Chemometrics-driven toxicity prediction and prioritization of diverse pesticides on birds for addressing hazardous effects

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

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

SHUBHA DAS

B. PHARM.

Registration No: **163665 of 2022-23**

Examination Roll No: **M4PHG24001**

Class roll number: **002211402022**

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 – 700032

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 work on **"Chemometrics-driven toxicity prediction and prioritization of diverse pesticides on birds for addressing hazardous effects"**.

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

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

NAME: **SHUBHA DAS**

EXAMINATION ROLL NUMBER: **M4PHG24001**

REGISTRATION NUMBER: **163665 of 2022-23**

THESIS TITLE: **"Chemometrics-driven toxicity prediction and prioritization of diverse pesticides on birds for addressing hazardous effects".**

SIGNATURE: Shubha Das

DATE: 27.08.2024

PLACE: Jadavpur University, Kolkata.

CERTIFICATE

Department of Pharmaceutical Technology

Jadavpur University

Kolkata - 700032

This is to certify that **Mr. SHUBHA DAS,** B. Pharm. (2018-22), has carried out the research work on the subject entitled ***"Chemometrics-driven toxicity prediction and prioritization of diverse pesticides on birds for addressing hazardous effects"*** *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 fulfilment 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 27/08/2024

DR. PROBIR KUMAR OJHA

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

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

27/8/24

(Prof. Dr. Amalesh Samanta)

Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

Prof. Amalesh Samanta, Ph.D.
Head
Dept. of Pharmaceutical Technology
Jadavpur University, Kolkata, India

Dipak Laha 28.8.24

(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 seniors Ankur Kumar, Arnab Bhattacharjee, Arkaprava Banerjee, Mainak Chatterjee. 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 friend Abhishek Samal and juniors Prodipta Bhattacharyya, and Pabitra Samanta. 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. Surajit Das, Mrs. Parul Das, and my sister Ms. Chaitali Das for all the love and inspiration without which my dissertation work would remain incomplete.

Shubha Das

Shubha Das

Examination Roll No: **M4PHG24001**

Department of Pharmaceutical Technology,

Jadavpur University,

Kolkata-700032

Preface

This dissertation is presented for the partial fulfilment of 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 PLS-based Quantitative Structure-Toxicity Relationship (QSTR) and q-RASTR (Quantitative Read-Across Structure-Toxicity Relationship) models using easily interpretable two-dimensional (2D) molecular descriptors for efficient prediction of toxicity of diverse organic compounds towards *Birds.* 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.

Quantitative Structure-Toxicity Relationship (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 and q-RASTR models were developed using different classes of simple 2D descriptors to estimate the toxicity of different organic compounds. We attempted to explore the toxicity profile of different organic pollutants to make a more realistic move toward 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. Chemometrics-driven prediction and prioritization of diverse pesticides on chickens for addressing hazardous effects on public health.

Study 2. First report on q-RASTR modeling of hazardous dose (HD_5) for acute toxicity of pesticides: An efficient and reliable approach towards safeguarding the sensitive avian species.

Study 3. Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across.

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

Chapter 6: References

Appendix: Reprints

In the "introduction" section, we have provided background information on various types of pesticides and toxicity toward bird species, humans, and the environment, as well as the different types of Quantitative Structures-Activity Relationship (QSAR) models. We have outlined the general QSAR procedure and conducted a brief survey of QSAR modeling for predicting the toxicity of chemicals and pharmaceuticals to humans and the environment. Additionally, we have discussed the applications of QSAR by governing and regulatory authorities are also discussed. The planned work has been discussed in the section of Present Work. The descriptors and methodologies have been given in the 'Materials and Methods' section while the results have been discussed in 'Results and Discussion' section. Finally, 'Conclusions' has been incorporated followed by 'References' and 'Reprints'. It is worth mentioning here that the author has already published the present work in referred journals like the *Journal of Hazardous Material* (Elsevier) and *Process Safety and Environmental Protection* (Elsevier) and also presented in different national and international seminars and conferences. Another research paper of the author has been communicated for publication in a journal. Reprints of the published papers and abstracts of the presentations have been enclosed.

Finally, the work done and presented in this dissertation constitutes a small part of the broad spectrum of envisaged works. Considering the stipulated time limit only some representative and relevant studies could be performed. Many other interesting aspects arising out of this work could have been investigated in a far more meaningful way, which can be planned in the future.

Abbreviations	Full forms	Abbreviations	Full forms
2-D	Two dimensional	**NOEL**	No observed effect level
AD	Applicability domain	**OECD**	Organization for Economic Co-operation and Development
Abs Max Pos-Max neg	The absolute difference between the MaxPos and MaxNeg values	**PPDB**	Pesticide properties database
BQ	Bobwhite quail	**PRESS**	Predictive residual sum of square
BSS	Best subset selection	**QSAR**	Quantitative structure-activity relationships
CA	Cluster analysis	**QSPR**	Quantitative structure-property relationship
CVsim	Coefficient of variation of the similarity values of the close source compounds	**QSTR**	Quantitative structure-toxicity relationship
CCC	Concordance correlation coefficient	**q-RASAR**	Quantitative Read Across Structure-Activity Relationship
CAS	Chemical abstracts service	**q-RASTR**	Quantitative Read Across Structure-Toxicity Relationship
EPA	Environmental protection agency	Q^2_{LOO}	Cross-validated correlation coefficient
EU	Europian union	**REACH**	Registration, Evaluation, Authorisation and Restrictions of Chemicals
ED	Euclidean distance	R^2_m	Root mean square
GA	Genetic algorithm	**RF**	Random forest
gm*Avg.Sim	Product of the values of gm and Avg. Sim	**RR**	Ridge regression
gm*SD Similarity	Product of the values of gm and SD similarity	**RMSE**	Root mean square error
HD	Hazardous dose	**SD**	Standard deviation
LOO	Leave one out	**SE**	Weighted standard error of the response values of the close source compounds
LV	Latent variable	**SAR**	Structure-Activity Relationship
LD	Lethal dose	**SVM**	Support vector machine
LC	Lethal concentration	**SVR**	Support vector regression
LOEL	Lowest observed effect level	**SMILES**	Simplified molecular input line entry system
LDA	Linear discriminant analysis	**SDEP**	standard deviation of error of prediction
MAE	Mean absolute error	**VIP**	Variable importance plot
MLR	Multiple linear regression	**WHO**	World health organization
MW	Molecular weight	$Y_{calc(train)}$	Calculated response value of training set
ML	Machine learning	$Y_{mean(train)}$	Average of all response of training set
MAPE	Mean absolute percentage error	$Y_{calc(test)}$	Calculated response value of test set

Contents

CHAPTER - 1

Introduction

1. INTRODUCTION

1.1 Toxicity

The term "toxicity" refers to the degree to which a chemical or specific combination of chemicals can harm an organism. In common use, the word is occasionally almost a synonym for poisoning. It is essential to understand that the effects of a toxin depend on the dosage. For example, drinking too much water can lead to water intoxication, while even highly toxic substances like snake venom have a threshold below which they do not cause harm [1]. Toxicity can manifest in various ways, such as disrupting the body's balance, causing irreversible damage to function or structure, or making an individual more susceptible to other chemicals, biological stress, and infections. Given that our society relies on various chemicals, itis crucial to understand how they interact with the environment and their potential toxic effects. Elevated levels of certain chemicals or prolonged exposure to them can result in significant harm to the affected organism, with the most severe outcome being potential death [2]. The severity of toxic effects depends on factors such as the type of chemical, its concentration, the duration of exposure, and the sensitivity of the organism. Some adverse effects may be subtle and go unnoticed, while others can be immediately life-threatening. Regulatory bodies and environmental agencies play a crucial role in monitoring and regulating the use of toxic chemicals to minimize the risks to human health and the environment.

1.1.1 Toxicity of pesticides

Pesticides are chemical compounds that are used to eliminate insects, rodents, fungi, and weeds. These consist of plant growth regulators, molluscicides, rodenticides, fungicides, insecticides, herbicides, nematicides, and other substances [3-4]. It plays important roles in commercial as well as food-based industrial processes, such as aquaculture, agriculture, food processing, and storage, and is typically employed to prevent infections spread by vectors [5]. Any living bodies, either animals or plants, which are harmful for humans or animals are known as pests. Pesticides are chemicals used to eradicate pests or stop them from growing. Various chemical compounds have been used since ancient times to control pests. Sulfur compounds and pyrethrum, a pesticide derived from the *Chrysanthemum cinerariaefolium* plant, have been utilized for over 2000 years [6-7]. The global pesticide consumption in 2019 was approximately 4.19 million metric tons, where China was by far the largest pesticide-consuming country (1.76 million metric tons), followed by the United States (408 thousand tons), Brazil (377 thousand tons), and Argentina (204 thousand tons) [8]. India is one of the major pesticide-producing countries in Asia, with an annual production of 90 thousand tons of organochlorine pesticides, including benzene hexachloride and DDT [9]. Pests, insects, diseases, and weeds can significantly reduce crop

production, making pesticides crucial for food production and processing. Warren (1998) [10] also noted a substantial increase in food production in the United States over the 20th century. Pesticides are used to increase agricultural output and food preservation while ignoring their associated risks. Overuse, exposure, and harmful consequences can all be mitigated by applying it judiciously and utilizing different pesticide categories (World Health Organization, 2009). Widespread pesticide usage has been associated with various detrimental effects, highlighting the need for effecting waste management strategies to address pesticide issues. Pesticide biodegradation offers an environmentally friendly solution for controlling pesticide pollution in the long term. Microorganisms play a significant role in breaking down pesticides and have various uses in promoting human welfare. Recent studies have shown that microorganisms isolated from sewage or soil have the potential to degrade pesticides. These microbes encompass bacterial, fungal strains, actinomycetes, algae, and more [11]. The entire process, including pesticide synthesis, manufacturing, environmental and health impacts, and pesticide biodegradation, is illustrated in **Figure 1.1**.

The use of pesticides has increased significantly in recent decades, particularly in agriculturally dependent developing countries. Due to the inherent characteristics, a significant portion of the applied dose continues to remain as remnants on crops and fields. 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. In light of the aforementioned, various regulatory authorities have emphasized the need for the toxicity evaluation of both new and existing pesticides [12]. The avian toxicity tests are essential for regulatory approval and licensing of the active ingredients of pesticides.

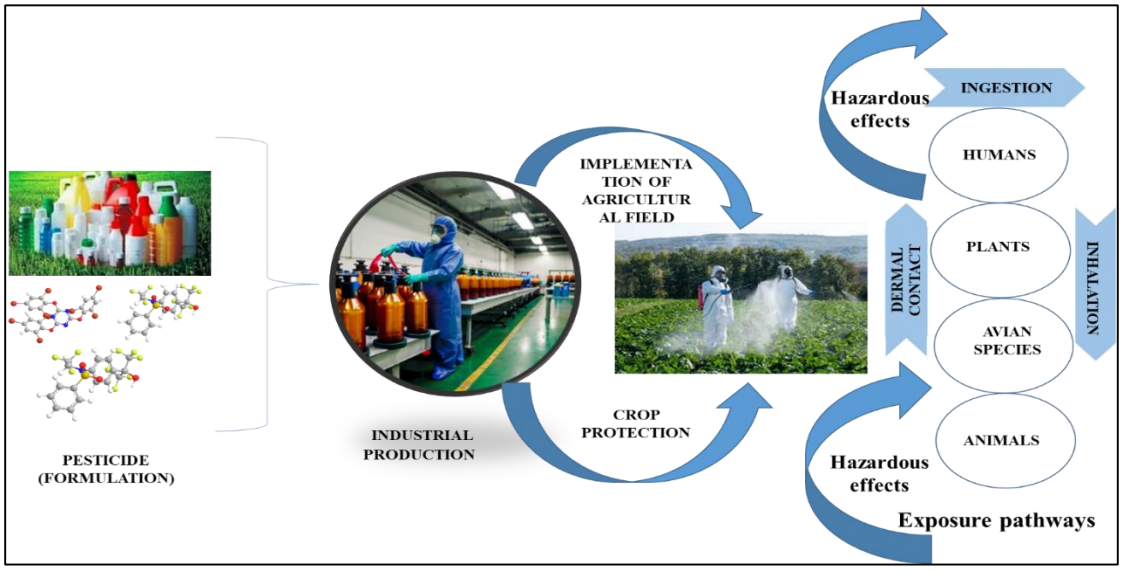


Figure 1.1. Thematic diagram of the synthesis, production, uses effects, and eco-friendly management of pesticides.

1.1.2 Classification of pesticides

Pesticides are a diverse group of substances that include insecticides, herbicides, fungicides, rodenticides, wood preservatives, garden chemicals, and household disinfectants. These chemicals are used to kill or protect against pests [13]. These pesticides differ in their physical, chemical, and identical properties from one class to other. Therefore, it is worthy to classify them based on their properties and study under their respective groups. Synthetic pesticides are manmade chemicals and do not occur in nature. They are categorized into various classes depending on the needs. Currently, there are three popular methods of pesticide classification suggested by Drum [14].These three popular methods of pesticide include (i) classification based on the mode of entry, (ii) classification based on pesticide function and the pest organism they kill, and (iii) classification based on the chemical composition of the pesticide.

1.1.2.1 Classification based on the mode of entry

The ways pesticides come in contact with or enter the target are called modes of entry. These include systemic, contact, stomach poisons, fumigants, and repellents.

1.1.2.1.1 Systemic pesticides

Systemic pesticides are chemicals that are absorbed by plants or animals and then spread to untreated parts of the organism. Systemic herbicides can move through the plant to reach areas that were not directly treated, such as leaves, stems, or roots, and effectively kill weeds even with partial spray coverage. They have the ability to penetrate plant tissues and move through the plant's vascular system to target specific pests. Some systemic insecticides are also applied to animals and move through their bodies to control pests like warble grubs, lice, or fleas. When applied to the root zone, systemic pesticides will travel throughout the plant, but if applied to the leaves, they will not move throughout the plant. Additionally, a few pesticides are considered locally systemic, affecting only a short distance in a plant from the point of contact. Examples of systemic pesticides include 2,4-dichlorophenoxyacetic acid (2,4-D) and glyphosate.

1.1.2.1.2 Non-systemic (contact) pesticides

Non-systemic pesticides, also known as contact pesticides, only work when they come into direct contact with the target pests. They enter the pests' bodies through the skin and cause death by poisoning. These pesticides do not spread through the plant's vascular system. Some examples of contact pesticides are paraquat and diquat dibromide.

1.1.2.1.3 Stomach poisoning and stomach toxicants

Pesticides that cause stomach poisoning enter pests' bodies through their mouth and digestive systems, leading to their death. Pests ingest these stomach poisons while feeding on leaves and other parts of the plants. The toxins can also be absorbed into the insect's body through the mouth

and digestive system. This method is particularly effective in controlling disease-carrying insects, such as mosquitoes and black flies, by applying bacteria or toxins to the water where larvae feed. These insecticides work by destroying the midgut or stomach of the larvae, ultimately killing them. An example of such a pesticide is malathion.

1.1.2.1.4 Fumigants

Fumigants are insecticides that function by vaporizing the target pests, possibly killing them. When these herbicides are used, toxic gasses are produced. Through spiracles, these vaporized insecticides enter the pests' bodies through their tracheal system (respiratory system) and poison them, killing them. When compressed under extreme pressure, some of their active constituents are liquids; yet, upon release, they transform into gasses. Other active chemicals are not formulated under pressure and, when enclosed in a regular container, are volatile liquids. Fruits, vegetables, and cereals are treated with fumigants to get rid of pests from stored goods. They play a crucial role in soil pest management as well.

1.1.2.1.5 Repellents

Repellents do not kill but are distasteful enough to keep pests away from treated areas/commodities. They also interfere with pest's ability to locate crops.

1.1.2.2 Classification based on pesticide function and pest organism they kill

In this method, pesticides are categorized based on the specific pest organism they target and are given names that reflect their activity. The group names of these pesticides come from the Latin word "cide," meaning "kill" or "killer," and are used as suffixes after the corresponding name of the pests they kill (**Table 1.1**). It's important to note that not all the pesticides end with the suffix "cide". Additionally, some pesticides are classified based on their function, such as growth regulators, defoliants, desiccants, repellents, attractants, and chemosterilants.

Table 1.1 Pesticide classification by target pests.

Type of pests	Function	Example
Insecticides	Kill insects and other arthropods	Aldicarb
Fungicides	Kill fungi (including blights, mildews, molds, and rusts)	Azoxystrobin
Herbicides	Kill weeds and other plants that grow where they are not wanted	Atrazine
Algaecides	Control or kill growth of algae	Copper sulfate
Bactericides	Kill bacteria or act against bacteria	Copper complexes
Rodenticides	Control mice and other rodents	Warfarin
Lervicides	Inhibits growth of larvae	Methoprene

Repellents	Repel pests by its taste or smell	Methiocarb
Virucides	Acts against viruses	Scytovirin
Avicides	Kill birds	Avitrol
Nematicides	Kill nematodes that act as parasites of plants	Aldicarb
Molluscicides	Inhibit or kill molluscs *i.e.* snails usually disturb the growth of plants or crops	Metaldehyde

1.1.2.3 Classification based on chemical composition of pesticides

The most common and useful method of classifying pesticides is based on their chemical composition and the nature of their active ingredients. This classification provides clues about the efficacy, as well as the physical and chemical properties of the pesticides. The information on the chemical and physical characteristics of pesticides is very useful for determining the mode of application, the precautions that need to be taken during application, and the application rates. Pesticides are classified into four main groups based on their chemical composition such as organochlorines, organophosphorus, carbamates, and pyrethroids [15]. The chemical-based classification of pesticides is rather complex. Modern pesticides are generally organic chemicals, including those of both the synthetic and plant origin, although some inorganic compounds are also used. The classification of pesticides is presented in **Figure 1.2.**

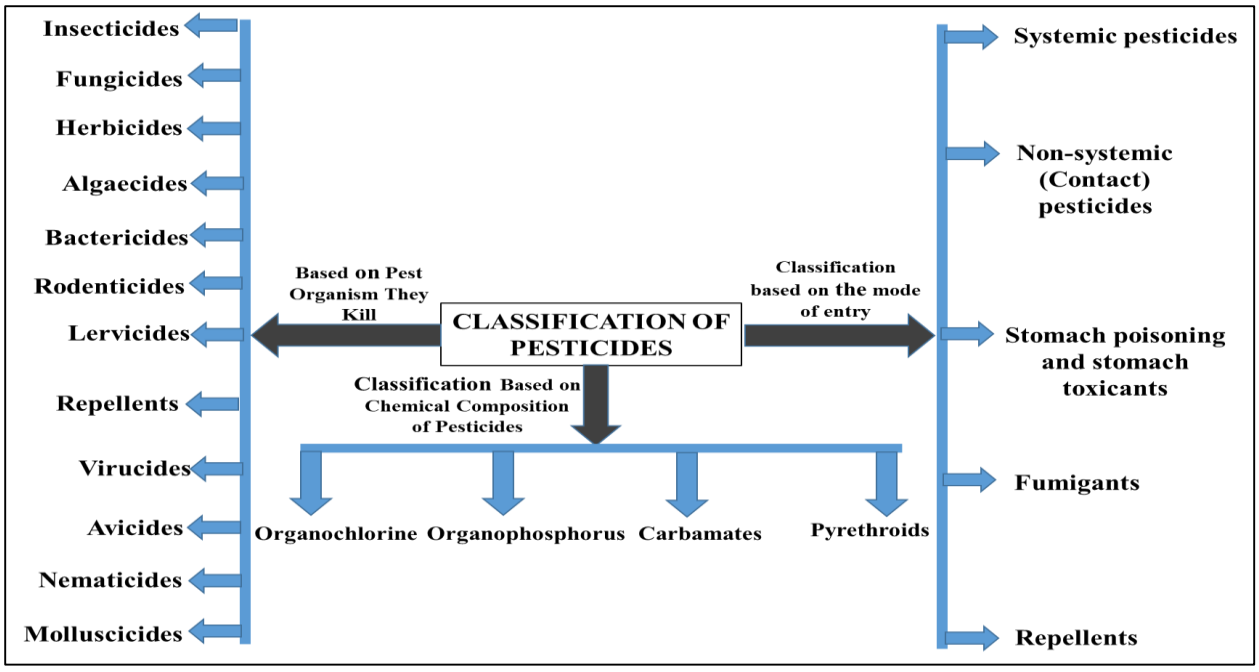


Figure 1.2. Classification of pesticides.

1.1.3 Effects of pesticides

Despite the beneficial results of using pesticides in agriculture and the public health sector, their use also invites harmful environmental and public health effects. Pesticides have a high biological activity and toxicity, making them unique among environmental contaminants. Most pesticides do not differentiate between pests and other incidental life forms, posing potential harm to humans, animals, and the environment when used incorrectly. It is estimated that 5,000–20,000 people die and about 500,000 to 1 million people are poisoned every year by pesticides [16]. At least half of the affected individuals and 75% of those who die due to pesticides are agricultural workers, while the rest are poisoned due to consuming contaminated food.

1.1.3.1 Potential impact on human health

It's important to be aware that pesticides can enter the human body in several ways. These include inhalation of polluted air, dust, and vapor containing pesticides, oral exposure through consuming contaminated food and water, and dermal exposure through direct contact with pesticides [17]. Pesticides are often sprayed onto fruits and vegetables and can end up in the soil and groundwater, which may then contaminate drinking water. Additionally, pesticide spray can drift and pollute the air. The harmful impact on human health depends on factors such as the toxicity of the chemicals, the duration, and the magnitude of exposure [18]. The toxicity of chemicals is influenced by the nature of the toxicant, routes of exposure (oral, dermal, and inhalation), dose, and the organism. Toxicity can manifest as either acute or chronic. Acute toxicity refers to the rapid development of harmful effects within a few hours or a day after absorption, while chronic toxicity results from long-term exposure. The toxicity of insecticides is often measured in terms of lethal dose 50% (LD_{50}) or lethal concentration 50% (LC_{50}). LD_{50} is the single exposure dose per unit weight of the organism required to kill 50% of the test population, expressed in milligrams per kilogram of body weight. LC_{50} is the concentration of the chemical in the external medium (usually air or water) causing 50% mortality in the test population and is expressed in parts per million (ppm).

1.1.3.2 Impacts on the environment

The widespread use and disposal of pesticides by farmers, institutions, and the general public create multiple potential sources of pesticides in the environment. These substances can have far-reaching effects, spreading through the air, being absorbed in the soil, or dissolving in water and ultimately reaching a much larger area than originally intended. Once released into the environment, pesticides can take on different fates. For instance, when pesticides are

sprayed on crops, they may travel through the air and end up in other parts in the environment, such as soil or water. Pesticides applied directly to the soil may wash off and make their way to nearby surface water through runoff, or seep down through the soil and reach lower layers and groundwater [19]. The impact of pesticides on the environment can vary from minor disruptions in the normal functioning of ecosystems to loss of species diversity. Pesticides can have both long-term residual effects and immediate, severe impacts. For example, many organochlorine pesticides persist in the environment for long periods, leading to contamination of groundwater, surface water, food products, air, and soil.

1.1.3.3 Impacts on avian species

Avian species hold a unique position in the ecosystem as one of the most diverse and evolutionary successful groups, especially in the tropics. Unfortunately, Europe has observed a significant loss of around 550 million birds over the last forty years. This decline is primarily attributed to the widespread use of pesticides and fertilizers in agriculture, as well as the effects of climate change, changes in forest cover, and urbanization. Pesticides are important for managing pests and improving crop productivity in modern agriculture, but they also pose risks to non-target organisms like birds, raising significant environmental concerns [20]. Birds play crucial roles in ecosystems by contributing to pest control, pollination, and seed dispersal, which are essential for biodiversity and environmental health [21]. However, exposure to pesticides can cause acute toxicity and long-term declines in avian populations, disrupting ecological balance and biodiversity [22]. Therefore, it's crucial to assess pesticide toxicity to manage the associated risks to avian species and maintain ecosystem balance.

1.1.4 Quantitative Structure-activity/property/toxicity relationship (QSAR/QSPR/QSTR) modeling and other *in-silico* approaches

The investigation of the properties of chemicals for toxicological prediction is often advised by governing bodies such as the Environmental Protection Agency (EPA), Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), European Chemicals Bureau (ECB), and European Food Safety Authority (EFSA). Computational tools such as read-across and QSAR are recommended for this purpose [23]. QSAR, in particular, is widely used to predict the toxicity of test chemicals. This technique involves developing a scientific model from a series of compounds with experimentally derived endpoint values. Due to its reproducibility, simplicity, and transferability, QSAR is widely employed. Current chemical risk assessment often relies on similarity-driven methods like Read-Across, which assumes that compounds with similar structures have comparable biological activities [24]. This makes emerging similarity-driven systems more suitable for consistent compound prediction. While

Read-Across predicts probe compounds more reliably than QSAR models, it has limitations in interpreting essential features [25]. To address this issue, a novel approach called Read-Across Structure-Activity Relationship (RASAR) was introduced, which combines the benefits of QSAR and Read-Across algorithms, resulting in better predictive ability and reduced mean absolute error (MAE) [26]. They utilized classification-based models that produced predictions on a graded scale. Banerjee and Roy [27] introduced q-RASAR modeling with descriptors based on similarity and error measures. The q-RASAR methodology utilizes descriptors based on similarity and error measures to develop simple, convenient, interpretable, and reproducible models with better predictivity. These q-RASAR models can be developed using statistical techniques like MLR, PLS, and other sophisticated machine learning (ML) techniques. Machine learning, which uses various algorithms for building models and making predictions using data, has shown potential for experimental studies. Commonly used machine learning algorithms include support vector machines (SVM), artificial neural networks (ANN), and others [28-29].

1.1.4.1 What is QSAR/QSPR/QSTR modeling?

QSAR modeling involves creating a mathematical relationship between a chemical response and the quantitative chemical attributes defining the features of related molecules. This study aims to establish a correlation between the behavior of a chemical (the "endpoint") and the quantitative chemical attributes that can be derived from the chemical structures through experimental or theoretical methods. Depending on the nature of the response being modeled, QSAR falls into three major classes: quantitative structure-property/activity/toxicity relationship (QSPR/QSAR/QSTR) studies, which consider modeling physicochemical property, biological activity, and toxicological data, respectively. The basic formalism of QSAR model can be mathematically represented as follows,

$$\text{Biological activity/property/toxicity} = f(\text{Chemical attributes}) \qquad (1.1)$$

The term "chemical attribute" refers to the features that define the behavior of the analyzed chemical compounds and control the response under study. These attributes are precise quantitative chemical information that can be derived from experimental analysis or theoretical algorithms. Considering the employment of a series of chemical information in presence-absence of physicochemical features, the QSAR equation for a specific response can be mathematically stated as follows:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + a_nX_n \qquad (1.2)$$

Since we are talking in terms of a mathematical correlation, such equations are better explained in terms of variables. Here, Y is the dependent variable representing the response being modeled, i.e., activity/property/toxicity while X_1, X_2,…, X_n are the independent variables denoting different structural features or physicochemical properties in the form of numerical quantities or descriptors and a_1, a_2,…, a_n are the contributions of individual descriptors to the response with a_0 being a constant. Hence, we can see that the physicochemical properties can not only be employed as a dependent or response variable giving a structure-property relationship, i.e., QSPR, but they might also be used as independent or predictor variables. QSAR studies may also use one response parameter as a predictor variable for modeling another type of endpoint, resulting in quantitative activity-activity relationship (QAAR), quantitative toxicity-toxicity relationship (QTTR), or quantitative property-property relationship (QPPR) modeling, as appropriate. While the modeled response should be quantitative to develop a regression model, it may also be categorical entities used for classification models. However, the predictor variables in QSAR modeling should always be quantitative. QSAR analysis focuses on quantifying chemical information and developing an interpretative relationship for a given response [30].

1.1.4.2 QSAR and regulatory perspectives

The use of QSAR techniques for developing predictive models is recognized and recommended by several international regulatory bodies. Different regulatory bodies address the following aspects for performing risk assessment of chemicals,

1. Assessment of chemical hazard: This includes identifying and characterizing the dose-response of hazards, as well as classifying and labelling the chemicals.
2. Assessment of exposure.
3. Assessment of hazard and exposure.
4. Identification of persistent, bioaccumulative, and toxic (PBT) as well as very persistent and very bioaccumulative (vPvB) chemicals.

Determining chemical toxicity typically involves a significant number of animal experiments to generate reliable chemical response data. Therefore, it is crucial for any hazard assessment strategy to seek suitable alternative methods to reduce animal experimentation. QSAR plays a significant role in this context, as it requires a comparatively smaller amount of response data and can predict responses for a large number of chemicals. The QSAR technique aligns with the '3R' principle of Russell and Burch – replacement, reduction, and refinement of animals in biological experiments. The major advantages of QSAR modeling in regulatory assessment include prioritization of chemicals and filling of data gaps. Furthermore, modeling of categorical data is important, as the toxicological response of chemicals can be categorized into groups or

classes, signifying different levels of hazards such as high, moderate, low, etc. Regulatory agencies advocating the use of QSAR as an alternative strategy to animal experiments include the European Centre for the Validation of Alternative Methods (ECVAM) of the European Union, the Office of Toxic Substances of the US Environmental Protection Agency (US-EPA), the Agency for Toxic Substances and Disease Registry (ATSDR), and the Council for International Organizations of Medical Sciences. The European Commission introduced the REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) regulations in 2006, aimed at systematically evaluating the toxicological hazards of existing and new chemicals (imported or produced) and identified QSAR as an alternative method for toxicity testing of animals. The Organization of Economic Cooperation and Development (OECD) proposed a set of five-point guidelines in 2004 for the development and validation of predictive QSAR models by its member countries. Over time, QSAR studies have become an essential part of global regulatory assessments, and various countries have developed their own 'expert systems' for determining chemical hazards. Expert systems are computational applications providing subject-matter expertise to non-experts by using logical reasoning. Different expert systems contain models on toxicological endpoints prepared and maintained by professional personnel, serving as trusted systems with a suitable user interface to easily test the toxicity or categorical hazard of any unknown or new chemical using the existing knowledge base.

1.1.4.3 Applications of QSAR

Chemicals are essential for a wide range of applications, from industrial and laboratory processes to household uses. QSAR is a valuable approach for monitoring the activity, properties, and toxicity of chemicals, with extensive applications across various fields. By fine-tuning the behavior of chemicals, QSAR can produce positive results for a large class of chemicals, including pharmaceuticals, agrochemicals, perfumeries, solvents, and more. The potential application of the QSAR technique is vast, as it can model chemicals in three main categories: those with health benefits (drugs, pharmaceuticals, food ingredients), those involved in industrial and laboratory processes (solvents, reagents), and those with hazardous outcomes (persistent organic pollutants, toxins, carcinogens). In addition to modeling biological activity and toxicity, QSAR is also used in the drug design process to monitor the pharmacokinetic profile of potential drug candidates, enhancing the efficacy of designed compounds within the biological system.

When assessing the toxicity of chemicals, two options are commonly considered: systematic toxicity evaluation and monitoring of ecotoxicological hazards. Drugs and pharmaceuticals can pose toxicity to specific organ systems (e.g. hepatotoxicity, nephrotoxicity, cardiovascular toxicity) and can also be concerning from an environmental perspective, as even trace amounts

of these compounds in wastewater streams can damage ecosystems [31]. Physiologically based pharmacokinetic (PBPK)
modeling is another area of interest, involving the modeling of chemicals such as volatile organic compounds (VOCs) using physicochemical and biochemical parameters.

In light of the growing health and environmental concerns, modern technologies are focused on establishing a sustainable and green ecosystem that promotes environmental friendliness in terms of efficiency, effectiveness, and safety. QSAR plays an encouraging role in achieving this environmental sustainability through the design and development of process-specific chemicals with reduced or no hazardous outcomes.

1.1.4.4 Descriptor

A QSAR model can be represented as a straightforward mathematical formula that correlates the physical, chemical, biological, and toxicological characteristics of molecules using a variety of quantitative factors that are obtained computationally or experimentally and are referred to as "descriptors." A number of chemometric techniques are used to link the descriptors with the experimental properties (response) in order to produce a statistically significant QSAR model. "Terms that characterize specific information of a studied molecule" are known as molecular descriptors. In order to correlate chemical structure with different physical attributes, chemical reactivity, or biological activity, these are the "numerical values associated with the chemical constitution." The resulting equation ought to offer substantial understanding of the fundamental structural requirements of the molecules that support the examined molecules' biological response [32]. In other words, the response of a chemical can be mathematically presented as the function of descriptors (**Eq. 1.3**).

Response (activity/property/toxicity)
= *f* (Molecular information extracted using chemical structure or physicochemical property)
= *f* (Descriptors) (1.3)

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

1. A descriptor should be relevant to a broad class of compounds.
2. A descriptor needs to show a negligible association with other descriptors and a correlation with the biological reactions under study.
3. The descriptor should be quickly calculated and unaffected by experimental characteristics.
4. Even with minor structural variations, a descriptor should yield distinct values for molecules with dissimilar structures.
5. Physical interpretability of a descriptor is necessary to identify the chemicals under study and identify their query properties.

1.1.4.5 Types of descriptors

Descriptors can be of different types depending on the method of their computation or determination: physicochemical (hydrophobic, steric, or electronic), structural (frequency of occurrence of substructure), topological, electronic (molecular orbital calculations), geometric (molecular surface area calculation), or simple indicator parameters (dummy variable). In a broader perspective descriptor, (basically physiochemical descriptors) can be classified can be two major groups 1) Substituent constant and 2) whole molecular descriptors [33,34]. Substituent constants are physiological descriptors which are deigned based on factors, which govern the physicochemical properties of chemical entities. Whole molecular descriptors are expansions of the substituent constant approach, but many of them are also derived from experimental approaches.

The descriptor may also be classified based on the dimensions. Different types of descriptors employed in the QSAR study are represented in **Table 1.2.**

Table 1.2. Different descriptors employed in the QSAR study based on dimensions.

Dimension of descriptors	Parameters
0D-descriptors	Constitutional indices, molecular property, and atom and bond count.
1D-descriptors	Fragment count, Fingerprints.
2D-descriptors	Topological parameters, structural parameters, physiochemical parameters, including thermodynamic descriptors.
3D-descriptors	Electronic parameters, spatial parameters, molecular shape analysis parameters, molecular field analysis parameters, and receptor surface analysis parameters.
4D-descriptors	Volsurf, GRID, Raptor, etc. derived descriptors
5D-descriptors	These descriptors considered induced-fit parameters and aimed to establish a ligand-based virtual or pseudoreceptor model. These can be explained as 4D-QSAR+ explicit representations of different induced fit models. Example- flexible protein docking
6D-descriptors	These are derived using representation of various solvation circumstances along with the information obtained from 5D descriptors. They can be explained as 5D-QSAR+ simultaneous consideration of different solvation models. Example- Quasar
7D- descriptors	They comprise real receptor or target based receptor model data.

1.1.4.6 Strategy for development of quantitative structure activity/property/toxicity relationship

QSAR steps

The strategy for developing quantitative structure-activity relationship (QSAR) in drug design involves multiple iterative steps based on statistical experimental design and multivariate data analysis. The ultimate goal is to design compounds or predict the toxicity of chemicals.

I. Generation of molecular descriptors from chemical structures

The chemical structures typically don't contain explicit information related to activity. This information needs to be extracted from the structure. Calculating descriptor values is generally straightforward due to the availability of many commercial and academic computer-aided molecular design (CAMD) packages that handle this calculation with ease. Different rationally designed molecular descriptors highlight various chemical properties present in the molecule's structure, and only those properties may have a more direct correlation with the activity.

II. Feature selection

In many applications, numerous molecular descriptors can be generated, often numbering in the hundreds or thousands. However, only a few of them are substantially correlated with the activity being studied. Additionally, many descriptors are correlated with each other, which can have adverse effects on various aspects of QSAR analysis. Some statistical methods require a significantly larger number of compounds than descriptors. Therefore, working with extensive descriptor sets necessitates large datasets.

III. Series design (selection of training set)

The selection of compounds for the training set is crucial in QSAR analysis. The most effective approach for selecting the training set is to consider relevant physicochemical descriptors and the principle of structural similarity. This process operates on the assumption that a molecule which is structurally similar to the molecules in the training set will be predicted accurately. This is because the model has captured common features of the training set molecules and is able to recognize them in the new molecule.

IV. Model construction

After selecting the relevant features, the final stage of building a QSAR model involves a feature mapping process, also known as the parameter estimation problem. The objective is to establish a mathematical relationship and estimate the model parameters. A variety of mapping function families can be utilized, such as linear ones (e.g., multiple linear regression, stepwise regression, partial least square regression) and non-linear ones (e.g., artificial neural network, random forest). Various methods can be used to train and obtain the optimal function.

V. Model validation

"The validation of a QSAR involves assessing the model's predictive ability, applicability domain, and mechanistic basis for a specific purpose. Before using a QSAR model to interpret and predict biological responses of untested compounds, it must be properly validated. In essence, there are four tools for assessing the validity of QSAR models. These are:

- **Randomization of Response into an array of Recorded Variables (Y scrambling)**

This procedure ensures that the model is not due to a chance. The most widely used approach to establish model robustness is Y scrambling (random permutation of response values, i.e., activities).

This process entails repeating the calculation with randomized activities and then evaluating the resulting statistics for their probability.

- **Cross-validation**

In recent times, the method known as cross-validation, or more accurately leave-one-out cross-validation (LOO), was developed. In this method, a single sample of size n is used. Each member of the sample is removed in turn, the full modeling method is applied to the remaining n-1 members, and the fitted model is applied to the holdback member.

- **Splitting of parent data set into training and validation sets**

Cross-validation gives a good estimate of how well the QSAR model can predict the activity values of new compounds. This is called internal validation because all the chemicals used belong to the same dataset. If there are enough compounds available, they can be divided into a training set and a separate validation set for external validation.

- **External validation using a designed validation set**

External validation using a designed validation set is a crucial aspect of any QSAR modeling. It is an absolute requirement for the development of a truly predictive QSAR model. True external datasets are rare for QSAR studies, and in cases where they are not available, the dataset is divided into training and test sets for appropriate validation.

1.1.4.7 Chemometric tools

Chemometrics is the chemical discipline that uses statistical methods to design optimal procedures, experiments, and objects, and to provide maximum chemical information by analyzing chemical data.

1.1.4.7.1 Various chemometric tool used in QSAR/ QSPR/QSTR

QSAR/QSPR/QSTR is a statistical approach correlating the response property, activity, or toxicity data with descriptors encoding chemical information. Such correlation may be derived either in regression-based approach (in case where the response property is quantitative and available on a continuous scale) or a classification-based approach (in cases where the response

property is graded or semi-quantitative). The most commonly used regression-based approaches are as follows:

- Multiple linear regression (MLR)
- Partial least square (PLS)

Some of the common classification-based approaches are as follows

- Linear discriminant analysis (LDA)
- Logistic regression

Machine learning tools such as artificial neural networks and support vector machines are very effective in developing predictive models, especially for handling high-dimensional and complex chemical information data that show a nonlinear relationship with the response(s) of the chemicals. This chapter will briefly discuss some of the more popular and commonly used chemometric tools. However, before applying any statistical model-building method, the QSAR/QSPR data table may need to be pre-treated followed by a suitable feature selection method.

- **Multiple linear regression (MLR)**

Multiple linear regression [35] or MLR is commonly used in QSAR due to its simplicity, transparency, reproducibility, and easy interpretability. The generalized expression of an MLR equation will be like the following,

$$Y = a_0 + a_1 \times X_1 + a_2 \times X_2 + a_3 \times X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + a_n \times X_n \qquad (1.4)$$

In the above expression, Y is the response or dependent variable, X1, X2, X_3…..X_n are descriptors (features or independent variables) present in the model with the corresponding regression coefficient $a_1, a_2, a_{\ldots..}a_n$, respectively, and a_0 is the constant term of the model. Each regression coefficient should be significant at $p<0.05$ which can be checked from the 't' test. The descriptors present in an MLR should not be much intercorrelated. For a statistically reliable model, the number of observations and descriptors should be maintained at 5:1.

- **Partial least square (PLS)**

When dealing with a large number of interrelated and noisy descriptors for a limited amount of data, PLS is a better choice over MLR. PLS is an extension of MLR and aims to extract latent variables (LVs) from the original variables, capturing as much of the underlying factor variation as possible while modeling the responses. In linear PLS, new variables (latent variables) are found, representing linear combinations of the original variables. When the number of LVs equals the number of variables, the PLS model is equivalent to the MLR model. It is important to rigorously test the predictive significance of each PLS component and stop adding new components when they become non-significant.

- **Linear discriminant analysis (LDA)**

LDA [36] can be used to separate two or more classes of objects, making it useful for classification problems. It performs a similar task to MLR by predicting an outcome when the response property has graded values and molecular descriptors, as well as continuous variables. LDA explicitly attempts to model the difference between the classes of data. In a two-group situation, the predicted membership is calculated by computing a discriminant function (DF) score for each case. Then, cases with DF values smaller than the cut-off value are classified as belonging to one group, while those with values larger are classified into the other group. The DF may take the following form:

$$DF= C_1 \times X_1 + C_2 \times X_2 + \ldots\ldots\ldots + C_m \times X_m + a \qquad (1.5)$$

where DF is the discriminate function, which is a linear combination (sum) of the discriminating variables, c is the discriminant coefficient or weight for that variable, X is respondent's score for that variable, a is a constant, m is the number of predictor variables. The c's are unstandardized discriminant coefficients analogous to the beta coefficients in the regression equation. This c's maximize the distance between the means of the criterion (dependent) variable. Good predictors tend to have large standardized coefficients. After using an existing set of data to calculate the DF and classify cases, any new cases (test samples) can then be classified. In a stepwise DF analysis, the model is built step-by-step. Specifically, at each step, all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

- **Logistic regression**

Logistic regression [37] is a statistical classification model that assesses the relationship between a categorical-dependent variable (having only two categories) and one or more independent variables. These independent variables are usually continuous, but not necessarily so. It uses probability scores as the predicted values of the dependent variable. Unlike linear regression, logistic regression does not assume a linear relationship between the dependent and independent variables. The independent variables do not need to be normally distributed, linearly related, or have equal variance within each group.

- **Cluster analysis**

Unlike LDA, cluster analysis [38] does not require prior knowledge about which elements belong to which clusters. Instead, the clusters are defined through an analysis of the data. Cluster analysis aims to maximize the similarity of cases within each cluster while maximizing the dissimilarity between initially unknown groups. Hierarchical cluster analysis identifies relatively homogeneous clusters of cases based on dissimilarities or distances among objects. The most common way to compute distances between objects in a multidimensional space is to calculate

the Euclidean distances or the squared Euclidean distance. The process starts with each case as a separate cluster and then sequentially combines the clusters, reducing the number of clusters at each step until only one cluster is left. The k-means clustering is a non-hierarchical method of clustering that is used when the number of clusters in the data is known. It is an unsupervised, centroid-based method. In general, the k-means method will produce exactly k different clusters. The method starts by defining k centroids, one for each cluster, and placing them as far away from each other as possible. The next step is to take each point in the dataset and associate it with the nearest centroid. When no point is pending, the positions of the k centroids are recalculated. This process is repeated until the centroids no longer move.

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

QSTR and read-across techniques have recently converged to form an emerging field known as read-across structure-toxicity relationship (RASTR). This approach combines the chemical similarity principles of read-across with supervised learning techniques similar to QSAR. RASTR has been used for both the classification modeling and quantitative predictions (q-RASTR) [39].

This modeling approach utilizes a combination of similarity and error-based descriptors. This method has been shown to have better predictive potential and lower Mean Absolute Error (MAE) as compared to both QSTR and read-across predictions. The strength of q-RASAR lies in its ability to incorporate both similarity and error measurement information into descriptors, creating models that are straightforward, interpretable, transferable, and replicable, with improved predictive accuracy [40].

- **q-RASTR descriptors**

Compound similarity is estimated using three different methods: Euclidean distance, Gaussian kernel similarity, and Laplacian kernel similarity. The RASAR descriptor **RA function** is a prediction function derived from read-across, created by averaging the response values of source compounds identified as having structurally analogous properties. The weighted standard deviation of activity near n source chemicals for a specific target compound is represented by the **SD activity** descriptor. **SE** stands for the weighted standard error associated with the activity values of the nearby n-source compounds for a given target compound. The descriptor **CVact** represents the coefficient of variation of the activity values among the nearby n-source compounds for a specific target compound. **MaxPos** signifies the maximum similarity score between the target compound and the training set, while **MaxNeg** quantifies the degree of resemblance between a target compound and a nearby 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 specific query compound is denoted as **Abs Max Pos-Max Neg**. The

descriptor **AvgSim** calculates the mean similarity value among n closely related source compounds for a specific target compound. The $\mathbf{g_m}$ (Banerjee-Roy coefficient) descriptor assesses the likelihood of whether the query compound is active or inactive, with values ranging from -1 to +1. **gm*Avg. Sim** and $\mathbf{g_m}$***SD_Similarity** descriptors are obtained by multiplying gm values with **Avg. Sim** and **SD_Similarity** values, respectively. **Pos. Avg. Sim** indicates the average similarity values among the n close source compounds with response values higher than the training set's mean response value, while **Neg. Avg. Sim** signifies the average similarity values among the n close source compounds with response values lower than the training set's mean response value [41-42].

1.1.4.8 Calculation of various statistical metrics to evaluate the quality of a model

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

1.1.4.8.1 Determination coefficient (R^2): This parameter is known as the determination coefficient or squared correlation coefficient. The squared correlation coefficient of a model can be obtained from the following equation,

$$R^2 = 1 - \frac{\Sigma\left(Y_{obs(train)} - Y_{calc(train)}\right)^2}{\Sigma\left(Y_{obs(train)} - \overline{Y}_{train}\right)^2} \quad (1.6)$$

In regression analysis, the goal is to minimize the sum of squared residuals (the differences between the observed and predicted values). A small sum of squared residuals indicates a good fit for the model. We expect most individual observed Y values to deviate significantly from the predicted Y values. In an ideal model, the sum of squared residuals is 0, and the R squared (R^2) value is 1. As the R^2 value deviates from 1, the model's fitting quality worsens. The square root of R^2 is the multiple correlation coefficient (R).

1.1.4.8.2 Leave-one-out cross-validation (Q^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*) [43] and the acceptable models have been further processed for the prediction of toxicity and/or property of the test set compounds. Cross-validated correlation coefficient R^2 (LOO−Q^2) is calculated according to the formula,

$$Q^2_{LOO} = 1 - \frac{\Sigma(Y_{obs(train)} - Y_{pred(train)})^2}{\Sigma(Y_{obs(train)} - \overline{Y}_{train})^2} \quad (1.7)$$

Here $Y_{obs(train)}$, $Y_{pred(train)}$, and $\overline{Y}_{train}$ and are the observed, predicted and the average value of the response variable of the training set. In this technique, a single compound is randomly omitted from the dataset in each cycle, and then a model is built using the remaining compounds. This process is repeated for every compound in the dataset. The model formed in this way is used to predict 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. Acceptable value of Q^2 is 0.5 with a maximum value of 1.0 and hence more the value is closer to 1, more will be the internal predictivity of the model.

1.1.4.8.3 Root mean square error of calibration

The root mean square error of calibration ($RMSE_C$) can be computed from the following expression,

$$RMSE_c = \sqrt{\frac{(\Sigma Y_{obs(train)} - Y_{calc(train)})^2}{n}} \quad (1.8)$$

The value of RMSEc should be low for a good model.

1.1.4.8.4 r_m^2(LOO)

It was shown that [44] squared cross-validated correlation coefficient alone might not indicate the true predictive capability of a model and hence a modified r^2 (r_m^2(LOO)) term was used to indicate the leave-one-out prediction capacity of the model for the training set compounds. The parameter r_m^2(LOO) is obtained from the following equation,

$$r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right) \quad (1.9)$$

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

1.1.4.8.5 Golbraikh and Tropsha criteria

Golbraikh and Tropsha [45] proposed several parameters for determining the external predictability of the QSAR model. An acceptable QSAR model should be close to ideal in order to exert high predictive ability. An ideal QSAR model should have a correlation coefficient (R) that is close to 1 between the observed (y) and predicted (y/) activities. According to Golbraikh and Tropsha, regressions of y against y' against y through the origin should be characterized by

either k or k' (slopes of the corresponding regression lines) being close to 1. Subsequently, the regression lines through the origin are defined by $y^{r0} = ky'$ and $y'^{r0} = k'y$, while the slopes k and k' are given by, respectively,

$$k = \frac{\sum y_i y_i'}{\sum y_i'^2} \tag{1.10}$$

$$k' = \frac{\sum y_i y_i'}{\sum y_i^2} \tag{1.11}$$

A stricter condition for the QSAR model to have high predictive ability was further proposed by Golbraikh and Tropsha. They showed that either of the squared correlation coefficients of these two regression lines (y against y' or y' against y through the origin) r_0^2or $r_0'^2$ (given by Eqs. (1.12) and (1.13), respectively) should be close to the value of r^2 for the developed model. The values of r^2 and r_0^2 indicate the squared correlation coefficients between the observed and the predicted activity values with and without intercept, respectively, while $r_0'^2$ represents the same information as r_0^2 does, but with inverted axes:

$$r_0^2 = 1 - \frac{\sum (y_i' - y_i^{r_0})^2}{\sum (y_i' - \overline{y'})^2} \tag{1.12}$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i'^{r_0})^2}{\sum (y_i - \overline{y})^2} \tag{1.13}$$

Based on Golbraikh and Tropsha criteria, the model will be acceptable if:

1. Q^2_{LOO} (train) > 0.5
2. R^2(test) > 0.6
3. $[(r^2-r_0^2)/ r^2] < 0.1$ or $[(r^2-r'^2_0)/ r^2]$
4. 1.15 > k > 0.85 or 1.15 > k'> 0.85

1.1.4.8.6 MAE-based criteria

In a recent study, Roy et al. [46] have shown that commonly used metrics like (Q^2_{F1}), (Q^2_{F2}), and (Q^2_{F3}) can often provide biased assessments of model predictivity. This is because these metrics are influenced by factors such as the response range and distribution of data. In this study, the authors have proposed a set of criteria that utilize the 'mean absolute error' (MAE) and the corresponding standard deviation (σ) of the predicted residuals to evaluate the external predictivity of the models.

$$MAE = \frac{1}{n} \times \sum \left| Y_{obs} - Y_{pred} \right| \tag{1.14}$$

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

- **Good prediction-** The criteria for good predictions is as follows,

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

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

- **Bad prediction-** The criteria for bad predictions is as follows,

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

Here, a value of MAE more than 15% of the training set range is considered high while an error of more than 25% of the training set range is judged as very high. The predictions that do not fall under either of the above two conditions may be considered moderate quality. The above criteria should be applied for judging the quality of test set predictions when the number of data points is at least 10 (statistical reliability) and there is no systematic error in model predictions (statistical applicability).

1.1.4.8.7 Q^2_{F1} or R^2_{pred}

Predictive R^2 (Q^2_{F1}) reflects the degree of correlation between the observed and predictive activity data of the test set.

$$Q^2_{F1} = 1 - \frac{\Sigma\left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\Sigma\left(Y_{obs(test)} - \overline{Y}_{train}\right)^2} \qquad (1.15)$$

Here, $Y_{obs(test)}$ and $Y_{pred(test)}$ are the observed and predicted activity data for the test set compounds, while $Y_{training}$ indicates the mean observed activity of the training set molecules. Thus, models with values of R^2_{pred} or Q^2_{F1} above the stipulated value of 0.5 are considered to be well predictive.

1.1.4.8.8 Q^2_{F2}

Another expression for the calculation of external Q^2 (i.e., Q^2_{F2}) is based on the prediction of test compounds proposed by Schurmann et.al [47] as given by Eq. (1.16)

$$Q^2_{F2} = 1 - \frac{\Sigma\left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\Sigma\left(Y_{obs(test)} - \overline{Y}_{test}\right)^2} \quad (1.16)$$

where $\overline{Y}_{test}$ refers to the mean observed data of the test set compounds and Q^2_{F2} differs from Q^2_{F1} in the mean value used in the denominator for calculation. When the two values approach each other, it can be inferred that the training set mean lies in close proximity to that of the test set, indicating that the test set used for prediction spans the whole response domain of the model. A threshold value 0.5 is defined for this parameter.

1.1.4.8.9 Q^2_{F3}

One more parameter, Q^2_{F3} with the threshold value of 0.5, used for external validation of a QSAR model, has been proposed by consonni et.al [48]. This parameter is defined as follows,

$$Q^2_{F3} = 1 - \frac{\left[\Sigma\left(Y_{obs(test)} - Y_{pred(test)}\right)^2\right]/n_{test}}{\Sigma\left(Y_{obs(training)} - \overline{Y}_{training}\right)^2/n_{train}} \quad (1.17)$$

Where, n_{train} refers to the number of compounds in the training set. Here, the summation in the numerator deals with the external test set, while that in the denominator runs over the training set compounds. Considering that the number of test and training objects are usually different, divisions by n_{test} and n_{train} make the two values comparable. However, although the value of Q^2_{F3} measures the model predictability, it is sensitive to training-set data selection and tends to penalize models fitted to a very homogeneous data set, even if predictions are close to the truth. Since this function includes information about the training set, it cannot be properly regarded as an external validation measure even if predictions are really obtained for the external test set.

1.1.4.8.10 Concordance correlation coefficient (CCC)

The concordance correlation coefficient (CCC) parameter [49] can also be calculated to check the model reliability by using the following equation:

$$\overline{\rho}_c = \frac{2\sum_{i=1}^{n}\left(x_{obs(test)} - \overline{x_{obs(test)}}\right)\left(y_{pred(test)} - \overline{y}_{pred(test)}\right)}{\sum_{i=1}^{n}\left(x_{obs(test)} - \overline{x_{obs(test)}}\right)^2 + \sum_{i=1}^{n}\left(y_{pred(test)} - \overline{y_{pred(test)}}\right)^2 + n\left(\overline{x_{obs(test)}} - \overline{y_{pred(test)}}\right)} \quad (1.18)$$

In the above equation, $X_{obs(test)}$ and $Y_{pred(test)}$ correspond to the observed and predicted values of the test compounds, n is the number of chemicals, and $\overline{X}_{obs(test)}$ and $\overline{Y}_{pred(test)}$ correspond to the averages of the observed and predicted values, respectively, for the test compounds. The ideal value of CCC should be equal to 1.

1.1.4.8.11 Y-randomization

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

i) Process randomization and

ii) Model randomization

In process randomization, random scrambling of the dependent response variables is performed accompanied with fresh selection of variables from the whole descriptor matrix and in model randomization scrambling or randomization of the response variable is performed within the descriptors present in an existing model. We have performed process as well as the model randomization of the genetic models. Y-randomization study has been performed to analyze and confirm whether the developed models are produced by any chance. Y-randomization plots are generated for developed models through the SIMCA-P software (https://www.sartorius.com/en/products/process-analytical-technology/data-analyticssoftware/mvda-software/simca).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 chance. The values of the $R^2y_{rand\ intercept}$ and Q^2y_{rand} intercept should not be more than 0.3 and 0.05 respectively.

1.1.4.8.12 Determination of model applicability domain (AD)

Applicability domain (AD) of a QSAR model can be described as the theoretical region in the chemical space defined by the chemical as well as response attributes of the model [50]. A definite domain of applicability enables reliability of predictive performance of a model. In other words, any QSPR 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, QSPR models are characterized by different domains. 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 Distance to model in X-space (DModX) approach for verifying the applicability domain of the best model developed for this study using Simca-P software [51].

$$\text{DModX} = \frac{\sqrt{\frac{SSEi}{K-A}}}{\sqrt{\frac{SSE}{(N-A-AO)(K-A)}}} \quad (1.19)$$

For observation i, in a model with A component, K variables, and N observations, SSE is the squared sum of the residuals. A0 is 1 if the model was centered and 0 otherwise. It is claimed that DModX is approximately F-distributed, so it can be used to check if an observation deviates significantly from a normal PLS model.

1.1.4.9 Model validation based on OECD guidelines

To authenticate the applicability of the developed QSTR models and to judge the reliability of the predictions made, the models were further analyzed based on the OECD guidelines [52]. Thus, the QSTR models developed in this work were validated based on these five guidelines laid down by the OECD. The compliance of the developed models with the OECD guidelines for applicability in regulatory purposes was assessed as follows:

Principle 1: A defined endpoint

The response parameter modeled in the present work for different datasets were measured under similar conditions. Thus, the QSTR models were developed in accordance with the 1st OECD principle.

Principle 2: An unambiguous algorithm

Various chemometric tools based on specific algorithms were employed for the calculation of the different categories of descriptors and subsequent QSTR model development using specific software packages. Thus, the model development pathway employed for the present studies follows a definite algorithm.

Principle 3: A defined domain of applicability

The domain of applicability of all the statistically significant QSTR models was analyzed using the standardization method. Thus, the selection of the best QSTR model was done in corroboration with this principle.

Principle 4: Appropriate measures of goodness-of-fit, robustness, and predictivity

All the developed models were rigorously validated using internal, external, and overall validation techniques. The quality of fitness and the predictive potential of the developed models were assessed based on the different validation metrics while the robustness of the models was judged using the randomization approach. The selection of the most significant models based on the acceptable values of the various validation metrics accounts for the compliance of the models with the 4^{th} guideline.

Principle 5: A mechanistic interpretation

All the descriptors appearing in the developed QSTR models could aptly define the essential structural attributes of the molecules imparting optimum endpoint values thus signifying suitable mechanistic interpretation of the developed models.

1.1.4.10 Software packages employed in the study

We have used different software's in this research work namely:

i. "AlvaDesc" software (https://www.alvascience.com/alvadesc/) was used for descriptor calculation.

ii."Best Subset Selection Modified" v2.1 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for model development.

iii. "Dataset Division GUI" v1.2 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to divide the dataset into training and test sets.

iv. "Minitab" v14 (https://www.minitab.com/en-us/) was used for model development.

v. "PLS_Single Y" v1.0 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to develop the PLS-based QSTR and q-RASTR models.

vi. "Read-Across-v4.1" (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for obtaining the optimized hyperparameters necessary for RASTR descriptor calculation.

vii."RASAR Descriptor Calculator" v2.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) was used for RASTR descriptors calculation.

viii. "Prediction Reliability Indicator" (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to evaluate the localization in AD of the test compounds to ascertain the reliability of prediction of final PLS-based q-RASTR model.

ix. "SIMCA-P" (https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca) was used for the randomization test.

CHAPTER - 2

Present work

2. PRESENT WORK

Over the past four decades, Europe has witnessed a staggering loss of approximately 550 million birds from its population with the predominant cause being the widespread use of pesticides and fertilizers in agricultural practices followed by climate change, changes in forest cover, and urbanization [53]. Pesticides play an indispensable role in modern agriculture by managing pests and improving crop productivity [54]. Nevertheless, they carry a double-edged sword, offering benefits while also posing risks to non-target organisms such as avian species, thereby raising significant environmental concerns in scientific research. Birds play crucial roles in ecosystems, aiding in pest control, pollination, and seed dispersal, which are vital for maintaining biodiversity, ecosystem equilibrium, and environmental health [55]. However, exposure to pesticides can result in acute toxicity and long-term declines in avian populations, thus disturbing ecological balance and biodiversity [56]. Consequently, assessing pesticide toxicity becomes imperative for managing the associated health risks to avian species and preserving ecosystem balance.

Traditional toxicity evaluations in birds involving *in-vivo* testing are costly, labor-intensive, time-consuming, alongside ethical concerns, and almost practically unfeasible for addressing a multitude of avian species [57]. The proliferation of new chemical entities and diverse pesticide formulations underscores the need for alternative methods to consistently predict the toxic effects of pesticides on avian species, wherein high-throughput computational approaches can offer promising solutions [58]. To explore the intrinsic characteristics of chemicals for toxicological prediction, regulatory institutions such as the Environmental Protection Agency (EPA), Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), European Chemicals Bureau (ECB), and European Food Safety Authority (EFSA) emphasize the use of in-silico based approaches i.e. Quantitative Structure-Toxicity Relationship (QSTR) and Read-across [59]. QSTR enables mathematical correlation of the physicochemical properties of chemicals with their biological activities. Read-across is a similarity-based approach employed to estimate toxicity by comparing a substance to a similar one with known toxicity, eliminating the need for supervised learning models. The q-RASTR (Quantitative Read Across Structure–Toxicity Relationship) approach is the amalgamation of the QSTR and Read-Across which incorporates the similarity and error-based estimations to improve prediction accuracy. A recent advancement in predictive modeling known as q-RASTR has emerged, offering improvements over traditional methods like QSTR and read-across predictions. QSTR relies solely on descriptor values for structural and physicochemical data of test compounds, but q-RASTR utilizes a combination of similarity and error-based descriptors [60]. This approach enhances predictive accuracy and reduces mean absolute error (MAE) compared to its predecessors. Additionally, q-

RASTR addresses the limitations of previous algorithms by incorporating information from close source neighbors in the training set into the descriptors of query/test compounds. This integration of training set data enables "prediction-inspired intelligent training", resulting in enhanced external predictivity for most scenarios. Machine learning is a growing technology that uses various algorithms for building models and making predictions using data. Support vector machines (SVM), artificial neural networks (ANN), and others are commonly used machine learning algorithms for numerous experimental studies [61].

Chemical toxicities are typically assessed using metrics such as the median lethal concentration (LC_{50}) or median lethal dose (LD_{50}), lowest observed effect level (LOEL), and no observed effect level, etc which vary based on factors such as the nature of the chemical, exposure pathways, and the species being tested. Typically, regulatory evaluations favor the lowest toxicity endpoint across species, yet this practice introduces bias towards compounds with sparse data. Conversely, comparing toxicity within the same species overlooks inherent variations in susceptibility among different species or chemical classes [62]. In instances where data scarcity impedes toxicity characterization, extrapolation techniques are employed to ensure a comprehensive understanding of a pesticide's impact on avian populations. An approach towards extrapolating laboratory toxicity data is the estimation of HD_5, also known as the fifth percentile of the LD_{50} distribution, by aggregating LD_{50} data across multiple species from laboratory experiments. The HD_5 value indicates a threshold where 50% mortality is expected for the most sensitive 5% of bird species. This distribution-based method ensures a comprehensive evaluation of pesticide toxicity, facilitating unbiased comparisons irrespective of data availability. To enable cross-species comparisons of toxicological susceptibility, the HD_5 calculation incorporates a body weight scaling factor, and adjustments are made to account for the heightened susceptibility of smaller species in lethality assessments [63]. Incorporating HD_5 alongside LD_{50} provides a safety buffer that will aid toxicologists and regulators in making informed decisions to protect the avian biodiversity.

Study 1. Chemometrics-driven prediction and prioritization of diverse pesticides on chickens for addressing hazardous effects on public health

In this work, we investigated the toxicity of several pesticides on chickens and developed a logical and trustworthy method for assessing ecotoxicological risk. Based on the OECD rules, we have developed q-RASTR models to predict pesticide ecotoxicity on bird species. RASTR combines the read-across and QSTR approaches to improve predictability. The pLOEL and pNOEL (the negative logarithm of Lowest Observed Effect Level and No Observed Effect Level values respectively) values have been used as endpoints in this study. NOEL is defined as the highest

dose of the toxicant that does not cause any toxicity or harm and LOEL stands for the lowest concentration of a substance that can cause an effect under specific exposure conditions. To successfully create the models, we used PLS for the initial model development. Further, RASAR descriptors were estimated using the optimal hyperparameters and incorporated to improve the external predictivity of the model. Additionally, Support vector machine and Ridge regression machine learning (ML) approaches were employed with the optimization of hyperparameters using cross-validation. The final test set predictions were then compared. After evaluating the test set predictions and interpretability, we have selected the PLS-based q-RASTR model as the final model. Using, globally accepted parameters, the robustness, reproducibility, and predictivity of the PLS-based q-RASTR models were thoroughly validated. It can be confidently affirmed that the models are reliable and accurate. The developed model was utilized to screen the Pesticide Properties Database (PPDB) to identify potential avian toxicants and promote the use of safer chemicals. The true predictive ability of the q-RASTR model was established by revalidating the real-world toxicity profiles of the most and least toxic screened compounds from the Pesticide Properties Database (PPDB).

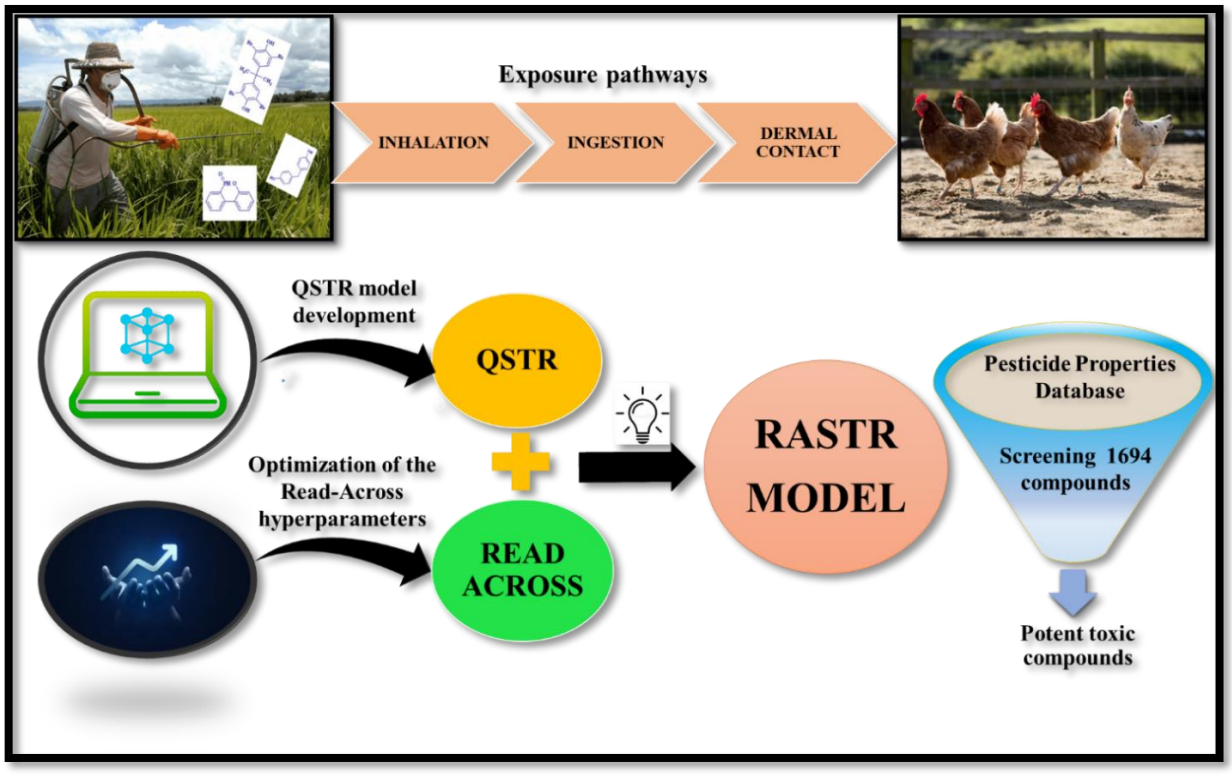


Figure 2.1. Graphical representation of study 1.

Study 2. First report on q-RASTR modeling of hazardous dose (HD_5) for acute toxicity of pesticides: An efficient and reliable approach towards safeguarding the sensitive avian species

The current work offers the first chemometric modeling for efficient prediction of HD_5 values pertaining to acute toxicity of pesticides towards avian species employing simple 2D molecular descriptors with ease of interpretability. q-RASTR approach was utilized to enhance the predictivity of the developed model. The structural features of pesticides closely associated with the modulation of toxicity towards multiple avian species were also highlighted. To address the practical applicability, the q-RASTR model was employed to analyze the Pesticide Properties Database (PPDB) to identify the safe and toxic compounds towards avian species to enable the adoption of safer chemical alternatives. To evaluate the actual external predictive performance of the q-RASTR model, real-world data was employed to validate the twenty most and least toxic pesticides identified through screening. By bridging the gap between computational predictions and real-world toxicological outcomes, this research endeavors to contribute significantly to the field of ecological risk assessment and the protection of avian biodiversity amidst the ever-increasing pesticide usage.

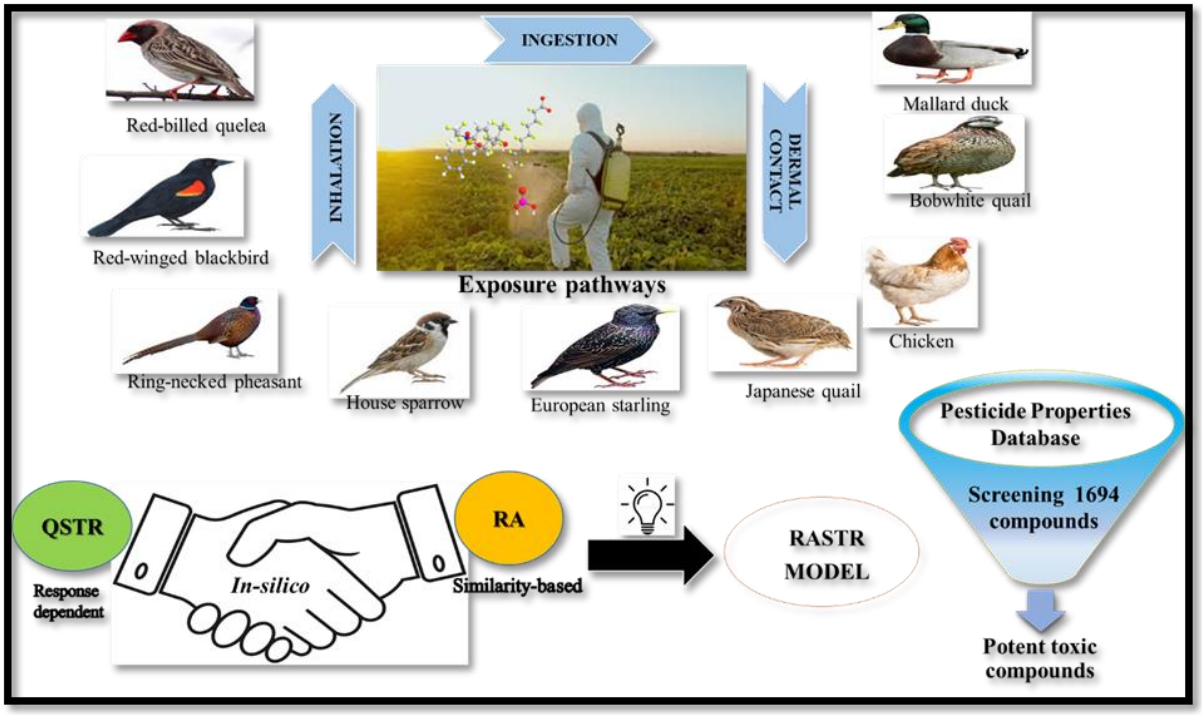


Figure 2.2. Graphical representation of study 2.

Study 3. Comprehensive Ecotoxicological Assessment of Pesticides on Multiple Avian Species: Employing Quantitative Structure-Toxicity Relationship (QSTR) Modeling and Read-Across

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, and JQ avian species following the OECD guidelines strictly [64]. 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) [65-66] 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, and JQ) 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 [67]. The robustness, reproducibility, and predictivity of QSTR models were thoroughly validated using globally accepted statistical parameters.

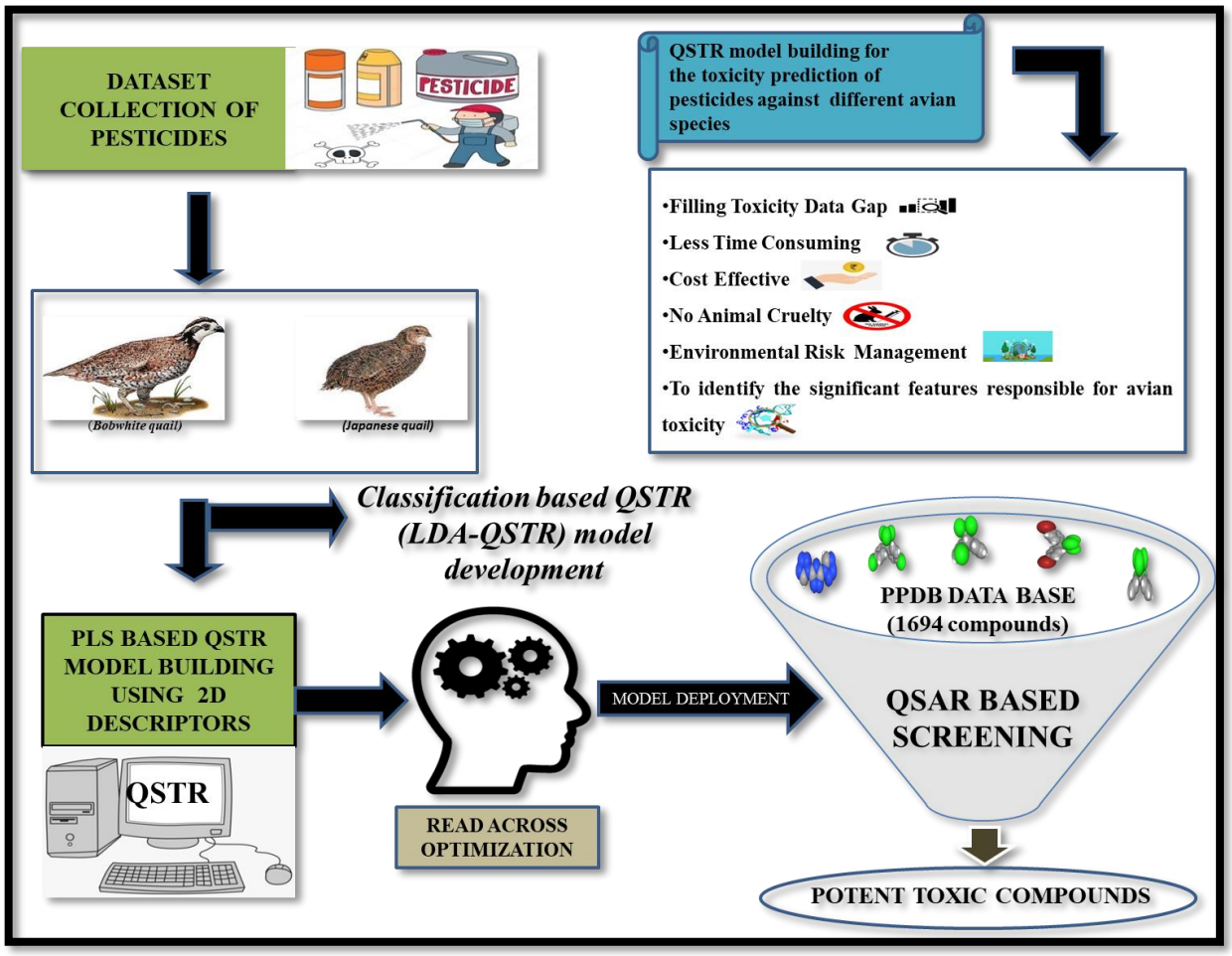


Figure 2.3. Graphical representation of study 3.

CHAPTER - 3

Materials and Methods

3. MATERIALS AND METHODS

This dissertation seeks to establish a clear and transparent methodological framework for constructing a predictive q-RASTR model, utilizing 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 promising features, and the assessment of the models' reliability and predictive capabilities. In the following sections, we provide comprehensive insights into the dataset used for the q-RASTR modeling. This includes a detailed presentation of the dataset, along with information about the activities and toxicity data it contains. These data are instrumental in facilitating our computational investigations and predictive modeling efforts. The research undertaken was organized into distinct components, each serving a specific purpose:

- ***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 present a general overview of the methodologies and techniques employed in the development of our q-RASTR model. This section outlines the strategies and tools we used to create predictive models for understanding the relationship between chemical structures and toxicity.

3.1 Study 1

3.1.1 Collection and curation of toxicity data of diverse pesticides

The required toxicity data of diverse pesticides against chicken *(Gallus gallus)* were retrieved from the ECOTOX repository (https://cfpub.epa.gov/ecotox/). The collected experimental toxicity data was expressed as LOEL and NOEL in micromolar (μM) concentration, which were transformed into molar concentrations and then their negative logarithmic equivalents (pLOEL and pNOEL) to reduce the data range. After excluding any outlier value(s), all available values for a particular chemical were averaged to generate a single value. We only included values that were numerically close to each other when calculating the average. After curating the primary data, we selected 43 pLOEL and 56 pNOEL compounds for modeling.

Table 3.1. Compounds name with respective experimental pLOEL values.

Sl.No	Compound	Exp pLOEL
1	(17beta) Estra-1,3,5(10) triene-3,17-diol	5.055
2*	4,4'-(1-Methylethylidene) bis [2,6-dibromophenol]	4.720
3	Phosphoric acid-triphenyl ester	5.000
4	1,2-Benzenedicarboxylic acid, 1,2-Bis(2-ethylhexyl) ester	3.301
5	2,2,3,3,4,4,5,5,6,6,6-Undecafluorohexanoic acid	5.000
6	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,12-Tricosafluorododecanoic acid	4.864
7*	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-Nonadecafluorodecanoic acid	4.823
8*	2,2,3,3,4,4,5,5,6,6,7,7,7-Tridecafluoroheptanoic acid	5.000
9	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,9-Heptadecafluorononanoic acid	4.585
10*	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,14-Heptacosafluorotetradecanoic acid	4.522
11*	(2R,3R)-2-(3,4-Dihydroxyphenyl)-3,5,7-trihydroxy-2,3-dihydro-4H-1-benzopyran-4-one	6.301
12	4,4'-Methylenebisphenol	4.488
13	Phosphoric acid-Diphenyl ester	5.000
14	4,4'-[2,2,2-Trifluoro-1-(trifluoromethyl)ethylidene] bis [phenol]	5.000
15*	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Heptadecafluoro-1-octanesulfonic acid	4.723
16	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,11-Heneicosafluoroundecanoic acid	4.923
17	4,4'-[1,4-Phenylenebis(1-methylethylidene)] bis[phenol]	4.903
18	2,2,3,3,4,4,5,5,5-Nonafluoropentanoic acid	4.301
19*	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Heptadecafluoro-1-octanesulfonic acid potassium salt	4.373
20	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-Heneicosafluoro-1-decanesulfonic acid	5.000
21*	2,2,3,3,4,4,5,5,6,6,6-undecafluoro-hexanoic acid	4.425
22	Tris[3-bromo-2,2-bis(bromomethyl) propyl] phosphate	4.000
23	2,2,3,3,4,4,5,5,6,6,7,7,7-Tridecafluoroheptanoic acid)	4.301
24	1,1'-(1-Methylethylidene) bis [3,5-dibromo-4-(2,3-dibromopropoxy) benzene	3.551
25*	1,1,2,2,3,3,4,4,5,5,6,6,7,7,7-Pentadecafluoro-1-heptanesulfonic acid	5.000
26	(8S,10S)-10-[(3-Amino-2,3,6-trideoxy-alpha-L-lyxo-hexopyranosyl) oxy]-7,8,9,10-tetrahydro-6,8,11-trihydroxy-8-(2-hydroxyacetyl)-1-methoxy-5,12-naphthacenedione	6.301
27	2,4,6-Tris(2,4,6-tribromophenoxy)-1,3,5-triazine	3.522
28*	3,4,5,6-Tetrabromo-1,2-benzenedicarboxylic acid 1,2-bis(2-ethylhexyl) ester	3.522
29*	1,1,2,2,3,3,4,4,4-Nonafluoro-1-butanesulfonic acid	5.187
30	Phosphoric acid-Isodecyl diphenyl ester	3.522
31	6H-Dibenz[c,e] [1,2] oxaphosphorin, 6-Oxide	4.000
32*	4,4'-Sulfonylbis[2-(prop-2-en-1-yl) phenol]	4.522

33	2-[4-(4-Chlorobenzoyl) phenoxy]-2-methyl propanoic acid, 1-Methylethyl ester	4.221
34	2-[[4-Chloro-6-[(2,3-dimethylphenyl) amino]-2-pyrimidinyl] thio] acetic acid	4.000
35	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,13-Pentacosafluorotridecanoic acid	4.301
36	1,1,2,2,3,3,4,4,5,5,6,6,6-Tridecafluoro-1-hexanesulfonic acid	5.691
37	4,4'-[Methylenebis(oxy-2,1-ethanediylthio)] bisphenol	5.000
38*	4-{4-[(Propan-2-yl) oxy] benzene-1-sulfonyl}phenol	4.522
39	4-[4-(4-Fluorophenyl)-5-(4-pyridinyl)-1H-imidazol-2-yl] phenol	5.000
40	N-[4-(1,1,1,3,3,3-Hexafluoro-2-hydroxypropan-2-yl) phenyl]-N-(2,2,2-trifluoroethyl) benzenesulfonamide	4.875
41	Ethanol	4.519
42	Hydrogen peroxide (H_2O_2)	5.552
43	1,4-diethyl 2-{[dimethoxy (sulfanylidene)-λ^5-phosphanyl]sulfanyl}butanedioate	3.397

*** Test set compounds**

Table 3.2. Compounds name with respective experimental pNOEL values.

Sl.No	Compound	Exp pNOEL
1	(1,1-Dimethylethyl) phenyldiphenyl ester, Phosphoric acid	4.782
2	(17beta) Estra-1,3,5(10) triene-3,17-diol	5.000
3*	(1R,2R,5S,6R,9R,10S)-rel-1,2,5,6,9,10-Hexabromocyclododecane	5.641
4	(2R,3R)-2-(3,4-Dihydroxyphenyl)-3,5,7-trihydroxy-2,3-dihydro-4H-1-benzopyran-4-one	6.301
5*	(3E)-2-Amino-4-methyl-5-phosphono-3-pentenoic acid 1-ethyl ester	4.000
6*	(5S)-10,11-dihydro-5-methyl-5H-Dibenzo[a,d]cyclohepten-5,10-imine (2Z)-2-butenedioate	5.301
7	1,1,2,2,3,3,4,4,4-Nonafluoro-1-butanesulfonic acid	4.282
8	1,1,2,2,3,3,4,4,5,5,6,6,6-Tridecafluoro-1-hexanesulfonic acid	4.698
9	1,1,2,2,3,3,4,4,5,5,6,6,7,7,7-Pentadecafluoro-1-heptanesulfonic acid	5.000
10	1,1,2,2,3,3,4,4,5,5,6,6,7,7,7-Pentadecafluoro-1-heptanesulfonic acid sodium	4.395
11	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Heptadecafluoro-1-octanesulfonic acid	4.637
12*	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Heptadecafluoro-1-octanesulfonic acid potassium	4.470
13	1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-Heneicosafluoro-1-decanesulfonic acid	4.338
14	1,1'-Oxybis[2,3,4,5,6-pentabromobenzene]	5.000
15	1,2,4,5-Tetrabromo-3,6-bis(2,3,4,5,6-pentabromophenoxy) benzene	5.721
16	10-[3-(4-Methyl-1-piperazinyl) propyl]-2-(trifluoromethyl)-10H-phenothiazine	6.154
17	2,2,3,3,4,4,4-Heptafluorobutanoic acid	5.000
18	2,2,3,3,4,4,5,5,5-Nonafluoropentanoic acid	4.363
19	2,2,3,3,4,4,5,5,6,6,6-Undecafluorohexanoic acid	5.148
20	2,2,3,3,4,4,5,5,6,6,6-undecafluoro-hexanoic acid	4.418
21	2,2,3,3,4,4,5,5,6,6,7,7,7-Tridecafluoroheptanoic acid	5.115

22	2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Pentadecafluorooctanoic acid	4.392
23	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-Nonadecafluorodecanoic acid	4.994
24	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,11-Heneicosafluoroundecanoic acid	5.070
25*	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,12-Tricosafluorododecanoic acid	4.999
26*	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,13-Pentacosafluorotridecanoic acid	4.363
27	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,14-Heptacosafluorotetradecanoic acid	4.323
28	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,15,15,16,16,16-Hentriacontafluorohexadecanoic acid	4.301
29	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,9-Heptadecafluorononanoic acid	4.580
30	2,4,6-Tris(2,4,6-tribromophenoxy)-1,3,5-triazine	3.522
31	2',7'-Dichloro-3',6'-dihydroxyspiro[isobenzofuran-1(3H),9'- 9H xanthen]-3-one	4.698
32*	2-[4-(4-Chlorobenzoyl) phenoxy]-2-methylpropanoic acid, 1-Methylethyl ester	4.307
33	2-Methyl-2-(methylthio)propanol O-[(methylamino) carbonyl] oxime	5.045
34	3-Methyl-3H-purin-6-amine	5.602
35*	4,4'-(1-Methylethylidene) bis[2,6-dibromophenol]	5.000
36	4,4'-(1-Methylethylidene) bis [2-methylphenol]	4.806
37	4,4'-(1-Methylethylidene) bisphenol	4.683
38*	4,4'-(1-Phenylethylidene) bis phenol	4.806
39*	4,4'-[1,3-Phenylenebis(1-methylethylidene)] bis phenol	5.107
40	4,4'-[1,4-Phenylenebis(1-methylethylidene)] bis [phenol]	5.028
41	4,4'-[2,2,2-Trifluoro-1-(trifluoromethyl)ethylidene] bis [phenol]	4.954
42*	4,4'-[Methylenebis(oxy-2,1-ethanediylthio)] bisphenol	5.000
43	4,4'-Methylenebisphenol	4.505
44	4,4'-Sulfonylbis[2-(prop-2-en-1-yl) phenol]	4.698
45	4-{4-[(Propan-2-yl) oxy] benzene-1-sulfonyl} phenol	4.522
46*	6-[4-[2-(1-Piperidinyl) ethoxy] phenyl]-3-(4-pyridinyl) pyrazolo [1,5-a] pyrimidine	4.698
47	6H-Dibenz[c,e][1,2]oxaphosphorin, 6-Oxide	4.000
48	Bis(tert-butylphenyl) phenyl phosphate	3.522
49	Hexabromocyclododecane	5.514
50	N-[(2S)-2-[[(1Z)-1-methyl-3-oxo-3-[4-(trifluoromethyl) phenyl]-1-propen-1-yl] amino]-3-[4-[2-(5-methyl-2-phenyl-4-oxazolyl) ethoxy] phenyl] propyl] propanamide	5.301
51	N-[4-(1,1,1,3,3,3-Hexafluoro-2-hydroxypropan-2-yl) phenyl]-N-(2,2,2-trifluoroethyl) benzenesulfonamide	5.193
52	Phosphoric acid-Diphenyl ester	4.187
53	Phosphoric acid-Isodecyl diphenyl ester	3.821
54	Phosphoric acid-Triphenyl ester	5.000
55	Tris(2,4-di-tert-butylphenyl) phosphate	5.000
56	Tris[3-bromo-2,2-bis(bromomethyl)propyl] phosphate	4.000

***Test set compounds**

3.1.2. Descriptor calculation

A single .sdf file of all the compounds was compiled which is essential to Alvadesc software for descriptor calculation. AlvaDesc software [68] was used to evaluate 2400 descriptors based on structural and physicochemical parameters. We removed the unnecessary descriptors columns using DataPreTreatmentGUI 1.2 software [69].

3.1.3 Dataset division and QSTR model development

Division of dataset is a crucial component of statistical modeling, particularly in the context of QSARs. The modeling data is divided into two parts, the training set for model development and the test set to validate the developed model. In this present study, different dataset division techniques such as the clustering technique, Euclidean-distance-based method, Kennard-stone-based method, activity property-sorted, and random-division methods were employed for dataset division into training and test sets. Among these techniques, the best result was obtained from the Kennard stone division method in case of the pLOEL endpoint and random selection in case of the pNOEL endpoint [65-70]. The training/test sets compounds for pLOEL endpoint and pNOEL endpoint are 30/13 and 44/12 respectively. And the divided training and test sets were also pre-treated using the tool dataPreTreatmentTrainTest1.0 (available from https://teqip.jdvu.ac.in/QSAR_Tools/). These final pre-treated training and test sets were used for further analysis. Preliminary multiple linear regression models were generated for two datasets using MINITAB software. After that, PLS (Partial Least Square) method was used to generate the final models for both datasets using the software PLS_Single Y_version 1.0 [65].

3.1.4 Read- Across and calculation of the RASTR descriptor

Optimizing hyperparameters (similarity-based algorithm; σ, γ, and number of close source compounds) is crucial for read-across prediction. The descriptor involved in the QSTR model was used to create sub-train and sub-test sets from the training data. We have chosen a Gaussian kernel-driven similarity, with σ=0.75; γ=0.75, and 9 close training compounds for pLOEL data points & Laplacian kernel-based similarity, with σ=0.25 and γ=0.25, and 4 close training compounds for pNOEL data points. During optimization, the hyperparameters were selected based on MAE-based (95%) criteria and external metrics (Q^2_{F1} and Q^2_{F2}). To perform q-RASTR modeling, similarity, and error-based RASTR descriptors were calculated for both training and test compounds with "RASAR Descriptor Calculator v2.0 tool using the optimized hyperparameters.

3.1.5 q- RASTR feature selection and model development

A total of 15 descriptors (**Table 3.3**) were computed based on three similarity-based approaches (Euclidean Distance-based, Gaussian Kernel similarity-based, and Laplacian Kernel similarity-

based) and a given set of source compounds for the individual training set and the test set. The calculated RASTR descriptors were integrated with the model descriptors and the combined pool was subjected to best subset selection using BestSubsetSelectionModified_v2.1 tool for model development. The final PLS-based q-RASTR model was developed with the best features using the PLS_Single Y_version 1.0 software.

Table 3.3 List of RASTR descriptors.

S.No.	RASTR descriptors	Definition
1	*RA function*	A composite function derived from Read-Across.
2	*MaxPos*	Similarity score of the closest positive source compound (with an observed response value greater than the mean activity of the training set).
3	*MaxNeg*	Similarity score of the closest negative source compound (with an observed response value less than the mean activity of the training set).
4	*Abs Maxpos-MaxNeg*	Absolute difference between the *MaxPos* and *Maxneg* levels.
5	*SE*	Weighted standard error of the close source compounds' response values.
6	*CVact*	Coefficient of variation of the close source compounds' observed response values.
7	*SD_Activity*	Weighted standard deviation of the close source compounds' observed response values.
8	*CVsim*	Coefficient and variation of the similarity values of the close source compounds.
9	*SD_similarity*	The standard deviation of the close source compounds' similarity levels.
10	*Pos.Avg.Sim*	The positive close source compounds' average similarity levels.
11	*Neg.Avg.Sim*	The negative close source compounds ' average similarity levels.
12	*Avg.Sim*	Average similarity level of the close source compounds.
13	g_m	A novel concordance measure also known as *Banerjee-Roy Coefficient*
14	*gm*SD_Similarity*	Product of the g_m and *SD similarity* levels
15	*gm*Avg.Sim*	Product of the g_m and *Avg.Sim* levels

3.1.6. 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 Ridge Regression (RR) using the Orange data mining tool. The hyperparameters were adjusted to tune the model for

optimal performance. The prediction qualities of the ML models were evaluated in terms of Q^2_{F1}, Q^2_{F2}, and MAE_{test} values.

3.1.7. *Statistical validation metrics and Y-randomization*

Validation metrics are the key parameters for the recognition of any predictive model. For internal validation (for the training set), we evaluated the model using various internationally accepted internal validation metrics including the determination coefficient (R^2) and leave-one-out cross-validated Q^2 (Q^2_{Loo}). R^2 and Q^2_{Loo} are the measures of goodness-of-fit, and robustness, respectively. In machine learning (SVM, RR) approaches, the root means squared error of calibration ($RMSE_C$) metric was also calculated by the Orange data mining tool. A lower RMSEc indicates a better model fit, showing that the model's predictions are, on average, closer to the true values. For external validation (for the test set), we calculated various globally accepted external validation metrics such as R^2_{Pred} or Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, MAE-based criteria, $\overline{r_m^2}$, Δr_m^2, and concordance correlation coefficient (CCC). External correlation coefficients such as Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} are well-known prediction indicators. In usual practice, the optimal value of these three measures (R^2_{Pred} or Q^2_{F1}, Q^2_{F2}, Q^2_{F3}) for model selection should be more than 0.5. Error measures such as mean absolute error (MAE_{test}) are frequently used to assess the accuracy of projected outputs, and they should be low for a strong model. The CCC measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. The concordance correlation coefficient (CCC) is an external validation measure proposed by Gramatica et.al. External validation is undertaken to ensure the predictability of the created model, and only the test set chemicals are employed for this purpose. Aside from traditional measures, r_m^2 metrics ($\overline{r^2_{m(test)}}$, $\Delta r_m^2{}_{test)}$ are calculated for external validation. When the $\overline{r^2_{m(test)}}$, values are quite good, the Δr_m^2 values may serve as an additional metric for judging the quality of predictions. The acceptability of the model was also checked using an external validation parameter proposed by Golbraikh and Tropsha. Based on Golbraikh and Tropsha criteria, the model will be acceptable if:

1. Q^2_{LOO} (train) > 0.5
2. R^2(test) > 0.6
3. $[(r^2\text{-}r_0^2)/ r^2] < 0.1$ or $[(r^2\text{-}r'_0{}^2)/ r^2]$
4. 1.15 > k > 0.85 or 1.15 > k'> 0.85

Y-randomization study was performed using "SIMCA-P" software to investigate the probability of chance occurrence in the final model. Herein, the response data are altered, without scrambling the descriptors, for a total of 100 times. After shuffling the original model is refitted to compute the R^2

and Q^2 values, and the intercept values of R^2 <0.3 and Q^2<0.05 indicate no chance of correlation in a statistically significant model [71].

3.1.8 *Screening of the Pesticide Properties DataBase* (PPDB)

We have collected 1903 chemical data from the Pesticide Properties DataBase (PPDB) which is accessible through the PPDB website (http://sitem.herts.ac.uk/aeru/ppdb/). KNIME curation was carried out using a KNIME workflow to eliminate any duplicates, inorganic salts, and mixtures [26]. As a result of the KNIME curation process, certain compounds have been eliminated. After curating the dataset, the enduring 1694 compounds were screened to verify model reliability. The descriptors of the molecules were calculated using the same procedure that was used in q-RASTR modeling as discussed earlier. The individual PLS q-RASTR models were used to make predictions, assisted by the PRI tool [17] which provided a reliable indication of the prediction's accuracy. The tool assesses the reliability of predictions using AD and furnishes qualitative prediction indicators categorized as 'Good', 'Moderate', and 'Bad. A detailed flow diagram of this study has been given in **Fig. 3.1.**

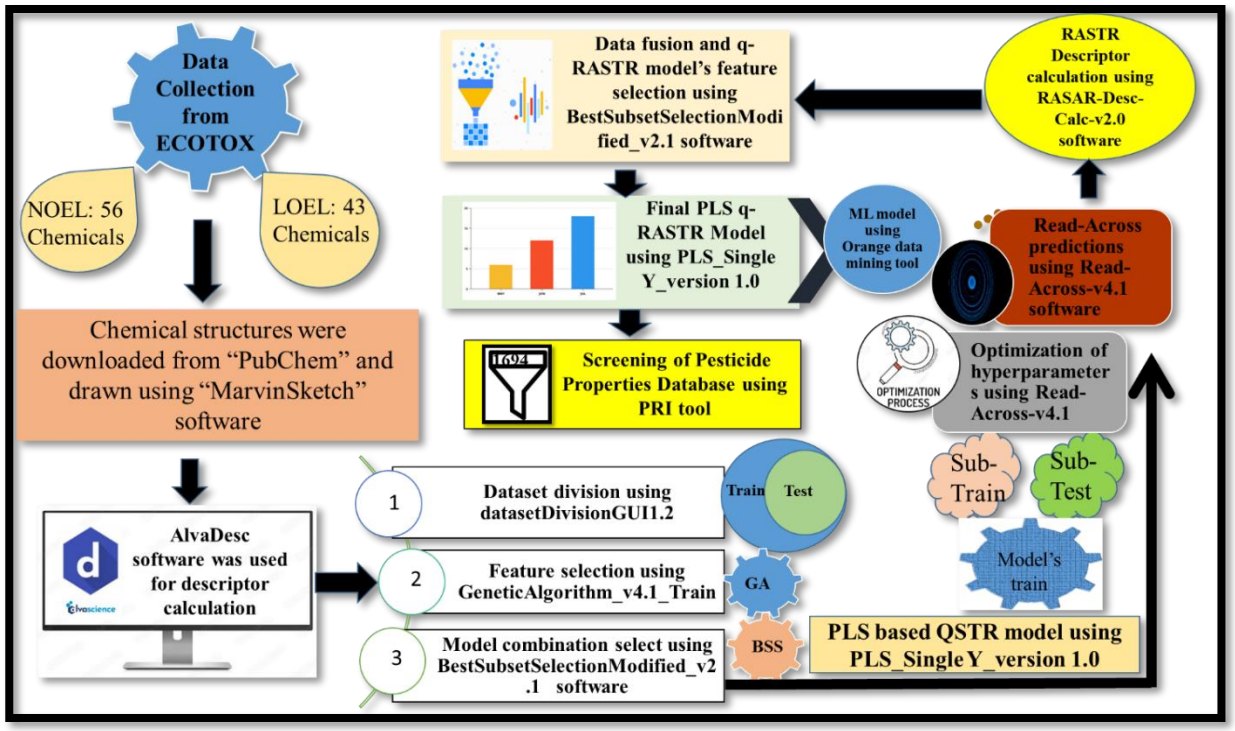


Figure 3.1. Schematic workflow of q-RASTR model development.

3.2. Study 2

3.2.1. Data collection and preparation

A set of non-cholinesterase inhibitors consisting of 733 pesticides was collected from the literature [72]. The data was curated to remove duplicates and treat missing or inconsistent values using

knime workflow (https://www.knime.com/cheminformatics-extensions). Some compounds were also omitted from the dataset due to the high residual values. After processing the data, we obtained toxicity information for 480 unique pesticides on avian species. The toxicity data are expressed as -log{HD_5(50%)} or pHD_5(50%) in molar units throughout the manuscript.

Table 3.4. Compounds name with respective experimental pHD_5 values.

Sl.No	Compound	pHD5
1	Methyl bromide	1.049
2	Dichlobutrazol	-0.452
3	Citronella oil	-0.229
4	Pretilachlor	-0.491
5	Azadirachtin	0.359
6	Methyl isocyanate	0.747
7	Fenvalerate	0.116
8	Methyl chloroform	-0.340
9*	Acibenzolat (CGA)	-0.043
10	Isobenzan	2.981
11	Triforine	-0.252
12	Benfuresate	-0.863
13*	Nemagon	1.154
14	Metconazole	0.542
15	Resmethrin	0.749
16	Dicofol	0.709
17	Trimethoxysilyl quats	0.478
18*	Flamprop-methyl	0.500
19*	Bromethalin	2.843
20	Carbendazim	-0.410
21	Butralin	-0.149
22*	Triazoxide	1.374
23	Diflufenzopyr (BAS 654)	-0.211
24	Rotenone	0.270
25	Coumatetralyl	0.180
26	Haloxyfop-P-methyl	0.446
27	Pyrithione	0.701
28	Chloroneb	-0.367
29*	Propisochlor	0.631
30	Furilazole	1.078
31	Fluroxypyr	0.088
32	Lignasan BLP	-0.369
33*	Glutaraldehyde	-0.224
34	Dithio-3-one,4,5-dichloro	0.814
35	Ferimzone	0.432
36	Oryzalin	0.820
37	Quizalofop-P-tefuryl	0.283
38	Hymexazol	-0.249
39*	Difethialone	3.241
40	Dichlorprop-P	0.758

41	Diflumetorim	0.586
42	Azoxystrobin	0.240
43	Dichloropropene	0.798
44	Metalaxyl	0.496
45	4-Chloro-3,5-xylenol	-0.246
46	Tolclofos-methyl	-0.207
47	Fipronil	2.473
48*	Pentachlorophenol (PCP)	0.720
49*	Chlormequat	0.360
50	Erioglaucine/tartrazine	0.504
51*	Cycloxydim	0.194
52	MCPA	0.709
53	Buprofezin	-0.348
54*	Sodium dichloro-S-tri-azine trione	-0.014
55	Diflubenzuron	-0.487
56	Alachlor	-0.088
57	Tolylfluanid	-0.143
58*	Anilofos	0.133
59	Thiazafluron	0.973
60	Hexazinone	-0.015
61	Diclofop-methyl	-0.364
62	3-lodo-2-propynyl butyl-carbamate	0.453
63	Bensulide	0.393
64	Oxyfluorfen	-0.230
65	Flucythrinate	0.216
66*	Metribuzin	0.708
67	Atrazine	-0.278
68*	Propyzamide	-0.457
69*	Fuberidazole	0.586
70*	Glufosinate-ammonium	-0.108
71	Prodiamine	0.127
72	Benoxacor	0.081
73	Azadioxabicyclooctane	-0.140
74	Isoproturon	-0.182
75*	Pyridate	0.213
76	Bioresmethrin	-0.175
77	Prochloraz	0.706
78	Thiazopyr	0.251
79*	Dikegulac-sodium	-0.137
80*	Bioallethron S-cyclo-pentenyl isomer	-0.236
81	Paradichlorobenzene	-0.104
82	Linuron	0.583
83	Sodium 2-mercaptoben-Zothiolate	-0.177
84*	Isocyanuric acid	-0.155
85	Zirame	0.611
86	Calcium polysulfide	0.392
87	Endrin	2.706
88	Tri-allate	0.066
89	Trifluralin	0.135

90	Copper triethanolamine	-0.117
91	Propaquizafop	0.308
92	Daminozide	-0.288
93*	Hydroxypropyl methane thiosulfonate	0.572
94	Fenclorim	0.669
95	Benalaxyl	-0.401
96	Sethoxydim	-0.168
97*	Triticonazole	0.136
98	Triadimenol	-0.274
99	2,3,6-TBA	0.477
100	Clofentezine	-0.212
101	Flubenzimine	-0.016
102	Tetradifon	-0.212
103*	Acrinathrin	0.540
104*	Ioxynil	1.047
105*	Nitrapyrin	-0.053
106	Cyfluthrin	-0.048
107*	Dicamba	0.550
108	Glyphosate	-0.138
109	Etoxazole	0.271
110	Hexythiazox	-0.136
111*	Chlorofenizon	-0.166
112	Kasugamycin	-0.008
113	2,4,5-T	0.655
114	Ethalfluralin	0.157
115	Cytokinin	-0.132
116*	Clodinafop-propargyl	0.295
117*	Dieldrin	1.963
118*	Fenoxycarb	-0.351
119	Oxabentrinil	-0.277
120	Dimethoxane	-0.024
121	Triethylhexahydro-s-triazine	0.557
122	Napropamide	0.541
123*	CGA 50 439	0.447
124	Sodium chlorite	-0.072
125	Acequinocyl (AKD-2023)	0.299
126*	Chlozolinate	-0.377
127	Sulcofuron-sodium	0.541
128*	Furalaxyl	-0.284
129	Ethidimuron	0.647
130	Parachlorometacresol	-0.098
131	Propamocarb	-0.233
132*	DCDMH (1,3-Dichloro-5,5-dimethylhydantoin)	-0.170
133	Toxaphene	1.598
134	Etridiazole	0.580
135	Myclobutanil	0.688
136	Bifenazate (D2341)	0.355
137	Hydramethylnon	0.346
138	Acifluorfen-sodium	0.559

139	Nonanoic acid	-0.218
140	Ametryn	-0.170
141	Butachlor	-0.156
142	Fenuron	-0.151
143	2,4-D Isooctyl ester	0.718
144*	Bilanafos	0.106
145	Potassium dimethylthio-carbamate	-0.084
146	Prosulfuron	0.420
147	Fluquinconazole	0.257
148	TCMTB	0.492
149	Thiabendazole	-0.113
150	Isoprothiolane	-0.169
151*	Methoxychlor	0.074
152	Sebuthylazine	-0.163
153*	Bromoxynil heptanoate	0.970
154*	Benfluralin	0.207
155	Oxadiazon	0.253
156	Fenoxaprop-P-ethyl	0.192
157	Tebufenozide	0.150
158*	Haloxyfop ethoxyethyl	0.321
159	Edifenphos	0.616
160	Difenacoum	1.679
161	2-Benzyl-4-chlorophenol	-0.125
162	Bifenox	-0.076
163	Cymoxanil	-0.072
164*	Busan 77	0.659
165	Metobromuron	0.274
166	Fluazifop-butyl	-0.289
167	Dodine (doguadine)	0.315
168	Propenamide	0.567
169	Vernolate	-0.218
170	Fenpropimorph	-0.141
171	Carbetamide	0.046
172*	Cyhalothrin	-0.030
173	Prallethrin	0.426
174	Captan	1.075
175	Penconazole	0.167
176*	2,4-D sodium	0.052
177	Trichloro-s-triazinetrione	0.200
178	Fenoxaprop-P	0.157
179	Mineral (including parafin)	0.065
180*	Dicloran	0.171
181	Fenbuconazol	0.096
182	Sulfluramid	1.493
183	Cloquintocet-mexyl	0.208
184	Forchlorfenuron	-0.023
185	Acetochlor	0.448
186	Tebuthiuron	0.492
187	Flutolanil	0.191

188	Epoxiconazol	0.152
189	Benzene Hexachloride	1.365
190	Flutriafol	-0.204
191	Flumetralin	0.275
192	Dithiopyr	0.186
193	Endosulfan	1.630
194	Thidiazuron	-0.222
195	BCDMH	0.109
196	Imazaquine	0.144
197	Lufenuron	0.390
198	Cosan 145	0.176
199*	Trinexapac-ethyl	0.117
200	Diflufenican	0.111
201	Metamitron	0.058
202*	Prometon	-0.067
203	Tralomethrin	0.358
204	Nicotine	2.193
205*	Nabam	0.012
206	Quintozene	0.063
207*	Triadimefon	-0.118
208	Chloroprop-sodium	-0.039
209*	Benazolin-ethyl	-0.211
210	Fenpiclonil	0.658
211*	Bis(trichloromethyl) sulfone	0.142
212*	Grotan	0.094
213	Lambda-Cyhalothrin	0.022
214	Azimsulfuron	0.258
215	Nitenpyram	0.214
216	MCPP Isooctyl ester	0.097
217	1,3-dibromo-5,5- dimethylhydantoin (DBDMH)	-0.008
218	Fluazinam	0.214
219*	Terbuthylazine	0.244
220	Bitertanol	-0.096
221	POE Isooctadecanol	0.270
222	Zineb	-0.004
223	Cycloate	-0.064
224	Ethirimol	0.014
225*	Fenothiocarb	0.249
226	Iprodione	0.319
227	Chlorthal-dimethyl	0.104
228	Haloxyfop	0.141
229	Dipropyl isocinchomero-nate	0.205
230*	SDDC	0.019
231	Napthaleneacetic acid	-0.047
232	PNMDC/DCDMC	-0.034
233	Neurolidol	-0.003
234	Pymetrozine	0.019
235*	Daimuron	0.063
236	Tefluthrin	0.370

237	Diiodomethyl p-tolyl sulfone	0.161
238*	Diniconazole	0.259
239	Cyprodinil	0.034
240	Clopyralid	-0.001
241	Tralkoxydim	0.054
242	Etofenprox	0.291
243	Flucycloxuron	0.400
244	DBNPA	1.024
245	Teflubenzuron	0.237
246	Bensulfuron-methyl	0.297
247	Oxadixyl	0.271
248	Gibberellic acid	0.122
249	2,4-D diolamine	0.505
250	Mecoprop	0.417
251	Mefenpyr-diethyl	0.286
252*	Hexaconazole	-0.095
253	Nicosamide-olamine	0.149
254	Fluoroglycofen-ethyl	0.086
255	Bronopol	0.610
256	Indole-3-butyric acid	-0.089
257*	Tetramethrin	0.079
258	Molinate	-0.058
259	Lindane	1.442
260	Potassium salts of fatty acids	-0.165
261	Quizalofop-ethyl	0.253
262	Pyrazosulfuron-ethyl	0.200
263*	Quinmerac	-0.020
264*	Fenpyroximate	0.307
265*	Quinclorac	0.071
266*	Tribufos	0.789
267	Strychnine	2.507
268*	Vinclozolin	-0.008
269	Heptachlor	2.032
270	Benzisothiazolin-3-one	0.324
271*	Cypermethrin	-0.143
272*	Warfarin	0.409
273	Cinosulfuron	0.331
274*	Flazasulfuron	0.324
275	Guanidine (iodine free base)	0.206
276	TDE	0.892
277	Primisulfuron-methyl	0.321
278	Dimepiperate	0.135
279	Triclosan	0.320
280	Cinmethylin	0.041
281	Fluometuron	0.081
282	Novaluron	0.408
283	Uniconazole	0.183
284	MCPA-thioethyl	-0.073
285	Triflusulfuron	0.310

286	Hexaflumuron	0.346
287	Ethametsulfuron-methyl	0.196
288*	Mepronil	0.112
289	Dichloropropene/	0.479
290	Imazapyr	0.067
291	Triflumuron	0.237
292	Flumequine	0.432
293*	Diuron	0.082
294	Pyridaben	0.116
295	Esfenvalerate	0.505
296	Pyriminobac-methyl	0.272
297*	Dazomet	0.483
298*	Terrazole	0.395
299*	Metiram	-0.074
300*	Trans-1,2-bis(n-propyl sulfonyl ethene	0.096
301	Amidosulfuron	0.172
302	Imibenconazole	0.245
303	Bromoxynil	1.106
304	Bromoxynil Phenol	1.106
305	Difenoconazol	0.293
306	Diethofencarb	0.058
307	Chloretazate	0.148
308*	SZI-121	0.198
309	Propiconazole	0.062
310*	Chlorsulfuron	-0.129
311*	Paclobutrazol	0.182
312	Dimethenamid (SAN	0.095
313	1,2-Benzenedicarbox-	0.276
314	Fluvalinate	0.237
315*	Tau-Fluvalinate	0.237
316	Flufenoxuron	0.323
317*	Ethofumesate	-0.218
318*	Bispyribac-sodium	0.217
319	Folpet	0.399
320*	Oxasulfuron	0.273
321	Fluoxypyr-meptyl	0.247
322*	Chlorhexidine diacetate	0.335
323	Capric acid/pelargonic	-0.115
324*	Farnesol	-0.003
325	Quizalofop	0.219
326	Norflurazon	0.365
327	Triflusulfuron-methyl	0.323
328	TFM (4-Nitro-3-[trifluoromethyl]phenol)	0.672
329	Terbacil	-0.083
330	6-Benzylaminopurine (N6-Benzuladenine)	0.084
331	4,4-Dimethyloxazolidine	0.042
332	Fenazaquin	0.197
333	Metazachlor	0.076
334*	Pefurazoate	0.177

335	Tebufenpyrad	0.205
336*	Fluridone	0.152
337	Metsulfuron	0.182
338	Thifensulfuron	0.189
339	Muscalure	0.143
340	Dimethipin	0.395
341	Triflumizole	0.074
342	Bromoxynil octanoate	1.396
343	Isouron	-0.041
344	Pencycuron	0.073
345*	Metosulam	0.278
346	Thenylchlor	0.144
347	Tridemorph	0.235
348	Bromonitrostyrene	0.675
349	2,4-D	0.221
350	Halosulfuron-methyl	0.221
351*	Paranitrophenol	0.317
352	Cycloprothrin	0.000
353*	Propineb	-0.333
354*	Clofencet	0.202
355*	Asulam sodium	-0.269
356	2,4-D Butotyl	0.141
357	BHAP (Bromohydroxya-	0.447
358	Chloramben	0.111
359	Cafenstrole	-0.138
360	Hydrogen cyanamide	0.165
361	Methoprene	0.207
362	Oxazolidine E	0.091
363	Imazethabenz-methyl	0.110
364*	Imazamethabenz-methyl	0.110
365	Nicosulfuron	0.295
366*	Polychlorocamphanes Potassium salt of oleic acid	0.058
367*	Flurazole	0.043
368	Mepanipyrim	-0.021
369	Copper sulfate (basic)	0.309
370*	Pyriproxyfen	0.189
371	Chlordecone	1.269
372	Codlemone	-0.137
373	Karbutilate	-0.317
374*	Imazapic (AC 263,222)	0.090
375	Oxine-copper	0.099
376	Thiophanate-methyl	-0.149
377*	Thiram	0.815
378	Monolinuron	0.073
379*	Aldrin	2.501
380	Pentoxazone	0.132
381	Tebuconazole	-0.052
382*	Bioban P-1487	0.291
383	Calcium tetrathiocarba-mate	0.280

384*	Benzyl benzoate	-0.039
385	Piperonyl butoxide	0.112
386	Pyrazophos	1.148
387	ADBAC	0.826
388	Halfenprox	0.339
389	Polyethoxylated aliphatic	0.065
390	Fluoroglycofen	0.292
391	Clomazone	-0.037
392	Diquat (dibromide)	1.015
393	Z-11-Hexadecanol	0.011
394	Bifenthrin	0.315
395	Lactofen	0.200
396	Esprocarb	0.138
397	Tetradec-11-en-1-yl	0.008
398	Endothall	0.805
399*	Dimethomorph	0.270
400*	Methyl nonyl ketone	-0.138
401	Bioallethrin	0.108
402	Zinc oxide	0.093
403	Phenothrin [(1R)-trans-isomer	0.080
404	2-(Octylthio)ethanol	-0.090
405	Flumetsulam	0.095
406	Fluxofenim	0.125
407*	Phenmedipham	0.001
408	Sodium 2-phenylphenate	0.163
409	2-Phenylphenol	0.163
410	Thifensulfuron-methyl	0.171
411*	Chlorfluazuron	0.349
412	Anthraquinone	0.033
413*	Chlorimuron-ethyl	0.234
414*	Rimsulfuron	0.429
415	Thiobencarb	0.059
416	N,N-Diethyl-M-Toluamide	0.078
417	Chloroxuron	-0.004
418	ZXI 8901	1.212
419	Butoxypolypropylene glycol	-0.138
420*	2,4-DB	0.145
421	Methabenzthiazuron	0.360
422	Flurprimidol	0.129
423	Alloxydim-sodium	0.053
424	Pyrimethanil	-0.019
425	Sulcotrione	0.258
426	Butroxydim	0.390
427	Phosacetim	2.962
428*	Desmedipham	0.065
429	Fomesafen	-0.041
430	Pyrazole	0.156
431	Allethrin	0.196
432	Brodifacoum	2.810

433	Metolachlor	0.070
434	Phenyl-indole-3-thiobutyrate	0.134
435*	Brofenprox	0.374
436*	DMPA	1.094
437	Propham	-0.032
438	Dichlone	0.071
439	Sodium dodecylbenzene-sulfonate	0.315
440*	Azafenidin	0.160
441*	Tribenuron	0.164
442	Silafluofen	0.326
443	Imazethapyr	0.112
444*	DDT	0.460
445	PHMB	-0.054
446	Imazosulfuron	0.246
447	Fluazuron	0.386
448	Procymidone	-0.351
449*	Flusilazole	0.314
450	DTEA	0.001
451	Bensultap	3.022
452*	Endothall (dimethylal-	0.687
453	Acetates of Z/E 8-dodecenyl and Z 8-dodecenol	-0.011
454	Pyrifenox	0.152
455	Flumiclorac-pentyl	0.210
456	Chlorpropham	0.044
457	Orbencarb	0.093
458	Fenoxaprop	-0.120
459	Tridec-4-en-1-yl acetate	0.050
460*	Triasulfuron	0.254
461*	TEPA	1.590
462	Clethodim	0.190
463	Methyl anthralinate	0.264
464	Tribenuron-methyl	0.180
465	Sulfometuron-methyl	-0.121
466	Flupyrsulfuron-methyl-	0.332
467	Irgarol	-0.013
468*	Fenoxaprop-ethyl	-0.085
469	Isoxaflutole	0.239
470	DMDM hydantoin	0.124
471	Sulfentrazone	0.171
472*	Diafenthiuron	0.392
473	Dimethirimol	0.014
474	Imazamox	0.154
475	Pebulate	0.023
476*	Isoxaben	0.156
477	Maneb	-0.215
478*	Iodine complex	0.127
479*	Prosulfocarb	0.033
480*	Bupirimate	-0.183

*** Test set compounds**

3.2.2. Descriptor calculation

A comprehensive set of molecular descriptors for each compound was calculated using alvaDesc software [64]. These descriptors included physicochemical properties, structural features, and electronic properties. Redundant and non-informative descriptors were eliminated based on correlation analysis and feature importance metrics.

3.2.3. Dataset division

The modeling process involves dividing the data into a training set for model development and a test set for model validation. In this study, various approaches were used for the data set division namely Kennard stone, activity property-based, and Euclidean distance methods using Dataset Division GUI 1.2 software (http://teqip.jdvu.ac.in/QSAR_Tools/). The optimal division was achieved using the activity property-based approach.

3.2.4. Feature selection and development of the QSTR model

Feature selection is a technique that reduces the dimensionality of the feature space by eliminating noisy and insignificant descriptors. To develop the robust, interpretable model, the choice of an appropriate descriptor is important. In the present study, we conducted a stepwise regression (using Minitab 14 software) and selected some descriptors. After removing the selected descriptors obtained from the first stepwise regression run, we repeated the stepwise regression using the remaining pool of descriptors. We repeated the same procedure and selected a manageable number of descriptors to create a reduced pool. The obtained reduced pool of descriptors was subjected to best subset selection to identify the most significant descriptors for model building using Best-Subset selection 2.1. software (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). The PLS regression approach was adopted to construct the final QSTR models.

3.2.5. Read-Across and calculation of the RASTR descriptor

Read-across approach is quite different from the QSAR/QSTR approach. Read-across assumes similar structural features in two compounds lead to the same biological activities. Optimization of hyperparameters for obtaining the read-across prediction is essential. The training set was divided into sub-train and sub-test sets for the weightage average prediction. Based on the quality of prediction for the validation set, laplacian kernel-driven similarity with σ=0.75; γ=0.75, and 10 close training compounds were chosen as hyperparameters. During the hyperparameters optimization, MAE-based (95%) criteria and external metrics (Q^2_{F1} and Q^2_{F2}) were used for the selection. To perform q-RASTR modeling, RASTR descriptors were calculated using "RASAR Descriptor Calculator v2.0" tool (available from:

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) for both the training and test compounds with optimized hyperparameters based on similarity and error.

3.2.6. q- RASTR feature selection and model development

A set of 15 descriptors was computed using optimized hyperparameters for the individual training set and the test set. The RASTR descriptors that were calculated earlier were combined with the model descriptors and the resulting pool was analyzed by the BestSubsetSelectionModified_v2.1 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) for q-RASTR model development. The best features were then used to create the final PLS-based q-RASTR model using the PLS_Single Y_version 1.0 software.

3.2.7. Statistical validation of the constructed model

This study employs various statistical validation approaches to measure robustness and prediction accuracy, establishing the significance and reliability of the constructed model using standard validation metrics. For statistical quality assessment as well as internal validation, we calculated metrics such as the determination coefficient (R^2), leave one out cross-validated correlation coefficient (Q^2_{LOO}), and MAE_{train}. Internal validation metrics are not true assessments of the robustness and predictivity of the model. Therefore, the developed models were validated using test set compounds employing various external validation parameters such as Q^2_{F1}, Q^2_{F2}, and MAE_{test}. The approved threshold value for $Q^2_{(LOO)}$, Q^2_{F1}, and Q^2_{F2} is 0.5.

3.2.8. Screening of the PPDB database

We obtained 1903 chemicals data from the PPDB database (http://sitem.herts.ac.uk/aeru/ppdb/). We used a KNIME workflow to curate the dataset, eliminating duplicates, inorganic salts, and mixtures. The dataset was curated and 1694 compounds were selected for screening to ensure the model's reliability. We calculated the molecules' descriptors using the same process as in q-RASTR modeling. We used the q-RASTR model and the PRI tool to make predictions and assess their reliability. The tool evaluates the quality of predictions using AD and provides qualitative prediction indicators, such as 'Good', 'Moderate', and 'Bad'.

3.2.9. Applicability domain (AD) study

AD is a specific region in chemical space where confident predictions can be made based on model descriptors and responses. To make an accurate prediction, the target compounds should closely resemble the training compounds in terms of structure. Therefore, it is crucial to validate the applicability domain for the statistical model as advised by OECD Principle 3 ("Validation of (Q)SAR Models - OECD," 2004). To adhere to the OECD guidelines, an applicability domain

analysis of the PLS-based q-RASTR model was conducted using DModX technique implemented in Simca-P software at a 99% confidence level.

3.2.10 Y-randomization study

Y-randomization study has been performed to analyze and confirm whether the developed models are produced by any chance. Y-randomization plots are generated for final PLS-based models through the SIMCA-P software. In randomization, the dependent variables are scrambled randomly while keeping the descriptor matrix constant, and by using the same set of variables from the original set, new models are 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 chance. The workflow for this entire study has been illustrated in **Figure. 3.2**.

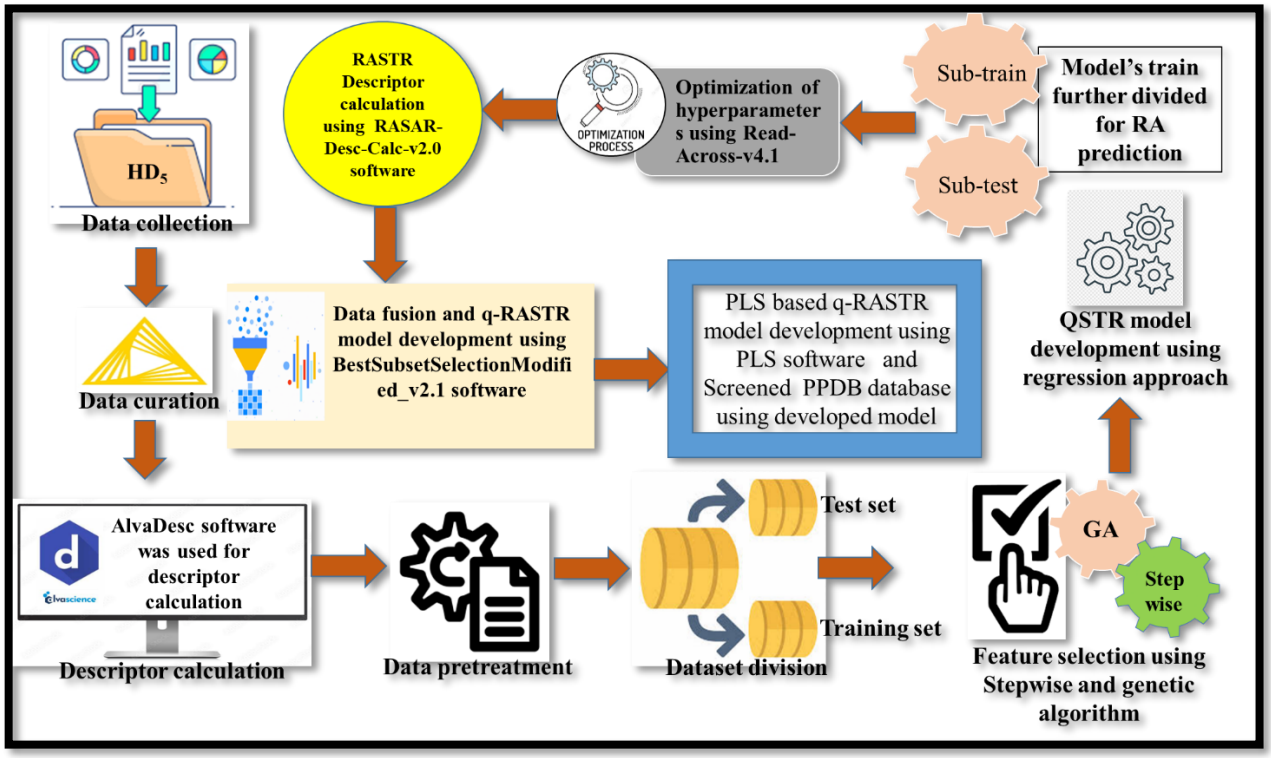


Figure 3.2. Schematic workflow of the q-RASTR model development.

3.3 Study 3

3.3.1. 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 [73] which was originally collected from the EPA, Ecotox database (http://cfpub.epa.gov/ecotox/). In this study; 556 pesticides for BQ and 117 pesticides for JQ, were taken for the development of the model. The toxicity endpoint values range from 0.082 to 4.957 in

BQ, and 0.162 to 4.968 in JQ. 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 [74]. 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 3.5. Compounds smile with respective experimental pLC_{50} values for BQ.

Sl.No	Canonical_smiles	pLC_{50}
1	COP(=O)(OC)OC(=CC(=O)N(C)C)C	4.261
2	COP(=S)(OC)Oc1ccc(cc1)[N+](=O)[O-]	4.958
3	COP(=S)(OC)Oc1ccc(SC)c(C)c1	4.133
4	CCN(CC)C(=O)\C(=C(/C)\OP(=O)(OC)OC)\Cl	4.097
5	CCOP(=S)(OCC)Oc1ccc(cc1)S(=O)C	3.945
6	CCOP(=O)(NC(C)C)Oc1ccc(SC)c(C)c1	3.902
7	CCCSP(=O)(OCC)SCCC	3.866
8*	CCCSP(=O)(OCC)Oc1ccc(Br)cc1Cl	3.817
9*	Nc1c(c(nn1-c1c(Cl)cc(cc1Cl)C(F)(F)F)C#N)S(=O)C(F)(F)F	3.716
10	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)cc2)cc1	3.705
11	CCOP(=S)(OCC)Oc1ccc2C(=C(Cl)C(=O)Oc2c1)C	3.630
12	CNC(=O)O\N=C(/SC)\C(=O)N(C)C	3.609
13	ClC1C=CC2C1C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	3.608
14	CCOP(=S)(OCC)Oc1cnccn1	3.582
15	COP(=S)(OC)Oc1nc(Cl)n(n1)C(C)C	3.569
16*	COP(=O)(N)SC	3.526
17	CCCSP(=S)(OCC)Oc1ccc(SC)cc1	3.513
18	CCOCn1c(c2ccc(Cl)cc2)c(C#N)c(Br)c1C(F)(F)F	3.490
19	CN(c1c(Br)cc(Br)cc1Br)c2c(cc(cc2C(F)(F)F)[N+](=O)[O-])[N+](=O)[O-]	3.440
20	CCOP(=S)(NC(C)C)Oc1ccccc1C(=O)OC(C)C	3.377
21*	CCOP(=S)(OCC)OC(Cl)C(Cl)(Cl)Cl	3.356
22*	[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)C(F)(F)F	3.356
23	CCOP(=S)(OCC)Oc1cc(C)nc(n1)C(C)C	3.337
24	CCOP(=O)(SC(C)CC)N1CCSC1=O	3.319
25*	CCOP(=S)(OCC)SCSC(C)(C)C	3.305
26*	CCOP(=S)(CC)Sc1ccccc1	3.268
27	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)C(F)(F)F	3.266
28	COP(=S)(OC)Oc1ccc(c(C)c1)[N+](=O)[O-]	3.247
29	CCOP(=S)(OC(C)C)Oc1cnc(nc1)C(C)(C)C	3.222

30	Clc1ccc(cc1)C(C(=O)C2C(=O)c3ccccc3C2=O)c4ccccc4	3.190
31	CNC(=O)CSP(=S)(OC)OC	3.184
32	CCOP(=S)(OCC)Oc1ccc(cc1)[N+](=O)[O-]	3.177
33	CCN(CC)c1nc(C)cc(OP(=S)(OC)OC)n1	3.169
34	COC1=NN(CSP(=S)(OC)OC)C(=O)S1	3.130
35	C1C2C(C(C1Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	3.093
36	CCNP(=S)(OC)O\C(=C\C(=O)OC(C)C)\C	3.038
37	CCOP(=S)(Oc1ccc(cc1)[N+](=O)[O-])c2ccccc2	2.967
38	CCOC(=O)CSc1nc(nn1C(=O)N(C)C)C(C)(C)C	2.884
39*	CCOP(=S)(OCC)SCSCC	2.844
40	COP(=S)(OC)SCN1N=Nc2ccccc2C1=O	2.813
41	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	2.802
42	CNC(=O)Oc1cccc2CC(C)(C)Oc12	2.796
43	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.764
44*	CCS(=O)CCSP(=O)(OC)OC	2.754
45	Clc1ccc(c(Cl)c1Cl)c2cccc(Cl)c2Cl	2.733
46	ClC1=C(Cl)C2(Cl)C3COS(=O)OCC3C1(Cl)C2(Cl)Cl	2.704
47	CCOP(=S)(OCC)SCCSCC	2.692
48*	CC(=O)CC(C1=C([O-])c2ccccc2OC1=O)c1ccccc1	2.692
49*	Clc1ccc(c(Cl)c1Cl)c2ccc(Cl)c(Cl)c2Cl	2.684
50	FC(F)(F)c1ccc(OCCCOc2c(Cl)cc(OCC=C(Cl)Cl)cc2Cl)nc1	2.644
51*	CC1(C)CNC(=NN=C(\C=C\c2ccc(cc2)C(F)(F)F)\C=C\c3ccc(cc3)C(F)(F)F)NC1	2.639
52	CCOP(=O)(OCC)SCCSCC	2.637
53	CC(C)(C)C(O)C(Oc1ccc(cc1)c2ccccc2)n3cncn3	2.621
54*	ClC(=C(c1ccc(Cl)cc1)c2ccc(Cl)cc2)Cl	2.586
55	CC(C1CC1)C(O)(Cn2cncn2)c3ccc(Cl)cc3	2.553
56	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	2.553
57*	C1(C(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl	2.518
58	CCS(=O)(=O)c1cccnc1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	2.508
59	CCCCCCCC(=O)Oc1c(Br)cc(cc1Br)C#N	2.487
60	CNC(=O)Oc1cc(C)c(N(C)C)c(C)c1	2.484
61*	CCC1CN(CCO1)c2ncc(cc2C#N)[N+](=O)[O-]	2.479
62	CNC(=O)Oc1cccc2OC(C)(C)Oc12	2.474
63	CC1(C)CCC(Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	2.472
64	CNC(=O)Oc1cc(C)c(SC)c(C)c1	2.457
65	C\C(=N/NC(=O)Nc1cc(F)cc(F)c1)c1ncccc1C([O-])=O	2.429
66*	C1CN2CC3=CCOC4CC(=O)N5C6C4C3CC2C61C7=CC=CC=C75	2.427
67	Clc1cc(Cl)cc(c1)c2cc(Cl)cc(Cl)c2	2.395
68	ClC(Cl)(Cl)C(NC=O)N1CCN(CC1)C(NC=O)C(Cl)(Cl)Cl	2.372
69	COC(=O)C=C(C)OP(=O)(OC)OC	2.351
70*	ClC1(Cl)C2(Cl)C3(Cl)C4(Cl)C(Cl)(Cl)C5(Cl)C(Cl)(C1(Cl)C35Cl)C24Cl	2.337
71	CCC(C)c1cccc(OC(=O)N(C)Sc2ccccc2)c1	2.330
72	CCCCSP(=O)(SCCCC)SCCCC	2.316

73	CC1(C)[C@H](\C=C(/Cl)\C(F)(F)F)[C@@H]1C(=O)O[C@H](C#N)c2cccc(Oc3ccccc3)c2	2.281
74*	Cc1cccc2sc3nncn3c12	2.277
75	CC1(C)CCCC(C1)=CC=O	2.275
76*	CCOP(=S)(OCC)SCN1C(=O)Oc2cc(Cl)ccc12	2.258
77	COP(=O)(OC)OC(Br)C(Cl)(Cl)Br	2.255
78*	C\C=C\C(=O)Oc1c(CCCCCC(C)C)cc(cc1[N+](=O)[O-])[N+](=O)[O-]	2.253
79	COP(=S)(OC)Oc1nc(Cl)c(Cl)cc1Cl	2.245
80	Nc1c(Cl)cc(cc1Cl)[N+](=O)[O-]	2.230
81	[O-][N+](=O)N=C1NCCN1Cc2ccc(Cl)nc2	2.221
82	[O-][N+](=O)NC1=NCCN1Cc1ccc(Cl)nc1	2.221
83*	COc1ccc(cc1NNC(=O)OC(C)C)c2ccccc2	2.208
84	CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(C(F)(F)F)Cl)C	2.207
85	[O-]C(=O)CF	2.200
86	CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C(C(Br)(Br)Br)Br)C	2.198
87	CC1=C(C(=O)CC1OC(=O)C2C(C2(C)C)C=C(C)C)CC=C	2.173
88	O=C1NSc2ccccc12	2.171
89	CC(C)(C)C(=O)C1C(=O)c2ccccc2C1=O	2.169
90	CNC(=O)ON=C(C)SC	2.169
91	ClC(Cl)C(c1ccc(Cl)cc1)c2ccc(Cl)cc2	2.167
92*	CON(C)C(=O)Nc1ccc(Cl)c(Cl)c1	2.166
93	COP(=O)(NC(=O)C)SC	2.156
94	CN1C(=O)ON(C1=O)c2ccc(Cl)c(Cl)c2	2.155
95*	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
96	Clc1ccc(c(Cl)c1)c2cccc(Cl)c2Cl	2.144
97*	Oc1c(Br)cc(cc1Br)C#N	2.139
98	CCCCCCCCCCCCCC[N+](C)(C)Cc1ccccc1	2.136
99	CN(C)C(=O)Nc1ccc(Cl)c(Cl)c1	2.130
100	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	2.124
101	Cc1cc(ccc1NC(=O)c1cccc(I)c1C(=O)NC(C)(C)CS(C)(=O)=O)C(F)(C(F)(F)F)C(F)(F)F	2.118
102	CC(C)[C@H](C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1)c3ccc(OC(F)F)cc3	2.118
103*	ClC(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	2.098
104*	[O-]S(=O)(=O)c1cc(Cl)ccc1Oc2ccc(Cl)cc2NC(=O)Nc3ccc(Cl)c(Cl)c3	2.094
105	OC(c1ccc(Cl)cc1)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.090
106	CCCCOCCOC(=O)C(C)Oc1cc(Cl)c(Cl)cc1Cl	2.086
107	COC(=O)c1ccc(I)cc1S(=O)(=O)[N-]C(=O)Nc2nc(C)nc(OC)n2	2.065
108	CCCC(=O)Oc1c(Br)cc(cc1Br)C#N	2.041

109*	COc1cnc(OC)n2nc(NS(=O)(=O)c3c(OCC(F)F)cccc3C(F)(F)F)nc12	2.040
110*	Clc1cccc(n1)C(Cl)(Cl)Cl	2.034
111	Cc1ccc2nc3SC(=O)Sc3nc2c1	2.033
112*	OC(=O)C1(CC1)C(=O)Nc2ccc(Cl)cc2Cl	2.026
113*	CC=C(C)C(=O)OC1CC(C2(COC3C2C1(C(C4(C3OC5C4=C(C(C5)C6=COC=C6)C)C)CC(=O)OC)C)C)OC(=O)C	2.026
114	CC1(C)C(C=C(Cl)Cl)C1C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2	2.023
115	COC(=O)c1ccc(CNS(=O)(=O)C)cc1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	2.021
116*	CCCCOCCOC(=O)COc1cc(Cl)c(Cl)cc1Cl	2.020
117*	CCC1C(CCC2(O1)CC3CC(O2)CC=C(CC(C=CC=C4COC5C4(C(C=C(C5O)C)C(=O)O3)O)C)C)C	2.019
118	CC1C(C(C(O1)OC2C(C(C(C(C2O)O)N=C(N)N)O)N=C(N)N)OC3C(C(C(C(O3)CO)O)O)NC)(C=O)O	2.015
119	CCCCCCCC[N+](C)(C)CCCCCCCC	2.013
120	CC(C)(C)c1ccc(OC2CCCCC2OS(=O)OCC#C)cc1	2.013
121	COC(=O)c1csc(C)c1S(=O)(=O)NC(=O)N2N=C(OC)N(C)C2=O	2.006
122	Cc1cc(Cl)ccc1OCC(=O)O	2.001
123	CN(\C=N\c1ccc(C)cc1C)\C=N\c2ccc(C)cc2C	1.979
124	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	1.978
125	Cc1c(COC(=O)C2C(\C=C(/Cl)\C(F)(F)F)C2(C)C)cccc1c3ccccc3	1.978
126	CCCCCCCCCCCCCCC[P+](CCCC)(CCCC)CCCC	1.977
127*	FC(OC(F)(F)F)C(F)(F)Oc1ccc(NC(=O)NC(=O)c2c(F)cccc2F)cc1Cl	1.977
128	CO\N=C(\C1=NOCCO1)/c2ccccc2Oc3ncnc(Oc4ccccc4Cl)c3F	1.976
129	CCCN(CCC)C(=O)SCC	1.976
130	CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC	1.975
131*	CCC(=O)Nc1ccc(Cl)c(Cl)c1	1.975
132	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2cc(NC=O)ccc2C(=O)N(C)C)n1	1.963
133	CNC(=O)N(C)c1nnc(s1)C(C)(C)C	1.961
134	CCNc1nc(Cl)nc(NC(C)(C)C#N)n1	1.958
135	CN1C(=NN(C1=O)C(=O)[N-]S(=O)(=O)C2=CC=CC=C2OC(F)(F)F)OC.[Na+]	1.957
136	CCOc1nc(F)cc2nc(nn12)S(=O)(=O)Nc3c(Cl)cccc3Cl	1.956
137	CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(Br)Br)C	1.954
138	CC(C)C(Nc1ccc(cc1Cl)C(F)(F)F)C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2	1.951
139	CC(C1CCC(C(O1)OC2C(CC(C(C2O)OC3C(C(C(CO3)(C)O)NC)O)N)N)N)NC	1.949
140*	CC(Oc1cc(Cl)c(Cl)cc1Cl)C(=O)O	1.949
141*	CCS(=O)(=O)c1nc2ccccn2c1S(=O)(=O)NC(=O)Nc3nc(OC)cc(OC)n3	1.948
142	CCN(Cc1c(F)cccc1Cl)c2c(cc(cc2[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.944
143	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	1.944
144	COC(=O)c1cccc(C)c1S(=O)(=O)NC(=O)Nc2nc(OCC(F)(F)F)nc(n2)N(C)C	1.943
145	OC(=O)COc1nc(Cl)c(Cl)cc1Cl	1.942

146	COc1cc(OC)n2nc(NS(=O)(=O)c3c(OC)nccc3C(F)(F)F)nc2n1	1.939
147*	CC1(C(C1C(=O)OC(C#N)C2=CC(=C(C=C2)F)OC3=CC=CC=C3)C=C(Cl)Cl)C	1.939
148	CCCCCCC(=O)Oc1c(Br)cc(cc1Br)C#N	1.934
149	CCOC(=O)C(O)(c1ccc(Cl)cc1)c2ccc(Cl)cc2	1.934
150*	COc1cc(OC)nc(NC(=O)NS(=O)(=O)Nc2ccccc2C(=O)N(C)C)n1	1.931
151	CS(=O)(=O)c1cc(ccc1C(=O)c2cnoc2C3CC3)C(F)(F)F	1.927
152	ClC1=C(Cl)C(Cl)(C(=C1Cl)Cl)C2(Cl)C(=C(Cl)C(=C2Cl)Cl)Cl	1.927
153	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
154	Cc1nn(C)c(Oc2ccccc2)c1\C=N\OCc3ccc(cc3)C(=O)OC(C)(C)C	1.926
155	CCSC(C)CC1CC(=C(C(=NOC\C=C\Cl)CC)C(=O)C1)O	1.926
156	CC(C)C(C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1)c3ccc(Cl)cc3	1.924
157	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2CCC(F)(F)F)n1	1.924
158*	Clc1ccc(CCC(Cn2cncn2)(C#N)c3ccccc3)cc1	1.920
159*	FC(F)(F)c1cnc(CCNC(=O)c2ccccc2C(F)(F)F)c(Cl)c1	1.919
160	CCOC(=O)COc1cc(c(F)cc1Cl)c2nn(C)c(OC(F)F)c2Cl	1.917
161*	CCOC(=O)C(C)OC(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.915
162	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2ncccc2C(=O)N(C)C)n1	1.914
163*	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
164*	COC1=CC(=NC(=N1)NC(=O)NS(=O)(=O)C2=C(N=C3N2C=CC=C3)Cl)OC	1.913
165	Cl\C=C\C[N+]12CN3CN(CN(C3)C1)C2	1.912
166	CON=C(C(=O)OC)c1ccccc1CON=C(C)c2cccc(c2)C(F)(F)F	1.908
167*	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2OCCCl)n1	1.905
168	CC(=O)CC(C1=C(O)c2ccccc2OC1=O)c3ccccc3	1.898
169	Oc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.894
170	CO\C=C(\C(=O)OC)/c1ccccc1Oc2cc(Oc3ccccc3C#N)ncn2	1.890
171	CON(C(=O)OC)c1ccccc1COc2ccn(n2)c3ccc(Cl)cc3	1.890
172	COC(=O)c1c(Cl)nn(C)c1S(=O)(=O)NC(=O)Nc2nc(OC)cc(OC)n2	1.889
173*	CCOc1nc(F)cc2nc(nn12)S(=O)(=O)Nc3c(Cl)cccc3C(=O)OC	1.884
174	COc1cc(OC)nc(Oc2cccc(Oc3nc(OC)cc(OC)n3)c2C(=O)[O-])n1	1.883
175*	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(Cl)c1COCC(F)(F)F	1.882
176	O=C(C(c1ccccc1)c2ccccc2)C3C(=O)c4ccccc4C3=O	1.880
177	CCCCCCCCSC(=O)Oc1cc(Cl)nnc1c2ccccc2	1.880
178	CN1C=C(C(=O)C(=C1)c2cccc(c2)C(F)(F)F)c3ccccc3	1.879
179*	CCCCCOC(=O)COc1cc(N2C(=O)C3=C(CCCC3)C2=O)c(F)cc1Cl	1.878
180*	Clc1ccc(c(Cl)c1)C2(Cn3cncn3)CC(Br)CO2	1.877
181	CCOc1ccc(cc1)C(C)(C)COCc2cccc(Oc3ccccc3)c2	1.877
182	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc2cccc(Oc3ccccc3)c2	1.877
183	Cc1ccc(cc1)S(=O)(=O)C(I)I	1.876
184	Cc1cc(C)c(C2=C(OC(=O)C(C)(C)C)C3(CCCC3)OC2=O)c(C)c1	1.874
185	CCOC(=O)OC1=C(C(=O)N[C@@]12CC[C@@H](CC2)OC)c3cc(C)ccc3C	1.873
186	CC(C)[C@H](C(=O)O[C@H](C#N)c1cccc(Oc2ccccc2)c1)c3ccc(Cl)cc3	1.873

187*	CC12C(C=CC3(C1C(C45C3CCC(C4)(C(=C)C5)O)C(=O)O)OC2=O)O	1.873
188	CCOC(=O)C(C)Oc1ccc(Oc2cnc3cc(Cl)ccc3n2)cc1	1.873
189	Cc1nn(C)c(O)c1C(=O)c2ccc(cc2S(=O)(=O)C)C(F)(F)F	1.868
190	CCOC(=O)C(Cl)Cc1cc(N2N=C(C)N(C(F)F)C2=O)c(F)cc1Cl	1.865
191	CCCCOC(=O)C(C)Oc1ccc(Oc2ccc(cn2)C(F)(F)F)cc1	1.865
192	COP(=O)(OC)OC(=CCl)c1cc(Cl)c(Cl)cc1Cl	1.864
193*	COc1ccc(cc1OC)\C(=C\C(=O)N2CCOCC2)\c3ccc(Cl)cc3	1.864
194	CCOc1nc(NC)nc(NC(=O)NS(=O)(=O)c2ccccc2C(=O)OC)n1	1.864
195	COC(=O)c1ccccc1CS(=O)(=O)NC(=O)Nc1nc(OC)cc(OC)n1	1.864
196*	CNC(=O)Oc1cc(C)c(C)c(C)c1	1.863
197	COc1cc(OC)nc(NC(=O)NS(=O)(=O)c2ncccc2C(F)(F)F)n1	1.863
198	CC(C)C(O)(c1ccc(OC(F)(F)F)cc1)c2cncnc2	1.860
199	CCCCNC(=O)OCC#CI	1.860
200	CCOc1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.859
201*	CSC(=O)c1c(CC(C)C)c(C(=O)SC)c(nc1C(F)F)C(F)(F)F	1.854
202	Clc1ccc(cc1)S(=O)(=O)c2cc(Cl)c(Cl)cc2Cl	1.853
203	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(c1)[N+](=O)[O-]	1.852
204*	CCOc1cc(ccc1C2COC(=N2)c3c(F)cccc3F)C(C)(C)C	1.851
205*	CC(C)OP(=S)(OC(C)C)SCCNS(=O)(=O)c1ccccc1	1.850
206*	CN1CSC(=S)N(C)C1	1.848
207	COC(=O)c1c(CC(C)C)c(C2=NCCS2)c(nc1C(F)F)C(F)(F)F	1.848
208	CCc1ccc(cc1)C(=O)NN(C(=O)c2cc(C)cc(C)c2)C(C)(C)C	1.848
209*	COC(=O)c1ccccc1S(=O)(=O)NC(=O)N(C)c2nc(C)nc(OC)n2	1.847
210	CC(=CC1C(C1(C)C)C(=O)OCC2=CC(=CC=C2)OC3=CC=CC=C3)C	1.846
211	COC(=O)c1cc(Cl)cc(N)c1Cl	1.843
212	CN(C)C(=O)Oc1nc(nc(C)c1C)N(C)C	1.843
213*	CCCCCCCCCCCCCCC=CCCCCCCCC	1.842
214	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])S(=O)(=O)N)[N+](=O)[O-]	1.841
215	COc1ccc(cc1)C(c2ccc(OC)cc2)C(Cl)(Cl)Cl	1.840
216*	OC(CN1NC=NC1=S)(Cc2ccccc2Cl)C3(Cl)CC3	1.839
217	CC(C)Oc1cc(N2N=C(OC2=O)C(C)(C)C)c(Cl)cc1Cl	1.839
218	CCCCC(O)(Cn1cncn1)c2ccc(Cl)cc2Cl	1.839
219	COC(=O)c1sccc1S(=O)(=O)NC(=O)Nc2nc(C)nc(OC)n2	1.838
220	CC1=NN(C(=O)N1C(F)F)c2cc(NS(=O)(=O)C)c(Cl)cc2Cl	1.838
221	CCCCCC(C)OC(=O)COc1ccc(Cl)c2cccnc12	1.837
222	Clc1ccc(cc1)c2ccccc2NC(=O)c3cccnc3Cl	1.837
223	Cc1c(ccc(c1C2=NOCC2)S(=O)(=O)C)C(=O)c3cnn(C)c3O	1.836
224	COC(=O)c1cc(Oc2ccc(Cl)cc2Cl)ccc1[N+](=O)[O-]	1.835
225*	CC(C(=O)OCC#C)OC1=CC=C(C=C1)OC2=C(C=C(C=N2)Cl)F	1.835
226	CC(Oc1ccc(Oc2ncc(Cl)cc2F)cc1)C(=O)OCC#C	1.835
227	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(C)nc(OC)n2	1.832
228*	COc1cc(OCC#C)ccc1CCNC(=O)C(OCC#C)c2ccc(Cl)cc2	1.831
229	COc1cc(C)c(C(=O)c2c(C)c(Br)ccc2OC)c(OC)c1OC	1.829

230*	CCC(CC)Nc1c(cc(C)c(C)c1[N+](=O)[O-])[N+](=O)[O-]	1.827
231	CCc1cc(C)cc(CC)c1C2=C(OC(=O)C(C)(C)C)N3CCOCCN3C2=O	1.827
232	CCCCN(CC)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.826
233	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.826
234	CC(COc1ccc(cc1)C(C)(C)C)OS(=O)OCCCl	1.826
235	Fc1cccc(F)c1C(=O)NC(=O)Nc2ccc(Cl)cc2	1.826
236	CN(C)C(=O)Nc1cccc(c1)C(F)(F)F	1.825
237	CC(=CC1C(C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2)C1(C)C)C	1.825
238*	CCc1nn(C)c(C(=O)NCc2ccc(cc2)C(C)(C)C)c1Cl	1.825
239	CCN(CC(=C)C)c1c(cc(cc1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.824
240*	CC1(OC(=O)N(Nc2ccccc2)C1=O)c3ccc(Oc4ccccc4)cc3	1.824
241	CCC(C)(CC)c1cc(NC(=O)c2c(OC)cccc2OC)on1	1.823
242	FC(F)(F)c1cnc(CNC(=O)c2c(Cl)cccc2Cl)c(Cl)c1	1.821
243	CCCCOCCOC(=O)COc1nc(Cl)c(Cl)cc1Cl	1.820
244*	BrCC(Br)(CCC#N)C#N	1.818
245	CCCCCCCCN1SC=CC1=O	1.815
246*	CC(C)(C)N1N=CC(=C(Cl)C1=O)SCc2ccc(cc2)C(C)(C)C	1.813
247	Nc1cc(Cl)cc(C(=O)[O-])c1Cl	1.812
248*	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc2nc(C)cc(C)n2	1.812
249	CCN(CC)c1c(cc(c(N)c1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.809
250	CCOC(=O)C(C)Oc1ccc(Oc2oc3cc(Cl)ccc3n2)cc1	1.809
251	CC(Oc1ccc(Oc2ncc(cc2Cl)C(F)(F)F)cc1)C(=O)O	1.809
252	CCCCOCCOC(=O)COc1ccc(Cl)cc1Cl	1.808
253	[O-]C(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.807
254*	CCC(C)(NC(=O)c1cc(Cl)c(C)c(Cl)c1)C(=O)CCl	1.807
255	COc1nc(C)nc(NC(=O)NS(=O)(=O)c2ccccc2Cl)n1	1.804
256*	CCCCOC(=O)C(C)OC1=CC=C(C=C1)OC2=C(C=C(C=C2)C#N)F	1.803
257	CC1(C)CCC(=Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	1.803
258	CC(C)(C)C(=O)C(Oc1ccc(Cl)cc1)n2cncn2	1.802
259	CS\C(=N\OC(=O)N(C)SN(C)C(=O)ON=C(C)SC)\C	1.800
260	Fc1cc2OCC(=O)N(CC#C)c2cc1N3C(=O)C4=C(CCCC4)C3=O	1.800
261	CCCCC(CC)COC(=O)c1nc(Cl)c(Cl)c(N)c1Cl	1.799
262*	CC(C)CCCCCCCCCCCCCCOCCO	1.799
263*	CC1C(SC(=O)N1C(=O)NC2CCCCC2)C3=CC=C(C=C3)Cl	1.798
264	COc1cccc(C(=O)NN(C(=O)c2cc(C)cc(C)c2)C(C)(C)C)c1C	1.798
265*	CCCCOCCOCCOCCc1cc2OCOc2cc1CCC	1.797
266	CON=C(C(=O)OC)c1ccccc1COc2ccccc2C	1.797
267	CC(C)CCCCCOC(=O)COc1ccc(Cl)cc1C	1.796
268	COCc1c(F)c(F)c(COC(=O)[C@@H]2[C@@H](C=CC)C2(C)C)c(F)c1F	1.796
269	CC(C)C1(C)N=C(NC1=O)c2nc3ccccc3cc2C(=O)O	1.794
270	C1C=CCC2C1C(=O)N(C2=O)SC(C(Cl)Cl)(Cl)Cl	1.793
271	ClC(Cl)C(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	1.793
272	CC(C)CC(C)c1sccc1NC(=O)c2cn(C)nc2C(F)(F)F	1.793

273	CCCN(CCC)c1c(cc(cc1[N+](=O)[O-])C(C)C)[N+](=O)[O-]	1.792
274	CC(COc1ccc(Oc2ccccc2)cc1)Oc3ccccn3	1.791
275	CCCCC(CC)COC(=O)C(C)OC1=C(C=C(C=C1)Cl)Cl	1.791
276	CC(C)CCCCCOC(=O)C(C)Oc1ccc(Cl)cc1Cl	1.791
277	CC(C)Oc1cccc(NC(=O)c2ccccc2C(F)(F)F)c1	1.790
278	CCCOCC(=Nc1ccc(Cl)cc1C(F)(F)F)n2ccnc2	1.789
279	CCc1ccc(cc1)C(C(Cl)Cl)c2ccc(CC)cc2	1.789
280	CC1CCCCN1CCCOC(=O)c2ccc(Cl)c(Cl)c2	1.788
281	CCCC1COC(Cn2cncn2)(O1)c3ccc(Cl)cc3Cl	1.785
282	CN(C)C(=S)SSC(=S)N(C)C	1.784
283	CCCCCCCCN1SC(=C(CI)C1=O)CI	1.784
284	COC(=O)C(C)Oc1ccc(Oc2ccc(Cl)cc2Cl)cc1	1.783
285	CCCCOCCOCCOCc1cc2OCOc2cc1CCC	1.780
286	CC1CC2=C(C1NC3=NC(=NC(=N3)N)C(C)F)C=C(C=C2)C	1.780
287	CCOC(=O)Nc1cccc(OC(=O)Nc2ccccc2)c1	1.779
288	CCOC(=O)CN(C(=O)CCl)c1c(CC)cccc1CC	1.778
289	CC(C)CCCC(C)C\C=C\C(=C\C(=O)OCC#C)\C	1.775
290	CC(=NNC(=O)Nc1cc(F)cc(F)c1)c1ncccc1C(O)=O	1.774
291	ClC(Cl)(Cl)SN1C(=O)c2ccccc2C1=O	1.773
292	CCCCC(CC)COC(=O)COc1ccc(Cl)cc1Cl	1.773
293*	CC(C)(C)C(O)C(Oc1ccc(Cl)cc1)n2cncn2	1.772
294*	CCOC(=O)C1=NOC(C1)(c2ccccc2)c3ccccc3	1.771
295*	COC(=O)c1c(Cl)c(Cl)c(C(=O)OC)c(Cl)c1Cl	1.771
296	[O-][N+](=O)c1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.771
297	CC(=CC1C(C1(C)C)C(=O)OCN2C(=O)C3=C(C2=O)CCCC3)C	1.771
298	CC(C)NC(=O)N1CC(=O)N(C1=O)c2cc(Cl)cc(Cl)c2	1.769
299*	CCOCN(C(=O)CCl)c1c(C)cccc1CC	1.767
300	CC(C)C1CCC(Cc2ccc(Cl)cc2)C1(O)Cn3cncn3	1.767
301	CC1C(OC(=O)C2C(C=C(C)C)C2(C)C)C=C(CC=CC=C)C1=O	1.767
302	N#CSCSC#N	1.766
303	CCCC(=NOCC)C1=C(O)CC(CC(C)SCC)CC1=O	1.766
304*	CC(C)N1\C(=N\C(C)(C)C)\SCN(C1=O)c2ccccc2	1.765
305	CCN1C(=CC(=O)C(=C1c2ccc(Cl)cc2)C(=O)[O-])C	1.765
306	CCCCCCCCCC[N+](C)(C)CCCCCCCCCC	1.764
307	COc1cc(OC)nc(Sc2cccc(Cl)c2C(=O)[O-])n1	1.763
308	Oc1cc(Cl)ccc1Oc2ccc(Cl)cc2Cl	1.763
309	Cc1ccn2nc(nc2n1)S(=O)(=O)Nc3c(F)cccc3F	1.763
310	CCc1cnc(C2=NC(C)(C(C)C)C(=O)N2)c(c1)C(=O)O	1.762
311	CC1=C(C(=CC=C1)C)N([C@H](C)C(=O)OC)C(=O)COC	1.762
312	CCCCC(Cn1cncn1)(C#N)c2ccc(Cl)cc2	1.762
313*	COC(=O)c1ccc(C)cc1C1=NC(=O)C(C)(N1)C(C)C	1.761
314*	CCC1=C(C(=O)[O-])C(=O)C=NN1c2ccc(Cl)cc2	1.761
315	COCC(=O)N(N1CCOC1=O)c2c(C)cccc2C	1.760

316*	CC1=C(SCCO1)C(=O)Nc2ccccc2	1.758
317	CCC(=NOC\C=C\Cl)C1=C(O)CC(CC1=O)C2CCOCC2	1.756
318	[S-]C(=S)NCCNC(=S)[S-]	1.753
319	CC(=CC1C(C1(C)C)C(=O)OCN2C(=O)CN(C2=O)CC#C)C	1.753
320	CCCCCCCCCCCC1C(O1)CCCCC(C)C	1.752
321	Cn1cc(C(=O)Nc2ccccc2C3CC3C4CC4)c(n1)C(F)F	1.752
322*	CN1COCN(Cc2cnc(Cl)s2)C1=N[N+](=O)[O-]	1.749
323	[S-]C1Nc2ccccc2S1	1.747
324*	CNC(=O)Oc1cccc(c1)\N=C\N(C)C	1.747
325	CCCCCCCCCC[N+](C)(C)CCCCCCC(C)C	1.745
326*	Clc1ccccc1Nc2nc(Cl)nc(Cl)n2	1.741
327*	CCOC1Oc2ccc(OS(=O)(=O)C)cc2C1(C)C	1.741
328	CCON=C(CC)C1=C(O)CC(CC1=O)c2c(C)cc(C)cc2C	1.740
329	CC1=CC(=C(N=C1)C2=NC(C(=O)N2)(C)C(C)C)C(=O)[O-]	1.739
330	CC(C)C1(C)NC(=NC1=O)c1ncc(C)cc1C([O-])=O	1.739
331	Fc1ccc(Oc2ccnc3cc(Cl)cc(Cl)c23)cc1	1.739
332	COCc1cnc(C2=NC(C)(C(C)C)C(=O)N2)c(c1)C(=O)O	1.739
333*	CC(C)N(C(C)C)C(=O)SCC(=C(Cl)Cl)Cl	1.734
334*	OC(c1ccc(Cl)cc1)(c2cncnc2)c3ccccc3Cl	1.733
335	CCc1cccc(CC)c1N(COC)C(=O)CCl	1.732
336	CC(C)=C[C@H]1[C@H](C(=O)O[C@@H]2CC(=O)C(CC=C)=C2C)C1(C)C	1.731
337	CSC1=NN=C(C(=O)N1N)C(C)(C)C	1.729
338	ClC(Cl)(Cl)S(=O)(=O)C(Cl)(Cl)Cl	1.729
339*	CC(=C[C@H]1[C@H](C(=O)O[C@@H]2CC(=O)C(=C2C)CC#C)C1(C)C)C	1.728
340	CCCCCCCCC=CCCCCCCCC(=O)[O-]	1.727
341	N#CSCSc1nc2ccccc2s1	1.724
342	COC(=O)Nc1cccc(OC(=O)Nc2cccc(C)c2)c1	1.723
343*	CC1=CC(=O)C(=NN1c2ccc(Cl)cc2)C(=O)[O-]	1.722
344	COc1cc(Cl)c(OC)cc1Cl	1.719
345	CN1SC=CC1=O	1.719
346	CC(C)C1(C)N=C(NC1=O)c2ncccc2C(=O)O	1.718
347	CCOCCOCCOC(=O)Nc1nc2ccccc2[nH]1	1.718
348	CC(C)COC(=O)COc1ccc(Cl)cc1C	1.711
349	Clc1c(Cl)c(C#N)c(Cl)c(C#N)c1Cl	1.709
350	Nc1c(Cl)c(F)nc(OCC(=O)O)c1Cl	1.708
351	CCC(CN1CCOCC1)[N+](=O)[O-]	1.707
352	CC1(OC(=O)N(C1=O)c2cc(Cl)cc(Cl)c2)C=C	1.707
353	CC1(C)N(Br)C(=O)N(Br)C1=O	1.707
354*	CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C	1.703
355	CN(C)C1=NC(=O)N(C2CCCCC2)C(=O)N1C	1.703
356	CC(C)(C)[C@H](O)C(=Cc1ccc(Cl)cc1)n1cncn1	1.703
357	CNC(=S)[S-]	1.702
358	CCCC\C=C\CCC=CCCCCCCOC(=O)C	1.698

359	OC(=O)CCCOc1ccc(Cl)cc1Cl	1.697
360	FC(F)(F)c1ccncc1C(=O)NCC#N	1.696
361	OCNC(=O)N(CO)C1N(CO)C(=O)N(CO)C1=O	1.695
362	CCOc1nc(ns1)C(Cl)(Cl)Cl	1.695
363	COc1ccc(cc1)C(O)(C2CC2)c3cncnc3	1.693
364	COC[C@H](C)N(C(=O)CCl)c1c(C)csc1C	1.691
365	COCC(C)N(C(=O)CCl)c1c(C)csc1C	1.691
366	CCCCC(CC)CN1C(=O)C2C3CC(C=C3)C2C1=O	1.690
367	CCOC(=O)C1CC(=O)C(=C(O)C2CC2)C(=O)C1	1.686
368	OC(=O)c1c(Cl)ccc2cc(Cl)cnc12	1.685
369	CC1(C)N(Cl)C(=O)N(Br)C1=O	1.684
370	Nc1c(Cl)c(Cl)nc(C(=O)O)c1Cl	1.684
371	CCN(CC)C(=O)C(C)Oc1cccc2ccccc12	1.684
372*	CSc1nc(NC(C)C)nc(NC(C)C)n1	1.684
373*	CC(C)OC(=O)COc1ccc(Cl)cc1Cl	1.683
374	OC(=O)Cc1c(Cl)ccc(Cl)c1Cl	1.680
375	CC(C)N1C(=O)c2ccccc2[N-]S1(=O)=O	1.680
376	CNC(=N[N+](=O)[O-])NCc1cnc(Cl)s1	1.679
377	CNC(NCc1cnc(Cl)s1)=N[N+]([O-])=O	1.679
378	FC1(F)Oc2cccc(c2O1)c3c[nH]cc3C#N	1.679
379	COP(=O)(OC)OC=C(Cl)Cl	1.678
380	OC(Cn1cncn1)(c2ccc(F)cc2)c3ccccc3F	1.676
381	Clc1ccc(C(Cn2ccnc2)OCC=C)c(Cl)c1	1.674
382	CCCCCCCCCCc1ccc(OCCO)cc1	1.673
383*	ClN1C(=O)N(Cl)C(=O)N(Cl)C1=O	1.667
384	Nc1nc(NCl)nc(n1)N(Cl)Cl	1.662
385	ClNc1nc(NCl)nc(NCl)n1	1.662
386	Cc1cc(Cl)ccc1OCCCC(=O)O	1.660
387	Cc1cc(Cl)ccc1OCCCC(=O)[O-]	1.658
388	COC(=O)Nc1nc2ccccc2[nH]1	1.658
389	CCCOP(=S)(OCCC)OP(=S)(OCCC)OCCC	1.658
390	ClC1=C(Cl)C(=O)c2ccccc2C1=O	1.657
391	CSc1nc(NC2CC2)nc(NC(C)(C)C)n1	1.654
392	Cc1cc(O)ccc1Cl	1.652
393	CCN(CC)C(=S)SCC(=C)Cl	1.651
394*	CCCOC(=O)c1ccc(nc1)C(=O)OCCC	1.650
395	C[C@H](c1ccccc1CNCCN)[N+](=O)[O-]	1.650
396	OCC(Br)(CO)[N+](=O)[O-]	1.649
397*	CN(Cc1ccc(Cl)nc1)C(=NC#N)C	1.649
398	CC(=CCC\C(=C\CCC(C)(O)C=C)\C)C	1.648
399	CC(=CCC\C(=C\CC\C(=C\CO)\C)\C)C	1.648
400	[O-][N+](=O)c1cc(c(Cl)c(c1Nc2ncc(cc2Cl)C(F)(F)F)[N+](=O)[O-])C(F)(F)F	1.646
401	Clc1cc(NC(=O)Nc2ccccc2)ccn1	1.644

402	O=C(Nc1ccccc1)Nc2cnns2	1.644
403*	COc1c(Cl)ccc(Cl)c1C(=O)[O-]	1.644
404	Clc1ccc(CN2CCSC2=NC#N)cn1	1.643
405*	Cc1cc(nc(Nc2ccccc2)n1)C3CC3	1.638
406	CCNc1nc(Cl)nc(NC(C)C)n1	1.635
407	NC(=O)C(Br)(Br)C#N	1.634
408	CC1(C)N(Br)C(=O)N(Cl)C1=O	1.633
409*	CC(Oc1ccc(Cl)cc1C)C(O)=O	1.633
410	Nc1c(Cl)c(Cl)nc(C(=O)[O-])c1Cl	1.631
411	CCCCCCCCC=CCCCOC(=O)C	1.631
412*	CCCCCCCC\C=C\CCCOC(C)=O	1.631
413	CC1(C)CON(Cc2ccccc2Cl)C1=O	1.630
414*	O=C(OCc1ccccc1)c2ccccc2	1.628
415	CCCCC=CCCCCCCCCCC=O	1.628
416*	CC1=NNC(=O)N(C1)\N=C\c2cccnc2	1.627
417	CC(C)N(C(=O)CCl)c1ccccc1	1.627
418	CSC(=O)c1cccc2nnsc12	1.624
419	CC1=C(C)S(=O)(=O)CCS1(=O)=O	1.624
420	CC(CN1COC(C)C1)OCOCOCO	1.622
421*	CCC12COCN1COC2	1.621
422	CC1CCCCC1NC(=O)Nc2ccccc2	1.616
423	Cc1cc(C)nc(Nc2ccccc2)n1	1.616
424*	CCNc1nc(Cl)nc(NC(C)(C)C)n1	1.612
425*	OC(=O)CCCc1c[nH]c2ccccc12	1.609
426	[O-][N+](=O)\C(=C\c1ccccc1)\Br	1.608
427*	CCNc1nc(NC(C)C)nc(SC)n1	1.607
428	Nc1nc(nc(C(=O)O)c1Cl)C2CC2	1.606
429	CCNc1nc(Cl)nc(NCC)n1	1.606
430	CNC(=O)Oc1cccc2ccccc12	1.605
431	COc1nc(NC(C)C)nc(NC(C)C)n1	1.603
432	C(Nc1[nH]cnc2ncnc12)c3ccccc3	1.603
433	CCCCCCCCCCCC(=O)[O-]	1.601
434	CN(C)C(=O)Nc1ccc(Cl)cc1	1.599
435*	CCN(CC)CCOCc1ccc(C)cc1	1.597
436	ClN1C(=O)[N-]C(=O)N(Cl)C1=O	1.595
437*	OC(=O)COc1ccc(Cl)cc1Cl	1.595
438	CCCCN(CC)C(=O)SCCC	1.592
439	OCCN1CN(CCO)CN(CCO)C1	1.591
440*	CNC(=O)O\N=C\C(C)(C)S(=O)(=O)C	1.591
441	Oc1ccc(Cl)cc1Cc2ccccc2	1.590
442	[O-]N1C=CC=CC1=S	1.590
443	CCCCN1Sc2ccccc2C1=O	1.589
444	CCSC(=O)N(CC(C)C)CC(C)C	1.588

445	CCCOC1=NN(C(=O)[N-]S(=O)(=O)c2ccccc2C(=O)OC)C(=O)N1C	1.585
446	CCCOC(=O)NCCCN(C)C	1.585
447	CCSC(=O)N(CC)C1CCCCC1	1.584
448	Oc1ccc(cc1)C(=O)CBr	1.583
449	CCOC(=O)Cc1cccc2ccccc12	1.581
450*	CC(C)OC(=O)Nc1cccc(Cl)c1	1.580
451	CC1(C)N(CO)C(=O)N(CO)C1=O	1.576
452	ClC1=C(Cl)C(=O)SS1	1.573
453*	Nc1cc(Cl)nc(C(=O)O)c1Cl	1.571
454	OC(=O)C1C2CCC(O2)C1C(=O)O	1.571
455*	Clc1cc(Cl)cc(c1)C2(CC(Cl)(Cl)Cl)CO2	1.570
456	CCOc1ccc2NC(C)(C)C=C(C)c2c1	1.570
457	NC(=O)Cc1cccc2ccccc12	1.569
458	CCCCCCCC1CCC(=O)O1	1.567
459	[O-]C(=O)C1C2CCC(O2)C1C(=O)[O-]	1.566
460	[O-][N+](=O)\C=C/1\NCCCS1	1.565
461*	C1=CC2C(C(C1O2)C(=O)[O-])C(=O)[O-]	1.561
462	CCC[N+](C)(C)CC[N+](C)(C)CCO	1.561
463	CCCSC(=O)N(CCC)CCC	1.559
464	CP(=O)(O)CCC(N)C(=O)[O-]	1.557
465*	c1ccc2[nH]c(nc2c1)c3cscn3	1.554
466	CC(Oc1cccc(Cl)c1)C(=O)O	1.553
467	CCCCCCCCCCCC(O)=O	1.552
468	CCNC(=O)NC(=O)\C(=N\OC)\C#N	1.547
469	CCCCCCCCN1CCCC1=O	1.546
470*	CC1(C)N(Cl)C(=O)N(Cl)C1=O	1.545
471	CCCN(CCC)c1c(cc(c(N)c1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	1.544
472	CC1(C)C(C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2)C1(C)C	1.543
473	CCN1CN(CC)CN(CC)C1	1.535
474	COC(=O)NC(=S)Nc1ccccc1NC(=S)NC(=O)OC	1.535
475	Cc1ncc([N+](=O)[O-])n1CCO	1.534
476*	CC(O)CSS(=O)(=O)C	1.532
477	CCSC(=O)N1CCCCCC1	1.528
478	OCOCC12COCN1COC2	1.528
479	CCCCOCCOC(=O)C(C)Oc1ccc(Cl)cc1Cl	1.525
480	CC(=O)Nc1cc(NS(=O)(=O)C(F)(F)F)c(C)cc1C	1.523
481	Clc1cccc(Cl)c1C#N	1.520
482	CCCCCCNC(=N)NC(=N)N	1.518
483	CN(C)C(=O)Nc1ccccc1	1.516
484	CCC(C)(CCC(C)C)C(=O)NC	1.514
485	OC(=O)CNCP(=O)(O)O	1.512
486	CC1(C)COCN1	1.493
487	COC(C)(C)CCCC(C)C\C=C\C(=C\C(=O)OC(C)C)\C	1.492

488	CC1CC(OC(=O)C)OC(C)O1	1.491
489	CNC1=C(Cl)C(=O)N(N=C1)c2cccc(c2)C(F)(F)F	1.482
490	CCCCCCCCCC(=O)C	1.482
491	Oc1ccccc1c2ccccc2	1.481
492	OCC(CO)(CO)[N+](=O)[O-]	1.480
493	[O-]c1ccccc1c2ccccc2	1.479
494	Nc1nc(N)nc(NC2CC2)n1	1.471
495*	CCC(C)Nc1c(cc(cc1[N+](=O)[O-])C(C)(C)C)[N+](=O)[O-]	1.470
496*	OCCN(CC[O-])CC[O-]	1.469
497	OC(=O)c1ccccc1C(=O)Nc2cccc3ccccc23	1.464
498*	CCCCCCCCSCCO	1.463
499*	CCCCNC(=O)n1c(NC(=O)OC)nc2ccccc12	1.463
500	OP(=O)(O)CCCl	1.461
501	CN(C)NC(=O)CCC(=O)O	1.455
502	CCc1cccc(C)c1N(C(C)COC)C(=O)CCl	1.453
503	CCCCCCCCC(=O)O	1.450
504	c1ccc2ccccc2c1	1.449
505	CCCCCCCCO	1.448
506	COCC(=O)N(C(C)C(=O)OC)c1c(C)cccc1C	1.446
507	Cc1c(F)c(F)c(COC(=O)C2C(\C=C(/Cl)\C(F)(F)F)C2(C)C)c(F)c1F	1.446
508*	Cc1cc(O)cc(C)c1Cl	1.445
509	CN1SC2=C(CCC2)C1=O	1.441
510	CCOP(=O)([O-])C(=O)N	1.432
511*	CC1=C(C(=O)Nc2ccccc2)S(=O)(=O)CCO1	1.427
512	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	1.417
513	CCOC(=O)NCCOc1ccc(Oc2ccccc2)cc1	1.416
514	CC(=O)Nc1ccc(O)cc1	1.409
515*	CC(C)(NC(=O)c1cc(Cl)cc(Cl)c1)C#C	1.408
516	Oc1ccc(cc1)[N+](=O)[O-]	1.394
517*	OCCNCO	1.385
518	CC(=C)C1CCC(=CC1)C	1.385
519	CC(C)N1C(=O)c2ccccc2NS1(=O)=O	1.381
520*	CC(Cl)(Cl)Cl	1.375
521	CC(Oc1ccc(Cl)cc1Cl)C(=O)O	1.375
522	CCCCOCC(C)O	1.372
523	O=C\C=C\c1ccccc1	1.371
524*	[S-]C(=NC#N)[S-]	1.366
525*	CN(C)C(=S)[S-]	1.364
526	BrCC(=O)OCc1ccccc1	1.360
527*	Cc1cc(C)n(CO)n1	1.351
528	COc1c(Cl)ccc(Cl)c1C(=O)O	1.344
529*	CCCCOC(=O)[C@@H](C)Oc1ccc(Oc2ccc(cn2)C(F)(F)F)cc1	1.343
530*	CS(=O)(=O)NC(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+]([O-])=O	1.341

531	CS(=O)(=O)[N-]C(=O)c1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1[N+](=O)[O-]	1.340
532*	OC(=O)c1ccccc1	1.337
533*	CCCCOC(=O)COc1ccc(Cl)cc1Cl	1.329
534	CC(Oc1cccc(Cl)c1)C(=O)[O-]	1.300
535	O=CCCCC=O	1.293
536	CCCC(C)C1(CC=C)C(=O)NC(=NC1=O)[O-]	1.285
537*	CC1=CC(=O)NO1	1.280
538*	OCNCC(=O)[O-]	1.268
539	Oc1nc(O)nc(O)n1	1.251
540	Nc1nc[nH]n1	1.226
541	CC(C(=O)O)O	1.205
542	CCCCCCCCCCO	1.200
543*	Clc1ccccc1c2nnc(nn2)c3ccccc3Cl	1.184
544*	CCCCCCCCCCCCCNC(N)=N	1.181
545	CCCCC(CC)COC(=O)COc1ccc(Cl)cc1C	1.153
546	CN(C)C(=O)C(c1ccccc1)c2ccccc2	1.124
547	[S-]C#N	1.065
548	C[N+]1(C)CCCCC1	1.058
549	CCC=C	1.050
550	O=C1NNC(=O)C=C1	1.050
551	ClC\C=C\Cl	1.045
552	NC(=O)N	1.029
553	NC#N	0.925
554	CCC(=O)O	0.870
555	Oc1ccc(c(c1)C(F)(F)F)[N+](=O)[O-]	0.714
556	C=CCN=C=S	0.083

***Test set compounds**

Table 3.6. Compounds smile with respective experimental pLC_{50} values for JQ.

Sl.No.	Canonical_smiles	pLC_{50}
1	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.795
2*	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	2.132
3*	COP(=S)(OC)Oc1ccc(cc1)S(=O)(=O)N(C)C	3.680
4*	COP(=S)(OC)Oc1ccc(SC)c(C)c1	3.510
5*	CCOP(=S)(OCC)Oc1ccc(cc1)[N+](=O)[O-]	3.821
6	CCOP(=S)(OCC)Oc1ccc2C(=C(Cl)C(=O)Oc2c1)C	3.207
7*	C1C2C(C(C1Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	3.068
8	C1(C(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl	2.835
9	CNC(=O)CSP(=S)(OC)OC	2.821
10	Nc1nc[nH]n1	1.226
11	COP(=O)(OC)OC=C(Cl)Cl	2.870
12	CNC(=O)Oc1cccc2ccccc12	1.605

13	CC(=O)C	0.162
14	COc1ccc(cc1)C(c2ccc(OC)cc2)C(Cl)(Cl)Cl	1.840
15	ClC(Cl)C(c1ccc(Cl)cc1)c2ccc(Cl)cc2	2.005
16	ClC(=C(c1ccc(Cl)cc1)c2ccc(Cl)cc2)Cl	2.371
17	CCc1ccc(cc1)C(C(Cl)Cl)c2ccc(CC)cc2	1.789
18	CC(Cl)(Cl)C(=O)O	1.456
19	ClC1C=CC2C1C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	3.604
20*	CCOP(=S)(OCC)SC1OCCOC1SP(=S)(OCC)OCC	1.837
21	CC(=C)C1CC2=C(O1)C=CC3=C2OC4COC5=CC(=C(C=C5C4C3=O)OC)OC	2.321
22	OC(=O)Cc1c(Cl)ccc(Cl)c1Cl	1.680
23	COP(=S)(OC)SCN1N=Nc2ccccc2C1=O	2.696
24*	Oc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	1.709
25*	CCC(C)c1cc(cc(c1O)[N+](=O)[O-])[N+](=O)[O-]	2.769
26	CC(Oc1cc(Cl)c(Cl)cc1Cl)C(=O)O	1.732
27	c1cc(c(cc1Cl)Cl)OCC(=O)O	1.646
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*	CC1OC(C)OC(C)OC(C)O1	1.707
32	CNC(=O)Oc1ccccc1OC(C)C	1.622
33	ClC1=C(Cl)C2(Cl)C3COS(=O)OCC3C1(Cl)C2(Cl)Cl	2.513
34	OC(c1ccc(Cl)cc1)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	2.612
35*	CCOP(=S)(OCC)Oc1ccc(cc1)S(=O)C	3.570
36*	CNC(=O)O\N=C\C(C)(C)SC	2.887
37	Clc1ccc(cc1)S(=O)(=O)c2cc(Cl)c(Cl)cc2Cl	1.853
38	ClC1=C(Cl)C(=O)c2ccccc2C1=O	1.657
39	CC(Oc1ccc(Cl)cc1Cl)C(=O)O	1.584
40*	CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC	2.191
41*	COP(=S)(OC)Oc1ccc(c(C)c1)[N+](=O)[O-]	2.799
42*	CCNc1nc(Cl)nc(NCC)n1	1.734
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	1.826
48	COP(=O)(OC)OC(=CC(=O)N(C)C)C	3.870
49*	[S-]C(=S)NCCNC(=S)[S-]	1.624
50*	CN(C)C(=O)Nc1ccc(Cl)cc1	1.599
51	CCOP(=S)(OCC)Oc1cnccn1	3.948
52	COP(=S)(OC)Oc1ccc(cc1)[N+](=O)[O-]	3.758
53	CCOP(=S)(OCC)SCSCC	3.115
54*	CCOP(=S)(OCC)SCCSCC	2.916
55*	COP(=O)(OC)OC(Br)C(Cl)(Cl)Br	2.458

56*	CCS(=O)CCSP(=O)(OC)OC	2.275
57	C1C2C=CC1C3C2C4(C(=C(C3(C4(Cl)Cl)Cl)Cl)Cl)Cl	4.031
58	CNC(=O)Oc1cc(C)c(N(C)C)c(C)c1	2.648
59	CN(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.669
60*	CON(C)C(=O)Nc1ccc(Cl)c(Cl)c1	1.697
61	CCOP(=S)(OCC)Oc1cc(C)nc(n1)C(C)C	3.811
62	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	1.886
63	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	2.416
64*	CCOP(=S)(CC)Sc1ccccc1	2.922
65	COP(=S)(OC)SCSc1ccc(Cl)cc1	1.998
66	C[N+](C)(C)CCCl	1.390
67	CCCS(=O)(=O)\C=C\S(=O)(=O)CCC	1.682
68	Clc1cccc(Cl)c1C#N	1.537
69	CNC(=O)Oc1cccc2CC(C)(C)Oc12	2.937
70*	CCNc1nc(Cl)nc(NC(C)C)n1	1.635
71	Nc1c(Cl)c(Cl)nc(C(=O)O)c1Cl	1.684
72*	CCCCOCCOC(=O)COc1ccc(Cl)cc1Cl	1.808
73	Clc1cccc(n1)C(Cl)(Cl)Cl	2.450
74	NC(=O)COC1CCC(Cl)CC1Cl	1.644
75	CNC(=O)Oc1cc(C)c(SC)c(C)c1	2.274
76	CCOP(=S)(Oc1ccc(cc1)[N+](=O)[O-])c2ccccc2	2.863
77	CCCC(C)c1cccc(OC(=O)NC)c1	1.646
78	ClC1(Cl)C2(Cl)C3(Cl)C4(Cl)C(Cl)(Cl)C5(Cl)C(Cl)(C1(Cl)C35Cl)C24Cl	2.038
79	CCCCOCCOC(=O)COc1cc(Cl)c(Cl)cc1Cl	1.852
80	CNC(=O)Oc1cc(C)c(C)c(C)c1	1.984
81	CCOP(=S)(OCC)Oc1nc(Cl)c(Cl)cc1Cl	3.069
82	CCCOP(=S)(OCCC)OP(=S)(OCCC)OCCC	1.879
83	COP(=S)(OC)Oc1ccc(Sc2ccc(OP(=S)(OC)OC)cc2)cc1	3.254
84	Clc1ccc(cc1)C(C(=O)C2C(=O)c3ccccc3C2=O)c4ccccc4	3.796
85	COP(=S)(OC)Oc1nc(Cl)c(Cl)cc1Cl	1.810
86*	CN(C)C=Nc1ccc(Cl)cc1C	2.051
87	CNC(=O)C=C(C)OP(=O)(OC)OC	4.968
88	Cc1ccc(N)cc1Cl	3.789
89	COC(=O)C=C(C)OP(=O)(OC)OC	2.794
90*	ClCC1(CCl)C(=C)C2(Cl)C(Cl)C(Cl)C1(Cl)C2(Cl)Cl	2.778
91	CC1C(OC(=O)C2C(C=C(C)C)C2(C)C)C=C(CC=CC=C)C1=O	1.818
92*	CCOP(=O)(OCC)SCCSCC	2.973
93	COP(=O)(N)SC	3.186
94	CC(=CC1C(C(=O)OCc2coc(Cc3ccccc3)c2)C1(C)C)C	1.831
95	Clc1ccc(c(Cl)c1Cl)c2ccc(Cl)c(Cl)c2Cl	2.218
96*	Clc1ccc(c(Cl)c1Cl)c2cccc(Cl)c2Cl	2.052
97*	CC1=CC(=C(C(=C1)OC(=O)NC)C)C	1.984
98*	Clc1cc(Cl)cc(c1)c2cc(Cl)cc(Cl)c2	1.780

99	CCN(CC)C(=O)\C(=C(/C)\OP(=O)(OC)OC)\Cl	3.527
100	CCCSP(=O)(OCC)SCCC	3.694
101	CC(=NOC(=O)NC)SC	1.715
102	CCOP(=O)(Sc1ccccc1)Sc2ccccc2	2.088
103*	CCCCOCCOC(=O)C(C)Oc1cc(Cl)c(Cl)cc1Cl	1.869
104	COP(=S)(Oc1cc(Cl)c(Br)cc1Cl)c2ccccc2	2.439
105*	CCOP(=O)(NC(C)C)Oc1ccc(SC)c(C)c1	3.711
106	COP(=O)(OC)OC(=CCl)c1cc(Cl)c(Cl)cc1Cl	1.864
107	CCOP(=S)(OCC)SCCl	1.672
108	CCC(C)c1cccc(OC(=O)N(C)Sc2ccccc2)c1	2.370
109	CN(\C=N\c1ccc(C)cc1C)\C=N\c2ccc(C)cc2C	2.212
110	COC(=O)C(C)Oc1ccc(Oc2ccc(Cl)cc2Cl)cc1	1.232
111	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc2cccc(Oc3ccccc3)c2	1.231
112	Clc1ccc(c(Cl)c1)c2cccc(Cl)c2Cl	1.687
113	COCC(=O)N(C(C)C(=O)OC)c1c(C)cccc1C	1.446
114	CP(=O)(O)CCC(N)C(=O)[O-]	1.557
115	COC(=O)c1ccc(I)cc1S(=O)(=O)[N-]C(=O)Nc2nc(C)nc(OC)n2	2.002
116	COc1ncc(F)c2nc(nn12)S(=O)(=O)Nc3c(F)cccc3F	1.857
117	CN\C(=N\[N+](=O)[O-])\NCC1CCOC1	1.607

***Test set compounds**

3.3.2. 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/). 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.

3.3.3. Dataset division

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. 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. We also employed modified k-medoid clustering by using Modified k-Medoid 1.3 (http://teqip.jdvu.ac.in/QSAR_Tools/). After that, the final selection of data-set division methods was done based on the statistical results. In this process of dataset division, the datasets are divided into 75:25 ratios of training and test sets compounds respectively.

3.3.4. 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. Primarily, we performed stepwise regression method and genetic algorithm (GA) for feature selection, and then we employed the regression-based partial least square (PLS) method through the Partial least squares v1.0 tool (http://teqip.jdvu.ac.in/QSAR_Tools/) for model building.

3.3.5. Validation metrics of QSTR models

A significant step in the creation of a QSTR model is statistical validation, which demonstrates its reliability and predictivity. Various internal validation parameters were calculated which involve 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} [75]. For both internal ($Q^2{}_{LOO}$) and external predictive parameters (Q^2_{F1}, Q^2_{F2}), the approved threshold value is 0.5.

3.3.6. 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. 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.

3.3.7. 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. 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.

3.3.8 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 [76]. 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 [77]. The values of the R^2y_{rand} intercept and Q^2y_{rand} intercept should not be more than 0.3 and 0.05 respectively.

3.3.9 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 [78]. 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.

3.3.10 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) [79-80].

3.3.11 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. A detailed flow diagram of this study has been given in **Figure 3.3.**

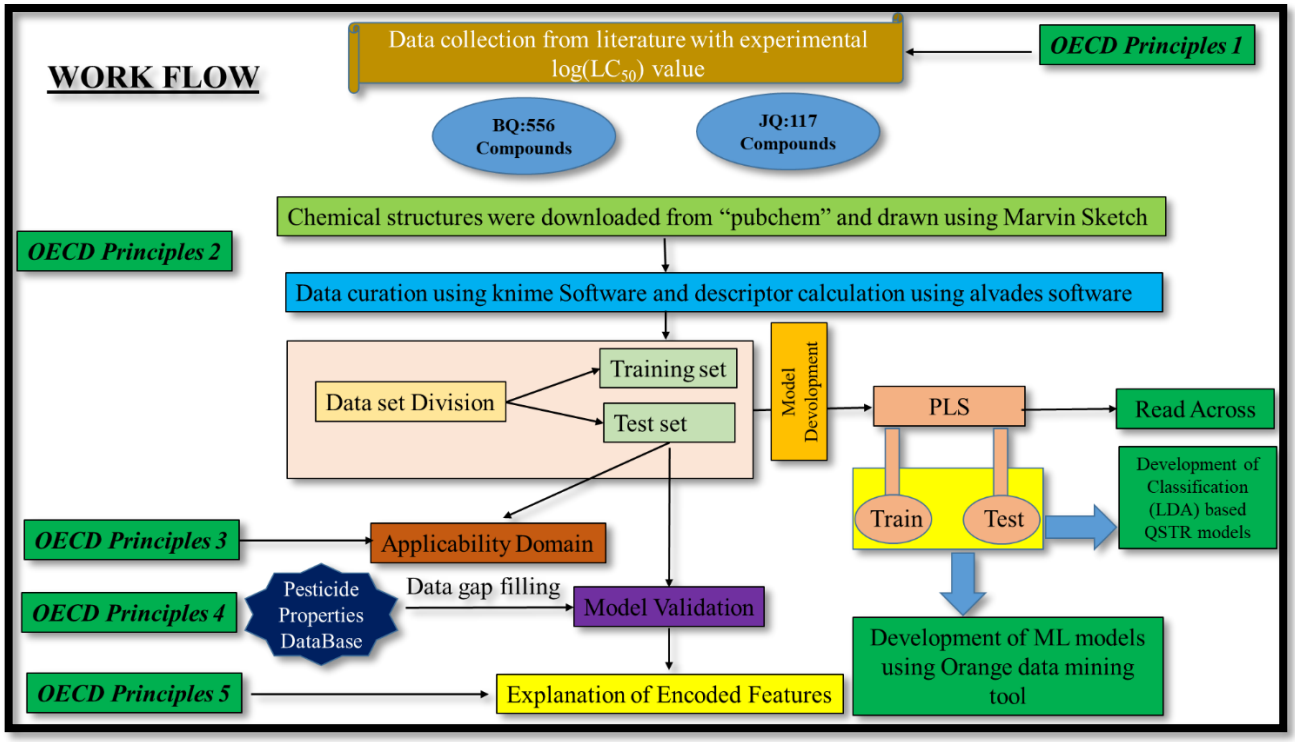


Figure 3.3. Workflow of QSTR model development.

CHAPTER - 4

Results and Discussion

4.1. Study 1

In this present study, we have developed QSTR and q-RASTR models for pLOEL and pNOEL endpoints using the PLS method and strictly obeying the OECD guidelines. We have additionally applied two different ML algorithms (SVM, RR) to check model performances.

4.1.1. PLS-based QSTR and q-RASTR models

The divided dataset is used to develop the QSTR and q-RASTR models for two endpoints (pLOEL and pNOEL) of chicken species. After the feature selection process, the PLS-based QSTR model was developed employing 3 and 5 descriptors with two and one latent variables for pLOEL (MODEL 1) and pNOEL (MODEL 2), respectively.

PLS-based QSTR model for pLOEL and pNOEL endpoints:

Model 1 (pLOEL endpoint):

$$\boldsymbol{pLOEL} = 4.75827 + 0.50323 \times \mathrm{NsOH} - 0.191 \times MaxsCH3 - 0.64324 \times B01[C - O]$$

Model 2 (pNOEL endpoint):

$$\boldsymbol{pNOEL} = 5.08369 + 0.16353 \times H - 050 + 0.35253 \times NsssN - 0.62789 \times B05[C - O] + 0.80035 \times B05[O - O] - 0.8449 \times B08[C - P]$$

After the development of the QSTR models, similarity and error-based RASTR descriptors were calculated for both training and test sets compounds of pLOEL and pNOEL endpoints models using "RASAR Descriptor Calculator v2.0 tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) with the optimized hyperparameters. After that, we clubbed the RASTR descriptors and Alvadesc descriptors for the final q-RASTR model development [81]. Finally, PLS-based q-RASTR models were developed using 3 and 4 descriptors with one and two latent variables as shown in model 3 and model 4 respectively for pLOEL and pNOEL endpoint models,

PLS-based q-RASTR model for pLOEL and pNOEL endpoints:

Model 3 (pLOEL endpoint):

$$\boldsymbol{pLOEL} = 5.1136 - 1.51275 \times SD\ similarity(GK) + 0.41951 \times NsOH - 0.75444 \times B01[C - O]$$

Model 4 (pNOEL endpoint):

$$\boldsymbol{pNOEL} = 5.78412 - 2.04509 \times SE(LK) + 1.18371 \times B05[O - O] - 0.74259 \times B02[C - O] + 0.03736 \times T(N..S)$$

Each model has been rigorously validated following the OECD protocols. The computed internal and external validation metrics along with the optimum number of latent variables have been shown in the following **Table 4.1.** The PLS-based q-RASTR models **3** and **4** show strong fit and predictability with uniform scattering observed along the line, going through the origin of Cartesian coordinates **(Figure. 4.1).**

Table 4.1. QSTR and q-RASTR model's statistical quality.

Validation Metrics	QSTR model's statistical quality		PLS q-RASTR model's statistical quality	
Model name	Model 1 (pLOEL)	Model 2 (pNOEL)	Model 3 (pLOEL)	Model 4 (pNOEL)
No of LVs	2	1	1	2
R^2(train)	0.748	0.669	0.734	0.603
Q^2_{LOO} (train)	0.672	0.582	0.665	0.526
$Q^2_{F1\ (test)}$	0.608	0.643	**0.844**	**0.762**
$Q^2_{F2\ (test)}$	0.577	0.640	**0.831**	**0.759**
$Q^2_{F3\ (test)}$	0.692	0.790	**0.877**	**0.860**
MAE_{test}	0.309	0.225	**0.214**	**0.195**
CCC	0.818	0.730	0.909	0.845
$\overline{r^2_{m(test)}}$	0.637	0.415	0.740	0.560
Δr^2_m(test)	0.035	0.318	0.136	0.220
MAE-based prediction quality	MODERATE	GOOD	GOOD	GOOD

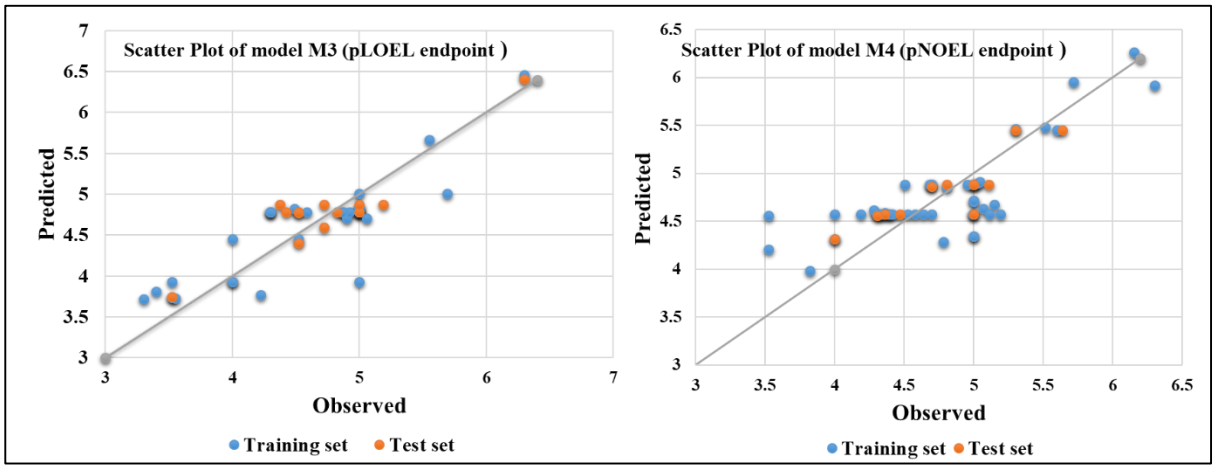


Figure 4.1. Scatter plots of developed models.

Here, we have seen that for both datasets, the external validation metrics were significantly improved for the PLS-based q-RASTR models as compared to the PLS-based QSTR models, indicating the significance of the RASTR descriptors. We have also validated all the models (PLS-based QSTR and q-RASTR models for the pLOEL and pNOEL endpoints) using

Golbraikh and Tropsha criteria and the results are given in **Tables 4.2-4.5.** The results showed that the PLS-based q-RASTR models for both endpoints are acceptable based on the Golbraikh and Tropsha's criteria [82]. Hence, we have generalized that the PLS-based q-RASTR models are better as compared to the corresponding QSTR models.

Table 4.2. Results of the final PLS-based q-RASTR (pLOEL) model obtained according to Golbraikh and Tropsha's criteria.

Sl.No	Parameters	PLS q-RASTR (pLOEL)	Remarks	Threshold value
1	Q^2_{LOO} (train)	0.665	Pass	Q^2_{LOO}>0.5
2	R^2(test)	0.844	Pass	R^2(test)>0.6
3	$[(r^2-r_0^2)/ r^2]$	0.001	Pass	<0.1
4	$[(r^2-r'^2_0)/ r^2]$	0.038	Pass	<0.1
5	k	0.986	Pass	0.85 < k <1.15
6	k'	1.011	Pass	0.85 < k'<1.15

Table 4.3. Results of the final PLS-based q-RASTR (pNOEL) model obtained according to Golbraikh and Tropsha's criteria.

Sl.No	Parameters	PLS q-RASTR (pNOEL)	Remarks	Threshold value
1	Q^2_{LOO} (train)	0.526	Pass	Q^2_{LOO}>0.5
2	R^2(test)	0.779	Pass	R^2(test)>0.6
3	$[(r^2-r_0^2)/ r^2]$	0.024	Pass	<0.1
4	$[(r^2-r'^2_0)/ r^2]$	0.269	Fail	<0.1
5	k	0.997	Pass	0.85 < k <1.15
6	k'	1.036	Pass	0.85 < k'<1.15

Table 4.4. Results of the final PLS-based QSTR (pLOEL) model obtained according to Golbraikh and Tropsha's criteria.

Sl.No	Parameters	PLS QSTR (pLOEL)	Remarks	Threshold value
1	Q^2_{LOO} (train)	0.672	Pass	Q^2_{LOO}>0.5
2	R^2(test)	0.733	Pass	R^2(test)>0.6
3	$[(r^2-r_0^2)/ r^2]$	0.060	Pass	<0.1
4	$[(r^2-r'^2_0)/ r^2]$	0.007	Pass	<0.1
5	k	0.960	Pass	0.85 < k <1.15
6	k'	1.036	Pass	0.85 < k'<1.15

Table 4.5. Results of the final PLS-based QSTR (pNOEL) model obtained according to Golbraikh and Tropsha's criteria.

Sl.No	Parameters	PLS QSTR (pNOEL)	Remarks	Threshold value
1	Q^2_{LOO} (train)	0.582	Pass	Q^2_{LOO}>0.5
2	R^2(test)	0.765	Pass	R^2(test)>0.6
3	$[(r^2-r_0^2)/ r^2]$	0.123	Fail	<0.1

4	$[(r^2-r'^2_0)/ r^2]$	1.023	Fail	<0.1
5	k	1.016	Pass	0.85 < k <1.15
6	k'	0.981	Pass	0.85 < k'<1.15

4.1.2. Results of ML-based q-RASTR model

As previously stated, we used two different ML algorithms to evaluate their effectiveness in model construction and prediction. Based on the internal validation, v-SVM was the best-performing model toward the pLOEL endpoint, and Ridge regression was the best-performing model toward the pNOEL endpoint based on internal and external validation metrics. In terms of external validation metric, Q^2_{F3} [28], the ability to efficiently predict the response values for the target (query) compounds, the best-performing models were the PLS-based q-RASTR models. Furthermore, the PLS-based q-RASTR models produce the lowest prediction error for the query set compounds, as indicated by the MAE_{test} value [29]. Thus, to assess the overall performance of the models for both endpoints, the PLS-based q-RASTR models are superior than QSTR models. The results of ML models are presented in **Table 4.6**.

Table 4.6. ML-based q-RASTR model's statistical quality.

Validation Metrics	ML model's statistical quality			
Model name	SVM (pLOEL)	SVM (pNOEL)	RR (pLOEL)	RR (pNOEL)
$R^2_{Loo\ (train)}$	0.831	0.695	0.776	0.758
$Q^2_{LOO\ (train)}$	0.746	0.585	0.746	0.604
RMSEc (train)	0.245	0.245	0.283	0.218
$Q^2_{F1\ (test)}$	0.742	0.718	0.725	0.653
$Q^2_{F2\ (test)}$	0.721	0.715	0.703	0.650
$Q^2_{F3\ (test)}$	0.797	0.835	0.784	0.796
MAE_{test} (test)	0.273	0.169	0.300	0.216
CCC	0.893	0.856	0.850	0.804
$\overline{r^2_{m(test)}}$	0.725	0.659	0.626	0.541
$\Delta r^2_m(test)$	0.101	0.071	0.033	0.148
Optimum hyperparameters	v-SVM Regression cost-0.50 Complexity bound-0.65 Kernel-Linear	v-SVM Regression cost-2.50 Complexity bound-0.70 Kernel-Linear	Alpha-0.001	Alpha-0.001

4.1.3. Regression coefficient plot

The plot describes the descriptor's positive/negative contribution towards the toxicity [30]. In this study, the descriptor NsOH contributed positively while the descriptors SD similarity (GK) and B01[C-O] contributed negatively toward the toxicity in case of **Model 3**. In case of **Model 4,** the descriptors B05[O-O], T(N..S) contributed positively while the descriptors SE(LK), B02[C-O] contributed negatively towards the toxicity. All the relevant plots have been provided in **Figures 4.2.-4.3.**

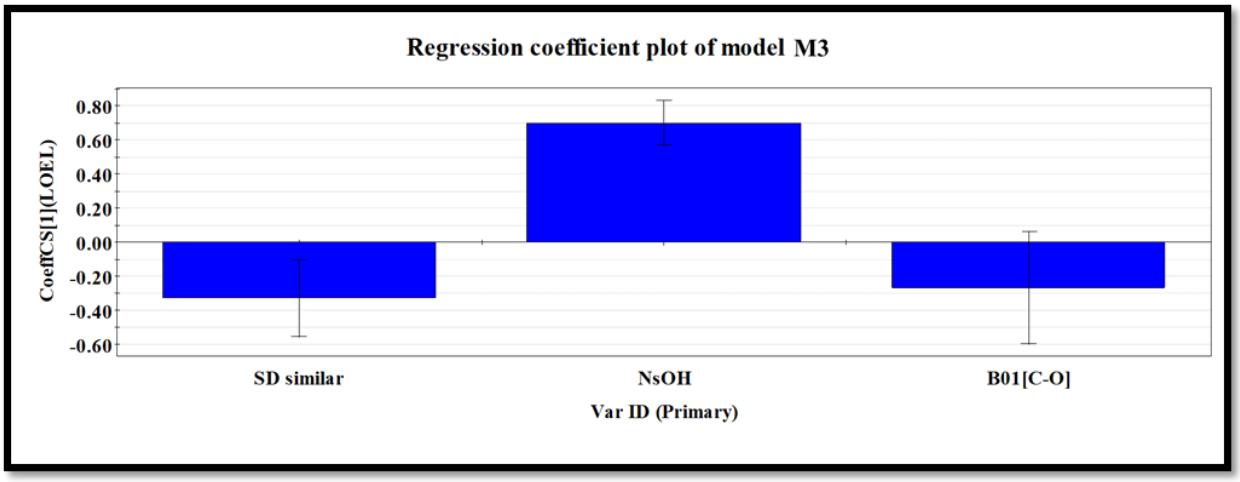


Figure 4.2. Regression coefficient plot of model M3.

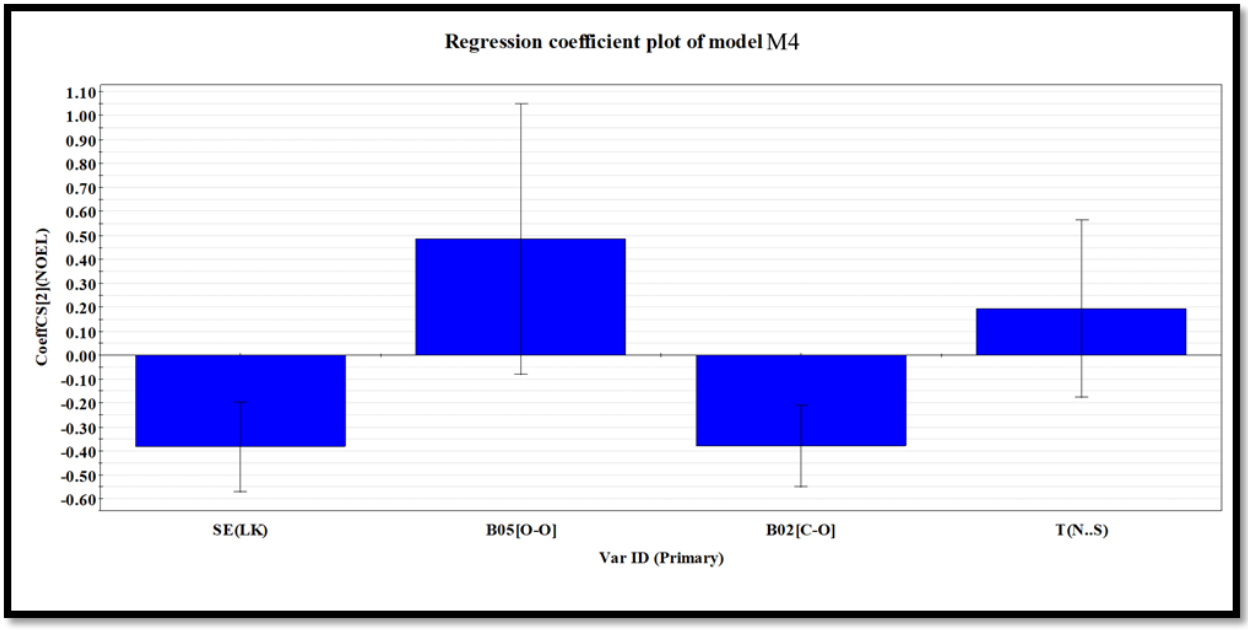


Figure 4.3. Regression coefficient plot of model M4.

4.1.4. Variable importance plot (VIP)

The respective descriptor contribution towards the model response is described by the variable importance plot, and the most and least important descriptors are recognized appropriately [31]. In this present study, NsOH and B02[C-O] depicting electronegativity and hydrophilicity were

identified as the most important descriptors for Model 3 and Model 4 respectively as shown in **Figures 4.4.-4.5.**

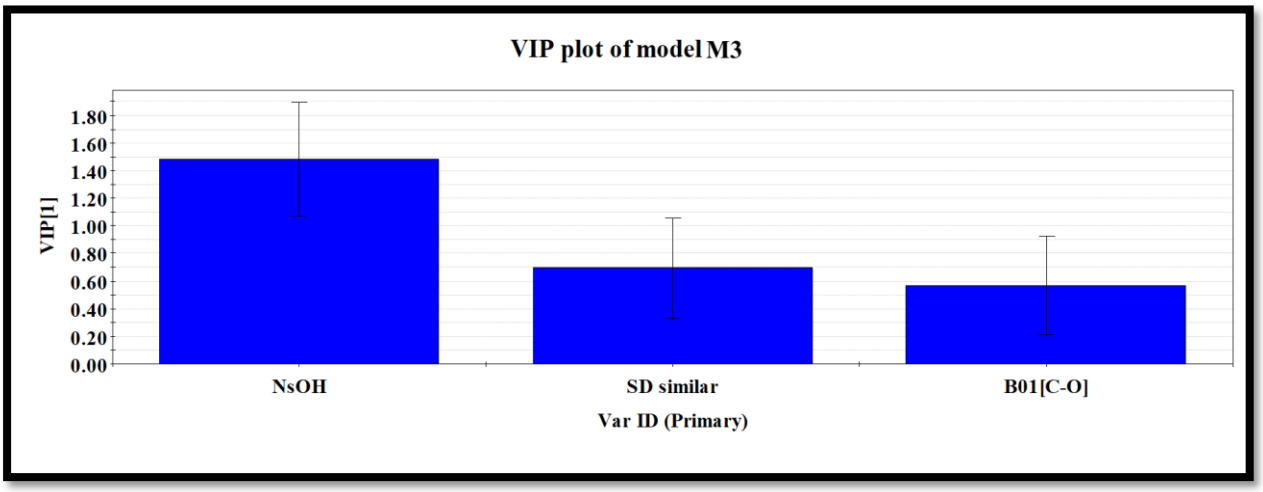


Figure 4.4. The variable_importance plot of model M3 (pLOEL).

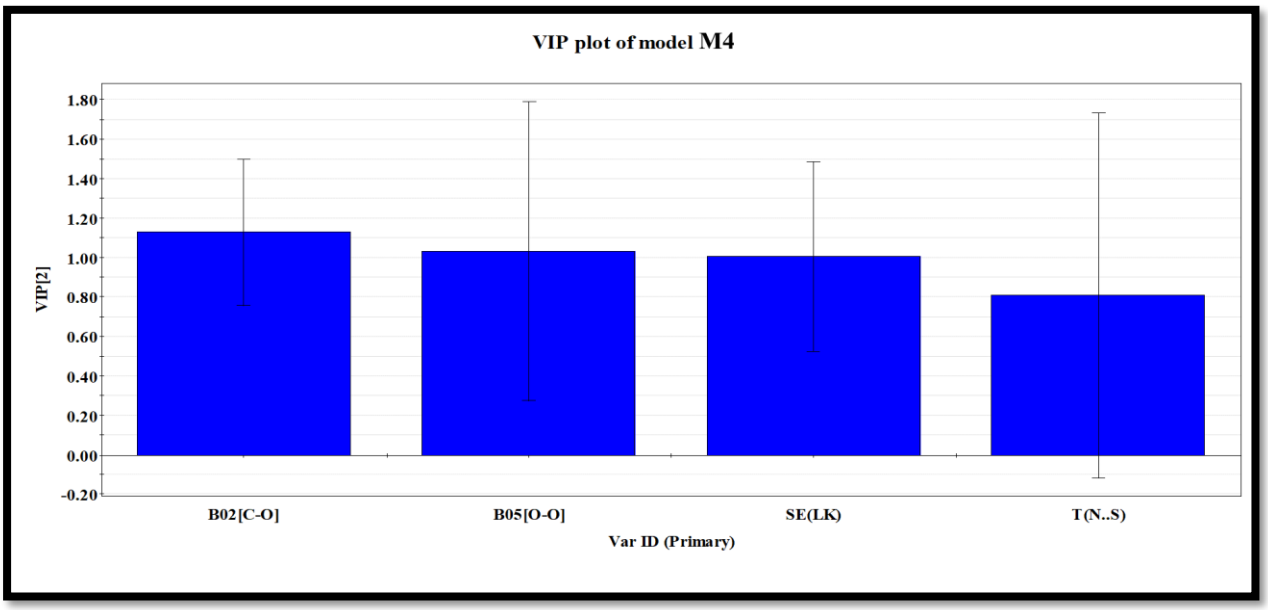


Figure 4.5. The variable_importance plot of model M4 (pNOEL).

4.1.5. Loading plot

The plot describes the correlation between the X and Y variables, illustrating the effect of various model descriptors. The first two components were used to create the loading plot. A descriptor is supposed to have a stronger effect on response value if it is situated far from the origin of the plot and near the modeled endpoint. All the relevant plots have been provided in **Figures 4.6-4.7.**

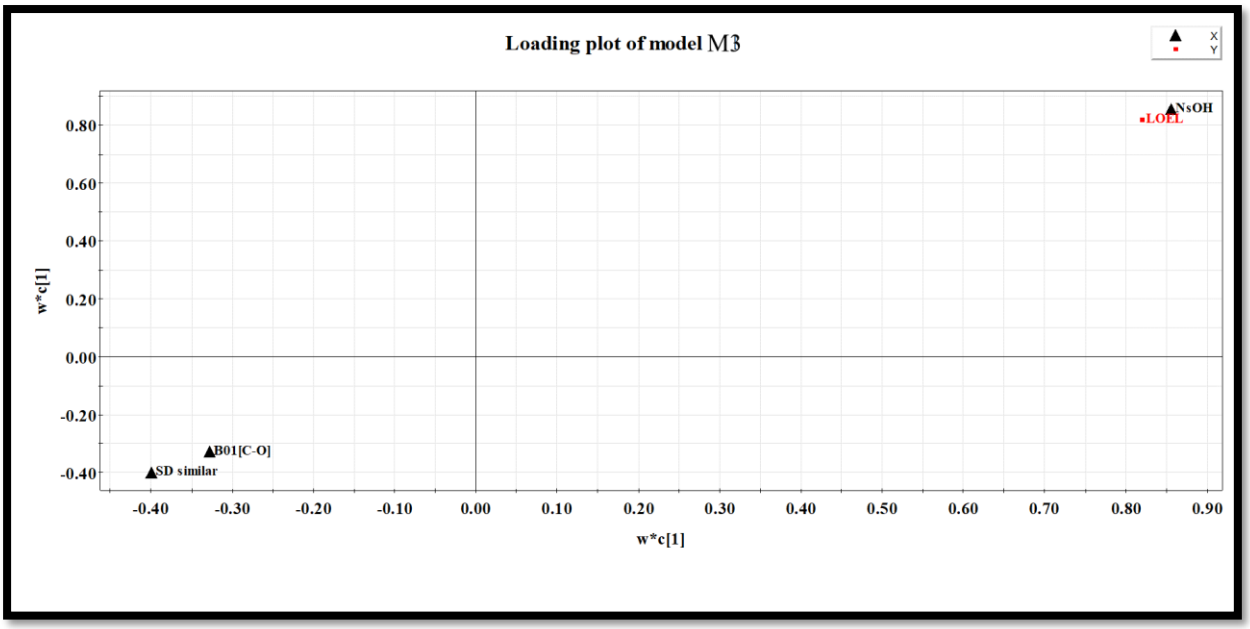


Figure 4.6. The loading plot of the model M3 (pLOEL).

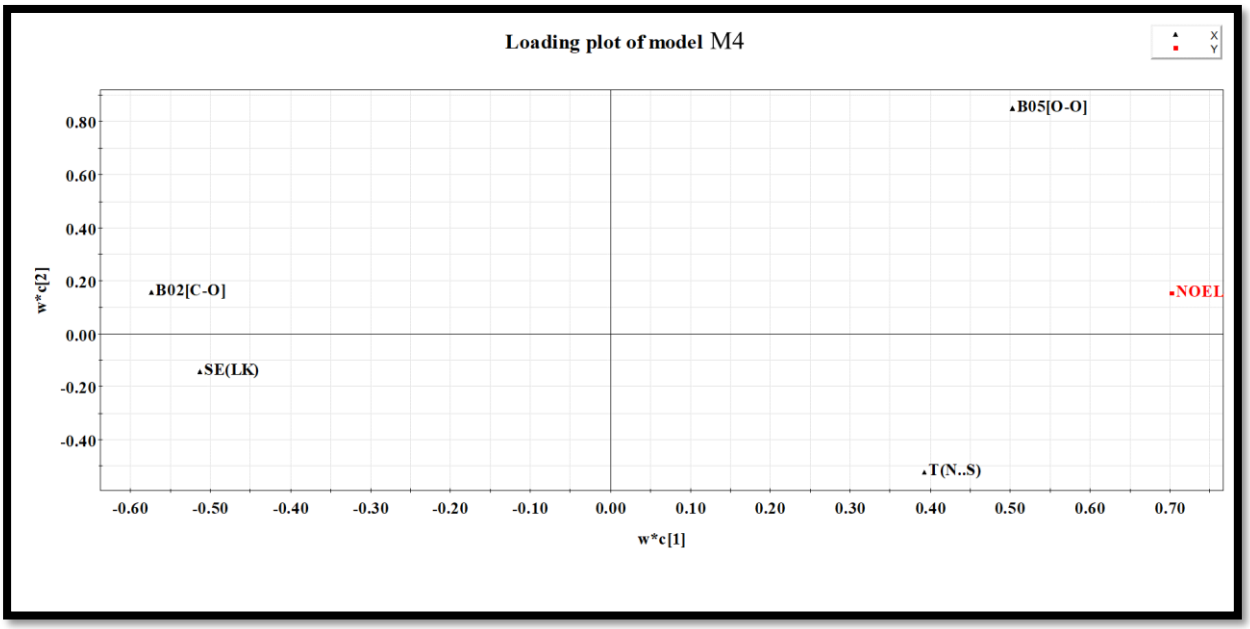


Figure 4.7. The loading plot of the model M4 (pNOEL).

4.1.6. Applicability domain (AD)

AD is the hypothetical region in chemical space specified by the respective model descriptors and responses where predictions may be made with confidence. To obtain a reliable prediction, the target compounds must have the highest structural similarity to the training compounds. As a result, validating the applicability domain is a fundamental prerequisite for every statistical model, as recommended by OECD principle 3 ("Validation of (Q)SAR Models - OECD,"

2004). To comply with the OECD guidelines, an applicability domain analysis of the created PLS-based q-RASTR model was done with SIMCA-P software using the DModX technique at a 99% confidence level.

$$\text{DModX} = \frac{\sqrt{\frac{SSEi}{K-A}}}{\sqrt{\frac{SSE}{(N-A-AO)(K-A)}}}$$

For observation i, in a model with A component, K variables, and N observations, SSE is the squared sum of the residuals. A0 is 1 if the model was centered and 0 otherwise. It is claimed that DModX is approximately F-distributed, so it can be used to check if an observation deviates significantly from a normal PLS model. The DModX (distance to model in X-space) plots for both the training and test sets have been showcased in **Figures 4.8-4.11.** (shows the AD plots of the Model 3 and Model 4). In this study, all the compounds from the training set (given in **Figure. 4.8.**) and test set (given in **Figure. 4.9.**) for the pLOEL endpoint model (model M3) are inside the applicability domain (below the D-Critical line) which indicates the reliability of predictions by the model. In the case of the pNOEL endpoint model (model M4), compounds 28 and 33 of the training set **(**given in **Figure. 4.10.)** are outside the applicability domain (above the D-critical line) due to the structural dissimilarity. All the compounds from the test set **(**given in **Figure. 4.11.)** of the pNOEL endpoint (model M4) are within the applicability domain.

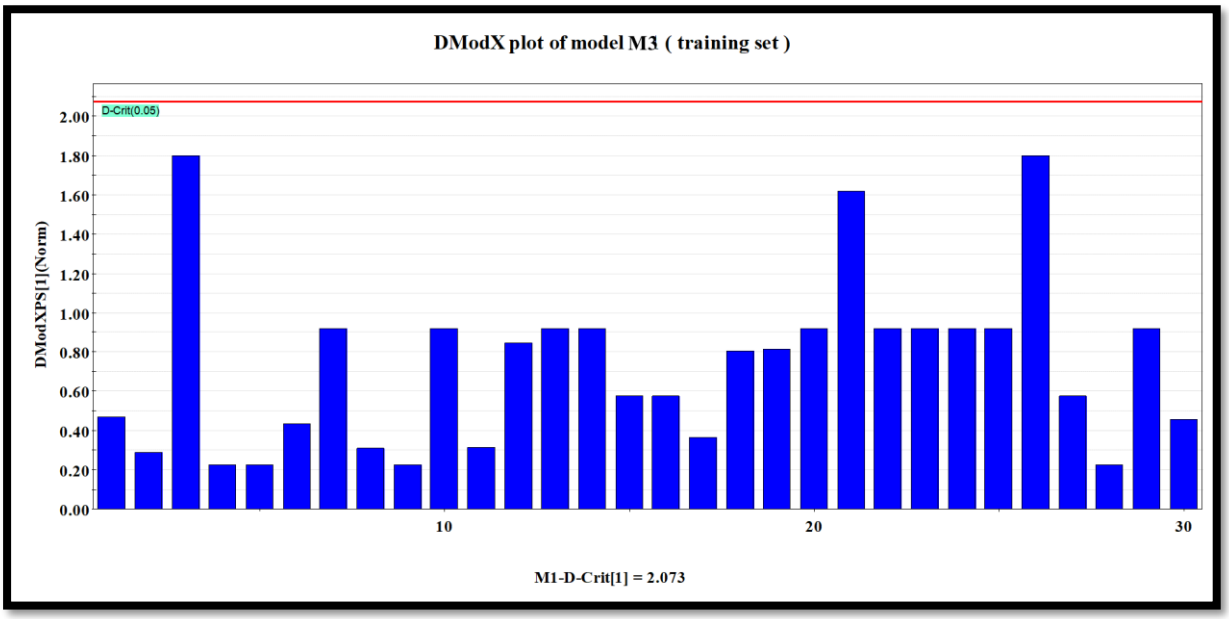


Figure 4.8. DModx plot (pLOEL) of the model M3 (training set).

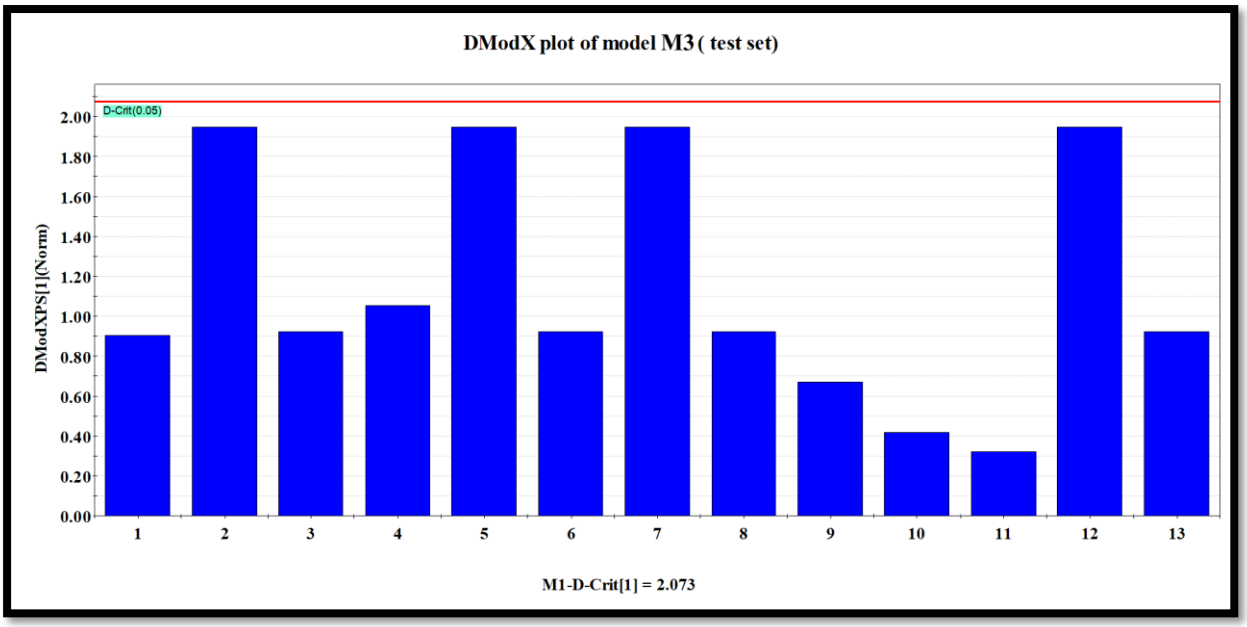


Figure 4.9. DModx plot (pLOEL) of the model M3 (test set).

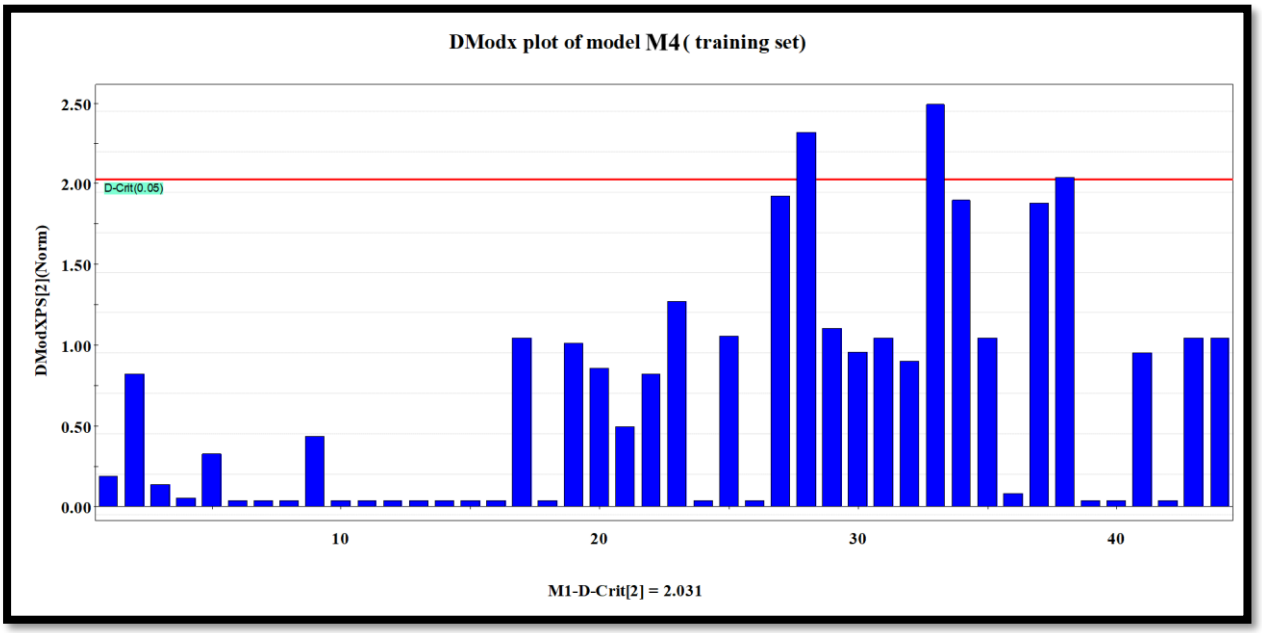


Figure 4.10. DModx plot (pNOEL) of the model M4 (training set).

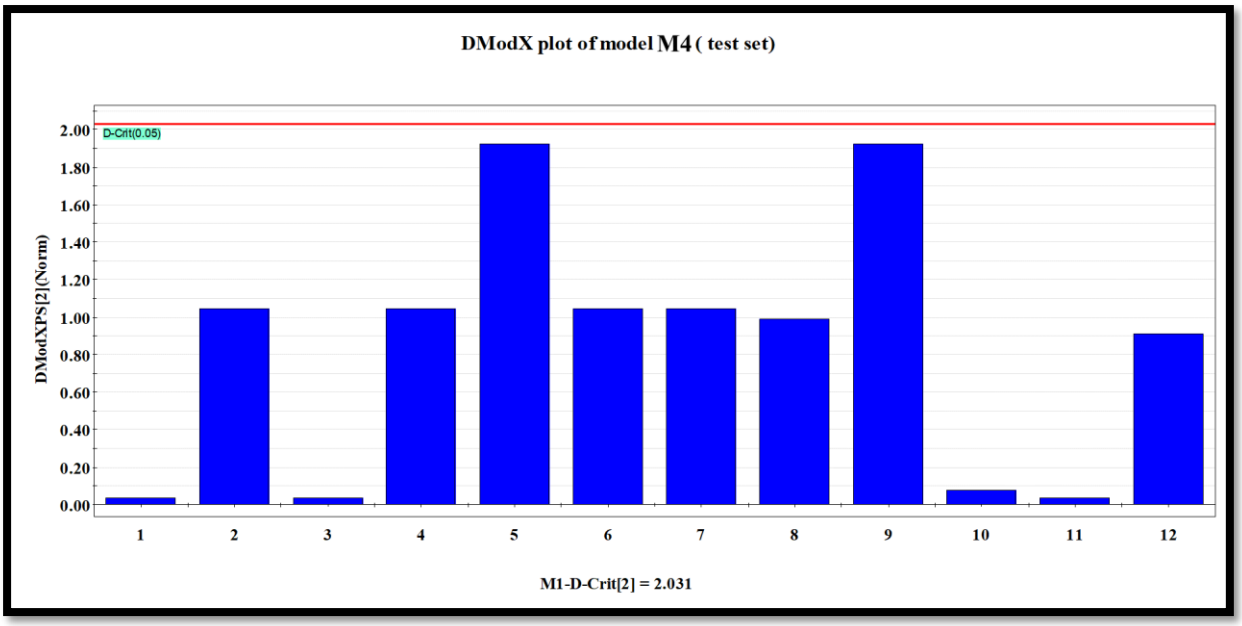


Figure 4.11. DModx plot (pNOEL) of the model M4 (test set).

4.1.7. Mechanistic interpretation

The details of the descriptors obtained from the M3 (pLOEL endpoint) and M4 models (pNOEL endpoint), their contribution, description, and probable mechanistic interpretation (according to OECD principle 5) are provided in **Table 4.7**.

4.1.7.1. Mechanistic interpretation of descriptors employed in Model M3 (pLOEL)

SD similarity (GK) is a RASTR descriptor that denotes the typical deviation of similarity levels among closely related compounds. It has a negative contribution to the toxicity endpoint. Higher standard deviation (SD) similarity shows that the distribution among the close source compounds is high thereby reducing prediction reliability as demonstrated in compound **30** and conversely shown in compound **3 (**depicted in **Figure. 4.12).**

The descriptor NsOH defines the number of atoms of type sOH in the compound and it contributes positively towards the toxicity endpoint. This fragment enhances the compound toxicity due to the presence of an electronegative atom (Oxygen) as demonstrated in compound **42** and the absence of this fragment decreases the toxicity as represented in compound **18 (**shown in **Figure. 4.12).**

The descriptor B01[C-O] is a 2D atom pair descriptor that shows the occurrence of C-O at topological distance 1 and gives negative contribution towards the endpoint. The presence of polar bond [C-O] increases the hydrophilicity of the compound [34] and thus toxicity will

decrease which is evidenced by compound **27** and vice versa in case of compound **36** (represented in **Figure. 4.12**).

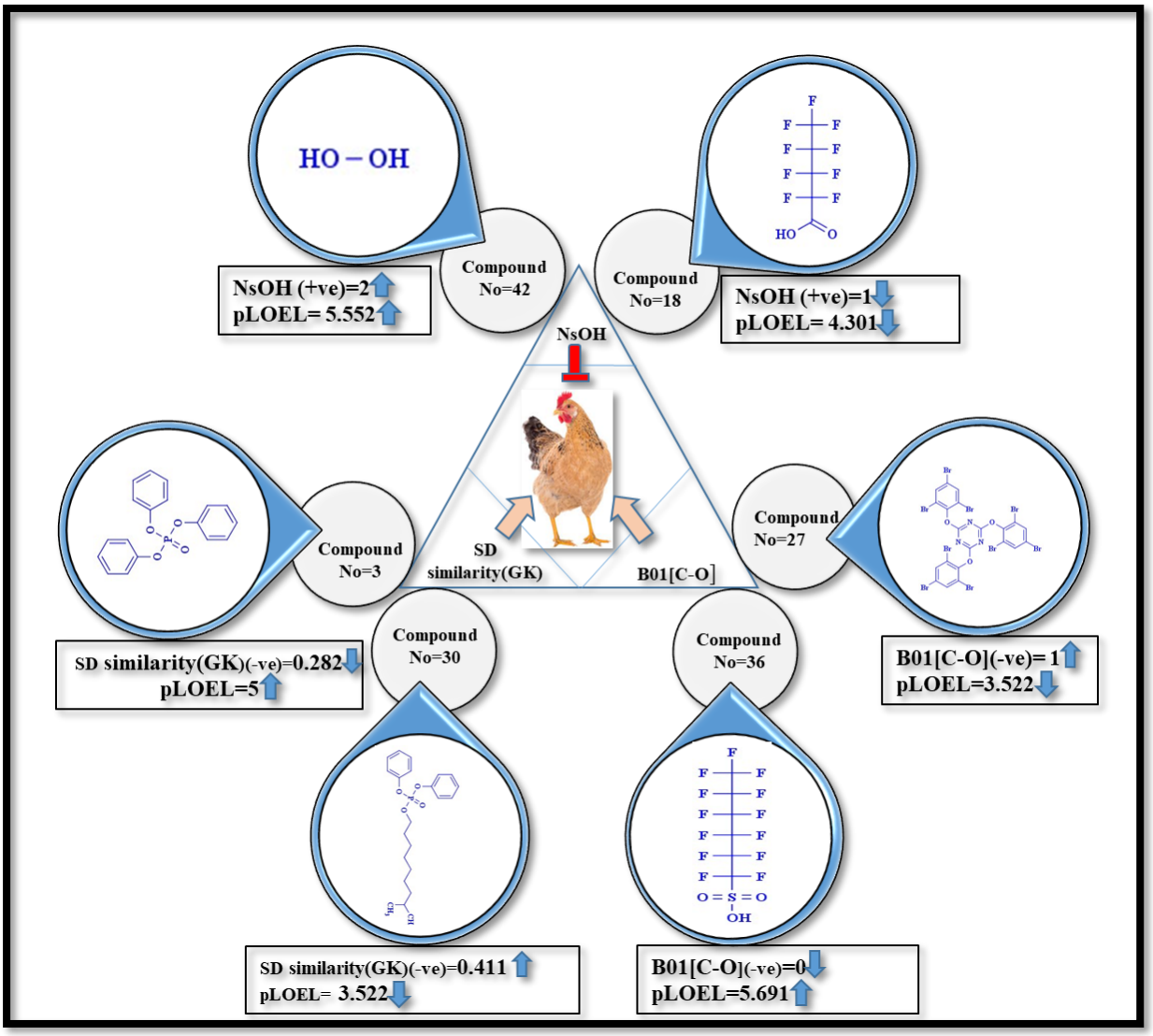


Fig 4.12. Contribution of the model descriptors towards pLOEL in chicken

4.1.7.2. Mechanistic interpretation of descriptors employed in Model M4 (pNOEL)

The SE(LK) is a RASTR descriptor that denotes the weighted standard error about the response values of adjacent source compounds. It shows negative contribution toward the endpoint. The presence of this high standard error based on the response values of the proximate source compound decreases the compound toxicity as demonstrated in compound **8** and the less standard error based on response enhances the toxicity as represented in compound **40** (given in **Figure. 4.13**).

The 2D atom pair descriptor, B05[O-O] shows the occurrence of two oxygen atoms at topological distance 5. The presence of two electronegative atoms increases the electronegativity rendering the compounds more electronegative [35]. The presence of large number of fragments in chemical structure will also increase the lipophilicity, ultimately enhancing the penetration ability of chemicals into the cell of the reference organism. Thus, the existence of oxygen atoms at the specified topological distance is associated with increased toxicity in pesticides as illustrated by compound **4**, while the opposite was characterized in compound **48** (provided in **Figure. 4.13**).

Another 2D atom pair descriptor, B02[C-O], indicates the occurrence of C-O at topological distance 2. It shows negative contribution toward the endpoint. This descriptor is related to hydrophilicity (oxygen is responsible for hydrogen bonding with water, and is easily excreted out from the body) [34]. Small fragments (occurrence of C-O at topological separation 2) are less lipophilic, as a result, toxicity will decrease which is evidenced by compound **30**, and the opposite was shown in compound **34** (represented in **Figure. 4.13**).

The T(N..S) descriptor denotes the summation of the topological distance between N..S and it contributed positively toward the endpoint. The occurrence of nitrogen and sulphur atoms in a compound increases its electronegativity, leading to oxidative stress and cell death [34-35]. Sulphur itself is toxic. Therefore, overall toxicity will increase as demonstrated in compound **33**. On the other hand, the compound containing less number of this fragment may exhibit less toxicity as shown in compound **53** (demonstrated in **Figure. 4.13**).

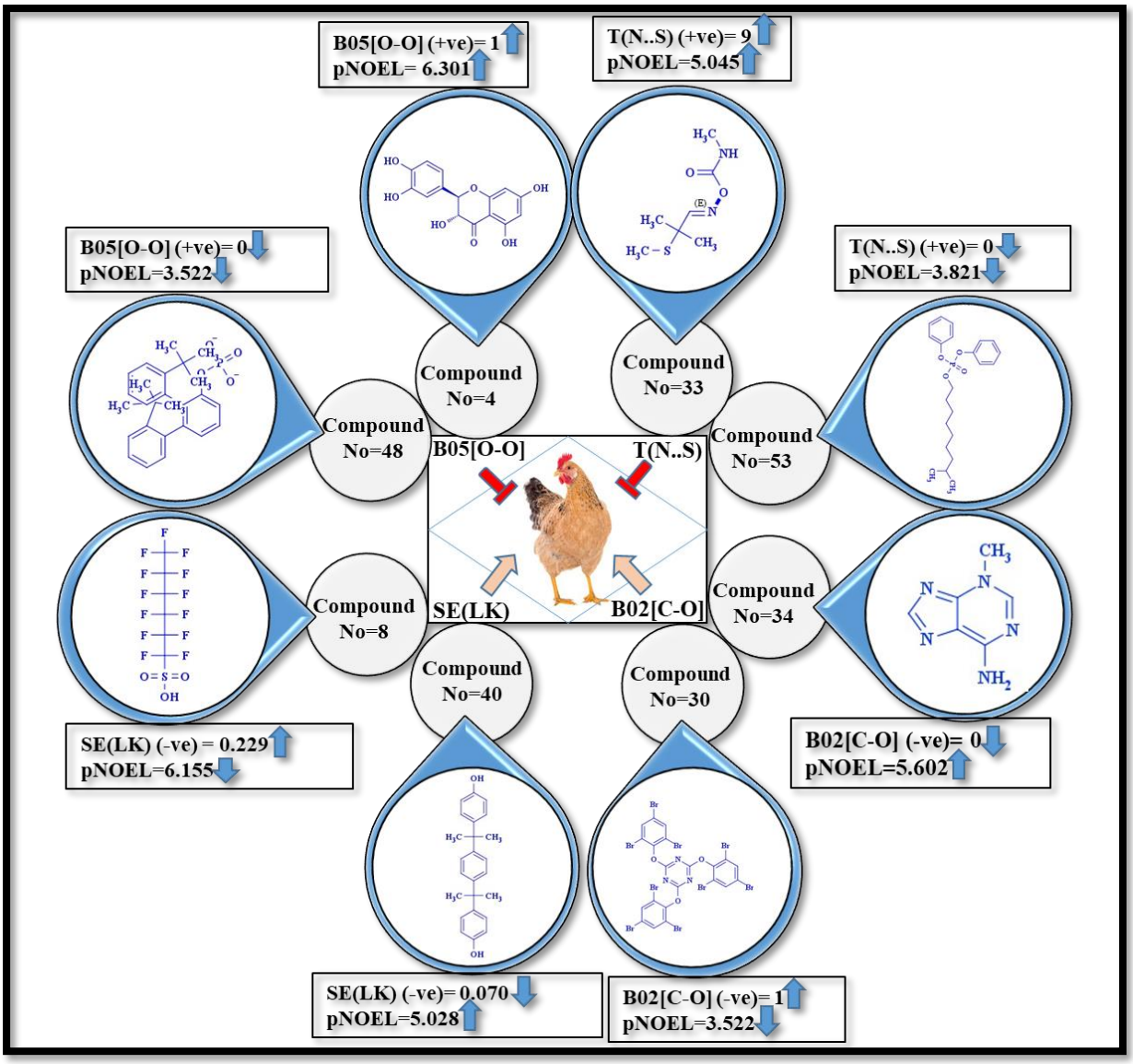


Figure 4.13 Contributions of the model descriptors towards pNOEL in chicken.

4.1.8. Pesticide Properties DataBase screening

The PPDB compounds were screened using developed models considering both the toxicity endpoints namely, pLOEL and pNOEL assisted by the Java-based tool "Prediction reliability indicator" (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). The applicability domain of the compounds was assessed to ascertain the reliability of the obtained prediction values and it was found that 100% and 55% of compounds lie within the chemical space of the developed pLOEL and pNOEL models respectively. The predicted pLOEL and pNOEL values of the respective compounds were cumulatively assessed. Then, based on the cumulative predictions, the top 20 and least 20 toxic compounds (compounds that are toxic for both pLOEL and pNOEL endpoints and lie within the AD of both models) with their CAS numbers, molecular weight, and pesticide groups have been provided in **Table 4.8.** Considering the top twenty

highest toxic compounds, our models' pLOEL and pNOEL prediction values were in complete coherence with the experimental toxicity data. From the results, it can be stated that our model predictions are correlated to real-world data and can be considered suitable for the identification of potential toxicants alongside less ones. Upon further validation, all predicted toxicities, demonstrate the practical applicability of the developed models.

Table 4.8. Twenty most and least toxic screened pesticides from the Pesticide Properties DataBase (PPDB).

Sl. No	Pesticide name (Group)	CAS no and Molecular mass	Safety and Hazards	Sources
Top 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Flumetsulam	98967-40-9 (Molecular mass-325.29)	Toxic to rats, rabbits, quail, ducks, and Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/91759#section=GHS-Classification&fullscreen=true
2	Dipyrithione	3696-28-4 (Molecular mass-252.31)	Environmental hazard, irritant	https://pubchem.ncbi.nlm.nih.gov/compound/3109#section=GHS-Classification&fullscreen=true
3	Sulfoxaflor	946578-00-3 (Molecular mass-277.27)	Environmental hazard, irritant	https://pubchem.ncbi.nlm.nih.gov/compound/16723172#section=GHS-Classification&fullscreen=true
4	Flusulfamide	106917-52-6 (Molecular mass-415.17)	Acute toxic to rats, mice, and Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/86268#section=GHS-Classification&fullscreen=true
5	Benzofluor	68672-17-3 (Molecular mass-299.33)	Threshold of Toxicological Concern (Cramer Class- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2711.htm
6	Nithiazine	58842-20-9 (Molecular mass-216.24)	Acute toxic to aves and irritants	https://pubchem.ncbi.nlm.nih.gov/compound/42853#section=EPA-Ecotoxicity&fullscreen=true
7	Perfluidone	37924-13-3 (Molecular mass-379.4)	Acute toxic to rats, rabbits, mice, and irritants	https://pubchem.ncbi.nlm.nih.gov/compound/37869#section=Acute-Effects&fullscreen=true
8	Fluensulfone	318290-98-1 (Molecular mass-291.70)	Acute toxic to fish and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/11534927#section=GHS-Classification&fullscreen=true
9	1,3-dinitrobenzene	99-65-0 (Molecular mass-168.12)	Acute toxic, Health hazard, and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/7452#section=GHS-Classification&fullscreen=true
10	Ampropylfos	16606-64-7 (Molecular mass-139.09)	Corrosive	https://pubchem.ncbi.nlm.nih.gov/compound/178368#section=GHS-Classification&fullscreen=true
11	Azoxybenzene	495-48-7 (Molecular mass-198.22)	Acute toxic to rats, mice, and rabbits	https://pubchem.ncbi.nlm.nih.gov/compound/10316#section=Acute-Effects&fullscreen=true

12	Benfluralin	1861-40-1 (Molecular mass-335.28)	Acute toxic to rats, mice, rabbits and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/2319#section=GHS-Classification&fullscreen=true
13	Benzamorf	12068-08-5 (Molecular mass-413.6)	Corrosive and Irritant	https://pubchem.ncbi.nlm.nih.gov/compound/20055166#section=GHS-Classification&fullscreen=true
14	Bis(methylmercury) sulphate	3810-81-9 (Molecular mass-527.31)	Threshold of Toxicological Concern (Cramer Class- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2716.htm
15	Bis-trichloromethyl sulfone	3064-70-8 (Molecular mass-300.80)	Acute toxic to rats, mice, rabbits and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/62478#section=GHS-Classification&fullscreen=true
16	Bromethalin	63333-35-7 (Molecular mass-577.9)	Acute toxic to rats, mice, dogs and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/44465#section=GHS-Classification&fullscreen=true
17	Butralin	33629-47-9 (Molecular mass-295.33)	Environmental hazard, Health hazard and Acute toxic to rats, rabbits	https://pubchem.ncbi.nlm.nih.gov/compound/36565#section=GHS-Classification&fullscreen=true
18	Cacodylic acid	75-60-5 (Molecular mass-138.00)	Acute toxic to rats, mice and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/2513#section=GHS-Classification&fullscreen=true
19	Chloropicrin	76-06-2 (Molecular mass-164.37)	Acute toxic to humans, rats and mice	https://pubchem.ncbi.nlm.nih.gov/compound/6423#section=GHS-Classification&fullscreen=true
20	Dicloran	99-30-9 (Molecular mass-207.01)	Environmental hazard, Health hazard and acute toxic to rat, mice	https://pubchem.ncbi.nlm.nih.gov/compound/7430#section=GHS-Classification&fullscreen=true
20 least screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Zarilamid	84527-51-5 (Molecular mass-238.67)	The predictive value for both endpoints indicates this pesticide is less toxic for both endpoints.	----------
2	Xylylcarb	2425-10-7 (Molecular mass-179.22)	Low toxic (Cramer Class): I	https://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2556.htm
3	Xylachlor	63114-77-2 (Molecular mass-239.77)	The test results show that metolachlor is practically non-toxic to birds. From the	https://www3.epa.gov/pesticides/chem_search/cleared_reviews/csr_PC-108801_21-Mar-94_205.pdf

			concept of structure-activity relationship, we can say xylachlor may also be non-toxic to birds.	
4	XMC	2655-14-3 (Molecular mass-179.22)	It has a low toxicity and is relatively stable	https://www.sciencedirect.com/science/article/abs/pii/S1386142521007654
5	Warfarin	81-81-2 (Molecular mass-308.35)	It is practically non-toxic	http://extoxnet.orst.edu/pips/warfarin.htm
6	Vinegar	90132-02-8 (Molecular mass-60.06)	Vinegar is used to promote the health of the birds	https://haithspro.wordpress.com/category/vinegar-bird-health/
7	Vinclozolin	50471-44-8 (Molecular mass-286.12)	Vinclozolin is practically nontoxic to birds	https://archive.epa.gov/pesticides/chemicalsearch/chemical/foia/web/pdf/113201/113201-142.pdf
8	Uniconazole	83657-22-1 (Molecular mass-291.81)	Uniconazole-p is non-toxic to birds	https://apvma.gov.au/sites/default/files/publication/14096-prs-uniconazole-p.pdf
9	Umifoxolaner	2021230-37-3 (Molecular mass-299.64)	Low toxic	https://www.ema.europa.eu/en/documents/assessment-report/nexgard-spectra-epar-public-assessment-report_en.pdf
10	Triticonazole	131983-72-7 (Molecular mass-317.82)	Triticonazole is non-toxic to pollinating insects	https://downloads.regulations.gov/EPA-HQ-OPP-2015-0602-0039/content.pdf
11	Triprene	40596-80-3 (Molecular mass-312.52)	Low toxic	https://hal.science/hal-00891905/document
12	Trimethacarb	12407-86-2 (Molecular mass-312.52)	Birds were not as sensitive to trimethacarb	https://escholarship.org/uc/item/91t7r9mv
13	Triisopropanolamine	122-20-3 (Molecular mass-191.27)	Practically non-toxic to birds, fish, honeybees	https://downloads.regulations.gov/EPA-HQ-OPPT-2013-0739-0140/attachment_1.pdf
14	Triflumuron	64628-44-0 (Molecular mass-358.70)	Triflumuron is not classified as toxic or highly toxic	http://dissemination.echa.europa.eu/Biocides/ActiveSubstances/1407-18/1407-18_Assessment_Report.pdf
15	Triflumizole	99387-89-0 (Molecular mass-345.75)	Triflumizole is categorized as being moderately	https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000QRHX.TXT

			toxic to fish	
16	Triflumezopyrim	1263133-33-0 (Molecular mass-398.34)	Triflumezopyrim was harmless to Anagrus nilaparvatae	https://pubmed.ncbi.nlm.nih.gov/29404868/
17	Trifloxystrobin	141517-21-7 (Molecular mass-408.37)	Trifloxystrobin is practically non-toxic to birds	https://www.apvma.gov.au/sites/default/files/publication/14081-prs-trifloxystrobin.pdf
18	Trifenofos	38524-82-2 (Molecular mass-363.63)	Profenofos has a moderate toxic	https://apps.who.int/pesticide-residues-jmpr-database/Document/123
19	Trifenmorph	1420-06-0 (Molecular mass-329.43)	Trifenmorph is hydrolysed at acid pH to relatively non - toxic compounds	http://erepository.uonbi.ac.ke/bitstream/handle/11295/21816/Benigna_Lethal%20and%20sub%20-%20lethal%20effects%20of%20dot%2C%20carbofuran%2C%20trifenmorph%20and%20niclosamide%20on%20oreochromis%20niger%20guther%20%281898%29.pdf?sequence=3&isAllowed=y
20	Tridiphane	58138-08-2 (Molecular mass-320.43)	The predictive value for both endpoints indicates this pesticide is less toxic for both endpoints.	----------

4.2 Study 2

4.2.1. PLS-based QSTR Model

A PLS-based QSTR model was developed using the PLS regression method with four latent variables from ten different features identified using the best subset selection tool against avian species. The developed PLS-based QSTR model is given below:

PLS-based QSTR model:

$$\textbf{Model M1} = 0.14334 + 0.17079*X5v + 0.56174*Br\text{-}094 - 0.28210*B07[C\text{-}C] + 1.85683*nPyrrolidines + 0.65279*F02[S\text{-}F] + 0.18407*C\text{-}003 - 0.39449*nCrq + 1.85133*B03[N\text{-}P] + 0.17754*nCXr - 0.31913*nR07$$

The model's performance has been thoroughly evaluated using rigorous internal and external validation methods following the guidelines of the OECD. The determination coefficient (R^2= 0.624) and leave one out cross-validated correlation coefficient (Q^2_{LOO}= 0.538) indicate the model's goodness of fit and robustness, whereas the mean absolute error of the training set

predictions (MAE_{train} = 0.247) indicates the predictive error. In addition to these, external validation metrics, such as the external predicted variance (Q^2_{F1}= 0.539 and Q^2_{F2}= 0.538), and mean absolute error of test set prediction (MAE_{test}= 0.268), which are the standard markers of good external predictability, have also been calculated.

4.2.2. PLS-based q-RASTR Model

In this study, we aimed to create a q-RASTR model to improve the external predictability of the corresponding QSTR model. To achieve this, we integrated the calculated read-across-based RASTR descriptor with the pull of ten alvaDesc descriptors. Using the best subset selection method, we obtained a new combination of descriptors. We then used PLS regression to model eight descriptors with the optimal number of latent variables (four LVs). The statistical metrics of the PLS-based QSTR and q-RASTR models are presented in **Table 4.9**. The resulting PLS-based q-RASTR model is given below:

PLS-based q-RASTR model:

Model M2 *=0.14648 + 0.53019* CVsim(LK) - 0.75982* SD similarity(LK) + 0.04142* gm*Avg.Sim + 0.12455* X5v -0.17495* B07[C-C] + 1.61151* nPyrrolidines + 0.08476* C-003 -0.46958* nCrq*

Table 4.9. Statistical quality of QSTR and q-RASTR model.

	Algorithm	**Training set**					**Test set**		
		N_{train}/ N_{test}	DES/ LVs	R^2	Q^2_{LOO}	$MAE_{(train)}$	Q^2_{F1}	Q^2_{F2}	$MAE_{(test)}$
Avian Species	PLS-QSTR	360/ 120	10/4	0.624	0.538	0.247	0.539	0.538	0.268
	PLS q-RASTR	360/ 120	8/4	0.623	0.569	0.247	0.541	0.540	0.261

The PLS-based q-RASTR model shown superior performance than the corresponding QSTR model in terms of internal validation metrics (Q^2_{Loo} =0.569) as well as external validation metrics (Q^2_{F1} =0.541, Q^2_{F2} =0.540). The model also produced the lowest prediction error for the test compounds as indicated by MAE_{test} = 0.261. The results of our study indicate that the best-performing models for predicting the response values of target compounds were found to be the PLS-based q-RASTR model. Y-randomization was carried out to investigate the chance occurrence of the developed model R^2y_{rand} and Q^2y_{rand} were found to be less than the standard threshold, which assures that the generated models were not obtained by any chance as depicted

in **Figure.4.14.** The PLS-based q-RASTR model's goodness-of-fit has been confirmed by evaluating the correlation between the observed and predicted values as shown in **Figure. 4.15.**

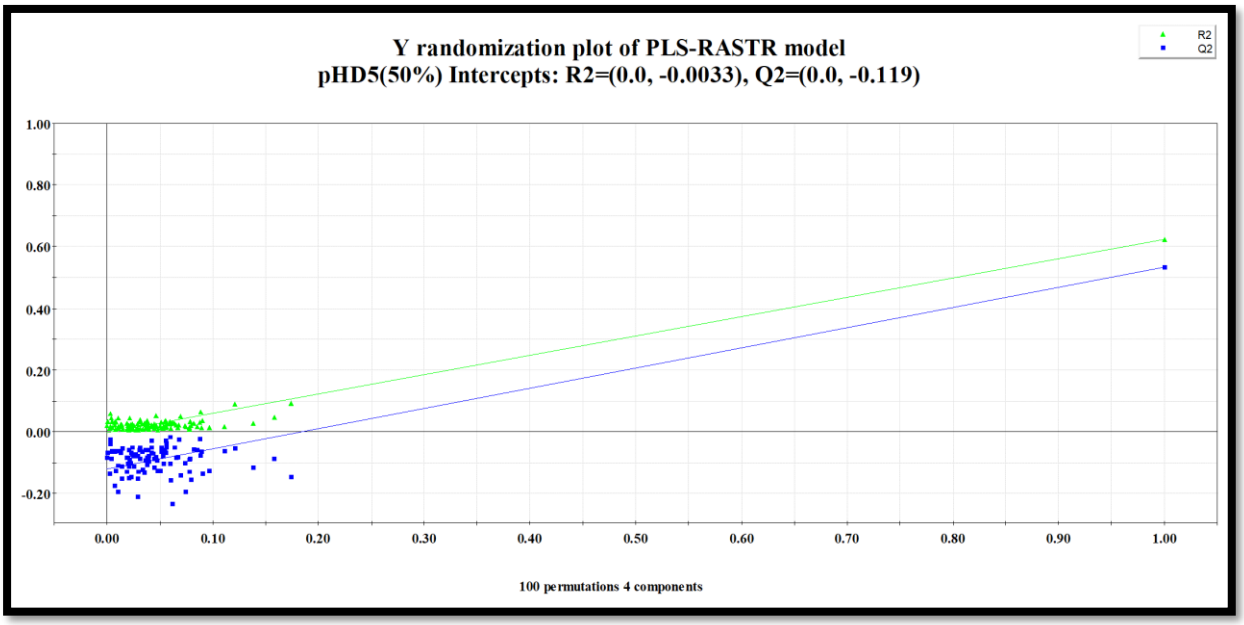


Figure 4.14. Y-randomization plot of PLS-based q-RASTR model.

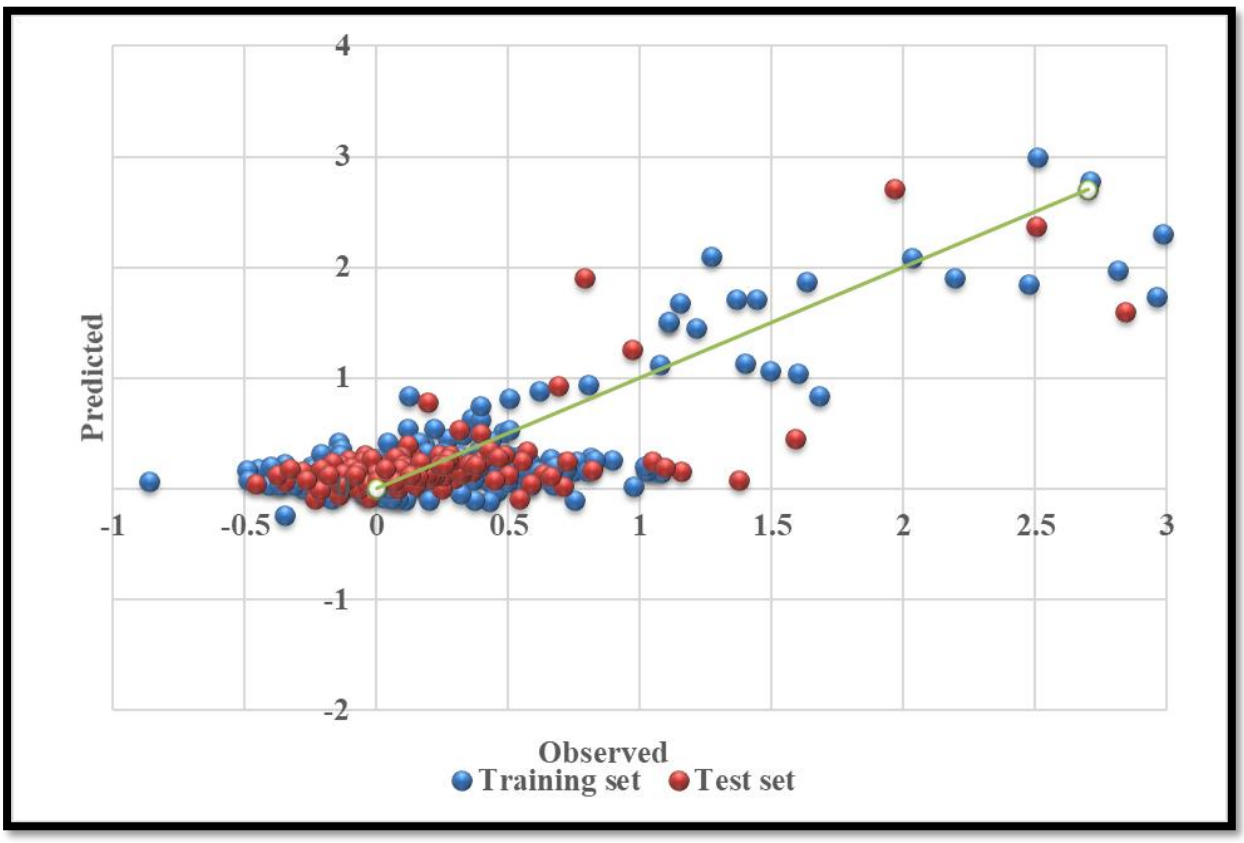


Figure 4.15. Scatter plot of established model.

4.2.3 PLS plots

The comprehensive use of PLS plots in SIMCA-P facilitated a detailed exploration of the dataset, providing valuable insights into the relationships between variables and the toxicity response. Each type of PLS plot played a crucial role in enhancing the understanding and reliability of the predictive models. The outcomes of these analyses contribute not only to model interpretation but also guide further refinement and optimization for robust predictive performance.

4.2.3.1 Regression coefficient plot

Employed SIMCA-P to generate regression coefficient plots, illustrating the contribution of each variable on the response variable. Evaluated the sign and magnitude of regression coefficients to discern the variables positively or negatively influencing the toxicity, offering valuable insights for understanding the underlying mechanisms. The descriptors CVsim(LK), gm*Avg.Sim, X5v, nPyrrolidines, and C-003 contributed positively towards the toxicity which indicates that the toxicity enhanced with increasing the numerical value of these descriptors while the descriptors SD similarity(LK), B07[C-C], and nCrq showed negative contribution towards the toxicity which indicated that the toxicity reduced with increasing the numerical value of these descriptors. The regression coefficient plot is provided in **Figure. 4.16.**

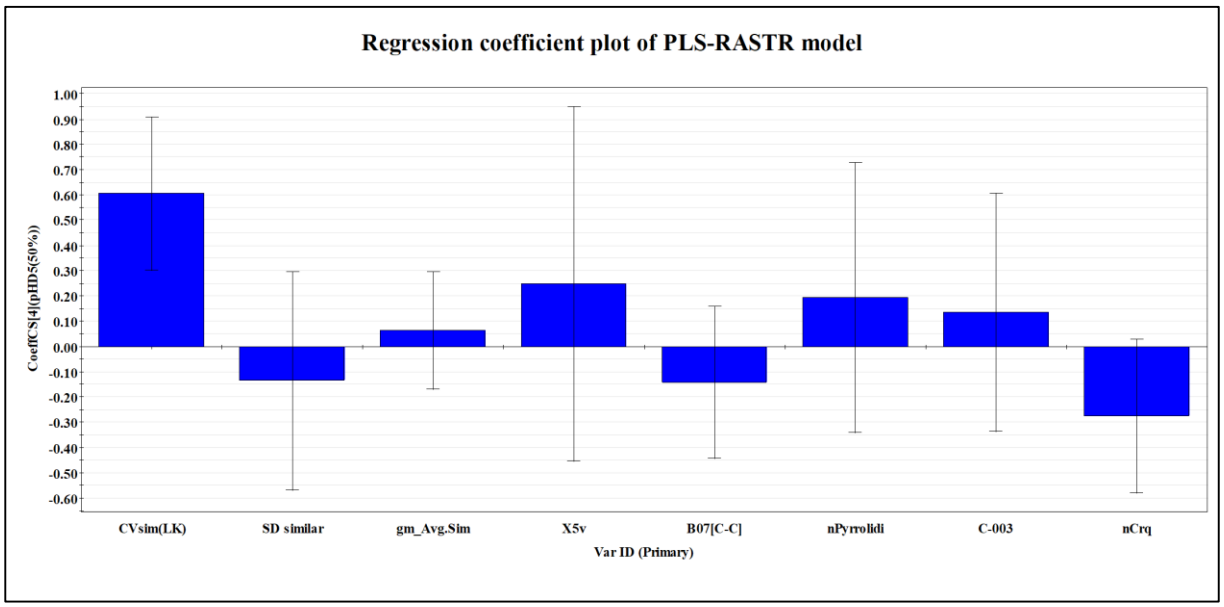


Figure 4.16. Regression coefficient plot of PLS-based q-RASTR model.

4.2.3.2. Variable importance plot (VIP)

Utilized the SIMCA-P software to generate VIP plots, quantifying the importance of each variable in explaining the variations observed in the data. Variables with higher VIP scores were considered more influential, aiding in feature selection and enhancing the interpretability of the predictive models. The influential descriptors toward toxicity of the developed model are CVsim(LK)> X5v> nCrq> SD similarity(LK)> C-003> nPyrrolidines> B07[C-C]> gm*Avg.Sim (arranged in higher to lower order as per their VIP score). The VIP plots are depicted in **Figure. 4.17.**

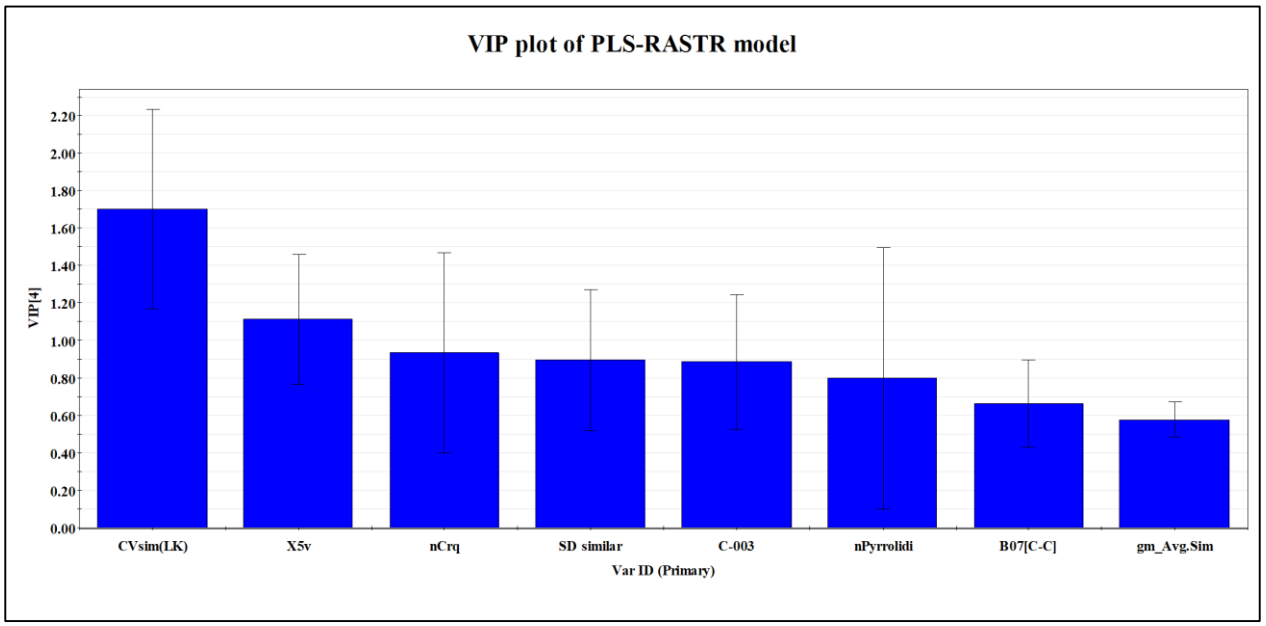


Figure. 4.17. Variable_importance plot PLS-based q-RASTR model.

4.2.3.3. Loading plot

The loading plot, portrayed in **Figure. 4.18.** identifies the relationship between the model's X-variables (independent variables) and Y-variables (dependent variables). The first two components of the developed model 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, CVsim(LK) descriptor is the most impactful variable for the PLS-based q-RASTR models as it is present farthest from the origin.

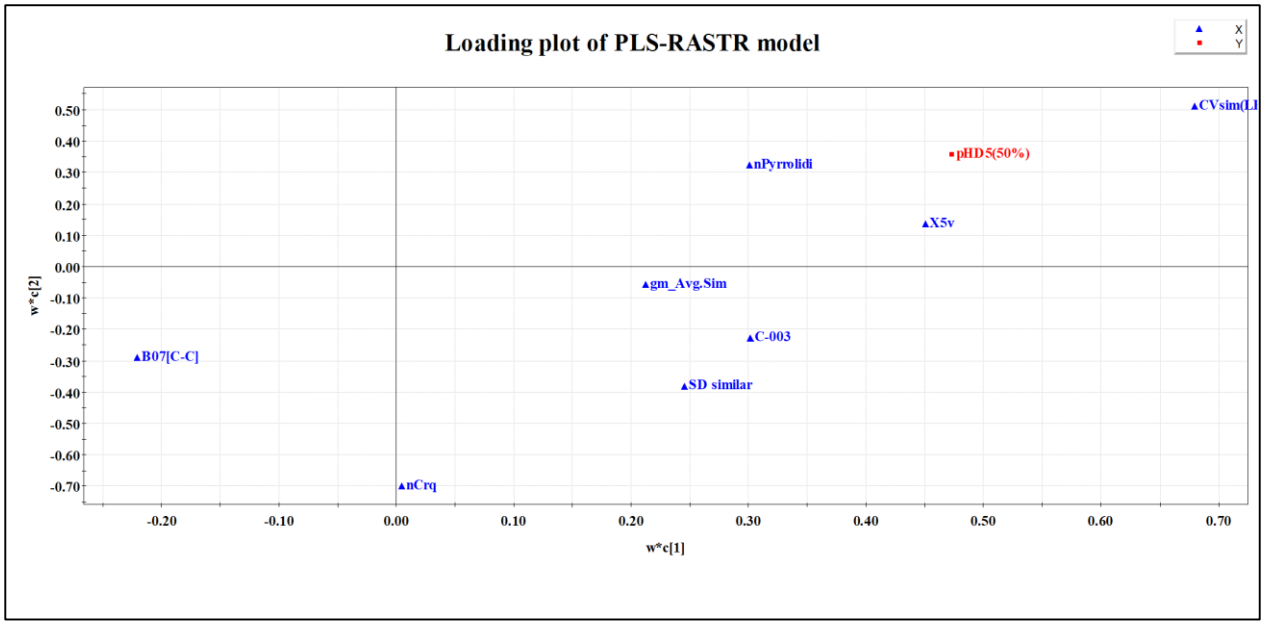


Figure 4.18. Loading plot of the model PLS-based q-RASTR model.

4.2.4. Mechanistic interpretation

The information regarding the descriptors gained from the developed model, their contribution, description, and probable mechanistic interpretation are provided in **Table 4.10** and presented in **Figure. 4.19**.

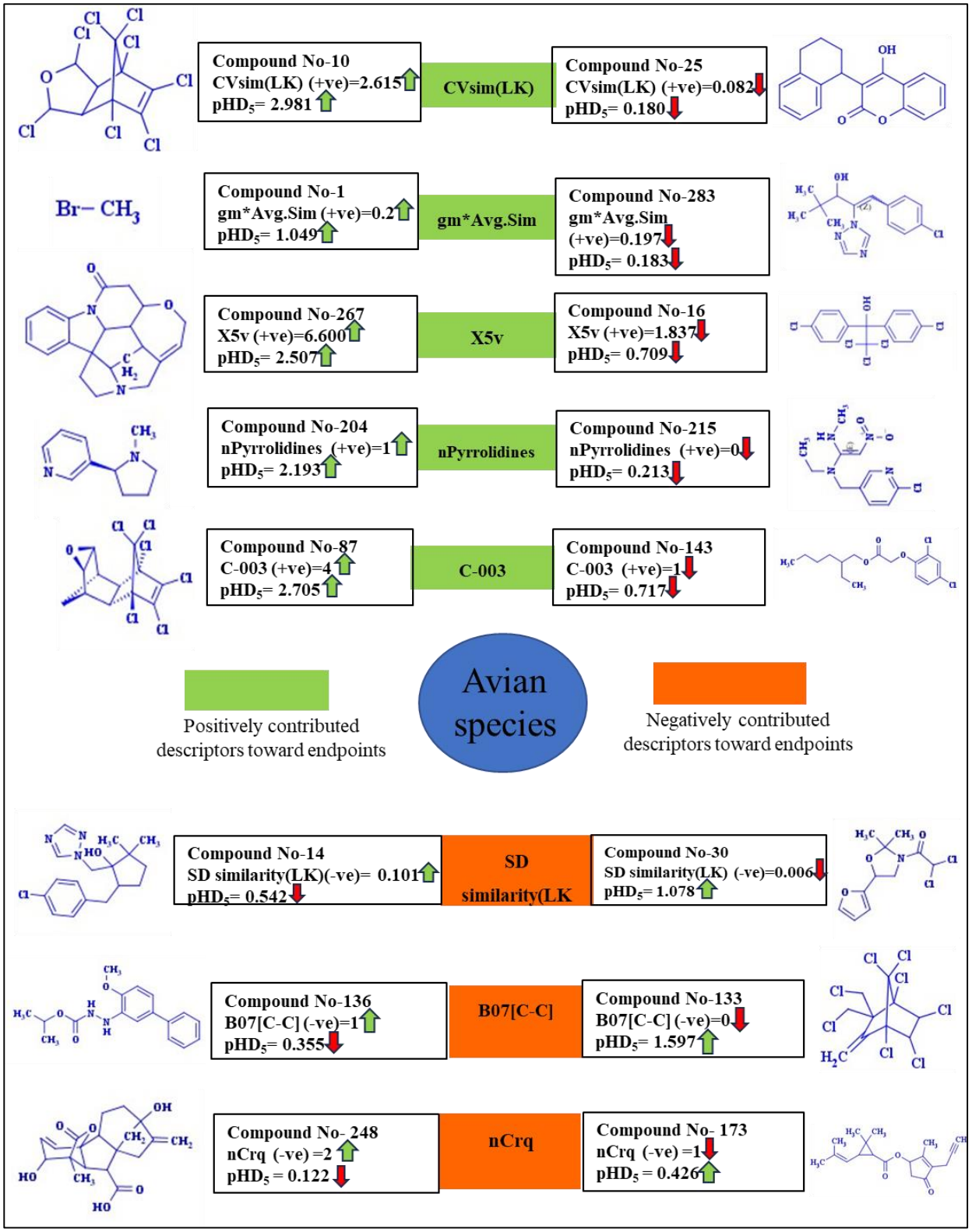


Figure. 4.19. Mechanistic introspection of modeled descriptors.

Table 4.10. Mechanistic introspection of modeled descriptors.

Sl. No.	Descriptor	Type	Description	Contribution
Various types of bird species				
1	CVsim(LK)	RASTR	Coefficient and variation of the similarity values of the close source compounds.	Positive
	Mechanistic introspection: The descriptor "CVsim (LK)" quantifies the coefficient of variation in similarity values for chemicals resembling the target molecule. A high value of this descriptor suggests that the target molecule exhibits significant deviations in similarity to related chemicals, which may correlate with increased toxicity towards the endpoint. For example, compound **10 (Isobenzan)** shows considerable variability in similarity measures among its close chemical analogs, indicating potential toxicity. In contrast, compound **25 (Coumatetralyl)** demonstrates a lower coefficient of variation in similarity values, suggesting reduced toxicity towards the endpoint.			
2	SD similarity(LK)	RASTR	The standard variation in similarity measures among closely related compounds	Negative
	Mechanistic introspection: The extensive variability observed among closely related source compounds is a significant factor that reduces the prediction reliability. This fact is evident in compound **14 (Metconazole),** whereas compound **30 (Furilazole)** does not demonstrate such phenomenon.			
3	gm*Avg.Sim	RASTR	Product of the g_m and *Avg.Sim* levels	Positive
	Mechanistic introspection: Increasing the numerical value of this variable enhanced the compound's toxicity (directly related to the toxicity as indicated by the positive regression coefficient) as represented in compound **1 (bromomethane)** and oppositely occurs in compound **283 (Uniconazole**).			
4	X5v	Connectivity indices	Valence connectivity index of order 5	Positive
	Mechanistic introspection:			

<table>
<tr><td></td><td colspan="4">The symbol X5v is commonly used to represent the valence connectivity index of order 5. However, based on the study of certain organic compounds, it can be inferred that X5v represents the extent of branching or molecular surface area. It has been observed that X5v has a positive correlation with the endpoint. This means when the numerical value of X5v increases (branching increases), the toxicity also increases [83] as traced in compound 267 (Strychnine) and conversely in compound 16 (Dicofol).</td></tr>
<tr><td rowspan="2">5</td><td>B07[C-C]</td><td>2D Atom Pairs</td><td>Presence/absence of C – C at topological distance 7</td><td>Negative</td></tr>
<tr><td colspan="4">Mechanistic introspection:
This descriptor is inversely correlated with the toxicity as indicated by its negative regression coefficient. Thus, an increased number of this fragment correlates with decreased toxicity, as illustrated by compound 136 (Bifenazate (D2341), while the opposite effect is observed in compound 133 (Toxaphene).</td></tr>
<tr><td rowspan="2">6</td><td>nPyrrolidines</td><td>Functional group counts</td><td>number of Pyrrolidines</td><td>Positive</td></tr>
<tr><td colspan="4">Mechanistic introspection:
The nPyrrolidines alvaDesc descriptor positively contributed to the endpoint. This suggests that the presence of pyrrolidine rings enhances toxicity, as exemplified by compound 204 (Nicotine), whereas the reverse effect is observed in compound 215 (Nitenpyram).</td></tr>
<tr><td rowspan="2">7</td><td>C-003</td><td>Atom-centred fragments</td><td>CHR_3</td><td>Positive</td></tr>
<tr><td colspan="4">Mechanistic introspection:
The positive regression coefficient for this descriptor suggests that it enhances the toxicity profile of the chemicals, as evidenced by compound 87 (Endrin) and conversely in compound 143 (2,4-D Isooctyl ester).</td></tr>
<tr><td rowspan="2">8</td><td>nCrq</td><td>Functional group counts</td><td>number of ring quaternary C(sp3)</td><td>Negative</td></tr>
<tr><td colspan="4">Mechanistic introspection:
Generally, the Sp^3 hybridized compound is more stable and less reactive due to the presence of a sigma bond. The less reactivity of any compound indicates that it is potentially less toxic. This feature has a positive contribution towards the response</td></tr>
</table>

	as depicted in compounds **248 (Gibberellic acid**) and conversely for compound **173 (Prallethrin).**

4.2.5. Pesticide Properties DataBase screening

The developed PLS-based q-RASTR model was utilized to screen the PPDB database using PRI tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). We have ensured the utmost reliability of the prediction values by thoroughly assessing the applicability domain of the compounds and discovering that 92.08% of the compounds fall within the chemical space of the developed model. The HD_5 values of the compounds were evaluated and the twenty highest and least toxic compounds have been provided with their respective CAS numbers, molecular weight, and pesticide groups in **Tables 4.11** and **4.12**. respectively. Our predictions were rigorously validated by corroborating them with the real-world experimental data available in the PubChem online repository, as well as in literature and references. We observed complete coherence between our predictions and the experimental toxicity data, particularly for the top twenty highest and least twenty toxic compounds. Therefore, we confidently state that our model predictions are reliable and can be considered highly suitable for identifying potential toxicants.

Table 4.11 Top twenty toxic screened pesticides from Pesticide Properties DataBase (PPDB).

Sl. No	**Pesticide name**	**CAS no and Molecular mass**	**Safety and Hazards**	**Sources**
Top 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Chlorbicyclen	103-17-3 (Molecular mass-269.19)	Acute toxic (Dermal, Inhalation)	https://pubchem.ncbi.nlm.nih.gov/compound/17357#section=GHS-Classification&fullscreen=true
2	Dialifos	10311-84-9 (Molecular mass-393.85)	Acute toxic to rats, mice, dogs, rabbits, and an Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/25146#section=GHS-Classification&fullscreen=true
3	Schradan	152-16-9 (Molecular mass-286.25)	Acute toxic to man, rats, etc.	https://pubchem.ncbi.nlm.nih.gov/compound/9037#section=GHS-Classification&fullscreen=true

4	Bromethalin	63333-35-7 (Molecular mass-577.9)	Acute toxic to rats, mice, dogs, and Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/44465#section=GHS-Classification&fullscreen=true
5	Imicyafos	140163-89-9 (Molecular mass-304.35)	Acute toxic, Irritant	https://pubchem.ncbi.nlm.nih.gov/compound/18772487#section=GHS-Classification&fullscreen=true
6	Bromocyclen	1715-40-8 (Molecular mass-393.75)	Inhalation May be harmful if inhaled.	https://www.hpc-standards.com/shop/ReferenceMaterials/Pesticides/Bromocyclen_Ethylacetate_1.htm
7	Phosalone	2310-17-0 (Molecular mass-367.8)	Acute toxic, Environmental Hazard, Irritant	https://pubchem.ncbi.nlm.nih.gov/compound/4793#section=GHS-Classification&fullscreen=true
8	Prothidathion	20276-83-9 (Molecular mass-358.4)	Threshold of Toxicological Concern (Cramer Class)- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2847.htm
9	Mazidox	7219-78-5 (Molecular mass-177.15)	Threshold of Toxicological Concern (Cramer Class)- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2870.htm
10	Pyrafluprole	315208-17-4 (Molecular mass-477.27)	(Cramer Class)- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/3071.htm
11	Diazinon	333-41-5 (Molecular mass-304.35)	Irritant, Health hazard, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/3017#section=GHS-Classification&fullscreen=true
12	Athidathion	19691-80-6 (Molecular mass-330.4)	Acute toxic (dermal, oral, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/88197#section=GHS-Classification&fullscreen=true
13	Azinphos-ethyl	2642-71-9 (Molecular mass-345.38)	Acute toxic to rat,dog and environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/17531#section=GHS-Classification&fullscreen=true
14	Fosmethilan	83733-82-8 (Molecular mass-367.8	Acute toxic to quail, bird-domestic and rat	https://pubchem.ncbi.nlm.nih.gov/compound/55138#section=Acute-Effects&fullscreen=true

15	Benzobicyclon	156963-66-5 (Molecular mass-446.96)	Toxic for aves	https://pubchem.ncbi.nlm.nih.gov/compound/11236201#section=Toxicity&fullscreen=true
16	Phosmet	732-11-6 (Molecular mass-317.33)	Acute toxic, Irritant, Health hazard, Environmental hazard	https://pubchem.ncbi.nlm.nih.gov/compound/12901#section=GHS-Classification&fullscreen=true
17	Tralopyril	122454-29-9 (Molecular mass-349.53)	Acute toxic, Health hazard,	https://pubchem.ncbi.nlm.nih.gov/compound/183559#section=GHS-Classification&fullscreen=true
18	Halacrinate	34462-96-9 (Molecular mass-312.55)	Acute toxic to rat	https://pubchem.ncbi.nlm.nih.gov/compound/114868#section=Acute-Effects&fullscreen=true
19	Fluazolate	174514-07-9 (Molecular mass-443.62)	Threshold of Toxicological Concern (Cramer Class)- High (class III)	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/326.htm
20	Bromophos	2104-96-3 (Molecular mass-366.00)	Environmental hazard and acute toxic to rat	https://pubchem.ncbi.nlm.nih.gov/compound/16422#section=Acute-Effects&fullscreen=true

Table 4.12 Least twenty screened pesticides from Pesticide Properties DataBase (PPDB).

Sl. No	**Pesticide name**	**CAS no and Molecular mass**	**Safety and Hazards**	**Sources (all references available in Supplementary 2)**
Least 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Isophorone	2104-96-3 (Molecular mass-366.00)	**Isophorone** does **not** affect the fertility or cause developmental toxicity in experimental animals (Rats, Mice)	https://www.inchem.org/documents/hsg/hsg/hsg91_e.htm
2	Empenthrin	54406-48-3 (Molecular mass-274.40)	Emperithrin has a **low** mammalian toxicity. It is **not** highly **toxic** to **birds**	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/1596.htm

3	Profluthrin	223419-20-3 (Molecular mass-330.32)	Metofluthrin, like other synthetic pyrethroids, is practically **non-toxic**	https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_PC-109709_01-Sep-06.pdf
4	Heptafluthrin	1130296-65-9 (Molecular mass-414.12)	The substance **has no implications** for human health, biodiversity or the environment	http://sitem.herts.ac.uk/aeru/ppdb/en/docs/Data_alerts_rules.pdf
5	Metofluthrin	240494-70-6 (Molecular mass-360.34)	The substance **has no implication**s for human health, biodiversity or the environment	http://sitem.herts.ac.uk/aeru/ppdb/en/docs/Data_alerts_rules.pdf
6	Epsilon-metofluthrin	240494-71-7 (Molecular mass-360.34)	**Epsilon-**momfluorothrin has **low** acute **toxicity**	https://echa.europa.eu/documents/10162/e81b30a1-400a-9fed-b8dd-362a3a54f08b
7	Imiprothrin	72963-72-5 (Molecular mass-318.37)	The chemical is practically **non-toxic** to **birds**	https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_PC-004006_01-Mar-98.pdf
8	Transfluthrin	118712-89-3 (Molecular mass-371.15)	Transfluthrin is **classified as practically non-toxic to birds and mammals**,	https://downloads.regulations.gov/EPA-HQ-OPP-2016-0581-0007/content.pdf
9	Tefluthrin	79538-32-2 (Molecular mass-371.15)	Tefluthrin is nontoxic to birds	https://www.sciencedirect.com/science/article/abs/pii/S0013935120308884
10	Kappa-tefluthrin	391634-71-2 (Molecular mass-418.7)	Nontoxic to mammals	https://patents.google.com/patent/EP3696177A1/en
11	Fenfluthrin	75867-00-4 (Molecular mass-389.16)	Practically non-toxic to slightly toxic when eaten by birds	http://npic.orst.edu/factsheets/cyfluthringen.html
12	Renofluthrin	352271-52-4 (Molecular mass-415.22)	No data found	----------

13	Meperfluthrin	915288-13-0 (Molecular mass-415.21)	Metofluthrin, like other synthetic pyrethroids, is practically **non-toxic**	https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_PC-109709_01-Sep-06.pdf
14	S-bioallethrin	28434-00-6 (Molecular mass-302.41)	**Bioallethrin** is less toxic to **birds** and honeybees.	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/80.htm
15	Bioallethrin	260359-57-7 (Molecular mass-302.41)	**Bioallethrin** is less toxic to **birds** and honeybees.	http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/80.htm
16	Allethrin	584-79-2 (Molecular mass-302.41)	**Allethrin** is practically **non-toxic** to **birds**	http://extoxnet.orst.edu/pips/allethri.htm
17	Momfluorothrin	609346-29-4 (Molecular mass-385.35)	**Momfluorothrin** is considered practically **non-toxic** to **birds** and mammals	https://downloads.regulations.gov/EPA-HQ-OPP-2013-0478-0020/content.pdf
18	Chloropralletherin	250346-55-5 (Molecular mass-341.23)	**Prallethrin** is of **low** mammalian **toxicity**	https://en.wikipedia.org/wiki/Prallethrin
19	Acrinathrin	101007-06-1 (Molecular mass-541.44)	It is **not** considered as harmful to **birds**	https://luxembourg.co.il/wp-content/uploads/2020/02/Rufast-1212.pdf
20	Formetanate hydrochloride	23422-53-9 (Molecular mass-257.8)	Toxic compounds	https://archive.epa.gov/pesticides/reregistration/web/html/formetanatehcl_fs.html

4.3 Study 3

In this study, we have developed PLS models utilizing the toxicity of pesticides ($LogLC_{50}$) on four different avians (BQ and JQ) 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 4.13**. Read-Across algorithm was employed to improve the model's external predictivity External predictivity was improved for both datasets (BQ, JQ) in Read-Across prediction, and results are provided in **Table 4.14**. 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) which demonstrated that the models were not formed by any chance. A visual representation of the correlation between observed and predicted toxicity values has been depicted in the scatter plot (provided in **Figure 4.20**). Additionally, we used two different ML algorithms namely support 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. The equations of the final developed models of BQ and JQ, 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:

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

Table 4.13. 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

Table 4.14. Read-across based predictions for four species.

Optimized settings	**Metrics**	**Ygk (Test)**
Bobwhite quail		
Ygk (Test) **σ = 0.25** **γ = 0.25** **No. of similar compounds =10**	Q^2_{F1}	0.690
	Q^2_{F2}	0.690
	$RMSE_P$	0.279
	MAE	0.179
Japanese quail		
Optimized settings	**Metrics**	**Ylk (Test)**
σ = 0.25 **γ = 0.25** **No. of similar compounds =10**	Q^2_{F1}	0.707
	Q^2_{F2}	0.698
	$RMSE_P$	0.394
	MAE	0.307

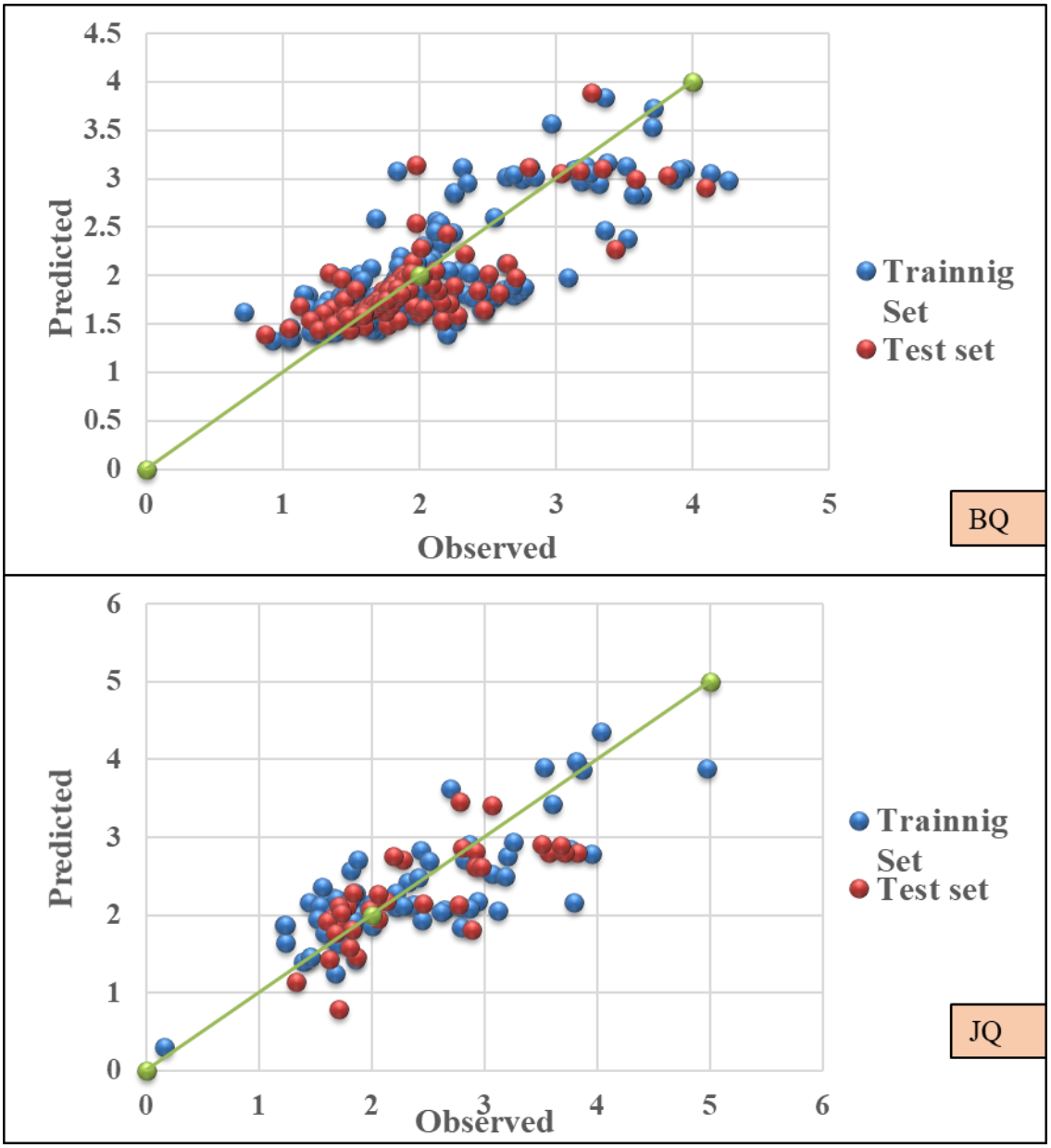


Figure 4.20. Scatter plots of developed models.

Several classification-based metrics have been computed with the PLS-based QSTR-read across models for all (BQ, and JQ) the avian species and reported in the following **Table 4.15** Good sensitivity, specificity, and accuracy values indicate the good classification ability of the model. The computed values of the Matthews correlation coefficient [49] indicate an acceptable prediction and an agreement between observed and predicted classification for all the developed models against avian species.

Table 4.15. 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

4.3.1. Regression coefficient plot

The descriptor's positive/negative contribution towards 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 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.

4.3.2. Variable importance plot (VIP)

The relative importance of model descriptors is illustrated with VIP [51]. Descriptors having the highest and lowest impact on avian species can be recognized from these plots. The significance of the variable is higher 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).

4.3.3. 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. Based on the loading plot; it is interpreted that the X-variables F02[C-P] and MW have more influence on the Y-variable as traced from the proximity with the response variable and the presence of

these features elevated pesticide toxicity towards BQ. Similarly, B01[O-P] are the most influential descriptors in the case of JQ.

4.3.4. Mechanistic interpretation of PLS models

Table 4.16. and **Figures 4.21.-4.22.** provide a detailed account of the model descriptors followed by mechanistic interpretations important to identify major structural and physicochemical features.

Table 4.16. Mechanistic analysis of model descriptors of all species.

S.no	Descriptor	Type	Function	Contribution
BQ oral pLC_{50}				
1	F02[C-P]	2D Atom pair	Frequency of carbon and phosphorus atoms at topological distance 2	+ve
	Mechanistic introspection Generally, the phosphate group is toxic. 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 **Figure. 4.21**).			
2	MW	Constitutional descriptor	Molecular weight	+ve
	Mechanistic introspection This descriptor is directly related to the molecular size and bulkiness of molecules. It may influence diffusion in biological membranes and fluid media. So the drug 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 as demonstrated in compound **381** and vice versa in compound **239** (given in **Figure. 4.21**).			
3	F09[S-F]	2D Atom pair	Frequency of sulfur and fluorine atoms at topological distance 9	+ve

<table>
<tr><td></td><td colspan="4">Mechanistic introspection
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 as shown in compound 23 and oppositely occurs in compound 523 (shown in Figure. 4.21).</td></tr>
<tr><td>4</td><td>B09[C-P]</td><td>2D Atom pair</td><td>Presence/absence of carbon and phosphorus atoms at topological distance 9</td><td>-ve</td></tr>
<tr><td></td><td colspan="4">Mechanistic introspection
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 Figure. 4.21).</td></tr>
<tr><td>5</td><td>F03[O-P]</td><td>2D Atom pair</td><td>Frequency of oxygen and phosphorus atoms at topological distance 3</td><td>-ve</td></tr>
<tr><td></td><td colspan="4">Mechanistic introspection
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 Figure. 4.21).</td></tr>
<tr><td>6</td><td>T(P..Cl)</td><td>2D Atom pair</td><td>Sum of topological distances between P..Cl</td><td>-ve</td></tr>
<tr><td></td><td colspan="4">Mechanistic introspection
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. This descriptor has a negative regression coefficient, indicating that the presence of this fragment will result in a decrease in pesticide toxicity profile, as</td></tr>
</table>

	exemplified by compound **243**, while it would have the opposite effect when present, as proven by compound **441** (provided in **Figure. 4.21)**.			
JQ oral pLC_{50}				
1	B01[O-P]	2D Atom pair	Presence/absence of O – P at topological distance 1	+ve
	Mechanistic introspection 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. This phenomenon is demonstrated in compound **81** and inversely occurs in compound **113** (shown in **Figure. 4.22**).			
2	X2A	Connectivity indices descriptor	Average connectivity index of order 2	-ve
	Mechanistic introspection X2A represents the degree of branching in molecules, which is inversely correlated with hydrophobic interaction as well as toxicity. 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 **Figure. 4.22**).			
3	B05[N-P]	2D Atom pair	Incidence of N – P at topological distance 5	+ve
	Mechanistic introspection 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. 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 **Figure. 4.22**).			
4	H-048	Atom-centered fragments	H attached to C2(sp3)/C1(sp2)/C0(sp)	-ve
	Mechanistic introspection H-048 has the potential to make compounds electronically conductive as well as hydrophilic. Hydrophilicity and toxicity are inversely related to each other. 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 **Figure. 4.22**).			
5	T(O..Cl)	2D Atom pair	Sum of topological distances between O..Cl	-ve

	Mechanistic introspection 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 **Figure. 4.22**).			
6	nBridgeHead	Ring descriptors	Number of bridgehead atoms	+ve
	Mechanistic introspection Usually, bridgehead atoms have a complex structure as well as toxic which is demonstrated in compound **19**. Conversely, the absence of bridgehead atoms makes the compound less toxic as shown in compound **110** (demonstrated in **Figure. 4.22**).			

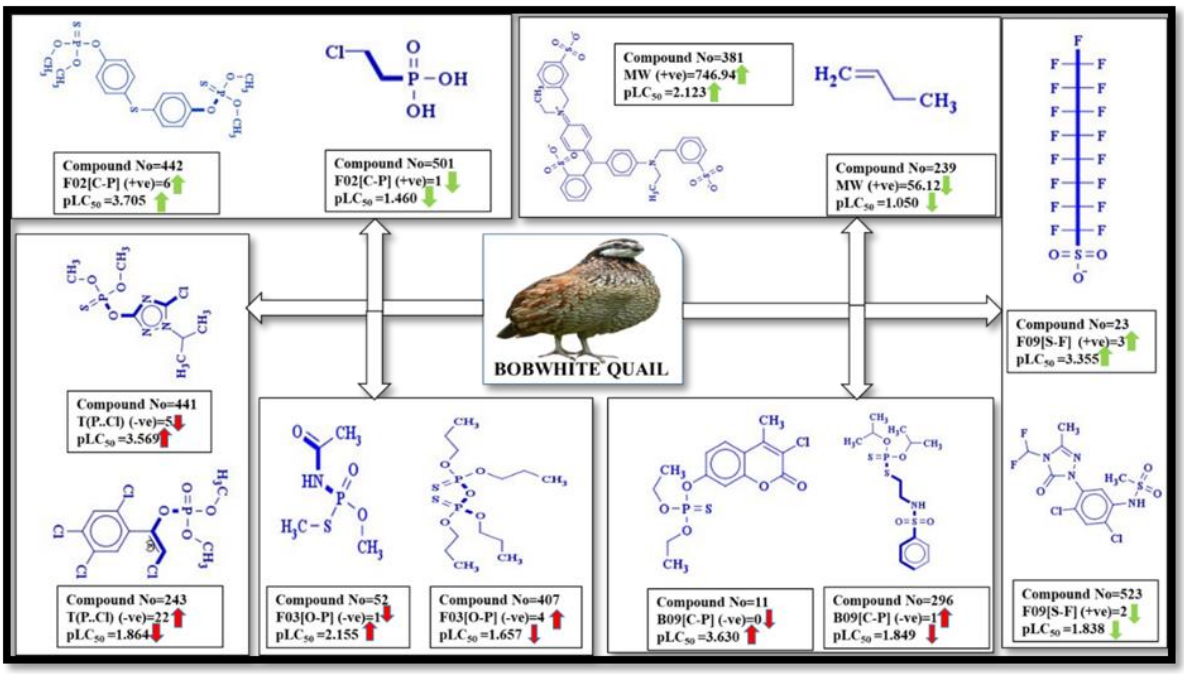


Figure 4.21. Positive and negative contribution of model descriptors towards .

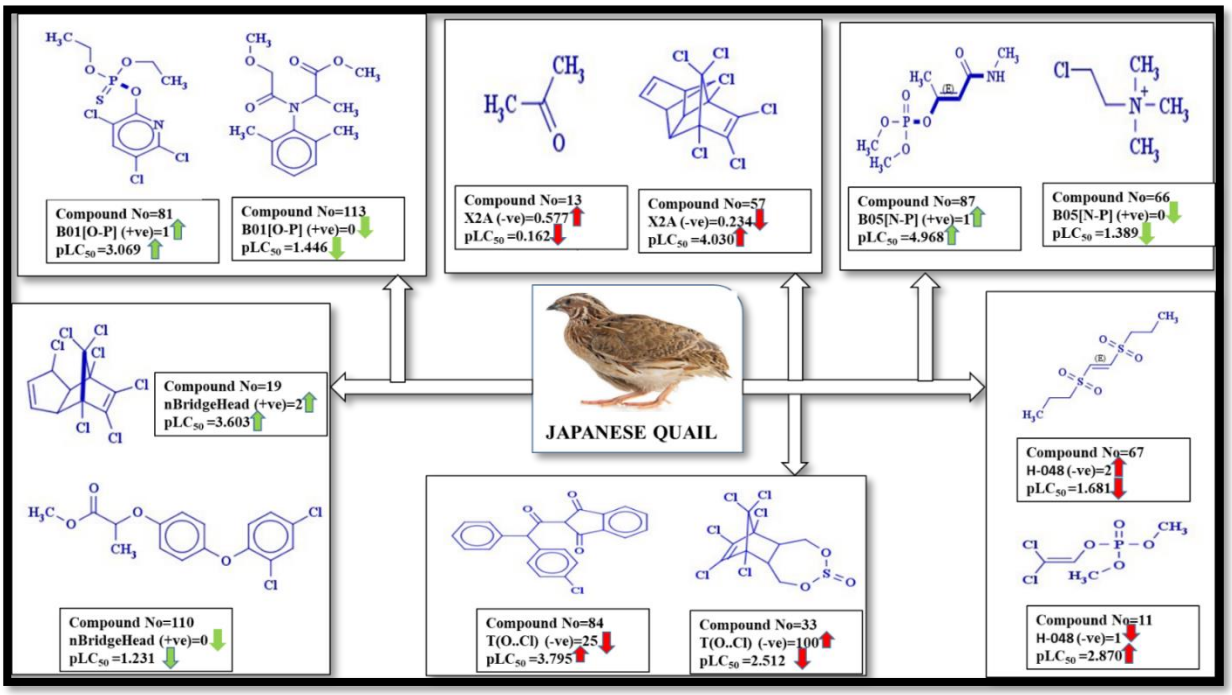


Figure 4.22. Positive and negative contribution of model descriptors towards JQ.

4.3.5. 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; ≥ 1.883; 2.236; 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 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. 4.17.** 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.

Table. 4.17. Top 20 and least 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB).

Sl. no.	Pesticide	Safety and Hazards	Sources
Top 20 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	Y Yu et al. [67]
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

Sl. no.	**Pesticide**	**Safety and Hazards**	**Sources**
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%20Chloroacetate%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

CHAPTER - 5

Conclusion

5. Conclusions

In the present work, we have utilized different 2D descriptors including both the ETA and non – ETA indices to develop our models. We have developed QSTR and q-RASTR models in our study to investigate the structural characteristics that cause acute oral toxicity in multiple avian species and interpret the descriptors mechanistically to determine how these structural characteristics influence acute oral toxicity in birds. The ecotoxicity of pesticides was regulated by various physicochemical and chemical properties such as lipophilicity, electronegativity, polarity, steric hindrance, and branching. The model developed in our study was rigorously validated by using both internal (using different internal validation metrics) and external (using different external validation metrics) validation strategies.

5.1. Study 1

This work reports the first PLS q-RASTR model for acute toxicity in chicken, the widely consumed source of animal protein. The study's importance lies in the direct link between chemical toxicity in chicken, human intake, and environmental damage. In this study, we can be concluded that the present research is significant and novel because of the following reasons:

I. By utilizing mathematical models, we got a comprehensive knowledge of how certain chemicals impact chicken species on a toxicological level. This knowledge is crucial in developing effective measures to protect the health of chicken species as well as human beings.

II. From this study, it was found that lipophilicity and electronegativity are responsible for the toxicity of pesticides towards chickens. On the other hand, polarity, hydrophilicity, and large numerical value of SE (LK) & SD similarity (GK) descriptors will reduce the toxicity of pesticides towards chickens.

III. The ability of models to identify specific features contributing to chicken toxicity will aid in creating safer, environmentally friendly chemicals.

IV. The developed q-RASTR model is robust and practical for toxicity & risk assessment.

V. The closeness of the acute toxicity prediction by the q-RASTR model with real-world data demonstrates its feasibility for screening acute toxicants in chickens.

VI. Models can be used for data-gap filling as well as predicting the toxicity of chemicals even before their synthesis.

5.2. Study 2

The current work demonstrates the suitability of amalgamation of RA and QSTR i.e. q-RASTR based model for efficient and reliable ecotoxicological risk assessment of diverse pesticides in avian species. The robustness, predictive ability, and reproducibility of the model were

meticulously evaluated by globally accepted internal and external validation metrics. As a critical step in ensuring the real-world applicability, the PLS-based q-RASTR model was deployed for reliable prediction of HD_5 values of the pesticides from the Pesticide Properties Database (PPDB), within the applicability domain. The high accuracy of the obtained predictions in comparison to the experimental toxicity data, demonstrated the true predictive capability of the q-RASTR model. Although LD_{50} is crucial for general comparisons, HD_5 provides a more cautious and safety-oriented approach, making it valuable for risk assessment and decision-making in developing effective measures to safeguard the health of avian species. Through the use of mathematical models, we have gained a comprehensive understanding of how certain chemicals affect avian species on a toxicological level. We found that the presence of high coefficient and variation of the similarity values of the close source compounds, product of the gm, and Avg.Sim levels, number of Pyrrolidines, and increases in branching influence the toxicity towards avian species. Conversely, the high distribution among the close source compounds, Presence/absence of C – C bonds at topological distance 7, and degree of saturation decrease the toxicity toward avian species. This approach offers a cost-effective and ethical alternative to traditional *in vivo* testing, aiding regulatory bodies, researchers, and industries in assessing the potential ecological risks associated with pesticide use.

5.3. Study 3

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.

This thesis presents a comprehensive investigation into the acute toxicity of pesticides in avian species, utilizing a variety of 2D descriptors, including ETA and non-ETA indices, to develop QSTR and q-RASTR models. These models enable a detailed understanding of the structural characteristics that influence toxicity, providing significant insights into the ecotoxicity of pesticides regulated by properties such as lipophilicity, electronegativity, polarity, steric hindrance, and branching. This thesis advances the field of ecotoxicology by providing novel, validated models that offer accurate predictions of pesticide toxicity in avian species. These models not only enhance our understanding of toxicological mechanisms but also contribute to the development of safer pesticides and more ethical research practices. The integration of chemometric tools and rigorous validation strategies ensures the reliability and applicability of these models in real-world scenarios, ultimately supporting the goal of protecting both avian species and human health from the adverse effects of pesticide exposure.

References

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. Zhan, H., Huang, Y., Lin, Z., Bhatt, P., and Chen, S. (2020). New insights into the microbial degradation and catalytic mechanism of synthetic pyrethroids. Environ. Res. 182:109138. doi: 10.1016/j.envres.2020.109138

4. Bhatt, P., Joshi, T., Bhatt, K., Zhang, W., Huang, Y., and Chen, S. (2021a). Binding interaction of glyphosate with glyphosate oxidoreductase and C–P lyase: Molecular docking and molecular dynamics simulation studies. J. Hazar. Mat. 409:124927. doi: 10.1016/j.jhazmat.2020.124927

5. Mieldazys, A., Mieldazys, R., Vilkevicius, G., and Stulginskis, A. (2015). Agriculture-Use of Pesticides/Plant Protection Products. Bilbao: EU-OSHA.

6. Gyawali, K., 2018. Pesticide uses and its effects on public health and environment. *Journal of Health Promotion*, *6*, pp.28-36.

7. Unsworth, J., 2010. History of pesticide use. International Union of pure and applied chemistry (IUPAC).

8. Fernández, L., 2021. Global pesticide use by country| Statista.

9. Khan, M.J., Zia, M.S. and Qasim, M., 2010. Use of pesticides and their role in environmental pollution. *World Acad Sci Eng Technol*, *72*, pp.122-128.

10. Warren, G.F., 1998. Spectacular increases in crop yields in the United States in the twentieth century. *Weed Technology*, *12*(4), pp.752-760.

11. Kafilzadeh, F., Ebrahimnezhad, M., and Tahery, Y. (2015). Isolation and identification of endosulfan-degrading bacteria and evaluation of their bioremediation in Kor River, Iran. Osong Public Health Res. Perspect. 6, 39–46. doi: 10.1016/j.phrp.2014.12.003

12. Marrs, T.C. and Ballantyne, B., 2004. Pesticide Toxicology and International Regulation.

13. Eldridge, B.F., 2008. Pesticide application and safety training for applicators of public health pesticides. *California Department of Public Health, Vector-Borne Disease Section*, *1616*.

14. Drum, C., 1980. Soil chemistry of pesticides, PPG industries. *Inc. USA*.

15. Buchel, K.H. ed., 1983. *Chemistry of pesticides* (pp. xii+-518).

16. Gupta, P.K., 2006. WHO/FAO Guidelines for cholinesterase-inhibiting pesticide residues in food. In *Toxicology of organophosphate & carbamate compounds* (pp. 643-654). Academic Press.

17. Sacramento, C.A., 2008. Department of pesticide regulation "What are the potential health effects of pesticides?". Community guide to recognizing and reporting pesticide problems, 352, pp.27-29.

18. Lorenz, E.S., 2009. Potential health effects of pesticides. Pesticide Safety Fact Sheets, Pennsylvania State University, College of Agricultural Sciences, Agricultural Research and Cooperative Extension, Pesticide Education Program.

19. Harrison, S.A., 1990. The fate of pesticides in the environment. *Agrochemical Fact Sheet*, *8*, pp.2-8.

20. Khan, B.A., Nadeem, M.A., Nawaz, H., Amin, M.M., Abbasi, G.H., Nadeem, M., Ali, M., Ameen, M., Javaid, M.M., Maqbool, R. and Ikram, M., 2023. Pesticides: impacts on agriculture productivity, environment, and management strategies. In *Emerging contaminants and plants: Interactions, adaptations and remediation technologies* (pp. 109-134). Cham: Springer International Publishing.

21. 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.

22. Sood, P., 2023. Pesticides Usage and Its Toxic Effects–A Review. *Indian Journal of Entomology*.

23. 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

24. Kovarich, S., Ceriani, L., Fuart Gatnik, M., Bassan, A. and Pavan, M., 2019. Filling data gaps by read-across: A mini review on its application, developments and challenges. Molecular Informatics, 38(8-9), p.1800121.

25. Luechtefeld, T., Marsh, D., Rowlands, C. and Hartung, T., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicological Sciences, 165(1), pp.198-212.

26. Chirico, N. and Gramatica, P., 2011. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *Journal of chemical information and modeling*, *51*(9), pp.2320-2335.

27. 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.

28. Mei, H., Zhou, Y., Liang, G. and Li, Z., 2005. Support vector machine applied in QSAR modelling. *Chinese Science Bulletin*, *50*, pp.2291-2296.

29. Wu, Z., Zhu, M., Kang, Y., Leung, E.L.H., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D. and Hou, T., 2021. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings in bioinformatics*, *22*(4), p.bbaa321.

30. Doddareddy, M.R., Lee, Y.J., Cho, Y.S., Choi, K.I., Koh, H.Y. and Pae, A.N., 2004. Hologram quantitative structure activity relationship studies on 5-HT6 antagonists. *Bioorganic & medicinal chemistry*, *12*(14), pp.3815-3824.

31. Luthy, R.G., Sedlak, D.L., Plumlee, M.H., Austin, D. and Resh, V.H., 2015. Wastewater-effluent-dominated streams as ecosystem-management tools in a drier climate. *Frontiers in Ecology and the Environment*, *13*(9), pp.477-485.

32. Guha, R. and Willighagen, E., 2012. A survey of quantitative descriptions of molecular structure. *Current topics in medicinal chemistry*, *12*(18), pp.1946-1956.

33. Todeschini, R. and Consonni, V., 2008. *Handbook of molecular descriptors*. John Wiley & Sons.

34. Livingstone, D.J., 2000. The characterization of chemical structures using molecular properties. A survey. *Journal of chemical information and computer sciences*, *40*(2), pp.195-209.

35. Snedecor GW, Cochran WG (1967) Statistical methods. Oxford and IBH, New Delhi.

36. Agresti A (1996) An introduction to categorical data analysis. Wiley, Hoboken.

37. Kar, S., Gajewicz, A., Puzyn, T. and Roy, K., 2014. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicology in Vitro*, *28*(4), pp.600-606.

38. Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Arnold, London.

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

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

41. 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.

42. 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.

43. Wu, J., Mei, J., Wen, S., Liao, S., Chen, J. and Shen, Y., 2010. A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. *Journal of computational chemistry*, *31*(10), pp.1956-1968.

44. 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.

45. Golbraikh, A. and Tropsha, A., 2002. Beware of q2!. *Journal of molecular graphics and modelling*, *20*(4), pp.269-276.

46. 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.

47. Schürmann, G., Ebert, R.U., Chen, J., Wang, B. and Kühne, R., 2008. External validation and prediction employing the predictive squared correlation coefficients test set activity mean vs training set activity mean. *J. Chem. Inf. Model*, *48*, pp.2140-2145.

48. Consonni, V., Ballabio, D. and Todeschini, R., 2009. Comments on the definition of the Q 2 parameter for QSAR validation. *Journal of chemical information and modeling*, *49*(7), pp.1669-1678.

49. Chirico N, Gramatica P (2011) Real External predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51:2320–2335.

50. 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.

51. 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*. DOI: 10.1039/D4EM00059E

52. Gordon, K., 2001. The OECD guidelines and other corporate responsibility instruments: a comparison.

53. Rosenberg, K.V., Dokter, A.M., Blancher, P.J., Sauer, J.R., Smith, A.C., Smith, P.A., Stanton, J.C., Panjabi, A., Helft, L., Parr, M. and Marra, P.P., 2019. Decline of the North American avifauna. *Science*, *366*(6461), pp.120-124.

54. Khan, B.A., Nadeem, M.A., Nawaz, H., Amin, M.M., Abbasi, G.H., Nadeem, M., Ali, M., Ameen, M., Javaid, M.M., Maqbool, R. and Ikram, M., 2023. Pesticides: impacts on agriculture productivity, environment, and management strategies. In *Emerging contaminants and plants: Interactions, adaptations and remediation technologies* (pp. 109-134). Cham: Springer International Publishing.

55. 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.

56. Sood, P., 2023. Pesticides Usage and Its Toxic Effects–A Review. *Indian Journal of Entomology*.

57. Yadav, N., Garg, V.K., Chhillar, A.K. and Rana, J.S., 2023. Recent advances in nanotechnology for the improvement of conventional agricultural systems: a review. *Plant Nano Biology*, p.100032.

58. Raj, A., Dubey, A., Malla, M.A. and Kumar, A., 2023. Pesticide pestilence: global scenario and recent advances in detection and degradation methods. *Journal of Environmental Management*, *338*, p.117680.

59. Das, S., Samal, A. and Ojha, P.K., 2024. Chemometrics-driven prediction and prioritization of diverse pesticides on chickens for addressing hazardous effects on public health. *Journal of Hazardous Materials*, p.134326.

60. 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.

61. Lane, T.R., Harris, J., Urbina, F. and Ekins, S., 2023. Comparing LD50/LC50 machine learning models for multiple species. *ACS Chemical Health & Safety*, *30*(2), pp.83-97.

62. Mineau, P., 1991. Difficulties in the regulatory assessment of cholinesterase-inhibiting insecticides.

63. Mineau, P., Collins, B.T. and Baril, A., 1996. On the use of scaling factors to improve interspecies extrapolation of acute toxicity in birds. *Regulatory Toxicology and Pharmacology*, *24*(1), pp.24-29.

64. 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/displaydocumentpdf/?cote=env/jm/ mono(2007)2&doclanguage=en (accessed September 15, 2014).

65. Karpov, P., Godin, G. and Tetko, I.V., 2020. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of cheminformatics*, *12*, pp.1-12. https://doi.org/10.1186/s13321-020-00423-w.

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

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

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

69. 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. https://doi.org/10.1016/j.chemolab.2015.07.007.

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

71. 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, 459, p.132129. https://doi.org/10.1016/j.jhazmat.2023.132129.

72. Luttik, R., Mineau, P. and Roelofs, W., 2005. A review of interspecies toxicity extrapolation in birds and mammals and a proposal for long-term toxicity data. *Ecotoxicology*, *14*, pp.817-832.

73. Luttik, R., Mineau, P. and Roelofs, W., 2005. A review of interspecies toxicity extrapolation in birds and mammals and a proposal for long-term toxicity data. *Ecotoxicology*, *14*, pp.817-832.

74. 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*.

75. 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. https://doi.org/10.1021/acs.jcim.6b00277.

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

77. 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. https://doi.org/10.1007/s11356-022-21937-w.

78. 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

79. Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

80. 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. https://doi.org/10.1016/0005-2795(75)90109-9.

81. Katritzky, A.R., Tatham, D.B. and Maran, U., 2001. Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. Journal of chemical information and computer sciences, 41(5), pp.1162-1176. https://doi.org/10.1021/ci010011r.

82. Golbraikh, A. and Tropsha, A., 2002. Beware of q2!. Journal of molecular graphics and modelling, 20(4), pp.269-276. https://doi.org/10.1016/S1093-3263(01)00123-1.

83. 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. https://doi.org/10.1016/j.chemosphere.2019.04.204.

Appendix

List of publications and reprints

List of Publications

A. Papers related to this dissertation:

1. **Shubha Das**, Abhishek 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,** 2024, 471, p.134326. **(Impact Factor-12.2)**

DOI: https://doi.org/10.1016/j.jhazmat.2024.134326

2. **Shubha Das**, Abhishek 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,** 2024, 188, pp.39-52**. (Impact Factor- 6.9).**

DOI: https://doi.org/10.1016/j.psep.2024.05.095

3. **Shubha Das**, Arnab Bhattacharjee, and Probir Kumar Ojha, "First report on q-RASTR modeling of hazardous dose (HD_5) for acute toxicity of pesticides: An efficient and reliable approach towards safeguarding the sensitive avian species" Environmental Science & Technology (Under Review) (**Impact Factor: 10.8**).

B. Papers not related to this dissertation:

1. Abhishek Samal, Shubha Das, and Probir Kumar Ojha, "*First report on Intelligent Consensus Prediction addressing Ecotoxicological effects of diverse pesticides against California quail. Journal: Chemosphere.*

2. Prodipta Bhattacharyya, Pabitra Samanta, Ankur Kumar, Shubha Das, and Probir Ojha *"Quantitative read-across structure-property relationship (q-RASPR): A novel approach to estimate the bioaccumulative potential for diverse classes of industrial chemicals in aquatic organisms"*. Journal: Environmental Science: Processes & Impacts.

Journal of Hazardous Materials 471 (2024) 134326

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Chemometrics-driven prediction and prioritization of diverse pesticides on chickens for addressing hazardous effects on public health

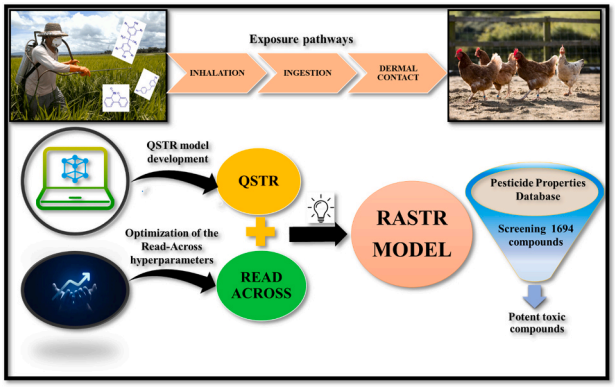
Shubha Das, Abhisek Samal, Probir Kumar Ojha [*,1]

Drug Discovery and Development Laboratory (DDD Lab), Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

HIGHLIGHTS

- This work reports the first PLS q-RASTR model for acute toxicity in chicken, the widely consumed source of animal protein.
- The developed q-RASTR model is robust and practical for toxicity & risk assessment.
- The models identify the essential features of chemicals associated with toxicity against chicken.
- The compliance between the predicted acute toxicity by the PLS q-RASTR model with real-world data demonstrates its feasibility for screening acute toxicants in chickens.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Pesticides
Acute toxicity
RASTR
Machine Learning
PPDB

ABSTRACT

The extensive use of various pesticides in the agriculture field badly affects both chickens and humans, primarily through residues in food products and environmental exposure. This study offers the first quantitative structure-toxicity relationship (QSTR) and quantitative read-across-structure toxicity relationship (q-RASTR) models encompassing the LOEL and NOEL endpoints for acute toxicity in chicken, a widely consumed protein. The study's significance lies in the direct link between chemical toxicity in chicken, human intake, and environmental damage. Both the QSTR and the similarity-based read-across algorithms are applied concurrently to improve the predictability of the models. The q-RASTR models were generated by combining read-across derived similarity and error-based parameters, alongside structural and physicochemical descriptors. Machine Learning approaches (SVM and RR) were also employed with the optimization of relevant hyperparameters based on the cross-validation approach, and the final test set prediction results were compared. The PLS-based q-RASTR models for NOEL and LOEL endpoints showed good statistical performance, as traced from the external validation metrics Q^2_{F1}: 0.762–0.844; Q^2_{F2}: 0.759–0.831 and MAE_{test}: 0.195–0.214. The developed models were further used to screen the Pesticide Properties DataBase (PPDB) for potential toxicants in chickens. Thus, established models can address eco-toxicological data gaps and development of novel and safe eco-friendly pesticides.

* Corresponding author.
E-mail addresses: probirojha@yahoo.co.in, pkojha.pharmacy@jadavpuruniversity.in (P.K. Ojha).
[1] ORCID id: 0000–0003-4796–3915

https://doi.org/10.1016/j.jhazmat.2024.134326
Received 30 January 2024; Received in revised form 9 April 2024; Accepted 15 April 2024
Available online 16 April 2024

1. Introduction

The most commonly consumed meat in the world is broiler chicken [1]. To fulfill the demand for meat, different types of bird diets (especially supplements) as well as other medicines are used for fast and healthy growth of chickens. These food supplements and medicines contain diverse types of pesticides and other chemicals. Pesticides are substances that are used to control or eliminate pests, such as insects, weeds, and fungi, in agriculture. While they can be effective in protecting crops, have the potential to impact both chickens and humans as well, primarily through residues in food products and environmental exposure [2,3]. One of the main concerns for humans is the presence of pesticide residues in food. If chickens consume feed containing pesticides, residues can be transferred to eggs and meat. Humans can then ingest these residues when consuming poultry products. Thus, consuming the meats of these chickens will affect the health of human beings too. There have been several concerns raised about the impact of pesticides on birds as well as on human beings. Such concerns arise due to the possible negative unintended impacts of pesticides on a variety of birds or the direct injurious effects of pesticides on human health [4]. Regulatory bodies have, therefore, underscored the need to carry out toxicity testing on current and new chemical pesticides to assess their impact on the environment [5]. Exposure to pesticides is severe and dangerous and can lead to death. While there are established techniques for evaluating avian toxicity through both *in vivo* and *in vitro* approaches, they are costly, time-consuming, and immoral [5]. To investigate the inherent properties of chemicals concerning toxicological prediction, governing bodies such as the Environmental Protection Agency (EPA), Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), European Chemicals Bureau (ECB), and European Food Safety Authority (EFSA) advise using computational tools such as read-across and QSAR [6]. Among the various *in-silico* techniques, QSAR is widely employed to predict the toxicity of test chemicals. By using this technique, a scientific model is developed from a compound series having experimentally derived endpoint values. Due to the reproducibility, simplicity, and transferability of the model, this technique is used widely. Current chemical risk assessment relies on similarity-driven methods like Read-Across, avoiding the need for mathematical models [7]. This approach assumes that compounds with similar structures have comparable biological activities, making emerging similarity-driven systems more suitable for consistent compound prediction. Often, Read-Across predicts probe compounds more reliably than QSAR models; however, one of the main limitations of Read-Across is that it lacks the ability to interpret essential features [8]. To overcome this problem, a novel approach, Read-Across Structure-Activity Relationship (RASAR), was introduced to combine the benefits of QSAR and Read-Across algorithms, which often results in better predictive ability and reduced mean absolute error (MAE) [9]. They utilized classification-based models that produced predictions on a graded scale. Banerjee and Roy [10] introduced q-RASAR modeling with descriptors based on similarity and error measures. The q-RASAR methodology utilizes similarity and error-based measures to produce simple, convenient, interpretable, and reproducible models with better predictivity. q-RASAR models can be developed using a variety of statistical techniques like MLR, PLS, etc. apart from sophisticated machine learning (ML) techniques. Machine learning is a growing technology that uses various algorithms for building models and making predictions using data. Support vector machines (SVM), artificial neural networks (ANN), and others are commonly used machine learning algorithms for numerous experimental studies [11,12]. There are various journals [13–15] present related to the *in-silico* prediction of acute toxicity of different species but concerning chicken, there are no *in-silico* reports available to date.

In this work, we investigated the toxicity of several pesticides on chickens and developed a logical and trustworthy method for assessing ecotoxicological risk. Based on the OECD rules, we have developed q-RASTR models to predict pesticide ecotoxicity on bird species. RASTR combines the read-across and QSTR approaches to improve predictability. The pLOEL and pNOEL (the negative logarithm of Lowest Observed Effect Level and No Observed Effect Level values respectively) values have been used as endpoints in this study. NOEL is defined as the highest dose of the toxicant that does not cause any toxicity or harm and LOEL stands for the lowest concentration of a substance that can cause an effect under specific exposure conditions. To successfully create the models, we used PLS for the initial model development. Further, RASAR descriptors were estimated using the optimal hyperparameters and incorporated to improve the external predictivity of the model. Additionally, support vector machine and Ridge regression machine learning (ML) approaches were employed with the optimization of hyperparameters using cross-validation. The final test set predictions were then compared. After evaluating the test set predictions and interpretability, we have selected the PLS-based q-RASTR model as the final model. Using, globally accepted parameters, the robustness, reproducibility, and predictivity of the PLS-based q-RASTR models were thoroughly validated. It can be confidently affirmed that the models are reliable and accurate. The developed model was utilized to screen the Pesticide Properties Database (PPDB) to identify potential avian toxicants and promote the use of safer chemicals. The true predictive ability of the q-RASTR model was established by revalidating the real-world toxicity profiles of the most and least toxic screened compounds from the Pesticide Properties Database (PPDB).

2. Methods and materials

2.1. Collection and curation of toxicity data of diverse pesticides

The required toxicity data of diverse pesticides against chicken *(Gallus gallus)* were retrieved from the ECOTOX repository (https://cfpub.epa.gov/ecotox/). The collected experimental toxicity data was expressed as LOEL and NOEL in micromolar (μM) concentration, which were transformed into molar concentrations and then their negative logarithmic equivalents (pLOEL and pNOEL) to reduce the data range. After excluding any outlier value(s), all available values for a particular chemical were averaged to generate a single value. We only included values that were numerically close to each other when calculating the average. After curating the primary data, we selected 43 pLOEL and 56 pNOEL compounds for modeling.

2.2. Descriptor calculation

A single .sdf file of all the compounds was compiled which is essential to Alvadesc software for descriptor calculation. AlvaDesc software [16] was used to evaluate 2400 descriptors based on structural and physicochemical parameters. We removed the unnecessary descriptors columns using DataPreTreatmentGUI 1.2 software [17].

2.3. Dataset division and descriptor selection

Division of dataset is a crucial component of statistical modeling, particularly in the context of QSARs. The modeling data is divided into two parts, the training set for model development and the test set to validate the developed model. In this present study, different dataset division techniques such as the clustering technique, Euclidean-distance-based method, Kennard-stone-based method, activity property-sorted, and random-division methods were employed for dataset division into training and test sets. Among these techniques, the best result was obtained from the Kennard stone division method in case of the pLOEL endpoint and random selection in case of the pNOEL endpoint [17,18]. The training/test sets compounds for pLOEL endpoint and pNOEL endpoint are 30/13 and 44/12 respectively. The divided training and test sets were also pre-treated using the tool dataPreTreatmentTrainTest1.0 (available from https://teqip.jdvu.ac.in/QS

AR_Tools/). These final pre-treated training and test sets were used for further analysis. Preliminary multiple linear regression models were generated for two datasets using MINITAB software. After that, PLS (Partial Least Square) method was used to generate the final models for both datasets using the software PLS_Single Y_version 1.0 [17].

2.4. Read – Across and calculation of the RASTR descriptor

Optimizing hyperparameters (similarity-based algorithm; σ, γ, and number of close source compounds) is crucial for read-across prediction [19]. The descriptor involved in the QSTR model was used to create sub-train and sub-test sets from the training data. We have chosen a Gaussian kernel-driven similarity, with σ = 0.75; γ = 0.75, and 9 close training compounds for pLOEL data points & Laplacian kernel-based similarity, with σ = 0.25 and γ = 0.25, and 4 close training compounds for pNOEL data points. During optimization, the hyperparameters were selected based on MAE-based (95%) criteria and external metrics (Q^2_{F1} and Q^2_{F2}). To perform q-RASTR modeling, similarity, and error-based RASTR descriptors were calculated for both training and test compounds with "RASAR Descriptor Calculator v2.0 tool using the optimized hyperparameters [17,20].

2.5. q- RASTR feature selection and model development

A total of 15 descriptors (Table S1 **in** supplementary information 2) were computed based on three similarity-based approaches (Euclidean Distance-based, Gaussian Kernel similarity-based, and Laplacian Kernel similarity-based) and a given set of source compounds for the individual training set and the test set [21]. The calculated RASTR descriptors were integrated with the model descriptors and the combined pool was subjected to best subset selection using BestSubsetSelectionModified_v2.1 tool [17] for model development. The final PLS-based q-RASTR model was developed with the best features using the PLS_Single Y_version 1.0 software.

2.6. 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 Ridge Regression (RR) using the Orange data mining tool [22]. The hyperparameters were adjusted to tune the model for optimal performance. The prediction qualities of the ML models were evaluated in terms of Q^2_{F1}, Q^2_{F2}, and MAE_{test} values.

2.7. Statistical validation metrics and Y-randomization

Validation metrics are the key parameters for the recognition of any predictive model. For internal validation (for the training set), we evaluated the model using various internationally accepted internal validation metrics including the determination coefficient (R^2) and leave-one-out cross-validated Q^2 (Q^2_{Loo}). R^2 and Q^2_{Loo} are the measures of goodness-of-fit, and robustness, respectively. In machine learning (SVM, RR) approaches, the root means squared error of calibration (RMSEC) metric was also calculated by the Orange data mining tool. A lower RMSEc indicates a better model fit, showing that the model's predictions are, on average, closer to the true values. For external validation (for the test set), we calculated various globally accepted external validation metrics such as R^2_{Pred} or Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, MAE-based criteria, $\overline{r^2_m}$, Δr^2_m and concordance correlation coefficient (CCC) [21]. External correlation coefficients such as Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} are well-known prediction indicators. In usual practice, the optimal value of these three measures (R^2_{Pred} or Q^2_{F1}, Q^2_{F2}, Q^2_{F3}) for model selection should be more than 0.5 [21,22]. Error measures such as mean absolute error (MAE_{test}) are frequently used to assess the accuracy of projected outputs, and they should be low for a strong model. The CCC measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. The concordance correlation coefficient (CCC) is an external validation measure proposed by Gramatica et.al. [9]. We have calculated the external coefficients (R^2_{Pred} or Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and CCC) using "PLS_Single Y" v1.0 software (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). External validation is undertaken to ensure the predictability of the created model, and only the test set chemicals are employed for this purpose. Aside from traditional measures, r^2_m metrics ($\overline{r^2_{m(test)}}$, $\Delta r^2_{m(test)}$) are calculated for external validation. When the $\overline{r^2_{m(test)}}$, values are quite good, the Δr^2_m values may serve as an additional metric for judging the quality of predictions [18]. The acceptability of the model was also checked using an external validation parameter proposed by Golbraikh and Tropsha [23,24]. Based on Golbraikh and Tropsha criteria, the model will be acceptable if:

1. Q^2_{LOO} (train) > 0.5.
2. R^2(test) > 0.6.
3. $[(r^2\text{-}r^2_0)/\ r^2] < 0.1$ or $[(r^2\text{-}r'^2_0)/\ r^2] < 0.1$
4. $1.15 > k > 0.85$ or $1.15 > k' > 0.85$.

Y-randomization study was performed using "SIMCA-P" software to investigate the probability of chance occurrence in the final model. Herein, the response data are altered, without scrambling the descriptors, for a total of 100 times. After shuffling the original model is refitted to compute the R^2 and Q^2 values, and the intercept values of $R^2 < 0.3$ and $Q^2 < 0.05$ indicate no chance of correlation in a statistically significant model [24,25].

2.8. Screening of the Pesticide Properties DataBase (PPDB)

We have collected 1903 chemical data from the Pesticide Properties DataBase (PPDB) which is accessible through the PPDB website (http://sitem.herts.ac.uk/aeru/ppdb/). KNIME curation was carried out using a KNIME workflow to eliminate any duplicates, inorganic salts, and mixtures [26]. As a result of the KNIME curation process, some compounds have been eliminated. After curating the dataset, the enduring 1694 compounds were screened to verify model reliability. The descriptors of the molecules were calculated using the same procedure that was used in q-RASTR modeling as discussed earlier. The individual PLS-based q-RASTR models were used to make predictions, assisted by the PRI tool [17] which provided a reliable indication of the prediction's accuracy. The tool assesses the reliability of predictions using AD and furnishes qualitative prediction indicators categorized as 'Good', 'Moderate', and 'Bad. A detailed flow diagram of this study has been given in Fig. 1.

2.9. Software used

We have used different software's in this research work namely:

i."AlvaDesc" software (available from https://www.alvascience.com/alvadesc/) was used for descriptor calculation.

ii."Best Subset Selection Modified" v2.1 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for model development.

iii."Dataset Division GUI" v1.2 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to divide the dataset into training and test sets.

iv. "Minitab" v14 (available from: https://www.minitab.com/en-us/) was used for model development.

v. "PLS_Single Y" v1.0 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to develop the PLS-based QSTR and q-RASTR models.

vi. "Read-Across-v4.1″ (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for obtaining the optimized hyperparameters necessary for RASTR descriptor calculation.

vii."RASAR Descriptor Calculator" v2.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) was used

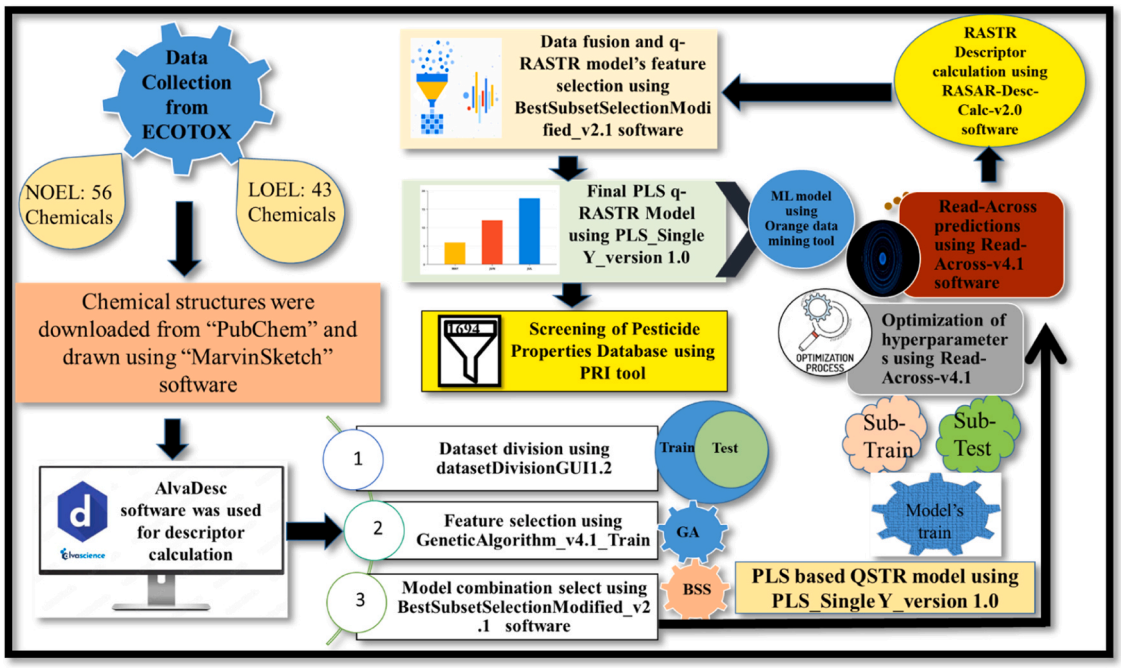


Fig. 1. Schematic workflow of q-RASTR model development.

for RASTR descriptors calculation.

viii. "Prediction Reliability Indicator" (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used to evaluate the localization in AD of the test compounds to ascertain the reliability of prediction of final PLS-based q-RASTR model.

ix. "SIMCA-P" (available from: https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca) was used for the randomization test.

3. Results and discussion

In this present study, we have developed QSTR and q-RASTR models for pLOEL and pNOEL endpoints using the PLS method and strictly obeying the OECD guidelines. We have additionally applied two different ML algorithms (SVM, RR) to check model performances.

3.1. PLS-based QSTR and q-RASTR models

The divided dataset is used to develop the QSTR and q-RASTR models for two endpoints (pLOEL and pNOEL) of chicken species. After the feature selection process, the PLS-based QSTR model was developed employing 3 and 5 descriptors with two and one latent variables for pLOEL (MODEL 1) and pNOEL (MODEL 2), respectively.

3.1.1. PLS-based QSTR model for pLOEL and pNOEL endpoints

Model 1 (pLOEL endpoint):

$$\boldsymbol{pLOEL} = 4.75827 + 0.50323 \times \mathrm{NsOH} - 0.191 \times MaxsCH3 - 0.64324 \times B01[C-O]$$

Model 2 (pNOEL endpoint):

$$\boldsymbol{pNOEL} = 5.08369 + 0.16353 \times H\_050 + 0.35253 \times NsssN - 0.62789 \times B05[C-O] + 0.80035 \times B05[O-O] - 0.8449 \times B08[C-P]$$

After the development of the QSTR models, similarity and error-based RASTR descriptors were calculated for both training and test sets compounds of pLOEL and pNOEL endpoints models using "RASAR Descriptor Calculator v2.0 tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) with the optimized hyperparameters. After that, we clubbed the RASTR descriptors and Alvadesc descriptors for the final q-RASTR model development [27]. Finally, PLS-based q-RASTR models were developed using 3 and 4 descriptors with one and two latent variables as shown in model 3 and model 4 respectively for pLOEL and pNOEL endpoint models,

3.1.2. PLS-based q-RASTR model for pLOEL and pNOEL endpoints

Model 3 (pLOEL endpoint):

$$\boldsymbol{pLOEL} = 5.1136 - 1.51275 \times SD \quad similarity(GK) + 0.41951 \times NsOH - 0.75444 \times B01[C-O]$$

Model 4 (pNOEL endpoint):

$$\boldsymbol{pNOEL} = 5.78412 - 2.04509 \times SE(LK) + 1.18371 \times B05[O-O] - 0.74259 \times B02[C-O] + 0.03736 \times T(N..S)$$

Each model has been rigorously validated following the OECD protocols. The computed internal and external validation metrics along with the optimum number of latent variables have been shown in the following Table 1**.** The PLS-based q-RASTR models **3** and **4** show strong fit and predictability with uniform scattering observed along the line, going through the origin of Cartesian coordinates **(**Fig. 2**).**

Here, we have seen that for both the datasets, the external validation metrics were significantly improved for the PLS-based q-RASTR models as compared to the PLS-based QSTR models, indicating the significance of the RASTR descriptors. We have also validated all the models (PLS-based QSTR and q-RASTR models for the pLOEL and pNOEL endpoints) using Golbraikh and Tropsha criteria and the results are given in Tables S2-S5 **(**Supplementary information 2**).** The results showed that the PLS-based q-RASTR models for both endpoints are acceptable based on the Golbraikh and Tropsha's criteria [24]. Hence, we have generalized that the PLS-based q-RASTR models are better as compared to the corresponding QSTR models.

3.2. Results of ML-based q-RASTR model

As previously stated, we used two different ML algorithms to evaluate their effectiveness in model construction and prediction. Based on

Table 1
QSTR and q-RASTR model's statistical quality.

Validation Metrics	QSTR model's statistical quality		PLS-based q-RASTR model's statistical quality	
Model name	Model 1 (pLOEL)	Model 2 (pNOEL)	Model 3 (pLOEL)	Model 4 (pNOEL)
No of LVs	2	1	1	2
R^2(train)	0.748	0.669	0.734	0.603
Q^2_{LOO} (train)	0.672	0.582	0.665	0.526
$Q^2_{F1\ (test)}$	0.608	0.643	**0.844**	**0.762**
$Q^2_{F2\ (test)}$	0.577	0.640	**0.831**	**0.759**
$Q^2_{F3\ (test)}$	0.692	0.790	**0.877**	**0.860**
MAE_{test}	0.309	0.225	**0.214**	**0.195**
CCC	0.818	0.730	0.909	0.845
$\overline{r^2_{m(test)}}$	0.637	0.415	0.740	0.560
Δr^2_m(test)	0.035	0.318	0.136	0.220
MAE-based prediction quality	MODERATE	GOOD	GOOD	GOOD

the internal validation, v-SVM was the best-performing model toward the pLOEL endpoint, and Ridge regression was the best-performing model towards the pNOEL endpoint based on internal and external validation metrics. In terms of external validation metric, Q^2_{F3} [28], the ability to efficiently predict the response values for the test set compounds, the best-performing models were the PLS-based q-RASTR models. Furthermore, the PLS-based q-RASTR models produce the lowest prediction error for the test set compounds, as indicated by the MAE_{test} value [29]. Thus, to assess the overall performance of the models for both endpoints, the PLS-based q-RASTR models are superior than QSTR models. The results of ML models are presented in Table 2.

3.3. Regression coefficient plot

The plot describes descriptor's positive/negative contribution towards the toxicity [30]. In this study, the descriptor NsOH contributed positively while the descriptors SD similarity (GK) and B01[C-O] contributed negatively towards the toxicity in case of **Model 3**. In case of **Model 4,** the descriptors B05[O-O], T(N.S) contributed positively while the descriptors SE(LK) and B02[C-O] contributed negatively towards the toxicity. All the relevant plots have been provided in Figs. S1-S2 in supplementary information 2.

3.4. Variable importance plot (VIP)

The respective descriptor contribution towards the model response is described by the variable importance plot, and the most and least important descriptors are recognized appropriately [31]. In this study, NsOH and B02[C-O] depicting electronegativity and hydrophilicity were identified as the most important descriptors for Model 3 and Model 4 respectively as shown in Figs. S3-S4 in supplementary information 2.

3.5. Loading plot

The plot describes the correlation between the X and Y variables [32], illustrating the effect of various model descriptors. The first two components were used to create the loading plot. A descriptor is supposed to have a stronger effect on response value if it is situated far from the origin of the plot and near the modeled endpoint. All the relevant plots have been provided in Figs. S5-S6 in supplementary information 2.

3.6. Applicability domain (AD)

AD is the hypothetical region in chemical space specified by the respective model descriptors and responses where predictions may be made with confidence [33]. To obtain a reliable prediction, the test

Table 2
ML-based q-RASTR model's statistical quality.

Validation Metrics	ML model's statistical quality			
Model name	SVM (pLOEL)	SVM (pNOEL)	RR (pLOEL)	RR (pNOEL)
$R^2_{Loo\ (train)}$	0.831	0.695	0.776	0.758
$Q^2_{LOO\ (train)}$	0.746	0.585	0.746	0.604
RMSEc (train)	0.245	0.245	0.283	0.218
$Q^2_{F1\ (test)}$	0.742	0.718	0.725	0.653
$Q^2_{F2\ (test)}$	0.721	0.715	0.703	0.650
$Q^2_{F3\ (test)}$	0.797	0.835	0.784	0.796
$MAE_{test\ (test)}$	0.273	0.169	0.300	0.216
CCC	0.893	0.856	0.850	0.804
$\overline{r^2_{m(test)}}$	0.725	0.659	0.626	0.541
Δr^2_m(test)	0.101	0.071	0.033	0.148
Optimum hyperparameters	v-SVM Regression cost-0.50 Complexity bound-0.65 Kernel- Linear	v-SVM Regression cost-2.50 Complexity bound-0.70 Kernel-Linear	Alpha-0.001	Alpha-0.001

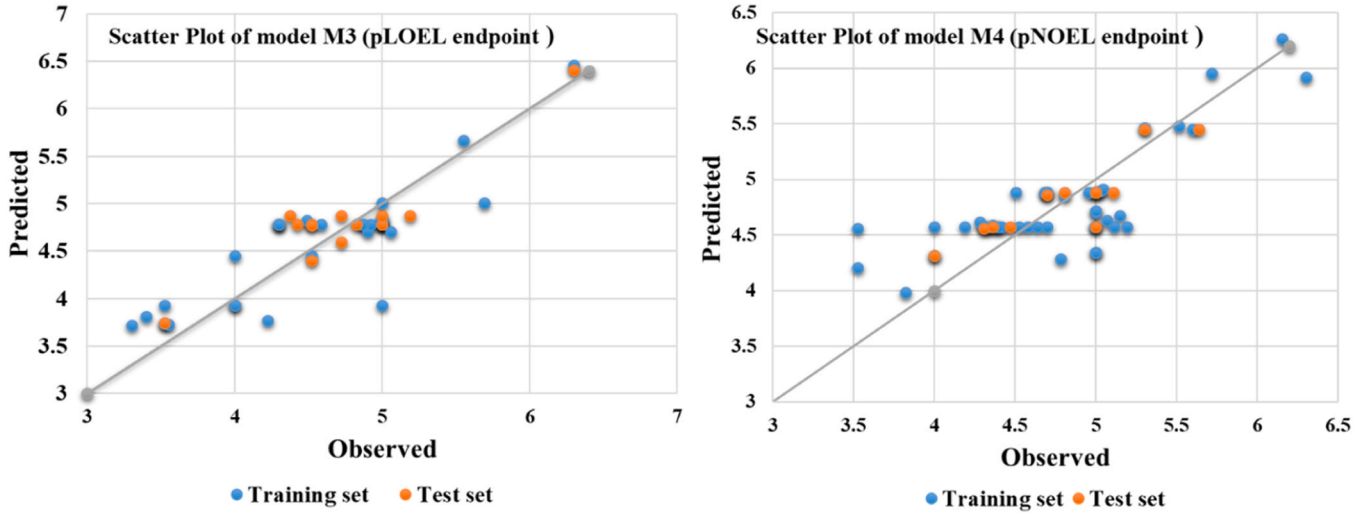


Fig. 2. Scatter plots of developed models.

compounds must have the highest structural similarity to the training compounds. As a result, validating the applicability domain is a fundamental prerequisite for every statistical model, as recommended by OECD principle 3 ("Validation of (Q)SAR Models - OECD," 2004). To comply with the OECD guidelines, an applicability domain analysis of the created PLS-based q-RASTR model was done with SIMCA-P software using the DModX technique at a 99% confidence level.

$$\mathrm{DModX} = \frac{\sqrt{\frac{SSEi}{K-A}}}{\sqrt{\frac{SSE}{(N-A-AO)(K-A)}}}$$

For observation i, in a model with A component, K variables, and N observations, SSE is the

squared sum of the residuals. A0 is 1 if the model was centered and 0 otherwise. It is claimed that DModX is approximately F-distributed, so it can be used to check if an observation deviates significantly from a normal PLS model. The DModX (distance to model in X-space) plots for both the training and test sets have been showcased in Figs. S7-S10 in supplementary information 2 (shows the AD plots of the Model 3 and Model 4). In this study, all the compounds from the training set (given in Fig. S7 in supplementary information 2) and test set (given in Fig. S8 in supplementary information 2) for the pLOEL endpoint model (model M3) are inside the applicability domain (below the D-Critical line) which indicates the reliability of predictions by the model. In the case of the pNOEL endpoint model (model M4), compounds 28 and 33 of the training set **(**given in Fig. S9 in supplementary information 2**)** are outside the applicability domain (above the D-critical line) due to the structural dissimilarity. All the compounds from the test set (given in Fig. S9 in supplementary information 2**)** of the pNOEL endpoint (model M4) are within the applicability domain.

3.7. Mechanistic interpretation

The details of the descriptors obtained from the M3 (pLOEL endpoint) and M4 models (pNOEL endpoint), their contribution, description, and probable mechanistic interpretation (according to OECD principle 5) are provided in Table 3.

3.7.1. Mechanistic interpretation of descriptors employed in Model M3 (pLOEL)

SD similarity (GK) is a RASTR descriptor that denotes the typical deviation of similarity levels among closely related compounds. It has a negative contribution to the toxicity endpoint. Higher standard deviation (SD) similarity shows that the distribution among the close source compounds is high thereby reducing prediction reliability as demonstrated in compound **30** and conversely shown in compound **3 (**depicted in Fig. 3).

The descriptor NsOH defines the number of atoms of type sOH in the compound and it contributes positively towards the toxicity endpoint. This fragment enhances the compound toxicity due to the presence of an electronegative atom (Oxygen) as demonstrated in compound **42** and the absence of this fragment decreases the toxicity as represented in compound **18** (shown in Fig. 3).

The descriptor B01[C-O] is a 2D atom pair descriptor that shows the occurrence of C-O at topological distance 1 and gives negative contribution towards the endpoint. The presence of polar bond [C-O] increases

Table 3
Mechanistic analysis of modeled descriptors.

S. NO	Descriptor	Type	Description	Contribution	Mechanistic introspection
CHICKEN - pLOEL					
1	SD similarity (GK)	RASTR	The typical deviation of similarity levels among closely related compounds	(-)ve	Higher standard deviation (SD) similarity shows that the distribution among the close source compounds is high thereby reducing prediction reliability as demonstrated in compound **30** and conversely shown in compound **3** (given in Fig. 3).
2	NsOH	Functional group counts	Number of atoms of type sOH	(+)ve	This fragment enhances the compound toxicity due to the presence of an electronegative atom (Oxygen) as demonstrated in compound **42** and in absence of this fragment decreases the toxicity as represented in compound **18** (given in Fig. 3).
3	B01[C-O]	2D Atom Pairs	Occurrence of C-O at topological separation of 1	(-)ve	In the case of B01[C-O] descriptor, the presence of polar bond [C-O] increases the hydrophilicity of the compound [34] and thus toxicity will decrease which is evidenced by compound **27** and vice versa in case of compound **36** (represented inFig. 3).
CHICKEN - pNOEL					
1	SE(LK)	RASTR	The weighted standard error pertains to the response values of adjacent source compounds.	(-)ve	The presence of this high standard error based on the response values of the proximate source compound decreases the compound toxicity as demonstrated in compound **8** and the less standard error based on response enhances the toxicity as represented in compound **40** (given in Fig. 4).
2	B05[O-O]	2D Atom Pairs	Occurrence of single bond oxygen-oxygen topological distance 5	(+)ve	The presence of two electronegative atoms increases the electronegativity rendering the compounds more electronegative[35]. The presence of large fragments in chemical structure will also increase the lipophilicity, ultimately enhancing the penetration ability of chemicals into the cell of reference organism. Thus existence of oxygen atoms at the specified topological distance is associated with increased toxicity in pesticides, as illustrated by compound **4**, while the opposite was characterized in compound **48** (provided in Fig. 4).
3	B02[C-O]	2D Atom Pairs	Occurrence of C-O at topological separation 2	(-)ve	This descriptor is related to hydrophilicity (oxygen is responsible for hydrogen bonding with water, and is easily excreted out from the body) [34, 35]. Small fragments (Occurrence of C-O at topological separation 2) are less lipophilic, as a result, toxicity will decrease which is evidenced by compound **30**, and the opposite was shown in compound **34** (represented in Fig. 4).
4	T(N.S)	2D Atom Pairs	Summation of topological separation between N.S	(+)ve	The occurrence of nitrogen and sulphur atoms in a compound increases its electronegativity, leading to oxidative stress and cell death [34]. Sulphur itself is toxic. Therefore, overall toxicity will increase as demonstrated in compound **33**. On the other hand, the compound containing less number of this fragment may exhibit less toxicity as shown in compound **53** (demonstrated in Fig. 4).

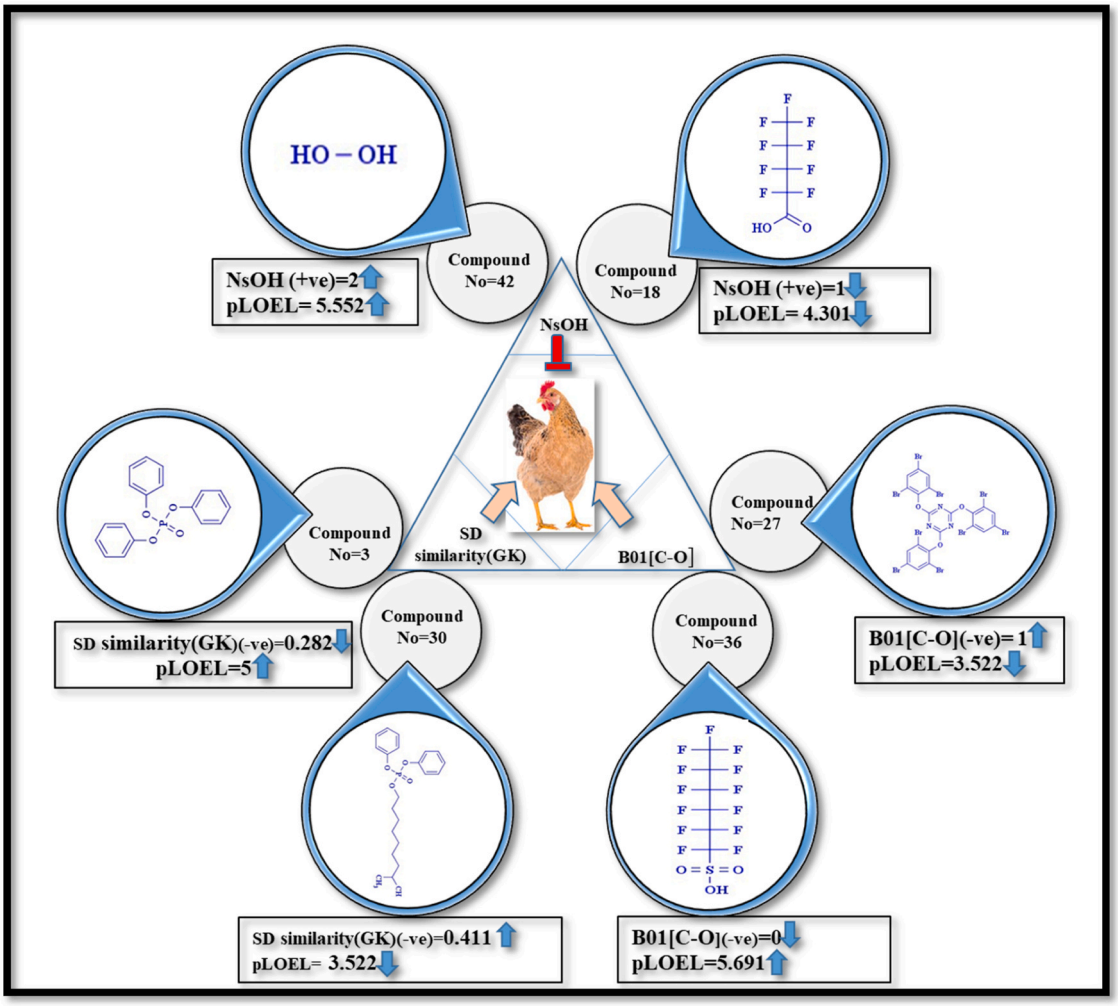


Fig. 3. Contribution of the model descriptors towards pLOEL in chicken.

the hydrophilicity of the compound [34] and thus toxicity will decrease which is evidenced by compound **27** and vice versa in case of compound **36** (represented in Fig. 3).

3.7.2. Mechanistic interpretation of descriptors employed in Model M4 (pNOEL)

The SE(LK) is a RASTR descriptor that denotes the weighted standard error about the response values of adjacent source compounds. It shows negative contribution toward the endpoint. The presence of this high standard error based on the response values of the proximate source compound decreases the compound toxicity as demonstrated in compound **8** and the less standard error based on response values of the proximate source compound enhances the toxicity as represented in compound **40** (given in Fig. 4).

The 2D atom pair descriptor, B05[O-O], shows the occurrence of two oxygen atoms at topological distance 5. The presence of two electronegative atoms increases the electronegativity rendering the compounds more electronegative [35]. The presence of large number of fragments in chemical structure will also increase the lipophilicity, ultimately enhancing the penetration ability of chemicals into the cell of the reference organism. Thus, the existence of oxygen atoms at the specified topological distance is associated with increased toxicity in pesticides as illustrated by compound **4**, while the opposite was characterized in compound **48** (provided in Fig. 4).

Another 2D atom pair descriptor, B02[C-O], indicates the occurrence of C-O at topological distance 2. It shows negative contribution toward the endpoint. This descriptor is related to hydrophilicity (oxygen is responsible for hydrogen bonding with water, and is easily excreted out from the body) [34]. Small fragments (occurrence of C-O at topological separation 2) are less lipophilic, as a result, toxicity will decrease which is evidenced by compound **30**, and the opposite was shown in compound **34** (represented in Fig. 4).

The T(N.S) descriptor denotes the summation of the topological distance between N.S and it contributed positively towards the endpoint. The occurrence of nitrogen and sulphur atoms in a compound increases its electronegativity, leading to oxidative stress and cell death [34,35]. Sulphur itself is toxic. Therefore, overall toxicity will increase as demonstrated in compound **33**. On the other hand, the compound containing less number of this fragment may exhibit less toxicity as shown in compound **53** (demonstrated in Fig. 4).

3.8. Pesticide Properties DataBase screening

The PPDB compounds were screened using developed models considering both the toxicity endpoints namely, pLOEL and pNOEL assisted by the Java-based tool “Prediction reliability indicator” (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). The applicability domain of the compounds was assessed to ascertain the reliability of the obtained prediction values and it was found that 100% and 55% of compounds lie within the chemical space of the developed pLOEL and pNOEL models respectively. The predicted pLOEL and pNOEL values (predicted value provided in Supplementary Information 1) of the

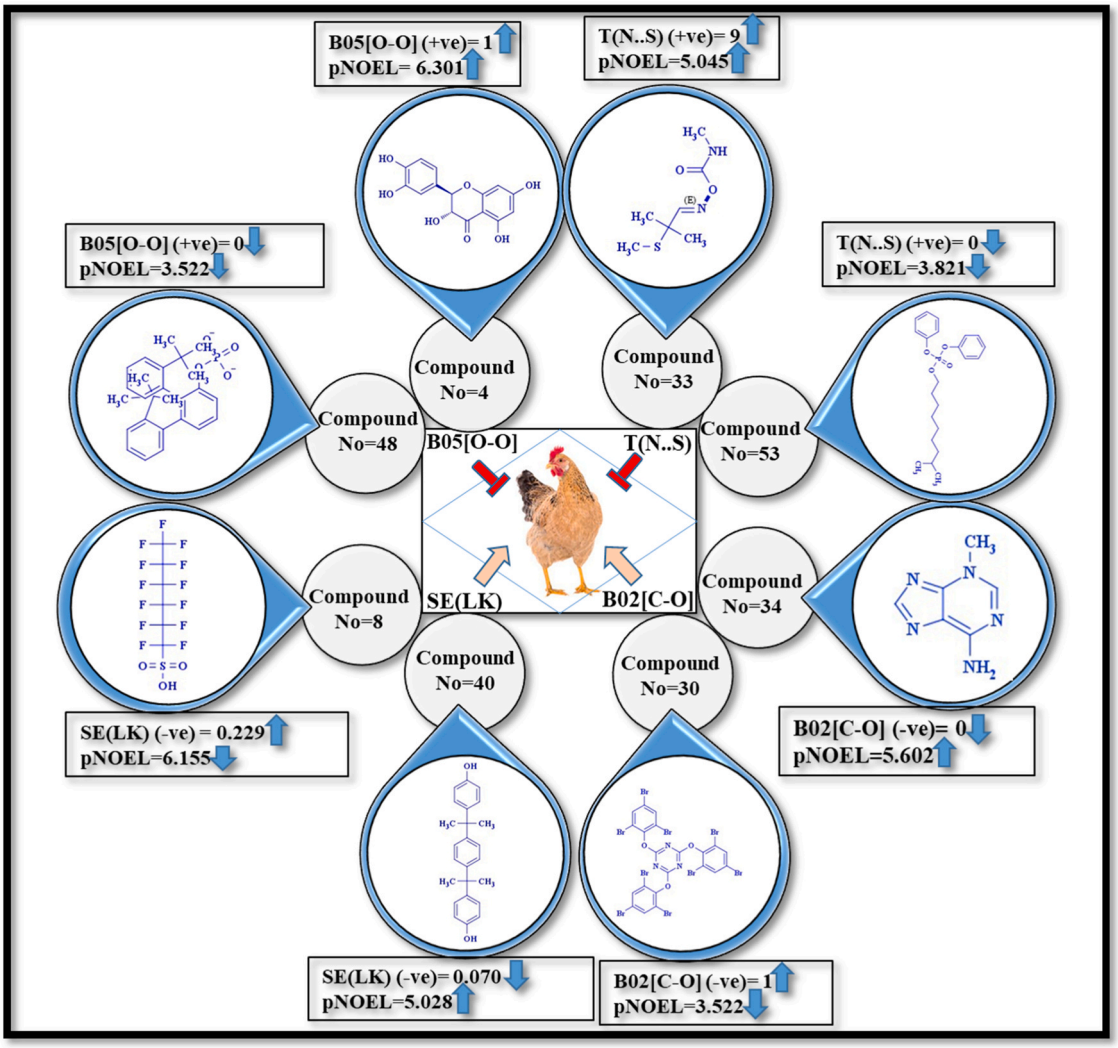


Fig. 4. Contributions of the model descriptors towards pNOEL in chicken.

respective compounds were cumulatively assessed. Then, based on the cumulative predictions, the top 20 and least 20 toxic compounds (compounds that are toxic for both pLOEL and pNOEL endpoints and lie within the AD of both models) with their CAS numbers, molecular weight, and pesticide groups have been provided in Table 4. Descriptor values of the top 20 and least 20 toxic pesticides are provided in supplementary information 1. Further, to validate our findings, an attempt was made to corroborate our predictions to the real-world experimental data available in the PubChem online repository, and literature and references of these findings are provided in Table S6 of Supplementary Information 2. Considering the top twenty highest toxic compounds, our models' pLOEL and pNOEL prediction values were in complete coherence with the experimental toxicity data. From the results, it can be stated that our model predictions are correlated to real-world data and can be considered suitable for the identification of potential toxicants alongside less ones. Upon further validation, all predicted toxicities, demonstrate the practical applicability of the developed models.

4. Conclusions

This work reports the first PLS-based q-RASTR model for acute toxicity in chicken, the widely consumed source of animal protein. The study's importance lies in the direct link between chemical toxicity in chicken, human intake, and environmental damage. In this study, we can be concluded that the present research is significant and novel because of the following reasons:

I. By utilizing mathematical models, we got a comprehensive knowledge of how certain chemicals impact chicken species on a toxicological level. This knowledge is crucial in developing effective measures to protect the health of chicken species as well as human beings.
II. From this study, it was found that lipophilicity and electronegativity are responsible for the toxicity of pesticides towards chickens. On the other hand, polarity, hydrophilicity, and large numerical value of SE (LK) & SD similarity (GK) descriptors will reduce the toxicity of pesticides towards chickens.
III. The ability of the models to identify specific features contributing to chicken toxicity will aid in creating safer, environmentally friendly chemicals.
IV. The developed q-RASTR models are robust and practical for toxicity & risk assessment.
V. The closeness of the acute toxicity prediction by the q-RASTR model with real-world data demonstrates its feasibility for screening acute toxicants in chickens.
VI. The models can be used for data-gap filling as well as predicting the toxicity of chemicals even before their synthesis.

Table 4
Twenty most and least toxic screened pesticides from the Pesticide Properties DataBase (PPDB).

Sl. No	Pesticide name (Group)	CAS no and Molecular mass	Safety and Hazards	Sources (all references available in Supplementary 2)
Top 20 toxic screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Flumetsulam	98967-40-9 (Molecular mass-325.29)	Toxic to rats, rabbits, quail, ducks, and Environmental hazard	**I**
2	Dipyrithione	3696-28-4 (Molecular mass-252.31)	Environmental hazard, irritant	**II**
3	Sulfoxaflor	946578-00-3 (Molecular mass-277.27)	Environmental hazard, irritant	**III**
4	Flusulfamide	106917-52-6 (Molecular mass-415.17)	Acute toxic to rats, mice, and Environmental hazard	**IV**
5	Benzofluor	68672-17-3 (Molecular mass-299.33)	Threshold of Toxicological Concern (Cramer Class- High (class III))	**V**
6	Nithiazine	58842-20-9 (Molecular mass-216.24)	Acute toxic to aves and irritants	**VI**
7	Perfluidone	37924-13-3 (Molecular mass-379.4)	Acute toxic to rats, rabbits, mice, and irritants	**VII**
8	Fluensulfone	318290-98-1 (Molecular mass-291.70)	Acute toxic to fish and environmental hazard	**VIII**
9	1,3-dinitrobenzene	99-65-0 (Molecular mass-168.12)	Acute toxic, Health hazard, and environmental hazard	**IX**
10	Ampropylfos	16606-64-7 (Molecular mass-139.09)	Corrosive	**X**
11	Azoxybenzene	495-48-7 (Molecular mass-198.22)	Acute toxic to rats, mice, and rabbits	**XI**
12	Benfluralin	1861-40-1 (Molecular mass-335.28)	Acute toxic to rats, mice, rabbits and environmental hazard	**XII**
13	Benzamorf	12068-08-5 (Molecular mass-413.6)	Corrosive and Irritant	**XIII**
14	Bis(methylmercury) sulphate	3810-81-9 (Molecular mass-527.31)	Threshold of Toxicological Concern (Cramer Class- High (class III))	**XIV**
15	Bis-trichloromethyl sulfone	3064-70-8 (Molecular mass-300.80)	Acute toxic to rats, mice, rabbits and environmental hazard	**XV**
16	Bromethalin	63333-35-7 (Molecular mass-577.9)	Acute toxic to rats, mice, dogs and environmental hazard	**XVI**
17	Butralin	33629-47-9 (Molecular mass-295.33)	Environmental hazard, Health hazard and Acute toxic to rats, rabbits	**XVII**
18	Cacodylic acid	75-60-5 (Molecular mass-138.00)	Acute toxic to rats, mice and environmental hazard	**XVIII**
19	Chloropicrin	76-06-2 (Molecular mass-164.37)	Acute toxic to humans, rats and mice	**XIX**
20	Dicloran	99-30-9 (Molecular mass-207.01)	Environmental hazard, Health hazard and acute toxic to rat, mice	**XX**
Least 20 screened pesticides from Pesticide Properties DataBase (PPDB)				
1	Zarilamid	84527-51-5 (Molecular mass-238.67)	The predictive value for both endpoints indicates this pesticide is less toxic for both endpoints.	**XXI**
2	Xylylcarb	2425-10-7 (Molecular mass-179.22)	Low toxic (Cramer Class): I	**XXII**
3	Xylachlor	63114-77-2 (Molecular mass-239.77)	The test results show that metolachlor is practically non-toxic to birds. From the concept of structure-activity relationship, we can say xylachlor may also be non-toxic to birds.	**XXIII**

(*continued on next page*)

Table 4 (*continued*)

Sl. No	Pesticide name (Group)	CAS no and Molecular mass	Safety and Hazards	Sources (all references available in Supplementary 2)
4	XMC	2655-14-3 (Molecular mass-179.22)	It has a low toxicity and is relatively stable	**XXIV**
5	Warfarin	81-81-2 (Molecular mass-308.35)	It is practically non-toxic	**XXV**
6	Vinegar	90132-02-8 (Molecular mass-60.06)	Vinegar is used to promote the health of the birds	**XXVI**
7	Vinclozolin	50471-44-8 (Molecular mass-286.12)	Vinclozolin is practically nontoxic to birds	**XXVII**
8	Uniconazole	83657-22-1 (Molecular mass-291.81)	Uniconazole-p is non-toxic to birds	**XXVIII**
9	Umifoxolaner	2021230-37-3 (Molecular mass-299.64)	Low toxic	**XXIX**
10	Triticonazole	131983-72-7 (Molecular mass-317.82)	Triticonazole is non-toxic to pollinating insects	**XXX**
11	Triprene	40596-80-3 (Molecular mass-312.52)	Low toxic	**XXXI**
12	Trimethacarb	12407-86-2 (Molecular mass-312.52)	Birds were not as sensitive to trimethacarb	**XXXII**
13	Triisopropanolamine	122-20-3 (Molecular mass-191.27)	Practically non-toxic to birds, fish, honeybees	**XXXIII**
14	Triflumuron	64628-44-0 (Molecular mass-358.70)	Triflumuron is not classified as toxic or highly toxic	**XXXIV**
15	Triflumizole	99387-89-0 (Molecular mass-345.75)	Triflumizole is categorized as being moderately toxic to fish	**XXXV**
16	Triflumezopyrim	1263133-33-0 (Molecular mass-398.34)	Triflumezopyrim was harmless to Anagrus nilaparvatae	**XXXVI**
17	Trifloxystrobin	141517-21-7 (Molecular mass-408.37)	Trifloxystrobin is practically non-toxic to birds	**XXXVII**
18	Trifenofos	38524-82-2 (Molecular mass-363.63)	Profenofos has a moderate toxic	**XXXVIII**
19	Trifenmorph	1420-06-0 (Molecular mass-329.43)	Trifenmorph is hydrolysed at acid pH to relatively non - toxic compounds	**XXXIX**
20	Tridiphane	58138-08-2 (Molecular mass-320.43)	The predictive value for both endpoints indicates this pesticide is less toxic for both endpoints.	**XL**

Environmental implications

The significance of this study lies in establishing a direct connection between chemical toxicity in chickens, human consumption, and environmental harm. Accurate assessment of compound toxicity is vital for managing various adverse effects such as carcinogenicity, genotoxicity, immunotoxicology, and reproductive toxicity. This is not only safeguards of avian species and public health but also addresses challenges like animal testing, time, and cost constraints. The developed PLS-based q-RASTR models emerges as a valuable tool, circumventing these limitations and enabling effective prediction of toxicity. The predictive models, along with the key structural insights gained in the present study, can contribute to develop environmentally friendly and safer chemicals, filling data gaps, and promoting the responsible use of ecotoxic substances.

Funding sources

No specific funding has been received by the author(s) for this work.

CRediT authorship contribution statement

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

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

SD and AS thankfully acknowledged for financial support from the AICTE, New Delhi in the form of a scholarship. PKO is thankful to DTC lab and Prof. Kunal Roy for providing technical assistance and guidance.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2024.134326.

References

[1] Serra, L., Bourdon, G., Estienne, A., Fréville, M., Ramé, C., Chevaleyre, C., et al., 2023. Triazole pesticides exposure impaired steroidogenesis associated to an increase in AHR and CAR expression in testis and altered sperm parameters in chicken. Toxicol Rep 10, 409–427. https://doi.org/10.1016/j.toxrep.2023.03.005.
[2] Fernández-Vizcaíno, E., de Mera, I.G.F., Mougeot, F., Mateo, R., Ortiz-Santaliestra, M.E., 2020. Multi-level analysis of exposure to triazole fungicides through treated seed ingestion in the red-legged partridge. Environ Res 189, 109928. https://doi.org/10.1016/j.envres.2020.109928.
[3] Geiger, F., Bengtsson, J., Berendse, F., Weisser, W.W., Emmerson, M., Morales, M. B., et al., 2010. Persistent negative effects of pesticides on biodiversity and biological control potential on European farmland. Basic Appl Ecol 11 (2), 97–105. https://doi.org/10.1016/j.baae.2009.12.001.
[4] Szabo, J.K., Khwaja, N., Garnett, S.T. Butchart, S.H., 2012. Global patterns and drivers of avian extinctions at the species and subspecies level. https://doi.org/10.1371/journal.pone.0047080.
[5] 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.
[6] Nicolotti, O., Benfenati, E., Carotti, A., Gadaleta, D., Gissi, A., Mangiatordi, G.F., et al., 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.
[7] Kovarich, S., Ceriani, L., Fuart Gatnik, M., Bassan, A., Pavan, M., 2019. Filling data gaps by read-across: a mini review on its application, developments and challenges. Mol Inform 38 (8-9), 1800121. https://doi.org/10.1002/minf.201800121.
[8] Luechtefeld, T., Marsh, D., Rowlands, C., Hartung, T., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 165 (1), 198–212. https://doi.org/10.1093/toxsci/kfy152.
[9] Chirico, N., Gramatica, P., 2011. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51 (9), 2320–2335. https://doi.org/10.1021/ci200211n.
[10] Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol Divers 26 (5), 2847–2862.
[11] Mei, H., Zhou, Y., Liang, G., Li, Z., 2005. Support vector machine applied in QSAR modelling. Chin Sci Bull 50, 2291–2296.
[12] Wu, Z., Zhu, M., Kang, Y., Leung, E.L.H., Lei, T., Shen, C., et al., 2021. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinforma 22 (4), bbaa321. https://doi.org/10.1093/bib/bbaa321.
[13] Roy, J.S., Gupta, K., Talapatra, S.N., 2016. QSAR modeling for acute toxicity prediction in rat by common painkiller drugs. Int Lett Nat Sci 52.
[14] Devillers, J., Flatin, J., 2000. A general QSAR model for predicting the acute toxicity of pesticides to Oncorhynchus mykiss. SAR QSAR Environ Res 11 (1), 25–43.
[15] Chen, S., Sun, G., Fan, T., Li, F., Xu, Y., Zhang, N., et al., 2023. Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): Assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods. Sci Total Environ 876, 162736.
[16] Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs 801–820.
[17] 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.
[18] Roy, K., Kar, S., Das, R.N., 2015. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.
[19] Banerjee, A., Chatterjee, M., De, P., Roy, K., 2022. Quantitative predictions from chemical read-across and their confidence measures. Chemom Intell Lab Syst 227, 104613. https://doi.org/10.1016/j.chemolab.2022.104613.
[20] Luechtefeld, T., Marsh, D., Rowlands, C., Hartung, T., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 165 (1), 198–212. https://doi.org/10.1093/toxsci/kfy152.
[21] Banerjee, A., Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity endpoints. Chem Res Toxicol 36 (3), 446–464. https://doi.org/10.1021/acs.chemrestox.2c00374.
[22] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., et al., 2013. Orange: data mining toolbox in Python. J Mach Learn Res 14 (1), 2349–2353.
[23] Rücker, C., Rücker, G., Meringer, M., 2007. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47 (6), 2345–2357. https://doi.org/10.1021/ci700157b.
[24] Golbraikh, A., Tropsha, A., 2002. Beware of q2! J Mol Graph Model 20 (4), 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1.
[25] Kumar, A., Kumar, V., Podder, T., Ojha, P.K., 2023. First report on ecotoxicological QSTR and I-QSTR modeling for the prediction of acute ecotoxicity of diverse organic chemicals against three protozoan species. Chemosphere, 139066. https://doi.org/10.1016/j.chemosphere.2023.139066.
[26] Chatterjee, M., 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. J Hazard Mater 459, 132129. https://doi.org/10.1016/j.jhazmat.2023.132129.
[27] Katritzky, A.R., Tatham, D.B., Maran, U., 2001. Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. J Chem Inf Comput Sci 41 (5), 1162–1176. https://doi.org/10.1021/ci010011r.
[28] Consonni, V., Ballabio, D., Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques. J Chemom 24 (3-4), 194–201. https://doi.org/10.1002/cem.1290.
[29] Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intell Lab Syst 152, 18–33. https://doi.org/10.1016/j.chemolab.2016.01.008.
[30] Gavaghan, C.L., Arnby, C.H., Blomberg, N., Strandlund, G., Boyer, S., 2007. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. J Comput-Aided Mol Des 21, 189–206.
[31] Yang, S., Kar, S., 2024. First report on chemometric modeling of tilapia fish aquatic toxicity to organic chemicals: Toxicity data gap filling. Sci Total Environ 907, 167991. https://doi.org/10.1016/j.scitotenv.2023.167991.
[32] Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct-Prop Relatsh (IJQSPR) 1 (1), 45–63.
[33] 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.
[34] 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.
[35] 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.

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

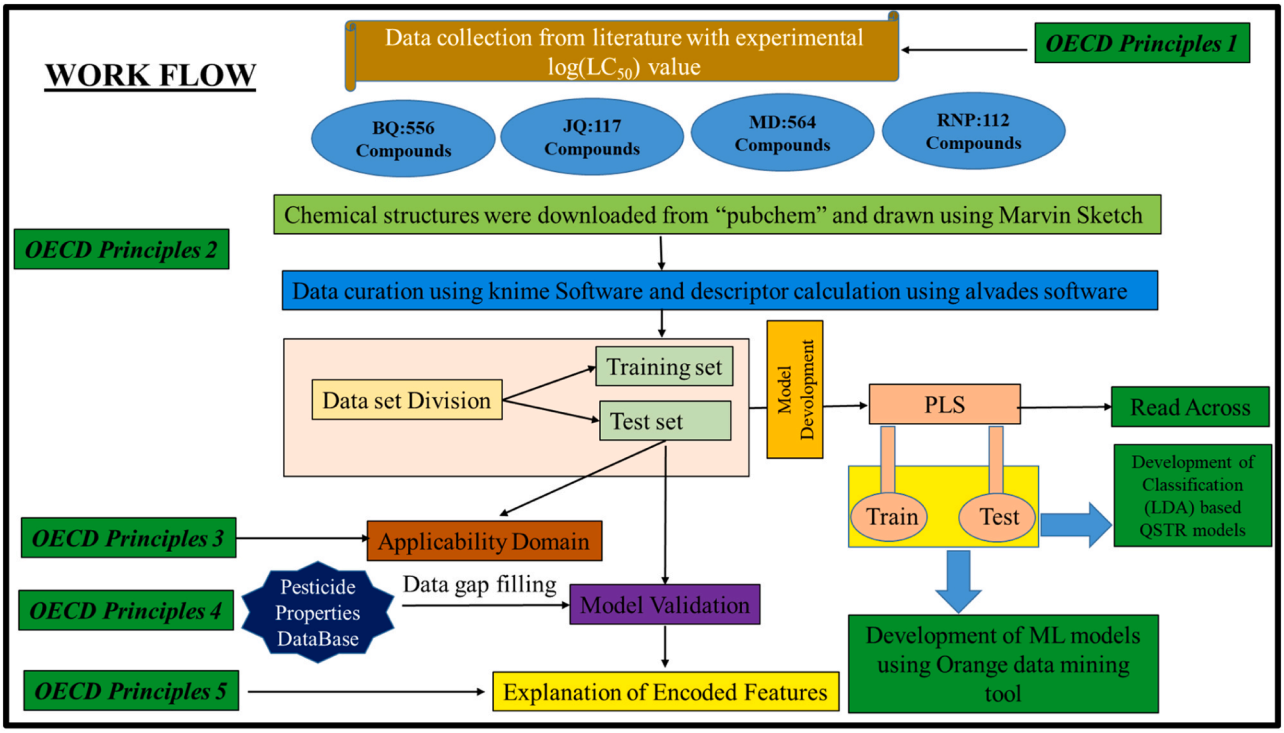


Fig. 1. Workflow of QSTR model development.

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:

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

Model JQ:

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

Model RNP:

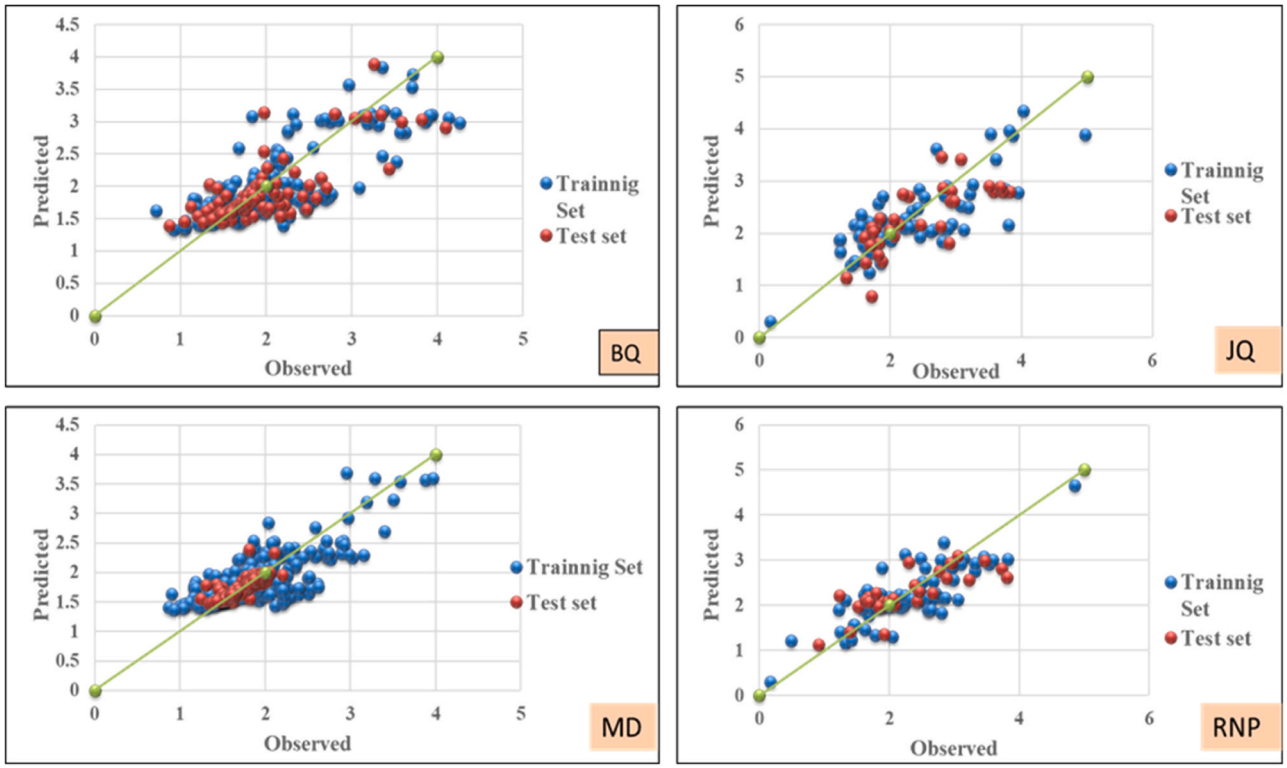


Fig. 2. Scatter plots of developed models.

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

Model MD:

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

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 $\geq$ 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.

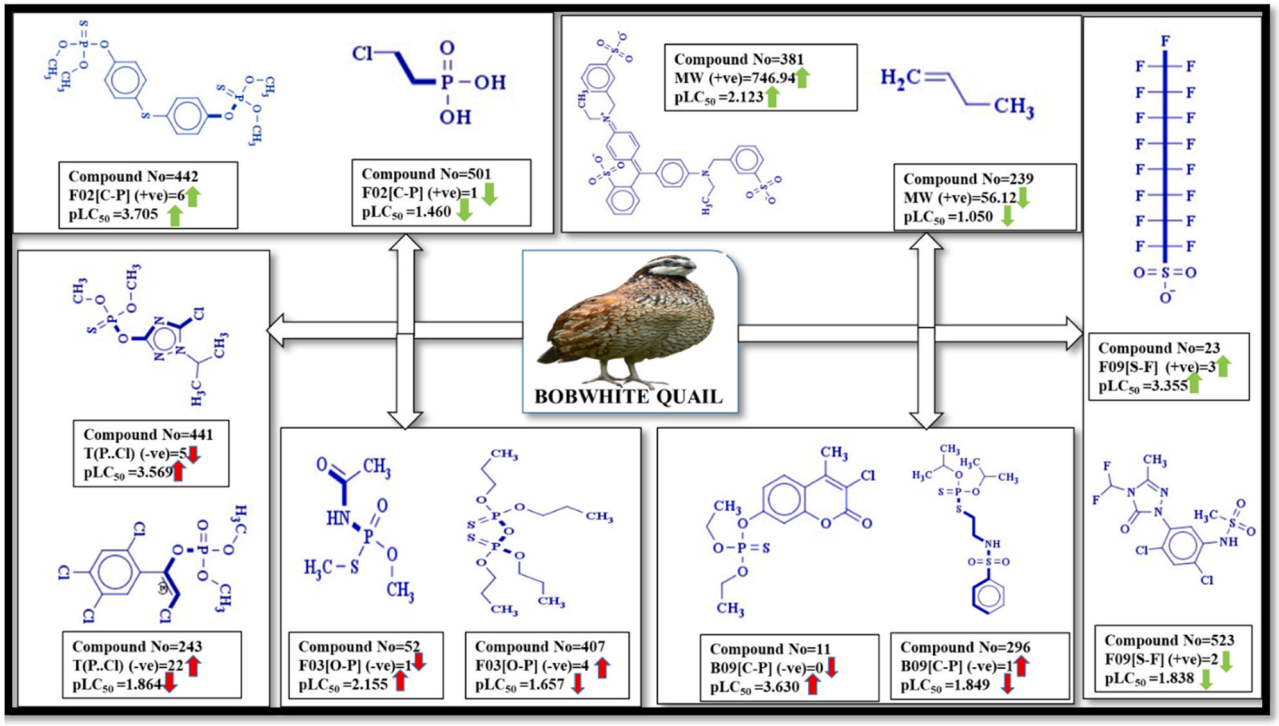


Fig. 3. Positive and negative contribution of model descriptors towards BQ.

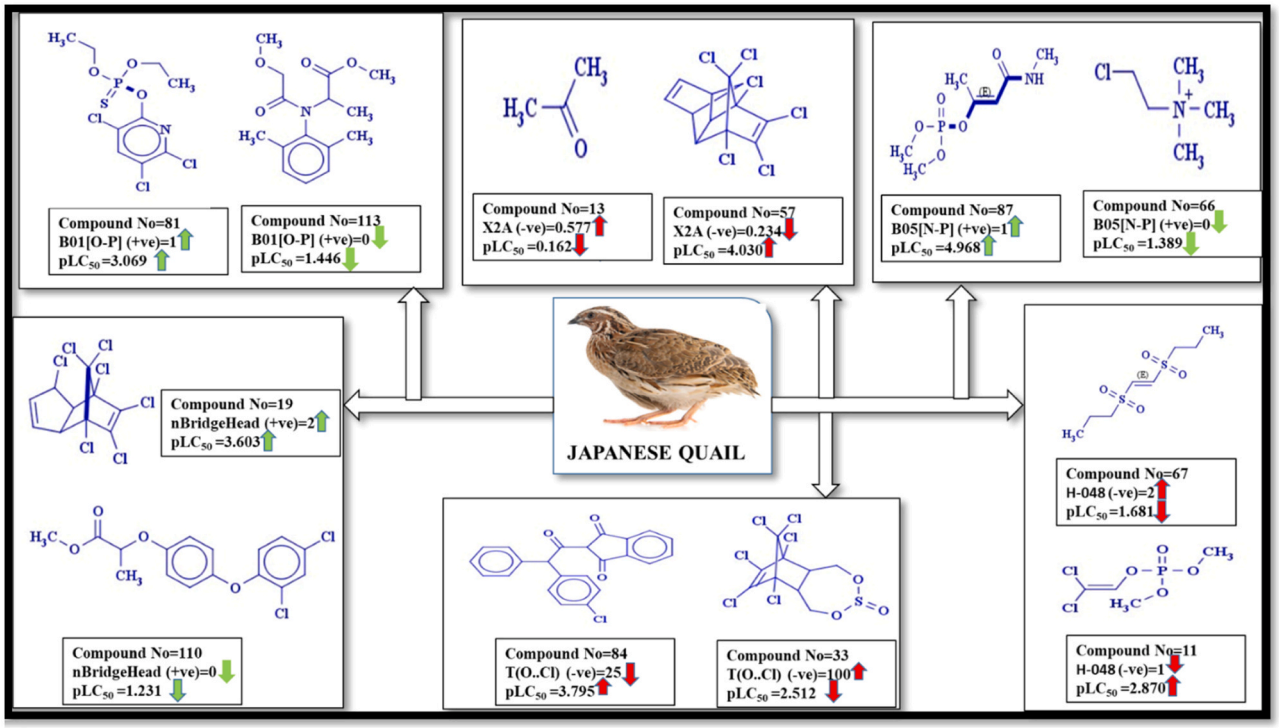


Fig. 4. Positive and negative contribution of model descriptors towards JQ.

(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

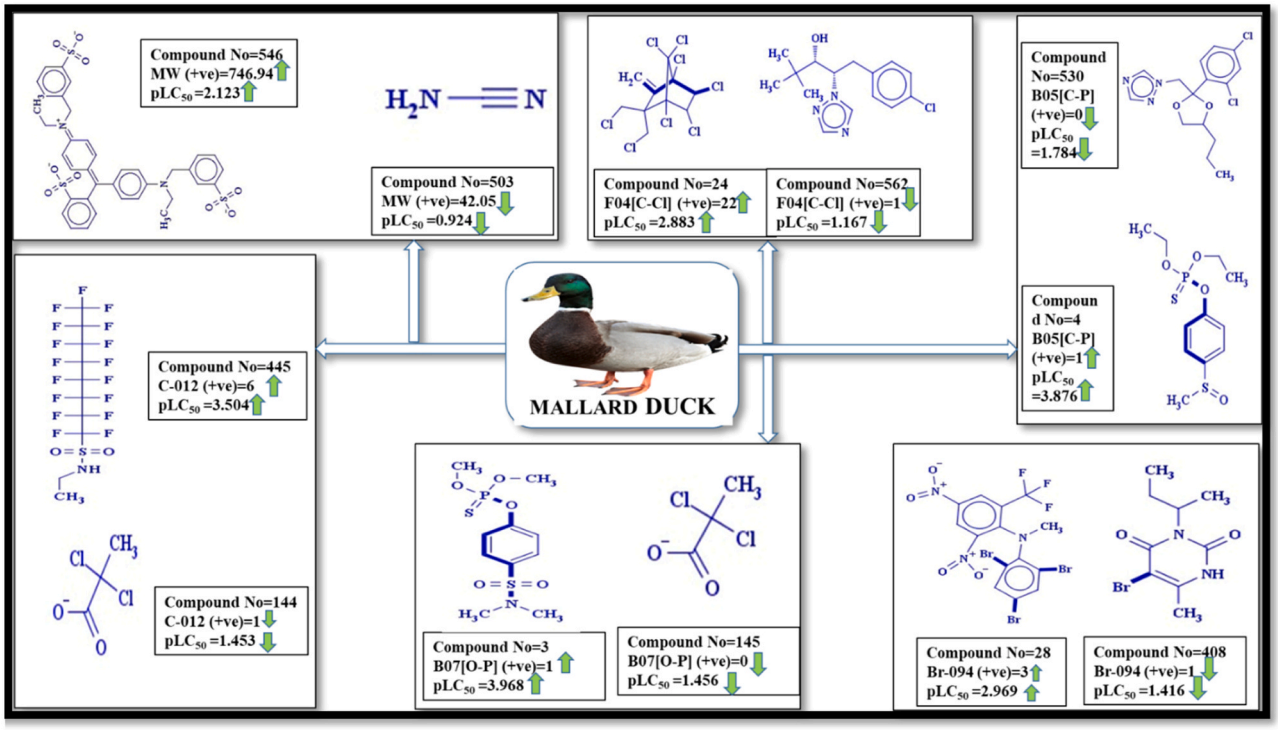


Fig. 5. Positive and Negative contribution of model descriptors towards MD.

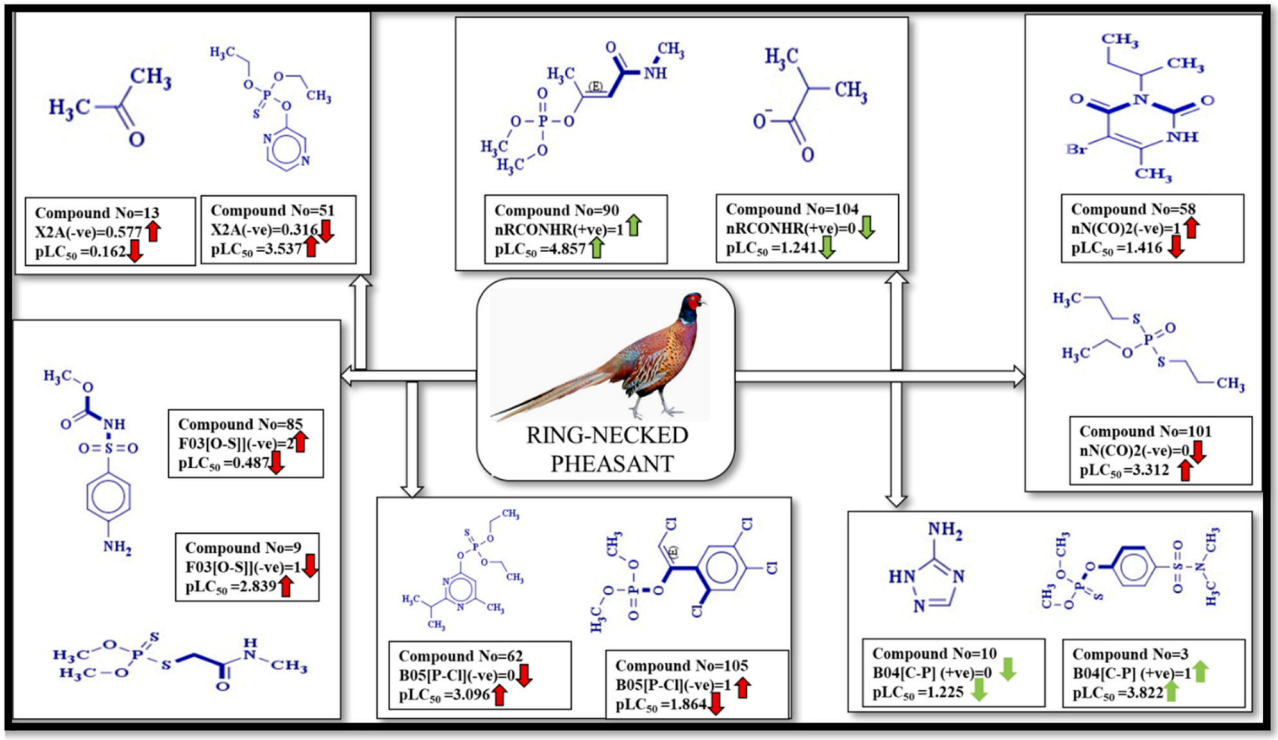


Fig. 6. Positive and Negative contribution of model descriptors towards RNP.

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	-	-	203	-	-	67	-	-
	MD		interspecies	-	-	143	-	-	60	-	-
	ZF		correlation	-	-	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.