migrant Red Kites Milvus milvus during a Water Vole Arvicola terrestris outbreak in eastern France and the associated risk of secondary poisoning by the rodenticide bromadiolone. Ibis, 154(1), pp.136-146.

94. Macdonald, J.W., 1963. Mortality in wild birds. Bird Study, 10(2), pp.91-108.

95. Cronin, M.T., Jaworska, J.S., Walker, J.D., Comber, M.H., Watts, C.D. and Worth, A.P., 2003. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environmental health perspectives, 111(10), pp.1391-1401.

96. Kumar, A., Ojha, P.K. and Ro y, K., 2024. First report on pesticide sub-chronic and chronic toxicities against dogs using QSAR and chemical read-across. SAR and QSAR in Environmental Research, 35(3), pp.241-263.

97. Basant, N., Gupta, S. and Singh, K.P., 2016. QSAR modeling for predicting reproductive toxicity of chemicals in rats for regulatory purposes. Toxicology research, 5(4), pp.1029-1038.

98. Ghosh, S., Kar, S. and Leszczynski, J., 2020. Chemometric modeling of the ecotoxicity of industrial chemicals to an avian species Anas Platyrhynchos. International Journal of Quantitative Structure-Property Relationships (IJQSPR), 5(2), pp.1-16.

99. Banerjee, A. and Roy, K., 2024. How to correctly develop q-RASAR models for predictive cheminformatics. Expert Opinion on Drug Discovery, pp.1-6.

100. Banerjee, A., Kar, S., Pore, S. and Roy, K., 2023. Efficient predictions of cytotoxicity of TiO2-

based multi-component nanoparticles using a machine learning-based q-RASAR approach. Nanotoxicology, 17(1), pp.78-93.

101. Ghosh, S., Chatterjee, M. and Roy, K., 2023. Quantitative Read-across structure-activity relationship (q-RASAR): A new approach methodology to model aquatic toxicity of organic pesticides against different fish species. Aquatic Toxicology, 265, p.106776. Wu, L., Yan, B., Han, J., Li, R., Xiao, J., He, S. and Bo, X., 2023.

102. Zhang, C., Cheng, F., Sun, L., Zhuang, S., Li, W., Liu, G., Lee, P.W. and Tang, Y., 2015. In silico prediction of chemical toxicity on avian species using chemical category approaches. Chemosphere, 122,pp.280-287.

103. O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. Journal of cheminformatics, 3(1), pp.1-14.

104. Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs, pp.801-820.

105. Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H. and Tropsha, A., 2012. Does rational selection of training and test sets improve the outcome of QSAR modeling? Journal of chemical information and modeling, 52(10), pp.2570-2578.

106. Ambure, P., Aher, R.B., Gajewicz, A., Puzyn, T. and Roy, K., 2015. “NanoBRIDGES” software: open access tools to perform QSAR and nano-QSAR modeling. Chemometrics and Intelligent Laboratory Systems, 147, pp.1-13.

107. Park, H.S. and Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert systems with applications, 36(2), pp.3336-3341.

108. Jillella, G.K., Ojha, P.K. and Roy, K., 2021. Application of QSAR for the identification of key molecular fragments and reliable predictions of effects of textile dyes on growth rate and biomass values of Raphidocelis subcapitata. Aquatic Toxicology, 238, p.105925.

109. Roy, P.P., Leonard, J.T. and Roy, K., 2008. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems, 90(1), pp.31-42.

110. Ojha, P.K. and Roy, K., 2011. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. Chemometrics and Intelligent Laboratory Systems, 109(2), pp.146-161.

111. Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), pp.109-130.

112. Todeschini, R., Ballabio, D. and Grisoni, F., 2016. Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models. Journal of Chemical Information and Modeling, 56(10), pp.1905-1913.

113. Banerjee, A., De, P., Kumar, V., Kar, S. and Roy, K., 2022. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across. Chemosphere, 309, p.136579.

114. Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A. and Roy, K., 2022. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environmental Science: Nano, 9(1), pp.189-203.

115. 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.

116. SIMCA-P, U.M.E.T.R.I.C.S., 2002. 10.0, info@ umetrics. com: www. umetrics. com, Umea.

117. Paul, R., Chatterjee, M. and Roy, K., 2022. First report on soil ecotoxicity prediction against Folsomia candida using intelligent consensus predictions and chemical read-across. Environmental Science and Pollution Research, 29(58), pp.88302-88317.

118. Dillon, W.R. and Goldstein, M., 1984. Multivariate analysis: Methods and applications. New York (NY): Wiley, 1984.

119. Morales Helguera, A., Perez Gonzalez, M., Dias Soeiro Cordeiro, M.N. and Cabrera Perez, M.A., 2008. Quantitative Structure− Carcinogenicity Relationship for Detecting Structural Alerts in Nitroso Compounds: Species, Rat; Sex, Female; Route of Administration, Gavage. Chemical research in toxicology, 21(3), pp.633-642.

120. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A. and Štajdohar, M., 2013. Orange: data mining toolbox in Python. the Journal of machine Learning research, 14(1), pp.2349-2353.

121. Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

122. Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), pp.442-451.

123. Kumar, A., Ojha, P.K. and Roy, K., 2024. The first report on the assessment of maximum acceptable daily intake (MADI) of pesticides for humans using intelligent consensus predictions. Environmental Science: Processes & Impacts, 26(5), pp.870-881.

124. Todeschini, R. and Consonni, V., 2008. Handbook of molecular descriptors. John Wiley &

Sons.

125. Cai, J., Luo, J., Wang, S. and Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300, pp.70-79.

126. Wild, D.J., 2005. MINITAB release 14.

127. Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N. and Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. Journal of chemical information and modeling, 52(2), pp.396-408.

128. 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.

129. Nath, A. and Roy, K., 2022. Chemometric modeling of acute toxicity of diverse aromatic compounds against Rana japonica. Toxicology in Vitro, 83, p.105427.

130. Roy, J. and Roy, K., 2021. Assessment of toxicity of metal oxide and hydroxide nanoparticles using the QSAR modeling approach. Environmental Science: Nano, 8(11), pp.3395-3407.

131. Kar, S. and Roy, K., 2010. First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals. Chemosphere, 81(6),pp.738-747.

132. Topliss, J.G. and Edwards, R.P., 1979. Chance factors in studies of quantitative structure-activity relationships. Journal of Medicinal Chemistry, 22(10), pp.1238-1244.

133. Paul, R., Chatterjee, M. and Roy, K., 2022. First r]eport on soil ecotoxicity prediction against Folsomia candida using intelligent consensus predictions and chemical read-across. Environmental Science and Pollution Research, 29(58), pp.88302-88317.

134. Lewis, K., Tzilivakis, J., Green, A. and Warner, D., 2006. Pesticide Properties Database (PPDB).

135. De, P., Kar, S., Ambure, P. and Roy, K., 2022. Prediction reliability of QSAR models: an overview of various validation tools. Archives of Toxicology, 96(5), pp.1279-1295.

136. Roy, K., Ambure, P. and Kar, S., 2018. How precise are our quantitative structure–activity relationship derived predictions for new query chemicals?. ACS omega, 3(9), pp.11392-11406.

137. TOXRIC: a comprehensive database of toxicological data and benchmarks. Nucleic Acids Research, 51(D1), pp.D1432-D1445.

138. Shahlaei, M., 2013. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chemical reviews, 113(10), pp.8093-8103.

139. Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A. and Roy, K., 2022. A novel

quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environmental Science: Nano, 9(1), pp.189-203.

140. Banerjee, A. and Roy, K., 2023. Machine-learning-based similarity meets traditional QSAR: "q-RASAR" for the enhancement of the external predictivity and detection of prediction confidence outliers in an hERG toxicity dataset. Chemometrics and Intelligent Laboratory Systems, 237, p.104829.

141. Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A. and Nicolotti, O., 2016. Applicability domain for QSAR models: where theory meets reality. International journal of quantitative structure-property relationships (IJQSPR), 1(1), pp.45.

142. Wu, Z., Li, D., Meng, J. and Wang, H., 2010. Introduction to SIMCA-P and its application. Handbook of partial least squares: concepts, methods and applications, pp.757-774.

143. Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), pp.442-451.

144. Akarachantachote, N., Chadcham, S. and Saithanu, K., 2014. Cutoff threshold of variable importance in projection for variable selection. Int J Pure Appl Math, 94(3), pp.307-322.

145. Hou, T.J., Zhang, W., Xia, K., Qiao, X.B. and Xu, X.J., 2004. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. Journal of chemical information and computer sciences, 44(5), pp.1585-1600.

146. Khan, K., Roy, K. and Benfenati, E., 2019. Ecotoxicological QSAR modeling of endocrine disruptor chemicals. Journal of hazardous materials, 369, pp.707-718.

147. Kar, S., Sanderson, H., Roy, K., Benfenati, E. and Leszczynski, J., 2020. Ecotoxicological assessment of pharmaceuticals and personal care products using predictive toxicology approaches. Green Chemistry, 22(5), pp.1458-1516.

148. Roy, J. and Roy, K., 2021. Assessment of toxicity of metal oxide and hydroxide nanoparticles using the QSAR modeling approach. Environmental Science: Nano, 8(11), pp.3395-3407.

149. M. Vervloet, Modifying Phosphate toxicity in chronic kidney disease, Toxins (Basel) 11 (9) (2019 Sep 9) 522.

150. Khan, K. and Roy, K., 2019. Ecotoxicological QSAR modelling of organic chemicals against Pseudokirchneriella subcapitata using consensus predictions approach. SAR and QSAR in Environmental Research, 30(9), pp.665-681.

151. Schultz, T.W., Yarbrough, J.W. and Koss, S.K., 2006. Identification of reactive toxicants: Structure–activity relationships for amides. Cell biology and toxicology, 22, pp.339-349.

152. Krishna, J.G., Ojha, P.K., Kar, S., Roy, K. and Leszczynski, J., 2020. Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. Nano Energy, 70, p.104537.

153. Ojha, P.K. and Roy, K., 2018. Chemometric modeling of odor threshold property of diverse aroma components of wine. RSC advances, 8(9), pp.4750-4760.

154. Vervloet, M., 2019. Modifying phosphate toxicity in chronic kidney disease. Toxins, 11(9), p.522.

155. Ghosh, S., Ojha, P.K., Carnesecchi, E., Lombardo, A., Roy, K. and Benfenati, E., 2020. Exploring QSAR modeling of toxicity of chemicals on earthworm. Ecotoxicology and Environmental Safety, 190, p.110067.

156. Wang, C., Wei, Z., Wang, L., Sun, P. and Wang, Z., 2015. Assessment of bromide-based ionic liquid toxicity toward aquatic organisms and QSAR analysis. Ecotoxicology and environmental safety, 115, pp.112-118.

157. Yuan, H., Wang, Y.Y. and Cheng, Y.Y., 2007. Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. Journal of Molecular Graphics and Modelling, 26(1), pp.327-335.

158. Sparling, D.W., Day, D. and Klein, P., 1999. Acute toxicity and sub-lethal effects of white phosphorus in mute swans, Cygnus olor. Archives of environmental contamination and toxicology, 36, pp.316-322.

159. Khan, K., Khan, P.M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K. and Benfenati, E., 2019. QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors. Chemosphere, 229, pp.8-17.

160. De, P., Kar, S., Roy, K. and Leszczynski, J., 2018. Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. Environmental Science: Nano, 5(11), pp.2742-2760.]

161. Roy, K. and Das, R.N., 2017. The “ETA” Indices in QSAR/QSPR/QSTR Research. In Pharmaceutical Sciences: Breakthroughs in Research and Practice (pp. 978-1011). IGI Global.

162. Roy, J., Ojha, P.K., Carnesecchi, E., Lombardo, A., Roy, K. and Benfenati, E., 2020. First report on a classification-based QSAR model for chemical toxicity to earthworm. Journal of hazardous materials, 386, p.121660.

163. Kumar, V., Banerjee, A. and Roy, K., 2024. Breaking the Barriers: Machine-Learning-Based c-RASAR Approach for Accurate Blood–Brain Barrier Permeability Prediction. Journal of Chemical

Information and Modeling.

164. Banerjee, A. and Roy, K., 2023. Prediction-inspired intelligent training for the development of classification read-across structure-activity relationship (c-RASAR) models for organic skin sensitizers: assessment of classification error rate from novel similarity coefficients. Chemical Research in Toxicology, 36(9), pp.1518-1531.

165. Pandey, S.K. and Roy, K., 2023. Development of a read-across-derived classification model for the predictions of mutagenicity data and its comparison with traditional QSAR models and expert systems. Toxicology, 500, p.153676.

Appendix

List of publications and Reprints

List of Publications

A. Papers related to this dissertation

1. Shubha Das, **Abhisek Samal**, Ankur Kumar, Vinayak Ghosh, Supratik Kar, and Probir Kumar Ojha. "Comprehensive ecotoxicological assessment of pesticides on multiple avian species: Employing quantitative structure-toxicity relationship (QSTR) modeling and read-across." **Process Safety and Environmental Protection 188 (2024): 39-52. (I.F- 6.9), DOI:** https://doi.org/10.1016/j.psep.2024.05.095
2. **Abhisek Samal**, Shubha Das, and Probir Kumar Ojha, "First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail. Journal: Environmental Toxicology and Pharmacology (2024). **(Under Review)**. **(I.F-4.2)**

B. Papers not related to this dissertation

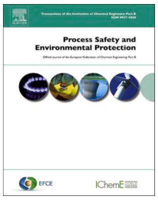
1. Shubha Das, **Abhisek Samal**, and Probir Kumar Ojha. "Chemometrics-driven prediction and prioritization of diverse pesticides on chickens for addressing hazardous effects on public health." Journal of Hazardous Materials 471 (2024): 134326.
2. Pabitra Samanta, Prodipta Bhattacharyya, **Abhisek Samal**, Ankur Kumar, Arnab Bhattacharjee, Probir Kumar Ojha. "Ecotoxicological risk assessment of active pharmaceutical ingredients (APIs) against different aquatic species leveraging intelligent consensus prediction and i-QSTTR modeling" Journal: Journal of Hazardous Materials.

Process Safety and Environmental Protection 188 (2024) 39–52

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.journals.elsevier.com/process-safety-and-environmental-protection



Comprehensive ecotoxicological assessment of pesticides on multiple avian species: Employing quantitative structure-toxicity relationship (QSTR) modeling and read-across

Shubha Das[a,1], Abhisek Samal[a,1], Ankur Kumar[a], Vinayak Ghosh[a], Supratik Kar[b], Probir Kumar Ojha[a,*]

[a] *Drug Discovery and Development Laboratory (DDD Lab), Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India*
[b] *Chemometrics and Molecular Modeling Laboratory, Department of Chemistry and Physics, Kean University, 1000 Morris Avenue, Union, NJ 07083, USA*

ARTICLE INFO

Keywords:
Pesticides
Avian species
Acute toxicity
2D-QSTR
Read-across

ABSTRACT

The rapid increase in the use of pesticides is driven by the growing demand in the agricultural sector. However, the widespread application of these pesticides and their inherent toxicity have significant repercussions on the ecosystem, particularly impacting animal and bird species. In this present study, we have developed four 2D quantitative structure-toxicity relationships (QSTRs) models for four different avian species using the largest number of available experimental data points to date employing the partial least squares (PLS) algorithm. Furthermore, we have also performed the read-across algorithm to improve the test set results. Based on the information derived from the models, it was found that hydrophilic characteristics, the presence of molecular branching and thio imide groups impact negatively to the pesticide toxicity, while the presence of phosphate group, presence of halogens viz. chlorine and bromine atoms, presence of hetero atoms, high molecular weight, presence of bridgehead atoms, presence of secondary aliphatic amide and fragments like RCONHR escalates avian toxicity. The developed QSTR models were further employed to predict the Pesticide Properties DataBase (PPDB) for all four avian species as a measure of data gap-filling and risk assessment. Thus, the developed models can be utilized for eco-toxicological data-gap filling, prediction of toxicity of untested pesticides as well as the development of novel and safe environmental-friendly pesticides.

1. Introduction

Pesticides encompass a wide range of chemicals, which are typically employed to control or kill pests viz. insects, rodents, fungi, weeds, etc. for effective crop management. The use of pesticides has increased significantly in recent decades, particularly in agriculturally dependent developing countries (Singh et al., 2014). Due to the inherent characteristics, a significant portion of the applied dose continues to remain as remnants on crops and fields (Basant et al., 2015). As a result, large amounts of pesticides have been found in crops, vegetation, and further edible products causing exposure to both animals and humans. According to reports, prolonged exposure to these substances can harm a person's nervous, endocrine, reproductive, immunological, cardiovascular, renal, and respiratory systems (Mostafalou and Abdollahi, 2013). In light of the aforementioned, various regulatory authorities have emphasized the need for the toxicity evaluation of both new and existing pesticides. The avian toxicity tests are essential for regulatory approval and licensing of the active ingredients of pesticides. Aves are significant for ecology and have a huge contribution to biodiversity by performing pollination of plants, rodent control, seed dispersal, and spreading

Abbreviations: BQ, Bobwhite quail; JQ, Japanese quail; MD, Mallard duck; RNP, Ring-necked pheasant; 2D descriptors, Two-dimensional descriptors; 2D-QSTR, Two dimensional- quantitative structure- toxicity relationship; AD, Applicability domain; DModx, Distance to model X; GA, Genetic algorithm; Log[LC_{50}], logarithmic value of the 50% Lethal concentration LC_{50}; OECD, The Organisation for Economic Cooperation and Development; PLS, partial least square; QSAR, Quantitative structure-activity relationship; QSTR, Quantitative structure-toxicity relationship; REACH, Registration, Evaluation, Authorisation, and Restrictions of Chemicals; RMSEP, root mean square error of prediction; EPA, Environmental Protection Agency; PPDB, Pesticide Properties DataBase.

[*] Corresponding author.
E-mail address: probirojha@yahoo.co.in (P.K. Ojha).

[1] These authors contributed equally

https://doi.org/10.1016/j.psep.2024.05.095
Received 20 October 2023; Received in revised form 16 April 2024; Accepted 21 May 2024
Available online 22 May 2024

nutrients (Mukherjee et al., 2021). According to today's scenario, one in every eight bird species faces extinction (Saxena et al., 2015). Therefore, birds are used as a model organism to evaluate toxicity. Oral toxicity testing is important for determining avian species' toxicological significance. Northern bobwhite quail *(Colinus virginianus)* [BQ], Japanese quail *(Coturnix japonica)* [JQ], ring-necked pheasant (*Phasianus colchicus*) [RNP], and mallard duck (*Anas platyrhynchos*) [MD] are the major test species as per OECD norms (OECD, 2010). The validated wet-lab techniques for the evaluation of compound toxicity towards avians are expensive, unethical, and require a significant amount of time and effort. So the relevant regulatory bodies encourage the employment of potential alternative strategies to achieve the objective. Regulatory agencies like the Environmental Protection Agency (EPA), European Food Safety Authority (EFSA), Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), and European Chemicals Bureau (ECB), have emphasized the potential of computational tools like QSTR, read-across, and alternative approaches for investigating the inherent characteristics of chemicals within the realm of toxicokinetics (Nicolotti et al., 2014; Pandey et al., 2020). Some alternatives in silico-based approaches were reported previously that offer significant improvements over single-output models for regulatory purposes (Speck-Planche et al., 2011; Speck-Planche et al., 2011, 2012; Speck-Planche, 2020; Jiang et al., 2020; Jain et al., 2021). Speck-Planche et al. (Speck-Planche et al., 2011) reported the discriminant model based on substructural descriptors for the rational design of new agrochemical fungicides. Speck-Planche et al. (Speck-Planche et al., 2011) also worked on new in-silico methods for the rational design of new insecticidal agents. Speck-Planche et al. (Speck-Planche et al., 2012) further reported the multi-species chemoinformatic methods for assessing the various ecotoxicological profiles in agrochemical fungicides. Speck-Planche et al. (Speck-Planche, 2020) also published a work regarding multi-scale QSAR methodology for simultaneous ecotoxicological modeling of pesticides. Jiang et al. (Jiang et al., 2020) worked on boosting tree-assisted multitask deep learning methods for small scientific datasets. A consensus multitask deep learning method was used to model multispecies acute toxic effects by Jain et al (Jain et al., 2021). Even other alternative modeling approaches based on machine learning (ML) tools that have demonstrated significant advancements, particularly in handling nonlinearity aspects and improving predictions were also reported earlier (Jiang et al., 2020; Jain et al., 2021; Halder et al., 2023; Samanipour et al., 2022). Halder et al. (Halder et al., 2023) reported the global models employing in-silico methods for predicting the ecotoxicity of endocrine disruptive chemicals. Samanipour et al. (Samanipour et al., 2022) worked on alternative methods for chemical prioritization using molecular descriptors and intrinsic fish toxicity of chemicals.

These *in silico* techniques examine significant structural features that are essential for predicting the biological activity, toxicity, and other characteristics of untested substances. Several research teams published *in silico* predictions of acute oral toxicity in various species, including rats, mice, and fish (Banjare et al., 2021; Song et al., 2011; Hamadache et al., 2016; Wang et al., 2021). But in the case of avian oral toxicity, very few in-silico reports are available (Basant et al., 2015; Mukherjee et al., 2021; Saxena et al., 2015; Banjare et al., 2021; Zhang et al., 2015; Podder et al., 2023).

Herein, we developed QSTR models to interpret the major structural and physicochemical features responsible for their toxicity followed by assessing the toxicity of external datasets in BQ, JQ, RNP, and MD avian species following the OECD guidelines strictly (OECD, 2007). Alternative tools, such as read-across, are widely used for hazard assessment to fill the data gaps. The read-across-based predictions assume that a molecule with an unreported experimental endpoint value should have a value similar to molecules that are structurally and/or biologically similar to the query molecule. So, we have conducted the read-across predictions to improve the test set results. The main motive for choosing the regression-based QSTR approach over others (e.g.: regarding its effectiveness, coping with chemical heterogeneity, and several different species) (Karpov et al., 2020; Jaganathan et al., 2022) was to develop a linear relationship between the descriptors and the defined endpoints (pLC_{50}) to identify the important features responsible for toxicity towards avian species (BQ, JQ, RNP, and MD) as well as data-gap filling. Classification-based approaches also excel in handling similar challenges, and both methodologies come with distinct advantages and disadvantages. For example, classification models are typically more robust to outliers and data errors than regression models. This is because classification models only focus on the categorical relationship between the input and output variables rather than the exact numerical relationship. On the other hand, regression models can identify the most important features or predictors driving the outcome variable. This information can be used to inform decision-making and guide further investigations. Sometimes, it may be beneficial to convert a classification problem into a regression problem or vice versa. By doing so, one can gain additional insights into the data and improve the accuracy of our predictions. Nevertheless, the decision to convert a problem type should be based on the specific problem at hand and the characteristics of the data. Additionally, we have also developed classification models as well as employed two different ML algorithms namely SVM, and RF to evaluate their effectiveness in model construction and prediction. The present work aimed to design a logical method to assess pesticide toxicity towards avians. Furthermore, screening of the Pesticide Properties DataBase (PPDB) was conducted to evaluate the avian toxicity following the prediction reliability assessment of the QSTR models by the PRI (prediction reliability indicator) tool (http://teqip.jdvu.ac.in/QSAR_Tools/) as a measure of data gaps filling and risk assessment (Kumar et al., 2023). The robustness, reproducibility, and predictivity of QSTR models were thoroughly validated using globally accepted statistical parameters.

2. Methods and materials

2.1. Preparation of dataset & curation

Here, we developed models using datasets with toxicity endpoint ($LC_{50;}$ defined as the lethal concentration in 50% population) for toxicity prediction in multiple avian species collected from literature (Zhang et al., 2015) which was originally collected from the EPA, Ecotox database (http://cfpub.epa.gov/ecotox/). In this study; 112 pesticides for RNP, 117 pesticides for JQ, 556 pesticides for BQ, and 564 pesticides for MD were taken for the development of the model. The toxicity endpoint values ranges from 0.082-4.957 in BQ, 0.162–4.968 in JQ, 0.27–4.67 in MD, and 0.162–4.857 in RNP. The two-dimensional structures of the pesticides were sketched using Marvin Sketch 5.5.0.1 (https://chemaxon.com) software with the addition of explicit hydrogen atoms as well as proper aromatization. The conversion of structure file formats was carried out using Open Babel v.2.3.2 (O'Boyle et al., 2011). Knime workflow (https://www.knime.com/cheminformatics-extensions) was employed for data curation which removes unwanted salts and duplicate compounds. Toxicity in an avian species characterized as an endpoint value (LC_{50}) was converted to millimolar (mM) concentration followed by converting to a negative logarithmic scale, pLC_{50}, for easy interpretation. Some compounds were omitted from the datasets due to high residual values.

2.2. Descriptor calculation & data pre-treatment

Descriptors are the numerical presentation in which we correlate the chemical structure with any physiochemical property/biological activity/ toxicity. In this work, a total of 9 classes of descriptors were calculated utilizing AlvaDesc 2.02 (https://www.alvascience.com/alvadesc/) software (Mauri, 2020). In each dataset, the defective and inter-correlated chemical descriptors were eliminated by V-WSP1.2 (http://teqip.jdvu.ac.in/QSAR_Tools/) software with a standard deviation less than 0.0001 or correlation coefficient greater than 0.95.

2.3. Dataset division

Dataset division is crucial for QSTR model development. Normally, training set compounds are used to develop the model and test set compounds for validation. The validation set is used to assess the model performance and fine-tune the parameters of the model. It tells us how well the model is learning and adapting, allowing for adjustments and optimizations to be made to the model's parameters and hyperparameters (the latter in the case of machine learning-based models) before it is finally tested. The test data set mirrors real-world data the model has never seen before, i.e.: a separate sample of unseen data. Its primary purpose is to offer a fair and final assessment of how the model would perform when it encounters new data in a live, operational environment. This is especially critical to evaluate models effectively along with preventing overfitting (Martin et al., 2012). We performed dataset division of four datasets by using rational methods such as the Kennard stone, activity property-based, and Euclidean distance based method using Dataset Division GUI 1.2 software as well as using random division method (Martin et al., 2012; Ambure et al., 2015). We also employed modified *k*-medoid clustering by using Modified *k*-Medoid 1.3 (http://teqip.jdvu.ac.in/QSAR_Tools/) (Park and Jun, 2009). After that, the final selection of data-set division methods was done based on the statistical results. The best results come in the Kennard stone method for the MD and JQ data set, the activity property-based method for the BQ dataset, and the random division method for the RNP dataset. In this process of dataset division, the datasets are divided into 75:25 ratios of training and test sets compounds respectively (Jillella et al., 2021).

2.4. Selection of features and model building

In the case of model building, feature selection is one of the vital steps by which we can find significant descriptors to boost the interpretability and predictive ability of the model (Roy et al., 2008). Primarily, we performed stepwise regression method and genetic algorithm (GA) for feature selection (Ojha and Roy, 2011) and then we employed the regression-based partial least square (PLS) (Wold et al., 2001) method through the partial least squares v1.0 tool (http://teqip.jdvu.ac.in/QSAR_Tools/) for model building.

2.5. Validation metrics of QSTR models

A significant step in the creation of a QSTR model is statistical validation, which demonstrates its reliability and predictivity (Roy et al., 2015a). Various internal validation parameters were calculated which involve determination coefficient (R^2), leave-one-out (LOO) cross-validated correlation coefficient (Q^2_{LOO}) to judge the reliability and importance of the model. External validation parameters demonstrate the predictivity of QSTR models. The model's external validation is determined using parameters such as Q^2_{F1} and Q^2_{F2} (Todeschini et al., 2016). For both internal (Q^2_{LOO}) and external predictive parameters (Q^2_{F1},Q^2_{F2}), the approved threshold value is 0.5.

2.6. Prediction using read-across algorithm

According to the fundamental tenet of read-across, substances with similar chemical structures will also have comparable attributes and it is not utilized in the model development process (Banerjee et al., 2022). Read-across prediction is a similarity-based non-testing technique that is widely used in eco-toxicological data-gap filling. Initially, the training set of the best model was split into sub-training and sub-test sets. These sets were again used to optimize the hyperparameters through Read-Across-v3.1 (http://teqip.jdvu.ac.in/QSAR_Tools/) software. After similarity-based sorting, similarity threshold values (0−1), various distance threshold values (1−0), and the numbers of most similar training compounds (2−10) were applied. The best setting of hyperparameters obtained from sub-training and sub-test was applied to the original training and test sets for the final prediction (Chatterjee et al., 2022).

2.7. Model's applicability domain study

The applicability domain (AD) of a QSAR model has been defined as the chemical structure and response space, considered by the properties of the molecules in the training set (Roy et al., 2015a). The AD expresses the fact that QSARs are undeniably associated with restrictions in the categories of physicochemical properties, chemical structures, and mechanisms of action for which the models can generate reliable predictions. In the current study, distance to the model in X-space (DModx) has been utilized for AD estimation of constructed PLS models which rely on residuals of response and predictive variables (Roy et al., 2015b).

2.8. Y-randomization study

Y-randomization study was carried out to check the chance correlation of the QSTR models with the help of SIMCA-P software (SIMCA-P, 2002). In the Y-randomization test, the descriptor matrix X is kept constant but only the vector Y is scrambled randomly, and a new model is developed using the same set of descriptors. The original model is considered as robust if its validation metrics are better than the random models (Paul et al., 2022). The values of the R^2y_{rand} intercept and Q^2y_{rand} intercept should not be more than 0.3 and 0.05 respectively.

2.9. Analysis of parametric assumptions of the developed models

To ensure that our model is reliable we carried out some diagnostic tests to check for the existence of multicollinearity, normal distribution, and homoscedasticity (Dillon and Goldstein, 1984; Morales Helguera et al., 2008). Multicollinearity is defined as predictor variables within a regression model that are highly correlated with each other, leading to inaccurate results in regression analysis. To identify multicollinearity, we used the variation inflation factor (VIF) which is a widely used metric. If the VIF is higher than 5, multicollinearity is considered to be present (Kim, 2019). In statistical regression models, exhibiting multicollinearity can lead to misleading results. For each modeled descriptor, we found that the VIF values were very close to 1. So, it can be concluded that all the independent variables are not collinear with the dependent variable. The function values follow a multidimensional normal distribution with a mean and covariance matrix that depends on the descriptor vectors. We have plotted the normal distribution curve for each (BQ, JQ, MD, and RNP) avian species and provided in Fig. S1 of supplementary information 2. Homoscedasticity refers to the equal variance of an error in a regression model was assessed using the Breusch-Pagan test in our study. A p-value of more than 0.05 indicates the homoscedasticity of the model. In our study, the calculated p-values were not less than 0.05 (0.093–0.209) for all the developed models. Therefore, we fail to reject the null hypothesis, and the model can be considered homoscedastic. All the statistical results of homoscedasticity and multicollinearity for each model are provided in Tables S1 and S2 of supplementary information 2.

2.10. Application of other machine learning (ML) algorithms

To estimate the prediction performance of other algorithms, we have employed two different state-of-the-art ML algorithms namely support vector machine (SVM) and random forest (RF) using the Orange data mining tool (Demšar et al., 2013, Senanayake et al., 2022). The hyperparameters were adjusted to tune the model for optimal performance. The prediction qualities of the ML models were evaluated in terms of R^2, $Q^2_{Loo,}$ and MAE values.

2.11. Classification based QSTR (LDA-QSTR) model development

In the present work, we have developed a classification-based linear discriminant analysis (LDA) QSTR model from the selected set of features and evaluated its performance for its predictive ability. The model development is done using ClassificationBasedQSAR_v1.0.0 tools (available at http://teqip.jdvu.ac.in/QSAR_Tools/). The model was extensively validated based on different internal and external classification metrics (area under the ROC curve (AUC), accuracy, precision, sensitivity, F-measure, and Matthews correlation coefficient (MCC)) (Fawcett, 2006; Matthews, 1975).

2.12. Screening of the Pesticide Properties DataBase

We have collected 1903 chemical data from Pesticide Properties DataBase (PPDB) available in (http://sitem.herts.ac.uk/aeru/ppdb/). Knime curation was done to remove duplicates, inorganic salts, and mixtures using the KNIME workflow. Due to the knime curation, some compounds were removed. After the curation, the remaining 1694 compounds were used for the screening process to check the developed model's reliability. The descriptors for these molecules were calculated using the same procedure as in the QSAR modeling process. The predictions were made through the use of individual PLS-based QSTR models with the help of the PRI (Prediction Reliability Indicator) tool (http://teqip.jdvu.ac.in/QSAR_Tools/). PRI tool categorizes the predictions into three distinct groups: good (composite score 3), moderate (composite score 2), and bad (composite score 1). Additionally, the tool determines the localization of compounds inside the AD. The screened compounds were ranked based on their predicted toxicity and the twenty highest and least toxic compounds which exhibited toxicity towards all four avian species were analysed. The results were further validated extensively based on experimental data reported previously, to establish the real-world applicability of the developed final PLS-based QSTR models. Detailed discussions on the results can be found in Section 3 (Roy et al., 2018). A detailed flow diagram of this study has been given in Fig. 1.

3. Results and discussion

In this study, we have developed PLS models utilizing the toxicity of pesticides ($LogLC_{50}$) on four different avians (BQ, JQ, MD, and RNP) employing a reduced pool of chemical descriptors. The created model's quality is measured by using different internal (R^2, Q^2_{LOO},) and external (Q^2_{F1}, Q^2_{F2},) statistical parameters. The results obtained from PLS models indicated the model's robustness, reliability, and predictivity. All the metrics obtained from QSTR models are depicted in Table 1. Read-across algorithm was employed to improve the model's external predictivity. External predictivity was improved for all three datasets (BQ, JQ, RNP) except MD in read-across prediction, and results are provided in Table 2. The obtained results from the Y-randomization test were found to be R^2= -0.01, Q^2= -0.0531, (for BQ), R^2 = 0.0194, Q^2 = -0.215 (for JQ), R^2 = -0.008, Q^2 = -0.0377 (for MD), and R^2 = 0.028, Q^2 = -0.213 (for RNP) which demonstrated that the models were not formed by any chance. AD study depicted that compounds **26**, **112**, and **113** in BQ, compounds **31** and **103** in JQ, compound **468** in MD, and compound **88** in RNP from the test set are outside the AD as depicted in **Figs: S1-S4 in** supplementary information 2. The tentative reasons or characteristics that designate certain compounds as outliers in each model (above the D-critical line) is due to some structural dissimilarity. As for example, in case of the BQ model; [O-P] fragment at topological distance 3 is absent for compounds 26,112 and 113; for the JQ model; nBridgeHead, [N-P] fragment at topological distance 5 and [O-P] fragment at topological distance 1 are absent; in the case of MD model; C-012, [O-P] fragment at topological distance 7, [C-P] fragment at topological distance 5 and [C-Cl] fragment at topological distance 4 are absent and lastly, for RNP model; nRCONHR, [C-P] fragment at topological distance 4, [P-Cl] fragment at topological distance 5, and [O-S] fragment at topological distance 3 is absent. We have developed new QSTR models without the identified outliers and checked the statistical metrics (**provided in** Table S3 **of** Supplementary Information 2). A visual representation of the correlation between observed and predicted toxicity values has been depicted in the scatter plot (provided in Fig. 2). Additionally, we used two different ML algorithms namely support

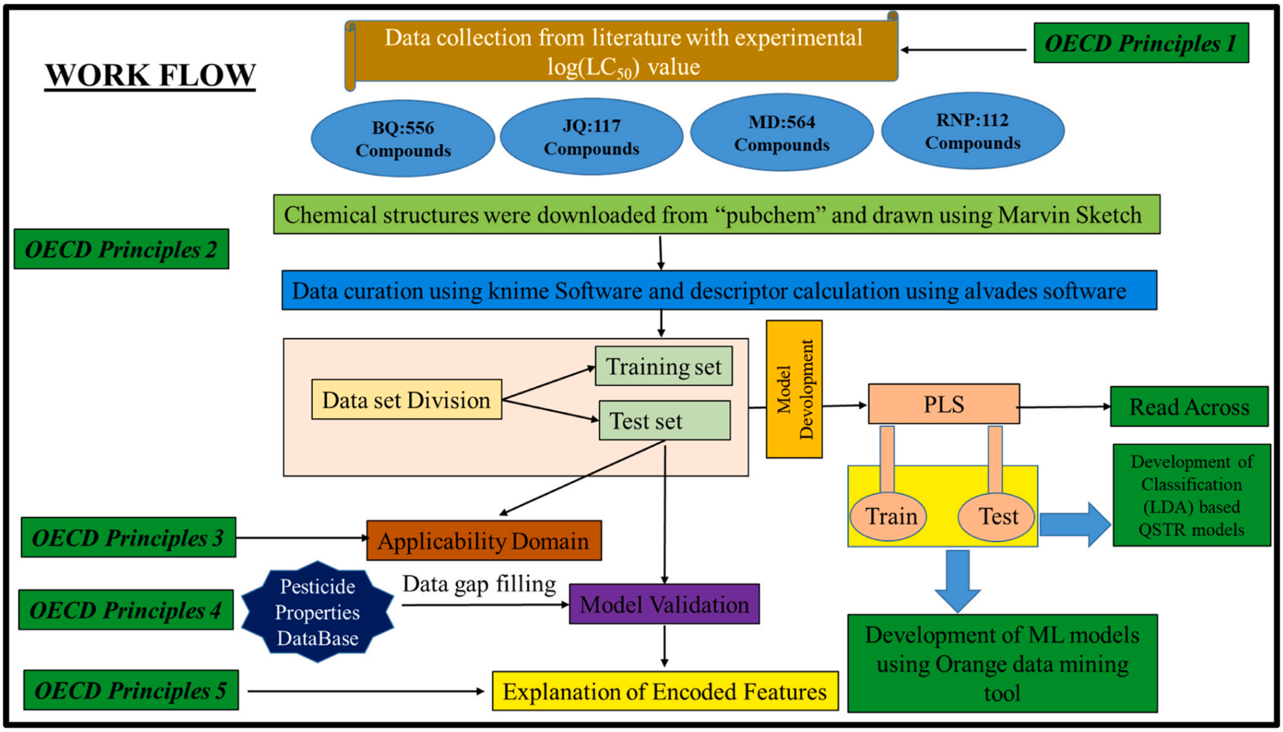

WORK FLOW
Data collection from literature with experimental log(LC50) value
OECD Principles 1
BQ:556 Compounds
JQ:117 Compounds
MD:564 Compounds
RNP:112 Compounds
Chemical structures were downloaded from "pubchem" and drawn using Marvin Sketch
OECD Principles 2
Data curation using knime Software and descriptor calculation using alvades software
Data set Division
Training set
Test set
Model Development
PLS
Read Across
Train
Test
Development of Classification (LDA) based QSTR models
OECD Principles 3
Applicability Domain
OECD Principles 4
Pesticide Properties DataBase
Data gap filling
Model Validation
Development of ML models using Orange data mining tool
OECD Principles 5
Explanation of Encoded Features

Fig. 1. Workflow of QSTR model development.

Table 1
Statistical parameter of developed PLS models.

Avian Species	Training set				Test set			
	N_{train}/N_{test}	LVs	R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	$MAE_{(test)}$	Quality$_{(test)}$
BQ	411/137	2	0.643	0.603	0.613	0.613	0.186	Good
JQ	77/34	2	0.630	0.552	0.534	0.519	0.403	Moderate
RNP	82/30	2	0.635	0.531	0.604	0.600	0.349	Moderate
MD	377/162	1	0.606	0.588	0.752	0.637	0.060	Good

Table 2
Read-across based predictions for four species.

Optimized settings	Metrics	Ygk (Test)
Bobwhite quail		
Ygk (Test)	Q^2_{F1}	0.690
σ = 0.25	Q^2_{F2}	0.690
γ = 0.25	RMSE$_P$	0.279
No. of similar compounds =10	MAE	0.179
Japanese quail		
Optimized settings	**Metrics**	**Ylk (Test)**
σ = 0.25	Q^2_{F1}	0.707
γ = 0.25	Q^2_{F2}	0.698
No. of similar compounds =10	RMSE$_P$	0.394
	MAE	0.307
Ring-necked pheasant		
Optimized settings	**METRICS**	**Ylk (Test)**
σ =0.5	Q^2_{F1}	0.714
γ =0.5	Q^2_{F2}	0.714
No. of similar compounds =10	RMSE$_P$	0.392
	MAE	0.290
Mallard duck		
Optimized settings	**METRICS**	**Yeuc (Test)**
σ =0.75	Q^2_{F1}	0.686
γ =0.75	Q^2_{F2}	0.540
No. of similar compounds =10	RMSE$_P$	0.114
	MAE	0.081

vector machine and random forest to evaluate their effectiveness in model construction and prediction. The PLS-based QSTR models with read-across predictions produce the lowest prediction error for the test set compounds, as indicated by the MAE$_{test}$ value compared to ML-based models against all of the avian species provided in Table S4 **of** Supplementary information 2**.** The equations of the final developed models of BQ, JQ, RNP, and MD are provided below:

Model BQ:

$$\boldsymbol{pLC50}\ (\boldsymbol{BQ}) = 1.25782 + 0.43538 \times \mathrm{F02[C-P]} + 0.00176 \times \mathrm{MW} + 0.5691 \times \mathrm{F09[S-F]} - 1.15994 \times \mathrm{B09[C-P]} - 0.55509 \times \mathrm{F03[O-P]} - 0.046 \times \mathrm{T(P..Cl)}$$

Model JQ:

$$\boldsymbol{pLC50}\ (\boldsymbol{JQ}) = 4.15712 + 0.74137 \times \mathrm{B01[O-P]} - 6.67929 \times \mathrm{X2A} + 1.18073 \times \mathrm{B05[N-P]} - 0.28037 \times \mathrm{H-048} - 0.00675 \times \mathrm{T(O..Cl)} + 0.44076 \times \mathrm{nBridgeHead}$$

Model RNP:

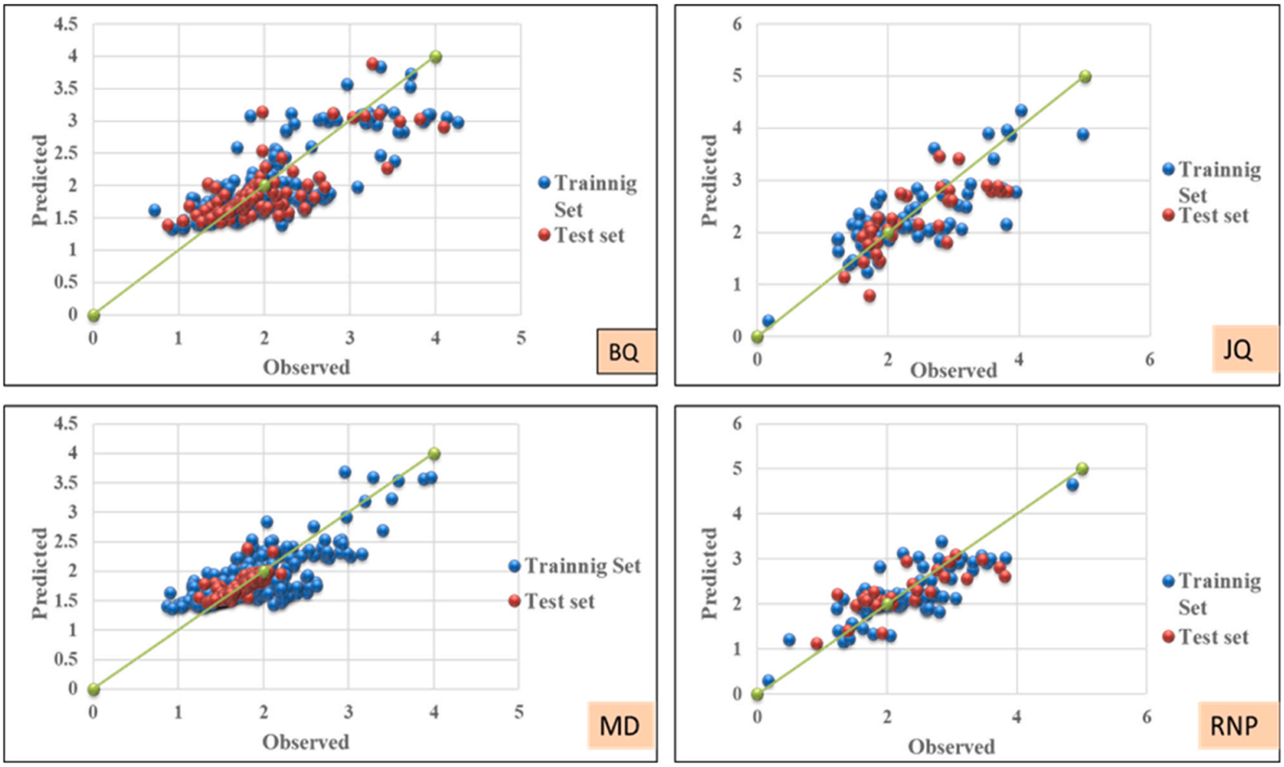


Fig. 2. Scatter plots of developed models.

$$\boldsymbol{pLC50}\ (\boldsymbol{RNP}) = 4.19704 - 6.73075 \times \mathrm{X2A} + 1.81161 \times \mathrm{nRCONHR} - 0.99523 \times \mathrm{nN(CO)2} + 0.84946 \times \mathrm{B04[C-P]} - 0.81404 \times \mathrm{B05[P-Cl]} - 0.42293 \times \mathrm{F03[O-S]}$$

Model MD:

$$\boldsymbol{pLC50}\ (\boldsymbol{MD}) = 1.31098 + 0.00138 \times \mathrm{MW} + 0.19812 \times \mathrm{C-012} + 1.25421 \times \mathrm{B07[O-P]} + 0.27204 \times \mathrm{Br-094} + 0.5788 \times \mathrm{B05[C-P]} + 0.01952 \times \mathrm{F04[C-Cl]}$$

Several classification-based metrics have been computed with the PLS-based QSTR-read across models for all (BQ, JQ, MD, and RNP) the avian species and reported in the following Table 3. Good sensitivity, specificity, and accuracy values indicate the good classification ability of the model. The computed values of the Matthews correlation coefficient (Matthews, 1975) indicate an acceptable prediction and an agreement between observed and predicted classification for all the developed models against avian species.

3.1. Regression coefficient plot

The descriptor's positive/negative contribution towards the toxicity is provided via a regression coefficient plot. In this investigation, the descriptors, F02[C-P], MW and F09[S-F]) contributed positively while the descriptors, B09[C-P], F03[O-P], and T(P.Cl) contributed negatively towards the toxicity of pesticides in case of BQ. In JQ, the descriptors which contributed positively toward the toxicity are B01[O-P], B05[N-P], nbridgehead and X2A, whereas the descriptors H-048 and T(O.Cl) contributed negatively towards the toxicity. In the case of MD, the descriptors MW, C-012, B07[O-P], Br-094, B05[C-P], and F04[C-Cl] contributed positively towards the toxicity. In case of RNP, the descriptors, nRCONHR and B04[C-P] contributed positively whereas the descriptors X2A, nN(CO)2, B05[P-Cl], and F03[O-S] contributed negatively towards the toxicity. All the relevant plots have been provided in Figs S5-S8 **in** supplementary information 2.

3.2. Variable importance plot (VIP)

The relative importance of model descriptors is illustrated with VIP (Akarachantachote et al., 2014). Descriptors having the highest and lowest impact on avian species can be recognized from these plots. The significance of the variable is higher if the VIP score is greater than 1. In VIP plot, the descriptors are presented concerning their significance (higher contribution to lower contribution) and their importance which is in the following order: F02[C-P], T(P.Cl), MW, B09[C-P], F03 [O-P], F09[S-F] (in case of BQ), B01[O-P], B05[N-P], X2A, nBridgeHead, H-048, T(O.Cl) (in case of JQ), B05[C-P], MW, B07[O-P], C-012, Br-094, F04[C-Cl)] (in case of MD) and B04[C-P], X2A, nRCONHR, F03[O-S], B05[P-Cl], Nn(CO)2 (in case of RNP) as depicted in **Figs: S9-S12 in** supplementary information 2.

3.3. Loading plot

The loading plot shows how the independent variables (descriptors) are related to the response variable. The first two components were used to create the loading plot. A descriptor is assumed to have a stronger effect on response value if it is located far from the origin of the plot. On the basis of the loading plot as shown in **Figs. S13-S16 in** supplementary information 2; it is interpreted that the X-variables F02[C-P] and MW have more influence to the Y-variable as traced from the proximity with response variable and the presence of these features elevated pesticide toxicity towards BQ. Similarly, B01[O-P], B05[C-P], and B04[C-P] are the most influential descriptors in the case of JQ, MD, and RNP respectively.

3.4. Mechanistic interpretation of PLS models

Table 4 and Figs. 3–6 provide a detailed account of the model descriptors followed by mechanistic interpretations important to identify major structural and physicochemical features.

3.5. Pesticide Properties DataBase screening

Pesticide Properties DataBase was screened through the developed models with the help of the software "PRI Tool_PLSversion" (available from http://teqip.jdvu.ac.in/QSAR Tools/) using the developed PLS models. The categorization threshold (mean value of the training set compound) for avian toxicity against BQ; JQ; MD; RNP ≥ 1.883; 2.236; 1.845; 2.191 respectively was applied for prioritization purposes. From the prediction, it was seen that maximum compounds are within the domain of applicability and show prediction quality as "good". The screened chemicals from the Pesticide Properties DataBase with their respective predicted toxicity against BQ, JQ, MD, and RNP are shown in supplementary information 1. The compounds were ranked in decreasing order of predicted toxicity for each avian species. The top 20 and least 20 toxic pesticides for all four avian species from the PPDB database are provided in Table 4. Further validation of the predicted toxicity of the selected pesticides revealed that apart from fluoroacetamide and sodium monofluoroacetate, all the predicted toxicity corroborated with the previous experimental findings, indicating the practical applicability of the developed models as shown in Table 5.

Table 3
Statistics of the classification-based QSTR models.

Sl no.	LDA-QSTR MODELS	AUC-ROC	SENSITIVITY	ACCURACY	PRECISION	F-MEASURE	MCC
1	BQ (train)	0.80	54.54	83.33	88.00	67.35	0.59
	BQ (test)	0.83	52.17	85.36	92.30	66.67	0.62
2	JQ (train)	0.82	62.50	80.76	86.95	72.73	0.60
	JQ (test)	0.80	75.00	84.84	81.81	78.26	0.66
3	MD (train)	0.88	75.00	83.59	82.60	78.62	0.65
	MD (test)	0.86	75.71	85.71	89.83	82.17	0.71
4	RNP (train)	0.83	63.88	79.74	88.46	74.19	0.60
	RNP (test)	0.87	76.92	84.84	83.33	80.00	0.67

Table 4
Mechanistic analysis of model descriptors of all species.

S. no	Descriptor	Type	Function	Contribution	Mechanistic introspection
BQ oral pLC$_{50}$					
1	F02[C-P]	2D Atom pair	Frequency of carbon and phosphorus atoms at topological distance 2	+ve	Generally, the phosphate group is toxic (Vervloet, 2019a).The presence of more phosphate groups in a molecule tends to increase its toxicity as evidenced in compound **442**. On the other hand, the presence of less number of these fragments in a compound may result in low toxicity values, as seen in compound **501** (depicted in Fig. 3).
2	MW	Constitutional descriptor	Molecular weight	+ve	This descriptor is directly related to the molecular size and bulkiness of molecules. It may influence diffusion in biological membranes and fluid media (Hou et al., 2004; Khan et al., 2019). So the chemicals may easily cross the biological membrane of species and retain in the body of reference species for a long time, which ultimately enhances the toxicity (Basant et al., 2015) as demonstrated in compound **381** and vice versa in compound **239** (given in Fig. 3).
3	F09[S-F]	2D Atom pair	Frequency of sulfur and fluorine atoms at topological distance 9	+ve	Lipophilic substances have a greater susceptibility to accumulation within the cells, resulting in a higher pesticide concentration inside the organism, which ultimately leads to enhanced toxic effects. The presence of two highly electronegative atoms (fluorine and sulfur) as well as a long carbon chain (lipophilicity) in a compound tend to make it more reactive and potentially more toxic (Mukherjee et al., 2021; Ghosh et al., 2020) as shown in compound **23** and oppositely occurs in compound **523** (shown in Fig. 3).
4	B09[C-P]	2D Atom pair	Presence/absence of carbon and phosphorus atoms at topological distance 9	-ve	The negative regression coefficient of this descriptor indicates that the presence of carbon and phosphorus atoms at the topological distance 9 may decrease the pesticide's toxicity towards avian species as shown in compound 296 while the absence of this fragment in a chemical may have higher toxicity values as shown in the case of compound **11** (described in Fig. 3).
5	F03[O-P]	2D Atom pair	Frequency of oxygen and phosphorus atoms at topological distance 3	-ve	The negative regression coefficient of this descriptor indicates that it inversely correlated with the pesticide's toxicity towards avian species. Thus, the presence of this fragment reduces the compound toxicity as demonstrated in compound **487** and the absence of this fragment enhances the toxicity as represented in compound **52** (given in Fig. 3).
6	T(P.Cl)	2D Atom pair	Sum of topological distances between P.Cl	-ve	The two-dimensional atom pair descriptor, T(P⋯Cl) accounts for the topological distances between phosphorus and chlorine atoms. Reduction of inductivity in chlorine substituents causes a decrease in electron density for the relevant compounds. Therefore, the incidence of the P–Cl bond in aromatic chemicals reduces the electron density of the aromatic ring, thus, electron-donor-acceptor interactions cannot happen easily between pesticides and the reference species (Ghosh et al., 2020). This descriptor has a negative regression coefficient, indicating that the presence of this fragment will result in a decrease in pesticide toxicity profile, as exemplified by compound **243**, while it would have the opposite effect when present, as proven by compound **441** (provided in Fig. 3).
JQ oral pLC$_{50}$					
1	B01[O-P]	2D Atom pair	Presence/absence of O – P at topological distance 1	+ve	The presence of two electronegative atoms (O and P) in a compound makes it more electronegative which leads to oxidative stress and the death of the reference species (Kumar et al., 2023; Roy and Roy, 2021). This phenomenon is demonstrated in compound **81** and inversely occurs in compound **113** (shown in Fig. 4).
2	X2A	Connectivity indices descriptor	Average connectivity index of order 2	-ve	X2A represents the degree of branching in molecules, which is inversely correlated with hydrophobic interaction as well as toxicity (Arvidsson et al., 1971; Roy and Das, 2013). Thus, the higher numerical value of this descriptor leads to a decrease in toxicity value as shown in compound **13** and vice versa occurs in compound **57** (given in Fig. 4).
3	B05[N-P]	2D Atom pair	Incidence of N – P at topological distance 5	+ve	The presence of two electronegative atoms (N and P) in a compound makes it more electronegative which leads to oxidative stress and the death of the reference species (Zhang et al., 2015; Roy and Roy, 2021). This phenomenon is demonstrated in compound **88.** On the other hand, the compound containing less number of this fragment may exhibit less toxicity as shown in compound **66** (demonstrated in Fig. 4).
4	H-048	Atom-centered fragments	H attached to C2(sp3)/C1(sp2)/C0 (sp)	-ve	H-048 has the potential to make compounds electronically conductive as well as hydrophilic (Kumar et al., 2013). Hydrophilicity and toxicity are inversely related to each other (Li et al., 2022). Thus the presence of a greater number of this descriptor in a molecule makes it less toxic as shown in compound **67**. On the other side, the presence of less number of hydrophilic groups in a molecule leads to an increase the toxicity as shown in compound **11** (depicted in Fig. 4)
5	T(O.Cl)	2D Atom pair	Sum of topological distances between O.Cl	-ve	The negative regression coefficient of this descriptor indicates that it is inversely correlated with the pesticide's toxicity towards avian species thus the presence of more of this fragment makes the compound less toxic as shown in compound **33** and conversely occurs in compound **84** (depicted in Fig. 4).

(*continued on next page*)

Table 4 (*continued*)

S. no	Descriptor	Type	Function	Contribution	Mechanistic introspection
6	nBridgeHead	Ring descriptors	Number of bridgehead atoms	+ve	Usually, bridgehead atoms have a complex structure as well as toxic (Kumar et al., 2023) which is demonstrated in compound **19**. Conversely, the absence of bridgehead atoms makes the compound less toxic as shown in compound **110** (demonstrated in Fig. 4).
MD oral pLC$_{50}$					
1	MW	Constitutional descriptor	Molecular weight	+ve	This descriptor is directly related to molecular bulkiness and lipophilicity (Hou et al., 2004; Khan et al., 2019). Usually, lipophilic compounds easily cross the lipophilic membrane of the reference species which ultimately leads to enhancement in toxicity as demonstrated in compound **546** and oppositely occurs in compound **503** (given in Fig. 5).
2	C-012	Atom-centered fragments	CR2X2 (X is a hetero atom (O, N, S, P, Se, or halogens) and R is a carbon-linked group)	+ve	This descriptor enhances the molecular size as well as the electronegativity of the compound due to the presence of heteroatom, which ultimately leads to enhancement in toxicity of diverse pesticides against avian species by incorporating oxidative stress (Kar et al., 2020) as demonstrated in compound **445,** and vice-versa occurs in compound **144** (depicted in Fig. 5).
3	B07[O-P]	2D Atom Pair	presence of O – P at topological distance 7	+ve	Oxygen and phosphorus are highly electronegative atoms and their presence makes the compound more toxic (due to increment in oxidative stress in reference species) (Roy and Roy, 2021). The presence of a long carbon chain (lipophilicity) also contributes to toxicity. This phenomenon is demonstrated in compound **3** and vice versa occurs in the case of compound **145** (illustrated in Fig. 5).
4	Br-094	Atom-centered fragments	Br attached to C1(sp2)	+ve	The Br-094 descriptor refers to the presence of the halogen group (bromine). Thus, the presence of more electronegative/halogen atoms (bromine) makes the compound more toxic as demonstrated in compound **28**. Conversely, absence of this atom/fragment tends to decrease the toxicity as shown in compound **408** (depicted in Fig. 5).
5	B05[C-P]	2D Atom pair	C – P situated at topological distance 5	+ve	The presence of the phosphate group enhances the toxicity of the compound (Vervloet, 2019b). This is evidenced in compound **4.** In opposition, absence of this fragment tends to decrease the toxicity as shown in compound **530** (provided in Fig. 5).
6	F04[C-Cl]	2D Atom pair	C – Cl situated at topological distance 4	+ve	This descriptor refers to the existence of a large electronegative atom such as chlorine, which has a high atomic refractivity and electronegativity (Khan and Roy, 2019). Thus, the presence of a greater number of this fragment results in high toxicity toward avian species as shown in compound **24** and vice versa occurs in compound **562** (provided in Fig. 5).
RNP oral pLC$_{50}$					
1	X2A	Connectivity indices descriptor	Average connectivity index of order 2	-ve	The negative regression coefficient of this descriptor indicates that higher numerical value of this descriptor leads to a decrease in toxicity as shown in compound **13** and vice versa in the case of compound **51** (given in Fig. 6). X2A is inversely correlated with hydrophobic interaction as well as toxicity (Arvidsson et al., 1971; Roy and Das, 2013).
2	nRCONHR	Functional group count	Presence of secondary aliphatic amides	+ve	Aliphatic amides are considered to be toxic as well as reactive (Schultz et al., 2006). The positive regression coefficient of this descriptor indicates that presence of this fragment may increase the toxicity as demonstrated in compound **90** and toxicity value may be decreased if the compounds have no such fragment as represented in compound **104** (shown in Fig. 6).
3	nN(CO)2	Functional group count	Number of imides (-thio)	-ve	Generally, this feature helps to facilitate hydrolysis of the compounds which facilitates quick excretion from the body of the reference organism resulting in a reduction of their toxic effects (Krishna et al., 2020) as demonstrated in compound **58** and the absence of this fragment tends to increase the toxicity as shown in compound **101** (illustrated in Fig. 6).
4	B04[C-P]	2D Atom pair	C – P situated at topological distance 4	+ve	The presence of an electronegative atom (like phosphorous) enhances the toxicity of the diverse pesticides by incorporating oxidative stress in avian species (Mukherjee et al., 2021; Kumar et al., 2024) as evidenced by compound **3.** On the other hand, the absence of this fragment leads to a decrease the toxicity as shown in compound **10** (described in Fig. 6).
5	B05[P-Cl]	2D Atom pair	Presence of P – Cl at topological distance 5	-ve	The negative regression coefficient of this descriptor indicates that presence of more number of this fragment reduces the toxicity as demonstrated in compound **105** and oppositely occurs in case of compound **62** (depicted in Fig. 6).
6	F03[O-S]	2D Atom pair	Frequency of oxygen and sulfur which are situated at topological distance 3.	-ve	This descriptor is directly related to the polarity (presence of polar bond) (Mukherjee et al., 2021) of the compound, as a result the hydrophilicity of the compound increase and thus toxicity will decrease which is evidenced by compound **85** and vice versa in case of compound **9**. (represented in Fig. 6).

3.6. Comparison with previous work

As the composition of the training and test sets, endpoints used, as well as the algorithms used for model development are not the same, we can't perform a rigorous comparison, so we have attempted to represent some simple comparative studies between the current work and previously reported literature. Mukherjee et al. (Mukherjee et al., 2021) developed the models using small data sets in comparison with current work. Basanta et al. (Basant et al., 2015) used tree-based approaches to build QSTR and i-QSTR models for various avian species. Banjare et al.

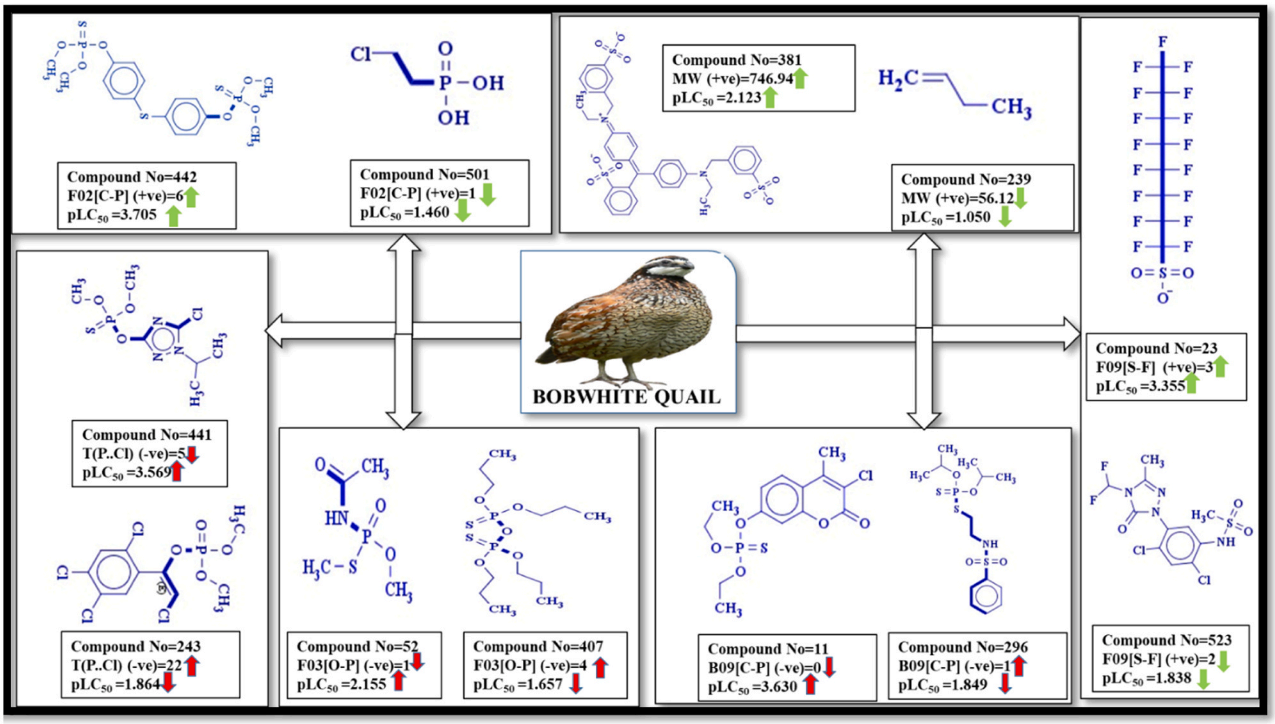


Fig. 3. Positive and negative contribution of model descriptors towards BQ.

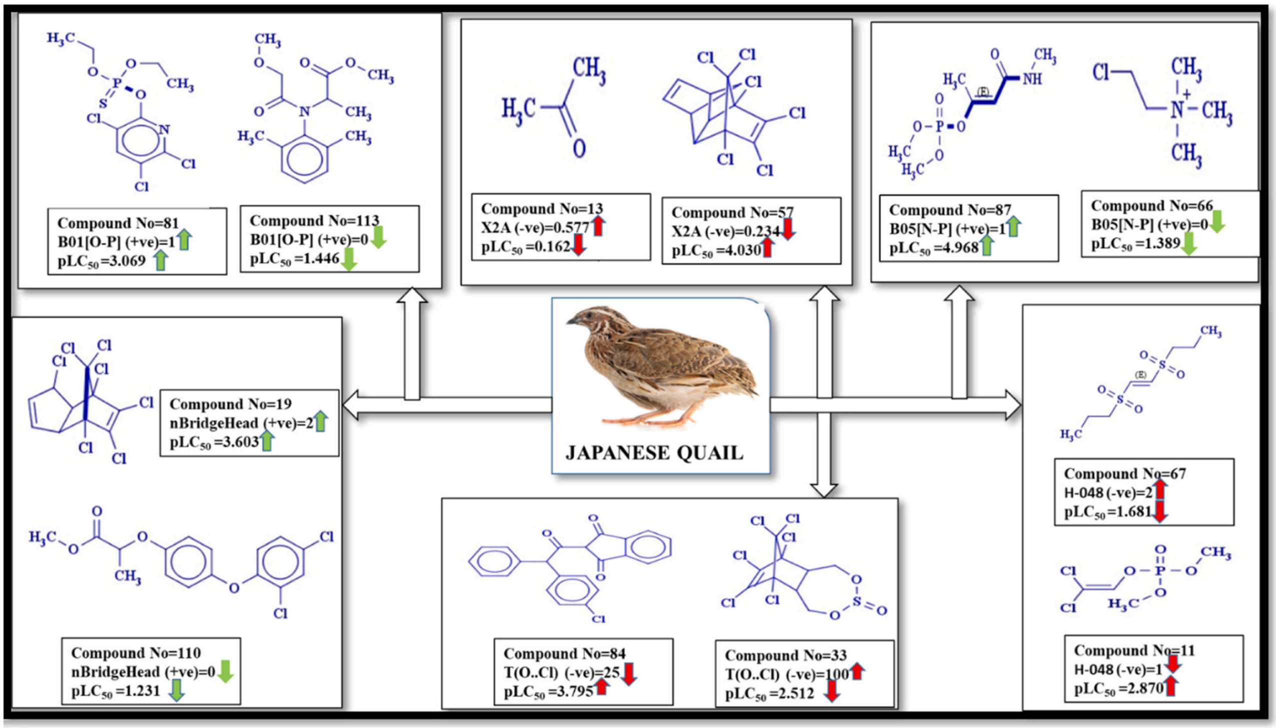


Fig. 4. Positive and negative contribution of model descriptors towards JQ.

(Banjare et al., 2021) presented QSTR and i-QSTR models for three avian species using a classification approach. Podder et al. (Podder et al., 2023; O'Boyle et al., 2011) developed a regression-based QSTR and i-QSTR models against multiple avian species (MD, BQ, and ZF). Leszczynski et al. (Kar and Leszczynski, 2020) reported ecotoxicity QSTR and i-QSTR modeling of chemicals to avian species. While regression models provide explicit quantitative predictions, classification approaches can be useful for data filtering at the outset of research. The current models are built using a regression-based method and a limited number of simple, 2D, and easily interpretable descriptors. In this work, we have tried to develop first PLS-based QSTR model considering LC_{50} as an endpoints to assess the toxicity of diverse pesticides against multiple avian species. Regression-based technique is an assertive and effective approach that can confidently tackle challenges

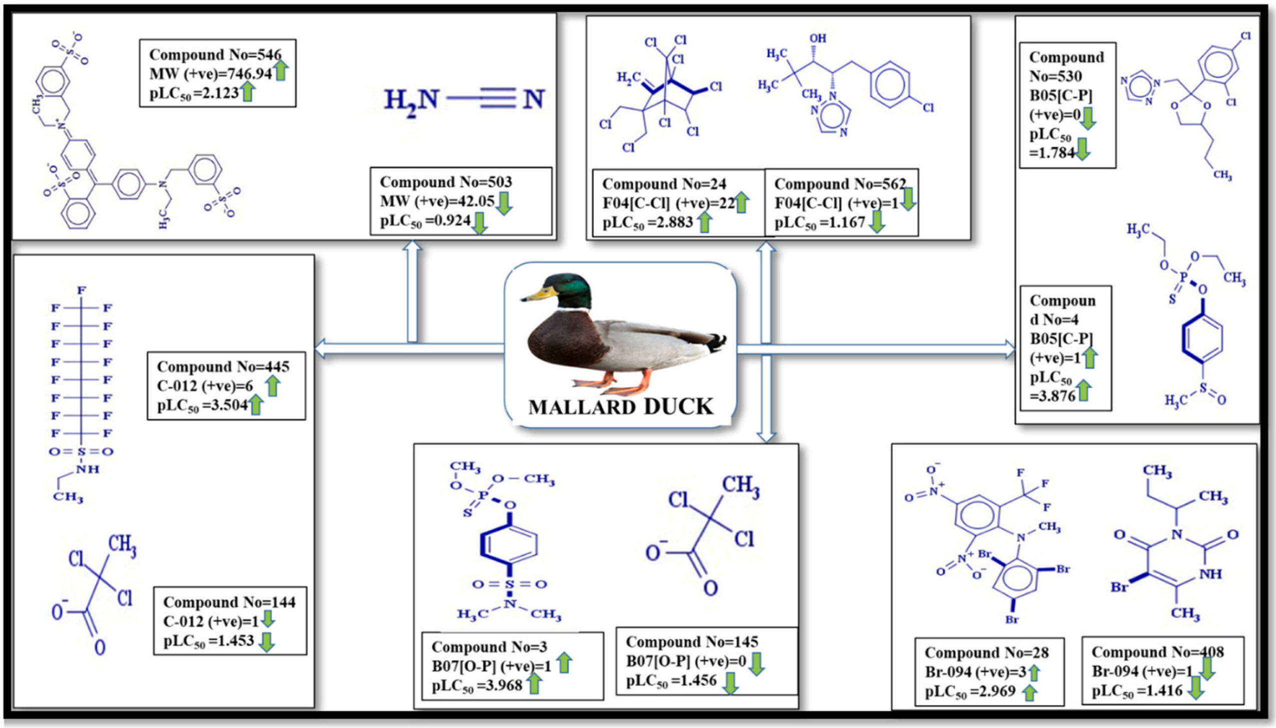

MALLARD DUCK
Compound No=546 MW (+ve)=746.94 pLC50 =2.123
Compound No=503 MW (+ve)=42.05 pLC50 =0.924
Compound No=24 F04[C-Cl] (+ve)=22 pLC50 =2.883
Compound No=562 F04[C-Cl] (+ve)=1 pLC50 =1.167
Compound No=530 B05[C-P] (+ve)=0 pLC50 =1.784
Compoun d No=4 B05[C-P] (+ve)=1 pLC50 =3.876
Compound No=445 C-012 (+ve)=6 pLC50 =3.504
Compound No=144 C-012 (+ve)=1 pLC50 =1.453
Compound No=3 B07[O-P] (+ve)=1 pLC50 =3.968
Compound No=145 B07[O-P] (+ve)=0 pLC50 =1.456
Compound No=28 Br-094 (+ve)=3 pLC50 =2.969
Compound No=408 Br-094 (+ve)=1 pLC50 =1.416

Fig. 5. Positive and Negative contribution of model descriptors towards MD.

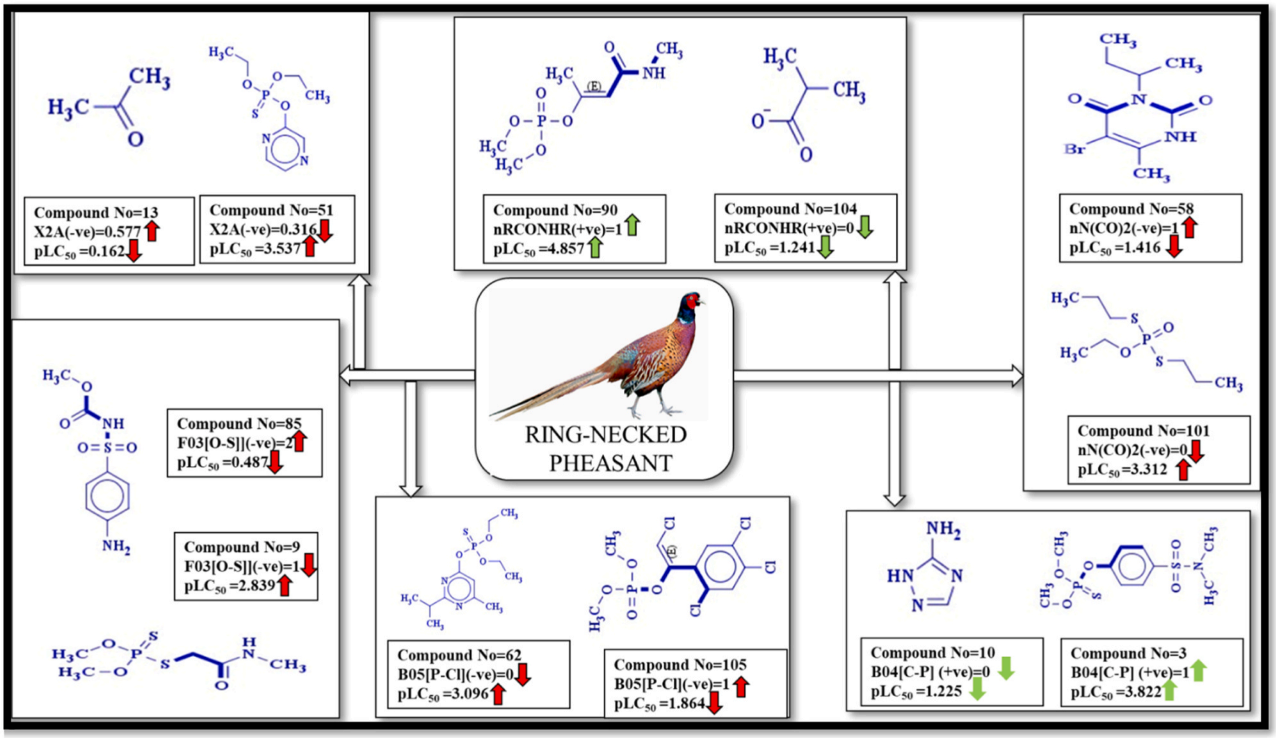

RING-NECKED PHEASANT
Compound No=13 X2A(-ve)=0.577 pLC50 =0.162
Compound No=51 X2A(-ve)=0.316 pLC50 =3.537
Compound No=90 nRCONHR(+ve)=1 pLC50 =4.857
Compound No=104 nRCONHR(+ve)=0 pLC50 =1.241
Compound No=58 nN(CO)2(-ve)=1 pLC50 =1.416
Compound No=101 nN(CO)2(-ve)=0 pLC50 =3.312
Compound No=85 F03[O-S]](-ve)=2 pLC50 =0.487
Compound No=9 F03[O-S]](-ve)=1 pLC50 =2.839
Compound No=62 B05[P-Cl](-ve)=0 pLC50 =3.096
Compound No=105 B05[P-Cl](-ve)=1 pLC50 =1.864
Compound No=10 B04[C-P] (+ve)=0 pLC50 =1.225
Compound No=3 B04[C-P] (+ve)=1 pLC50 =3.822

Fig. 6. Positive and Negative contribution of model descriptors towards RNP.

such as descriptor inter-correlation, high levels of noise, collinearity, and a large number of descriptors. In the present work, we have developed the models using large datasets of different avian species. So, it has a wide domain of applicability compared to previous studies. Additionally, we used read-across algorithm to enhance the external predictivity and it is widely used for data-gap filing as well as widely accepted and recommended by regulatory bodies Apart from the previous studies, and consequently read-across prediction shows a better result than the previous model except for MD. Apart from the previous studies, we get additionally some new findings (specifically observation) which are related to pesticide toxicity towards avian species such as presence of C-012 (CR2X2), B07[O-P] (Presence/absence of O–P at

Table 5
Top 20 and least 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB).

Sl. no.	Pesticide	Safety and Hazards	Sources
Top 20 most toxic screened pesticides from Pesticide Properties DataBase (PPDB).			
1	Imicyafos	Acute toxic, Irritant.	https://pubchem.ncbi.nlm.nih.gov/compound/18772487#section=Safety-and-Hazards&fullscreen=true
2	Pirimiphos-ethyl	Acute toxic, Environmental Hazard.	https://pubchem.ncbi.nlm.nih.gov/compound/31957#section=Safety-and-Hazards&fullscreen=true
3	Quinothion	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/89714#section=Toxicity&fullscreen=true
4	Pirimiphos-methyl	Irritant, Health hazard, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/34526#section=Safety-and-Hazards&fullscreen=true
5	Etrimfos	Irritant, Environmental Hazard	https://pubchem.ncbi.nlm.nih.gov/compound/37995#section=Safety-and-Hazards&fullscreen=true
6	Buminafos	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/39966#section=Toxicity&fullscreen=true
7	Diazinon	Irritant, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/3017#section=Safety-and-Hazards&fullscreen=true
8	Quintiofos	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/72069#section=Toxicity&fullscreen=true
9	Phoxim	Irritant, Health hazard, and Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/9570290#section=Safety-and-Hazards&fullscreen=true
10	Inezin	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/30772#section=Toxicity&fullscreen=true
11	Dufulin	Oxidative stress inducer	(Yu et al., 2021).
12	Chlorphoxim	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/5360461#section=Safety-and-Hazards&fullscreen=true
13	Pyridaphenthion	Irritant	https://pubchem.ncbi.nlm.nih.gov/compound/8381#section=Safety-and-Hazards&fullscreen=true
14	Triazophos	Acute toxic, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/32184#section=Safety-and-Hazards&fullscreen=true
15	Isoxathion	Acute toxic, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/29307#section=Safety-and-Hazards&fullscreen=true
16	Naftalofos	Acute toxic	https://pubchem.ncbi.nlm.nih.gov/compound/15148#section=Safety-and-Hazards&fullscreen=true
17	Quinalphos	Acute toxic, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/26124#section=Safety-and-Hazards&fullscreen=true
18	Butamifos	Irritant, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/37419#section=Safety-and-Hazards&fullscreen=true
19	Sulprofos	Acute toxic, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/37125#section=Safety-and-Hazards&fullscreen=true
20	Edifenphos	Acute toxic, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/28292#section=Safety-and-Hazards&fullscreen=true
Least 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB).			
1	Ferbam	non-toxic	https://www3.epa.gov/pesticides/chem_search/reg_actions/reregistration/fs_PC-034801_01-Sep-05.pdf
2	Hexylene glycol	less toxic	https://hpvchemicals.oecd.org/ui/handler.axd?id=3c2a8190–8500–467c-af27-a636e6636c38
3	Bisthiosemi	moderate toxic	https://www.drugfuture.com/toxic/dir/5061.html
4	Choline chloride	less toxic	http://sitem.herts.ac.uk/aeru/iupac/Reports/161.htm
5	Glutaraldehyde	less toxic	https://archive.epa.gov/pesticides/reregistration/web/pdf/glutaraldehyde-red.pdf
6	Fumaric acid	less toxic	https://www.sciencedirect.com/science/article/pii/S0095955315310854
7	Lime sulphur	less toxic	https://www.ams.usda.gov/sites/default/files/media/Lime%20Sulfur%20Evaluation%20TR.pdf
8	Methyl isobutyl ketone	less toxic	https://www.epa.gov/sites/default/files/2016–09/documents/methyl-isobutyl-ketone.pdf
9	Sodium tetrathiocarbonate	moderate toxic	https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/thiocarbonate
10	1,2-dichloropropane	less toxic	https://wedocs.unep.org/bitstream/handle/20.500.11822/29625/HSG76.pdf?sequence=1&isAllowed=y
11	Metam	less toxic	https://archive.epa.gov/pesticides/chemicalsearch/chemical/foia/web/pdf/039003/039003–028.pdf
12	Methylene bisthiocyanate	less toxic	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2905.htm
13	Bentonite	Nontoxic	https://digitalfire.com/hazard/bentonite+toxicity#:~:text=Bentonite%20is%20a%20ground%20naturally,flush%20to%20remove%20the%20particles.
14	Butanethiol	moderate toxic	https://pubchem.ncbi.nlm.nih.gov/compound/1-Butanethiol
15	Sodium monochloroacetate	moderate toxic	https://tera.org/OARS/Sodium%20Chloroacetat%20(3926–62–3)%20WEEL%202016%20public%20comment.pdf
16	Fluoroacetamide	high toxic	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/338.htm
17	Sodium monofluoroacetate	high toxic	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/3160.htm
18	Propylene glycol	less toxic	https://downloads.regulations.gov/EPA-HQ-OPP-2013–0218–0007/content.pdf
19	Peroxyacetic acid	moderate toxic	https://www.federalregister.gov/documents/2000/12/01/00–30679/peroxyacetic-acid-exemption-from-the-requirement-of-a-tolerance#:~:text=Because%20of%20the%20low%20toxicity,not%20pose%20a%20dietary%20risk
20	2-hydrazinoethanol	moderate toxic	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2803.htm

topological distance 7), Br-094 (Br attached to C1(sp2)), B05[C-P] (Presence/absence of C–P at topological distance 5), F04[C-Cl] (Frequency of C–Cl at topological distance 4) and nRCONHR (number of secondary amides (aliphatic)) enhances the pesticides toxicity towards avian species; on the other hands, presence of nN(CO)2 (number of imides (-thio)) and B05[P-Cl] (Presence/absence of P–Cl at topological distance 5) reduces the pesticides toxicity towards avian species. Furthermore, our work highlighted some extra features not mentioned in the previous studies, which are useful for pesticide toxicity assessment viz. molecular weight, presence of heteroatom, presence of bridgehead atoms, secondary aliphatic amide, and molecular refractivity. On the other hand, features like molecular branching and the presence of thio imides contribute negatively towards the toxicity. The PPDB database was screened using developed models to show the predictivity as well as application in the real-world data of the developed models. The current study's comparison to previously published studies is depicted in Table 6.

4. Conclusion

In summary, this study employs a range of chemometric tools to predict pesticide toxicity for four different avian species. The research focuses on creating robust and easily interpretable QSTR models based on OECD principles. The study's statistical validation parameters consistently demonstrate the strength and reliability of the constructed PLS-based QSTR-read across models. External validation metrics, employing the read-across algorithm, show slightly superior performance in predicting toxicity, except for the mallard duck dataset. Additionally, we have developed classification models and employed two Machine Learning algorithms SVM and RF to evaluate their effectiveness in constructing models and making predictions. The PLS-based QSTR models with read-across predictions produce better statistical results (such as the lowest prediction error for the test set compounds, as indicated by the MAE_{test} value) as compared to ML-based models against all of the avian species.

Furthermore, this research develops regression-based models, surpassing previous studies in terms of the dataset's size, the variety of avian species examined, domain of applicability features responsible for toxicity, model quality, algorithm used as well as the endpoint (LC_{50}). The findings highlight the significance of electronegativity, molecular weight, imide count, lipophilicity, and steric effects in avian toxicity. Additional findings (descriptors) such as C-012, B07[O-P], Br-094, B05[C-P], F04[C-Cl], nRCONHR, $nN(CO)_{2,}$ and B05[P-Cl] were observed in this study which is related to pesticides toxicity towards avian species. Notably, the presence of C-P fragments at specific topological distances and electronegative groups intensifies toxicity, while features like branching and hydrogen bond acceptor characteristics reduce it.

The validation of the predicted toxicity of the screened compounds by experimental data demonstrated the reliability and feasibility of applying the developed models for screening pesticides, offering valuable support to researchers striving to design eco-friendly and safe chemical pesticides. They effectively bridge gaps in toxicity data and simplify the evaluation of novel pesticides for various bird species. Moreover, these models significantly reduce the time, resources, costs, and the need for animal testing, aligning with the principles of reduction, refinement, and replacement (RRR) in research practices.

Funding sources

The author(s) received no specific funding for this work.

CRediT authorship contribution statement

Shubha Das: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Writing – original draft. **Ankur Kumar:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – review & editing. **Abhisek Samal:** Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft. **Supratik Kar:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Vinayak Ghosh:** Conceptualization, Data curation, Investigation, Methodology, Writing – review & editing. **Probir Kumar Ojha:** Conceptualization, Investigation, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

Table 6
Comparison table with previous work.

Source	Organisms used in this study	Defined endpoint	Model	LV	Features	Training set			Test set		
						N_{train}	R^2	Q^2_{Loo}	N_{test}	Q^2_{F1}	Q^2_{F2}
In this present study	BQ	LC_{50}	PLS-Read across	2	6	411	0.64	0.60	137	0.61–0.69	0.61–0.69
	JQ			2	6	77	0.63	0.55	34	0.53–0.70	0.51–0.69
	RNP			2	6	82	0.63	0.53	30	0.60–0.71	0.60–0.71
	MD			1	6	377	0.60	0.58	162	0.71–0.75	0.63–0.68
(Mukherjee et al., 2021)	BQ	LD_{50}	PLS	3	10	103	0.65	0.58	25	0.64	0.64
	JQ			2	3	–	0.73	0.59	–	–	–
	RNP			2	4	22	0.76	0.60	7	0.64	0.64
	MD			2	7	49	0.65	0.56	13	0.65	0.57
	HS			1	2	–	0.91	0.86	–	0.94	0.88
Mazzatorta *et al (* Kim, 2019*)*.	BQ	LD_{50}	GA-SVM	–	–	94	–	–	19	—	—
Podder et al (O'Boyle et al., 2011).	BQ	LD_{50}	MLR	-	7	278	0.715–0.719	0.694–0.700	88	0.722–0.732	0.722–0.732
	MD			-	8	182	0.689–0.708	0.626–0.695	65	0.620–0.639	0.620–0.638
	ZF			-	5	40	0.754–0.758	0.697–0.722	13	0.787–0.830	0.786–0.829
(Banjare et al., 2021).	BQ	LD_{50}	GA-LDA along with interspecies correlation	-	-	203	-	-	67	-	-
	MD			-	-	143	-	-	60	-	-
	ZF			-	-	31	-		12	-	
(Basant et al., 2015).	BQ	LD_{50}	Tree-based QSAR approaches	-	-	98	-	-	33	-	-
(Kar and Leszczynski, 2020).	BQ	LD_{50}	GFA-PLS	3	5	41	0.67	0.63	15	0.70	0.68
	MD			2	5	42	0.75	0.67	14	0.88	0.87
	RNH			3	4	20	0.89	0.80	7	0.87	0.87

LV: Latent variable; PLS: Partial least square; SVM: Support vector machine.

the work reported in this paper.

Acknowledgments

SD and AS are thankfully acknowledged for financial assistance from the AICTE, New Delhi in the form of a scholarship. AK thanks the GPC Regulatory India Private Limited for financial support in the form of a project assistant (GPC Regulatory India Private Limited sponsored research, Ref No-P-1/RS/171/22, date-07-09.2022). S.K. wants to thank the administration of Dorothy and George Hennings College of Science, Mathematics and Technology (HCSMT) of Kean University for providing research opportunities through release time for research and resources.

Author contributions

The manuscript was written with the contributions of all authors. All authors have approved the final version of the manuscript.

Additional contents

Supporting information 1

SMILES of the whole dataset compounds and corresponding toxicity values against BQ, JQ, MD, and RNP Avian species

Supporting information 2

Different PLS plots; Fig. (S1-S16) of the individual QSTR models for *BQ, JQ, MD, and RNP*

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.psep.2024.05.095.

References

Akarachantachote, N., Chadcham, S., Saithanu, K., 2014. Cutoff threshold of variable importance in projection for variable selection. Int J. Pure Appl. Math. *94* (3), 307–322. https://doi.org/10.12732/ijpam.v94i3.2.

Ambure, P., Aher, R.B., Gajewicz, A., Puzyn, T., Roy, K., 2015. NanoBRIDGES" software: open access tools to perform QSAR and nano-QSAR modeling. Chemom. Intell. Lab. Syst. *147*, 1–13. https://doi.org/10.1016/j.chemolab.2015.07.007.

Arvidsson, E.O., Green, F.A., Laurell, S., 1971. Branching and Hydrophobic Bonding: partition equilibria and serum albumin binding of palmitic and phytanic acids. J. Biol. Chem. *246* (17), 5373–5379. https://doi.org/10.1016/S0021-9258(18)619179.

Banerjee, A., De, P., Kumar, V., Kar, S., Roy, K., 2022. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across. Chemosphere *309*, 136579. https://doi.org/10.1016/j.chemosphere.2022.136579.

Banjare, P., Singh, J., Roy, P.P., 2021. Predictive classification-based QSTR models for toxicity study of diverse pesticides on multiple avian species. Environ. Sci. Pollut. Res. *28* (14), 17992–18003. https://doi.org/10.1007/s11356-020-11713-z.

Basant, N., Gupta, S., Singh, K.P., 2015. Predicting toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. J. Chem. Inf. Model. *55* (7), 1337–1348. https://doi.org/10.1021/acs.jcim.5b00139.

Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A., Roy, K., 2022. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ. Sci.: Nano *9* (1), 189–203. https://doi.org/10.1039/D1EN00725D.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., 2013. Orange: data mining toolbox in Python. J. Mach. Learn. Res. 14 (1), 2349–2353.

Dillon, W.R., Goldstein, M., 1984. Multivariate analysis: Methods and applications, 1984. Wiley, New York (NY).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Ghosh, S., Ojha, P.K., Carnesecchi, E., Lombardo, A., Roy, K., Benfenati, E., 2020. Exploring QSAR modeling of toxicity of chemicals on earthworm. Ecotoxicol. Environ. Saf. 190, 110067 https://doi.org/10.1016/j.ecoenv.2019.110067.

Halder, A.K., Moura, A.S., Cordeiro, M.N.D., 2023. Predicting the ecotoxicity of endocrine disruptive chemicals: multitasking in silico approaches towards global models. Sci. Total Environ. *889*, 164337 https://doi.org/10.1016/j.scitotenv.2023.164337.

Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L., Moussa, C.S., 2016. A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. J. Hazard. Mater. 303, 28–40. https://doi.org/10.1016/j.jhazmat.2015.09.021.

Hou, T.J., Zhang, W., Xia, K., Qiao, X.B., Xu, X.J., 2004. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. J. Chem. Inf. Comput. Sci. *44* (5), 1585–1600. https://doi.org/10.1021/ci049884m.

Jaganathan, K., Tayara, H., Chong, K.T., 2022. An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors. Pharmaceutics 14 (4), 832. https://doi.org/10.3390/pharmaceutics14040832.

Jain, S., Siramshetty, V.B., Alves, V.M., Muratov, E.N., Kleinstreuer, N., Tropsha, A., Nicklaus, M.C., Simeonov, A., Zakharov, A.V., 2021. Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. J. Chem. Inf. Model. 61 (2), 653–663. https://doi.org/10.1021/acs.jcim.0c01164.

Jiang, J., Wang, R., Wang, M., Gao, K., Nguyen, D.D., Wei, G.W., 2020. Boosting tree-assisted multitask deep learning for small scientific datasets. J. Chem. Inf. Model. 60 (3), 1235–1244. https://doi.org/10.1021/acs.jcim.9b01184.

Jillella, G.K., Ojha, P.K., Roy, K., 2021. Application of QSAR for the identification of key molecular fragments and reliable predictions of effects of textile dyes on growth rate and biomass values of Raphidocelis subcapitata. Aquat. Toxicol. *238*, 105925 https://doi.org/10.1016/j.aquatox.2021.105925.

Kar, S., Leszczynski, J., 2020. Is intraspecies QSTR model answer to toxicity data gap filling: Ecotoxicity modeling of chemicals to avian species. Sci. Total Environ. 738, 139858 https://doi.org/10.1016/j.scitotenv.2020.139858.

Kar, S., Sanderson, H., Roy, K., Benfenati, E., Leszczynski, J., 2020. Ecotoxicological assessment of pharmaceuticals and personal care products using predictive toxicology approaches. Green. Chem. *22* (5), 1458–1516. https://doi.org/10.1039/C9GC03265G.

Karpov, P., Godin, G., Tetko, I.V., 2020. Transformer-CNN: swiss knife for QSAR modeling and interpretation. J. Chemin.-. *12* (1), 12. https://doi.org/10.1186/s13321-020-00423-w.

Khan, K., Roy, K., Benfenati, E., 2019. Ecotoxicological QSAR modeling of endocrine disruptor chemicals. J. Hazard. Mater. *369*, 707–718. https://doi.org/10.1016/j.jhazmat.2019.02.019.

Khan, K., Roy, K., 2019. Ecotoxicological QSAR modelling of organic chemicals against Pseudokirchneriella subcapitata using consensus predictions approach. SAR QSAR Environ. Res. *30* (9), 665–681. https://doi.org/10.1080/1062936X.2019.1648315.

Kim, J.H., 2019. Multicollinearity and misleading statistical results. Korean J. Anesth. 72 (6), 558–569. https://doi.org/10.4097/kja.19087. Epub 2019 Jul 15. PMID: 31304696; PMCID: PMC6900425.

Krishna, J.G., Ojha, P.K., Kar, S., Roy, K., Leszczynski, J., 2020. Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. Nano Energy *70*, 104537. https://doi.org/10.1016/j.nanoen.2020.104537.

Kumar, V., Gupta, M.K., Singh, G., Prabhakar, Y.S., 2013. CP-MLR/PLS directed QSAR study on the glutaminyl cyclase inhibitory activity of imidazoles: rationales to advance the understanding of activity profile. J. Enzym. Inhib. Med. Chem. *28* (3), 515–522. https://doi.org/10.3109/14756366.2011.654111.

Kumar, A., Ojha, P.K., Roy, K., 2023. QSAR modeling of chronic rat toxicity of diverse organic chemicals. Comput. Toxicol. *26*, 100270 https://doi.org/10.1016/j.comtox.2023.100270.

Kumar, A., Ojha, P.K., Roy, K., 2024. Chemometric modeling of the lowest observed effect level (LOEL) and no observed effect level (NOEL) for rat toxicity. Environ. Sci.: Adv. https://doi.org/10.1039/D3VA00265A.

Li, J., Wu, Y., Yu, X., Zheng, X., Xian, J., Li, S., Shi, W., Tang, Y., Chen, Z.S., Liu, G., Yao, S., 2022. Isolation, bioassay and 3D-QSAR analysis of 8-isopentenyl flavonoids from Epimedium sagittatum maxim. as PDE5A inhibitors. Chin. Med. *17* (1), 1–18.

Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H., Tropsha, A., 2012. Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (10), 2570–2578. https://doi.org/10.1021/ci300338w.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Et. Biophys. Acta (BBA)-Protein Struct. 405 (2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicol. QSARs 801–820.

Morales Helguera, A., Perez Gonzalez, M., Dias Soeiro Cordeiro, M.N., Cabrera Perez, M. A., 2008. Quantitative structure− carcinogenicity relationship for detecting structural alerts in nitroso compounds: species, rat; sex, female; route of administration, Gavage. Chem. Res. Toxicol. 21 (3), 633–642. https://doi.org/10.1021/tx700336n.

Mostafalou, S., Abdollahi, M., 2013. Pesticides and human chronic diseases: evidences, mechanisms, and perspectives. Toxicol. Appl. Pharmacol. *268* (2), 157–177. https://doi.org/10.1016/j.taap.2013.01.025.

Mukherjee, R.K., Kumar, V., Roy, K., 2021. Ecotoxicological QSTR and QSTTR modeling for the prediction of acute oral toxicity of pesticides against multiple avian species. Environ. Sci. Technol. *56* (1), 335–348. https://doi.org/10.1021/acs.est.1c05732.

Nicolotti, O., Benfenati, E., Carotti, A., Gadaleta, D., Gissi, A., Mangiatordi, G.F., Novellino, E., 2014. REACH and in silico methods: an attractive opportunity for medicinal chemists. Drug Discov. Today 19, 1757–1768. https://doi.org/10.1016/j.drudis.2014.06.027.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. Chemin.-. *3* (1), 1–14. https://doi.org/10.1186/1758-2946-3-33.

OECD; Environment Health and Safety Publications Series on Testing and Assessment No. 69. Guidance Document On The Validation Of (Quantitative) Structure-Activity

Relationship [(Q) SAR] Models; 2007. Accessed from http://search.oecd.org/officialdocuments/displaydo cumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en (accessed September 15, 2014).

OECD, 2010. Test No. 223: Avian Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, Effects on Biotic Systems. OECD Publishing, Paris, France.

Ojha, P.K., Roy, K., 2011. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. Chemom. Intell. Lab. Syst. *109* (2), 146–161. https://doi.org/10.1016/j.chemolab.2011.08.007.

Pandey, S.K., Ojha, P.K., Roy, K., 2020. Exploring QSAR models for assessment of acute fish toxicity of environmental transformation products of pesticides (ETPPs) (No.). Chemosphere 252, 126508. https://doi.org/10.1016/j.chemosphere.2020.126508.

Park, H.S., Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. *36* (2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039.

Paul, R., Chatterjee, M., Roy, K., 2022. First report on soil ecotoxicity prediction against Folsomia candida using intelligent consensus predictions and chemical read-across. Environ. Sci. Pollut. Res. *29* (58), 88302–88317. https://doi.org/10.1007/s11356-022-21937-w.

Podder, T., Kumar, A., Bhattacharjee, A., Ojha, P.K., 2023. Exploring regression-based QSTR and i-QSTR modeling for ecotoxicity prediction of diverse pesticides on multiple avian species. Environ. Sci.: Adv. 2 (10), 1399–1422. https://doi.org/10.1039/D3VA00163F.

Roy, K., Ambure, P., Kar, S., 2018. How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? ACS Omega *3* (9), 11392–11406. https://doi.org/10.1021/acsomega.8b01647.

Roy, K., Das, R.N., 2013. QSTR with extended topochemical atom (ETA) indices. 16. Development of predictive classification and regression models for toxicity of ionic liquids towards Daphnia magna. J. Hazard. Mater. 254, 166–178. https://doi.org/10.1016/j.jhazmat.2013.03.023.

Roy, K., Kar, S., Ambure, P., 2015b. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. *145*, 22–29. https://doi.org/10.1016/j.chemolab.2015.04.013.

Roy, K., Kar, S., Das, R.N., 2015a. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.

Roy, P.P., Leonard, J.T., Roy, K., 2008. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemom. Intell. Lab. Syst. *90* (1), 31–42. https://doi.org/10.1016/j.chemolab.2007.07.004.

Roy, J., Roy, K., 2021. Assessment of toxicity of metal oxide and hydroxide nanoparticles using the QSAR modeling approach. Environ. Sci.: Nano *8* (11), 3395–3407. https://doi.org/10.1039/D1EN00733E.

Samanipour, S., O'Brien, J.W., Reid, M.J., Thomas, K.V., Praetorius, A., 2022. From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization. Environ. Sci. Technol. *57* (46), 17950–17958. https://doi.org/10.1021/acs.est.2c07353.

Saxena, A.K., Devillers, J., Bhunia, S.S., Bro, E., 2015. Modelling inhibition of avian aromatase by azole pesticides. SAR QSAR Environ. Res. *26* (7-9), 757–782. https://doi.org/10.1080/1062936X.2015.1090749.

Schultz, T.W., Yarbrough, J.W., Koss, S.K., 2006. Identification of reactive toxicants: Structure–activity relationships for amides. Cell Biol. Toxicol. 22, 339–349. https://doi.org/10.1007/s10565-006-0079-z.

Senanayake, N.M., Carter, J.L.W., Bowman, C.L., et al., 2022. A data-driven framework to select a cost-efficient subset of parameters to qualify sourced materials. Integr. Mater. Manuf. Innov. 11, 339–351. https://doi.org/10.1007/s40192-022-00266-3.

SIMCA-P, U.M.E.T.R.I.C.S., 2002. 10.0, info@ umetrics. com: www. umetrics. com, Umea.

Singh, K.P., Gupta, S., Basant, N., Mohan, D., 2014. QSTR modeling for qualitative and quantitative toxicity predictions of diverse chemical pesticides in honey bee for regulatory purposes. Chem. Res. Toxicol. *27* (9), 1504–1515. https://doi.org/10.1021/tx500100m.

Song, I.S., Cha, J.Y., Lee, S.K., 2011. Prediction and analysis of acute fish toxicity of pesticides to the rainbow trout using 2D-QSAR. Anal. Sci. Technol. 24 (6), 544–555. https://doi.org/10.5806/AST.2011.24.6.544.

Speck-Planche, A., 2020. Multi-scale QSAR approach for simultaneous modeling of ecotoxic effects of pesticides. Ecotoxicol. QSARs 639–660. https://doi.org/10.1007/978-1-0716-0150-1_26.

Speck-Planche, A., Kleandrova, V.V., Luan, F., Cordeiro, M.N.D., 2012. Predicting multiple ecotoxicological profiles in agrochemical fungicides: a multi-species chemoinformatic approach. Ecotoxicol. Environ. Saf. 80, 308–313. https://doi.org/10.1016/j.ecoenv.2012.03.018.

Speck-Planche, A., Natalia Dias Soeiro Cordeiro, M., Guilarte-Montero, L., Yera-Bueno, R., 2011. Current computational approaches towards the rational design of new insecticidal agents. Curr. Comput. -Aided Drug Des. 7 (4), 304–314. https://doi.org/10.2174/157340911798260359.

Speck-Planche, A., Guilarte-Montero, L., Yera-Bueno, R., Rojas-Vargas, J.A., García-López, A., Uriarte, E., Molina-Pérez, E., 2011. Rational design of new agrochemical fungicides using substructural descriptors. Pest Manag. Sci. 67 (4), 438–445. https://doi.org/10.1002/ps.2082.

Todeschini, R., Ballabio, D., Grisoni, F., 2016. Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models. J. Chem. Inf. Model. *56* (10), 1905–1913. https://doi.org/10.1021/acs.jcim.6b00277.

Vervloet, M., 2019b. Modifying Phosphate toxicity in chronic kidney disease. Sep 9 Toxins 11 (9), 522. https://doi.org/10.3390/toxins11090522.

Vervloet, M., 2019a. Modifying phosphate toxicity in chronic kidney disease. Toxins *11* (9), 522. https://doi.org/10.3390/toxins11090522.

Wang, L.L., Ding, J.J., Pan, L., Fu, L., Tian, J.H., Cao, D.S., Jiang, H., Ding, X.Q., 2021. Quantitative structure-toxicity relationship model for acute toxicity of organophosphates via multiple administration routes in rats and mice. J. Hazard. Mater. 401, 123724 https://doi.org/10.1016/j.jhazmat.2020.123724.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. *58* (2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Yu, Y., Zhu, Y., Yang, J., Zhu, W., Zhou, Z., Zhang, R., 2021. Effects of Dufulin on Oxidative Stress and Metabolomic Profile of Tubifex. Metabolites *11* (6), 381. https://doi.org/10.3390/metabo11060381.

Zhang, C., Cheng, F., Sun, L., Zhuang, S., Li, W., Liu, G., Lee, P.W., Tang, Y., 2015. In silico prediction of chemical toxicity on avian species using chemical category approaches. Chemosphere *122*, 280–287. https://doi.org/10.1016/j.chemosphere.2014.12.001.