q-RASPR modelling of HPLC retention time and fish toxicity of diverse classes of pesticides

Thesis submitted in partial fulfillment for the requirements of the Degree of

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

Shilpayan Ghosh

(B.Pharm)

Registration No. **160253 of 2021-2022**

Examination Roll. No**. M4PHC23021**

Under the Guidance of

DR. KUNAL ROY

Professor

Drug Theoretics & Cheminformatics Laboratory

Department of Pharmaceutical Technology

Jadavpur University

Kolkata – 700 032

India

2023

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research as part of my work on "q-RASPR modelling of HPLC retention time and fish toxicity of diverse classes of pesticides".

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

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

NAME: **SHILPAYAN GHOSH**

EXAMINATION ROLL NUMBER: **M4PHC23021**

REGISTRATION NUMBER: **160253 of 2021-2022**

THESIS TITLE: "**q-RASPR modelling of HPLC retention time and fish toxicity of diverse classes of pesticides"**

SIGNATURE WITH DATE:

CERTIFICATE

Department of Pharmaceutical Technology

Jadavpur University

Kolkata - 700 032

This is to certify that **Mr. Shilpayan Ghosh**, B. Pharm. (NSHM knowledge campus, kolkata), has carried out the research work on the subject entitled **"q-RASPR modelling of HPLC retention time and fish toxicity of diverse classes of pesticides"** under my supervision in Drug Theoretics & Cheminformatics Laboratory in the Department of Pharmaceutical Technology of this university. He has incorporated his findings into this thesis of the same title, being submitted by him, in partial fulfillment of the requirements for the degree of Master of Pharmacy of Jadavpur University. He has carried out this research work independently and with proper care and attention to my entire satisfaction.

_____

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

_____ _____

(Prof. Sanmoy Karmakar)
Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

(Prof. Saswati Mazumdar)
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata

ACKNOWLEDGEMENTS

*I deem it a pleasure and privilege to work under the guidance of **Dr. Kunal Roy**, Professor, Drug Theoretics & Cheminformatics Laboratory, 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. Kunal Roy for his valuable guidance throughout the work that enabled me to complete the work. With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, perpetual assistance and co-operation from my seniors Mainak Chatterjee, Priyanka De, Arkaprava Banerjee, Sapna Kumari Panday, Joyita Roy and Vinay Kumar. Their constant support and helpful suggestions have helped me to accomplish this work in time. I would like to express my special thanks to my juniors Subham Panday and Doelima Bera who all have extended their helping hands and friendly cooperation all through my work.*

I am thankful to the authority of Jadavpur University and Head of the Department Prof. Dr. Sanmoy Karmakar 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.Haru Ghosh and Mrs. Shima Ghosh for all the love and inspirations without which my dissertation work would remain incomplete.

_____

SHILPAYAN GHOSH

Examination Roll No.: **M4PHC23021**

Department of Pharmaceutical Technology,

Jadavpur University,

Kolkata-700032

PREFACE

This dissertation is presented for the partial fulfillment of the degree of Master of Pharmacy (M. Pharm). The time span for this present study has almost taken over one year. This present study has been explored through the development of predictive in silico models on retention time (log t_R) of pesticides and the aquatic toxicity [Log (1/LC50)] of organic pesticides toward various fish species, using "Read-across v4.1" and "RASAR descriptor v2.1" tools. Currently, different in silico approaches are being implemented for predicting the toxicity of industrial chemicals as a rational alternative technique to animal testing. The application of statistical models to predict physicochemical and biological properties started with linear regression models developed by Hansch in the early 1960s. Since the appearance of computer-aided drug design (CADD) studies, the term "Quantitative Structure-Activity Relationships (QSARs)" has become quite popular in the field of medicinal, synthetic, and environmental chemistry. Development of predictive "Quantitative Structure-Property Relationship (QSPR)" and "Quantitative Structure-Toxicity Relationship (QSTR)" models enables the assessment of the toxicological hazard of different chemical compounds from the chemical information extracted by employing descriptors.

In contrast to the QSPR/QSAR/QSTR, which uses various statistical algorithms to fit the structural information of chemicals with the properties they exert, read-across is another similarity-based predictive method that uses straightforward algebraic calculations to predict the properties of query chemicals from structural or functional analogues. Additionally, environmental toxicology and regulatory decision-making frequently employ this technique. Even in the absence of a large amount of experimental data, read-across can still be utilized to make predictions that are both quantitative and qualitative. Additionally, researchers are developing combinatorial modelling strategies. "Read-across structure-activity relationship" (RASAR), which uses similarity parameters obtained from chemical fingerprints by the traditional read-across algorithm to build supervised QSAR for the prediction of exerted activity/property/toxicity, was reported by Luechtefeld et al. in 2018.

Any product or mixture intended to prevent, get rid of, deter, or control a pest is referred to as a pesticide. Pesticides are used to regulate the growth of plants in addition to acting as

defoliants or desiccants and stabilizers of nitrogen. These compounds are therefore used to control a range of pests and disease carriers, such as mosquitoes, ticks, rats, and mice. Other applications for insecticides in agriculture include the control of weeds and insect infestations. The use of pesticides in modern agriculture has become essential. Modern pesticides are playing an ever-more-important role in defending plants and commodities against pests, controlling the proliferation of undesirable target organisms, and limiting the spread of disease as agricultural science and technology advance quickly. Pesticides can be hazardous to non-target aquatic and terrestrial species as well as people when used improperly, in excess, or without proper dose instructions (Yu and Zeng, 2022). Due to the simple entry of agricultural waste into water bodies through sewage systems, pesticides and their residues mostly impact aquatic creatures. The bioaccumulation and biomagnification of pesticides through aquatic animals can also endanger for human health. Therefore, it is necessary to assess the potential risk of pesticides to human health and environmental species.

Fishes are frequently employed as model species to assess aquatic toxicity in water contamination, particularly under EC Regulation 1107/2009 (European Pesticide Regulation No. 1107/2009). Before pesticides and their residues can be sold, registrants must examine their environmental safety. As a result, fish have been routinely used as model species to evaluate the aquatic toxicity of industrial pollutants. A typical animal test is used to measure acute fish toxicity, providing LC50 (lethal concentration for population values) according to OECD test no:203, but there is an increasing need to reduce or substitute animal trials for regulatory purposes (Tunkel et al., 2005). However, they are expensive, time-consuming, and involve the use of animal models. In silico innovative approaches may be used to address the complications involved with experiments with the application of computer modelling techniques such as quantitative structure-activity relationships (QSAR), read-across prediction, and read-across structure activity relationships (RASAR). This study is the accompaniment of the 3Rs (replacement, refinement and reduction of animals in research) reducing animal testing. An appropriately statistically validated model can produce toxicity data of chemicals within the applicability domain of the model, and such predictions can be used as an alternative or supplement of the experimental/observed results. Again, the application of the developed RASAR model provides supervised mathematical models from which knowledge regarding the significance of modelled features and their contribution toward property prediction can be extracted. As a result, the benefits and advantages of

combinatorial modelling make this methodology more efficient in the field of cheminformatics.

In the present section of Case Study-1, a q-RASPR model has been developed with the experimental retention time data (log t_R) of 823 environmentally relevant pesticide residues (identified in foods and vegetable products). Due to its potential as a toxicity indicator, the retention time obtained from reverse-phase HPLC analysis was utilized as the endpoint of the dataset. To circumvent the complexities of conformational analysis and energy minimization, this work relied on simple 2D descriptors. The RASAR descriptors were derived from the 2D descriptors identified in this work by the developed QSPR model. A comparison of the q-RASPR with previously published work revealed an improvement in external predictivity and a reduction in prediction errors; the external predictivity of the q-RASPR model in terms of Q^2_{F1} value outperformed the quality of the previously reported MLR model. As a result, it has the potential to be employed as a suitable alternative strategy for predicting retention time and identifying toxicity.

In Case Study-2, the primary goal of this study was to develop regression-based in silico models for the aquatic prediction of various organic pesticides. In this direction, we have developed q-RASAR models based on experimental data [Log (1/LC50)] of organic pesticides to various fish species, including Rainbow trout (*Oncorhynchus mykiss*), Lepomis (*Lepomis macrochirus*), and Miscellaneous (*Pimephales promelas* and *Brachydanio rerio*). The experimental Log 1/LC50 value is taken as the potential indicator of acute aquatic toxicity in this study (Braunbeck et al., 2005). To encode the structural properties of pesticide compounds, preliminary 0D-2D descriptors were successfully applied, avoiding the additional complexity involved with higher-order descriptor (3D-7D) calculation. During the QSAR and q-RASAR modelling, the hydrophobicity of organic pesticides was discovered as a toxicity mediating component that positively increases overall aquatic toxicity in this study. We have developed the preliminary QSAR models from the respective datasets, have strictly validated those models, and then applied the conjoined q-RASAR algorithm to further enhance the external predictability of those models in terms of Q^2_{F1} and Q^2_{F2} values and also reduce the errors of test set predictions (MAE_{test}). To the best of our knowledge, no RASAR model has yet been published in the literature for predicting the aquatic toxicity of organic pesticides to different fish species. As a result, this method is an intentionally added new

method of pesticide risk assessment that will assist us in providing more accurate aquatic toxicity forecasts and understanding the toxicity mechanisms without harming animals.

The following analyses have been performed in this dissertation:

1) Predictive q-RASPR modeling of the retention time (log t_R) of pesticide residues present in foods and vegetables

2) Quantitative Read-Across Structure-Activity Relationship (q-RASAR): A New Approach Methodology to Model Aquatic Toxicity of Organic Pesticides Against Three Different Fish Species

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

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

ABBREVIATION

2D QSAR	Two-dimensional QSAR	MLC (LC50)	Median Lethal Concentration
3Rs	Replacement, Refinement and Reduction	EC	Effective Concentration
QSAR	Quantitative Structure Activity Relationships	MEC (EC50)	Median Effective Concentration
q-RASAR	Quantitative Read-across Structure Activity Relationships	MED (ED50)	Median Effective Dose
HPLC	High performance Liquid Chromatography	BT	Baseline Toxicity
ECA	European Chemical Agency	MLD (LD50)	Median Lethal Dose
USEPA	United states environmental protection agency	MED (ED50)	Median Effective Dose
QSPR	Quantitative Structure Property Relationships	AF	Application Factor
q-RASPR	Quantitative Read-across Structure Property Relationships	TI	Therapeutic Index
QSTR	Quantitative Structure Toxicity Relationships	ED	Euclidean Distance
SAR	Structure Activity Relationships	GK	Gaussian kernel
LR	Linear Regression	LK	Laplacian kernel
MLR	Multiple Linear Regression	OECD	Organization for Economic, Co-operation and Development
PCA	Principal Component Analysis	SMILES	Simplified Molecular Input Line Entry System
PCR	Principal Component Regression	REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
ANN	Artificial Neural Networks	LOO	Leave-One-Out

KNN	K-Nearest Neighbours	EMEA	European Medicines Evaluation Agency
SOM	Self-Organizing Map	EU	European Union
RP-HPLC	Reverse phase High Performance Liquid Chromatography	FDA	Food and Drug Administration
GA	Genetic Algorithm	LC	Lethal Concentration
BSS	Best Subset Selection	VIP	Variable Importance Plot
PLS	Partial Least Square	RMSEC	Root mean square error of calibration
SVS	Stepwise Variable Selection	RMSEP	Root mean square error of prediction
VSS	Variable Subset Selection	CCC	Concordance Correlation Coefficient
FA	Factor Analysis		
LDA	Linear Discriminant Analysis		
DA	Discriminant Analysis		

Contents

CHAPTER - 1

INTRODUCTION

"I eliminated coffee and fish from my diet, the pesticides in coffee and fish, as well as the mercury in the latter, are considered possible contributors to birth defects in fetal tissue."

-Constance Marie

1. INTRODUCTION

Pesticides are necessary tools used for regulating destructive and invasive pests in agriculture, forestry, and the landscape. Pesticides include all materials that are used to prevent, destroy, repel, attract or reduce pest organisms. All substances used to stop, eliminate, deter, attract, or lessen pest organisms are considered as pesticides. Among the most well-known pesticides include insecticides, herbicides, fungicides, and rodenticides. Others include growth regulators, plant defoliants, surface disinfectants, and various pool chemicals. All pesticides used in Canada must be registered by Health Canada in accordance with federal law (AY- Sangodoyin et al., 1996). This covers every kind of pesticide, whether synthetic or naturally occurring, utilized in both conventional and organic farming practices.

Pesticides Name:

Three names are used to identify pesticides: the chemical name, the trade name, and the active ingredient.

The pest-controlling chemical is the **active ingredient**. On a pesticide label, the name of the active ingredient is known as the common name. An illustration would be the herbicide active chemical known as **glyphosate**.

The brand name of pesticide is known as the **trade name**, according to the producer. The name which stands out the most on the front of a pesticide level. The same active component may be included in pesticides with various trade names, and some pesticides have multiple active ingredients. This is exemplified by **Roundup**, which is one of the many trade names for glyphosate-containing herbicide products.

Scientists refer to an active ingredient by its **chemical name**, which is the name of its chemical composition. The chemical name of the active component in Roundup is **N-(phosphonomethyl) glycine**.

Classification of pesticides based on Mode of Entry:

Pesticides that are absorbed by plants or animals and go to untreated tissues are known as **systemic pesticides**. Since the **non-systemic pesticides** act on the target pests when they come into touch, they are also known as contact pesticides.

Pesticides that cause stomach poisoning and enter a pest's body through their mouth and digestive tract and result in death by poisoning, these are termed as **stomach poisoning pesticides.**

Pesticides are also known as **fumigants** which act or may kill the target pests by producing vapour.

Repellents may not actually kill, but they are unpleasant enough to deter pests from treated areas/commodities and also they hinder pest's ability to find crops.

Classification based on Chemical Composition of Pesticides:

Most common and useful method for classification of pesticides is dependent on their chemical composition and nature of active ingredients. The efficacy, physical characteristics, and chemical composition of the various pesticides are revealed by this form of classification.

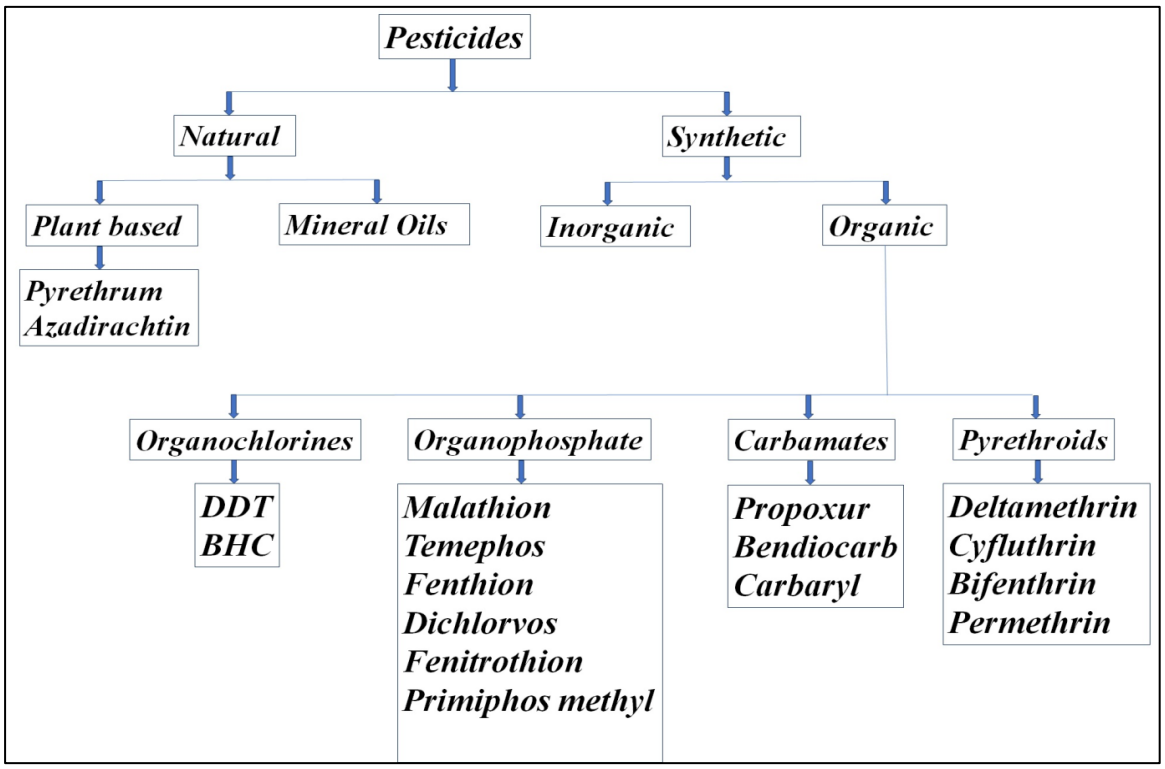


Fig.1.1. Flow diagram of classifying pesticides based on chemical composition

Mechanism of action (MOA) of pesticides:

Pesticides can be categorized based on how they control the target pest or how they work. This is also referred as the primary site of action. For instance, one insecticide may have an impact on an insect's nervous system, while another may have an impact on moulting.

Another herbicide may interfere with the plants capacity to turn light into food, while one herbicide may mimic the plants growth regulators. MOA of a pesticide is indicated by a group number or letter on the label. The group number for products with a similar MOA will have the same group number.

Chemicals originating from animals, bacteria, fungi, and plants are also referred to as **biopesticides.** Microbial pesticides are those that have a living bacterium, fungus, virus, protozoan, or alga as its active ingredient. Bacillus thuringiensis (Bt), a common biopesticide with many strains to specifically manage caterpillars, mosquitoes, or beetles, is a type of **biopesticide**.

Nowadays, pesticides have become an indispensable part of the modern agricultural system. With the rapid advancement of agricultural research and technology, contemporary pesticides are playing an increasingly crucial role in protecting plants and commodities from pests, regulating the overgrowth of undesired target organisms, and preventing disease transmission. When pesticides are used on crops, residues may remain, which might remain in food and enter the food chain. Pesticide residues have detrimental impacts on both humans and pollinators as a result of their toxicity and deleterious consequences. However, the release of pesticides also affects all the areas of natural resources, including the environment, water, soil, and animals. Also, pesticide mishandling has historically resulted in major difficulties for both human and animal health as well as the environment. Pesticide residues in food and the environment must be reduced, and more information about the risks of pesticides and their residues must be provided by regulators. In this regard, national and international authorities as well as environmental organizations have advocated for strict regulation of public health and environmental protection. Pesticide extraction from food products (e.g., fruits, vegetables, grains, etc.) is regarded as an important stage in pesticide residue analysis because it serves as the foundation for pesticide detection at the trace level. For the proper risk assessment of pesticides and their residues, experimental approaches, such as solvent extraction, high- performance liquid chromatography (HPLC), mass spectrometry, etc., are widely used. As a result of the intricacy of the process, various aspects must be optimized during pesticide extraction, which is time-consuming and labor-intensive. To avoid the associated complexities of experimental approaches, *in silico* new approach methodologies can be used. The use of computational modeling, like quantitative structure−activity relationships (QSARs), read-across, machine learning models, etc., is

recommended for the toxicity predictions of different chemicals against eco- toxicological end points by different chemical regulatory agencies.

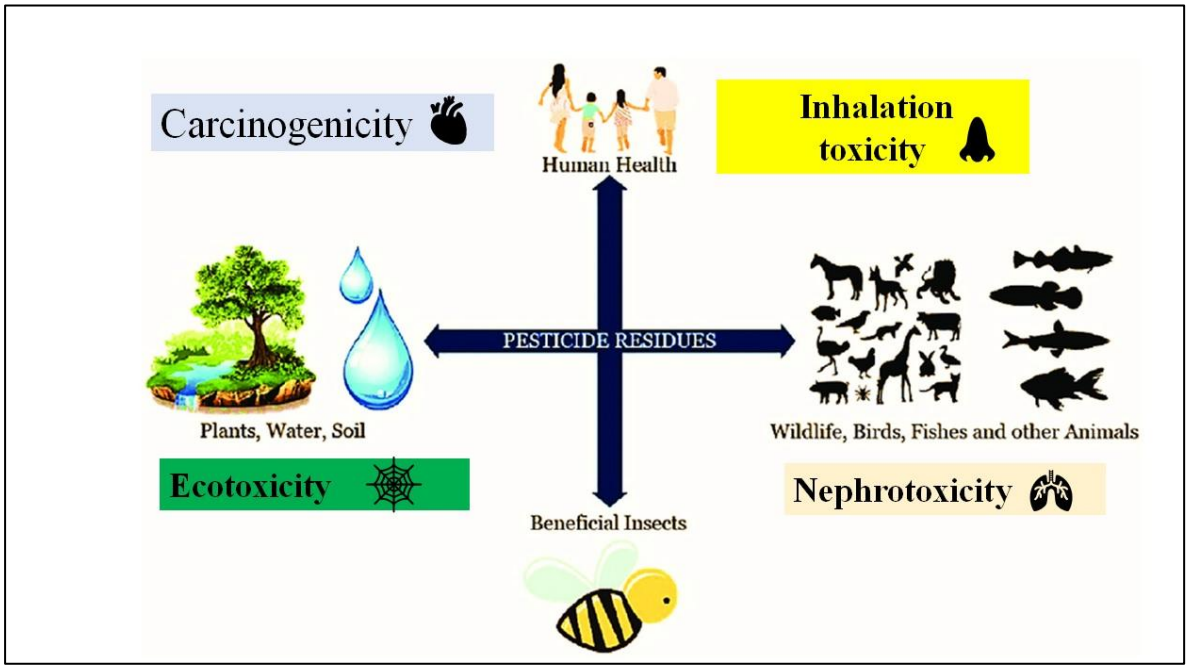


Fig.1.2. Impact of Pesticide residues on human health, insects and environment

Besides this, the non-specificity, improper usage, and unspecified dosage of pesticides cause toxicity to non-target aquatic and soil species as well as human beings. Aquatic organisms are mostly affected by pesticides and their residues due to the easy passage of agricultural waste into water bodies through sewage systems. Human health can also be at threat due to the bioaccumulation and biomagnification of pesticides via aquatic species. Therefore, it is essential to identify the potential risk of pesticides to human health and environmental species. Fishes are commonly used as model species to evaluate aquatic toxicity in water pollution; specifically, under EC Regulation 1107/2009 (European Pesticide Regulation No. 1107/2009). Registrants must assess the environmental safety of pesticides and corresponding residues before entering into the market. So, fishes have been widely utilized as model species to assess the aquatic toxicity of industrial chemicals. As per the OECD test no:203, a typical animal test is used to determine acute fish toxicity, yielding *LC*50 (lethal concentration for population values). Still, there is an increasing need to decrease or replace animal experiments for regulatory purposes. Also, the experimental procedure for assessing fish acute toxicity is time-consuming and labour-intensive. To overcome these experiment-associated complexities, in silico innovative approach methodologies might be adopted. The use of computational modelling, like Quantitative structure-activity relationships (QSAR),

read-across prediction, machine-learning models etc. can be used to predict the aquatic toxicity of chemicals. These approaches are efficient, cost-effective, fast, and comparatively less troublesome. An exceptional advantage of in silico technique over the experimental approach is the ability to predict toxicity/activity before even synthesizing the compound. Regulatory agencies such as European Chemical Agency, US-Environmental Protection Agency etc. are also promoting the use of these computer-based new approach methodologies to limit the usage of animals in acute toxicity testing.

The efficacy of a chemical compound can be quantified by precise experimental techniques. Currently, an approximate existence of about 30 million chemical compounds (Benfenati.et al.,2012) including pharmaceuticals, food products, preservatives, colouring agents, pesticides, varnishes, cosmetics, cleansing agents, fuels, textiles, cements, explosives etc. has been reported in the market (European Commission, 2006). Even though these chemicals might be beneficial substances for their intended use, a comprehensive evaluation report on their harmful/undesirable outcomes is not widely available. Hence, observing the behavior of such a large number of chemicals should unavoidably include some rational approach which is less time and money consuming without compromising the eventual goal. The application of predictive modelling technique is highly efficient and productive to resolve such issues.
With the passage of time, the association of physical and biological features with chemistry on the foundation of absolute mathematical perceptions has paved the way for the development of a useful method known as the "Quantitative Structure-Activity Relationship (QSAR)."The sole goal of this strategy is to investigate the chemical characteristics responsible for their behavioral manifestation. On the other hand, this type of study uses quantitative data gleaned from many chemistry facts to attempt to link various aspects of molecules, such as biological activity, physicochemical property, or toxicity.

1.1. Quantitative structure-activity relationship (QSAR) analysis

The principles of the QSAR study presume that the chemicals existing in nature contain in their structures information for their chemical, physical, biological or toxicological properties which can be properly defined by different numerical relationships. Eq. 1.1 defines the response shown by molecules to be a numerical equation of chemical attributes.

Chemical Response = f (Chemical attributes) = f (Structure, Property) (1.1)

The term "response" in Eq. 1.1 refers to any activity, physicochemical characteristic, or toxicity shown by a molecule, whereas "molecular features" refers to the quantitative data obtained from the molecules using exact experimental/theoretical approaches. The methodologies or techniques are named recognized based on the nature of the response. Therefore, three comprehensive classes—QSAR, QSPR, and QSTR—can be identified, these classes represent the response as biological activity (e.g., anti-bacterial, antiviral, anticholinergic, anticancer, anti-Parkinson's, anti-Alzheimer's, anti-diabetic, anti-malarial, antitubercular, etc.); physicochemical property (boiling point, melting point, lipophilicity, viscosity, molar refractivity, aqueous solubility, glass transition temperature etc.) along with toxicity (systemic such as nephrotoxicity, cardiac toxicity, hepatotoxicity, pulmonary toxicity; non-systemic such as aquatic toxicity etc.). Though, we shall denote the term 'QSAR' for QSAR/QSPR/QSTR studies in a broader way.

The key objective of a QSAR study is to develop a numerical relationship between the response and molecular features of a set of chemicals. Hereafter, this is all about evolving a mathematical equation for a set of compounds between a Y and some X variables, where Y is the dependent variable located in the left-hand side and the X variables are independent variables located in the right-hand side of the equation. The Y variable or response is also known as 'endpoint' whereas the X molecular features are known as 'predictor variables', and a QSAR equation can be expressed as follows (Eq. 1.2):

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots a_nx_n \quad (1.2)$$

In Eq. 1.2, Y is the response variable (dependent) being modeled for a set of n number of molecules with the predictor variables (independent) x_1, x_2, x_3... x_n etc. having coefficient values a_1, a_2, a_3... a_n respectively, and a_0 is a constant. With the known values of Y and x variables the coefficients (a_1, a_2, a_3... a_n) together with the constant (a_0) can be determined from a precise mathematical equation associating activity, physicochemical property or toxicity of chemicals with their molecular features. Interestingly in interspecies toxicity modelling, one species' toxicity endpoint is used as an independent variable for another species toxicity endpoint modelling (dependent variable).

1.1.1. Brief history of QSAR

Over a century ago, the stone of QSAR was laid in 1863, when Cros (Cros, 1863) discovered the existence of an inverse relationship between water solubility and toxicity of a chemical. The hazardous potential of alcohols towards mammals is primarily increased as the water solubility of those alcohols decreases. Soon after, in 1868, Crum-Brown and Fraser (Brown and Fraser, 1868) explained that how different chemical substituents influence the physiological properties of a molecule. Then in the 1890s, Hans Horst Meyer found that the toxic potential of organic molecules was influenced by their lipophilicity (Borman, 1990; Lipnick, 1991).Then, a linear relationship between lipophilicity (e.g., oil-water partition coefficients) and biological activity was investigated. Louis Hammett (Hansch et al., 1991) investigated the relationship between the electrical characteristics of distinct organic bases and acids and their reactivity and equilibrium constants. These preliminary researches laid the groundwork for the modern QSAR by proposing how molecular characteristics influenced an endpoint (i.e., chemical property and biological activity) of a molecule. In 1962, Corwin H. Hansch and colleagues (Hansch et al., 1962) formally introduced the term "QSAR" and laid the groundwork by investigating the "structure-activity relationship (SAR)" of various plant growth hormones and insecticides and their dependence on "Hammett constants and lipophilicity (Gallup et al., 1952).

The Free-Wilson model (Free and Wilson, 1964) is an efficient and simple approach to quantitatively describe SAR. It uses the presence or lack of functional groups or substituents as molecular descriptors to describe differences between a group of congeneric molecules. In contrast to Hansch analysis, where physical qualities are in a correlation with biological activities, it is only a mathematical approach that directly correlates molecular features with biological activity/properties (Kubinyi, 1988). However, from a theoretical as well as a practical perspective, both of these approaches are closely related (Kubinyi, 1988).In several research, these models were joined to form a single framework that included Free-Wilson type parameters to describe the contributions of various structural traits and physicochemical qualities to reflect the effect of substituents on biological activity (Kubinyi, 1988; Wei et al., 2001). Many researches, specially Hansch and his co-workers (Verma and Hansch, 2009; Hansch et al., 2002; Kurup et al., 2000; Gao et al., 1999; Selassie et al. 2002; Kurup et al., 2001; Hansch and Gao, 1997; Kurup et al., 2001; Hansch et al., 1996; Hadjipavlou-Litina et al., 2004; Garg et al., 1999, 2003) on the SAR of inhibitors of enzymes, showed that this integrated model provides an excellent performance for ‘classical QSAR’ (Hansch, 2011).Some dissimilarities to Free-Wilson model have been recently established and found

suitable to apply in 'fragment based drug design' (Eriksson et al., 2014; Chen et al., 2013; Radoux et al., 2016).

The area of QSAR modelling has developed gradually and this includes two fundamental evolutions as follows:

1) Paradigm move from the 'classical' to the 'non-classical QSAR' method (Fujita and Winkler, 2016). The first one is based on a short set of congeneric molecules that generally have a single mechanism of action while the last one is based on a huge non-congeneric and heterogeneous dataset and that may encompass some mechanism of actions.

2) Paradigm movement of QSAR models (Nantasenamat et al., 2009; 2010; Cherkasov et al., 2014) that studies the SAR of "several compounds against a single target protein" to the "proteochemometric model" (Cortes-Ciriano et al., 2015; Qiu et al., 2016) (also as "computational chemogenomics") that examines the SAR of "several compounds against several target proteins".

1.1.2. The objectives of QSAR study

There are several practical purposes of a QSAR and these methods are exploited broadly in many conditions. The basic objectives of QSAR include the following:

1) Prediction of new analogues of molecules with better properties.

2) To simply understand and explore the mechanism of actions.

3) To optimize a lead compound with reduced toxicity (Tong et al., 2005).

4) To reduce the cost, time and manpower necessity by evolving a more active and effective molecules employing a scientifically convenient and productive way.

To accomplish the above-mentioned objectives, it is essential to have a thorough knowledge of the following characteristics:

(i) Thorough knowledge of the mechanism of action of the compounds.

(ii) Several parameters influence the experimental conditions of the molecules.

(iii)A detailed inspection of molecular structures and their properties."Quantitative structure - activity relationship (QSAR)" is a multidisciplinary research area lying in the intersection of chemistry, biology, machine learning, and statistics. By the prediction of the crucial structural features needed to obtain a molecule with optimized activity/toxicity/property, the QSAR approach offers a good platform for the synthesis of a relatively small number of desired chemical compounds with improved activity/toxicity/property.

1.1.3. Molecular descriptors

The term "molecular descriptor" refers to the precise mathematical information about a molecule of interest. According to Todeschini and Consonni *"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiments."*(Todeschini and Consonni, 2008).In other words, the modeled response (activity/property/toxicity) is represented as a function of numerical or quantitative values of structural traits or features known as descriptors for a QSAR model (shown in Eq. 1.3).Cheminformatics techniques rely on the creation of chemical reference spaces in which a novel molecule can be predicted using a built QSAR model. The chemical spaces are remarkably dependent on the use of in silico descriptors of a molecule structure of interest, chemical or physical properties, or other features.

Response (activity/property/toxicity) = f (chemical structure or property information) = f (descriptors) (1.3)

The type of descriptors used and their ability to encrypt the structural properties of the molecules associated with the response are pivotal factors in determining the superiority of a QSAR model. The descriptors may be physicochemical (electronic, steric or hydrophobic), structural (depending on the occurrence frequency of a substructure), geometric (based on the calculation of molecular surface area), electronic (based on the calculation of molecular orbitals), topological, or simple indicator variables (replicated parameters). An efficiency of a descriptor is dependent profoundly on the cases being studied. The molecular properties, in particular, are dependent on a precise endpoint. The finest combinations of features that create a descriptor perfect for the development of a QSAR model are listed below:

1) An important characteristic is the capacity to plot descriptor values back to the structure to visualize (Segall et al., 2009). These visualizations are practical when descriptor values are related with structural features.

2) It is also necessary for a descriptor to have several kinds of physical interpretability in order to encrypt the requisite properties of the molecules of interest.

3) A descriptor should be appropriate for a broad range of compounds.

4) A descriptor should be interrelated with the structural attributes for a definite end point and show minor redundancy with other descriptors.

5) A descriptor must produce different values for structurally diverse molecules, despite the structural diversity is minor. On the other words the descriptor should show negligible errors. Moreover, to the errors, a descriptor should be continuous. It indicates that small structural alterations should result a little deviation in descriptor value.

6) A descriptor being computed quickly should not be dependent on the experimental results, which might be considered more appropriate than the computationally comprehensive another one which depends solely on experimental properties.

1.1.3.1. Classification of descriptors

Descriptors can be categorized depending on the process of their determination or computation, i.e., physicochemical (electronic, steric, or hydrophobic), structural (depending on the occurrence frequency of a substructure), geometric (based on the calculation of molecular surface area), electronic (based on the calculation of molecular orbitals), topological, or simple indicator variables (replicated parameters).In a wider sense, descriptors (especially, physicochemical descriptors) can be divided into two main classes: (1) whole molecular descriptors and (2) substituent constants (Todeschini and Consonni, 2008; Livingstone, 2000).

1) Whole molecular descriptors: Extension of substituent constants method.

2) Substituent constants: Physicochemical descriptors that are established depending on the physicochemical properties of molecules.

1.1.3.2. 2D descriptors

1.1.3.2.1. Physicochemical descriptors

These descriptors are resultant from the several physiological attributes, having a relationship with physicochemical characteristics.Alteration of the physicochemical traits leads to changes in absorption, distribution, metabolism, and excretion. The important physicochemical factors influencing the biological activity of chemicals and drugs include

steric, hydrophobicity, partition coefficient and electronic character of the entire molecules as well as the presence of substituted groups in the molecule.(Taylor, 1991)

1.1.3.2.1.1 Partition coefficient

The relative affinity of a molecule for an aqueous or lipid (oil) media is important for its activity, as partitioning phenomena have an impact on absorption, distribution, metabolism, and excretion processes (Taylor, 1991; Rekker, 1977; Hansch et al., 1995).The most popular molecular structure descriptor to signify this property, which is the logarithm of the partition coefficient, P (logP) between n-octanol and water:

$$P = [C]_{octanol} / [C]_{aqueous} \quad (1.4)$$

Here $[C]_{octanol}$ is the concentration of a solute in the lipid phase (n-octanol) and $[C]_{aqueous}$ is the concentration of the solute in the aqueous phase. Pharmaceutical molecules with P>1 that are hydrophobic or lipophilic compounds for which P<1 are hydrophilic in nature. The term "lipophilicity" refers to the affinity of a drug moiety or molecule for a lipophilic environment. It is usually determined by its distribution behavior in a biphasic system, either liquid-liquid (e.g., partition coefficient in n-octanol/water) or solid-liquid (retention on reversed-phase high-performance liquid chromatography (RP-HPLC)). AlogP is a lipophilicity descriptor, which is calculated from the log of partition coefficient by Ghose & Crippen's method (Ghose et al., 1998).

1.1.3.2.1.2. Hydrophobic substitution constant (π)

Hydrophobicity (Taylor, 1991; Rekker, 1977) is a phenomenon generally obtained by non-polar groups or molecules in an aqueous medium, which is caused by the tendency of water to discard non-polar molecules. The hydrophobicity of the molecules in the series can be illustrated on a relative scale with the hydrophobic substituent constant π. The value for the substituent X has been expressed through the following equation:

$$\log P_{R-X} = \log P_{R-H} + \pi_X \quad (1.5)$$

In this equation, $\log P_{R-X}$ and $\log P_{R-H}$ are partition coefficients of substituted and un-substituted compounds selectively. The π_X variable indicates the difference in lipophilicity, which is resultant by the replacement of H in RH by substituent X.

1.1.3.2.1.3. Hammett electronic constant (σ)

There are two types of electronic constants i.e. $\sigma_m$andσ_p.Due to the steric interaction with the reaction center, it is not appropriate for -o substituents (Hammett, 1935). It is denoted by the following equation:

$$\log k_X = p\sigma + \log k_h \qquad (1.6)$$

The reaction rate constants for the substituents x and h are indicated by k_x and k_h.The term σ is the electronic constant and ρ is fit to the set of analogs being studied. A positive value signifies withdrawal of electrons while a negative value signifies release of electrons.

1.1.3.2.1.4. Steric parameters

The size of the substituents is frequently associated with the biological activity of the individual molecules in a homologous series of pharmacological compounds (Rekker, 1977; Hansch and Leo, 1979). Bulky substituents may affect the intermolecular processes that lead to drug activity. A number of steric substituent constants have been used to quantitatively symbolize the steric parameter of structure of a drug.

1.1.3.2.2. Topological descriptors

Topological descriptors are computed based on molecular graphical symbolization, therefore they do not require any calculations of physicochemical qualities or the time-consuming calculations needed to produce the "quantum chemical descriptors". The 2D topological graph, which indicates where the atoms and connected bonds are located and it is necessary for the structural representation. The arrangement is based on the graphical theory, in which the atoms are represented by the vertices and the edges of the molecular structure, respectively (Roy and Das, 2014).

1.1.3.2.2.1. Wiener index (W)

It consists of the total of all the chemical bonds that exist between all of the heavy atom pairs in the molecule (Wiener, 1947). According to graph theory, it can be defined as the total length of all minimum pathways connecting every pair of vertices in a molecular graph. It can be expressed as follows:

$$W = \frac{1}{2}\sum_i \sum_j \delta_{ij} \qquad (1.8)$$

δ_{ij} is the shortest path distance between vertices i and j.

1.1.3.2.2.2. Zagreb index (Zagreb)

It is expressed by the sum of the squares of vertex degrees δ_i^2 (Gutman and Trinajstić, 1972).

$$Zagreb = \sum_i \delta_i^2 \qquad (1.9)$$

The molecular branching for an isomeric group of molecules is related to the Zagreb index.

1.1.3.2.2.3. Balaban index (J)

The Balaban index is expressed by the following equation (Balaban, 1982):

$$J = \frac{M}{\mu + 1}\sum_{all\ edges} \left(\delta_i \delta_j\right)^{-0.5} \qquad (1.10)$$

Where M = number of edges, μ = cyclomatic number, δ_i and δ_j are the vertex-distance degrees of the adjacent vertices, and δi is expressed by $\delta_i = \sum_{j=1}^{A} b_{ij}$, where A = number of vertices. This index is calculated from the distance matrix of the molecular graph.

1.1.3.2.2.4. Molecular connectivity indices

Using the atomic vertex degree in the H-suppressed molecular graph, one can compute the molecular connectivity indices. Greek letter "χ chi" is used to symbolize them.

1.1.3.2.2.4.1. Randic connectivity index

This is also known as the branching index or the connectivity index, and it was the first introduced connectivity index. It is defined as:

$$\chi_R = 1_\chi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij.} \left(\delta_i \delta_j\right)^{-0.5} \qquad (1.11)$$

Here, n=total number of vertices in the molecular graph, a_{ij}=adjacency matrix elements, and $\delta_i and \delta_j$ indicate vertex degree, the number of other vertices joined to vertex i and j

respectively. The term "edge connectivity" is defined as $(\delta_i\delta_j)^{-0.5}$ applied to every pair of adjacent edges or 1st order vertices. Though, it can also be applied for >2 adjacent vertices. This connectivity index is primarily associated with molecular branching.

1.1.3.2.2.4.2. Kier and Hall's connectivity index

It is based on Randic's principle, which has been developed by a general scheme for computing zero and higher-order connectivity descriptors. Kier and Hall's connectivity index is also known as molecular connectivity indices (Kier and Hall, 1977; 1986). The equations have been represented below for zero, first, second and higher order indices.

$$0\chi = \sum_{i=1}^{n} \delta_i^{-0.5} \quad (1.12) \qquad 1\chi = \sum_{b=1}^{B} (\delta_i\delta_j)_b^{-0.5} \quad (1.13)$$

$$2\chi = \sum_{k=1}^{2p} (\delta_i\delta_l\delta_j)_k^{-0.5} \quad (1.14) \qquad m\chi_t = \sum_{k=1}^{K} \left(\prod_{i=1}^{n} \delta_i\right)_k^{-0.5} \quad (1.15)$$

The last equation shows a general equation for higher order indices where *k* runs over the *m*-th order sub-graphs containing *n* vertices and *B* edges. The total number of appearing *m*-th order sub-graph is K. "χ" signifies the product of the values of simple vertex degrees (δ), whereas a subscript *t,* i.e., χt indicates a contiguous type of specific sub-graphs. The term ^{2}P refers to the second order index, i.e., 2χ denotes a path length of 2 with 3 vertices. Similarly, for higher order, ^{m}P will be inserted with the appropriate graph fragment type t.

1.1.4. Classification based on the type of employed methods

Quantitative structure-activity relationships (QSAR) have been used to develop relationships between the physicochemical properties of chemical substances and their biological activities in order to obtain a reliable statistical model for predicting the activities of new chemical entities. Sometimes QSAR is classified into the following two categories, such as Linear methods [Linear regression (LR), multiple linear regression (MLR), partial least-squares (PLS), and principal component analysis/regression (PCA/ PCR)] and Non-linear methods [Artificial neural networks (ANN), k-nearest neighbors (kNN), and Bayesian neural networks etc.] (Verma et al., 2010).

1.1.5. Methodology of QSAR

The four fundamental steps of QSAR analysis include - 1) Data preparation, 2) Data processing, 3) Data validation, and 4) Data interpretation (Roy et al., 2015). These steps have been defined in the following:

1.1.5.1. Data preparation

Initially, to maintain the uniformity of data, the endpoint is transformed to an obligatory unit (micromolar or millimolar). Then the requisite chemical structures are drawn by employing several popular software like Marvin Sketch, Chem Sketch, Chem Draw etc. or the structures can be downloaded from online public data bases like PubChem, ChemSpider etc. The energy minimization and conformational analysis are done if necessary. Afterward, the file containing the structures is subjected for descriptor calculation, and then data pretreatment can be performed to eliminate noisy data, constants etc. Finally, the descriptors comprising dissimilar variables and a single column of endpoint or response (activity/ property/ toxicity) are assembled in a single worksheet which is called 'QSAR data matrix'. An extra column representing the name or serial numbers of the molecules can be included for fast and easy identification of any molecule/compound.

1.1.5.2. Data processing: This segment contains three steps

a) Dataset division

A robust, sound, and well predictive QSAR model generation is the main objective of a QSAR analysis, which can be achieved through the proper division of the dataset into a training set (employed to develop a model) and a test set (employed for validation of the developed model). Nevertheless, the most comprehensible technique to select a training set is dependent on important chemical similarity and physicochemical descriptors. A large number of compounds is selected in the training set which is employed in model development. The algorithm is based on the principle is that a structurally similar molecule to the training set molecules can be predicted confidently because the model has learned the features that are shared by the training set molecules and is capable to search them in a new compound. The selection of the training and test sets will be in such a way that the test set compounds will fall within the structural domain of the training set molecules. Else, if there is an alteration in the structural features of the test set molecules, it will lead towards a bad prediction quality. Several methods of dataset division include – "k-means clustering", "Modified k-medoid

clustering", "Euclidean distance (ED) based method", "Kennard-Stone method", "Activity/Property based division", "Principal Component Analysis (PCA)", "Kohonen's Self-Organizing Map (SOM)", "D-optimal design", "Sphere exclusion" etc. (Leonard and Roy, 2006; Roy, 2007).

b) Feature selection

A feature selection process decreases the original feature space into a low-dimensional feature subspace by directly eliminating the noise and non-significant input features (Cai et al., 2018) which helps for better interpretability in QSAR modeling, thereby enhancing the predictive capacity of the model (Saxena and Prathipati, 2003). Usually, several feature selection algorithms can be integrated with one or more model development approaches under a similar interface so that the operator can choose the best possible combination of features and build models employing them concurrently. Several feature selection approaches employed in QSAR analysis includes "Stepwise Variable Selection", "Genetic Algorithm (GA)", "Best Subset Selection (BSS)", "Variable Subset Selection", and "Factor Analysis" etc. Usually, few of them are noticeably interrelated with the endpoint/response. Moreover, descriptors being intercorrelated have negative influences on some traits of QSAR study. A primary requirement of several statistical techniques is that the number of data points should be considerably higher than the number of descriptors/variables.

c) Model development

This step indicates that the best selected structural features are to be assembled in a single model using an explicit formalism. After finishing descriptor computation, evolution of QSAR model is accomplished by feature mapping method. The objective is to make a precise mathematical equation between the descriptors and the response/endpoint being studied. Several techniques like "Multiple Linear Regression (MLR)", "Partial Least Squares (PLS)" etc. are employed to develop regression-based models. While "Linear Discriminant Analysis (LDA)" is employed for the development of classification-based models. The variable selection techniques are conducted by statistical assessment of the resultant QSAR model built employing chosen variables as "Stepwise-MLR", "GA-MLR", "Genetic PLS (G/PLS)", PLS followed by discriminant analysis (PLS-DA). Lastly the best model is selected for auxiliary study based on several model validation metrics (Roy, 2007).

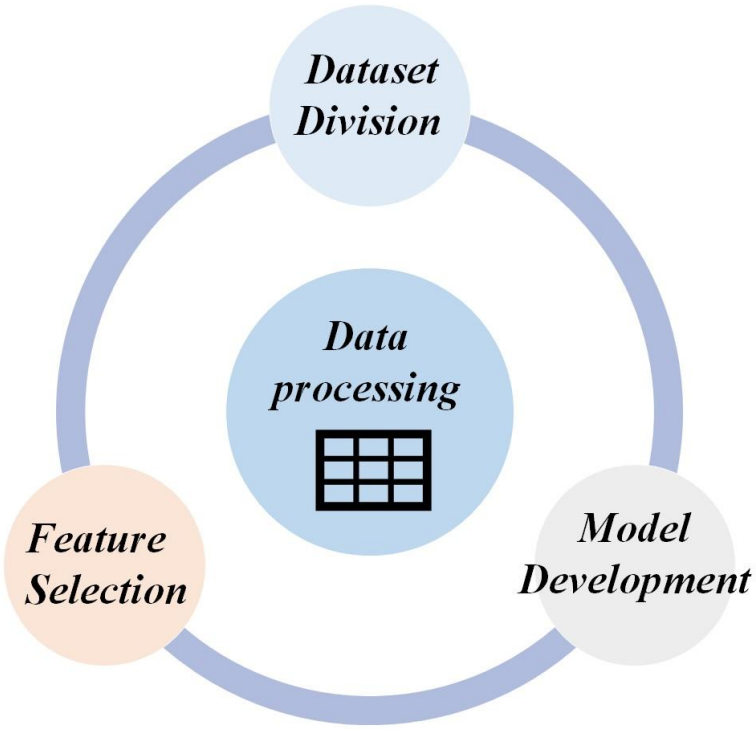


Fig.1.3. Schematic Representation of Data processing for QSAR Model Development

1.1.5.3 Model validation

Statistical validation is an inevitable part of QSAR modeling. As per the OECD principle 4, the QSAR model should be a good fit with the training data, sufficiently robust and externally predictive. Thus, to comply with the norm, we have computed several internal and external validation metrics for assessing the quality of developed models in this study. The internal validation metrics such as determination coefficient (R^2), adjusted R^2 (R^2_{Adj}), and cross-validated correlation coefficient (Q^2_{LOO}), mean absolute error of training set (MAE_{Train}) have been computed from the training data, whereas, external validation metrics like external correlation coefficient (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}), mean absolute error of test set (MAE_{train}) etc. have been computed with the test data. The determination coefficient (R^2 or R^2_{Adj}) and cross-validated correlation coefficient (Q^2_{LOO}) are the measurement parameters of goodness of fit and robustness, mean absolute errors (MAE_{Train} or MAE_{Test}) are the measures of error of predictions, and the external correlation coefficients (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}) are the measures of the external predictability of the model. The threshold limits for external validation metrics ($Q^2_{F1}/Q^2_{F2}/Q^2_{F3} \geq 0.5$) along with internal validation metrics ($R^2/R^2_{adj} \geq 0.6$, $Q^2_{LOO} \geq 0.5$) have been mentioned as per the criterion of Goldbaikh and Tropsha for defining the statistical quality of the regression-based models (Roy et al., 2015; Roy and Mitra, 2011).

1.1.5.4. Model interpretation

After the QSAR model development and validation, the molecular features are interpreted rigorously. The formation of the relationship between the structural attributes and the response variable aids in understanding the mechanism of action. Consequently, integrating the observation and experimental results from the developed and validated model, one can simply describe the behavioral exhibition of molecules. This information is valuable to assess the toxicity of chemical pollutants towards designing and evolving effective derivatives **(Roy et al., 2015).**

1.2. Environmental toxicity by diverse organic chemicals/pesticides

The most important routes for entrance of chemicals into the environment are via fecal matters and urine of the patients, liberation of drugs such as antibiotics with other pharmaceuticals used as growth regulators in livestock (Wollenberger et al., 2000) during excessive farming, from manufacture and leachate from landfill areas where unwanted, expired or spoiled drugs have been disposed of in household garbage, with household waste (Leclercq et al., 2009). Other sources like direct application in aquatic farming, run-off from the application of sewage sludge, manure run-off and manure on farmland as fertilizers, via landfill leaching finally, via hospital effluent (Jones et al., 2001). Studying the published publications and available scientific reports, it is shown that acute toxic effects of chemicals on various types of creatures in environment are due to their bio-accumulative, persistent and toxic effects. The rising concern over the discharge of active compounds used by humans or personal care products or veterinary medicines (Halling-Sørensen et al., 1998) into waste water, soil and air of our surrounding environment has encouraged the introduction of risk assessment guidelines in the European Union (EU) by the “European Medicines Evaluation Agency (EMEA)” and in the United States by the “Food and Drug Administration (FDA)” and also other countries.

1.2.1. Aquatic toxicology

“Aquatic toxicology is the study of the effects of manufactured chemicals and other anthropogenic and natural materials and activities on aquatic organisms at various levels of organization, from sub-cellular through individual organisms to communities and ecosystems” (Rand and Petrocelli, 1985). Aquatic toxicology is a interdisciplinary area which merges aquatic ecology, toxicology and aquatic chemistry (Rand and Petrocelli, 1985). This

area of study comprises marine water, freshwater and sediment environments. Common tests comprise of standardized acute and chronic toxicity tests 24–96 hours duration (acute test) to 7 days or longer (chronic tests). These tests evaluate endpoints i.e., reproduction, growth, survival, that are determined at each concentration gradient, together with a control test (US EPA, 2001). Generally employing chosen creatures with ecologically pertinent sensitivity to toxic chemicals and a well-recognised research context. These creatures are easy to obtain or culture in laboratory and easy to handle (US EPA, n.d.).

1.2.1.1. Aquatic toxicity test methods

Aquatic toxicology assays/tests are performed to afford quantitative and qualitative data on deleterious (adverse) impacts of a toxicant on aquatic organisms. Toxicity assays can be employed towards assessment of damaging potential to an aquatic environment and to produce a database to access the effects of toxicants. Aquatic toxicology assays can be done in the field or in a lab. Field assays usually denote exposure to various species whereas lab assays generally denote exposure to a single species. A dose–response relationship is frequently employed with a sigmoidal curve at a selected criteria or end-point to estimate the toxic effects (i.e. adverse effects or death to an organism). Concentration is plotted the x-axis while % inhibition/response on the y-axis (Rand and Petrocelli, 1985). The conditions for endpoints or effects being tested can include sublethal and lethal effects (Rand and Petrocelli, 1985). There are several classes of toxicity assays that can be studied on different test species. Various species vary in their vulnerability to chemicals, mostly due to the variations in genetic factors, accessibility, metabolic rate, dietary factors, excretion rate, sex, health, age and stress level of the organism. Some commonly used aquatic test species are listed below (US EPA, n.d.).

1) **Rainbow trout (*Oncorhynchus mykiss*):** It belongs to the family – Salmonidae. Nudds et al. published an article titled "Rainbow trout provide the first experimental evidence for adherence to a distinct Strouhal number during animal oscillatory propulsion." **(Nudds et al., 2014).**
2) **Lepomis (*Lepomis macrochirus*):** It belongs to the family – Centrarchidae. Lukenbach et al. reported that it plays a key role within the food chain of its freshwater ecosystem (Lukenbach et al.,2013).
3) **Fathead minnow (*Pimephales promelas*):** It belongs to cyprinidae family. It is used because a large number of offspring produced and its relative hardiness. Steroidal

chemicals increase the blood plasma vitellogenin levels in male fathead minnows. Vitellogenin levels are a sign to determine if compounds have any osteogenic activity to fish or not (Panter et al., 1998).

4) **Zebrafish:** Danio rerio belongs to Cyprinidae family and it is indigenous to south Asian lakes. The advantage of using Danio rerio embryos is that their high fertility and reproduction rate, their transparent property and their phenotypic properties can be easily identified (Hill et al., 2005). Also, some drugs and pollutants with small molecular structures can penetrate through the skin of zebra fish embryos and show their toxicities (Milan et al., 2003; McLeish et al., 2010). Also, the zebrafish and its embryos are valuable in the field of drug discovery and development due to its genetic resemblance with humans (Howe et al., 2013).

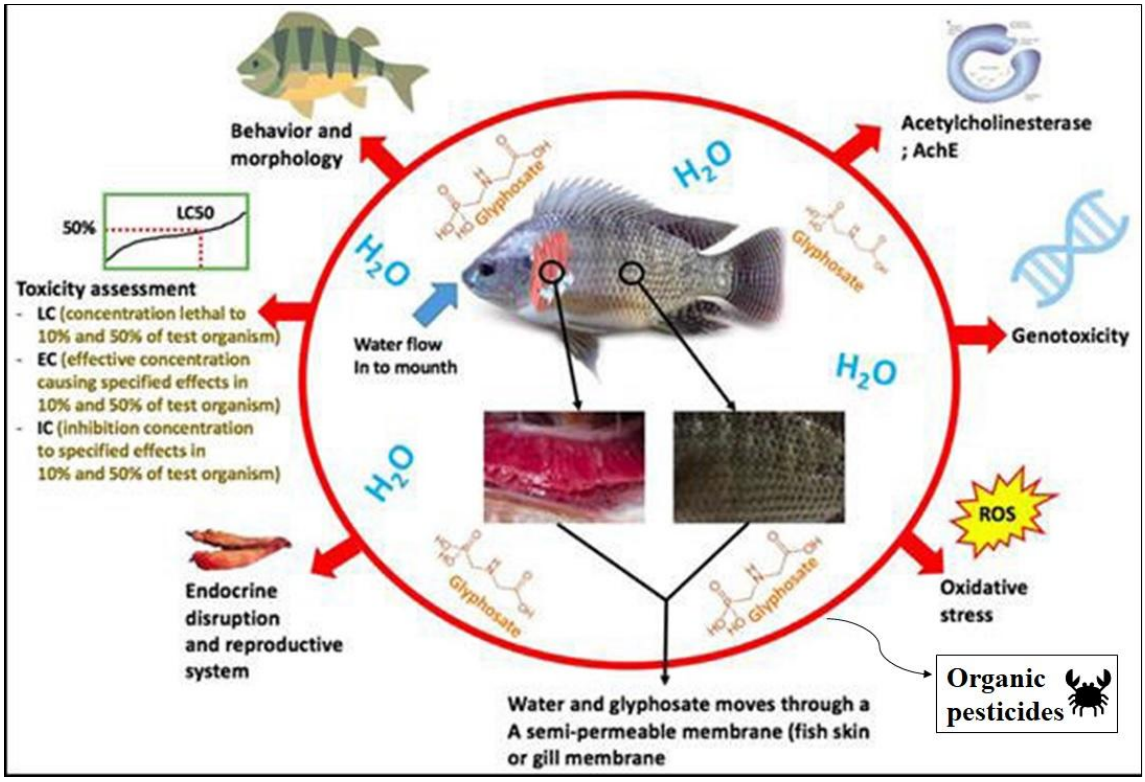


Fig.1.4. Aquatic toxicity of Fish Species

1.2.1.2. Definitions of the toxicity endpoints and some terminology

Median Lethal Concentration (LC50) – The concentration of a compound that destroys 50% of the test subjects throughout the experimental period.

Median Effective Concentration (EC50) – The concentration of a chemical that is expected to have one or more specified effects in 50% of the test subjects.

Half maximal inhibitory concentration (IC50) – The concentration of a compound that inhibits 50% of a specific biological or biochemical component of the test subjects throughout the experimental period.

Baseline toxicity – It denotes to narcosis which is suppression in biological activity by the toxic chemicals accumulated in the species.

Median Lethal Dose (LD50) – It is the amount of an ingested chemical that kills 50% of the test subjects.

Median Effective Dose (ED50) – It is the dose of a medication that shows a desired pharmacologic effect in 50% of the test subjects that takes the medication.

Bio magnification – "The method by which the concentration of a chemical in the tissues of an organism increases as it passes through several levels in the food web."

Application Factor (AF) – "An empirically derived "safe" concentration of a chemical."

Bio monitoring – The regular use of living species to analyze environmental alterations over time.

Therapeutic index (TI) **-** It is determined in animals as median lethal dose of a drug (LD50) divided by the median effective dose (ED50). It quantifies the relative safety of a drug.

1.3. Quantitative Read-across analysis

In the absence of experimental data, in silico approaches and models are being utilized more frequently to anticipate chemical attributes for hazard identification and hazard characterization. There are numerous in silico models that can be utilized separately or together. While these models have significant advantages for toxicologists, risk assessors, and the global scientific community, but. the absence of a standardized framework for the integration of in silico data can cause uncertainty and even conflicts between models and users, even for the same compounds. Numerous techniques for integrating in silico data have been put forth in this context, either on a statistical or case-specific basis.

Read-across constitutes another strategy for deriving reference points or points of departure for hazard characterization of untested chemicals, from the available experimental data for structurally-similar compounds, mostly using expert judgment (Benfenati et al.,2019). Read-across based predictions seem more suitable for tiny data sets because they are not a

rigorous statistical method. This is mostly (though not completely) due to the fact that read-across approaches, in contrast to the QSAR approach, require less empirical data to conduct trustworthy and accurate predictions utilized to fill in data gaps. A number of software systems have recently been created to assist specialists in this endeavor by providing a codified and organized method. A similar process might make it easier to integrate in silico model with read-across data.

1.3.1. Methodology of Quantitative Read-across algorithm

Chatterjee and co-workers reported a read-across methodology where they used Euclidean distance (ED), Gaussian kernel function (GK), and Laplacian kernel (LK) function for estimating the similarity between source and target compounds(Chatterjee et al., 2022). The ED-based similarity, GK and LK- functional similarity have been estimated by the mathematical formula represented in the following equations 1, 2, and 3, respectively:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1.16)$$

where, Euclidean distance between two points X and Y has been represented yd(X, Y), whereas, X_i is the i^{th} descriptor of compound X, and Y_i is the i^{th} descriptor of compound Y, where i = 1, 2, 3, ..., n. The ED-based similarity has been computed by subtracting the scaled (0-1) distance from 1 because distance is the reciprocal of similarity.

$$GK(X,Y) = f = e^{-\|X_i - Y_i\|/2\sigma^2} \quad (1.17)$$

where, Gaussian kernel function similarity *(f)* between two compounds X and Y, which has been represented by GK(X, Y)and $\| X_i - Y_i \|^2$ is the Euclidean norm or L^2 norm, which can be measured by squaring the Euclidean distance, and σ indicates the width of Gaussian kernel function and can never become zero.

$$LK(X,Y) = \hat{e} = e^{(-y\|X-Y\|_1)} \quad (1.18)$$

where, LK(X, Y) is the Laplacian kernel function similarity (κ) between two compounds X and Y; γ signifies a non-zero and positive number which determines the performance of the Laplacian kernel function.

1.4. Quantitative Read-across structure activity (q-RASAR) analysis

The q-RASAR model is a combination of read-across prediction and QSAR approach where the read-across derived non-linear prediction function, similarity, error, and concordance measures (RASAR descriptors)may be combined with the important structural and physicochemical features to build the linear QSAR model (Luechtefeld et al., 2018)(Banerjee and Roy, 2022). A q-RASAR model may be developed from both discrete and continuous data to form both classification and regression-based models.

1.4.1. Methodology of Quantitative Read-across structure activity algorithm

To calculate the RASAR descriptors, optimization of the read-across algorithm is a prior step. We have used the previous training and test sets selected by QSAR analysis with the identified 2D descriptors for determining the optimum read-across method. The training and test sets have been used as the source and target compounds, respectively, for q-RASAR analysis. However, the RASAR descriptor calculation for training set chemicals is a little bit tricky because the training set has been simultaneously employed as the source and target compounds. Therefore, to avoid the chance of overfitting, the training RASAR descriptor calculation has been performed using the "leave-same-out" algorithm, where the same training compound with a query compound is identified and eliminated before weighted averaging.

1.4.2. Descriptions of q-RASAR descriptors employed in q-RASAR analysis:

RA function: The RA function is a read-across-derived prediction function obtained by the tool "RASAR-Desc-Calc-v2.0" employing the read-across based similarity estimation algorithm. It contains information from all of the structural descriptors and process variables and functions similarly as a composite variable (Banerjee and Roy, 2022).

Pos.Avg.Sim.: It measures the average similarity of positive close source compounds (with respect to the mean of the source compounds) for a query compound.

SD Similarity: It indicates standard deviation of the similarity values of the selected close source compounds for each query compound.

SD Activity: It definesstandard deviation of the (observed) activity values of the selected close source compounds for each query compound.

MaxPos: It indicates maximum similarity level to the positive close source compounds (based on source set observed mean).

MaxNeg: It indicates maximum similarity level to the negative close source set compounds (based on source set observed mean).

AbsDif: It defines absolute difference between MaxPos and MaxNeg.

CV_activity: It signifies coefficient of variation of the response.

CHAPTER - 2

PRESENT WORK

"*Not only is the risk of the pesticide less, but the consequences of the exposures would be very different.*"

-Bob Peterson

2. PRESENT WORK

The experimental determination of environmental parameters (e.g., toxic effects, bio concentration, soil sorption, biotransformation and biodegradation) of commercially available chemicals is a costly and time consuming process. Since there is large number of chemicals presently in common applications and new chemicals are registered at a very high rate, it is evident that our human and material resources are inadequate to gather experimentally even basic information on environmental fate and effects for all these chemicals. Therefore, it is crucial to create quantitative models that could quickly and accurately predict the environmental behaviour of a vast array of pesticides. Quantitative structure-activity relationship (QSAR) modelling, which forecasts the biological activity from chemical structure, is a key component of the chemistry-based predictive toxicology method. Such techniques have proven their effectiveness when employed to well-defined toxicity end points of chemical space.

Similar to environmental toxicity, predicting the toxicity or side effects of chemicals or medicines is an important step during the early phases of drug development to prevent significant wastage at later stages. In silico toxicity prediction of pharmaceuticals provides a method to screen initially the candidate molecules for prospective synthesis and development, even though experimental toxicity assessment using animal models is unavoidable for drug candidates at an advanced stage of drug research and development.

Pharmaceutical industries perform toxicity screening in earlier stages for drug induced toxicity due to following reasons: (1) Huge number of compounds needed for the in vivo studies, (2) lacking suitable high-throughput in vitro assays and (3) insufficiency of animal models and in vitro methods to extrapolate correctly human toxicities. To resolve these inconveniences, in silico (computational) toxicity prediction tools offer as one of the key approaches to explore the potentially toxic effects of drugs on humans even before the compounds are being physically marketed (Kahn et al., 2007).

In this context, the emergence of eco-toxicity, aquatic toxicity, environmental fate parameters of various chemicals, drug-induced toxicity, and health effects of organic pesticides has prompted various government and non-government chemical regulatory organizations to recommend the application of theoretical, non-animal, non-experimental, alternatives, and particularly, in silico methods to provide information on the fundamental physicochemical

and fate properties of chemicals as well as their ecological and human health effects before they come into the market for public consumption. In order to assess these features, advanced software for computer-aided toxicity prediction has been associated to make the application of relationship between chemical structure and biological activity (toxicity). Therefore, as a non-experimental method, the application of quantitative structure–toxicity relationship (QSTR) as well as quantitative structure–property relationship (QSPR) is very important towards reduction of time and cost involved in a new chemical and drug design, development and discovery process.

In the current study, primarily two-dimensional (2D) approaches have been used to create the QSAR models. The majority of the descriptors used in this work are classified as two-dimensional (2D) descriptors, which are highly effective in avoiding the computational complexity of conformational analysis, energy minimization, and alignment issues as well as these descriptors are very much reliable to encode the structural features due to their simplicity, reproducibility, and better interpretability. We have also employed the two dimensional (2D) structural aspects of the chemicals with similarity and associated parameters obtained from the read across prediction tool for enhancement the quality of toxicity prediction through the development of q-RASAR/q-RASPR models. In contrast to read-across, RASAR provides supervised mathematical models from which one can extract knowledge about the relevance of modeled features and their contribution towards property prediction. Therefore, the benefits and advantages of combinatorial modeling make this methodology more efficient in the field of cheminformatics. Then the developed models were rigorously validated using various stringent internal and external validation tests as recommended by OECD guidelines to show goodness-of-fit, robustness, predictive capacity and soundness of these models.

Two different case studies were performed for the present work:

2.1. Case study 1

Pesticides are now a necessary component of the modern agriculture system. Modern pesticides are becoming more important in preventing disease transmission, defending crops and goods from pests, and controlling the proliferation of undesirable target organisms due to the rapid improvement of agricultural research and technology. In this study, the retention

time of chemicals/pesticides detected from reverse-phase HPLC can be considered as a potential indicator of cellular toxicity. The retention time of chemicals depends upon their own lipophilicity, where the polarity of the mobile phase is constant (Soares, J. et. al) .Lipophilicity bears a directly proportional relationship with toxicity (more lipophilic chemicals can easily pass through the lipophilic plasma membrane and also stay in the lipid cells for a long time) as well. Therefore, we can extrapolate the proportional relationship between toxicity and retention time also (Hongmao, S. et. al). This assumption can be exemplified by the following observations: silafluofen, pyridalyl, and ivermectin (with tR values of 11.05, 10.74, and 10.71 h, respectively, and X log P values of 8.2, 7.59, 121 and 5.83, respectively). These compounds are reported to be very toxic to aquatic life, causing danger in specific target organ toxicity as well as reproductive toxicity. Again, compounds, like cyromazine, ethylene thiourea, and maleic hydrazide, have low tR values (1.14, 1.08, and 1.06 h, respectively) and lower X log P values (−0.06, −0.66, and −0.84, respectively), and they are relatively less toxic than the previous compounds. We have attempted q-RASPR modeling in the present work for pesticide residues to detect toxicity levels in terms of retention time (log t_R) as a result of its simple, cost-effective, fast, and reproducible nature. The entire computational research is devoid of animal experimentation, which complies with the 3R strategy (replacement, reduction, and refinement of animal experimentation) of the REACH legislation.

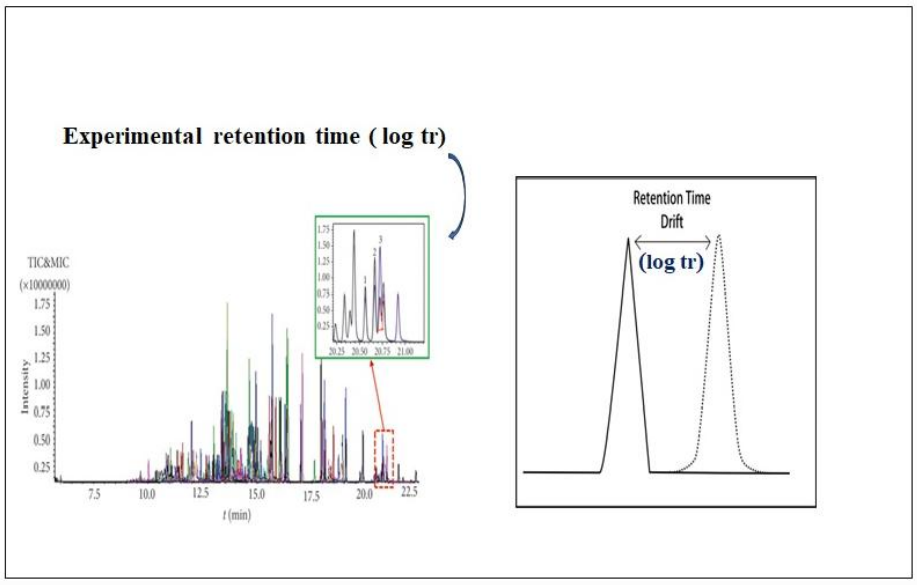


Fig. 2.1. Detection of experimental retention time (**log t_r)** of pesticide residues.

The current study describes the development of a q-RASPR model using experimental retention time data (log t_R) in the HPLC experiments of 823 environmentally significant pesticide residues collected from a large compound database. The 2D structures of the aforementioned chemicals were retrieved (.sdf format) from the PubChem database and were cross-checked with another popular chemical database ChemSpider. After manual checking of all of the structures, chemical curation was performed using a chemical curation workflow in the KNIME platform to remove inorganic salts, ions, radicals, repetitive compounds, and mixtures from the data set. The structures of the molecules were drawn in Marvin Sketch software (version 14.10.27) (https://www.chemaxon.com) with proper aromatization and addition of explicit hydrogen atoms. Only 2D descriptors were computed, including connectivity indices, E-state indices, 2D atom pairs, constitutional, functional, molecular property-based descriptors, ring descriptors, atom-centered fragments and extended topochemical atom (ETA) indices by employing Alvadesc software. To model the retention time (log t_R) end point, 0D−2D descriptors have been used along with the read-across-derived similarity descriptors. In the present work, the sorted activity-based division technique has been used by employing (75-25) division of data points as training and test set for the development of q-RASPR model. Selection of descriptors was done by the genetic algorithm (GA) approach and the selected pool of descriptors was subjected to Best Subset Selection (BSS) v2.1" for the development of QSPR model. Then, we have followed the read-across optimization to develop the final q-RASPR model by the selection of RASAR descriptors along with previously identified 2D descriptors. Finally, the developed q-RASPR model was rigorously validated by various internal and external validation metrics as recommended by the Organization for Economic Co-operation and Development (OECD). Further validation was also performed by utilizing (process and model) Y-randomization tests. Additionally Regression coefficients plot, VIP, AD, Loading, Score plots etc. are analyzed to confirm the statistical quality of the presented models.

2.2. Case study 2

Aquatic organisms are mostly affected by pesticides and their residues due to the easy passage of agricultural waste into water bodies through sewage systems. Human health can also be at threat due to the bioaccumulation and bio-magnification of pesticides via aquatic species. Therefore, it is essential to identify the potential risk of pesticides to human health and environmental species. In particular, European Pesticide Regulation No. 1107/2009 (EC

Regulation 1107/2009) calls for the use of fish as model species to assess aquatic toxicity in water contamination. Prior to releasing a pesticide onto the market, registrants must evaluate the environmental safety of the substance and any associated residues. Fish have thus been used frequently as model species to evaluate the aquatic toxicity of industrial pollutants. According to OECD test no. 203, a typical animal test is used to assess acute fish toxicity and produces LC50 (lethal concentration for population values), although there is an increasing need to reduce or replace animal studies for regulatory purposes (Tunkel et al., 2005). Additionally, the experimental process for determining fish acute toxicity requires a lot of time and effort (Yu and Zeng, 2022). In silico innovative approach methodologies may be used to get beyond these experiment-related complications. Acute toxicity is usually assessed with short-term exposure of fish to a series of concentrations of chemical doses and the concentration that is lethal to 50% of the test fish is calculated and expressed as LC50 value. Thus, the experimental [Log (1/LC50)] value is taken as the potential indicator of acute aquatic toxicity in this study (Braunbeck et al., 2005).

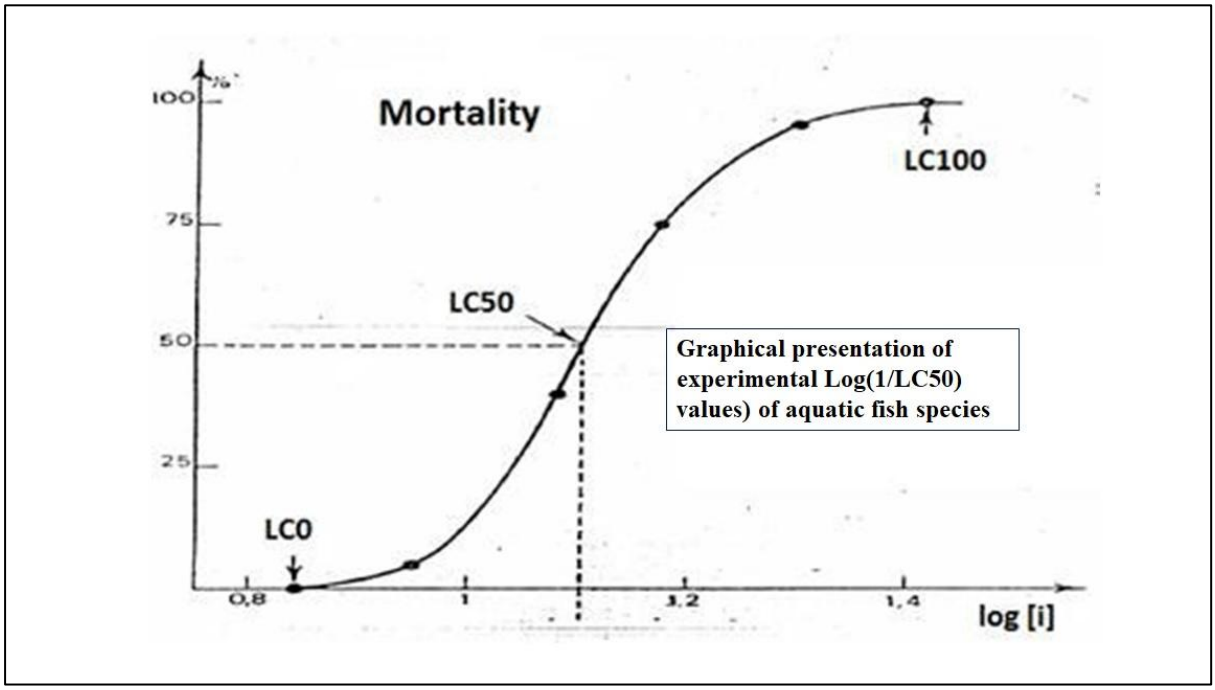


Fig. 2.2. Detection of experimentally Log (1/LC50) values of aquatic fish species

The present study describes the development of q-RASAR models using experimental data (Log 1/LC50) of organic pesticides to various fish species, including Rainbow trout (RT: *Oncorhynchus mykiss*: 715 data points), Lepomis (LP: *Lepomis macrochirus*: 136 data points), and Miscellaneous fish species (Misc: *Pimephales promelas*, *Brachydanio rerio*: 226

data points). After the manual checking of all the structures, chemical curation was performed manually by excluding the compounds having an unknown chemical structure, inorganic compounds, metal complexes, salts containing organic polyatomic counterions, mixtures, and substances of unknown or variable composition (UVCB) for modelling purposes. Furthermore, in RT (rainbow trout) and others datasets, we also excluded the compounds having higher hi values based on the leverage calculation for modeling (Adeniji et al., 2019). Finally, we have selected 715 data points for Rainbow trout (RT), 136 data points for Lepomis (LP) and 226 data points for Miscellaneous fish species to develop the models in this study. We have associated the 0D-2D descriptors to encode the structural features due to their simplicity, reproducibility, and better interpretability. In this work, we have employed the division of training set (~75%) for model development and test set (~25%) for model validation purpose using Euclidean distance-based division for all the datasets. Features selection was done by stepwise selection approach to extract the important features based on partial F statistic. Then the selected pool of features was subjected to Best Subset Selection (BSS) v2.1" for documentation of the best subset or combination of descriptors. Then we have followed "Read-across" optimization to identify the best functional similarity based approach for calculating the RASAR descriptors. Finally, we have employed "Partial Least Squares (PLS)" regression to develop the q-RASAR models for all the datasets. Then the models were rigorously validated by using several internal and external statistical metrics. Additionally Regression coefficients plots, VIP, Applicability Domain (AD), Loading and Y-Randomization of the models were also determined.

CHAPTER - 3

MATERIALS AND METHODS

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

– **Marie Curie** (Nobel laureate physicist and chemist)

3. MATERIALS AND METHODS

The present dissertation was performed with the objective to implement a transparent methodological framework for the development of predictive q-RASAR/q-RASAPR models employing simply interpretable two dimensional (2D) molecular descriptors along with identified RASAR descriptors based on Read-across optimization. We have attempted to maintain explicitness towards computation of the descriptors, thinning of the entire matrix of variables, selection of important features and also judgment of predictive ability and robustness of the models. Therefore, we have reported in details with their property or toxicity data and methods applied to accomplish this computational (in silico) study namely, q-RASAR modelling of retention time and aquatic toxicity endpoints. This section has been split into the following segments:

1) Details of the datasets including SMILES along with the experimental/observed end points.

2) General principle of methodologies employed to develop the q-RASAR/q-RASPR models along with study wise precise explanation of methodologies exploited in each case study.

3.1. Details of the datasets including SMILES along with the experimental/observed end points (log t_R)

3.1.1. Dataset 1 (case study 1)

The retention time (logt_R) of pesticidal compounds in a reverse-phase HPLC analysis has a direct relationship with lipophilicity which could be related to the ecotoxicity potential of the compounds. This study describes the development of a quantitative read-across structure-property relationship (q-RASPR) model using experimental retention time data (logt_R) in the HPLC experiments of 823 environmentally significant pesticide residues collected from a large compound database. The chemical name, canonical smiles along with retention time (logt_R) end points of 823 pesticide residues have been tabulated in the following **Table 3.1.**

Table 3.1. Chemical name, canonical smiles and experimental retention time ($\log t_R$) of 823 pesticides. (* Test set compounds)

ID	Name	Canonical smiles	$\log t_R$
1	2,4-Dichlorophenoxyacetic acid	OC(=O)COc1ccc(cc1Cl)Cl	6.75
2*	2,4-Dichlorophenoxybutyric acid	CCC(Oc1ccc(cc1Cl)Cl)C(O)=O	8.14
3*	2,6-Dichlorobenzamide	NC(=O)c1c(cccc1Cl)Cl	3.53
4	2-aminobenzimidazole	Nc1nc2ccccc2[nH]1	2.55
5*	2-Hydroxypropazine	CC(C)NC1=NC(=O)N=C(NC(C)C)N1	4.95
6	2-Hydroxypropoxycarbazone	COC(=O)c1ccccc1S(=O)(=O)NC(=O)N1N=C(OCC(C)O)N(C)C1=O	4.33
7	2-Imidazolidinethione	S=C1NCCN1	1.08
8	2-Mercaptobenzothiazole	S=C1Nc2ccccc2S1	5.87
9	2-Naphthyloxy acetic acid	OC(=O)COc1ccc2ccccc2c1	6.58
10	2-Pivaloyl-1,3-indandione	CC(C)(C)C(=O)C1C(=O)c2ccccc2C1=O	9.08
11*	3,4,5-Trimethylphenyl methyl carbamate	CNC(=O)Oc1cc(c(c(c1)C)C)C	7.08
12	3-Hydroxycarbofuran	CNC(=O)Oc1cccc2c1OC(C)(C)C2O	4.55
13*	3-Hydroxycycloate cis/trans	CCSC(=O)N(CC)C1CCCC(O)C1	7.055

14	3-Ketocarbofuran	CNC(=O)Oc1cccc2c1OC(C)(C)C2=O	5.44
15	4-Hydroxycycloate cis/trans	CCSC(=O)N(CC)C1CCC(O)CC1	7.02
16	5-Hydroxythiabendazol	Oc1ccc2nc([nH]c2c1)-c1c[s]cn1	3.17
17	Abamectin B1a	CCC(C)C1OC2(CC3CC(CC=C(C)C(OC4CC(OC)C(OC5CC(OC)C(O)C(C)O5)C(C)O4)C(C)C=CC=C4COC5C(O)C(=CC(C(=O)O3)C45O)C)O2)C=CC1C	10.1
18	Acephate	COP(=O)(NC(C)=O)SC	1.87
19*	Acetamiprid	CN(Cc1ccc(nc1)Cl)C(C)=NC#N	4.65
20	Acetamiprid metabolite 6-chloronicotinic acid	OC(=O)c1ccc(nc1)Cl	4.11
21	Acetamiprid metabolite IM-1-2	CN(Cc1ccc(nc1)Cl)C(C)=NC(N)=O	2.52
22	Acetamiprid metabolite IM-1-4	CNCc1ccc(nc1)Cl	1.37
23*	Acetamiprid metabolite IM-2-1	CC(NCc1ccc(nc1)Cl)=NC#N	4.54
24	Acetamiprid metabolite IM-2-2	CC(NCc1ccc(nc1)Cl)=NC(N)=O	1.87
25	Acetochlor	CCOCN(C(=O)CCl)c1c(cccc1CC)C	8.24
26	Acetochlor metabolite EMA	CCc1cccc(c1N)C	5.84
27	Acetochlor metabolite tert sulfinyl acetic acid	CCOCN(C(=O)CS(=O)CC(O)=O)c1c(cccc1CC)C	6.42

28*	Acibenzolar-S-methyl	CSC(=O)c1cccc2nn[s]c12	7.65
29	Acifluorfen amine methyl ester LS-82-5282	COC(=O)c1cc(ccc1N)Oc1ccc(cc1Cl)C(F)(F)F	9.4
30	Acifluorfen amine	Nc1ccc(cc1C(O)=O)Oc1ccc(cc1Cl)C(F)(F)F	8.72
31	Acifluorfen methyl ester	COC(=O)c1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)C(F)(F)F	8.94
32	Aclonifen	Nc1c(c(ccc1N(=O)=O)Oc1ccccc1)Cl	8.5
33	Acrinathrin	CC1(C)C(C=CC(=O)OC(C(F)(F)F)C(F)(F)F)C1C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	10.05
34	Akton	CCOP(=S)(OCC)OC(=CCl)c1cc(ccc1Cl)Cl	9.51
35	Alachlor	CCc1cccc(c1N(COC)C(=O)CCl)CC	8.24
36	Alachlor hydroxy	CCc1cccc(c1N(COC)C(=O)CO)CC	7.34
37	Alachlor metabolite 2,6-diethylaniline	CCc1cccc(c1N)CC	7.17
38*	Allethrin	CC(C)=CC1C(C(=O)OC2CC(=O)C(=C2C)CC=C)C1(C)C	9.455
39	Allidochlor	ClCC(=O)N(CC=C)CC=C	5.63
40	Alloxydim	CCCC(=NOCC=C)C1=C(O)C(C(=O)OC)C(C)(C)CC1=O	8.36
41*	Ametoctradin	CCCCCCCCc1c([n]2ncnc2nc1CC)N	9.05

42	Ametryn	CCNc1nc(nc(n1)SC)NC(C)C	7.3
43	Amicarbazone	CC(C)C1=NN(C(=O)NC(C)(C)C)C(=O)N1N	6.05
44	Amidosulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)N(C)S(C)(=O)=O)OC	6.41
45	Aminocarb	CNC(=O)Oc1ccc(c(c1)C)N(C)C	2.33
46*	Amiphos	COP(=S)(OC)SCCNC(C)=O	5.07
47*	Amisulbrom	CN(C)S(=O)(=O)n1cnc(n1)S(=O)(=O)n1c(c(c2ccc(cc12)F)Br)C	9.15
48*	Amitraz metabolite dimethylaniline	Cc1ccc(c(c1)C)N	3.74
49	Ancymidol	COc1ccc(cc1)C(O)(C1CC1)c1cncnc1	6.33
50	Anilofos	COP(=S)(OC)SCC(=O)N(C(C)C)c1ccc(cc1)Cl	8.58
51	Antimycin A	CCCCCCC1C(OC(=O)CC(C)C)C(C)OC(=O)C(NC(=O)c2cccc(c2O)NC=O)C(C)OC1=O	9.9
52	Aramite	CC(COc1ccc(cc1)C(C)(C)C)OS(=O)OCCCl	9.38
53	Aspon	CCCOP(=S)(OCCC)OP(=S)(OCCC)OCCC	9.75
54	Asulam	COC(=O)NS(=O)(=O)c1ccc(cc1)N	2.38
55	Atraton	CCNc1nc(nc(n1)OC)NC(C)C	6.1

56*	Atrazine	CCNc1nc(nc(n1)NC(C)C)Cl	6.93
57*	Atrazine-desethyl	CC(C)Nc1nc(nc(n1)Cl)N	5
58	Azaconazole	Clc1ccc(c(c1)Cl)C1(Cn2cncn2)OCCO1	7.23
59	Azadirachtin	COC(=O)C1(O)OCC23C(CC(OC(C)=O)C4(COC(C(O)C(C)(C12)C12OC2(C)C2CC1OC1OC=CC21O)C34)C(=O)OC)OC(=O)C(C)=CC	6.65
60	Azafenidin	Clc1cc(c(cc1OCC#C)N1N=C2CCCCN2C1=O)Cl	7.08
61*	Azamethiphos	COP(=O)(OC)SCN1C(=O)Oc2cc(cnc12)Cl	5.98
62	Azimsulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1c(cnn1C)-c1nnn(n1)C)OC	6.84
63	Azinphos-ethyl	CCOP(=S)(OCC)SCN1N=Nc2ccccc2C1=O	8.06
64	Azinphos-methyl	COP(=S)(OC)SCN1N=Nc2ccccc2C1=O	7.29
65	Aziprotryne	CSc1nc(nc(n1)N=[N+]=[N-])NC(C)C	7.8
66	Azoxystrobin	COC=C(C(=O)OC)c1ccccc1Oc1cc(ncn1)Oc1ccccc1C#N	7.355
67	Azoxystrobin metabolite R234886	COC=C(C(O)=O)c1ccccc1Oc1cc(ncn1)Oc1ccccc1C#N	6.96
68	Azoxystrobin metabolite R402173	OC(=O)c1ccccc1Oc1cc(ncn1)Oc1ccccc1C#N	6.87
69	Barban	ClCC#CCOC(=O)Nc1cccc(c1)Cl	7.84

70*	Beflubutamid	CCC(Oc1ccc(c(c1)C(F)(F)F)C(=O)NCc1ccccc1	8.54
71	Benalaxyl	COC(=O)C(C)N(C(=O)Cc1ccccc1)c1c(cccc1C)C	8.65
72	Benazolin	OC(=O)CN1C(=O)Sc2cccc(c12)Cl	5.65
73	Bendiocarb	CNC(=O)Oc1cccc2c1OC(C)(C)O2	6.22
74	Benfluralin	CCCCN(CC)c1c(cc(cc1N(=O)=O)C(F)(F)F)N(=O)=O	9.76
75	Benfuracarb	CCOC(=O)CCN(SN(C)C(=O)Oc1cccc2c1OC(C)(C)C2)C(C)C	9.19
76	Benfuresate	CCS(=O)(=O)Oc1ccc2c(c1)C(C)(C)CO2	8.43
77*	Benodanil	Ic1ccccc1C(=O)Nc1ccccc1	6.83
78	Benoxacor	CC1COc2ccccc2N1C(=O)C(Cl)Cl	7.38
79	Bensulfuron-methyl	COC(=O)c1ccccc1CS(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	7.3
80	Bensulide	CC(C)OP(=S)(OC(C)C)SCCNS(=O)(=O)c1ccccc1	8.4
81	Bensultap	CN(C)C(CSS(=O)(=O)c1ccccc1)CSS(=O)(=O)c1ccccc1	8.31
82	Benthiavalicarb-isopropyl	CC(C)OC(=O)NC(C(C)C)C(=O)NC(C)c1nc2ccc(cc2[s]1)F	7.85
83	Benzadox	OC(=O)CONC(=O)c1ccccc1	3.2

84	Benzanilide	O=C(Nc1ccccc1)c1ccccc1	6.41
85	Benzobicyclon	CS(=O)(=O)c1ccc(c(c1)Cl)C(=O)C1=C(Sc2ccccc2)C2CCC(C2)C1=O	7.81
86	Benzofenap	Cc1ccc(cc1)C(=O)COc1c(c(nn1C)C)C(=O)c1ccc(c(c1Cl)C)Cl	9.14
87*	Benzoguanamine	Nc1nc(nc(n1)-c1ccccc1)N	3.83
88	Benzovindiflupyr metabolite SYN-546039	Cn1cc(c(n1)C(F)F)C(=O)Nc1cccc2c1C1CC(O)C2C1=C(Cl)Cl	6.77
89	Benzovinidiflupyr	Cn1cc(c(n1)C(F)F)C(=O)Nc1cccc2c1C1CCC2C1=C(Cl)Cl	8.62
90	Benzoximate	CCON=C(OC(=O)c1ccccc1)c1c(ccc(c1OC)Cl)OC	8.82
91*	Benzoylprop-ethyl	CCOC(=O)C(C)N(C(=O)c1ccccc1)c1ccc(c(c1)Cl)Cl	8.72
92	Bicyclopyrone metabolite SYN-545910	OCCOCc1nc(ccc1C(O)=O)C(F)(F)F	4.31
93	Bicyclopyrone	COCCOCc1nc(ccc1C(O)=C1C(=O)C2CCC(C2)C1=O)C(F)(F)F	7.17
94	Bifenazate	COc1ccc(cc1NNC(=O)OC(C)C)-c1ccccc1	8.04
95	Bifenox	COC(=O)c1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)Cl	8.9
96*	Bifenthrin	Cc1c(cccc1-c1ccccc1)COC(=O)C1C(C=C(Cl)C(F)(F)F)C1(C)C	10.56
97	Bioresmethrin	CC(C)=CC1C(C(=O)OCc2coc(c2)Cc2ccccc2)C1(C)C	10.165

98*	Bispyribac	COc1cc(nc(n1)Oc1cccc(c1C(O)=O)Oc1nc(cc(n1)OC)OC)OC	7.58
99*	Bistrifluron	Fc1cccc(c1C(=O)NC(=O)Nc1cc(cc(c1Cl)C(F)(F)F)C(F)(F)F)F	9.72
100	Bitertanol	CC(C)(C)C(O)C(Oc1ccc(cc1)-c1ccccc1)n1cncn1	8.8
101*	Bixafen	Cn1cc(c(n1)C(F)F)C(=O)Nc1ccc(cc1-c1ccc(c(c1)Cl)Cl)F	8.42
102	Boscalid	Clc1ccc(cc1)-c1ccccc1NC(=O)c1cccnc1Cl	7.69
103	Brodifacoum	OC1=C(C2CC(Cc3ccccc23)c2ccc(cc2)-c2ccc(cc2)Br)C(=O)Oc2ccccc12	10.02
104	Bromacil	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	6.16
105*	Bromadiolone	OC(CC(c1ccccc1)C1=C(O)c2ccccc2OC1=O)c1ccc(cc1)-c1ccc(cc1)Br	9.13
106*	Bromfenvinphos	CCOP(=O)(OCC)OC(=CBr)c1ccc(cc1Cl)Cl	8.76
107	Bromfenvinphos-methyl	COP(=O)(OC)OC(=CBr)c1ccc(cc1Cl)Cl	8.12
108	Bromobutide	CC(C)(C)C(Br)C(=O)NC(C)(C)c1ccccc1	8.2
109	Bromophos-ethyl	CCOP(=S)(OCC)Oc1cc(c(cc1Cl)Br)Cl	10.02
110	Bromophos	COP(=S)(OC)Oc1cc(c(cc1Cl)Br)Cl	9.54
111	Bromuconazole cis/trans	Clc1ccc(c(c1)Cl)C1(CC(Br)CO1)Cn1cncn1	8.23

112*	Bupirimate	CCCCc1c(nc(nc1OS(=O)(=O)N(C)C)NCC)C	8.27
113	Buprofezin	CC(C)N1C(=O)N(CSC1=NC(C)(C)C)c1ccccc1	9.33
114	Butachlor	CCCCOCN(C(=O)CCl)c1c(cccc1CC)CC	9.46
115	Butafenacil	CN1C(=O)N(C(=O)C=C1C(F)(F)F)c1ccc(c(c1)C(=O)OC(C)(C)C(=O)OCC=C)Cl	8.04
116	Butamifos	CCOP(=S)(NC(C)CC)Oc1cc(ccc1N(=O)=O)C	8.78
117*	Butamifos oxon	CCOP(=O)(NC(C)CC)Oc1cc(ccc1N(=O)=O)C	7.93
118	Butralin	CCC(C)Nc1c(cc(cc1N(=O)=O)C(C)(C)C)N(=O)=O	9.79
119	Buturon	CC(C#C)N(C)C(=O)Nc1ccc(cc1)Cl	7.01
120	Butylate	CCSC(=O)N(CC(C)C)CC(C)C	9.25
121*	Cadusafos	CCOP(=O)(SC(C)CC)SC(C)CC	9
122*	Cafenstrole	CCN(CC)C(=O)n1cnc(n1)S(=O)(=O)c1c(cc(cc1C)C)C	7.95
123*	Captafol	ClC(Cl)C(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	8.03
124	Carbanilide	O=C(Nc1ccccc1)Nc1ccccc1	6.72
125	Carbanolate	CNC(=O)Oc1cc(c(cc1Cl)C)C	7.08

126	Carbaryl	CNC(=O)Oc1cccc2ccccc12	6.45
127	Carbendazim	COC(=O)Nc1nc2ccccc2[nH]1	3.42
128	Carbetamide	CCNC(=O)C(C)OC(=O)Nc1ccccc1	5.79
129*	Carbofuran	CNC(=O)Oc1cccc2c1OC(C)(C)C2	6.24
130*	Carbofuran phenol	CC1(C)Cc2cccc(c2O1)O	8.47
131	Carbophenothion	CCOP(=S)(OCC)SCSc1ccc(cc1)Cl	9.69
132	Carbophenothion oxon sulfone	CCOP(=O)(OCC)SCS(=O)(=O)c1ccc(cc1)Cl	7.25
133	Carbophenothion sulfoxide	CCOP(=S)(OCC)SCS(=O)c1ccc(cc1)Cl	8.47
134	Carboxin	CC1=C(SCCO1)C(=O)Nc1ccccc1	6.4
135*	Carboxin sulfoxide	CC1=C(C(=O)Nc2ccccc2)S(=O)CCO1	4.75
136	Carfentrazone-ethyl	CCOC(=O)C(Cl)Cc1cc(c(cc1Cl)F)N1N=C(C)N(C(F)F)C1=O	8.45
137*	Carpropamid	CCC1(C(C)C1(Cl)Cl)C(=O)NC(C)c1ccc(cc1)Cl	8.63
138*	Cartap	CN(C)C(CSC(N)=O)CSC(N)=O	0.97
139	Chinomethionate	Cc1ccc2nc3c(nc2c1)SC(=O)S3	9.27
140	Chlorbromuron	CON(C)C(=O)Nc1ccc(c(c1)Cl)Br	7.74

141	Chlorbufam	CC(OC(=O)Nc1cccc(c1)Cl)C#C	7.6
142	Chlordimeform	CN(C)C=Nc1ccc(cc1C)Cl	3.97
143	Chlorfenapyr	CCOCn1c(c(c(c1C(F)(F)F)Br)C#N)-c1ccc(cc1)Cl	9.13
144	Chlorfenvinphos	CCOP(=O)(OCC)OC(=CCl)c1ccc(cc1Cl)Cl	8.65
145*	Chlorfluazuron	Fc1cccc(c1C(=O)NC(=O)Nc1cc(c(c(c1)Cl)Oc1ncc(cc1Cl)C(F)(F)F)Cl)F	9.87
146	Chloridazon	NC1=C(Cl)C(=O)N(N=C1)c1ccccc1	4.54
147	Chlorimuron-ethyl	CCOC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)Cl	7.7
148	Chlormephos	CCOP(=S)(OCC)SCCl	8.23
149	Chlorophacinone	Clc1ccc(cc1)C(C(=O)C1C(=O)c2ccccc2C1=O)c1ccccc1	8.44
150	Chlorotoluron	CN(C)C(=O)Nc1ccc(c(c1)Cl)C	6.78
151	Chloroxuron	CN(C)C(=O)Nc1ccc(cc1)Oc1ccc(cc1)Cl	8.02
152	Chlorpyrifos	CCOP(=S)(OCC)Oc1nc(c(cc1Cl)Cl)Cl	9.56
153	Chlorpyrifos-methyl	COP(=S)(OC)Oc1nc(c(cc1Cl)Cl)Cl	8.91
154*	Chlorpyrifos-oxon	CCOP(=O)(OCC)Oc1nc(c(cc1Cl)Cl)Cl	8.33
155	Chlorsulfuron	COc1nc(nc(n1)NC(=O)NS(=O)(=O)c1ccccc1Cl)C	6.4

156	Chlorthiophos	CCOP(=S)(OCC)Oc1cc(c(cc1Cl)SC)Cl	9.71
157*	Chlorthiophos oxygen analog 2,5-isomer	CCOP(=O)(OCC)Oc1cc(c(cc1Cl)SC)Cl	8.69
158	Chlorthiophos sulfone	CCOP(=S)(OCC)Oc1cc(c(cc1Cl)S(C)(=O)=O)Cl	8.36
159	Chlorthiophos sulfoxide 2,5 isomer	CCOP(=S)(OCC)Oc1cc(c(cc1Cl)S(C)=O)Cl	8.55
160*	Chromafenozide	Cc1cc(cc(c1)C(=O)N(NC(=O)c1ccc2c(c1C)CCCO2)C(C)(C)C)C	8.2
161	Cinidon-ethyl	CCOC(=O)C(Cl)=Cc1cc(ccc1Cl)N1C(=O)C2=C(CCCC2)C1=O	9.3
162	Cinmethylin	CC(C)C12CCC(C)(O1)C(C2)OCc1ccccc1C	9.4
163*	Cinosulfuron	COCCOc1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(n1)OC)OC	5.82
164	Clethodim	CCSC(C)CC1CC(=C(C(=O)C1)C(CC)=NOCC=CCl)O	9.08
165	Clethodim, 5-hydroxy sulfone	CCC(=NOCC=CCl)C1=C(O)CC(O)(CC(C)S(=O)(=O)CC)CC1=O	6.67
166	Clodinafop-propargyl	CC(Oc1ccc(cc1)Oc1ncc(cc1F)Cl)C(=O)OCC#C	8.43
167	Clofencet	CCC1=C(C(O)=O)C(=O)C=NN1c1ccc(cc1)Cl	6.22
168	Clofentezine	Clc1ccccc1-c1nnc(nn1)-c1ccccc1Cl	8.8
169*	Clomazone	CC1(C)CON(Cc2ccccc2Cl)C1=O	7.42

170	Clomeprop	CC(Oc1ccc(c(c1Cl)C)Cl)C(=O)Nc1ccccc1	9.33
171	Cloquintocet-mexyl	CCCCCC(C)OC(=O)COc1ccc(c2cccnc12)Cl	9.43
172	Cloquintocet-mexyl metabolite CGA-153433	OC(=O)COc1ccc(c2cccnc12)Cl	5.15
173*	Cloransulam-methyl	CCOc1nc(cc2nc(n[n]12)S(=O)(=O)Nc1c(cccc1C(=O)OC)Cl)F	6.4
174	Clothianidin	CN=C(NCc1cnc([s]1)Cl)NN(=O)=O	4.16
175	Coumafuryl	CC(=O)CC(c1ccco1)C1=C(O)Oc2ccccc2C1=O	6.98
176	Coumaphos	CCOP(=S)(OCC)Oc1ccc2c(c1)OC(=O)C(=C2C)Cl	8.6
177	Coumaphos-oxon	CCOP(=O)(OCC)Oc1ccc2c(c1)OC(=O)C(=C2C)Cl	7.33
178	Coumatetralyl	OC1=C(C2CCCc3ccccc23)C(=O)Oc2ccccc12	8.02
179	Crimidine	CN(C)c1cc(nc(n1)Cl)C	5.1
180	Crotoxyphos	COP(=O)(OC)OC(C)=CC(=O)OC(C)c1ccccc1	7.72
181	Crufomate	CNP(=O)(OC)Oc1ccc(cc1Cl)C(C)(C)C	8.39
182	Cumyluron	CC(C)(NC(=O)NCc1ccccc1Cl)c1ccccc1	7.98
183	Cyanazine	CCNc1nc(nc(n1)NC(C)(C)C#N)Cl	5.87

184	Cyanofenphos	CCOP(=S)(Oc1ccc(cc1)C#N)c1ccccc1	8.56
185	Cyanophos	COP(=S)(OC)Oc1ccc(cc1)C#N	7.02
186	Cyantraniliprole	CNC(=O)c1cc(cc(c1NC(=O)c1cc(nn1-c1ncccc1Cl)Br)C)C#N	6.55
187	Cyazofamid	CN(C)S(=O)(=O)n1c(nc(c1-c1ccc(cc1)C)Cl)C#N	8.21
188	Cybutryne	CSc1nc(nc(n1)NC(C)(C)C)NC1CC1	8.15
189	Cyclanilide	OC(=O)C1(CC1)C(=O)Nc1ccc(cc1Cl)Cl	7.89
190*	Cyclaniliprole	CC(NC(=O)c1cc(cc(c1NC(=O)c1cc(nn1-c1ncccc1Cl)Br)Br)Cl)C1CC1	8.25
191	Cycloate	CCSC(=O)N(CC)C1CCCCC1	9.03
192	Cyclohexamide	CC1CC(C)C(=O)C(C1)C(O)CC1CC(=O)NC(=O)C1	5.7
193	Cycloprothrin	CCOc1ccc(cc1)C1(CC1(Cl)Cl)C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	9.78
194	Cyclosulfamuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)Nc1ccccc1C(=O)C1CC1)OC	8
195*	Cycloxydim	CCCC(=NOCC)C1=C(O)CC(CC1=O)C1CCCSC1	9.05
196	Cycluron	CN(C)C(=O)NC1CCCCCCC1	7.24
197	Cyenopyrafen	Cc1nn(c(c1C)C(OC(=O)C(C)(C)C)=C(C#N)c1ccc(cc1)C(C)(C)C)C	9.86

198	Cyflufenamid	Fc1ccc(c(c1F)C(NC(=O)Cc1ccccc1)=NOCC1CC1)C(F)(F)F	8.79
199	Cyflumetofen	COCCOC(=O)C(C#N)(C(=O)c1ccccc1C(F)(F)F)c1ccc(cc1)C(C)(C)C	9.14
200	Cyfluthrin	CC1(C)C(C=C(Cl)Cl)C1C(=O)OC(C#N)c1ccc(c(c1)Oc1ccccc1)F	9.79
201	Cyhalofop-butyl	CCCCOC(=O)C(C)Oc1ccc(cc1)Oc1ccc(cc1F)C#N	9
202	Cymiazole	CN1C=CSC1=Nc1ccc(cc1C)C	4.92
203	Cymoxanil	CCNC(=O)NC(=O)C(=NOC)C#N	4.83
204	Cypermethrin	CC1(C)C(C=C(Cl)Cl)C1C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	9.94
205	Cyphenothrin	CC(C)=CC1C(C(=O)OC(C#N)c2cccc(c2)Oc2ccccc2)C1(C)C	10.06
206	Cyprazine	CC(C)Nc1nc(nc(n1)NC1CC1)Cl	6.98
207	Cyproconazole	CC(C1CC1)C(O)(Cn1cncn1)c1ccc(cc1)Cl	8.1
208	Cyprodinil	Cc1cc(nc(n1)Nc1ccccc1)C1CC1	8.43
209	Cyprofuram	Clc1cccc(c1)N(C1CCOC1=O)C(=O)C1CC1	6.66
210	Cyprosulfamide	COc1ccccc1C(=O)NS(=O)(=O)c1ccc(cc1)C(=O)NC1CC1	5.96
211*	Cyromazine	Nc1nc(nc(n1)NC1CC1)N	1.14

212*	Cythioate	COP(=S)(OC)Oc1ccc(cc1)S(N)(=O)=O	5.57
213	Daimuron	Cc1ccc(cc1)NC(=O)NC(C)(C)c1ccccc1	7.88
214*	Daminozide	CN(C)NC(=O)CCC(O)=O	1.12
215	Deisopropylatrazine	CCNc1nc(nc(n1)Cl)N	3.84
216	Deltamethrin	CC1(C)C(C=C(Br)Br)C1C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	9.96
217*	Demeton-S	CCOP(=O)(OCC)SCCSCC	7.46
218	Demeton-S-methyl sulfone	CCS(=O)(=O)CCSP(=O)(OC)OC	3.4
219	Desmedipham	CCOC(=O)Nc1cccc(c1)OC(=O)Nc1ccccc1	7.1
220	Desmethyl formamido pirimicarb	CN(C)C(=O)Oc1nc(nc(c1C)C)N(C)C=O	6.14
221	Diafenthiuron	CC(C)c1cc(cc(c1NC(=S)NC(C)(C)C)C(C)C)Oc1ccccc1	9.79
222	Di-allate	CC(C)N(C(C)C)C(=O)SCC(Cl)=CCl	9.11
223	Diamidafos	CNP(=O)(NC)Oc1ccccc1	4.8
224	Diazinon	CCOP(=S)(OCC)Oc1cc(nc(n1)C(C)C)C	8.67
225	Diazinon oxon	CCOP(=O)(OCC)Oc1cc(nc(n1)C(C)C)C	7.38
226	Dicamba	COc1c(ccc(c1C(O)=O)Cl)Cl	4.9

227	Dichlofenthion	CCOP(=S)(OCC)Oc1ccc(cc1Cl)Cl	9.49
228	Dichlofluanid	CN(C)S(=O)(=O)N(SC(F)(Cl)Cl)c1ccccc1	8.12
229	Dichlormate	CNC(=O)OCc1ccc(c(c1)Cl)Cl	6.8
230	Dichlormid	ClC(Cl)C(=O)N(CC=C)CC=C	6.31
231*	Dichlorprop-P	CC(Oc1ccc(cc1Cl)Cl)C(O)=O	7.42
232*	Dichlorvos	COP(=O)(OC)OC=C(Cl)Cl	6.16
233	Diclobutrazol	CC(C)(C)C(O)C(Cc1ccc(cc1Cl)Cl)n1cncn1	8.51
234	Diclocymet	CC(NC(=O)C(C#N)C(C)(C)C)c1ccc(cc1Cl)Cl	8.19
235	Diclofop-methyl	COC(=O)C(C)Oc1ccc(cc1)Oc1ccc(cc1Cl)Cl	9.35
236*	Diclomezine	Cc1c(cc(cc1Cl)C1=NNC(=O)C=C1)Cl	8.36
237	Diclosulam	CCOc1nc(cc2nc(n[n]12)S(=O)(=O)Nc1c(cccc1Cl)Cl)F	6.63
238*	Dicrotophos	COP(=O)(OC)OC(C)=CC(=O)N(C)C	4.01
239	Dicyclanil	Nc1nc(nc(c1C#N)N)NC1CC1	2.05
240	Diethatyl-ethyl	CCOC(=O)CN(C(=O)CCl)c1c(cccc1CC)CC	8.47
241	Diethofencarb	CCOc1ccc(cc1OCC)NC(=O)OC(C)C	7.48

242	Difenacoum	OC1=C(C2CC(Cc3ccccc23)c2ccc(cc2)-c2ccccc2)C(=O)Oc2ccccc12	9.59
243	Difenoconazole	CC1COC(Cn2cncn2)(O1)c1ccc(cc1Cl)Oc1ccc(cc1)Cl	8.92
244	Difenoxuron	COc1ccc(cc1)Oc1ccc(cc1)NC(=O)N(C)C	7.07
245*	Difenzoquat	Cn1c(cc([n+]1C)-c1ccccc1)-c1ccccc1	9.1
246	Diflovidazin	Fc1cccc(c1-c1nnc(nn1)-c1ccccc1Cl)F	8.45
247	Diflubenzuron	Fc1cccc(c1C(=O)NC(=O)Nc1ccc(cc1)Cl)F	8.3
248	Diflufenican	Fc1ccc(c(c1)F)NC(=O)c1cccnc1Oc1cccc(c1)C(F)(F)F	9
249	Diflufenzopyr	CC(=NNC(=O)Nc1cc(cc(c1)F)F)c1ncccc1C(O)=O	6.3
250	Dimefluthin	COCc1c(c(c(c(c1F)F)COC(=O)C1C(C=C(C)C)C1(C)C)F)F	9.87
251	Dimefox	CN(C)P(F)(=O)N(C)C	4.05
252*	Dimefuron	CN(C)C(=O)Nc1ccc(c(c1)Cl)N1N=C(OC1=O)C(C)(C)C	7.34
253	Dimepiperate	CC(C)(SC(=O)N1CCCCC1)c1ccccc1	9.07
254	Dimethachlor	COCCN(C(=O)CCl)c1c(cccc1C)C	7.29
255*	Dimethametryn	CCNc1nc(nc(n1)SC)NC(C)C(C)C	8.39
256	Dimethenamid	COCC(C)N(C(=O)CCl)c1c(c[s]c1C)C	7.71

257	Dimethirimol	CCCCC1=C(C)NC(=NC1=O)N(C)C	5.7
258	Dimethoate	CNC(=O)CSP(=S)(OC)OC	4.51
259*	Dimethomorph	COc1ccc(cc1OC)C(=CC(=O)N1CCOCC1)c1ccc(cc1)Cl	7.66
260	Dimethyltolylsulfonyldiamide (DMST)	CN(C)S(=O)(=O)Nc1ccc(cc1)C	6.32
261	Dimethylvinphos	COP(=O)(OC)OC(=CCl)c1ccc(cc1Cl)Cl	8.02
262*	Dimetilan	CN(C)C(=O)Oc1cc(n(n1)C(=O)N(C)C)C	4.88
263*	Dimoxystrobin	CNC(=O)C(=NOC)c1ccccc1COc1cc(ccc1C)C	8.45
264*	Diniconazole	CC(C)(C)C(O)C(=Cc1ccc(cc1Cl)Cl)n1cncn1	8.96
265*	Dinitramine	CCN(CC)c1c(cc(c(c1N(=O)=O)N)C(F)(F)F)N(=O)=O	8.84
266	Dinotefuran	CN=C(NCC1CCOC1)NN(=O)=O	2.44
267*	Dinotefuran metabolite UF	CNC(=O)NCC1CCOC1	1.85
268	Dioxabenzofos	COP1(=S)OCc2ccccc2O1	7
269	Dioxacarb	CNC(=O)Oc1ccccc1C1OCCO1	4.51
270*	Dioxathion	CCOP(=S)(OCC)SC1OCCOC1SP(=S)(OCC)OCC	9.18
271*	Diphacinone	O=C(C(c1ccccc1)c1ccccc1)C1C(=O)c2	7.32

		ccccc2C1=O	
272	Diphenamid	CN(C)C(=O)C(c1ccccc1)c1ccccc1	7.29
273	Diphenamid desmethyl	CNC(=O)C(c1ccccc1)c1ccccc1	6.76
274	Diphenylamine	N(c1ccccc1)c1ccccc1	8.03
275	Dipropetryn	CCSc1nc(nc(n1)NC(C)C)NC(C)C	8.39
276	Disulfoton	CCOP(=S)(OCC)SCCSCC	8.88
277*	Disulfoton sulfone	CCOP(=S)(OCC)SCCS(=O)(=O)CC	6.88
278	Disulfoton sulfoxide	CCOP(=S)(OCC)SCCS(=O)CC	6.81
279	Ditalimfos	CCOP(=S)(OCC)N1C(=O)c2ccccc2C1=O	8.06
280	Dithiopyr	CSC(=O)c1c(c(c(nc1C(F)F)C(F)(F)F)C(=O)SC)CC(C)C	9.14
281	Diuron	CN(C)C(=O)Nc1ccc(c(c1)Cl)Cl	7.1
282	Dodemorph	CC1CN(CC(C)O1)C1CCCCCCCCCCC1	7.3
283	Doramectin	COC1CC(OC(C)C1O)OC1C(C)OC(CC1OC)OC1C(C)C=CC=C2COC3C(O)C(=CC(C(=O)OC4CC(CC=C1C)OC1(C4)OC(C(C)C=C1)C1CCCCC1)C23O)C	10.39
284	Edifenphos	CCOP(=O)(Sc1ccccc1)Sc1ccccc1	8.51

285	Emamectin B1a	CCC(C)C1OC2(CC3CC(CC=C(C)C(OC4CC(OC)C(OC5CC(OC)C(NC)C(C)O5)C(C)O4)C(C)C=CC=C4COC5C(O)C(=CC(C(=O)O3)C45O)C)O2)C=CC1C	9.47
286	Endothal	OC(=O)C1C2CCC(O2)C1C(O)=O	2.23
287	EPN	CCOP(=S)(Oc1ccc(cc1)N(=O)=O)c1ccccc1	8.99
288	Epoxiconazole	Fc1ccc(cc1)C1(Cn2cncn2)OC1c1ccccc1Cl	8.21
289	Eprinomectin B1a	CCC(C)C1OC2(CC3CC(CC=C(C)C(OC4CC(OC)C(OC5CC(OC)C(NC(C)=O)C(C)O5)C(C)O4)C(C)C=CC=C4COC5C(O)C(=CC(C(=O)O3)C45O)C)O2)C=CC1C	10
290*	EPTC	CCCN(CCC)C(=O)SCC	8.43
291*	Esfenvalerate	CC(C)C(C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1)c1ccc(cc1)Cl	10.02
292*	Esprocarb	CCN(C(C)C(C)C)C(=O)SCc1ccccc1	9.36
293	Etaconazole	CCC1COC(Cn2cncn2)(O1)c1ccc(cc1Cl)Cl	8.23
294	Ethaboxam	CCNc1nc(c([s]1)C(=O)NC(C#N)c1ccc[s]1)CC	6.73
295	Ethametsulfuron-methyl	CCOc1nc(nc(n1)NC(=O)NS(=O)(=O)c1ccccc1C(=O)OC)NC	6.52
296	Ethidimuron	CCS(=O)(=O)c1nnc([s]1)N(C)C(=O)NC	4.12

297*	Ethiofencarb	CCSCc1ccccc1OC(=O)NC	6.66
298*	Ethiofencarb sulfone	CCS(=O)(=O)Cc1ccccc1OC(=O)NC	3.97
299	Ethiofencarb sulfoxide	CCS(=O)Cc1ccccc1OC(=O)NC	4.14
300	Ethiolate	CCSC(=O)N(CC)CC	7.26
301	Ethion	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	9.43
302*	Ethion dioxon	CCOP(=O)(OCC)SCSP(=O)(OCC)OCC	6.92
303	Ethiozin	CCSC1=NN=C(C(=O)N1N)C(C)(C)C	6.88
304	Ethiprole	CCS(=O)c1c(n(nc1C#N)-c1c(cc(cc1Cl)C(F)(F)F)Cl)N	7.66
305	Ethirimol	CCCCC1=C(C)NC(=NC1=O)NCC	5.68
306	Ethofumesate	CCOC1Oc2ccc(cc2C1(C)C)OS(C)(=O)=O	7.58
307	Ethoprophos	CCCSP(=O)(OCC)SCCC	8.25
308	Ethoxysulfuron	CCOc1ccccc1OS(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	7.81
309*	Ethyl 1-naphthaleneacetate	CCOC(=O)Cc1cccc2ccccc12	8.15
310	Ethylchlozate	CCOC(=O)Cc1[nH]nc2ccc(cc12)Cl	7.44
311	Etobenzanid	CCOCOc1ccc(cc1)C(=O)Nc1cccc(c1Cl)Cl	8.55

312*	Etofenprox	CCOc1ccc(cc1)C(C)(C)COCc1cccc(c1)Oc1ccccc1	10.44
313	Etofenprox 4-hydroxy metabolite	CCOc1ccc(cc1)C(C)(C)COCc1cccc(c1)Oc1ccc(cc1)O	9.61
314	Etofenprox alpha-CO metabolite	CCOc1ccc(cc1)C(C)(C)COC(=O)c1cccc(c1)Oc1ccccc1	10.25
315	Etoxazole	CCOc1cc(ccc1C1COC(=N1)c1c(cccc1F)F)C(C)(C)C	9.71
316	Etridiazole	CCOc1nc(n[s]1)C(Cl)(Cl)Cl	8.4
317	Etrimfos	CCOc1cc(nc(n1)CC)OP(=S)(OC)OC	8.58
318*	Famoxadone	CC1(OC(=O)N(Nc2ccccc2)C1=O)c1ccc(cc1)Oc1ccccc1	8.66
319	Famphur	COP(=S)(OC)Oc1ccc(cc1)S(=O)(=O)N(C)C	6.88
320	Famphur oxon	COP(=O)(OC)Oc1ccc(cc1)S(=O)(=O)N(C)C	5.25
321*	Fenamidone	CSC1=NC(C)(C(=O)N1Nc1ccccc1)c1ccccc1	7.66
322*	Fenamiphos	CCOP(=O)(NC(C)C)Oc1ccc(c(c1)C)SC	8.36
323	Fenamiphos sulfone	CCOP(=O)(NC(C)C)Oc1ccc(c(c1)C)S(C)(=O)=O	6.43
324*	Fenamiphos sulfoxide	CCOP(=O)(NC(C)C)Oc1ccc(c(c1)C)S(C)=O	6.32
325	Fenarimol	OC(c1ccc(cc1)Cl)(c1cncnc1)c1ccccc1Cl	8.17

326	Fenazaquin	CC(C)(C)c1ccc(cc1)CCOc1ncnc2ccccc12	10.14
327	Fenbuconazole	Clc1ccc(cc1)CCC(Cn1cncn1)(C#N)c1ccccc1	8.29
328	Fenbuconazole metabolite RH-9129	Clc1ccc(cc1)C1CC(Cn2cncn2)(C(=O)O1)c1ccccc1	7.87
329*	Fenchlorphos oxon	COP(=O)(OC)Oc1cc(c(cc1Cl)Cl)Cl	8.21
330*	Fenchlorphos	COP(=S)(OC)Oc1cc(c(cc1Cl)Cl)Cl	9.47
331	Fenfuram	Cc1occc1C(=O)Nc1ccccc1	6.43
332*	Fenhexamid	CC1(CCCCC1)C(=O)Nc1ccc(c(c1Cl)Cl)O	8.14
333	Fenobucarb	CCC(C)c1ccccc1OC(=O)NC	7.55
334	Fenothiocarb	CN(C)C(=O)SCCCCOc1ccccc1	8.4
335	Fenoxanil	CC(C)C(C)(NC(=O)C(C)Oc1ccc(cc1Cl)Cl)C#N	8.4
336	Fenoxaprop-ethyl	CCOC(=O)C(C)Oc1ccc(cc1)Oc1nc2ccc(cc2o1)Cl	9.17
337*	Fenoxycarb	CCOC(=O)NCCOc1ccc(cc1)Oc1ccccc1	8.4
338*	Fenpiclonil	Clc1cccc(c1Cl)-c1c[nH]cc1C#N	7.44
339	Fenpropathrin	CC1(C)C(C(=O)OC(C#N)c2cccc(c2)Oc2ccccc2)C1(C)C	9.8
340*	Fenpropidin	CC(CN1CCCCC1)Cc1ccc(cc1)C(C)(C)C	7.44

341	Fenpropimorph	CC(CN1CC(C)OC(C)C1)Cc1ccc(cc1)C(C)(C)C	7.7
342	Fenpyrazamine	CC(C)N1N(C(=C(C1=O)c1ccccc1C)N)C(=O)SCC=C	7.95
343	Fenpyroximate	Cc1nn(c(c1C=NOCc1ccc(cc1)C(=O)OC(C)(C)C)Oc1ccccc1)C	9.84
344	Fensulfothion	CCOP(=S)(OCC)Oc1ccc(cc1)S(C)=O	7.1
345	Fensulfothion oxon sulfone	CCOP(=O)(OCC)Oc1ccc(cc1)S(C)(=O)=O	5.48
346	Fensulfothion oxone	CCOP(=O)(OCC)Oc1ccc(cc1)S(C)=O	5.48
347*	Fensulfothion sulfone	CCOP(=S)(OCC)Oc1ccc(cc1)S(C)(=O)=O	7.18
348	Fenthion	COP(=S)(OC)Oc1ccc(c(c1)C)SC	8.58
349	Fenthion oxon sulfone	COP(=O)(OC)Oc1ccc(c(c1)C)S(C)(=O)=O	4.87
350*	Fenthion sulfone	COP(=S)(OC)Oc1ccc(c(c1)C)S(C)(=O)=O	6.58
351*	Fenthion sulfoxide	COP(=S)(OC)Oc1ccc(c(c1)C)S(C)=O	6.43
352	Fenuron	CN(C)C(=O)Nc1ccccc1	4.33
353	Ferimzone	Cc1cc(nc(n1)NN=C(C)c1ccccc1C)C	7.44
354	Fipronil	Nc1c(c(nn1-c1c(cc(cc1Cl)C(F)(F)F)Cl)C#N)S(=O)C(F)(F)F	8.4

355*	Fipronil amide	NC(=O)c1nn(c(c1S(=O)C(F)(F)F)N)-c1c(cc(cc1Cl)C(F)(F)F)Cl	7.25
356	Fipronil monochloro-	Nc1c(c(nn1-c1ccc(cc1Cl)C(F)(F)F)C#N)S(=O)C(F)(F)F	8.1
357	Fipronildesufinyl amide	NC(=O)c1nn(c(c1C(F)(F)F)N)-c1c(cc(cc1Cl)C(F)(F)F)Cl	6.62
358*	Fipronil-sulfide	Nc1c(c(nn1-c1c(cc(cc1Cl)C(F)(F)F)Cl)C#N)SC(F)(F)F	8.51
359	Fipronil-sulfone	Nc1c(c(nn1-c1c(cc(cc1Cl)C(F)(F)F)Cl)C#N)S(=O)(=O)C(F)(F)F	8.66
360	Flamprop-M-isopropyl	CC(C)OC(=O)C(C)N(C(=O)c1ccccc1)c1ccc(c(c1)Cl)F	8.66
361	Flamprop-M-methyl	COC(=O)C(C)N(C(=O)c1ccccc1)c1ccc(c(c1)Cl)F	7.99
362	Flazasulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1ncccc1C(F)(F)F)OC	7.25
363	Flonicamid	FC(F)(F)c1ccncc1C(=O)NCC#N	3.33
364	Florasulam	COc1ncc(c2nc(n[n]12)S(=O)(=O)Nc1c(cccc1F)F)F	5.33
365	Fluacryprim	COC=C(C(=O)OC)c1ccccc1COc1cc(nc(n1)OC(C)C)C(F)(F)F	8.95
366	Fluazifop	CC(Oc1ccc(cc1)Oc1ccc(cn1)C(F)(F)F)C(O)=O	7.51
367	Fluazifop-butyl	CCCCOC(=O)C(C)Oc1ccc(cc1)Oc1ccc(cn1)C(F)(F)F	9.25

368	Fluazinam	FC(F)(F)c1cnc(c(c1)Cl)Nc1c(cc(c(c1N(=O)=O)Cl)C(F)(F)F)N(=O)=O	9.43
369*	Flubendiamide	Cc1cc(ccc1NC(=O)c1cccc(c1C(=O)NC(C)(C)CS(C)(=O)=O)I)C(F)(C(F)(F)F)C(F)(F)F	8.43
370	Flucarbazone	COC1=NN(C(=O)NS(=O)(=O)c2ccccc2OC(F)(F)F)C(=O)N1C	5.17
371	Flucetosulfuron	COCC(=O)OC(C(C)F)c1ncccc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	7.4
372	Flucycloxuron	Fc1cccc(c1C(=O)NC(=O)Nc1ccc(cc1)CON=C(C1CC1)c1ccc(cc1)Cl)F	9.58
373	Flucythrinate	CC(C)C(C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1)c1ccc(cc1)OC(F)F	9.61
374*	Fludioxonil	FC1(F)Oc2cccc(c2O1)-c1c[nH]cc1C#N	7.77
375	Flufenacet	CC(C)N(C(=O)COc1nnc([s]1)C(F)(F)F)c1ccc(cc1)F	8.17
376	Flufenoxuron	Fc1cc(ccc1NC(=O)NC(=O)c1c(cccc1F)F)Oc1ccc(cc1Cl)C(F)(F)F	9.64
377	Flufenpyr-ethyl	CCOC(=O)COc1cc(c(cc1Cl)F)N1N=CC(=C(C)C1=O)C(F)(F)F	8.32
378*	Flumethrin	CC1(C)C(C=C(Cl)c2ccc(cc2)Cl)C1C(=O)OC(C#N)c1ccc(c(c1)Oc1ccccc1)F	10.4
379	Flumetralin	CCN(Cc1c(cccc1Cl)F)c1c(cc(cc1N(=O)=O)C(F)(F)F)N(=O)=O	9.76
380	Flumetsulam	Cc1cc[n]2nc(nc2n1)S(=O)(=O)Nc1c(cccc1F)F	4.33
381	Flumiclorac-pentyl	CCCCCOC(=O)COc1cc(c(cc1Cl)F)N1C(=O)C2=C(CCCC2)C1=O	9.28

382*	Flumioxazin	Fc1cc2c(cc1N1C(=O)C3=C(CCCC3)C1=O)N(CC#C)C(=O)CO2	7.29
383	Flumorph	COc1ccc(cc1OC)C(=CC(=O)N1CCOCC1)c1ccc(cc1)F	7.29
384*	Fluometuron	CN(C)C(=O)Nc1cccc(c1)C(F)(F)F	6.69
385	Fluopicolide	FC(F)(F)c1cnc(c(c1)Cl)CNC(=O)c1c(cccc1Cl)Cl	7.84
386	Fluopyram	FC(F)(F)c1cnc(c(c1)Cl)CCNC(=O)c1ccccc1C(F)(F)F	8.06
387	Fluopyram-benzamide	NC(=O)c1ccccc1C(F)(F)F	4.1
388*	Fluoridamid	CC(=O)Nc1ccc(c(c1)NS(=O)(=O)C(F)(F)F)C	6.13
389	Fluorochloridone	FC(F)(F)c1cccc(c1)N1CC(CCl)C(Cl)C1=O	7.95
390	Fluoroglycofen-ethyl	CCOC(=O)COC(=O)c1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)C(F)(F)F	9.06
391	Fluoxastrobin	CON=C(C1=NOCCO1)c1ccccc1Oc1ncnc(c1F)Oc1ccccc1Cl	8.06
392	Flupyradifurone	FC(F)CN(Cc1ccc(nc1)Cl)C1=CC(=O)OC1	4.64
393	Flupyrsulfuron-methyl	COC(=O)c1ccc(nc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC)C(F)(F)F	7.55
394	Fluquinconazole	Fc1ccc2c(c1)C(=O)N(c1ccc(cc1Cl)Cl)C(=N2)n1cncn1	8.03
395*	Flurecol-butyl	CCCCOC(=O)C1(O)c2ccccc2-c2ccccc12	8.17

396	Flurenol	OC(=O)C1(O)c2ccccc2-c2ccccc12	5.16
397	Fluridone	CN1C=C(c2ccccc2)C(=O)C(=C1)c1cccc(c1)C(F)(F)F	7.36
398	Fluroxypyr	Nc1c(c(nc(c1Cl)OCC(O)=O)F)Cl	5.6
399*	Fluroxypyr-meptyl	CCCCCCC(C)OC(=O)COc1nc(c(c(c1Cl)N)Cl)F	9.65
400	Flurprimidol	CC(C)C(O)(c1ccc(cc1)OC(F)(F)F)c1cncnc1	7.92
401*	Flurtamone	CNC1=C(C(=O)C(O1)c1ccccc1)c1ccccc(c1)C(F)(F)F	7.58
402	Flusilazole	C[Si](Cn1cncn1)(c1ccc(cc1)F)c1ccc(cc1)F	8.4
403*	Fluthiacet-methyl	COC(=O)CSc1cc(c(cc1Cl)F)N=C1SC(=O)N2CCCCN12	8.47
404*	Flutianil	COc1ccccc1N1CCSC1=C(Sc1cc(ccc1F)C(F)(F)F)C#N	8.55
405	Flutolanil	CC(C)Oc1cccc(c1)NC(=O)c1ccccc1C(F)(F)F	7.78
406	Flutriafol	OC(Cn1cncn1)(c1ccc(cc1)F)c1ccccc1F	6.95
407	Fluxapyroxad	Cn1cc(c(n1)C(F)F)C(=O)Nc1ccccc1-c1cc(c(c(c1)F)F)F	7.85
408	Fluxapyroxad metabolite M700F008	FC(F)c1n[nH]cc1C(=O)Nc1ccccc1-c1cc(c(c(c1)F)F)F	7.7
409	Fluxapyroxad metabolite M700F048	OCC1OC(C(O)C(O)C1O)n1cc(c(n1)C(F)F)C(=O)Nc1ccccc1-c1cc(c(c(c1)F)F)F	7.18

410	Fluxofenim	FC(F)(F)C(=NOCC1OCCO1)c1ccc(cc1)Cl	8.69
411	Fomesafen	CS(=O)(=O)NC(=O)c1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)C(F)(F)F	7.55
412	Fonofos	CCOP(=S)(CC)Sc1ccccc1	8.66
413*	Fonofos oxon	CCOP(=O)(CC)Sc1ccccc1	7.29
414	Foramsulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1c(cccc1C(=O)N(C)C)NC=O)OC	6.43
415	Forchlorfenuron	Clc1cc(ccn1)NC(=O)Nc1ccccc1	7.03
416*	Fosthiazate	CCOP(=O)(SC(C)CC)N1CCSC1=O	6.73
417	Fuberidazole	c1ccc2[nH]c(nc2c1)-c1ccco1	4.29
418	Fumesate	CC1(C)C(O)Oc2ccc(cc12)OS(C)(=O)=O	5.41
419	Furalaxyl	COC(=O)C(C)N(C(=O)c1ccco1)c1c(cccc1C)C	7.58
420	Furametpyr	CC1OC(C)(C)c2cccc(c12)NC(=O)c1c(nn(c1Cl)C)C	6.95
421*	Furathiocarb	CCCCOC(=O)N(C)SN(C)C(=O)Oc1ccccc2c1OC(C)(C)C2	9.29
422*	Furilazole	CC1(C)OC(CN1C(=O)C(Cl)Cl)c1ccco1	6.78
423*	Furmecyclox	CON(C1CCCCC1)C(=O)c1cc(oc1C)C	8.62
424	Gibberellin A4	CC12C(O)CCC3(OC1=O)C1CCC4CC1(CC4=C)C(C23)C(O)=O	7.55

425*	Gibberellin A7	CC12C(O)C=CC3(OC1=O)C1CCC4CC1(CC4=C)C(C23)C(O)=O	7.36
426	Griseofulvin	COc1cc(c2c(c1Cl)OC1(C(C)CC(=O)C=C1OC)C2=O)OC	6.95
427	Halfenprox	CC(C)(COCc1cccc(c1)Oc1ccccc1)c1ccc(cc1)OC(F)(F)Br	10.69
428	Halofenozide	CC(C)(C)N(NC(=O)c1ccc(cc1)Cl)C(=O)c1ccccc1	7.66
429	Halosulfuron-methyl	COC(=O)c1c(nn(c1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC)C)Cl	7.81
430	Haloxyfop	CC(Oc1ccc(cc1)Oc1ncc(cc1Cl)C(F)(F)F)C(O)=O	8.36
431	Haloxyfop-P-methyl	COC(=O)C(C)Oc1ccc(cc1)Oc1ncc(cc1Cl)C(F)(F)F	8.99
432	Heptenophos	COP(=O)(OC)OC1=C(Cl)C2C=CCC12	7.21
433	Hexaconazole	CCCCC(O)(Cn1cncn1)c1ccc(cc1Cl)Cl	8.75
434*	Hexaflumuron	FC(F)C(F)(F)Oc1c(cc(cc1Cl)NC(=O)NC(=O)c1c(cccc1F)F)Cl	9.03
435	Hexazinone	CN(C)C1=NC(=O)N(C2CCCCC2)C(=O)N1C	6.24
436	Hexythiazox	CC1C(SC(=O)N1C(=O)NC1CCCCC1)c1ccc(cc1)Cl	9.58
437	Icaridin	CCC(C)OC(=O)N1CCCCC1CCO	7.44
438	Imazalil	Clc1ccc(c(c1)Cl)C(Cn1ccnc1)OCC=C	6.99

439	Imazamethabenz-methyl	COC(=O)c1ccc(cc1C1=NC(C)(C(C)C)C(=O)N1)C	6.24
440	Imazamox	COCc1cnc(c(c1)C(O)=O)C1=NC(C)(C(C)C)C(=O)N1	4.79
441*	Imazapic	CC(C)C1(C)N=C(NC1=O)c1ncc(cc1C(O)=O)C	4.94
442	Imazapyr	CC(C)C1(C)N=C(NC1=O)c1ncccc1C(O)=O	4.06
443	Imazaquin	CC(C)C1(C)N=C(NC1=O)c1nc2ccccc2cc1C(O)=O	6.13
444	Imibenconazole	Clc1ccc(cc1)CSC(Cn1cncn1)=Nc1ccc(cc1Cl)Cl	9.44
445	Imicyafos	CCCSP(=O)(OCC)N1CCN(CC)C1=NC#N	5.94
446*	Imidacloprid	Clc1ccc(cn1)CN1CCNC1=NN(=O)=O	4.14
447*	Imidacloprid desnitro	NC1=NCCN1Cc1ccc(nc1)Cl	2.51
448	Imidacloprid desnitro-olefin	Nc1nccn1Cc1ccc(nc1)Cl	2.47
449*	Imidacloprid urea	Clc1ccc(cn1)CN1CCNC1=O	4.1
450	Imidacloprid, 5-hydroxy	OC1CN=C(NN(=O)=O)N1Cc1ccc(nc1)Cl	3.71
451*	Imiprothrin	CC(C)=CC1C(C(=O)OCN2C(=O)CN(CC#C)C2=O)C1(C)C	8.1
452*	Inabenfide	OC(c1ccccc1)c1cc(ccc1NC(=O)c1ccncc1)Cl	7.58

453	Indanofan	CCC1(CC2(CO2)c2cccc(c2)Cl)C(=O)c2ccccc2C1=O	8.21
454	Indaziflam	CC(F)c1nc(nc(n1)NC1C(C)Cc2ccc(cc12)C)N	8.25
455	Indole-3-acetic acid	OC(=O)Cc1c[nH]c2ccccc12	4.98
456	Indoxacarb	COC(=O)N(C(=O)N1COC2(Cc3cc(ccc3C2=N1)Cl)C(=O)OC)c1ccc(cc1)OC(F)(F)F	8.99
457	Iodosulfuron-methyl	COC(=O)c1ccc(cc1S(=O)(=O)NC(=O)Nc1nc(nc(n1)OC)C)I	7.1
458	Ipconazole	CC(C)C1CCC(Cc2ccc(cc2)Cl)C1(O)Cn1cncn1	9.1
459	Ipfencarbazone	CC(C)N(C(=O)N1C=NN(C1=O)c1ccc(cc1Cl)Cl)c1ccc(cc1F)F	8.55
460*	Iprobenfos	CC(C)OP(=O)(OC(C)C)SCc1ccccc1	8.52
461*	Iprodione	CC(C)NC(=O)N1CC(=O)N(C1=O)c1cc(cc(c1)Cl)Cl	8.32
462	Iprovalicarb	CC(C)OC(=O)NC(C(C)C)C(=O)NC(C)c1ccc(cc1)C	8.1
463	Isazophos	CCOP(=S)(OCC)Oc1nc(n(n1)C(C)C)Cl	8.03
464	Isocarbamid	CC(C)CNC(=O)N1CCNC1=O	5.04
465	Isocarbophos	COP(N)(=S)Oc1ccccc1C(=O)OC(C)C	7.1
466	Isofenphos	CCOP(=S)(NC(C)C)Oc1ccccc1C(=O)OC(C)C	8.88

467	Isofenphos oxon	CCOP(=O)(NC(C)C)Oc1ccccc1C(=O)OC(C)C	8.06
468	Isofenphos-des-N-isopropyl	CCOP(N)(=S)Oc1ccccc1C(=O)OC(C)C	7.55
469	Isofenphos-methyl	COP(=S)(NC(C)C)Oc1ccccc1C(=O)OC(C)C	8.39
470*	Isofetamid	CC(C)Oc1ccc(c(c1)C)C(=O)C(C)(C)NC(=O)c1[s]ccc1C	8.1
471	Isonoruron	CN(C)C(=O)NC1CCC2C3CCC(C3)C12	7.66
472*	Isopropalin	CCCN(CCC)c1c(cc(cc1N(=O)=O)C(C)C)N(=O)=O	10
473	Isoprothiolane	CC(C)OC(=O)C(C(=O)OC(C)C)=C1SCCS1	7.79
474	Isoproturon	CC(C)c1ccc(cc1)NC(=O)N(C)C	7.07
475	Isopyrazam	CC(C)C1C2CCC1c1c(cccc21)NC(=O)c1cn(nc1C(F)F)C	8.99
476	Isotianil	Clc1n[s]c(c1Cl)C(=O)Nc1ccccc1C#N	7.62
477	Isouron	CN(C)C(=O)Nc1cc(on1)C(C)(C)C	6.35
478	Isoxaben	CCC(C)(CC)c1cc(on1)NC(=O)c1c(cccc1OC)OC	7.76
479	Isoxadifen-ethyl	CCOC(=O)C1=NOC(C1)(c1ccccc1)c1ccccc1	8.47
480	Isoxaflutole	CS(=O)(=O)c1cc(ccc1C(=O)c1cnoc1C1CC1)C(F)(F)F	7.04

481	Isoxathion	CCOP(=S)(OCC)Oc1cc(on1)-c1ccccc1	8.77
482	Ivermectin B1a	CCC(C)C1OC2(CCC1C)CC1CC(CC=C(C)C(OC3CC(OC)C(OC4CC(OC)C(O)C(C)O4)C(C)O3)C(C)C=CC=C3COC4C(O)C(=CC(C(=O)O1)C34O)C)O2	10.71
483	Karbutilate	CN(C)C(=O)Nc1cccc(c1)OC(=O)NC(C)(C)C	6.2
484	Kresoxim-methyl	CON=C(C(=O)OC)c1ccccc1COc1ccccc1C	8.51
485	Lactofen	CCOC(=O)C(C)OC(=O)c1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)C(F)(F)F	9.35
486*	Lambda-cyhalothrin	CC1(C)C(C=C(Cl)C(F)(F)F)C1C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	9.84
487	Lenacil	O=C1NC2=C(CCC2)C(=O)N1C1CCCCC1	6.99
488	Leptophos	COP(=S)(Oc1cc(c(cc1Cl)Br)Cl)c1ccccc1	10.21
489	Leptophos oxon	COP(=O)(Oc1cc(c(cc1Cl)Br)Cl)c1ccccc1	9.1
490*	Linuron	CON(C)C(=O)Nc1ccc(c(c1)Cl)Cl	7.58
491*	Lufenuron	FC(C(F)(F)F)C(F)(F)Oc1cc(c(cc1Cl)NC(=O)NC(=O)c1c(cccc1F)F)Cl	9.39
492	Malathion	CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC	7.84
493	Malathion oxon	CCOC(=O)CC(SP(=O)(OC)OC)C(=O)OCC	6.32

494*	Maleic hydrazine	O=C1NNC(=O)C=C1	1.06
495	Mandestrobin	CNC(=O)C(OC)c1ccccc1COc1cc(ccc1C)C	8.62
496	Mandipropamid	COc1cc(ccc1OCC#C)CCNC(=O)C(OCC#C)c1ccc(cc1)Cl	7.73
497*	MCPA-2-Ethylhexyl	CCCCC(CC)COC(=O)COc1ccc(cc1C)Cl	10.36
498	MCPB	Cc1cc(ccc1OCCCC(O)=O)Cl	8.17
499*	Mecarbam	CCOC(=O)N(C)C(=O)CSP(=S)(OCC)OCC	8.21
500	Medetomidine	CC(c1cnc[nH]1)c1cccc(c1C)C	5.87
501	Mefenacet	CN(C(=O)COc1nc2ccccc2[s]1)c1ccccc1	7.99
502	Mefenpyr-diethyl	CCOC(=O)C1=NN(c2ccc(cc2Cl)Cl)C(C)(C1)C(=O)OCC	8.73
503*	Mefluidide	CC(=O)Nc1cc(c(cc1C)C)NS(=O)(=O)C(F)(F)F	6.24
504*	Mepanipyrim	CC#Cc1cc(nc(n1)Nc1ccccc1)C	8.06
505*	Meperfluthrin	COCc1c(c(c(c(c1F)F)COC(=O)C1C(C=C(Cl)Cl)C1(C)C)F)F	9.84
506	Mephosfolan	CCOP(=O)(OCC)N=C1SCC(C)S1	6.13
507	Mepronil	CC(C)Oc1cccc(c1)NC(=O)c1ccccc1C	7.84
508	Mesosulfuron-methyl	COC(=O)c1ccc(cc1S(=O)(=O)NC(=O)	6.84

		Nc1nc(cc(n1)OC)OC)CNS(C)(=O)=O	
509	Mesotrione	CS(=O)(=O)c1ccc(c(c1)N(=O)=O)C(=O)C1C(=O)CCCC1=O	4.98
510*	Metaflumizone	FC(F)(F)Oc1ccc(cc1)NC(=O)NN=C(Cc1ccc(cc1)C#N)c1cccc(c1)C(F)(F)F	9.32
511*	Metalaxyl	COCC(=O)N(C(C)C(=O)OC)c1c(cccc1C)C	7.1
512	Metalaxyl alanine metabolite CGA-94689	COCC(=O)N(C(C)C(=O)OC)c1c(cccc1CO)C	5.93
513	Metamitron	CC1=NN=C(c2ccccc2)C(=O)N1N	4.37
514*	Metazachlor	Cc1cccc(c1N(Cn1cccn1)C(=O)CCl)C	6.99
515	Metconazole	CC1(C)CCC(Cc2ccc(cc2)Cl)C1(O)Cn1cncn1	8.77
516	Methabenzthiazuron	CNC(=O)N(C)c1nc2ccccc2[s]1	6.95
517*	Methacrifos	COC(=O)C(C)=COP(=S)(OC)OC	6.92
518*	Methamidophos	COP(N)(=O)SC	1.56
519	Methfuroxam	Cc1oc(c(c1C)C(=O)Nc1ccccc1)C	7.29
520*	Methidathion	COC1=NN(CSP(=S)(OC)OC)C(=O)S1	7.21
521*	Methiocarb	CNC(=O)Oc1cc(c(c(c1)C)SC)C	7.66
522	Methiocarb sulfone	CNC(=O)Oc1cc(c(c(c1)C)S(C)(=O)=O)C	4.79

523	Methiocarb sulfoxide	CNC(=O)Oc1cc(c(c(c1)C)S(C)=O)C	4.45
524	Methomyl	CNC(=O)ON=C(C)SC	3.25
525	Methomyl oxime	CSC(C)=NO	2.19
526	Methomyl sulfone	CNC(=O)ON=C(C)S(C)(=O)=O	2.35
527*	Methoprotryne	COCCCNc1nc(nc(n1)SC)NC(C)C	7.32
528	Methyldymron	CN(C(=O)NC(C)(C)c1ccccc1)c1ccccc1	8.06
529	Metobromuron	CON(C)C(=O)Nc1ccc(cc1)Br	6.84
530*	Metofluthrin	COCc1c(c(c(c(c1F)F)COC(=O)C1C(C=CC)C1(C)C)F)F	9.61
531	Metolachlor	CCc1cccc(c1N(C(C)COC)C(=O)CCl)C	8.32
532	Metolcarb	CNC(=O)Oc1cccc(c1)C	5.79
533*	Metominostrobin	CNC(=O)C(=NOC)c1ccccc1Oc1ccccc1	7.25
534	Metosulam	COc1cc([n]2nc(nc2n1)S(=O)(=O)Nc1c(ccc(c1Cl)C)Cl)OC	6.32
535	Metoxuron	COc1ccc(cc1Cl)NC(=O)N(C)C	5.45
536	Metrafenone	COc1ccc(c(c1C(=O)c1c(cc(c(c1OC)OC)OC)C)C)Br	8.84
537*	Metribuzin	CSC1=NN=C(C(=O)N1N)C(C)(C)C	6.08

538	Metsulfuron-methyl	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(n1)OC)C	6.09
539	Mevinphos, cis-	COC(=O)C=C(C)OP(=O)(OC)OC	5.21
540	Mexacarbate	CNC(=O)Oc1cc(c(c(c1)C)N(C)C)C	4.83
541*	Milbemectin A4	CCC1OC2(CCC1C)CC1CC(CC=C(C)CC(C)C=CC=C3COC4C(O)C(=CC(C(=O)O1)C34O)C)O2	10.11
542	Mobam	CNC(=O)Oc1cccc2[s]ccc12	6.24
543	Molinate	CCSC(=O)N1CCCCCC1	7.99
544	Momfluorothrin	COCc1c(c(c(c(c1F)F)COC(=O)C1C(C=C(C)C#N)C1(C)C)F)F	8.47
545	Monocrotophos	CNC(=O)C=C(C)OP(=O)(OC)OC	3.63
546	Monolinuron	CON(C)C(=O)Nc1ccc(cc1)Cl	6.58
547	Monuron	CN(C)C(=O)Nc1ccc(cc1)Cl	5.94
548	Moxidectin	CON=C1CC2(CC3CC(CC=C(C)CC(C)C=CC=C4COC5C(O)C(=CC(C(=O)O3)C45O)C)O2)OC(C1C)C(C)=CC(C)C	10.55
549	MPMC	CNC(=O)Oc1ccc(c(c1)C)C	6.54
550	Myclobutanil	CCCCC(Cn1cncn1)(C#N)c1ccc(cc1)Cl	7.93
551	N,N-Diethyl-m-toluamide	CCN(CC)C(=O)c1cccc(c1)C	7.1

552	N-1-Naphthylphthalamic acid	OC(=O)c1ccccc1C(=O)Nc1cccc2ccccc12	6.2
553	Naled	COP(=O)(OC)OC(Br)C(Cl)(Cl)Br	7.21
554	Napropamide	CCN(CC)C(=O)C(C)Oc1cccc2ccccc12	8.21
555	Neburon	CCCCN(C)C(=O)Nc1ccc(c(c1)Cl)Cl	8.43
556	Nicosulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1ncccc1C(=O)N(C)C)OC	6.02
557	Nicotine	CN1CCCC1c1cccnc1	1.12
558	Nitenpyram	CCN(Cc1ccc(nc1)Cl)C(NC)=CN(=O)=O	3.1
559*	Nithiazine	O=N(=O)C=C1NCCCS1	2.03
560	Nitralin	CCCN(CCC)c1c(cc(cc1N(=O)=O)S(C)(=O)=O)N(=O)=O	8.36
561	Norea	CN(C)C(=O)NC1CC2CC1C1CCCC21	7.66
562	Norflurazon	CNC1=C(Cl)C(=O)N(N=C1)c1cccc(c1)C(F)(F)F	7.14
563	Novaluron	FC(OC(F)(F)F)C(F)(F)Oc1ccc(cc1Cl)NC(=O)NC(=O)c1c(cccc1F)F	9.06
564	Noviflumuron	FC(C(F)(F)F)C(F)(F)Oc1c(cc(c(c1Cl)F)NC(=O)NC(=O)c1c(cccc1F)F)Cl	9.51
565	Nuarimol	OC(c1ccc(cc1)F)(c1cncnc1)c1ccccc1Cl	7.63
566	Octhilinone	CCCCCCCCN1SC=CC1=O	8.32

567	Ofurace	Cc1cccc(c1N(C1CCOC1=O)C(=O)CCl)C	6.24
568*	Omethoate	CNC(=O)CSP(=O)(OC)OC	2.39
569	Orbencarb	CCN(CC)C(=O)SCc1ccccc1Cl	8.81
570	Orthosulfamuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)Nc1ccccc1C(=O)N(C)C)OC	7.03
571	Orysastrobin	CNC(=O)C(=NOC)c1ccccc1CON=C(C)C(=NOC)C(C)=NOC	7.88
572*	Oryzalin	CCCN(CCC)c1c(cc(cc1N(=O)=O)S(N)(=O)=O)N(=O)=O	8.22
573*	Oryzalin dimethyl	CCCN(CCC)c1c(cc(cc1N(=O)=O)S(=O)(=O)N(C)C)N(=O)=O	8.88
574	Oxabetrinil	N#CC(=NOCC1OCCO1)c1ccccc1	7.47
575	Oxadiargyl	CC(C)(C)C1=NN(C(=O)O1)c1cc(c(cc1Cl)Cl)OCC#C	8.77
576*	Oxadiazon	CC(C)Oc1cc(c(cc1Cl)Cl)N1N=C(OC1=O)C(C)(C)C	9.43
577*	Oxadixyl	COCC(=O)N(N1CCOC1=O)c1c(cccc1C)C	5.75
578	Oxamyl	CNC(=O)ON=C(SC)C(=O)N(C)C	2.99
579*	Oxamyl-oxime	CSC(=NO)C(=O)N(C)C	2.28
580	Oxasulfuron	Cc1cc(nc(n1)NC(=O)NS(=O)(=O)c1ccccc1C(=O)OC1COC1)C	5.9

581	Oxathiapiprolin	Cc1cc(nn1CC(=O)N1CCC(CC1)c1nc(c[s]1)C1=NOC(C1)c1c(cccc1F)F)C(F)(F)F	7.77
582	Oxaziclomefone	CC1=C(C(=O)N(CO1)C(C)(C)c1cc(cc(c1)Cl)Cl)c1ccccc1	9.25
583	Oxycarboxin	CC1=C(C(=O)Nc2ccccc2)S(=O)(=O)CCO1	4.91
584	Oxydemeton-methyl	CCS(=O)CCSP(=O)(OC)OC	3.29
585	Oxyfluorfen	CCOc1cc(ccc1N(=O)=O)Oc1ccc(cc1Cl)C(F)(F)F	9.28
586*	Paclobutrazol	CC(C)(C)C(O)C(Cc1ccc(cc1)Cl)n1cncn1	7.81
587	Parathion oxon	CCOP(=O)(OCC)Oc1ccc(cc1)N(=O)=O	6.89
588*	Pebulate	CCCCN(CC)C(=O)SCCC	8.94
589*	Penconazole	CCCC(Cn1cncn1)c1ccc(cc1Cl)Cl	8.62
590	Pencycuron	Clc1ccc(cc1)CN(C1CCCC1)C(=O)Nc1ccccc1	8.88
591	Pendimethalin	CCC(CC)Nc1c(cc(c(c1N(=O)=O)C)C)N(=O)=O	9.6
592	Penflufen	CC(C)CC(C)c1ccccc1NC(=O)c1c(nn(c1F)C)C	8.55
593	Penoxsulam	COc1cnc([n]2nc(nc12)NS(=O)(=O)c1c(cccc1C(F)(F)F)OCC(F)F)OC	6.54
594	Pentanochlor	CCCC(C)C(=O)Nc1ccc(c(c1)Cl)C	8.43

595*	Penthiopyrad	CC(C)CC(C)c1[s]ccc1NC(=O)c1cn(nc1C(F)(F)F)C	8.58
596	Pentoxazone	CC(C)=C1OC(=O)N(C1=O)c1cc(c(cc1F)Cl)OC1CCCC1	9.25
597	Perfluidone	Cc1cc(ccc1NS(=O)(=O)C(F)(F)F)S(=O)(=O)c1ccccc1	6.81
598	Permethrin	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc1cccc(c1)Oc1ccccc1	10.4
599	Pethoxamid	CCOCCN(C(=O)CCl)C(=C(C)C)c1ccccc1	8.21
600	Phenmedipham	COC(=O)Nc1cccc(c1)OC(=O)Nc1cccc(c1)C	7.24
601	Phenothiazine	N1c2ccccc2Sc2ccccc12	8.3
602	Phenothrin	CC(C)=CC1C(C(=O)OCc2cccc(c2)Oc2ccccc2)C1(C)C	10.36
603	Phenthoate	CCOC(=O)C(SP(=S)(OC)OC)c1ccccc1	8.43
604	Phorate sulfone	CCOP(=S)(OCC)SCS(=O)(=O)CC	6.92
605*	Phorate sulfoxide	CCOP(=S)(OCC)SCS(=O)CC	6.81
606*	Phorate	CCOP(=S)(OCC)SCSCC	8.8
607	Phosalone	CCOP(=S)(OCC)SCN1C(=O)Oc2cc(ccc12)Cl	8.77
608	Phosfolan	CCOP(=O)(OCC)N=C1SCCS1	5.45
609*	Phosmet	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=	7.34

		O	
610*	Phosphamidon	CCN(CC)C(=O)C(Cl)=C(C)OP(=O)(OC)OC	5.87
611*	Phoxim	CCOP(=S)(OCC)ON=C(C#N)c1ccccc1	8.73
612	Picolinafen	Fc1ccc(cc1)NC(=O)c1cccc(n1)Oc1ccccc(c1)C(F)(F)F	9.35
613	Picoxystrobin	COC=C(C(=O)OC)c1ccccc1COc1cccc(n1)C(F)(F)F	8.4
614	Pinoxaden	CCc1cc(cc(c1C1=C(OC(=O)C(C)(C)C)N2CCOCCN2C1=O)CC)C	8.82
615	Piperonyl butoxide	CCCCOCCOCCOCc1cc2c(cc1CCC)OCO2	9.43
616	Piperophos	CCCOP(=S)(OCCC)SCC(=O)N1CCCCC1C	8.98
617	Pirimicarb	CN(C)C(=O)Oc1nc(nc(c1C)C)N(C)C	5.98
618	Pirimicarb-desmethyl	CNc1nc(c(c(n1)OC(=O)N(C)C)C)C	4.33
619	Pirimiphos-ethyl	CCOP(=S)(OCC)Oc1cc(nc(n1)N(CC)CC)C	9.36
620	Pirimiphos-ethyl-oxon	CCOP(=O)(OCC)Oc1cc(nc(n1)N(CC)CC)C	8.43
621	Pirimiphos-methyl	CCN(CC)c1nc(cc(n1)OP(=S)(OC)OC)C	8.8
622	Prallethrin	CC(C)=CC1C(C(=O)OC2CC(=O)C(=C2C)CC#C)C1(C)C	9.03

623	Pretilachlor	CCCOCCN(C(=O)CCl)c1c(cccc1CC)CC	9.14
624	Primisulfuron-methyl	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC(F)F)OC(F)F	7.77
625*	Probenazole	C=CCOC1=NS(=O)(=O)c2ccccc12	5.86
626	Prochloraz	CCCN(CCOc1c(cc(cc1Cl)Cl)Cl)C(=O)n1ccnc1	8.75
627	Procyazine	CC(C)(Nc1nc(nc(n1)NC1CC1)Cl)C#N	6.09
628	Prodiamine	CCCN(CCC)c1c(cc(c(c1N(=O)=O)N)C(F)(F)F)N(=O)=O	9.47
629	Profenophos	CCCSP(=O)(OCC)Oc1ccc(cc1Cl)Br	9.25
630*	Profluralin	CCCN(CC1CC1)c1c(cc(cc1N(=O)=O)C(F)(F)F)N(=O)=O	9.8
631*	Progesterone	CC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	8.73
632*	Prohexadione	CCC(=O)C1C(=O)CC(CC1=O)C(O)=O	5.29
633	Promecarb	CNC(=O)Oc1cc(cc(c1)C(C)C)C	7.84
634	Prometon	COc1nc(nc(n1)NC(C)C)NC(C)C	6.92
635	Prometryn	CSc1nc(nc(n1)NC(C)C)NC(C)C	7.92
636*	Prometryn metabolite GS 11354	CSc1nc(nc(n1)NC(C)C)N	5.65
637	Propachlor	CC(C)N(C(=O)CCl)c1ccccc1	7.08

638	Propamocarb	CCCOC(=O)NCCCN(C)C	2.58
639	Propanil	CCC(=O)Nc1ccc(c(c1)Cl)Cl	7.62
640*	Propaphos	CCCOP(=O)(OCCC)Oc1ccc(cc1)SC	8.56
641	Propaquizafop	CC(Oc1ccc(cc1)Oc1cnc2cc(ccc2n1)Cl)C(=O)OCCON=C(C)C	9.32
642	Propargite	CC(C)(C)c1ccc(cc1)OC1CCCCC1OS(=O)OCC#C	9.69
643	Propazine	CC(C)Nc1nc(nc(n1)NC(C)C)Cl	7.58
644	Propetamphos	CCNP(=S)(OC)OC(C)=CC(=O)OC(C)C	7.99
645	Propiconazole	CCCC1COC(Cn2cncn2)(O1)c1ccc(cc1Cl)Cl	8.69
646	Propisochlor	CCc1cccc(c1N(COC(C)C)C(=O)CCl)C	8.62
647	Propoxur	CNC(=O)Oc1ccccc1OC(C)C	6.17
648	Propoxycarbazone	CCCOC1=NN(C(=O)NS(=O)(=O)c2ccccc2C(=O)OC)C(=O)N1C	5.79
649*	Propyrisulfuron	CCCc1ccc2nc(c([n]2n1)S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC)Cl	7.81
650*	Propyzamide	CC(C)(NC(=O)c1cc(cc(c1)Cl)Cl)C#C	7.84
651	Proquinazid	CCCOC1=Nc2ccc(cc2C(=O)N1CCC)I	9.91
652*	Prosulfocarb	CCCN(CCC)C(=O)SCc1ccccc1	9.17

653	Prosulfuron	COc1nc(nc(n1)NC(=O)NS(=O)(=O)c1ccccc1CCC(F)(F)F)C	7.51
654	Prothioconazole	OC(CN1NC=NC1=S)(Cc1ccccc1Cl)C1(Cl)CC1	8.69
655	Prothioconazole-Desthio	OC(Cc1ccccc1Cl)(Cn1cncn1)C1(Cl)CC1	8.29
656	Prothiophos	CCCSP(=S)(OCC)Oc1ccc(cc1Cl)Cl	10.14
657	Proxel	O=C1NSc2ccccc12	4.45
658	Prynachlor	CC(C#C)N(C(=O)CCl)c1ccccc1	6.81
659	Pymetrozine	CC1=NNC(=O)N(C1)N=Cc1cccnc1	2.11
660	Pyracarbolid	CC1=C(CCCO1)C(=O)Nc1ccccc1	6.24
661*	Pyraclofos	CCCSP(=O)(OCC)Oc1cnn(c1)-c1ccc(cc1)Cl	8.77
662	Pyraclonil	CN(CC#C)c1c(cnn1-c1nn2c(c1Cl)CCCC2)C#N	6.84
663*	Pyraclostrobin	CON(C(=O)OC)c1ccccc1COc1ccn(n1)-c1ccc(cc1)Cl	8.69
664	Pyraflufen-ethyl	CCOC(=O)COc1cc(c(cc1Cl)F)-c1nn(c(c1Cl)OC(F)F)C	8.58
665	Pyrasulfotole	CN1NC(=C(C1=O)C(=O)c1ccc(cc1S(C)(=O)=O)C(F)(F)F)C	5.68
666*	Pyrazolynate	Cc1ccc(cc1)S(=O)(=O)Oc1c(c(nn1C)C)C(=O)c1ccc(cc1Cl)Cl	8.84

667	Pyrazophos	CCOC(=O)c1c[n]2nc(cc2nc1C)OP(=S)(OCC)OCC	8.77
668*	Pyrazosulfuron-ethyl	CCOC(=O)c1cnn(c1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC)C	7.95
669*	Pyrazoxyfen	Cc1nn(c(c1C(=O)c1ccc(cc1Cl)Cl)OCC(=O)c1ccccc1)C	8.58
670	Pyrethrin II	COC(=O)C(C)=CC1C(C(=O)OC2CC(=O)C(=C2C)CC=CC=C)C1(C)C	9.06
671	Pyribencarb	COC(=O)NCc1cc(ccc1Cl)C(C)=NOCc1cccc(n1)C	7.58
672	Pyribenzoxim	COc1cc(nc(n1)Oc1cccc(c1C(=O)ON=C(c1ccccc1)c1ccccc1)Oc1nc(cc(n1)OC)OC)OC	9.25
673*	Pyributicarb	COc1cccc(n1)N(C)C(=S)Oc1cccc(c1)C(C)(C)C	9.5
674*	Pyridaben	CC(C)(C)N1N=CC(=C(Cl)C1=O)SCc1ccc(cc1)C(C)(C)C	10.06
675	Pyridalyl	FC(F)(F)c1ccc(nc1)OCCCOc1c(cc(cc1Cl)OCC=C(Cl)Cl)Cl	10.74
676	Pyridaphenthion	CCOP(=S)(OCC)OC1=NN(C(=O)C=C1)c1ccccc1	7.95
677	Pyridate	CCCCCCCCSC(=O)Oc1cc(nnc1-c1ccccc1)Cl	10.29
678	Pyrifenox	CON=C(Cc1cccnc1)c1ccc(cc1Cl)Cl	8.21
679	Pyrifluquinazon	CC(=O)N1C(=O)N(Cc2cc(ccc12)C(F)(C(F)(F)F)C(F)(F)F)NCc1cccnc1	8.06
680	Pyriftalid	COc1cc(nc(n1)Sc1cccc2c1C(=O)OC2C)OC	7.4

681	Pyrimethanil	Cc1cc(nc(n1)Nc1ccccc1)C	7.4
682	Pyrimidifen	CCOCCc1ccc(c(c1C)C)OCCNc1ncnc(c1Cl)CC	9.2
683*	Pyriminobac-methyl	CON=C(C)c1cccc(c1C(=O)OC)Oc1nc(cc(n1)OC)OC	7.37
684	Pyrimisulfan	COCc1cccc(c1NS(=O)(=O)C(F)F)C(O)c1nc(cc(n1)OC)OC	7.14
685*	Pyriofenone	COc1cc(c(c(c1OC)OC)C(=O)c1c(c(cnc1OC)Cl)C)C	8.87
686	Pyriproxyfen	CC(COc1ccc(cc1)Oc1ccccc1)Oc1ccccn1	9.43
687	Pyrithiobac	COc1cc(nc(n1)Sc1cccc(c1C(O)=O)Cl)OC	6.99
688*	Pyroquilon	O=C1CCc2cccc3c2N1CC3	6.09
689*	Pyroxasulfone	Cn1nc(c(c1OC(F)F)CS(=O)(=O)C1=NOC(C)(C)C1)C(F)(F)F	6.92
690	Pyroxsulam	COc1cc([n]2nc(nc2n1)NS(=O)(=O)c1c(nccc1C(F)(F)F)OC)OC	6.09
691	Quinalphos	CCOP(=S)(OCC)Oc1cnc2ccccc2n1	8.47
692	Quinclorac	OC(=O)c1c(ccc2cc(cnc12)Cl)Cl	5.1
693	Quinmerac	Cc1cnc2c(ccc(c2C(O)=O)Cl)c1	4.49
694	Quinoclamine	NC1=C(Cl)C(=O)c2ccccc2C1=O	5.83
695	Quinoxyfen	Fc1ccc(cc1)Oc1ccnc2cc(cc(c12)Cl)Cl	9.58

696	Quizalofop	CC(Oc1ccc(cc1)Oc1cnc2cc(ccc2n1)Cl)C(O)=O	8.21
697	Quizalofop-ethyl	CCOC(=O)C(C)Oc1ccc(cc1)Oc1cnc2cc(ccc2n1)Cl	9.21
698	Rimsulfuron	CCS(=O)(=O)c1cccnc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	6.46
699	Rotenone	COc1cc2c(cc1OC)C1C(CO2)Oc2c3c(ccc2C1=O)OC(C3)C(C)=C	8.32
700	Saflufenacil	CC(C)N(C)S(=O)(=O)NC(=O)c1cc(c(cc1Cl)F)N1C(=O)C=C(N(C)C1=O)C(F)(F)F	7.32
701	Schradan	CN(C)P(=O)(OP(=O)(N(C)C)N(C)C)N(C)C	5.13
702*	Sebuthylazine	CCNc1nc(nc(n1)NC(C)CC)Cl	7.55
703	Sebuthylazine-desethyl	CCC(C)Nc1nc(nc(n1)Cl)N	5.97
704*	Secbumeton	CCNc1nc(nc(n1)OC)NC(C)CC	6.925
705*	Sedaxane	Cn1cc(c(n1)C(F)F)C(=O)Nc1ccccc1C1CC1C1CC1	8.02
706	Sethoxydim	CCCC(=NOCC)C1=C(O)CC(CC(C)SCC)CC1=O	9.27
707	Siduron	CC1CCCCC1NC(=O)Nc1ccccc1	7.58
708	Silafluofen	CCOc1ccc(cc1)[Si](C)(C)CCCc1ccc(cc1)Oc1ccccc1)F	11.05
709	Silthiofam	Cc1[s]c(c(c1C)C(=O)NCC=C)[Si](C)(C)C	8.5

710*	Simazine	CCNc1nc(nc(n1)NCC)Cl	6.16
711	Simeconazole	C[Si](C)(C)CC(O)(Cn1cncn1)c1ccc(cc1)F	8.2
712*	Simeton	CCNc1nc(nc(n1)OC)NCC	5.1
713	Simetryn	CCNc1nc(nc(n1)SC)NCC	6.49
714*	Spinetoram	CCOC1C(OC)C(C)OC(OC2CC3CCC4C5CC(=O)OC(CC)CCCC(OC6CCC(C(C)O6)N(C)C)C(C)C(=O)C5=CC4C3C2)C1OC	9.13
715	Spinosyn A	CCC1CCCC(OC2CCC(C(C)O2)N(C)C)C(C)C(=O)C2=CC3C4CC(CC4C=CC3C2CC(=O)O1)OC1OC(C)C(OC)C(OC)C1OC	8.83
716	Spinosyn D	CCC1CCCC(OC2CCC(C(C)O2)N(C)C)C(C)C(=O)C2=CC3C4CC(CC4C(=CC3C2CC(=O)O1)C)OC1OC(C)C(OC)C(OC)C1OC	9.09
717*	Spirodiclofen	CCC(C)(C)C(=O)OC1=C(C(=O)OC21CCCCC2)c1ccc(cc1Cl)Cl	9.87
718	Spiromesifen	Cc1cc(c(c(c1)C)C1=C(OC(=O)CC(C)(C)C)C2(CCCC2)OC1=O)C	9.68
719	Spiromesifen alcohol	Cc1cc(c(c(c1)C)C1=C(O)C2(CCCC2)OC1=O)C	7.5
720*	Spirotetramat	CCOC(=O)OC1=C(C(=O)NC21CCC(CC2)OC)c1cc(ccc1C)C	8.09
721	Spirotetramat cis-enol	COC1CCC2(CC1)NC(=O)C(=C2O)c1cc(ccc1C)C	6.68

722	Spirotetramat cis-keto-hydroxy	COC1CCC2(CC1)NC(=O)C(O)(C2=O)c1cc(ccc1C)C	7.13
723	Spirotetramat monohydroxy	COC1CCC2(CC1)NC(=O)C(C2O)c1cc(ccc1C)C	6.05
724	Spiroxamine	CCCN(CC)CC1COC2(CCC(CC2)C(C)(C)C)O1	7.85
725	Starlicide	Cc1ccc(cc1Cl)N	5.88
726*	Strychnine	O=C1CC2OCC=C3CN4CCC56C4CC3C2C6N1c1ccccc51	3.59
727	Sulcotrione	CS(=O)(=O)c1ccc(c(c1)Cl)C(=O)C1C(=O)CCCC1=O	5.63
728	Sulfentrazone	CC1=NN(C(=O)N1C(F)F)c1cc(c(cc1Cl)Cl)NS(C)(=O)=O	6.37
729	Sulfometuron-methyl	COC(=O)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(cc(n1)C)C	6.2
730	Sulfosulfuron	CCS(=O)(=O)c1nc2cccc[n]2c1S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	7.24
731*	Sulfotep	CCOP(=S)(OCC)OP(=S)(OCC)OCC	8.54
732*	Sulfoxaflor	CC(c1ccc(nc1)C(F)(F)F)S(C)(=O)=NC#N	4.83
733	Sulprofos	CCCSP(=S)(OCC)Oc1ccc(cc1)SC	9.61
734	Tau-fluvalinate	CC(C)C(Nc1ccc(cc1Cl)C(F)(F)F)C(=O)OC(C#N)c1cccc(c1)Oc1ccccc1	10.17
735	TDCPP	ClCC(CCl)OP(=O)(OC(CCl)CCl)OC(CCl)CCl	8.57

736*	Tebuconazole	CC(C)(C)C(O)(CCc1ccc(cc1)Cl)Cn1cncn1	8.57
737	Tebufenozide	CCc1ccc(cc1)C(=O)NN(C(=O)c1cc(cc(c1)C)C)C(C)(C)C	8.43
738	Tebufenpyrad	CCc1nn(c(c1Cl)C(=O)NCc1ccc(cc1)C(C)(C)C)C	9.33
739*	Tebupirimfos	CCOP(=S)(OC(C)C)Oc1cnc(nc1)C(C)(C)C	9.38
740*	Tebutam	CC(C)N(Cc1ccccc1)C(=O)C(C)(C)C	8.28
741	Tebuthiuron	CNC(=O)N(C)c1nnc([s]1)C(C)(C)C	6.29
742	Tecloftalam	OC(=O)c1c(c(c(c(c(c1C(=O)Nc1cccc(c1Cl)Cl)Cl)Cl)Cl)Cl)Cl	8.18
743	Teflubenzuron	Fc1cccc(c1C(=O)NC(=O)Nc1cc(c(c(c1F)Cl)F)Cl)F	9.46
744	Tefuryltrione	CS(=O)(=O)c1ccc(c(c1COCC1CCCO1)Cl)C(=O)C1C(=O)CCCC1=O	6.83
745*	Tembotrione	CS(=O)(=O)c1ccc(c(c1COCC(F)(F)F)Cl)C(=O)C1C(=O)CCCC1=O	7.02
746*	Temephos	COP(=S)(OC)Oc1ccc(cc1)Sc1ccc(cc1)OP(=S)(OC)OC	9.35
747	Temephos sulfone	COP(=S)(OC)Oc1ccc(cc1)S(=O)(=O)c1ccc(cc1)OP(=S)(OC)OC	8.24
748	Tepraloxydim	CCC(=NOCC=CCl)C1=C(O)CC(CC1=O)C1CCOCC1	8
749	Terbacil	CC1=C(Cl)C(=O)N(C(=O)N1)C(C)(C)C	6.3

750	Terbufos	CCOP(=S)(OCC)SCSC(C)(C)C	9.35
751	Terbufos oxon sulfone	CCOP(=O)(OCC)SCS(=O)(=O)C(C)(C)C	5.82
752	Terbufos sulfone	CCOP(=S)(OCC)SCS(=O)(=O)C(C)(C)C	7.5
753*	Terbufos sulfoxide	CCOP(=S)(OCC)SCS(=O)C(C)(C)C	7.54
754	Terbumeton	CCNc1nc(nc(n1)OC)NC(C)(C)C	6.95
755	Terbumeton-desethyl	CCC(C)Nc1nc(nc(n1)OC)N	5.52
756	Terbuthylazine	CCNc1nc(nc(n1)NC(C)(C)C)Cl	7.6
757	Terbuthylazine-desethyl	CC(C)(C)Nc1nc(nc(n1)Cl)N	6.42
758	Terbutryn	CCNc1nc(nc(n1)SC)NC(C)(C)C	7.9
759	Tetrachlorvinphos	COP(=O)(OC)OC(=CCl)c1cc(c(cc1Cl)Cl)Cl	8.46
760	Tetraconazole	FC(F)C(F)(F)OCC(Cn1cncn1)c1ccc(cc1Cl)Cl	8.2
761	Tetraethyl pyrophosphate	CCOP(=O)(OCC)OP(=O)(OCC)OCC	5.95
762*	Tetrahydrophthalimide	O=C1NC(=O)C2C=CCCC12	2.12
763*	Tetramethrin	CC(C)=CC1C(C(=O)OCN2C(=O)C3=C(CCCC3)C2=O)C1(C)C	9.31
764	Thenylchlor	COc1cc[s]c1CN(C(=O)CCl)c1c(cccc1C)C	8.17

765	Thiabendazole	c1ccc2[nH]c(nc2c1)-c1c[s]cn1	4.02
766	Thiacloprid	Clc1ccc(cn1)CN1CCSC1=NC#N	5.06
767	Thiamethoxam	CN1COCN(Cc2cnc([s]2)Cl)C1=NN(=O)=O	3.43
768	Thiazopyr	COC(=O)c1c(c(c(nc1C(F)F)C(F)(F)F)C1=NCCS1)CC(C)C	8.54
769	Thidiazuron	O=C(Nc1ccccc1)Nc1cnn[s]1	6.12
770	Thiencarbazone-methyl	COC(=O)c1c[s]c(c1S(=O)(=O)NC(=O)N1N=C(OC)N(C)C1=O)C	5.52
771	Thifensulfuron-methyl	COC(=O)c1[s]ccc1S(=O)(=O)NC(=O)Nc1nc(nc(n1)OC)C	5.97
772*	Thifluzamide	Cc1nc(c([s]1)C(=O)Nc1c(cc(cc1Br)OC(F)(F)F)Br)C(F)(F)F	8.28
773	Thiobencarb	CCN(CC)C(=O)SCc1ccc(cc1)Cl	8.83
774*	Thiocyclam	CN(C)C1CSSSC1	1.91
775	Thiodicarb	CSC(C)=NOC(=O)N(C)SN(C)C(=O)ON=C(C)SC	6.61
776	Thiofanox sulfone	CNC(=O)ON=C(CS(C)(=O)=O)C(C)(C)C	4.41
777*	Thiofanox sulfoxide	CNC(=O)ON=C(CS(C)=O)C(C)(C)C	4.21
778	Thionazin	CCOP(=S)(OCC)Oc1cnccn1	7.06
779*	Thiophanate	CCOC(=O)NC(=S)Nc1ccccc1NC(=S)NC(=O)OCC	7.07

780	Thiuram	CN(C)C(=S)SSC(=S)N(C)C	5.9
781	Tiadinil	Cc1ccc(cc1Cl)NC(=O)c1[s]nnc1C	7.98
782	Tolclofos-methyl	COP(=S)(OC)Oc1c(cc(cc1Cl)C)Cl	8.83
783	Tolfenpyrad	CCc1nn(c(c1Cl)C(=O)NCc1ccc(cc1)Oc1ccc(cc1)C)C	9.38
784	Topramezone	CN1NC=C(C1=O)C(=O)c1ccc(c(c1C)C1=NOCC1)S(C)(=O)=O	4.58
785	Tralkoxydim	CCON=C(CC)C1=C(O)CC(CC1=O)c1c(cc(cc1C)C)C	9.53
786	Triadimefon	CC(C)(C)C(=O)C(Oc1ccc(cc1)Cl)n1cncn1	7.94
787*	Triadimefon metabolite BAY KWG 1323	CC(C)(CO)C(=O)C(Oc1ccc(cc1)Cl)n1cncn1	6.91
788	Triadimefon metabolite BAY KWG 1342	CC(C)(CO)C(O)C(Oc1ccc(cc1)Cl)n1cncn1	6.87
789	Triadimenol	CC(C)(C)C(O)C(Oc1ccc(cc1)Cl)n1cncn1	8.06
790	Triallate	CC(C)N(C(C)C)C(=O)SCC(Cl)=C(Cl)Cl	9.64
791*	Triamiphose	CN(C)P(=O)(N(C)C)n1nc(nc1N)-c1ccccc1	7.03
792	Triasulfuron	COc1nc(nc(n1)NC(=O)NS(=O)(=O)c1ccccc1OCCCl)C	6.01
793*	Triazamate	CCOC(=O)CSc1nc(nn1C(=O)N(C)C)C(C)(C)C	8.13

794	Triazophos	CCOP(=S)(OCC)Oc1ncn(n1)-c1ccccc1	7.96
795*	Triazoxide	[O-][n+]1nc(nc2ccc(cc12)Cl)-n1ccnc1	6.72
796	Tribenuron-methyl	COC(=O)c1ccccc1S(=O)(=O)NC(=O)N(C)c1nc(nc(n1)OC)C	6.87
797	Tribufos	CCCCSP(=O)(SCCCC)SCCCC	10.09
798	Trichlorfon	COP(=O)(OC)C(O)C(Cl)(Cl)Cl	4.48
799	Triclocarban	Clc1ccc(cc1)NC(=O)Nc1ccc(c(c1)Cl)Cl	8.99
800*	Tricyclazole	Cc1cccc2[s]c3nnc[n]3c12	5.36
801	Tridemorph	CCCCCCCCCCCCCN1CC(C)OC(C)C1	8.8
802	Trietazine	CCNc1nc(nc(n1)N(CC)CC)Cl	8.17
803	Trifloxystrobin	CON=C(C(=O)OC)c1ccccc1CON=C(C)c1cccc(c1)C(F)(F)F	9.02
804	Trifloxysulfuron	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1ncccc1OCC(F)(F)F)OC	6.91
805*	Triflumizole	CCCOCC(=Nc1ccc(cc1C(F)(F)F)Cl)n1ccnc1	9.16
806*	Triflumuron	FC(F)(F)Oc1ccc(cc1)NC(=O)NC(=O)c1ccccc1Cl	8.7
807	Trifluralin	CCCN(CCC)c1c(cc(cc1N(=O)=O)C(F)(F)F)N(=O)=O	9.76
808	Triflusulfuron-methyl	COC(=O)c1cccc(c1S(=O)(=O)NC(=O)Nc1nc(nc(n1)N(C)C)OCC(F)(F)F)C	7.65

809	Triforine	ClC(Cl)(Cl)C(NC=O)N1CCN(CC1)C(NC=O)C(Cl)(Cl)Cl	7.4
810	Trinexapac-ethyl	CCOC(=O)C1CC(=O)C(C(=O)C1)=C(O)C1CC1	7.13
811	Triphenyl phosphate	O=P(Oc1ccccc1)(Oc1ccccc1)Oc1ccccc1	8.66
812	Triticonazole	CC1(C)CCC(=Cc2ccc(cc2)Cl)C1(O)Cn1cncn1	8.17
813	Tritosulfuron	COc1nc(nc(n1)C(F)(F)F)NC(=O)NS(=O)(=O)c1ccccc1C(F)(F)F	7.17
814	Uniconazole	CC(C)(C)C(O)C(=Cc1ccc(cc1)Cl)n1cncn1	8.39
815	Valifenalate	COC(=O)CC(NC(=O)C(NC(=O)OC(C)C)C(C)C)c1ccc(cc1)Cl	7.91
816	Vamidothion	CNC(=O)C(C)SCCSP(=O)(OC)OC	4.52
817	Vamidothion sulfone	CNC(=O)C(C)S(=O)(=O)CCSP(=O)(OC)OC	3.4
818	Verbenone	CC1=CC(=O)C2CC1C2(C)C	6.39
819	Vernolate	CCCSC(=O)N(CCC)CCC	8.94
820*	Vinclozolin metabolite M2	CC(O)(C=C)C(=O)Nc1cc(cc(c1)Cl)Cl	8.17
821	Warfarin	CC(=O)CC(c1ccccc1)C1=C(O)c2ccccc2OC1=O	7.69
822*	XMC	CNC(=O)Oc1cc(cc(c1)C)C	6.68

823	Zoxamide	CCC(C)(NC(=O)c1cc(c(c(c1)Cl)C)Cl)C(=O)CCl	8.68

3.1.2. Dataset 2 (case study 2)

The present work deals with 715 datapoints for Rainbow trout (RT), 136 datapoints for Lepomis (LP) and 226 datapoints for Miscellaneous fish species to develop the q-RASAR models for toxicity prediction towards aquatic species in this study. The canonical smiles along with experimental values of aquatic toxicity end point [Log(1/LC50)] of aforementioned datapoints for RT, LP, and Miscellaneous species have been represented in the following **Tables 3.2, 3.3, and 3.4** respectively.

Table 3.2. Reported aquatic toxicity [Log(1/LC50)] data of curated 715 datapoints of RT(Rainbow trout) dataset. (* Test set compounds)

ID. No.	Canonical smiles	Log(1/LC50)
1*	Cl\C(=C/C2C(C(=O)OCc1c(F)c(F)c(c(F)c1F)C)C2(C)C)C(F)(F)F	9.84
2	Cl\C(=C/C3C(C(=O)OCc2cccc(c1ccccc1)c2C)C3(C)C)C(F)(F)F	9.21
4	N#CC(OC(=O)C1C(/C=C(/C)C)C1(C)C)c3cccc(Oc2ccccc2)c3	9.04
5	Cl/C(Cl)=C/C3C(C(=O)OC(C#N)c2ccc(F)c(Oc1ccccc1)c2)C3(C)C	8.97
6*	O=C(OCc1cc(oc1)Cc2ccccc2)C3C(\C=C(/C)C)C3(C)C	8.74
7	O=C2C(\Cl)=C(\SCc1ccc(cc1)C(C)(C)C)/C=N\N2C(C)(C)C	8.72
8	S=C(NC(C)(C)C)Nc2c(cc(Oc1ccccc1)cc2C(C)C)C(C)C	8.74
9*	Clc3cc(ccc3N[C@@H](C(=O)OC(C#N)c2cccc(Oc1ccccc1)c2)C(C)C)C(F)(F)F	8.80
11*	C1[C@@H]2[C@H]3[C@@H]([C@H]1[C@H]4[C@@H]2O4)[C@]5(C(=C([C@@]3(C5(Cl)Cl)Cl)Cl)Cl)Cl	8.50
12	CC(=CC1C(C1(C)C)C(=O)OCc2c(c(c(c(c2F)F)COC)F)F)C#N	8.51

13	Clc1cc(Cl)ccc1OP(=S)(OCC)OCC	8.40
14	Clc1c(OP(=S)(OCC)OCC)nc(Cl)c(Cl)c1	8.43
18	c12c(sc(n1)SCSC#N)cccc2	8.06
19	O=C(OC(C#N)c2cccc(Oc1ccccc1)c2)C3C(C)(C)C3(C)C	8.18
20*	O(c1ccccc1)c2cc(ccc2)COCC(c3ccc(OCC)cc3)(C)C	8.14
21	Clc1ccc(cc1)CC(Cl)C(=O)OC	7.94
22	CC(=CC1C(C1(C)C)C(=O)OCC2=CC(=CC=C2)OC3=CC=CC=C3)C	8.11
23	CC1(C(C1C(=O)O[C@@H](C#N)c2cccc(c2)Oc3ccccc3)C=C(Cl)Cl)C	8.17
24*	CC1(C(C1C(=O)OC(C#N)c2cccc(c2)Oc3ccccc3)C=C(Cl)Cl)C	8.17
25	ClC1C(Cl)C(Cl)C(Cl)C(Cl)C1Cl	8.00
26	[O-][N+](=O)c1cc(cc(c1O)C(C)(C)C)[N+]([O-])=O	7.85
27*	Clc1ccc(cc1)C(C(=O)OC(C#N)c3cccc(Oc2ccccc2)c3)C(C)C	8.07
28	S=P(OP(=S)(OCC)OCC)(OCC)OCC	7.95
29	ClC(Cl)(Cl)[N+]([O-])=O	7.53
30*	Clc3cc(ccc3Oc2ccc(NC(=O)NC(=O)c1c(F)cccc1F)c(F)c2)C(F)(F)F	8.00
31*	S=P(OCC)(OCC)Oc1nc2ccccc2nc1	7.78
32	CCOP(=S)(OCC)Oc1nc(Cl)n(n1)C(C)C	7.72
33	O=C(OC(C(F)(F)F)C(F)(F)F)\C=C\C3C(C(=O)OC(C#N)c2cccc(Oc1ccccc1)c2)C3(C)C	7.95
34	Clc2ccc(c1c(C#N)c(Br)c(n1COCC)C(F)(F)F)cc2	7.77
35	O=P(OCC)(Oc1ccc(SC)c(c1)C)NC(C)C	7.51
37	O=C2OC(C(=O)N2Nc1ccccc1)(c4ccc(Oc3ccccc3)cc4)C	7.53
38	c1([P@@](Oc2c(cc(Br)c(c2)Cl)Cl)(OC)=S)ccccc1	7.57

39	c1([Hg]OC(C)=O)ccccc1	7.49
40	C1([C@H]([C@@H]1C(O[C@@H]1C(=C(CC#C)C(C1)=O)C)=O)\C=C(\C)C)(C)C	7.40
41	O=P(O/C(=C/C(=O)OC)C)(OC)OC	7.27
42	c3ccccc3Oc(c2)cccc2COC(=O)C1C(C1(C)C)C=C(Cl)Cl	7.50
43	S=P(OCC)(SCSCC)OCC	7.30
44	O=C(Oc1c(cc(cc1C(C)CC)[N+]([O-])=O)[N+]([O-])=O)OC(C)C	7.37
45	FC(F)(F)c1cccc(c1)\C(=N\OCc2ccccc2C(=N\OC)\C(=O)OC)C	7.41
46	O=C(O\C1=C(\C(=O)OC12CCCC2)c3c(cc(cc3C)C)C)CC(C)(C)C	7.36
47*	Clc1cc(I)c(Cl)cc1OP(=S)(OC)OC	7.41
48	S(P(SCCCC)SCCCC)CCCC	7.24
49	O=C(OCC)C(SP(=S)(OC)OC)CC(=O)OCC	7.26
50	Cl\C1=C(/Cl)C2(Cl)C(Cl)(Cl)C1(Cl)C4C2C3OC3C4Cl	7.29
52	OO	6.19
53	Clc2c(nn(c2C(=O)NCc1ccc(cc1)C(C)(C)C)C)CC	7.16
54	CCOc1cc(OP(=S)(OC)OC)nc(CC)n1	7.09
55*	Clc2c(NC(CC)c1ccc(OC(F)F)cc1)ncnc2C	7.12
56	Clc1=c(Cl)c(Cl)=c(Cl)c(Cl)=c1Cl	6.98
57	O=C(OCCCC)N(SN(C(=O)Oc2cccc1c2OC(C1)(C)C)C)C	7.11
58	Clc2ccc(\C(=C(/Cl)Cl)c1ccc(Cl)cc1)cc2	7.00
59	CC(C)(CC)C(=O)O\C2=C(\C(=O)OC12CCCCC1)c3cc(Cl)cc(Cl)c3	7.07
60	O=C(OCC)CCN(SN(C(=O)Oc2cccc1c2OC(C1)(C)C)C)C(C)C	7.03
61*	Clc1c(C#N)c(Cl)c(C#N)c(Cl)c1Cl	6.84
62	O=C(OCN1C(=O)CN(C1=O)CC#C)C2C(\C=C(/C)C)C2(C)C	6.92

63*	CCOP(=S)(OCC)OC1=NN(C=N1)C2=CC=CC=C2	6.92
64	c12c([nH]c(n2)NC(=O)OC)ccc(c1)Sc1ccccc1	6.87
65*	O=C(O)C(=O)O.S1SSCC(N(C)C)C1	6.83
66*	S1SSCC(N(C)C)C1	6.66
67	O=C(OC(C)(C)C)c1ccc(cc1)CO\N=C\c3c(nn(c3Oc2ccccc2)C)C	7.01
68	NC1=CC=CC=C1	6.36
69*	O=S(OCC#C)OC2CCCCC2Oc1ccc(cc1)C(C)(C)C	6.91
71	O=C(NC)\C(=N\OC)c1c(cccc1)COc2cc(ccc2C)C	6.88
73	S=C(SSC(=S)N(C)C)N(C)C	6.72
74	O=Nc2ccc(NNC(=O)c1ccccc1)cc2	6.70
75	O=C(Oc1c(cc(cc1C(C)CC)[N+]([O-])=O)[N+]([O-])=O)\C=C(/C)C	6.81
77	Brc1ccc(cc1)c2ccc(cc2)C6Cc3ccccc3C(C=4C(=O)c5c(SC=4O)cccc5)C6	7.02
78	C1C(CC2=CC=CC=C2C1C3=C(OC4=CC=CC=C4C3=O)O)C5=CC=C(C=C5)C6=CC=C(C=C6)Br	7.01
80	CN=C=S	6.14
81	COP(=S)(OC)SCSC1=CC=C(C=C1)Cl	6.75
83	n1c(c(c(n2ncnc12)N)CCCCCCCC)CC	6.63
84	O=C1/C=C\SN1CCCCCCCC	6.52
85*	c1(c(cc(cc1CCCCCCCC)[N+](=O)[O-])[N+](=O)[O-])O	6.65
86	N#CC=1S\C2=C(/SC=1C#N)C(=O)c3c(C2=O)cccc3	6.63
87	FC(F)(F)c2nc(OCc1ccccc1C(=C\OC)/C(=O)OC)ccc2	6.69
88	Clc1cc(Cl)ccc1n1nc(C(O)=O)nc1C(Cl)(Cl)Cl	6.67
89*	S=P(OCC)(OCC)SCN1\N=N/c2ccccc2C1=O	6.64

90	C(c1nc(n(n1)c1c(cc(cc1)Cl)Cl)C(Cl)(Cl)Cl)(=O)OCC	6.70
91	[O-][N+](=O)c1cc(cc([N+]([O-])=O)c1N(CCCC)CC)C(F)(F)F	6.62
92*	[O-][N+](=O)c1cc(cc([N+]([O-])=O)c1N(CCC)CCC)C(F)(F)F	6.58
93	CCOP(=S)(OCC)OC(C(Cl)(Cl)Cl)Cl	6.58
94	FC(F)(F)c1ccc(cc1)COc2ccc(cc2)C6Cc3ccccc3C(C=4C(=O)c5c(OC=4O)cccc5)C6	6.78
95*	c1cc(c(cc1c1c(ccc(c1)F)NC(=O)c1cn(nc1C(F)F)C)Cl)Cl	6.64
96	Clc1nc(nc(Cl)n1)Nc2ccccc2Cl	6.46
97	OC1=CC=C(C=C1)O	6.06
99	S=P(Oc1ccc(cc1)[N+]([O-])=O)(OCC)c2ccccc2	6.47
100	c12ccccc1cccc2	6.07
101	O=P(OC)(OC)SCN1c2ncc(Cl)cc2OC1=O	6.45
102*	O=P(SC(CC)C)(OCC)SC(CC)C	6.32
103	Clc2ccc(NC(=O)NC(=O)c1c(F)cccc1F)cc2	6.38
104	O=C(O\N=C(/C(C)(C)C)CSC)NC	6.23
105	CCN(CC(=C)C)C1=C(C=C(C=C1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	6.39
106	C(C=C)=C	5.56
107	O=C\C=C	5.54
108	Clc1cc(cc(Cl)c1C)C(=O)NC(C(=O)CCl)(C)CC	6.32
109	FC(F)(F)c1ccc(cc1)\C=C\C(=N/N/C2=N/CC(C)(C)CN2)/C=C/c3ccc(cc3)C(F)(F)F	6.46
110	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1O	6.19
111	ClCC(=O)N(c1ccccc1)C(C)C	6.07

113	O(c1c(cc(cc1)Cl)Cl)c1ccc(O)cc1	6.13
114	COC(=O)NC1=NC2=CC=CC=C2N1	6.00
115	O=C(OC)\C(=N\OC)c1c(cccc1)COc2ccccc2C	6.22
116	Clc1cc(Cl)c(Cl)c(Cl)c1Cl	6.12
117*	ClC(Br)(Cl)C(Br)OP(=O)(OC)OC	6.29
118	CCOP(=S)(CC)OC1=CC(=C(C=C1Cl)Cl)Cl	6.22
119	O=C2O\C(=N/N2c1c(Cl)cc(Cl)c(OCC#C)c1)C(C)(C)C	6.21
120	CC(C)(C)C(=O)C1C(=O)c2ccccc2C1=O	6.01
121	Fc1nc(OCC(=O)OC(C)CCCCCC)c(Cl)c(c1Cl)N	6.21
122	C(=N/NC(Nc1ccc(cc1)OC(F)(F)F)=O)(\Cc1ccc(cc1)C#N)c1cc(C(F)(F)F)ccc1	6.34
123	N#Cc3cncc3c1cccc2OC(F)(F)Oc12	6.03
124	N1(C(c2ccccc2C1=O)=O)SC(Cl)(Cl)Cl	6.09
125	C(SC#N)SC#N	5.73
126	Clc2cc(ccc2Oc1ccc([N+]([O-])=O)c(OCC)c1)C(F)(F)F	6.16
127	Fc3ccc(Oc1c2c(Cl)cc(Cl)cc2ncc1)cc3	6.06
128	FC(F)(F)c2nc1cc(Cl)c(Cl)cc1n2C(=O)Oc3ccccc3	6.11
129*	Clc1ccc(cc1)CN(C(=O)Nc2ccccc2)C3CCCC3	6.04
130	c1(ccc(O)cc1)CCCCCCCCC	5.87
131	c1c(N(C)C(=O)[C@H](Oc2ccc(cc2)Oc2nc3ccc(cc3o2)Cl)C)c(ccc1)F	6.16
132	Clc2cc(Cl)ccc2Oc1ccc(OC(C(=O)OC)C)cc1	6.04
133	CC1=CC(=CC(=C1)OC(=O)NC)C(C)C	5.83
134	c1(ccc(C(C)(C)C)cc1)OC[C@@H](O[S@@](OCCCl)=O)C	6.02
135	O=C2N(C(=N/C(C)(C)C)/SCN2c1ccccc1)C(C)C	5.97

136	C1(C(OCC)=O)=NOC(C1)(c1ccccc1)c1ccccc1	5.94
138*	CC(C)OC(=O)C(C1=CC=C(C=C1)Br)(C2=CC=C(C=C2)Br)O	6.09
139*	ClCC(=O)N(c1c(cccc1CC)C)COCC	5.87
140	Clc2cccc(Cl)c2C(=O)NCc1ncc(cc1Cl)C(F)(F)F	6.03
141	CCC(C)NC1=C(C=C(C=C1[N+](=O)[O-])C(C)(C)C)[N+](=O)[O-]	5.90
142	Clc1c([N+]([O-])=O)c(Cl)c(Cl)cc1Cl	5.85
143	[O-][N+](=O)C1=CC(=CC=C1)[N+]([O-])=O	5.66
144	OC(=S)c1cccc2nnsc12	5.69
145	S=P(OC)(OC)O\C=C(/C)C(=O)OC	5.78
146	S=P(OC)(OC)Oc1nc(nc(c1)C)N(CC)CC	5.88
147	Clc1c(OP(=S)(OC)OC)nc(Cl)c(Cl)c1	5.90
148	O=C(n1nc(nc1SCC(=O)OCC)C(C)(C)C)N(C)C	5.86
149	O=P(Sc1ccccc1)(Sc2ccccc2)OCC	5.86
150	Clc4ccccc4Oc3ncnc(Oc2ccccc2C(=N\OC)/C=1OCCON=1)c3F	6.02
151*	Clc2ccc(OP(=S)(/N=C(/N)C)Oc1ccc(Cl)cc1)cc2	5.93
152	Clc1ccc(cc1)C(c2ccccc2)C(=O)C4C(=O)c3ccccc3C4=O	5.92
153	Clc1ccc(cc1)C(O)(c2ccc(Cl)cc2)C(=O)OC(C)C	5.88
154	O=C(OC)\C(=C\OC)c3ccccc3Oc2ncnc(Oc1c(C#N)cccc1)c2	5.93
155*	Clc1cc(cc(Cl)c1N)[N+]([O-])=O	5.63
156*	ClC(Cl)(Cl)CC1(OC1)c2cc(Cl)cc(Cl)c2	5.82
157*	Clc1cc(O)c(Cl)cc1Cl	5.61
158	c3ccc2cc1ccccc1cc2c3	5.56
159	c1c(cc(c(c1Cl)OCCCOc1cccc(n1)C(F)(F)F)Cl)OC\C=C(\Cl)Cl	5.99

160	O=C(OCC)\C=C(\C=C\CC(C)CCCC(C)C)C	5.73
161	C1=CCC2C(C1)C(N(C2=O)SC(Cl)(Cl)Cl)=O	5.78
162*	Clc1cc(Cl)ccc1OP(=S)(OCC)SCCC	5.81
163*	Clc1cc(c(O)cc1)Cc2cc(Cl)ccc2O	5.69
164*	c1(c(cccc1)c1cc(c(c(c1)F)F)F)NC(=O)c1cn(nc1C(F)F)C	5.84
165	n2c1c(cccc1)nc2c3ncsc3	5.56
166*	CCOP(=S)(OCC)SC(CCl)N1C(=O)C2=CC=CC=C2C1=O	5.85
167	[O-][N+](=O)c2ccc(Oc1ccc(cc1[N+]([O-])=O)C(F)(F)F)cc2	5.77
168*	O=C(O\N=C\C(SC)(C)C)NC	5.52
169	FC(F)(F)c2c(\N=C(\n1ccnc1)COCCC)ccc(Cl)c2	5.78
170*	Clc1nc(nc(Cl)c1)c2ccccc2	5.57
171	Clc2ccc1c(OC(=O)N1CSP(=S)(OCC)OCC)c2	5.77
172	FC(F)(F)Oc1ccc(cc1)N(C(=O)OC)C(=O)N3/N=C4/c2c(cc(Cl)cc2)C[C@@]4(OC3)C(=O)OC	5.91
173	O=C(Oc1cc(c(SC)c(c1)C)C)NC	5.54
174	O=C(OCC)NCCOc2ccc(Oc1ccccc1)cc2	5.66
175*	c12c(c(ccc1C)C)cccc2	5.37
176	Clc2cc(Cl)ccc2Oc1cc(C(=O)OC)c([N+]([O-])=O)cc1	5.71
177	Clc3ccc(/N=C(\SCc1ccc(Cl)cc1)Cn2ncnc2)c(Cl)c3	5.79
178*	C1=CC(=CC(=C1)OC2=CC=CC(=N2)C(=O)NC3=CC=C(C=C3)F)C(F)(F)F	5.74
179	Clc1cc(cc(Cl)c1OP(=S)(OC)OC)C	5.64
180	COP(=O)(C(C(Cl)(Cl)Cl)O)OC	5.57
181*	Clc1ccc(cc1)C(O)(c2ccc(Cl)cc2)C(=O)OCC	5.67

182	n1c(C#CC)cc(nc1Nc2ccccc2)C	5.48
183	O=C2N(Nc1ccccc1)C(\SC)=N/[C@]2(c3ccccc3)C	5.62
184	COP(=S)(OC)Oc1cc(Cl)c(Cl)cc1Cl	5.64
185*	FC(F)(F)c1cc(c(c([N+]([O-])=O)c1N)N(CC)CC)[N+]([O-])=O	5.63
186	CC(=O)CP(=O)(OC)OC	5.34
187	O=S(=O)(SCC(N(C)C)CSS(=O)(=O)c1ccccc1)c2ccccc2	5.75
188	Clc1ccc(c(Cl)c1C)C(=O)c3c(nn(c3OCC(=O)c2ccc(cc2)C)C)C	5.75
189	Clc1ccc(c(Cl)c1)C(=O)c3c(nn(c3OCC(=O)c2ccccc2)C)C	5.71
190	Clc1cccc(c1Cl)c2c(C#N)cnc2	5.47
191	S=P(OC)(OC)Oc1ccc(SC)c(c1)C	5.54
192*	Brc2ccc(OC)c(C(=O)c1c(cc(OC)c(OC)c1OC)C)c2C	5.70
193	FC(F)(F)c1cc(c(c([N+]([O-])=O)c1N)N(CCC)CCC)[N+]([O-])=O	5.63
194	O=C(SCc1ccccc1)N(CCC)CCC	5.48
195*	CC(NC1=NC(NC2CC2)=NC(SC)=N1)(C)C	5.47
196	O=P(OP(=O)(OCC)OCC)(OCC)OCC	5.52
197	c1cccc(CO\N=C(\C(=N\OC)\C(C)=N\OC)C)c1\C(C(=O)NC)=N/OC	5.64
198*	ClCC(=O)N(c1c(cccc1CC)CC)CCOCCC	5.54
199	O=C(NC)C(OC)c1c(cccc1)COc2cc(ccc2C)C	5.52
200	CC(C)N(C(C)C)C(=O)SCC(Cl)=C(Cl)Cl	5.50
201	O=C(N(c1c(cccc1C)C)C(C(=O)OC)C)COC	5.46
202*	S(SC)C	4.99
204	O=C(Oc1cc(c(c(c1)C)C)C)NC	5.29
205*	C(C(N)=O)(C#N)(Br)Br	5.39

206	O=C(SCCCCCCCC)Oc2cc(Cl)nnc2c1ccccc1	5.54
207	O1C[C@@H](C)O[C@@]1(Cn1ncnc1)c1c(Cl)cc(Oc2ccc(Cl)cc2)cc1	5.57
208	O=S(=O)(NCCSP(=S)(OC(C)C)OC(C)C)c1ccccc1	5.56
209	S(c1nc(nc(n1)NC(C)(C)C)NCC)C	5.34
210	IC	5.10
211*	Clc1ccc(c(Cl)c1)C(CCC)Cn2ncnc2	5.39
212	N(C(SCC(=C)Cl)=S)(CC)CC	5.29
213	C([N+](CCCCCCCCCC)(C)C)CCCCCCCCC.[ClH-]	5.50
214	c12[C@@](c3c(cccc3)c1ccc(c2)Cl)(C(OC)=O)O	5.37
215	O=C(OCCOCC)C(Oc2ccc(Oc1ncc(cc1Cl)C(F)(F)F)cc2)C	5.56
216*	n1(cncn1)C[Si](c1ccc(F)cc1)(C)c1ccc(cc1)F	5.42
217	CC(C)OC1=C(C=C(C(=C1)N2C(=O)OC(=N2)C(C)(C)C)Cl)Cl	5.46
218	Clc2cc(ccc2NS(=O)(=O)c1ccc(Cl)c(c1)C(F)(F)F)[N+]([O-])=O	5.53
219	ClCC(=O)N(c1c(cccc1CC)C)C(COC)C	5.36
220	S=P(Oc1cc(c(cc1)[N+]([O-])=O)C)(OC)OC	5.33
221	c1(c(cc(Cl)c(c1)Cl)Cl)OCC(O)=O	5.29
222	O1N(CCC1)C([C@@H](C)Oc1ccc(cc1)Oc1c(cc(cn1)Cl)Cl)=O	5.47
223	O=C(Nc1ccc(O)c(Cl)c1Cl)C2(C)CCCCC2	5.35
224	COc1ccc(Cl)cc1Cl	5.10
225	n1cc(c(c(c1OC)C(=O)c1c(c(c(cc1C)OC)OC)OC)C)Cl	5.40
226	Clc1ccc(c(Cl)c1)C(OC\C=C)Cn2ccnc2	5.30
227	Clc1ccc(cc1)CCC(C#N)(c2ccccc2)Cn3ncnc3	5.35
228	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	5.40

229	c1cc(ccc1)c2ccccc2	5.01
230*	C(c1ccccc1)(c1ccccc1)(c1ccccc1)N1CCOCC1	5.34
231*	S=P(Oc1ccc(cc1)[N+]([O-])=O)(OCC)OCC	5.29
232	Fc1cccc(F)c1COCC3(O\N=C(\c2sccc2C)C3)C	5.34
233	n1(ncnc1)C(\[C@H](O)C(C)(C)C)=C/c1c(cc(cc1)Cl)Cl	5.31
234	Clc2ccc(/C=C(/n1ncnc1)C(O)C(C)(C)C)c(Cl)c2	5.31
235	O=S(C(\C=N\OC(=O)NC)(C)C)C	5.11
236	c12c(cccc2)s[nH]c1=O	4.98
237	ClC4(Cl)C(c1ccc(OCC)cc1)(C(=O)OC(C#N)c3cccc(Oc2ccccc2)c3)C4	5.48
238	BrC1CC(OC1)(c2c(Cl)cc(Cl)cc2)Cn3ncnc3	5.35
239	O=C(c1c(onc1)C2CC2)c3ccc(cc3S(=O)(=O)C)C(F)(F)F	5.33
240	O=C(Oc1cccc(c1)NC(=O)OC)Nc2cc(ccc2)C	5.24
242	c1(N(CC(OCC)=O)C(CCl)=O)c(cccc1CC)CC	5.23
243*	S=C1/N=C\NN1CC(O)(Cc2ccccc2Cl)C3(Cl)CC3	5.27
244	C=O	4.21
245*	Clc1ccc(c(Cl)c1)N3C(=N\c2c(cc(F)cc2)C3=O)/n4ncnc4	5.30
246*	c1(c(ccc(c1)N)Cl)Cl	4.92
247	O=C1\C3=C(/NC(=O)N1C2CCCCC2)CCC3	5.07
248	Clc1ccc(cc1)C(C#N)(CCCC)Cn2ncnc2	5.16
249*	ClCC(=O)N(C\C=C)C\C=C	4.94
250	COc1c(ccc(c1F)c2cc(c(c(n2)C(=O)OC)Cl)N)Cl	5.23
251*	Clc1ccc(cc1)CC2C(O)(C(CC2)(C)C)Cn3ncnc3	5.18
252	O(c2ccc(c1ccccc1)cc2)C(n3ncnc3)C(O)C(C)(C)C	5.20

253	ClCC(=O)N(\C(=C(\C)C)c1ccccc1)CCOCC	5.13
254*	c1(Nc2ccccc2)ccccc1	4.89
255	c1(CS(C2=NOC(C)(C)C2)(=O)=O)c(C(F)(F)F)nn(c1OC(F)F)C	5.25
256	CNc1nc(NC(C)C)nc(SC)n1	4.99
257	CC(C)Oc1ccc(C(=O)C(C)(C)NC(=O)c2sccc2C)c(C)c1	5.20
258	C1(C(Nc2ccccc2)=O)=C(OCCS1)C	5.01
259	O=C3\C4=C(/C(=O)N3c2c(F)cc1OCC(=O)N(c1c2)CC#C)CCCC4	5.19
260	c12c(ccc(c1)OCC(O)=O)cccc2	4.93
261*	Fc2ccc(N(C(=O)c1ccccc1)[C@@H](C(=O)OC(C)C)C)cc2Cl	5.18
262	n1c(cc(nc1Nc2ccccc2)C3CC3)C	4.97
263*	CCSc1nc(NC(C)C)nc(NC(C)C)n1	5.02
264*	c1(nc(=O)[nH]c(n1)NCC)NC(C)(C)C	4.93
265	Clc1ccc(c(Cl)c1)C1(OCC(O1)CC)Cn1ncnc1	5.12
266	Clc1cc(Cl)c(O)cc1	4.79
267	c1nc(Cl)c(cc1)C(=O)Nc1c(cccc1)c1ccc(cc1)Cl	5.10
268	C(=O)(OCC)COc1c(cc(c(c1)n1ncc(c(c1=O)C)C(F)(F)F)F)Cl	5.18
269*	S=P(Oc1ccc(cc1)[N+](=O)[O-])(OC)OC	4.99
270*	c1(/[N+](=N/c2ccccc2)/[O-])ccccc1	4.87
271	ClC=CCCl	4.60
272	O=C1c2ccccc2C(=O)C1C(=O)C(c1ccccc1)c1ccccc1	5.08
273	Clc1ccc(N(C(=O)CSP(=S)(OC)OC)C(C)C)cc1	5.12
274	C1[C@@H](C)O[C@@H](CN1CCCCCCCCCCCC)C	5.00
275	O=C1OC(C(=O)N1c1cc(Cl)cc(Cl)c1)(C=C)C	5.00

276	Clc1ccc(cc1)C(OCC#C)C(=O)NCCc1ccc(c(c1)OC)OCC#C	5.15
277	CCCCCCCCSCCO	4.82
278*	C(CCCCCC)CCC(=O)C	4.75
279	O=Cc1ccco1	4.50
280	S=P(OCC)(OCC)Oc1nc(nc(c1)C)C(C)C	4.99
281*	Clc1cc(ccc1N)C(F)(F)F	4.80
282	O=C(/C(=N/OC(=O)NC)/SC)N(C)C	4.85
283	Clc1ccc(NC(=O)N(OC)C)cc1Cl	4.90
284	Clc1cc(OC)c(Cl)cc1OC	4.80
285	O=C(/C=C(/c1ccc(Cl)cc1)\c1ccc(OC)c(OC)c1)N1CCOCC1	5.06
286*	FC(F)c1nc(c(C2=NCCS2)c(c1C(=O)OC)CC(C)C)C(F)(F)F	5.07
287	Clc1ccccc1n1nnn(c1=O)C(=O)N(C1CCCCC1)CC	5.01
289	Clc1cc(Cl)ccc1OCCCC(=O)O	4.85
290	Clc1ccc(Cl)c(N)c1	4.67
291	S=P(OC)(SCCSCC)OC	4.82
292	ClCC(=O)N(c1c(cccc1C)C)CCOC	4.82
293	BrC	4.39
294	CCc1cccc(c1N(C(C)COC)C(=O)CCl)C	4.86
295	c1(ccc(OCCOCCO)cc1)CCCCCCCCC	4.89
296	CC(C)(C)C(=O)C(n1cncn1)Oc1ccc(cc1)Cl	4.86
297	Clc1ccc(cc1)C(O)(c1ccccc1Cl)c1cncnc1	4.91
298	c1(c(ccc(c1)C)[N+](=O)[O-])O[P@@](=S)(N[C@@H](CC)C)OCC	4.91
299	O=C(SCC)N(CC(C)C)CC(C)C	4.71

300	O=C(Oc1cccc(/N=C/N(C)C)c1)NC	4.71
301	Clc1cc(c(OCCCC(=O)O)cc1)C	4.73
302*	Clc1ccc(cc1)CCC(O)(C(C)(C)C)Cn1ncnc1	4.84
303*	Clc1nc(nc(n1)NC(C)C)NCC	4.68
304	O=C(SCC)N(CC)C1CCCCC1	4.68
305*	O=C(SCCC)N(CCC)CCC	4.65
306	Clc1c(C(=O)OC)c(Cl)c(Cl)c(C(=O)OC)c1Cl	4.85
307	Clc1cc(C(=O)NC(C#C)(C)C)cc(Cl)c1	4.74
308	O=C(N(c1c(cccc1C)C)[C@@H](C(=O)OC)C)Cc1ccccc1	4.82
309	CN(C(=O)Nc1cc(c(cc1)Br)Cl)OC	4.77
310	S(c1nc(nc(n1)NC(C)C)NCC)C	4.66
311	n1(c(c(c(=O)n1C(C)C)c1ccccc1C)N)C(=O)SCC=C	4.80
312	Clc1ccc(NC(=O)CC)cc1Cl	4.61
313	N#CCON=C(C#N)c1ccccc1	4.52
314	n12c(=NCC2)ssc1=S	4.49
315	O=P(O/C(=C(\Cl)/C(=O)N(CC)CC)/C)(OC)OC	4.70
316	O=C(N1C(C)CCCC1)CSP(=S)(OCCC)OCCC	4.77
317	CCCC(=O)N(CSP(=S)(OC)OC)c1ccccc1Cl	4.79
318	C(=O)(C(Cl)Cl)N1C(O[C@@H](C1)c1occc1)(C)C	4.65
319*	O=P(O/C(=C/C(=O)N(C)C)/C)(OC)OC	4.58
320	C(CCCC/C=C\CCCC)CCCOC(=O)C	4.60
321	O=C(Nc1ccccc1)c1ccccc1I	4.70
322	ClC(Cl)(Cl)c1nc(Cl)ccc1	4.55

323	C([C@@H]1CCC(=CC1)C)(C)(C)O	4.37
324	O=C(Oc1cc(c(c(c1)C)S(=O)C)C)NC	4.56
325*	O=C(N(CC)CC)C(Oc1cccc2ccccc12)C	4.61
326	O=C(OC(C)C)C(=C1SCCS1)C(=O)OC(C)C	4.63
327*	CCCC(=O)c1c(C)c(c(C)cc1C)C1CC(=C(C(=O)C1)/C(=N/OCC)/CC)O	4.76
328	c1(Oc2ccc([N+](=O)[O-])cc2)c(cc(Cl)cc1)Cl	4.61
329	O=P(O/C(=C/C(=O)NC)/C)(OC)OC	4.50
330	Clc1ccc(cc1)C(c1ccc(Cl)cc1)C(Cl)(Cl)Cl	4.70
331*	N#Cc1c(Cl)cccc1Cl	4.38
332	Clc1cc(cc(Cl)c1)N1C(=O)C2(CC2(C1=O)C)C	4.60
333*	n1nc2sc3cccc(c3n2c1)C	4.41
334*	S=P(OCC)(OCC)Oc1nn(c2ccccc2)c(=O)cc1	4.66
335	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)Nc1ccccc1C(=O)C1CC1	4.74
336*	c1ccc(cc1)C(CC(c1ccc(cc1)c1ccc(cc1)Br)O)c1c(oc2ccccc2c1=O)O	4.82
338	c1(nc(nc(n1)SC)NCCCOC)NC(C)C	4.53
339	FC(F)(F)c1cc(ccc1)n1ncc(c(Cl)c1=O)NC	4.57
340	O=C(O)c1ccc(CO/N=C/c2c(Oc3ccccc3)n(C)nc2C)cc1	4.65
341	O=C(Nc1nc(OC)cc(Cl)n1)NS(=O)(=O)c1ccccc1C(=O)OCC	4.69
343*	CC(C)(C)N(C(=O)c1ccccc1)NC(=O)c1ccc(cc1)Cl	4.59
344	Clc1c(N)c(Cl)c(nc1Cl)C(=O)O	4.44
345	O=C(NC)CSP(=O)(OC)OC	4.37
346*	c1(c(ccc(c1)C)O)C	4.12
347	c1cc(cc2sc(nc12)[C@H](NC(=O)[C@H](NC(=O)O)C(C)C)C)F	4.56

348*	O=C(OC(C)C)Nc1cc(OCC)c(OCC)cc1	4.46
349*	FC(F)C(=O)O	3.98
350*	c1cc(cc2c1nc([C@@H](C)NC(=O)[C@@H](C(C)C)NC(=O)OC(C)C)s2)F	4.58
351	O=C(c1ccccc1C)Nc1cccc(OC(C)C)c1	4.43
352*	C(=C(\N(Cc1cnc(cc1)Cl)CC)/NC)\[N+](=O)[O-]	4.43
353*	O=C(Oc1c(c(=O)n2n1CCOCC2)c1c(cc(cc1CC)C)CC)C(C)(C)C	4.59
354	n1c(cc(nc1Nc1ccccc1)C)C	4.28
355	O=C(O)C1(CC1)C(=O)Nc1ccc(Cl)cc1Cl	4.40
358	Clc1cc(Sc2ccc(Cl)cc2)c(Cl)cc1Cl	4.47
359*	FC(F)(F)c1cc(c2c(=O)c(cn(c2)C)c2ccccc2)ccc1	4.45
360*	Clc1cc(c(/N=C/N(C)C)cc1)C	4.23
361*	O=C1NC(=O)C2C1CC=CC2	4.10
362	CC(C)Nc1nc(nc(n1)OC)NC(C)C	4.27
363*	O=S(=O)(Oc1cc2c(OCC2(C)C)cc1)CC	4.32
364*	O=C(Nc1nc(cc(n1)C)C)NS(=O)(=O)c1ccccc1C(=O)OC	4.46
366	O=C(Oc1cc(c(c(c1)C)N(C)C)C)NC	4.20
367	O(c1nc(nc(n1)NC(C)(C)C)NCC)C	4.21
368	CC(C(=O)O)Oc1c(Cl)cc(Cl)c(Cl)c1	4.26
369*	Clc1ccc(OC(C)CON/C(=C\2/C(=O)CC(C3CCCSC3)CC2=O)/CCC)cc1	4.49
370	c1(c2cc[n+](C)cc2)cc[n+](C)cc1.[Cl-].[Cl-]	4.23
371	Clc1ccccc1CN1OCC(C1=O)(C)C	4.19
372	C1=CC2C3C=CC(C3)C2C1	3.92

373	O=C(SCC)N1CCCCCC1	4.07
374	O=S(CCSP(=O)(OC)OC)CC	4.16
375*	Clc1nc(nc(n1)NC(C)C)NC(C)C	4.12
377	O(c1ccc2c(C(=CC(N2)(C)C)C)c1)CC	4.08
378	N(CCCC)CCCC	3.86
379	O(c1nc(nc(n1)NC(C)CC)NCC)C	4.10
381	CC(C)C(c1ccc(cc1)OC(F)(F)F)(c1cncnc1)O	4.23
382*	Clc1cc(ccc1OC)NC(=O)N(C)C	4.08
383	c1cc(c2cc[n+](cc2)C)cc[n+]1C	3.99
384*	O=C1C(=C(C)C(OC(=O)C2C(C=C(C)C)C2(C)C)C1)CC=C	4.20
385	O=C(Nc1ccccc1)Nc1snnc1	4.06
386	O=S1(=O)C(=C(OCC1)C)C(=O)Nc1ccccc1	4.13
387*	Clc1cc(NC(=O)N(C)C)ccc1C	4.03
388*	O=c1c(n[nH]c(Cl)c1)c1ccccc1	4.01
389	O=C(NC(C)C)CSP(=S)(OCC)OCC	4.15
390	[C@H]1(CC=C(C)C(=O)C1)C(=C)C	3.87
391*	c12c3[n+](CC[n+]1cccc2)cccc3	3.94
392*	c1(O[C@@H](C(=O)O)C)cc(ccc1)Cl	3.98
393	C(C(=O)N(C)C)CCCCCCCC	3.98
394	c1c(OC)nc(Oc2c(c(ccc2)/C(=N/OC)/C)C(=O)OC)nc1OC	4.23
395	O=C(c1ccc(Cl)cc1Cl)Cn1ccnc1	4.08
396*	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)c1ncccc1C(F)(F)F	4.27
397	[O-][N+](=O)c1cc(cc([N+](=O)[O-])c1N(CCC)CCC)S(=O)(=O)C	4.20

398	CC(C)OC(=O)NC(C(=O)N[C@@H](C)c1ccc(C)cc1)C(C)C	4.15
399	Clc1ccc(c(Cl)c1)CC(n1ncnc1)C(O)C(C)(C)C	4.14
400	Clc1ncc(cc1)CN1/C(=N/C#N)/SCC1	4.01
401*	O=S1C(=C(OCC1)C)C(=O)Nc1ccccc1	4.00
402	O=C1/C(=C(\NOCC=CCl)/CC)/C(=O)CC(CC(SCC)C)C1	4.16
403	O=P(O)(CC[C@@H](C(=O)O)N)C	3.83
404	O=C(O/N=C(\C(SC)C)/C)NC	3.82
405	O=C1OCCC1N(C(=O)CCl)c1c(cccc1C)C	3.99
406*	Clc1c(ccc(Cl)c1NS(=O)(=O)c1nc2nc(OC)cc(OC)n2n1)C	4.15
407	O=C(Nc1cc(ccc1)C(F)(F)F)N(C)C	3.89
408	CCC1CCCC(C(C(=O)C2=CC3C4CC(CC4C=CC3C2CC(=O)O1)OC1C(C(C(C(O1)C)OC)OC)OC)C)OC1CCC(C(O1)C)NC	4.38
410	N(C(=O)Nc1ccc(cc1)C(C)C)C	3.78
411*	c12c(CC(O1)(C)C)cccc2O	3.71
412	O=C(N(c1c(cccc1C)C)C(C(=O)OC)C)c1occc1	3.97
413	Cc1c(c(ccc1)C)N(C(C)C(=O)OC)C(=O)c1ccc(cc1)C(=O)c1ccccc1	4.11
414	O=S(=O)(Nc1ccc(cc1)C)N(C)C	3.79
415	O=C1/C(=C(/O)\C2CC2)/C(=O)CC(C(=O)OCC)C1	3.86
416	O=S(=O)(C(/C=N/OC(=O)NC)(C)C)C	3.80
417*	C(C(=O)[O-])NCP(=O)(O)O.[H]	3.65
418	O=C1c2ccccc2C(=O)N1	3.59
419	O=S(=O)(O)O.N.N	3.54
420	Nc1c(Cl)c(F)nc(O)c1Cl	3.70
421*	c1(Cc2ccc(N)cc2)ccc(N)cc1	3.71

422	O=BOB(OB(OB=O)O[Na])O[Na]	3.70
423	c1ccc(cc1)n1c(=O)c(c(cn1)N)Cl	3.73
424*	Clc1ccc(NC(=O)N(OC)C)cc1	3.71
425	Brc1ccc(NC(=O)N(OC)C)cc1	3.78
426	c12c(cccc1CC(=O)N)cccc2	3.62
427	CC1=C(CCCO1)C(=O)Nc1ccccc1	3.68
428	O=c1[nH]c(c(Cl)c(=O)n1C(C)(C)C)C	3.67
429	O=c1c2c(oc(O)c1C1c3ccccc3CCC1)cccc2	3.78
430	Clc1cc(c(OCC(=O)O)cc1)C	3.60
431	B(O)(O)O	3.09
432	C(=O)(c1c(cccc1Cl)Cl)NCc1ccc(cc1)C(F)(F)F	3.84
433	O=S1(=O)C(=C(C)S(=O)(=O)CC1)C	3.60
434	Clc1cc(ccc1Oc1cc(C(=O)O)c([N+](=O)[O-])cc1)C(F)(F)F	3.83
435	OC(=O)Cc1cccc2ccccc12	3.52
436	CN	2.74
437	Clc1cc(OC)c(nn1)c1ccccc1	3.57
438	O=c1nc([nH]c(c1CCCC)C)NCC	3.54
439*	O=c1c2c(oc(O)c1C(c1ccccc1)CC(=O)C)cccc2	3.68
440	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)c1ncccc1C(=O)N(C)C	3.80
441*	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)Cc1ccccc1C(=O)OC	3.79
442	[Br-].[Br-].c1cccc2c3[n+](cccc3)CC[n+]12	3.71
443	CCC(=O)O	3.04
444	O=C(O)c1c(F)cccc1F	3.36

445	C(=C\Cl)/C(=O)O	3.19
446	O=C(c1cc(ccc1)C)N(CC)CC	3.43
447	CCCC(CC)(C(=O)NC)CCC	3.41
448*	O=C1c2ccccc2C(=O)c2ccccc12	3.46
449*	CC(C(=O)O)Oc1ccc(cc1)Oc1nc2ccc(cc2o1)Cl	3.66
450	O=C(O)C(Oc1ccc(Oc2nc3ccc(Cl)cc3o2)cc1)C	3.66
451	O=c1c(nnc(SC)n1N)C(C)(C)C	3.46
452	Oc1cccc(N)c1	3.16
453	O=C(OC(C)CCCCC)COc1ccc(Cl)c2cccnc12	3.65
454	c1(c2ccccc2)cc([n+](n1C)C)c1ccccc1	3.52
455*	O=C(O)c1ccccc1C(=O)Nc1cccc2ccccc12	3.58
456*	O=C(O)C1C(C(=O)O)C2OC1CC2	3.38
457*	O=C(O)C(SP(=S)(OC)OC)CC(=O)OCC	3.58
458	COc1cc(nc(n1)Sc1cccc2c1C(=O)OC2C)OC	3.59
460	O=C(OC)c1cccc(Cl)c1NS(=O)(=O)c1nc2cc(F)nc(OCC)n2n1	3.70
461	O=C(O)c1c(Cl)ccc2cc(cnc12)C	3.41
462	c1(S(=O)(=O)Nc2nc3n(c(cc(OC)n3)OC)n2)c(C(F)(F)F)ccnc1OC	3.70
463*	P(=O)(O)(O)O	3.05
464	Cc1cc(Cl)ccc1OCC(=O)O	3.35
465*	COc1nc(nc(OC)c1)NC(=O)NS(=O)(=O)c1c(nc2ccccn12)S(=O)(=O)CC	3.71
466*	O=c1nc[nH]c2ccccc12	3.21
467*	c1ccc(cc1)NC(=O)C(=O)O	3.25
468	O(c1nc(nc(OC)c1)N)C	3.20

469	N1(/C(=N\C#N)/N(CC1)CC)[P@@](=O)(SCCC)OCC	3.49
470	O=C(Nc1c(C)cccc1C)COC	3.29
471	[Cl-].O=C(OCCC)NCCC[NH+](C)C.[H]	3.36
472	C(CN)O.c1(c(ccc(n1)Cl)Cl)C(=O)O	3.40
473	O=C(O)c1c(ncc(c1)COC)C1=NC(C(=O)N1)(C(C)C)C	3.49
474	c1(c(c(c(c(c1Cl)Cl)C(=O)O)Cl)Cl)C(=O)O	3.48
475*	C1CCCC[N+]1(C)C	3.06
476*	COc1cc(nc(n1)NC(=O)NS(=O)(=O)c1cc(ccc1C(=O)O)CNS(=O)(=O)C)OC	3.69
477*	Clc1c(nc(Cl)cc1N)C(=O)O	3.32
478*	O=C1N(C(C)C)S(=O)(=O)Nc2ccccc12	3.38
479	[Ca+2].[O-]C(=O)[O-]	3.00
480	c1(NC(=O)O[C@@H](C(=O)NCC)C)ccccc1	3.37
481	n1c(nc(nc1N)NC1CC1)N	3.22
482*	Clc1ccc(Cl)c(c1OC)C(=O)O	3.34
483	O=C(NCC#N)c1cnccc1C(F)(F)F	3.36
484	Fc1cccc(F)c1NS(=O)(=O)c1nc2c(F)cnc(OC)n2n1	3.56
485	Cc1cc(=O)[nH]o1	3.00
486*	O=C(O)c1c(nc2ccccc2c1)C1=NC(C(=O)N1)(C(C)C)C	3.49
487*	Clc1cc(c(O[C@@H](C(=O)O)C)cc1)C	3.33
488*	O=C(N(c1c(cccc1C)C)[C@@H](C(=O)OC)C)COC	3.45
490	[Ca+2].O=C1/C(=C(\[O-])/CC)/C(=O)CC(C(=O)O)C1.[O-]/C(=C\1/C(=O)CC(C(=O)O)CC1=O)/CC	3.67
491*	FC(F)(F)CCc1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(OC)n1)C	3.62

492*	Cc1nc(nc(n1)OC)NC(=O)NS(=O)(=O)c1ccccc1OCCCl	3.60
494	Oc1n(C)ncc1C(=O)c1ccc(c(c1C)C1=NOCC1)S(=O)(=O)C	3.56
495*	Clc1cc(Cl)ccc1Oc1cc(C(=O)O)c([N+](=O)[O-])cc1	3.52
496	O=c1nc([nH]c(c1)C)C(C)C	3.18
497*	Clc1ccc(Cl)c(C(=O)O)c1O	3.32
498	O=S(=O)(Nc1ccccc1)N(C)C	3.30
499*	Clc1cc(Cl)ccc1Oc1ccc(OC(C(=O)O)C)cc1	3.51
500	O=C(O)c1cnccc1C(F)(F)F	3.28
501*	Clc1cc(cnc1C(=O)O)C(F)(F)F	3.35
502	O=C(O)c1ccccc1C(=O)O	3.22
503	O=C(O)c1ccccc1C(=O)N	3.22
504*	O=C1NCCCN1	3.00
505	n1c(cc(nc1N)C)C	3.09
506*	C1(=O)/C(=C(/O)\C2CC2)/C(=O)CC(C1)C(=O)O	3.35
507	O=C(O)[C@H](Oc1ccc(Oc2nc3ccc(Cl)cc3nc2)cc1)C	3.54
508	Clc1c(cnn(c1=O)C)N	3.20
509	CS(=O)(=O)c1c(COCC(F)(F)F)c(Cl)c(C(=O)C2C(=O)CCCC2=O)cc1	3.64
510	Clc1ccc(cc1)C(NC(=O)[C@@H](NC(=O)OC(C)C)C(C)C)CC(=O)OC	3.60
511	C1C[C@H]2[C@]34CC[C@@H]([C@@](C)([C@H]3[C@@H]([C@@]32CC(=C)[C@@H]1C3)C(=O)O)C(=O)O4)O	3.52
512	C[C@@H](N(C(=O)COC)c1c(C)cccc1C)C(=O)O	3.42
513	C[C@@H](N(C(=O)COC)c1c(C)cccc1C(=O)O)C(=O)O	3.47
514*	c1c([nH]nc1C(=O)O)C(F)(F)F	3.26

515	FC(F)(F)c1cc(ccc1)Oc1cccc(n1)C(=O)O	3.45
516	OC(c1cncnc1)(c1ccc(OC)cc1)C1CC1	3.41
517*	N(C)C(=O)Nc1ccc(Oc2ccc(Cl)cc2)cc1	3.44
518*	[O-][N+](=O)N/C(=N\C)/NCC1CCOC1	3.31
519*	C(=O)(c1c(cc(cc1)C)C1=N[C@@](C(=O)N1)(C(C)C)C)O	3.44
520*	C(CCCCCCCCCCC)C1=C(C(=O)c2c(cccc2)C1=O)OC(=O)C.CC	3.62
521	O=S(=O)(c1ccccc1OCCOC)NC(=O)Nc1nc(OC)nc(OC)n1	3.62
522	Fc1cccc(F)c1C(=O)NC(=O)Nc1ccc(cc1)CO/N=C(\C1CC1)/c1ccc(Cl)cc1	3.68
523	COC(=O)c1cc(C)ccc1C1=NC(C(=O)N1)(C)C(C)C.COC(=O)c1ccc(C)cc1C1=NC(C(=O)N1)(C)C(C)C	3.76
524	O=S(=O)(Nc1cc(c(cc1C)C)NC(=O)C)C(F)(F)F	3.49
526	COCCO[Si](CCCl)(OCCOC)OCCOC	3.50
527	Clc1ccc(cc1)c1cc(c(s1)c1ccccc1)c1ccccc1	3.54
528	O=C(O)c1c(ncc(c1)C)C1=NC(C(=O)N1)(C(C)C)C	3.44
529	O=c1[nH][nH]c(=O)cc1	3.05
530	O=C(Nc1nc(cc(OC)n1)OC)NS(=O)(=O)c1ncccc1OCC(F)(F)F	3.63
531	s1c(c(S(=O)(=O)NC(=O)n2c(=O)n(c(n2)OC)C)c(C(=O)OC)c1)C	3.57
532*	Clc1ncc(s1)CN/C(=N/C)/N[N+](=O)[O-]	3.38
533	Fc1cc(cc(F)c1)NC(=O)N/N=C(/c1ncccc1C(=O)O)\C	3.50
534	Clc1c(N)cc(nc1C(=O)O)c1ccc(Cl)c(OC)c1F	3.49
535*	Clc1cc(Cl)ccc1O[C@@H](C(=O)O)C	3.33
536	Clc1ccc(cc1)C(C(C)C)C(=O)OC(c1cccc(Oc2ccccc2)c1)C(=O)N	3.60
537	O=C(NP(=O)(OC)SC)C	3.22

538	Clc1cccc(Cl)c1NS(=O)(=O)c1nc2cc(F)nc(OCC)n2n1	3.57
539	O=C(OC)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(OC)n1)C	3.53
540	O=P(SC(CC)C)(OCC)N1C(=O)SCC1	3.40
541	O=S(=O)(N)c1sccc1	3.15
542	Clc1c(OCC(=O)O)nc(Cl)c(Cl)c1	3.34
543	O=C1O[C@@]23C=C[C@H](O)C1([C@H]2[C@H](C(=O)O)[C@@]12[C@H]3CCC(O)(C(=C)C1)C2)C	3.46
544	Cc1cccc(C)c1O	3.01
545	OC(=O)c1sccc1S(=O)(=O)NC(=O)Nc1nc(OC)nc(C)n1	3.49
546	c12c(S(=O)(=O)N(C1=O)C)cccc2	3.22
547	Clc1c(N)cc(nc1C(=O)O)c1ccc(Cl)c(O)c1F	3.42
548	O=c1n(N)c(nn1C(=O)NC(C)(C)C)C(C)C	3.30
549	CC(C)C1(C(=O)NC(=N1)c1c(cc(cn1)COC)C(=O)O)C	3.40
550*	O[P@@H](=O)OCC	2.96
551*	Clc1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(OC)n1)C	3.47
552	OS(=O)(=O)CC(=C(Cl)Cl)Cl	3.27
553	O=C1NCCN1	2.85
554	O=C1NS(=O)(=O)c2ccccc12	3.17
555	c1(c(nc(Cl)c(c1)Cl)O)Cl	3.20
556	Clc1ncc(s1)CN1/C(=N/[N+](=O)[O-])/N(C)COC1	3.37
557*	Cc1nc(N)nc(OC)n1	3.05
558	O=C(OC)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(nc(OCC)n1)NC	3.51
559*	O=S(=O)(c1c(C(=O)OC)c(Cl)nn1C)NC(=O)Nc1nc(OC)cc(OC)n1	3.52
560*	O=C(N(C)C)/C(=N/O)/SC	3.09

561	Cc1ccc(cc1N=C=O)N=C=O	3.12
562	O=c1c(n[nH]c(=O)n1N)C(C)(C)C	3.14
563	ClC(Cl)C(=O)N(CC=C)CC=C	3.17
564*	O=P(O)(O)C(N)CC	2.98
565	O=C(O)c1ccccc1S(=O)(=O)NC(=O)Nc1nc(C)nc(OC)n1	3.39
567	FC(F)(F)c1ccccc1C(=O)O	3.07
568	Clc1cc(ccc1Oc1cc(C(=O)NS(=O)(=O)C)c([N+](=O)[O-])cc1)C(F)(F)F	3.41
569*	O=C1/C(=C(\NOCC)/CCC)/C(=O)CC(CC(SCC)C)C1	3.28
570*	CCC(C)OC(=O)N1CCCCC1CCO	3.12
571*	O=C(O)c1ccc(cc1Cl)S(=O)(=O)C	3.12
572	O=c1n(c(nnc1c1ccccc1)C)N	3.03
573*	O=S(=O)(c1c(C(=O)OCC)cnn1C)NC(=O)Nc1nc(OC)cc(OC)n1	3.36
574	O(C(=O)/C(=C\c1c(ccc(N2C(=O)C3=C(C2=O)CCCC3)c1)Cl)/Cl)CC	3.29
575	c1(cc(ccc1)C(=O)O)C(F)(F)F	2.97
576	O=C1/C(=C(/NOCC)\CCC)/C(=O)CC(C2CCCSC2)C1	3.17
577	O=C(c1ccc(cc1Cl)S(=O)(=O)C)C1C(=O)CCCC1=O	3.16
578*	O=C(N)c1c(Cl)cccc1Cl	2.90
579*	O=C(N)P(=O)(O)OCC	2.71
580	P(=O)(C(=O)N)(OCC)[O-].[NH4+]	2.75
582	O=S(=O)(c1ncccc1S(=O)(=O)CC)N	2.90
583	O=c1nc(n(c(=O)n1C1CCCCC1)C)N(C)C	2.90
584*	C(NCC(=O)N(S(=O)(=O)C)C)P(=O)(O)O	2.91
585	O=C(O)c1c(ncc(c1)CC)C1=NC(C(=O)N1)(C(C)C)C	2.93

586	O=C(OCC#C)/C=C(/C=C/CC(C)CCCC(C)C)\C	2.90
587	[Na+].O=C(Nc1nc(cc(OC)n1)OC)[N-]S(=O)(=O)c1nc(ccc1C(=O)OC)C(F)(F)F	3.02
588	S=C1NCCN1	2.32
589	[nH]1cnnc1	2.14
590*	O=P(O)(O)CN	2.33
591	N(N=O)(C)C	2.12
592	OC(=O)c1ccccc1S(=O)(=O)NC(=O)N(C)c1nc(OC)nc(C)n1	2.83
593	Cc1ccc(cc1)n1c2cccc(c2c(=O)c(n1)C(=O)O)OCCOC	2.79
594*	CC(C)CC(=O)C	2.22
595	OC(=O)c1cccc(C)c1S(=O)(=O)NC(=O)Nc1nc(N(C)C)nc(OCC(F)(F)F)n1	2.82
596	O=C(OC)c1cccc(c1S(=O)(=O)NC(=O)Nc1nc(nc(OCC(F)(F)F)n1)N(C)C)C	2.83
597	Cc1nc(nc(n1)OC)N(C)C(=O)NS(=O)(=O)c1ccccc1C(=O)OC	2.73
598	O=C(O)C(Oc1ccc(Oc2ncc(cc2Cl)C(F)(F)F)cc1)C	2.66
599	Clc1ccc(n2ncc(=O)c(c2CC)C(=O)O)cc1	2.45
600*	[nH]1cnnc1N	1.92
601	CCC1OCC(COc2ccc(Oc3ccccc3)cc2)O1	2.48
602	C(=C(C)C)CNc1ncnc2c1[nH]cn2	2.31
603	c12c(NCc3ccco3)ncnc1[nH]cn2	2.33
604*	C/C(=C\CNc1c2c([nH]cn2)ncn1)/CO	2.34
605*	O=c1ccc(=O)[nH][nH]1	2.05
606	N(c1cc(c(F)cc1)Cl)(C(=O)c1ccccc1)[C@@H](C(=O)OC)C	2.22

607	CC(=O)C	0.86
608	C(C)O	0.54
609	OCC(O)CO	0.23
610	Cl/C(Cl)=C/[C@H]3[C@@H](C(=O)O[C@H](C#N)c2cccc(Oc1ccccc1)c2)C3(C)C	9.03
611	Cl/C(Cl)=C/C3C(C(=O)OC(C#N)c2cccc(Oc1ccccc1)c2)C3(C)C	8.78
612	Cl/C(Cl)=C/C2C(C(=O)OCc1c(F)c(F)cc(F)c1F)C2(C)C	8.72
613	Cl\C1=C(/Cl)C2(Cl)C3C4CC(C3C1(Cl)C2(Cl)Cl)C5OC45	8.72
614	C1([C@@H]([C@@H]1\C=C(/C)C)C(O[C@@H](\C(=C\CC)C)C#C)=O)(C)C	8.21
615	P(Oc1c(cc(Cl)c(n1)Cl)Cl)(OC)(OC)=O	8.11
616	n3c1c(cccc1)c(OCCc2ccc(cc2)C(C)(C)C)nc3	7.91
617*	O=C(OC)N(OC)c1ccccc1COc3nn(c2ccc(Cl)cc2)cc3	7.81
618	ClC1(C(Cl)2Cl)C(C=CC3Cl)C3C2(Cl)C(Cl)=C1Cl	7.73
619	O=C1SC(=N/N1CSP(=S)(OC)OC)\OC	7.48
620	[O-][N+](=O)c1cc(cc([N+]([O-])=O)c1N(CCCl)CCC)C(F)(F)F	7.47
621	Clc3ccccc3c1nnc(nn1)c2c(Cl)cccc2	7.31
622	O=C1\C3=C(/C(=O)N1COC(=O)C2C(\C=C(/C)C)C2(C)C)CCCC3	7.32
623	S=P(OC)(OC)SCN1\N=N/c2ccccc2C1=O	7.20
625	[Sn](c1ccccc1)(c1ccccc1)(c1ccccc1)Cl	7.02
627	Brc1cc(C#N)cc(Br)c1OC(=O)CCCCCCC	6.99
630*	ClC(Cl)(Cl)C(c1ccc(OC)cc1)c2ccc(OC)cc2	6.82
631	O=C\1c6c(OC(/O)=C/1C5c2ccccc2CC(c4ccc(c3ccccc3)cc4)C5)cccc6	6.84
632	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)Cl	6.66

633*	CCCSP(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl	6.67
634	Cl/C2=C(\Cl)C3(Cl)C1CC(Cl)C(Cl)C1C2(Cl)C3(Cl)Cl	6.66
635	FC(F)Oc1c(Cl)c(nn1C)c2cc(OCC(=O)OCC)c(Cl)cc2F	6.62
637	[O-][N+](=O)c1c(c(cc([N+]([O-])=O)c1NC(CC)CC)C)C	6.31
638	O=C(n1c2ccccc2nc1NC(=O)OC)NCCCC	6.23
639	N1(C([C@@H]2CC=CC[C@@H]2C1=O)=O)SC(Cl)(Cl)Cl	6.21
641	COP(=S)(OC)SCN1C(=O)C2=CC=CC=C2C1=O	6.14
643	Clc1cc(c(OCC(=O)SCC)cc1)C	5.91
644	FC(F)Oc1ccc(cc1)C(C(=O)OC(C#N)c3cc(Oc2ccccc2)ccc3)C(C)C	6.15
645*	c1(C(Nc2c(cc([N+](=O)[O-])cc2)Cl)=O)c(ccc(c1)Cl)O	5.98
647	Clc1cc(C(OP(=O)(OC)OC)=[C@H]Cl)c(Cl)cc1Cl	5.93
648	ClC(Cl)C(=O)NC1=C(Cl)C(=O)c2ccccc2C1=O	5.85
649	S=P(OCC)(OCC)Oc1nn2cc(c(nc2c1)C)C(=O)OCC	5.89
650	Clc1cc(Cl)ccc1OC(C(=O)O)C	5.67
651	O=P(O\C=C(/Cl)Cl)(OC)OC	5.60
652*	O=[N+]([O-])c1cc(cc(C(=O)OC(C)C)c1)C(=O)OC(C)C	5.72
653*	FC(F)(F)c2ccccc2C(=O)C(C#N)(c1ccc(cc1)C(C)(C)C)C(=O)OCCOC	5.85
654	Clc2c(c(ccc2Oc1ccccc1)[N+]([O-])=O)N	5.60
656	N(=C/N(\C=N\c1ccc(cc1C)C)C)\c2ccc(cc2C)C	5.60
657	O=C(OC(C)C)\C=C(\C=C\CC(CCCC(OC)(C)C)C)C	5.61
658	[O-][N+](=O)c1cc(cc([N+]([O-])=O)c1N(CCC)CCC)C(C)C	5.59
659*	O1CCOC1CO\N=C(\C(F)(F)F)c1ccc(Cl)cc1	5.56
660	n2c1c(cccc1)nc2c3occc3	5.31

661	C1(=CC1)C	4.75
662	ClC=CCO	4.97
663*	FC(F)(F)c3ccc(F)c(F)c3C(\NC(=O)Cc1ccccc1)=N\OCC2CC2	5.60
664	O=C(OCCCCC)COc1cc(c(F)cc1Cl)N2C(=O)\C3=C(/C2=O)CCCC3	5.59
665	c12[C@@](c3ccccc3c1ccc(c2)Cl)(C(O)=O)O	5.35
666	Fc1c(F)c(c(F)c(F)c1COC(=O)C2C(/C=C/C)C2(C)C)COC	5.48
667	c1(c(c(CC)ccc1)N(COC(C)C)C(CCl)=O)C	5.34
668	O=C(OCCCC)[C@H](Oc2ccc(Oc1ncc(cc1)C(F)(F)F)cc2)C	5.43
669	Clc1ccc(cc1)CC2C(O)(C(CC2)C(C)C)Cn3ncnc3	5.35
670*	O=C(Oc1cccc2OC(Oc12)(C)C)NC	5.16
671	P(OCC)(=O)(OCC)SCCSCC.P(OCC)(OCC)(OCCSCC)=S	5.51
672*	ClCC(=O)N(c1c(cccc1CC)CC)COC	5.18
673	FC(F)(F)c2cc(OC(C(=O)NCc1ccccc1)CC)ccc2F	5.28
674*	c12c(c(ccc1)O)nccc2	4.86
675	Clc1nc(nc(n1)NC(C)(C)C)NCC	5.02
677	ClC(Cl)(Cl)c1nc(OCC)sn1	5.01
678	Clc2c(Cl)cc(N1OC(=O)N(C1=O)C)cc2	5.02
679	CNC(=O)Oc1ccccc1C2OCCO2	4.92
680	Fc3cccc(F)c3C/2=N/C(c1c(OCC)cc(cc1)C(C)(C)C)CO\2	5.11
681	[O-][N+](=O)c1cc(cc([N+]([O-])=O)c1N(CCC)CCC)S(=O)(=O)N	5.08
682	FC(F)(F)c1cccc(c1)N2C(=O)C(Cl)C(C2)CCl	5.02
683	Fc1ccc(cc1)C3(OC3c2ccccc2Cl)Cn4ncnc4	5.02
684	Clc1ccc(c(Cl)c1)C(O)(CCCC)Cn2ncnc2	4.97

685	CC1(C)CC\C(=C/C2=CC=C(Cl)C=C2)C1(O)CN1C=NC=N1	4.95
686	c1(c2c(cccc2)O)ccccc1	4.63
687	O=C(c1cc(cc(c1)C)C)N(NC(=O)c2cccc(OC)c2C)C(C)(C)C	4.94
688	c1(cc(ccc1)\N=C\N(C)C)OC(NC)=O.Cl	4.78
689	C[C@@H]1CCC[C@H](/C=C/C(=O)O[C@]23[C@@H](/C=C/C1)[C@@H](C(=C)[C@H]([C@H]2[C@@H](NC3=O)Cc4ccccc4)C)O)O	5.03
690*	O=C2c1ccccc1C(=O)C2C(=O)CC(C)C	4.66
691*	FC(F)(F)c1c(N)ccc(Cl)c1	4.57
692	O=C(SC(c1ccccc1)(C)C)N2CCCCC2	4.66
693	Cc1csc(C)c1N([C@@H](C)COC)C(=O)CCl	4.64
694*	O=P(OC)(SCCSCC)OC	4.56
695	O(C2CC1(OC2(C)CC1)C(C)C)Cc3ccccc3C	4.62
696	FC(F)(F)c3cc(C=1C(=O)C(OC=1NC)c2ccccc2)ccc3	4.68
697	O1CCOC1CO\N=C(/C#N)c1ccccc1	4.51
698*	Clc1c(c(Cl)ccc1Cl)CC(=O)O	4.50
699	O=C(NC1CCCCC1C)Nc2ccccc2	4.46
700	O=C(N(c1c(cccc1C)C)Cn2nccc2)CCl	4.51
701	Clc2cc1c(OC(=O)N1)cc2	4.26
702	Clc1cc(C(=O)O)c(Cl)c(N)c1	4.31
703	n1(c(cc(n1)C)OC(N(C)C)=O)C(C)C	4.30
704	c1ccccc1C=O	3.98
705	CC1=CC(=C(C=C1)N=C2N(C=CS2)C)C	4.26
706	O=C3N2c1c(cccc1CC2)CC3	4.12
707	Clc2ccc(/C=C(/n1ncnc1)[C@@H](O)C(C)(C)C)cc2	4.29

708	O=C(NC)N(c1nc2ccccc2s1)C	4.14
709*	FC(F)(F)c1cc(ccc1)Oc2ncccc2C(O)=O	4.21
710	c1(nc(Cl)nc(NC(C)(C)C)n1)N	4.05
711*	Clc1ccc(cc1)C(O)(C(C)C2CC2)Cn3ncnc3	4.19
712	Fc1ccc(N)cc1	3.75
713	Clc2ccc(OC(n1ncnc1)C(O)C(C)(C)C)cc2	4.14
714	Clc2cc(ccc2Oc1cc(C(=O)OCC(=O)O)c([N+]([O-])=O)cc1)C(F)(F)	4.24
715	O=P(OC)(SC)N	3.75
717	O=C(NC)CSP(=S)(OC)OC	3.88
718	Clc3c(OCC#C)cc(N1/N=C2\N(C1=O)CCCC2)c(Cl)c3	4.01
719	O=C(N(C=O)C)CSP(=S)(OC)OC	3.83
720	O=C(OCC)Nc1cccc(O)c1	3.63
721	C(C)(O)=O	3.13
722*	C[C@H](C(=O)O)Oc1ccc(cc1)Oc2c(cc(cn2)C(F)(F)F)Cl	3.86
723*	O[As](=O)(O)O	3.43
724	O=C(OC)c1sccc1S(=O)(=O)NC(=O)Nc2nc(nc(OC)n2)C	3.84
725	Clc1ccc(NC(=O)N)cc1	3.37
726	O1C(OC(OC(OC1C)C)C)C	3.37
727	[O-][N+](=O)NC/1=N/CCN\1Cc2cnc(Cl)cc2	3.49
728	O=C(N(c1nnc(s1)C(C)(C)C)C)NC	3.42
729	C([C@@H](Oc1ccccc1)CC)(=O)O	3.29
730	O=C(N(C)C)C(c1ccccc1)c2ccccc2	3.39
731*	Clc1c(nc(Cl)cc1)C(=O)O	3.28

732	O=C(Nc1nc(cc(OC)n1)OC)NS(=O)(=O)c2cc(NC=O)ccc2C(=O)N(C)C	3.66
733	O=C(Nc1nc(cc(OC)n1)OC)NS(=O)(=O)c2cc(CNS(C)(=O)=O)ccc2C(=O)OC	3.70
734*	O=C2N(/N=C/c1cccnc1)C\C(=N/N2)C	3.34
735*	OC(=O)[C@@H](C)Oc1ccc(cc1)Oc2ccc(cc2F)C#N	3.48
736*	O=P(OCC)(Oc1cc(c(cc1)S(=O)C)C)NC(C)C	3.50
737*	O=C/1NC(=O)N\N=C\1C(C)(C)C	3.23
738	NC=1/C=N\NC(=O)C=1Cl	3.16
739*	C1(CC(C(C(C1)=O)C(CC)=O)=O)C(=O)O	3.33
740	c1(ccc(C#N)cc1)C(O)=O	3.17
742	O=C(O)c1c(Cl)ccc2cc(Cl)cnc12	3.38
743*	FC(F)(F)c1ncc(cc1)C(C)S(=O)(=NC#N)C	3.44
744	Cc1cc(cc(C)c1)C(=O)N(NC(=O)c3ccc2OCCCc2c3C)C(C)(C)C	3.57
745	C1(\NCCCS1)=C/[N+](=O)[O-]	3.16
746	n1c(nc(nc1N)N)N	3.02
747*	C(c1c(cc(cc1)S(=O)(=O)C)[N+](=O)[O-])(=O)O	3.31
748	O=S(=O)(c1ccccc1Cl)N	3.19
749	OC(=O)C(/C#N)=N\OC	2.99
750	O=C(Oc1cccc(c1)NC(=O)N(C)C)NC(C)(C)C	3.32
751	CO/C=C(\c1ccccc1Oc2cc(ncn2)Oc3ccccc3C#N)/C(=O)O	3.41
752	CC(C(=NOC(=O)NC)C)S(=O)(=O)C	3.12
753	O=C(Nc1ccccc1)N(C)C	2.91
754	Fc1cccc(F)c1NS(=O)(=O)c2nc3nc(ccn3n2)C	3.04

755	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)N(C)S(=O)(=O)C	3.06
756	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)c2ncccc2S(=O)(=O)CC	3.04
757	O[C@@H]1CC(\C(=C)CC1)=C\C=C2/CCC[C@]3([C@H]2CC[C@@H]3[C@H](C)CCCC(C)C)C	2.59
758	C(CC)CO	1.79
759	OCCO	0.53

Table 3.3. Reported aquatic toxicity [Log(1/LC50)] data of curated 136 datapoints of LP (Lepomis) dataset. (* Test set compounds)

ID. No.	Canonical smiles	Log(1/LC50)
1	Cl\C(=C/C3C(C(=O)OC(C#N)c2cccc(Oc1ccccc1)c2)C3(C)C)C(F)(F)F	8.99
2	c1([C@@H](OC([C@@H](Nc2c(cc(C(F)(F)F)cc2)Cl)C(C)C)=O)C#N)cc(Oc2ccccc2)ccc1	8.75
3	ClC2(Cl)CC1C(C2(C(Cl)C1Cl)CCl)(C(Cl)Cl)CCl	7.97
4	Cl\C3=C(/Cl)C4(Cl)C2C(C1/C=C\C2C1)C3(Cl)C4(Cl)Cl	7.90
5	Clc1cc(c(F)c(Cl)c1F)NC(=O)NC(=O)c2c(F)cccc2F	7.77
6	ClC1[C@@]2(Cl)C(C)(C)C(=C)[C@@]1(Cl)C(Cl)(Cl)C2(Cl)Cl	7.65
7*	C1([C@@]2([C@@H]3[C@@H]([C@@H]4C=C[C@H]3C4)[C@]1(Cl)C(=C2Cl)Cl)Cl)(Cl)Cl	7.48
8	c12c(CC(O2)(C)C)cccc1OC(N(SN(CCCC)CCCC)C)=O	7.40
9	O=C(OCc1cc(oc1)Cc2ccccc2)C3C(\C=C(/C)C)C3(C)C	7.30
10	Clc1cccc(F)c1CN(c2c([N+]([O-])=O)cc(cc2[N+]([O-])=O)C(F)(F)F)CC	7.26
11	CCCN(CC1CC1)c2c(cc(cc2[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]	7.18

12	O=C2\C(=C(\C)C(OC(=O)C1C(/C=C(/C)C)C1(C)C)C2)C\C=C	7.10
13	Brc1cc(C#N)cc(Br)c1OC(=O)CCCCCC	7.13
14	O=C1Sc2nc3ccc(cc3nc2S1)C	6.85
15*	S=P(OCC)(SCCSCC)OCC	6.85
16	[O-][N+](=O)c1c(c([N+]([O-])=O)c(Cl)c(c1)C(F)(F)F)Nc2ncc(cc2Cl)C(F)(F)F	6.93
17	O=C(Oc1c(cc(cc1C(C)CCCCCC)[N+]([O-])=O)[N+]([O-])=O)/C=C/C	6.81
18	S=P(Oc1ccc(cc1)S(=O)C)(OCC)OCC	6.63
19*	Clc2ccc1nc(n[n+]([O-])c1c2)n3ccnc3	6.51
20	C[C@H](C(=O)OC)Oc1ccc(cc1)Oc2c(cc(cn2)C(F)(F)F)Cl	6.63
21*	c1(ncc(O[P@@](=S)(OC(C)C)OCC)cn1)C(C)(C)C	6.55
22	Clc2cc(ccc2Oc1cc(C(=O)OC(CC)C(=O)OCC)c(cc1)[N+]([O-])=O)C(F)(F)F	6.68
23	CCNc1nc(CC)c(s1)C(=O)NC(C#N)c2cccs2	6.50
24	O=C(Oc1cc(c(cc1)N(C)C)C)NC	6.28
25	O=C(OCC)[C@H](Oc3ccc(Oc1nc2ccc(Cl)cc2o1)cc3)C	6.28
26*	O=C(OCC#C)[C@H](Oc2ccc(Oc1ncc(Cl)cc1F)cc2)C	6.22
27*	O=C(OCC)[C@H](Oc3ccc(Oc1nc2ccc(Cl)cc2nc1)cc3)C	6.25
28	N#C/C(=N\OP(=S)(OCC)OCC)c1ccccc1	6.13
29	[O-][N+](=O)c1cc([N+]([O-])=O)cc(c1O)C	5.94
30	c1c2cccc([O-])c2ncc1.[Cu+2].c1c2c(ncc1)c(ccc2)[O-]	6.17
31	O=C(Oc1cccc(c1)NC(=O)OCC)Nc2ccccc2	6.08
32	O(c1ncccc1)C(COc3ccc(Oc2ccccc2)cc3)C	6.08
33	O=P(SCCC)(OCC)SCCC	5.88

34*	CC(=CC1C(C(=O)OCc2ccc(C)cc2C)C1(C)C)	5.93
35	N(=C(\N)N)\CCCCCCCCCNCCCCCCCCC/N=C(\N)N	5.93
36	CCCCOCN(C1=C(C=CC=C1CC)CC)C(=O)CCl	5.85
37*	Clc1ccc(cc1)C(O)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	5.86
38	O=C(OCCCC)C(Oc2ccc(Oc1ncc(cc1)C(F)(F)F)cc2)C	5.86
39	O=C2N(c1cc(Cl)cc(Cl)c1)C(=O)CN2C(=O)NC(C)C	5.78
40*	O=C(OC(C)C)NNc1cc(ccc1OC)c2ccccc2	5.71
41	C1=CC(=CC=C1OS(=O)(=O)C2=CC=C(C=C2)Cl)Cl	5.64
42	c1(OP(OC)(OC)=S)cc(c([N+](=O)[O-])cc1)Cl	5.62
43	CCCCSP(=O)(SCCCC)SCCCC	5.64
44	Clc1ccc(cc1Cl)C(=O)OCCCN2C(CCCC2)C	5.63
45	CCCCC1=C(N=C(N=C1OS(=O)(=O)N(C)C)NCC)C	5.50
46	[N-]=[N+]=N/c1nc(nc(SC)n1)NC(C)C	5.35
47	CC(C)Nc1nc(Cl)nc(NC2CC2)n1	5.25
48*	CCOP(=S)(OCC)Oc1cc(Cl)c(SC)cc1Cl	5.44
49	O=C(ON=C(SC)C)N(SN(C(=O)ON=C(SC)C)C)C	5.40
50	c1(c(cc(O)cc1C)C)Cl	4.99
51	c1(c(O[P@@](NC(C)C)(OCC)=S)cccc1)C(OC(C)C)=O	5.28
52	c1cc(ccc1CC(CN2CCCCC2)C)C(C)(C)C	5.16
53*	O=C(N(c1ccc(F)cc1)C(C)C)COc2nnc(s2)C(F)(F)F	5.23
54	O2[C@H](CN(CC(C)Cc1ccc(cc1)C(C)(C)C)C[C@H]2C)C	5.12
55	c1(c(ccc(c1)Cl)O)C	4.79
56	Clc1cc(c(OCC(=O)O)cc1)CO	4.97

57	O=C(OCC)C(Oc3ccc(Oc1nc2ccc(Cl)cc2nc1)cc3)C	5.12
58*	Clc2cccc1SC(=O)N(c12)CC(=O)OCC	4.99
59	O=C(c1cc(cc(c1)C)C)N(NC(=O)c2ccc(cc2)CC)C(C)(C)C	5.07
60	Clc1ccc(cc1)C3SC(=O)N(C(=O)NC2CCCCC2)[C@H]3C	5.04
61*	FC(F)C(F)(F)OCC(c1ccc(Cl)cc1Cl)Cn2ncnc2	4.94
62	CC(C)Oc1cccc(c1)NC(=O)c1ccccc1C(F)(F)F	4.78
63*	O=C1/C(=C(\NOCC)/CC)/C(=O)CC(c2c(cc(cc2C)C)C)C1	4.73
64	ClC(Cl)C(=O)N1c2c(OCC1C)cccc2	4.60
65	Clc1ccc(C(=NOC)Cc2cccnc2)c(Cl)c1	4.65
66	O1CC(OC21CCC(CC2)C(C)(C)C)CN(CC)CCC	4.62
67*	O=C(SCCC)N(CCCC)CC	4.44
68	FC(F)(F)c1nc(Cl)sc1C(=O)OCc1ccccc1	4.58
69	OCc1ccccc1	4.03
70	c1(ccncc1)N	3.96
71*	Cc1c(sc(c1C(=O)NCC=C)[Si](C)(C)C)C	4.39
72*	Fc1ccc(cc1)C(O)(c1ccccc1Cl)c1cncnc1	4.42
73	O=C(SCC)N(CCC)CCC	4.13
74*	Fc1nc(OCC(=O)O)c(Cl)c(c1Cl)N	4.25
75	C(CCl)(Cl)(Cl)Cl	4.02
76*	C(CBr)Br	4.02
77*	O=C(N(C(C)C)Cc1ccccc1)C(C)(C)C	4.09
78	Cc1c(C)cccc1	3.73
79	C(CC=O)CC=O	3.65

80	Clc1cccc2sc(=O)n(c12)CC(=O)O	3.96
81	N#C/C(=N\OC)/C(=O)NC(=O)NCC	3.83
82	Clc1cc(c(Cl)cc1OC(F)(F)C(F)C(F)(F)F)NC(=O)NC(=O)c1c(F)cccc1F	4.25
83	Oc1c(Br)cc(cc1Br)C#N	3.98
84	c1ccc(c(c1)C(Cn1cncn1)(c1ccc(cc1)F)O)F	3.96
85	c1ccccc1NC(=O)c1c(C)nc(C)s1	3.83
86*	CCC(C)n1c(=O)c(c([nH]c1=O)C)Br	3.86
87*	N(CCN(CC(=O)O)CC(=O)O)(CC(=O)O)CC(=O)O	3.85
88*	NC(C)CC	3.20
89	n1cc(Cl)cc(F)c1Oc1ccc(O[C@H](C)C(=O)O)cc1	3.61
90*	CC/C(=N\OC/C=C/Cl)/C1=C(CC(CC1=O)C1CCOCC1)O	3.62
91	C(=O)(/C=C(/C=C/C[C@@H](CCCC(C)(C)OC)C)\C)SCC	3.56
92	c1(S(=O)(=O)NC(=O)OC)ccc(N)cc1	3.40
93*	c1(C)c(c(O)n(C)n1)C(=O)c1c(cc(cc1)C(F)(F)F)S(=O)(=O)C	3.59
94*	O=S(=O)(Nc1c(Cl)cc(Cl)c(n2nc(n(c2=O)C(F)F)C)c1)C	3.62
95	Cn1nc(c2cc(OCC(=O)O)c(Cl)cc2F)c(Cl)c1OC(F)F	3.59
96*	Clc1ccc(NC(=O)N(C)C)cc1	3.30
97*	[As](=O)(C)(C)O	3.14
98	Clc1cc(cc(Cl)c1OC(F)(F)C(F)F)NC(=O)NC(=O)c1c(F)cccc1F	3.66
99*	C(C(=O)O)(C)(Cl)Cl	3.13
100	O=C(OC1COC1)c1c(cccc1)S(=O)(=O)NC(=O)Nc1nc(cc(n1)C)C	3.56
101	CS(=O)(=O)c1cc(c(cc1)C(=O)C1C(=O)CCCC1=O)[N+](=O)[O-]	3.45
102	C1CC1c1nc(c(c(n1)N)Cl)C(=O)O	3.25

103	Clc1ccc(OCC(=O)O)cc1	3.02
104	c1ccc2cc(ccc2c1)OCC(CO)(CO)CCl	3.19
105	ClCC(Cl)C	2.61
106*	Clc1ccc(cc1)C(c1c(=O)c2c(oc1O)cccc2)CC(=O)C	3.00
107	C(OCCO)CCC	2.35
108	Clc1cc(c(Cl)cc1Cl)S(=O)(=O)c1ccc(Cl)cc1	2.61
109	O=c1oc(nn1c1c(Cl)cc(NC(=O)N(C)C)cc1)C(C)(C)C	2.53
110*	C1COCCO1	1.40
111	C(CO)OC	0.88
112	C1([C@@H]([C@@H]1\C=C(\C(F)(F)F)Cl)C(=O)O[C@@H](c1cc(ccc1)Oc1ccccc1)C#N)(C)C	10.11
113	S=P(OCC)(SCSC(C)(C)C)OCC	7.86
114*	[Sn](O[Sn](CCCC)(CCCC)CCCC)(CCCC)(CCCC)CCCC	7.89
115	O=C(c1ccccc1Cl)NC(=O)Nc2ccc(OC(F)(F)F)cc2	7.23
116	S=P(Sc1ccccc1)(OCC)CC	6.94
117	O=C(Oc1c(cc(cc1C(C)CCCCCC)[N+]([O-])=O)[N+]([O-])=O)/C=C/C	6.81
118	Clc1c(c(Cl)c(Cl)c(Cl)c1Cl)[N+]([O-])=O	6.47
119	CC2(C)CC1=C(C(=CC=C1)OC(NC)=O)O2	6.09
120	O=C(OCC1OCCC1)C(Oc4ccc(Oc2nc3c(nc2)cc(Cl)cc3)cc4)C	6.27
121	S=C1SCN(C)CN1C	5.73
122	Clc2ccc(N(C(=O)c1ccccc1)C(C(=O)OCC)C)cc2Cl	5.87
123*	O=C2N(c1cc(Cl)cc(Cl)c1)C(=O)CN2C(=O)NC(C)C	5.78
124	C/C(=N\OC(=O)NC)SC	5.41
125	O=C(OCCCC)[C@H](Oc2ccc(Oc1ccc(C#N)cc1F)cc2)C	5.48

126	FC(F)(F)c2nc(sc2C(=O)Nc1c(Br)cc(OC(F)(F)F)cc1Br)C	5.61
127	c12c([nH]c(n2)NC(OC)=O)ccc(c1)[S@@](c1ccccc1)=O	5.07
128	O=C(Oc1ccccc1OC(C)C)NC	4.53
129	Clc2ccc(OS(=O)(=O)c1ccccc1)cc2	4.50
130	S=P(Oc1ccc(SC)cc1)(OCC)SCCC	4.47
131	Clc1cc(Cl)c(cc1)C2(OCCO2)Cn3ncnc3	4.22
132	O=C(OC(C)C)Nc1ccccc1	3.75
133*	N#CN	2.99
134	Clc1nc(nc(n1)NCC)NCC	3.35
135	c1(c(S(=O)(=O)NC(Nc2nc(nc(OC)n2)C)=O)cc(cc1)I)C(=O)O	3.69
136	O=C(NN(C)C)CCC(=O)O	3.03

Table 3.4. Reported aquatic toxicity [Log(1/LC50)] data of curated 226 datapoints of Miscellaneous dataset. (* Test set compounds)

ID. No.	Canonical smiles	Log(1/LC50)
1*	Cl/C(Cl)=C/[C@H]3[C@@H](C(=O)O[C@@H](C#N)c2ccc(F)c(Oc1ccccc1)c2)C3(C)C	9.81
2	Clc1ccc(cc1)[C@H](C(=O)O[C@@H](C#N)c3cccc(Oc2ccccc2)c3)C(C)C	9.62
3	Cl\C1=C(/Cl)C2(Cl)C(Cl)(Cl)C1(Cl)C4C2C3OC3(Cl)C4Cl	8.33
4	Cl/C1=C(\Cl)C2(Cl)C(Cl)(Cl)C1(Cl)C3C2C(Cl)OC3Cl	8.31
5	C1([C@@H](\C=C\C)[C@@H]1C(OCc1c(c(c(C)c(F)c1F)F)F)=O)(C)C	8.06
6*	BrC(F)(F)Oc1ccc(cc1)C(C)(C)COCc3cc(Oc2ccccc2)ccc3	8.13
7	Cl\C1=C(/Cl)C3(Cl)C(Cl)(Cl)C1(Cl)C2C3COS(=O)(=O)OC2	7.63

8	COP(=S)(OC)Oc1ccc(S(=O)C)c(C)c1	7.47
9*	ClC=1C(=O)C(\Cl)=C(\Cl)C(=O)C=1Cl	7.39
10	CCOP(=S)(OCC)Oc1cc(C)nc(n1)N(CC)CC	7.22
11*	OC1=C(C=C(C(=C1CC2=C(C(=CC(=C2Cl)Cl)Cl)O)Cl)Cl)Cl	7.29
12	O=P(OCC)(SCCC)Oc1cn(nc1)c2ccc(Cl)cc2	7.11
13	c1(c(c(cc(c1)[N+](=O)[O-])[N+](=O)[O-])O)[C@@H](CCC)C	6.93
14*	CCC(C)c1cc(cc([N+](=O)[O-])c1O)[N+](=O)[O-]	6.74
15	C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)N	6.68
16	O[Sn](C1CCCCC1)(C2CCCCC2)C3CCCCC3	6.81
17	S=P(Oc1nccnc1)(OCC)OCC	6.44
18*	Clc1cc(OP(=S)(OC)OC)c(Cl)cc1Br	6.66
19*	CCOP(=S)(OCC)OC1=CC(=C(C=C1Cl)Br)Cl	6.34
20*	O=C(OCCO\N=C(/C)C)C(Oc3ccc(Oc1nc2ccc(Cl)cc2nc1)cc3)C	6.37
21*	O=C(OC)\C(=C\OC)c1ccccc1COc2nc(OC(C)C)nc(c2)C(F)(F)F	6.34
22	Clc1ccc(NC(=O)N(CCCC)C)cc1Cl	5.93
23*	COC(=S)SSC(=S)OC	5.80
24	C(=O)(CCl)N(c1c(cccc1CC)CC)COC/C=C\C	5.81
25	Clc1ccc(cc1Cl)C(OC(=O)C)C(Cl)(Cl)Cl	5.83
26	N(=N/c1ccccc1)\c2ccccc2	5.56
29	Clc1cc(NC(=O)OCC#CCCl)ccc1	5.69
30	CC1=NOC(=O)C1=NNc2ccccc2Cl	5.64
31	O=S(=O)(n1c(c(Cl)nc1C#N)c2ccc(cc2)C)N(C)C	5.76
32*	CCCCOC(=O)c1ccccc1C(=O)OCCCC	5.67

33	C(CCCCCC\C=C/CCC)O	5.43
34*	C(CCCCCCC)CCCCCCO	5.50
35	C=C(\[C@H]1C/C=C(/C)CC1)C	5.29
36	Cl[C@H]1[C@@H]([C@@H]([C@@H](Cl)[C@@H]([C@H]1Cl)Cl)Cl)Cl	5.55
37	O=C(Nc1onc(c1)C(CC)(C)CC)c2c(OC)cccc2OC	5.58
38	S(=O)(=O)(F)F	5.06
40	c1(ccccc1C(F)(F)F)C(NCCc1c(cc(cn1)C(F)(F)F)Cl)=O	5.61
41	CCN(CC)C(=O)SCC1=CC=C(C=C1)Cl	5.42
42*	O=C(OCC)CCN(SN(C(=O)O\N=C(/SC)C)C)Cc1ccccc1	5.61
43*	c1(snc(c1Cl)Cl)C(=O)Nc1c(cccc1)C#N	5.47
44	C(CC\C=C\CCCC)COC(C)=O	5.26
45*	C=1(C(C)C)CCC(C)=CC1	5.07
46*	O=S(=O)(c1ncn(n1)C(=O)N(CC)CC)c2c(cc(cc2C)C)C	5.47
47	c1(c(cc(Cl)c(c1)Cl)Cl)\N=N\Sc1ccc(Cl)cc1	5.40
48	C(O\N=C(/SCCC#N)C)(=O)NC	5.13
49	C(c1ccc(Cl)cc1)(c1ccc(Cl)cc1)(C)O	5.24
50	Clc1n(nc(c1C(=O)Nc2cccc3c2C(OC3(C)C)C)C)C	5.33
51	S=P(Oc2noc(c1ccccc1)c2)(OCC)OCC	5.27
52	O=P(OCC)(Oc1c(Cl)cc(Cl)cc1)Oc1c(Cl)cc(Cl)cc1	5.39
53*	Clc1ccc(OC)c(c1OC)C(=N\OCC)\OC(=O)c2ccccc2	5.32
54*	c1([N+]([O-])=O)c(O[P@@](NC(C)C)(=S)OC)ccc(c1)C	5.20
55	O=C(Sc1ccc(OS(=O)(=O)C)cc1)NC	5.13
56	O2C(CN(C1CCCCCCCCCCC1)CC2C)C	5.11

57	Clc1cc(Cl)ccc1C(OP(=O)(OC)OC)=[C@H]Cl	5.16
58	C(CCCCC)CCCCO	4.82
59	Clc1cc(Cl)c(cc1)N2/N=C(/C(=O)OCC)CC2(C(=O)OCC)C	5.19
60	Fc2ccc(N(C(=O)c1ccccc1)C(C(=O)O)C)cc2Cl	5.13
61	ClCSP(=S)(OCC)OCC	4.97
62	O=C(OCC)C(SP(=S)(OC)OC)c1ccccc1	5.11
63	Clc1ccc(c(Cl)c1)C2(OCC(O2)CCC)Cn3ncnc3	5.12
64	c12c(OC(NC)=O)cccc1cccc2	4.89
65	C(CCCC\C=C\CC)CCCOC(C)=O	4.86
66	O=C(Oc1ccccc1CSCC)NC	4.82
67	Clc1ccccc1CSC(=O)N(CC)CC	4.88
68	O1C(CN(CCCCCCCCCCCCC)CC1C)C	4.94
69	O=C(Nc1ccccc1)C(Oc3ccc2c(cccc2)c3)C	4.92
70	c1(N[C@@H](NC=O)C(Cl)(Cl)Cl)cc(c(Cl)cc1)Cl	4.96
71	O=C(N(c1c(cccc1C)C)C(C(=O)OC)C)Cc2ccccc2	4.94
72	O=C(OC\C=C)C(OC(=O)c1cc(ccc1Cl)N2C(=O)\C=C(/N(C2=O)C)C(F)(F)F)(C)C	5.09
73	Oc1cccc2ccccc12	4.53
74	CCCC(=O)OC(C(Cl)(Cl)Cl)P(=O)(OC)OC	4.82
75	O=C2c1ccccc1C(=O)C2(CC)CC3(OC3)c4cccc(Cl)c4	4.83
76*	c1(cc(Cl)cc(Cl)c1)C(C)(C)N1C(C(=C(OC1)C)c1ccccc1)=O	4.88
77	O1c2cc(c(cc2OC1)COCCOCCOCCCC)CCC	4.81
78	Oc1ccc(Cl)cc1	4.38
79	Clc1nc(nc(n1)N(CC)CC)NCC	4.62

80	O=C(NC(c1ccc(Cl)cc1)C)C2(CC)C(C)C2(Cl)Cl	4.78
81	Clc1cc(Cl)ccc1OC(C(=O)NC(C#N)(C)C(C)C)C	4.75
82	O=C(N(c1ccccc1)C)COc2nc3ccccc3s2	4.70
83	Clc1ccc(NC(=O)N(C)C)cc1Cl	4.54
84	Cc1nnsc1C(=O)Nc1ccc(C)c(Cl)c1	4.58
85	C1(CCCCC1)(NCCCC)P(=O)(OCCCC)OCCCC	4.70
86	Clc1cc(NC(=O)OC(C)C)ccc1	4.45
87*	c1(cc(c(C)cc1)C)OC(NC)=O	4.38
88	CN(C)C(=O)SCCCCOc1ccccc1	4.51
89	S=P(Oc1ccc(C#N)cc1)(OC)OC	4.47
90*	c1(c(cc(O)cc1)C)[N+](=O)[O-]	4.26
91*	COP1(=S)OCc2ccccc2O1	4.39
92*	Clc2nccc(NC(=O)Nc1ccccc1)c2	4.45
93	C1[C@@H](O1)CCl	4.01
94	c1(C(NC([C@@H](C(C)(C)C)Br)=O)(C)C)ccccc1	4.51
95	Clc1nc(cc(n1)N(C)C)C	4.25
96*	c1(Oc2ccc([N+](=O)[O-])cc2)c(cc(Cl)cc1Cl)Cl	4.51
97	Clc1nc(nc(n1)NC(C#N)(C)C)NCC	4.39
99	C(CCCC\C=C/CC)CCCOC(C)=O	4.35
100	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)c2c(Cl)nc3ccc(nn23)CCC	4.66
101*	ClC(=[C@H]Cl)CSC(=O)N(C(C)C)C(C)C	4.43
102	C(C#N)=C	3.72
103*	COc1ccc(Oc2ccc(NC(=O)N(C)C)cc2)cc1	4.43

104	O=S(=O)(Oc2cc1c(OC(OCC)C1(C)C)cc2)C	4.42
105	O=C(Nc1ccccc1)c2c(occ2)C	4.26
107	c1(cc(ccc1)C)OC(NC)=O	4.14
108	N1(C(CCC1)=O)CCCCCCCC	4.21
110	CN(C(=O)NC(C)(C)c1ccccc1)c2ccccc2	4.28
111	O=C(OC)c1cc(Cl)ccc1Cl	4.15
112	O=P(SCc1ccccc1)(OC(C)C)OC(C)C	4.29
113	CC(C)(C)C(=O)C(Oc1ccc(Cl)cc1)n2ccnc2	4.29
114	O=C(Nc1ccc(cc1)C(C)C)N(C)C	4.06
115	Clc1cc(Cl)c(N)cc1	3.91
116	ClCC(Br)CBr	4.07
117	CCS(=O)CSP(=S)(OC(C)C)OC(C)C	4.18
118*	OCC([N+]([O-])=O)(Br)CO	4.00
119	O=C(Oc1ccccc1C(C)C)NC	3.94
121	CCOC(=O)C1(C)OC(=O)N(C1=O)c2cc(Cl)cc(Cl)c2	4.08
122	[nH]1cnc(c1)[C@@H](C)c1cccc(c1C)C	3.82
123	Clc1c(c(c(Cl)c(Cl)c1Cl)C(=O)O)C(=O)Nc2cccc(Cl)c2Cl	4.17
124	c1cnccc1C(=O)Nc1ccc(Cl)cc1[C@@H](O)c1ccccc1	4.05
125*	C(C)=O	3.16
127	CCS(=O)CC(C)SP(=O)(OC)OC	3.81
128	O([C@@H](CCl)C)[C@@H](CCl)C	3.63
129	NC(=S)c1c(Cl)cccc1Cl	3.70
130	C(Cl)(Cl)(Cl)Cl	3.55

131	O=S(=O)(c1nnc(s1)N(C(=O)NC)C)CC	3.78
132*	O=C(\C=C(\c1ccc(F)cc1)c2ccc(OC)c(OC)c2)N3CCOCC3	3.92
133	c1(C(C)C)ccc(C)cc1	3.45
134	C(CCl)(CCl)Cl	3.46
135*	O=P(OCC)(/N=C1\SCC(S1)C)OCC	3.69
136	C1=CC=C2C(=C1)N=CS2	3.32
137	c1cc(ncc1CN(CC(F)F)C2=CC(=O)OC2)Cl	3.61
138*	O=C(Nc1nc(cc(OC)n1)OC)NS(=O)(=O)Oc2ccccc2OCC	3.70
139	c1(c(OS(c2ccc(C)cc2)(=O)=O)n(C)nc1C)C(c1c(cc(Cl)cc1)Cl)=O	3.68
140*	CC(=NOC(=O)Nc1ccccc1)C	3.32
141	CC(Cl)(Cl)Cl	3.15
142	N(C(OCCC)=O)CCCN(C)C	3.29
144*	O=C(O)c2cc(Oc1ccccc1)ccc2	3.34
145*	O=C(O)C(SP(=S)(OC)OC)CC(=O)O	3.44
146*	ClCCP(=O)(O)O	3.16
147	Clc1cc(Cl)ccc1OCC(=O)O	3.34
148	c1cccc(c1NS(=O)(=O)NC(Nc1nc(cc(n1)OC)OC)=O)C(=O)N(C)C	3.63
149	O=C(Oc1nc(nc(c1C)C)N(C)C)N(C)C	3.38
150*	Clc1ncc(cc1)CN(\C(=N\C#N)C)C	3.35
151	CCCCCCCCCCCCCC(=O)OC(C)C	3.43
152*	Clc1ccc(NC(=O)C(C)(C)CCC)cc1	3.38
153*	C(CCCCCCCCCC)CCCOC(C)=O	3.41
154	O1C(C)CNCC1C	3.06

155	c1c(cccc1)Oc1cc(CCC[Si](c2ccc(cc2)OCC)(C)C)ccc1F	3.61
158	C1=CC=[N+]2C(=C1)N(C(=C(C2=O)C3=CC(=CC=C3)C(F)(F)F)[O-])CC4=CN=CN=C4	3.60
159	O=C(c1ccc(cc1)S(=O)(=O)NC(=O)c2ccccc2OC)NC3CC3	3.55
160	Clc1ccccc1C(=N\OP(=S)(OCC)OCC)/C#N	3.44
161	c1(NC(NC)=O)ccccc1	2.94
162	c1(ccccc1)C(O)=O	2.83
163	COC1=C(C=CC(=C1)OC2=C(C=C(C=C2)Cl)Cl)[N+](=O)[O-]	3.12
164*	C1(CC(=CC(C1)=O)C)(C)C	2.76
166	O=C2OCCN2N(C(=O)COC)c1c(cccc1C)C	2.97
167	Cc1c(Cl)cc(cc1Cl)c2ccc(=O)[nH]n2	2.93
168	O=C(O)C3(O)c1ccccc1c2c3cccc2	2.85
169	O=C1OCc2c1c(Cl)c(Cl)c(Cl)c2Cl	2.93
170*	C(C(C(=O)O)N)SC(=O)N	2.66
171*	n1nc(nnc1c1c(cccc1F)F)c1c(cccc1)Cl	2.88
172*	C1(CCCCC1)=O	2.27
173	CC(SCCSP(=O)(OC)OC)C(=O)NC	2.69
174	CC(O)CO	2.03
175	C(CC(O)=O)(CC(O)=O)(C(O)=O)O	2.35
176	c1(C[N+](CC(Nc2c(cccc2C)C)=O)(CC)CC)ccccc1.c1cccc(c1)C([O-])=O	2.65
178	C(CC)CS	1.91
179	FC(F)(F)C([O-])=O	1.97
180	O=C(NC)Nc1nc2ccccc2s1	2.20
181	C(C)#N	1.40

182*	O=C1/C(C(=O)C(C(=O)OC)C(C)(C)C1)=C(/NOC\C=C)CCC	2.21
183*	ClCC(O)CO	1.71
185	ClC(Cl)(Cl)C(O)=O	1.68
186	C1(N(CCC1)C)=O	1.39
187	C[C@@H](CC(C)(C)O)O	1.14
188	S=C(N)N	0.88
189	C1OCOCO1	0.80
190	C(N)(N)=O	0.54
191	c1(CS\C(=N/c2cccnc2)SCCCC)ccc(C(C)(C)C)cc1	1.23
192	C(C)(N)=O	0.35
193	O=C(N(C(=O)OCC)C)CSP(=S)(OCC)OCC	7.74
194	c1(nn(c(c1Cl)C(=O)NCc1ccc(cc1)Oc1ccc(cc1)C)C)CC	6.89
195*	COP(=O)(OC)OC1=C(Cl)C2C=CCC12	6.65
196	O=C\2c1c(cccc1)C(=O)C(/Cl)=C/2N	6.52
197	Fc1ccc(c(F)c1)NC(=O)c3cccnc3Oc2cccc(c2)C(F)(F)F	6.60
198*	O=C(c1c(nn(c1F)C)C)Nc2ccccc2C(C)CC(C)C	6.49
199*	n1(cc(C(Nc2ccsc2[C@@H](C)CC(C)C)=O)c(n1)C(F)(F)F)C	6.09
200	ClC1=C(Cl)C(=O)c2ccccc2C1=O	5.86
201	CC(=O)O[Sn](C1=CC=CC=C1)(C2=CC=CC=C2)C3=CC=CC=C3	6.11
202	O=C(OCC)C(Oc3ccc(Oc1nc2ccc(Cl)cc2o1)cc3)C	5.88
203	C=C(\C1C/C=C(/C)CC1)C	5.29
204	C1(N(C(C=2CCCCC12)=O)c1c(cc(c(c1)O[C@@H](C#C)C)Cl)F)=O	5.54
205	Cc1ccc2c(c1)C(C(C2)C)Nc3nc(nc(n3)N)C(C)F	5.48

206	OCCCCCCCCCCCC	5.27
207*	CCOP(=O)(OCC)OC(=CCl)C1=C(C=C(C=C1)Cl)Cl	5.51
208	O=C(SCc1ccccc1)N(C(C)C(C)C)CC	5.24
209	n1(c(n(C(F)F)c(n1)C)=O)c1cc(C[C@@H](C(=O)OCC)Cl)c(cc1F)Cl	5.41
210	CN(C)C(CSC(=O)N)CSC(=O)N	5.17
211*	O=C(Oc1ccccc1C(C)CC)NC	5.09
212	O(C(OCC)=O)C=1[C@@]2(CC[C@H](CC2)OC)NC(=O)C1c1cc(ccc1C)C	5.28
213	Clc1ccc(Cl)cc1SCSP(=S)(OCC)OCC	5.03
214	c1(C(OCc2ccccc2)=O)ccccc1	4.66
215*	c1cc(F)ccc1N1C(=O)C(Cl)=C(Cl)C1=O	4.67
216*	CCN(CC)C(=O)[C@@H](C)Oc1cccc2c1cccc2	4.61
217	O=C(SCc1ccccc1)N(C(C)CC)C(C)CC	4.54
218	Clc1c(C(=O)O)c(Cl)ccc1Cl	4.42
219	Clc1cc(ccc1[C@@H](C)NC([C@@H](C#N)C(C)(C)C)=O)Cl	4.55
220	O=C(Nc1nc(OC)cc(OC)n1)NS(=O)(=O)c2c(Cl)nc3ccccn23	4.62
221	C(COC)(=O)O[C@@H]([C@@H](C)F)c1c(cccn1)S(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC	4.69
222	CCOc1ccc(cc1)C(COCc2cccc(Oc3ccc(Cl)cc3)c2)C(F)(F)F	4.65
223	C(CCCCCCCCCC)CCCCC(O)=O	4.33
224	c1(ccccc1)CCl	4.01
225	O=C(NC)C(=N\OC)\c2ccccc2Oc1ccccc1	4.20
226	CCNc1nc(NC(C)(C)C#N)nc(SC)n1	4.13
227	O=C3O/C(C(=O)N3c2c(F)cc(Cl)c(OC1CCCC1)c2)=C(/C)C	4.22

228*	FCC(=O)N	3.28
229	O=C(Oc1cc(cc(c1)C)C)NC	3.65
230	O=S(=O)(O)O	3.37
231	CCCCc1c(C)nc(nc1O)N(C)C	3.70
232*	c1(CNC(=O)NC(c2ccccc2)(C)C)ccccc1Cl	3.78
233	CC(N(C(=O)COC)c1c(C)cccc1C)C(=O)O	3.42
234	OP(O)=O	2.91
235*	c1(c(cccc1Cl)Cl)C(O)=O	3.17
236	ClCCCl	2.86
237*	Clc1cc(c(OC(C(=O)O)C)cc1)C	2.95
238	O=CC=O	1.88
240*	FC(F)(F)c1nnc(s1)N(C(=O)NC)C	2.38
241	O=C(O)C21OC(OC1C3OC(OCC3O2)(C)C)(C)C	1.74

3.2. General principles of the methodologies employed to develop the q-RASAR/q-RASPR models along with study wise precise explanation of methodologies exploited in each case study.

3.2.1. Dataset curation

Dataset (Case study 1)

The present work deals with the q-RASPR modeling of diverse classes of pesticides with a defined endpoint $\log t_R$ of 823 pesticides taken from the previous literature (Wang.et. al., 2019). There are fair representations of cyclic (alicyclic and aromatic; homocyclic and heterocyclic structures) and acyclic (straight and branched chain structures) chemicals in the collected dataset. The two-dimensional structures of the aforesaid chemicals were retrieved

(*.sdf format) from PubChem database and were cross-verified with another popular chemical database ChemSpider. After manual checking of all the structures, chemical curation was performed by using a chemical curation workflow in the Knime Platform to remove inorganic salts, ions, radicals, repetitive compounds, and mixtures from the dataset. However, all the 823 compounds passed the curation and the curated structures were selected for modeling in this study; the reported retention time ($\log t_R$) data of curated 823 compounds has been represented in the following **Table 3.1.**

Dataset (case study 2)

The present work deals with the novel q-RASAR modeling approach using experimental data (Log 1/LC50) of organic pesticides to various fish species, including Rainbow trout (RT: *Oncorhynchus mykiss*: 829 data points), Lepomis (LP: *Lepomis macrochirus*: 151 data points), and (Miscellaneous data set: *Pimephales promelas*, *Brachydanio rerio*: 278 data points) accordingly taken from the previous literature (Li et al., 2017). Most of the 2D structures of the aforementioned chemicals were obtained (*.sdf format) from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) and cross-checked with another popular chemical database, ChemSpider (http://www.chemspider.com). Also, the remaining 2D structures were retrieved manually and we compared them with the chemical names, and the Chemical Abstract Service (CAS) numbers using online databases (ChemCell) (https://github.com/cdd/chemcell); MarvinView (Marvin 17.28.0, 2017) and ChemIDplusAdvanced (https://chem.nlm.nih.gov/chemidplus/). After the manual checking of all the structures, chemical curation was performed manually by excluding the compounds having an unknown chemical structure, inorganic compounds, metal complexes, salts containing organic polyatomic counterions, mixtures, and substances of unknown or variable composition (UVCB) for modelling purposes. Furthermore, in RT (rainbow trout) and others datasets, we also excluded the compounds having higher hi values based on the leverage calculation for modeling (Adeniji et al., 2019). Finally, we have selected 715 data points for Rainbow trout (RT), 136 data points for Lepomis (LP) and 226 data points for Miscellaneous fish species to develop the models in this study. The reported (Log 1/LC50) values of these data points in RT, LP and Miscellaneous datasets have been tabulated in the following **Tables 3.2, 3.3, and 3.3.**

3.2.2. Molecular descriptors calculation

Dataset (Case study 1)

Molecular descriptors are the mathematically encoded numerical representations of various structural and physicochemical features of chemical entities which are used to correlate the structural aspects of chemicals with the exerted molecular properties (activity/property/toxicity) in QSAR studies (Todeschini et.al., 2000). There are various types of descriptors which are used to develop QSAR models. The 0D – 7D descriptors can be used for modeling and the level of complexity increases from 0D to 7D **(**Roy et. al.,2015). In this work, we have developed our models utilizing only 2D descriptors; they are extremely effective for avoiding the intricacy of alignment issues as well as the computational complexity of conformational analysis and energy minimization **(**Chatterjee et. al.,2022). The 2D structures of the molecules were drawn using Marvin Sketch software (v14.10.27) (https://chemaxon.com/marvin)with proper aromatization and explicit hydrogen addition. In this study, nine classes of 2D descriptors namely, electro-topochemical atom descriptors, molecular properties, constitutional index, functional group count, ring descriptors, atom-centred E-state indices, 2D atom pairs, connectivity index, atom-centred fragments have been computed by employing AlvaDesc software **(**Mauri et. al.,2020). These descriptor classes were selected based on their easy interpretability and usefulness in previous studies. After descriptor calculation, the inter-correlated descriptors (inter-correlation cut-off >0.95) and the descriptors with the least variability in values (variance cut-off <0.0001) were removed by the in-built pre-treatment tool of AlvaDesc software **(**Mauri et. al.,2020). As an additional measure, the flawed and extraneous data were eliminated by employing an in-house tool namely 'DataPreTreatmentGUI1.2_3March2016' before using the data in the model development, and a pool of 931 2D descriptors have been used in the preliminary modeling.

Dataset (case study 2)

Molecular descriptors are used to correlate the structural characteristics of chemicals with the exerted molecular properties (activity/property/toxicity) in most of the QSAR studies (Todeschini and Consonni, 2000). There are several types of descriptors which are used to build QSAR models. We can use 0D-7D descriptors for modelling, but the level of complexity increases from0D to 7D descriptors (Roy et al., 2015). To avoid the conformational complexity that arises due to the inclusion of higher-order of descriptors (3D-7D descriptors), we incorporated only 0D-2D descriptors in the present work. These

preliminary descriptors (0D – 2D) are not only useful for avoiding the difficulties of alignment issues and computational complexity of conformational analysis and energy minimization, but they are also extremely effective to mark the structural features for the ease of interpretability and reproducibility in comparison to 3D descriptors (Chatterjee and Roy, 2022). Also, the computation of 2D descriptors is very fast and cost-effective which enhances the convenience. There are many previous studies where activity/property/toxicity has been efficiently correlated with the structural features encoded by the 2D descriptors (Chatterjee and Roy, 2023). The 2D structures of chemicals were drawn in Marvin Sketch software (version 14.10.27) (Chemaxon. *Marvin*; https://chemaxon.com/marvin accessed Feb 27, 2023) with relevant aromatization and explicit hydrogen addition. Selected classes of descriptors with specific physicochemical meanings (namely, electro-topochemical atom index descriptors (ETA), molecular properties, constitutional index, functional groups count, ring descriptors, atom type E-state indices, 2D atom pairs, connectivity index, and atom-centered fragments) have been calculated using Alvadesc software (Mauri, 2020) for simplifying the interpretation. After the calculation of descriptors, the redundant and flawed descriptors (highly intercorrelated/least variability) were removed from the pool by applying an intercorrelation cutoff of >0.95 and variance cutoff of <0.1 in the in-built pretreatment operation of AlvaDesc software for all the datasets.

3.2.3. Dataset division

Dataset (case study 1)

Dataset division is an essential part of QSAR model development. As a standard procedure, the dataset is divided into training and test sets; the training set is solely employed for model development whereas the test set is exclusively used for external validation of the developed model. In the present work, the sorted activity-based division technique has been used by employing the "datasetDivisionGUI1.2_19Feb2019" tool (accessible from http://dtclab.webs.com/software-tools) for 2D q-RASPR modeling. It is a response-based division technique which divides a complete dataset into clusters; the data points inside a cluster are similar to one another but different from those found in other clusters (Roy, 2015). Although there is no stringent ratio for dataset division, 70:30 to 80:20 proportions of training and test data points have been found in many previously published QSAR studies. We chose around 25% of the data points from each cluster to serve as test set compounds (N_{test}=205) and the remaining 75% to serve as the training set compounds (N_{train}=618) after organizing

the entire dataset according to the cluster number and the corresponding activity levels. This method ensures a uniform representation from all the clusters into training and test sets for feature selection and QSPR model development.

Dataset (case study 2)

Dataset division is a necessary part of QSAR model development. Following the standard procedure, the datasets were divided into training and test sets. The training sets were exclusively employed for model development and internal validation and the test sets were effectively used for the external validation of the developed models. In the current work, we have employed Euclidean distance-based division for RT and LP datasets using "datasetDivisionGUI1.2_19Feb2019" tool (accessible from http://dtclab.webs.com/software-tools), and a manual sorted activity-based division for Miscellaneous dataset. In RT and LP datasets, we used a 75:25 proportion of training and test data points (RT: 4-3 distance-based division; LP: 4-1 distance-based division) whereas, in the third dataset (Miscellaneous: 4-4 distance-based division), a 75:25 proportion of training and test data points were used. The Euclidean distance-based approach is a descriptor-based (X-based) division strategy, which has been used for all the datasets in this study. Another pretreatment operation was performed using "DataPretreatment Train-Test v1.0" tool for removing the redundant descriptors obtained after dataset division. Therefore, we have moved forward towards the feature selection process with 392 descriptors in "RT dataset", 389 descriptors in "LP dataset" and 348 descriptors in "Miscellaneous dataset".

3.2.4. Feature selection and development of QSPR/QSAR model

Dataset (case study 1)

In this present study, we have developed a QSPR model using the retention time ($\log t_R$) of pesticide residues as the response variable for the model development. The feature selection was started with 931 2D descriptors and the response variable $\log t_R$. We have employed the genetic algorithm (GA) approach for feature selection. The GA technique was primarily applied by Rogers and Hopfinger (Rogers et. al.,1994) in QSAR investigations as a very effective tool with numerous advantages over other variable selection methods. It examines a variety of potential answers concurrently, each of which probes a unique area of the vector space delineated by computed descriptors. When performing QSPR modeling, GA uses a fitness function based on MAE-based criteria to pick significant descriptors (variables). We

have employed an in-house tool namely "GeneticAlgorithm_v4.1_Train" (available from: https://dtclab.webs.com/software-tools) to identify the most correlated descriptors with the response variable. After identifying the important descriptors, we used the "Partial Least Squares (PLS) regression" to build the preliminary QSPR models. The PLS regression approach is an extended version of the "Multiple Linear Regression (MLR)" method that allows us to study strongly correlated, noisy, and collinear data as well as a large number of X-variables (dos Santos.et.al.,2023). The PLS regression has been carried out with a java-based software tool "PLS_SingleY_version 1.0" (available from: https://dtclab.webs.com/software-tools).

Dataset (case study 2)

In this present study, we have developed three QSAR models for RT, LP and Miscellaneous datasets using the aquatic toxicity endpoint (Log 1/LC50) of organic pesticides. We have employed the stepwise selection approach in multiple iterations for extracting the important descriptors based on partial F statistic (F-for-inclusion = 4; and F-for-exclusion= 3.9 as stepping criterion) for all the datasets (Khan et al., 2019); thereby, reduced pools of descriptors (RT: 14 descriptors; LP: 24 descriptors; Miscellaneous: 22 descriptors) were obtained. With these reduced pools of descriptors, all possible subset selection was performed using "BestSubsetSelection v2.0" tool (available at http://teqip.jdvu.ac.in/ QSAR_Tools/) and the final combinations of descriptors were selected based on the computed cross-validated correlation coefficient (Q^2_{LOO}) as a measure of robustness. Finally, the selected combinations of descriptors of all datasets were subjected to partial least squares (PLS) regression (Wold et al., 2001) using "PLS_Single_Y v1.0" tool (available at http://teqip.jdvu.ac.in/QSAR_Tools/) to develop the preliminary QSAR models with an optimum number of latent variables. The PLS regression strategy, which is an extended form of the "multiple linear regression (MLR)", permits us to examine the data that are highly correlated, noisy, collinear, and contain a significant number of X variables (dos Santos et al., 2023). Acceptance of QSAR models for prediction significantly depends on the performance which can be determined using a variety of globally recognized validation metrics as per the OECD guidelines. The developed models for aquatic toxicity [Log (1/LC50)] of organic pesticides towards RT, LP and miscellaneous fish species were rigorously validated using various stringent internal and external validation tests for all the datasets (Roy and Mitra, 2011). The training sets were rigorously validated using internal validation metrics like leave-one-out (LOO) cross-validation (Q^2_{LOO}) and mean absolute error (MAE_{train}), while the test sets were rigorously

validated using external validation metrics like Q^2_{F1}, Q^2_{F2} and mean absolute error (MAE_{test}) (Roy and Mitra, 2011).

3.2.5. Development of q-RASPR/q-RASAR model

Dataset (case study 1)

The q-RASPR is a conjuncture of read-across and QSPR modeling where the read-across tool derived similarity, error, and concordance measures (RASAR-descriptors) are used along with the previously identified descriptors to develop the QSAR-based final predictive models (Banerjee and Roy,2022). To calculate the RASAR descriptors, optimization of the read-across algorithm is a prior step. We have used the previous training and test sets with the identified 2D descriptors for determining the optimum read-across method. The training and test sets have been used as the source and target compounds respectively, for read-across predictions. "Read-across v4.1" tool (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) has been employed for the optimization where Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity have been used for identifying the most similar compounds of target chemical for further weighted average predictions (Chatterjee.et.al.,2022). We have used the default setting, i.e., sigma =1, gamma = 1, no. of close training compounds = 10, distance threshold = 1, similarity threshold = 0 for method optimization with sub-training and sub-test sets (obtained after dividing the training set in 75:25 ratio).As per the result of optimization (provided in **Table 3.5**), Laplacian kernel function similarity read-across is the least error-prone method (as suggested by the least MAE value: **0.691**) for this dataset and has been used for RASAR descriptor calculation. Finally, the RASAR descriptors were calculated by using the "RASAR descriptor v2.1" tool (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) after employing Laplacian kernel function similarity estimation as the optimized algorithm with gamma = 1, no. of closed training compounds = 10, and similarity threshold = 0 setting. The training and test sets have been used as the source and target compounds for the RASAR descriptor calculation of test set chemicals. However, the RASAR descriptor calculation of training set chemicals is a little bit tricky; the training set has been simultaneously used as the source and target compounds here. A "leave-same-out algorithm" (Banerjee and Roy,2023) has been applied in the training RASAR descriptor calculation where the same training chemical with query compound is identified and removed before further procedures.

Finally, the pooled combination of RASAR descriptors and previously identified structural and physicochemical descriptors has been subjected to best subset selection. Based on the cross-validated correlation coefficient (Q^2_{LOO}) of combinations, eight descriptors were selected and modeled using PLS regression with optimum latent variables. The final PLS q-RASPR model was thoroughly assessed by both internal and external validations (Roy et. al.,2015). For a better understanding of the modeling, a schematic representation has been given in the following **Figure 3.1**.

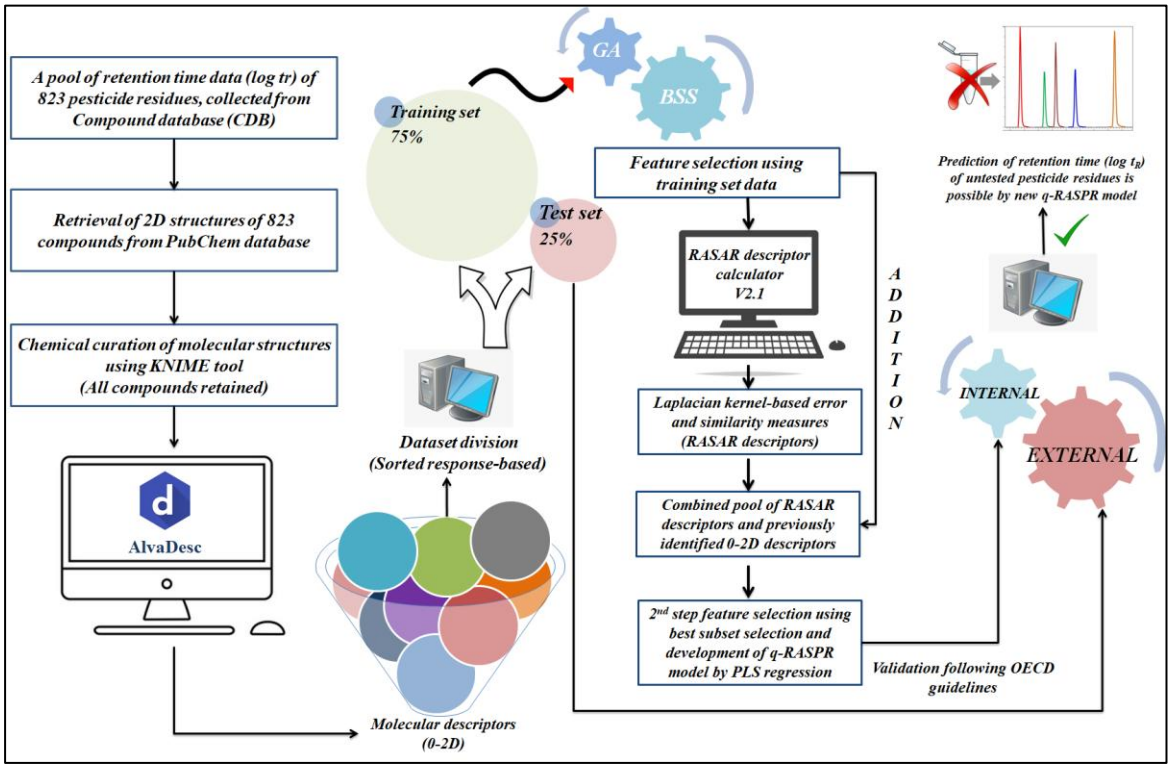


Fig.3.1. Flow diagram of q-RASPR model development (GA= genetic algorithm, BSS= best subset selection)

Table 3.5. Optimization of similarity technique used for RASAR descriptor calculation

Type of Similarity	MAE	Parameters of similarity measurement
Euclidean distance based similarity	0.741	No.of similarity =10; Distance threshold = 1
Gaussian kernel function similarity	0.694	Sigma value =1; No.of similarity =10; Similarity threshold =0
Laplacian kernel function similarity	**0.691(least)**	Gamma value = 1; No.of similarity = 10; Similarity threshold = 0

Dataset (case study 2)

The q-RASAR model is a combination of read-across prediction and QSAR approach, where, the read-across derived non-linear prediction function, similarity, error, and concordance measures (RASAR descriptors are combined with the important structural and physicochemical features to build the linear QSAR model (Banerjee and Roy, 2022; Luechtefeld et al., 2018)). Although the q-RASAR model may be developed from both discrete and continuous data to form classification and regression-based models, in this study only regression-based q-RASAR models have been reported for all the datasets. Before calculating the RASAR descriptors, optimization of the read-across algorithm is a prerequisite. Chatterjee and co-workers reported a read-across methodology where they used Euclidean distance (ED), Gaussian kernel function (GK), and Laplacian kernel (LK) function for estimating the similarity between source and target compounds (https://doi.org/10.1016/j.chemolab.2022.104613). The ED-based similarity, GK and LK-function similarity have been estimated by the mathematical formula represented in the following equations 1.16, 1.17, and 1.18, respectively:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \ldots (1.16)$$

where, Euclidean distance between two points X and Y has been represented yd(X, Y), whereas, X_i is the i^{th} descriptor of compound X, and Y_i is the i^{th} descriptor of compound Y, where i = 1, 2, 3, ..., n. The ED-based similarity has been computed by subtracting the scaled (0-1) distance from 1 because distance is the reciprocal of similarity.

$GK(X,Y) = f = e^{-\|X_i - Y_i\|/2\sigma^2}$....(1.17)

where, Gaussian kernel function similarity *(f)* between two compounds X and Y, which has been represented by GK(X, Y)and $\| X_i - Y_i \|^2$ is the Euclidean norm or L^2 norm, which can be measured by squaring the Euclidean distance, and σ indicates the width of Gaussian kernel function and can never become zero.

$LK(X,Y) = \hat{e} = e^{(-\gamma\|X-Y\|_1)}$.....(1.18)

Where, LK(X, Y) is the Laplacian kernel function similarity (κ) between two compounds X and Y; γ signifies a non-zero and positive number which determines the performance of the Laplacian kernel function.

Those three methods have been used, and the optimal one has been identified for the corresponding RT, LP and Miscellaneous datasets respectively, as represented in the **Tables 3.6, 3.7, and 3.8**. To do so, we have performed the read-across prediction with corresponding training and test sets using the default setting of read-across (sigma = 1, gamma = 1, number of close training compounds = 10, distance threshold = 1, similarity threshold = 0) in the "Read-across v4.1" tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). Only the identified descriptors of QSAR modeling step have been used for read-across method optimization. The best method was selected based on the least mean absolute error of the test set value (MAE_{Test}).Finally, the RASAR descriptors were calculated using the "RASAR descriptor v2.1" tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home)by employing the optimized method of the corresponding dataset. For calculating the RASAR descriptors of the test set chemicals, the training sets have been used as the source compounds, while the test sets have been employed as the target compounds. However, the RASAR descriptor calculation of training set chemicals is a little bit tricky because the training set has been simultaneously employed as the source and target compounds. Therefore, to avoid the chance of overfitting, the training RASAR descriptor calculation has been performed using the "leave-same-out" algorithm(Chatterjee et al., 2023), where the same training compound with a query compound is identified and eliminated before weighted averaging.

For developing the final q-RASAR models, the RASAR descriptors were pooled along with the previously identified structural and physicochemical descriptors for all possible subset selection. Two different combinations of eight descriptors were identified for RT and LP

datasets, and a combination of five descriptors was identified for Miscellaneous dataset based on the cross-validated correlation coefficient (Q^2_{LOO}). The final q-RASAR models were developed with the identified descriptors (from the pool of RASAR, physicochemical, structural) by employing PLS regression with an optimum number of latent variables. These q-RASAR models were thoroughly assessed using both internal and external validations as recommended by OECD guidelines (OECD, 2004), and all of the results have been analyzed in the Results and Discussion part. A schematic representation has been provided in (**Fig. 3.2**) for a better understanding of the modeling of all datasets.

3.2.6. Statistical Validation

Statistical validation is an inevitable part of QSAR modeling. As per the OECD principle 4, the QSAR model should be a good fit with the training data, sufficiently robust and externally predictive. Thus, to comply with the norm, we have computed several internal and external validation metrics for assessing the quality of developed models in this study. The internal validation metrics such as determination coefficient (R^2), adjusted R^2 (R^2_{Adj}), and cross-validated correlation coefficient (Q^2_{LOO}), mean absolute error of training set (MAE_{train}) have been computed from the training data, whereas, external validation metrics like external correlation coefficient (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}), mean absolute error of test set (MAE_{train}) etc. have been computed with the test data. The determination coefficient (R^2 or R^2_{Adj}) and cross-validated correlation coefficient (Q^2_{LOO}) are the measurement parameters of goodness of fit and robustness, mean absolute errors (MAE_{Train} or MAE_{Test}) are the measures of error of predictions, and the external correlation coefficients (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}) are the measures of the external predictability of the model. The threshold limits for external validation metrics ($Q^2_{F1}/Q^2_{F2}/Q^2_{F3} \geq 0.5$) along with internal validation metrics ($R^2/R^2_{adj} \geq 0.6$, $Q^2_{LOO} \geq 0.5$) have been mentioned as per the criterion of Goldbaikh and Tropsha for defining the statistical quality of the regression-based models (Roy et al., 2015; Roy and Mitra, 2011).The important validation metrics which are used in this study, have been computed by the following mathematical equations (1.19-1.26):

Internal validation metrics:

$$R^2 = 1 - \frac{\sum(Y_{obs} - Y_{calc})^2}{\sum(Y_{obs} - \bar{Y}_{train})^2} \ldots\ldots (1.19)$$

$$R^2_{adj} = \frac{\{(n-1) \times R^2\} - p}{n - p - 1} \ldots\ldots (1.20)$$

$$Q^2_{Loo} = 1 - \frac{\sum\left(Y_{obs(train)} - Y_{Pred(train)}\right)^2}{\sum\left(Y_{obs(train)} - \bar{Y}_{train}\right)^2} \ldots\ldots (1.21)$$

$$MAE_{Train} = \frac{\sum|Y_{Obs(Train)} - Y_{Pred(Train)}|}{n_{Train}} \ldots\ldots (1.22)$$

External validation metrics:

$$Q^2_{F1} = 1 - \frac{\sum\left(Y_{obs(test)} - Y_{Pred(test)}\right)^2}{\sum\left(Y_{obs(test)} - \bar{Y}_{train}\right)^2} \ldots\ldots (1.23)$$

$$Q^2_{F2} = 1 - \frac{\sum\left(Y_{obs(test)} - Y_{Pred(test)}\right)^2}{\sum\left(Y_{obs(test)} - \bar{Y}_{test}\right)^2} \ldots\ldots (1.24)$$

$$Q^2_{F3} = 1 - \frac{\left[\sum\left(Y_{obs(test)} - Y_{pred(test)}\right)^2\right]/n_{test}}{\left[\sum\left(Y_{obs(train)} - \bar{Y}_{train}\right)^2\right]/n_{train}} \ldots\ldots (1.25)$$

$$MAE_{test} = \frac{\sum\left|Y_{obs(test)} - Y_{Pred(test)}\right|}{n_{test}} \ldots\ldots (1.26)$$

In this study, we have also performed the Y-randomization test to ensure that the developed q-RASAR models are not the result of any chance correlation. Validation in Y-randomization is accomplished by permuting the response values (Y) 100 times with respect to the X matrix, which remains unchanged. The Y-randomization test has been performed using "SIMCA-P software" (Wu et al., 2010) in this study. As per the protocol, the acceptable value of R^2 and Q^2 intercepts of the Y-randomization plot must be below 0.2 and 0.05 respectively**.**

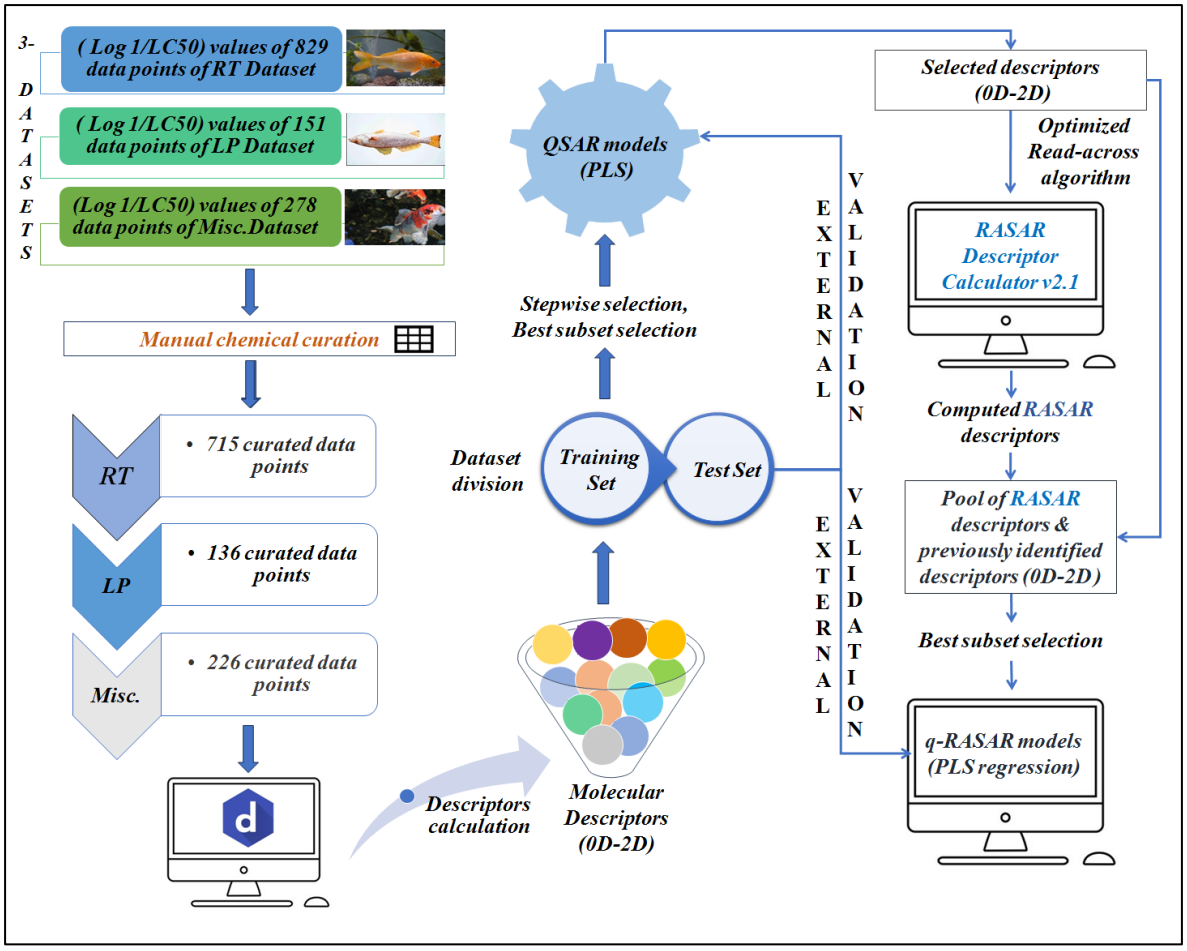


Figure 3.2. Schematic diagram of q-RASAR model development of RT, LP and Miscellaneous fish species (Misc. : Miscellaneous)

Table 3.6. Optimization of similarity technique used for RASAR descriptor calculation for RT (Rainbow trout) dataset

Type of Similarity	MAE	Parameters of similarity measurement
Euclidean distance based similarity	**0.725 (least)**	No.of similarity =10; Distance threshold = 1
Gaussian kernel function similarity	0.727	Sigma value =1; No.of similarity =10; Similarity threshold =0
Laplacian kernel function similarity	0.728	Gamma value = 1; No.of similarity = 10; Similarity threshold = 0

Table 3.7. Optimization of similarity technique used for RASAR descriptor calculation for LP (Lepomis) dataset

Type of Similarity	MAE	Parameters of similarity measurement
Euclidean distance based similarity	0.927	No.of similarity =10; Distance threshold = 1
Gaussian kernel function similarity	**0.841 (least)**	Sigma value =1; No.of similarity =10; Similarity threshold =0
Laplacian kernel function similarity	0.872	Gamma value = 1; No.of similarity = 10; Similarity threshold = 0

Table 3.8. Optimization of similarity technique used for RASAR descriptor calculation for Miscellaneous dataset

Type of Similarity	MAE	Parameters of similarity measurement
Euclidean distance based similarity	0.908	No.of similarity =10; Distance threshold = 1
Gaussian kernel function similarity	0.850	Sigma value =1; No.of similarity =10; Similarity threshold =0
Laplacian kernel function similarity	**0. 820(least)**	Gamma value = 1; No.of similarity = 10; Similarity threshold = 0

CHAPTER - 4

RESULT AND DISCUSSION

"*Everything should be made as simple as possible, but not simpler*."

– Sir Albert Einstein

4. RESULTS AND DISCUSSIONS

4.1. Predictive q-RASPR modeling of the retention time (log t_R) of pesticide residues present in foods and vegetables

4.1.1. Developed QSPR model for the retention time (log t_R)

A QSPR model has been developed in this study by employing PLS regression with five latent variables (LVs) from the identified eight different features by the best subset selection tool. The developed model has been represented in the following equation **(1):**

$$\text{Log}\, t_R = 14.36 + 0.13 \times O - 060 + 0.43 \times F05[N - S] + 1.04 \times LOGPcons - 15.04 \times GD - 1.57 \times nRNR2 - 0.089 \times nBM + 0.50 \times B02[O - S] - 7.47 \times Me \quad \textbf{(1)}$$

The performance of the model has been assessed using strict internal and external validation following the OECD guidelines. Computed values of internal validation metrics such as determination coefficient (R^2= 0.804), adjusted R^2 (R^2_{adj} = 0.789), leave-one-out cross-validated correlation coefficient (Q^2_{LOO} = 0.795) and r^2_m metrics of training set ($\overline{r_m^2}(Train) = 0.708, \Delta r_m^2(Train) = 0.162$) signify the goodness of fit and robustness of the model, whereas, the mean absolute error of training set predictions (MAE_{Train}= 0.611) points out the error of predictions. Along with these, the external validation metrics i.e., external predicted variance (Q^2_{F1} = 0.82, Q^2_{F2} = 0.82), mean absolute error of test set predictions (MAE_{Test} = 0.67), $\overline{r_m^2}(Test) = 0.72, \Delta r_m^2(Test) = 0.15$, concordance correlation coefficient (CCC = 0.90) etc. have also been calculated which are the classic markers of good external predictability. The identified structural descriptors have been tabulated in the following **Table 4.1** along with the definition and class.

Table 4.1. Identified structural descriptors with definition and class

Descriptor	Class	Definition
O-060	Atom-cantered fragments	It indicates Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X type of atom-cantered fragments.
F05[N-S]	2D atom pairs	It indicates the frequency of

		N-S atom pair at topological distance 5.
LOGPcons	Molecular properties	It indicates the octanol-water partition coefficient (log P).
GD	Constitutional indices	It is termed as the graph density, expressed by- $GD = 2nBo/nSK(nSK-1)$ where *nBo* is the number of edges in the H-suppressed graph, *nSK* is the number of vertices in the H-suppressed graph.
nRNR2	Functional group counts	It indicates the number of tertiary amines (aliphatic).
nBM	Constitutional indices	It indicates the number of multiple bonds present in a molecule

4.1.2. Developed q-RASPR model for the retention time ($logt_R$)

Our objective was to develop a q-RASPR model and enhance the external predictability of the corresponding QSPR model further in this study. Thus, we combined the pool of eight structural descriptors along with the computed read-across-based RASAR descriptors, and finally obtained the new combination of descriptors using the best subset selection method.PLS regression has been utilized to model the final descriptors with the optimum number of latent variables (5 LVs) and the developed PLS q-RASPR model has been shown in the following equation **(2)**:

$$\text{Log}\, t_R = 6.90771 - 3.91482 \times \text{CVact(LK)} + 0.73673 \times \text{Pos. Avg. Sim} + 0.30878 \times \text{F05[N - S]} + 0.88941 \text{ x LOGPcons} - 10.04724 \times \text{GD} - 0.86658 \times \text{nRNR2} - 0.07316 \times \text{nBM} + 0.53026 \times \text{B02[O - S]} \quad (2)$$

This final model has been rigorously validated employing several internal and external validation metrics following the OECD guidelines and the computed metrics have been given in the following **Table 4.2**.

Table 4.1. Computed validation metrics of the PLS q-RASPR model

Internal validation metrics	
R^2	0.81
R^2_{adj}	0.81
Q^2_{LOO}	0.81
MAE_{Train}	0.59
$\overline{r_m^2(Train)}$	0.73
$\Delta r_m^2(Train)$	0.15
External validation metrics	
Q^2_{F1}	0.84
Q^2_{F2}	0.84
$\overline{r_m^2(Test)}$	0.75
$\Delta r_m^2(Test)$	0.14
MAE_{Test}	0.63

The acceptable range of internal validation metrics like R^2, R^2_{adj}, Q^2_{LOO}, and r_m^2 metrics ($\overline{r_m^2}(Train), \Delta r_m^2(Train)$) justify the model as a robust and good fit, whereas, the performance of the model on external compounds has been justified by the acceptable range of computed external validation metric values (Q^2_{F1}, Q^2_{F2}). Apart from the conventional

validation procedure, we have performed the Y-randomization of the model with 100 times random shuffling of response values (Log t_R). The Y-randomization plot has been shown in the following **Figure 4.1A**, where the Q^2 and R^2 values of the original model have been plotted against 100 new Q^2 and R^2 values of the randomly developed model. The intercepts of trend lines obtained by R^2 and Q^2 plots (R^2_Y and Q^2_Y) have been calculated, and these are -0.012 and 0.0946 respectively. Both of these values are less than the threshold limits (R^2_Y = 0.3, Q^2_Y = 0.05) which confirm that the model was not the result of any chance correlation. An observed vs predicted scatter plot has also been produced with training and test set data, and has been given in **Figure 4.1B**. The uniform scattering and equal distribution of training and test data points surrounded by the trend line indicate the goodness of fit of the generated model as well as how well the observed values correspond with the predicted ones.

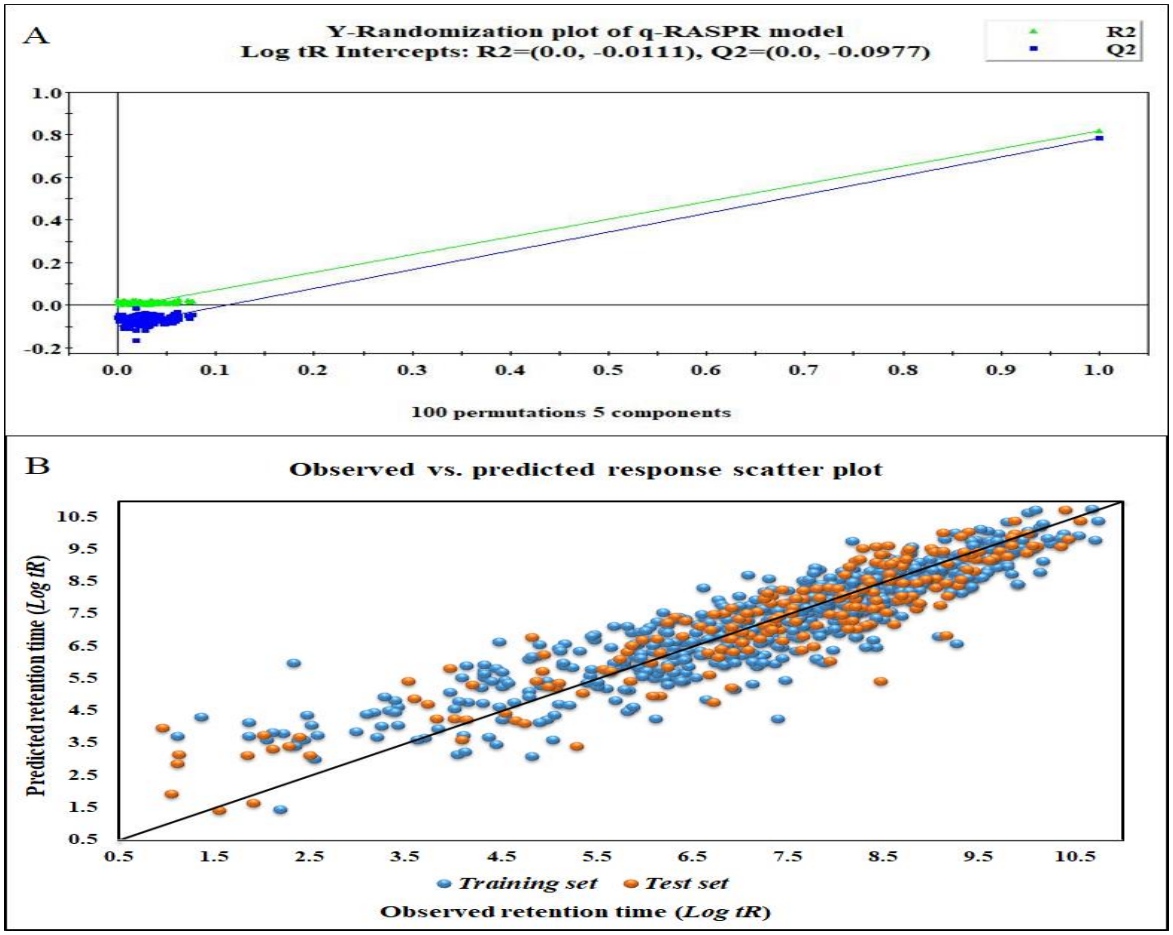


Figure 4.1. (**A**) Y-Randomization plot of the q-RASPR model; (**B**) Observed vs predicted response (Logt_R) in the scatter plot (q-RASPR model)

4.1.3. PLS plots

4.1.3.1. Coefficient plot

The descriptors appearing in the final q-RASPR model can be grouped into two classes depending on their regression coefficient values. The modeled descriptors are plotted against their corresponding regression coefficient in Y-axis to obtain the coefficient plot (Wold et al., 2001). Descriptors with positive regression coefficients are represented as upward columns whereas the descriptors with negative regression coefficients are drawn as downward columns from the X-axis. Here in this study, the coefficient plot has been obtained by using SIMCA-P software and has been presented in the following **Figure 4.2A**. According to the coefficient plot, LOGPcons, F05[N-S], B02[O-S], and *Pos.Avg.Sim* are positively contributing descriptors, whereas GD,nBM, nRNR2 and *CVactivity (LK)* are negatively contributing features of the developed model.

4.1.3.2. Variable Importance plot (VIP)

Variable importance plot (VIP) defines the relative importance of the modeled descriptors towards the endpoint ($logt_R$) predictions by a particular regression model (Akarachantachote et al., 2014). It is essentially a column plot where the VIP scores of each descriptor are plotted on the Y-axis and the modeled descriptors are plotted on the X-axis in descending order of importance from left to right. As per the standard practice, the descriptors with VIP score >1 are considered to be the most important for endpoint prediction. We have generated the VIP of developed q-RASPR by utilizing SIMCA-P software, which has been showcased in the following **Figure 4.2B**. According to the VIP, the order of the relative importance of contributing descriptors towards the response values appearing in the model is LOGPcons>CVactivity (LK)>GD>nBM>Pos.Avg.Sim>B02[O-S]> nRNR2> F05[N-S]. LOGPcons, CVactivity (LK), and GD are the most important descriptors which impact the endpoint prediction to a great extent. Apart from these, nBM and Pos.Avg.Sim.are the descriptors having moderate importance towards the response ($logt_R$) prediction because their VIP scores are close to 1, and they are located in the middle region of the plot. B02[O-S], nRNR2, and F05[N-S] have the least importance as they have very low VIP scores (<1) and are located far right in the plot.

4.1.3.3. Score plot

A score plot depicts the distribution of compounds in the latent variable space (up to 2 LVs) as determined by the scores (De.P.et.al.,2018). The scores of the first two components t1 and t2 have been shown below in **Figure 4.2C**. The ellipse represents the model's application domain, as described by Hotelling's t^2. Hotelling's t^2 is a multivariate version of the t-test developed by Student. It serves as a screening for compounds that adhere to multivariate normality. By this plot, compounds which are located near each other have similar kind of characteristics or properties, whereas, compounds which are far from each other have dissimilar properties with respect to the retention time (log t_R) of pesticide residues. Compounds which are close to the centre possess average properties and we can conclude that compound no.: **7**, **62**, **251**, **525**, **and 797** are identified as outliers according to the plot because these compounds are situated at the outside of the eclipse.

4.1.3.4. Loading plot

The loading plot of this q-RASPR has also been generated with SIMCA-P software. This plot analyses the loading of both descriptors and response of a regression model considering up to the 1^{st} two latent variable spaces; therefore, the knowledge of the interrelationship between X (descriptor) and Y (response) variables can be obtained from the loading plot (Wold et al., 2001). The distance of the X-variables from the origin correlates with the relevance of the descriptor. As per the loading plot given in **Figure 4.2D**, we can see that LOGPcons, Pos.Avg.Sim, CVactivity (LK) and GD descriptors provided the greatest impact to predict the endpoint (logt_R) as these are located far away from the origin, and nBM, B02[O-S], nRNR2 and F05[N-S] descriptors displayed the less impact on the prediction of endpoint(logt_R) as these are placed near the origin. It can also be inferred from the loading plot that the LOGPcons is the most correlated descriptor with the response variable logt_R and this can be justified by the VIP also.

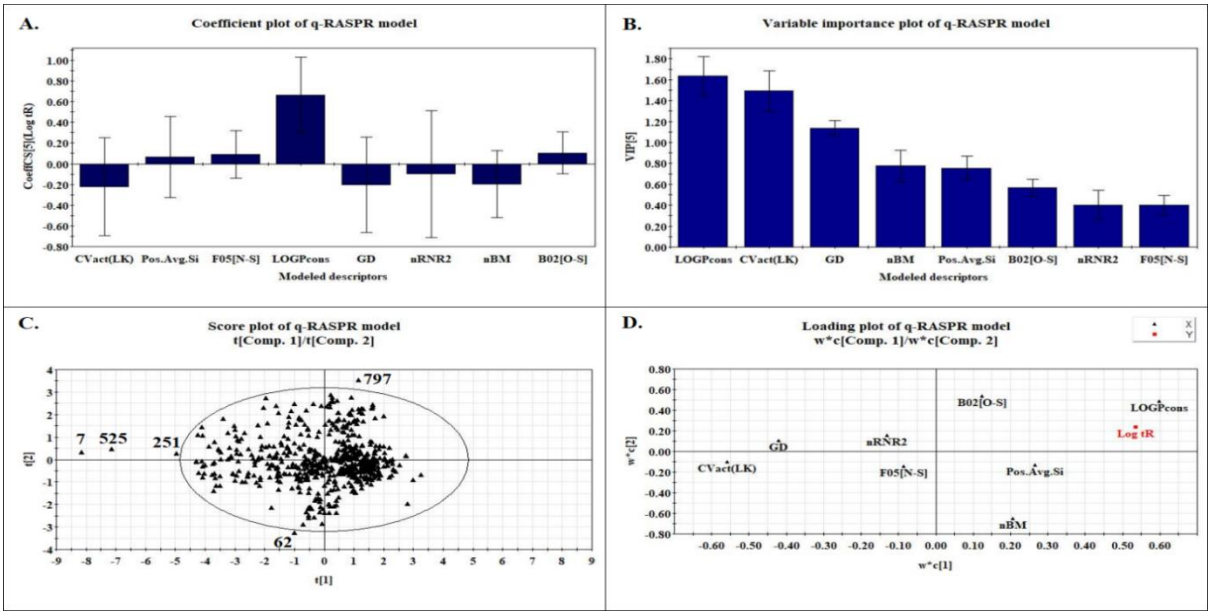


Figure 4.2. (**A**) Coefficient plot of q-RASPR model; (**B**) Variable importance plot (VIP) of q-RASPR model; (**C**) Score plot of q-RASPR model; (**D**) Loading plot of q-RASPR model

4.1.4. Insights on the modeled descriptors

4.1.4.1. LOGPcons descriptor

LOGPcons (computed octanol-water partition coefficient) is the most important descriptor of the final q-RASPR model as seen in the VIP, a molecular property descriptor which essentially transfers the knowledge of the hydrophobicity of molecules. As observed in the coefficient plot, LOGPcons is a positively contributing descriptor which has a linear correlation with the retention time endpoint (logt_R). In the data set, a compound like Brodifacoum (Compound **103**) with a large planer hydrophobic moiety in its structure has larger LOGPcons (LOGPcons = 7.3839) and higher retention time (logt_R = 10.02) than small hydrophilic compound like Cymoxanil (compound **203**; LOGPcons = -0.7992, logt_R = 4.83) as observed in **Figure 4.3**. Hydrophobicity of chemicals potentiates cellular toxicity due to the ease of passing lipid cell membrane (Vesterkvist et.al.,2012). Therefore, these observations justify the linear correlation between retention time (Log*tR*) and hydrophobicity, and thus to cellular toxicity also.

4.1.4.2. CVactivity(LK) descriptor

CVactivity(LK) is the 2nd most important and the 1st RASAR descriptor of this q-RASPR model that describes the coefficient of variation of the observed responses of selected close source compounds for each query compound. It is a negatively contributing descriptor of the developed q-RASPR model as shown in the coefficient plot (**Figure 4.2A**). It has been observed that the compound with less retention time value has more *CVactivity(LK)* (compound **525**; logt_R = 2.19, *CVactivity(LK)* = 0.6669), whereas, compound with more retention time has less *CVactivity(LK)* (compound **602**; logt_R = 10.36, *CVactivity(LK)* = 0.02569) value which justifies the negative contribution of *CVactivity(LK)* towards model endpoint prediction.

4.1.4.3. GD descriptor

In the final q-RASPR model, the third most contributing descriptor towards retention time (logt_R) is GD. It is a constitutional descriptor that is termed as graph density. It is mathematically derived by the following equation (1.19):

$$GD = 2nBo/nSK(nSK - 1) \qquad (1.19)$$

Here, *nBo* is the number of edges in the H-suppressed graph; *nSK* is the number of vertices in the H-suppressed graph (Yukawa, T et.al.,2020). We have shown two different compounds namely ethylene thiourea (compound no. **7**) and eprinomectin B1a (compound no. **289**) in **Figure 4.3**. Compound **289** contains higher number of vertices than compound **7** and thus *nSK (nSK-1)* portion of the equation has a greater value which decreases the resultant value of GD in compound **289**. As seen in the regression coefficient plot, GD has a negative contribution towards endpoint prediction. It is observed in the depicted compounds also; the retention time of compound **289** (GD = 0.034) is significantly higher (Log t_R = 10) than compound **7** (Log t_R = 1.08) with less GD value.

4.1.4.4. nBM descriptor

In the resultant model, the fourth most contributing descriptor towards retention time (log tR) is the constitutional descriptor nBM which indicates the number of multiple bonds present in a molecule. As shown by the regression coefficient plot, it has a negative impact on the endpoint value, which indicates a decrease in the retention time (Log t_R) value with an increase in the descriptor value and vice-versa. For instance, compound 194 in Figure 4.3 has

16 multiple bonds and less retention time ($logt_R$=8.00) than compound 805 with 12 multiple bonds ($logt_R$=9.16).

4.1.4.5. Pos.Avg.Sim descriptor

It is the 2nd RASAR descriptor of the q-RASPR model which measures the average similarity of positive close source compounds (with respect to the mean of the source compounds) for a query compound. It shows a positive regression coefficient in the model equation which indicates a linear relationship between ***Pos.Avg.Sim*** and modeled response, and can be justified by the observations in our dataset. Compound **802** has more retention time ($logt_R$ = 8.17) as well as *Pos.Avg.Sim* (0.7190) values, whereas, Compound **702** has less retention time ($logt_R$ = 4.48) and *Pos.Avg.Sim* (0.0) values.

4.1.4.6. B02[O-S] descriptor

B02[O-S] is a 2D atom pair descriptor which indicates the presence or absence of N-S atom pair at the topological distance 2. As shown in the regression coefficient plot, B02[O-S] has a positive regression coefficient i.e., it linearly correlates with the retention time ($logt_R$). For instance, compound **501** possesses O-S atom pair at the topological distance 2 in its structure as depicted in **Figure 4.3** and has a greater retention time ($logt_R$ = 7.99) than compound **16** ($logt_R$ = 3.17) which does not contain O-S atom pair at the topological distance 2.

4.1.4.7. nRNR2 descriptor

nRNR2 is a functional group count descriptor which indicates the number of tertiary aliphatic amine groups in a compound. It shows a lower VIP score and thus possesses a lower influence towards the endpoint prediction. The negative regression coefficient indicates that this descriptor is inversely correlated with the response values. If we compare two training compounds namely triamethoxam (compound **767**) and eprinomectin B1a (compound **289**) (**Figure 4.3**), it can be observed that higher tertiary amine group containing compound triamethoxam (nRNR2 = 2) has lower retention time ($logt_R$ = 3.43) than eprinomectin B1a which is devoid of tertiary amine (nRNR2 = 0; $logt_R$= 10). Therefore, these observations justify the inverse relationship between nRNR2 and the endpoint.

4.1.4.8. F05[N-S] descriptor

This is a 2D atom pair descriptor, indicating the frequency of N-S atom pairs at the topological distance 5, as illustrated in **Figure 4.3**. According to the regression coefficient plot, F05[N-S] linearly correlates with the retention time (logt_R), i.e., if the number or frequency of N-S at the topological distance 5 in a molecule increases, the response value (logt_R) will also increase for that molecule and vice-versa. For instance, compound **62** with a higher number of N-S atom pairs at the topological distance 5 (F05[N-S]=4) has a comparatively higher retention time (logt_R=6.84) than compound **822** with no N-S atom pair (F05[N-S]=0) at the topological distance 5, having low retention time (Log t_R= 6.68) (**Figure 4.3**). This implies that the presence of more fragments containing N-S atom pairs at the topological distance 5 increases the polar surface area as observed in **Figure 4.3**. Thus, polar surface area is an important factor to control lipophilicity as well as toxicity in terms of retention time (log t_R) as seen in the work of Yukawa and Naven (Yukawa, T et.al.,2020) which justifies our observations.

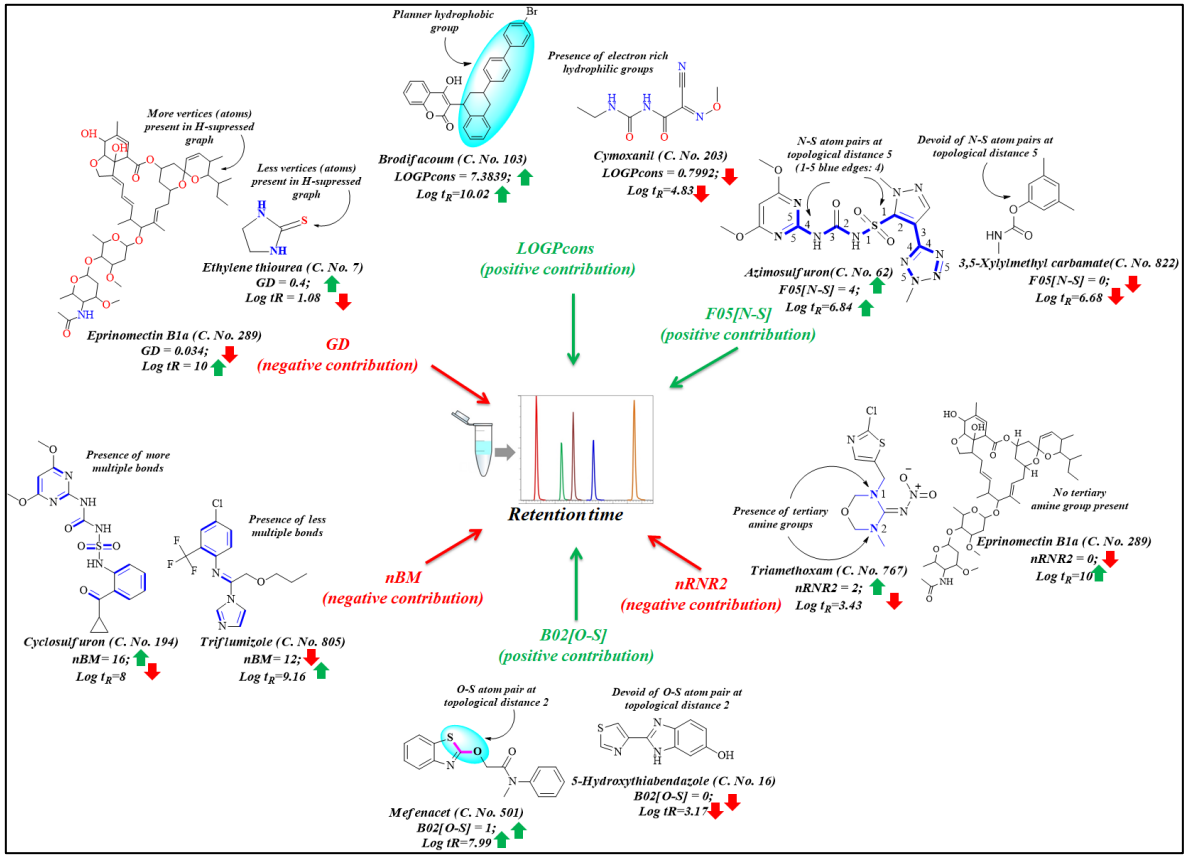


Figure 4.3. Insights of identified 0-2D descriptors used in the q-RASPR model

4.1.5. Applicability domain analysis of the PLS q-RASPR model

Applicability domain is the theoretical region in descriptor space in which the prediction obtained by the model is reliable (Gadaleta, D et.al.,2016). The model's Applicability domain (AD) was evaluated using the DModX (distance to model in X space) approach in the SIMCA-P software at the 99% confidence level with D-critical value 0.009999. The AD plots of training and test sets and the corresponding outliers have been shown in **Figure 4.4A and 4.4B** respectively. We found 2.59% of training data (26 outliers; compound no. **7, 21, 44, 62, 142, 202, 257, 282, 341, 392, 397, 435, 450, 457, 508, 557, 558, 570, 638, 715, 716, 724, 766, 767, 801, 809**) and 4.39% of test data (9 outliers; compound no. **19, 28, 340, 446, 447, 668, 714, 726, 774**)as outliers for the developed q-RASPR model. Considering the variety of molecules, the identification of any common structural traits for the outliers was difficult. However, from a statistical standpoint, we can approximate that the standard deviation of X-residuals is higher for outliers than for molecules lying within the applicability domain of the model.

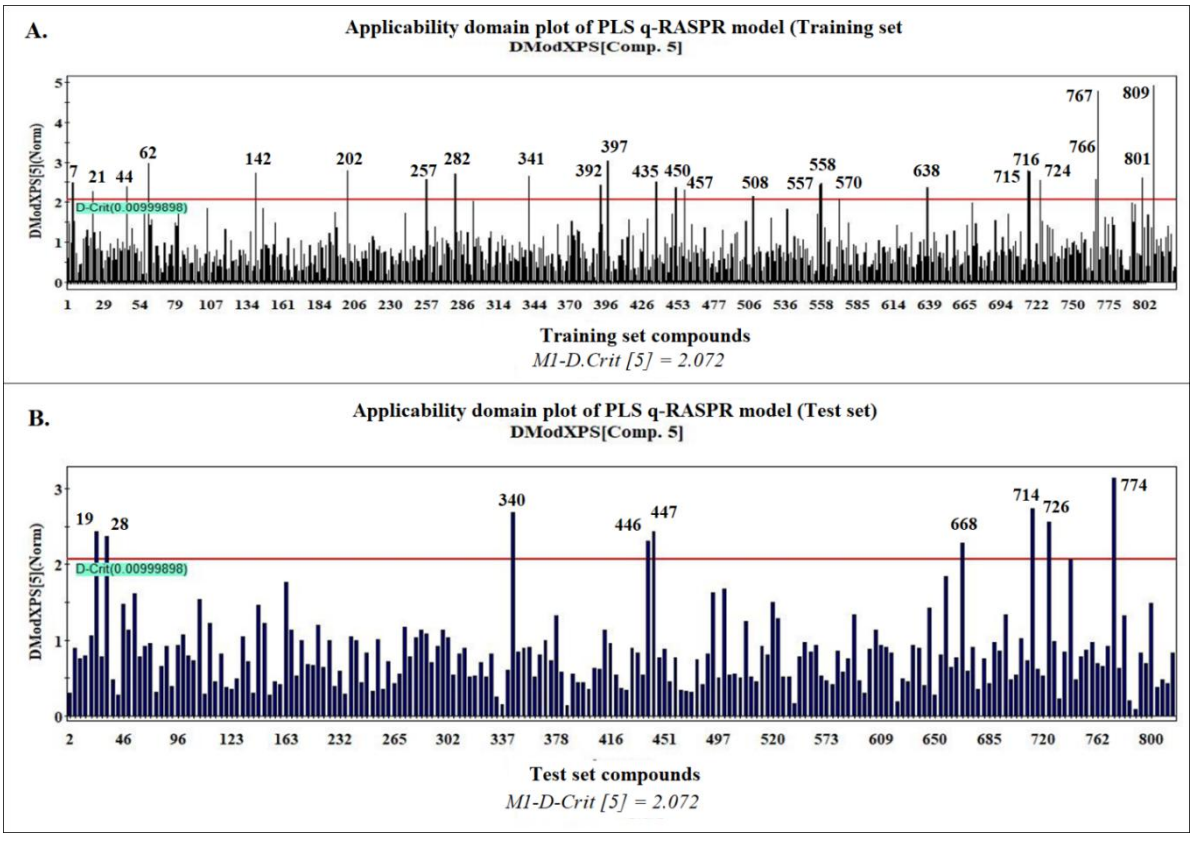


Figure 4.4. (A) Applicability domain (AD) plot of test set; (**B**) Applicability domain (AD) plot of training set

4.1.6. Comparison of developed QSPR and q-RASPR models with previously published work

In this present study, both QSPR and q-RASPR models have been developed for the prediction of retention time ($\log t_R$) of pesticide residues in the HPLC analysis. Here, we have extrapolated the knowledge of retention time towards the eco-toxicity potential of pesticide residues because of their proportional relationships with lipophilicity. The quality of q-RASPR model in terms of goodness of fit, robustness, and external predictivity is better than the QSPR model as suggested by the tabulated validation metrics (in **Table 4.3**) of both the models.

Table 4.3. Validation metrics of developed QSPR, q-RASPR, and previously reported MLR models

Work done by-	Type of models	No. of data points	Training set statistics		Test set statistics		No. of descriptors
			R^2	Q^2_{LOO}	Q^2_{F1}	MAE_{Test}	
Rojaset. al. (**2021**)	MLR	nTrain = 275, nVal = 275, nTest = 273	0.87	0.86	0.75	0.62	5
Present study	QSPR model (PLS)	nTrain = 618, nTest = 205	0.80	0.80	0.82	0.67	6 (5 LVs)
	q-RASPR model (PLS)	nTrain = 618, nTest = 205	0.82	0.81	**0.84**	0.63	8 (5 LVs)
	q-RASPR model (PLS)*	nTrain = (275+275) =550; nTest = 273	0.82	0.81	**0.82**	**0.60**	8 (5 LVs)

*q-RASPR developed with training set (training + validation compounds of previous work by Wang, et al.), LVs = latent variables

Previously, Rojas *et.al 2021* reported a food-informatics model where they took 275 data points as the training and validation sets each, and 273 data points as the test set. They employed five conformation-independent descriptors to correlate with the retention time endpoint using MLR algorithm. The reported internal and external validation metrics of the previous model have been tabulated in **Table 4.3**. The previous model was developed with a remarkably lower number of training set compounds (nTrain = 275) in comparison to the present q-RASPR model (nTrain = 618) and thus, the previous model had somewhat better R^2 (0.87) and Q^2_{LOO} (0.86) values. However, both the QSPR and q-RASPR model shave better external predictivity in terms of Q^2_{F1} (QSPR = 0.82, q-RASPR = 0.84) than the previous MLR (Q^2_{F1} of previous MLR = 0.75). For a proper comparison of external predictivity between the two works, we have repeated the q-RASPR by taking the combination of the reported training and validation set compounds (from the work by Rojas et al 2021 as the training set. The same set of descriptors and the same number of LVs (as used in the developed q-RASPR model) have been used for the prediction of retention time ($\log t_R$) of test set compounds (same as previous work) by the q-RASPR model. It has been observed from the computed Q^2_{F1} (q-RASPR* = 0.82, previous MLR = 0.75) and MAE_{Test} (q-RASPR* = 0.60, previous MLR = 0.62) values in **Table 4.3**, that the external predictivity and error of predictions of the q-RASPR model have held a better place than the previous model.

4.2. Quantitative Read-Across Structure-Activity Relationship (q-RASAR): A New Approach Methodology to Model Aquatic Toxicity of Organic Pesticides Against Three Different Fish Species

4.2.1. Developed QSAR models for aquatic toxicity (Log 1/ LC50) endpoints

Three different QSAR models for RT, LP and Miscellaneous datasets have been developed in the present study by using PLS regression with five, three, and four latent variables (LVs) respectively. Three sets of descriptors (eight descriptors for each dataset) have been identified by best subset selection in this study and have been employed to model the toxicity endpoint (Log (1/LC50)) of RT, LP, and Miscellaneous datasets respectively. We have defined the modeled descriptors along with their class in **Table 4.5**. The developed models for RT, LP, and Miscellaneous datasets have been reported in the following equations (**1**), (**2**), and (**3**) respectively:

For the RT dataset:

$$Log(1/LC50) = 3.89 + 0.35 \times \text{LOGPcons} - 0.012 \times \text{SAdon} + 0.026 \times \text{D/Dtr03} + 0.36 \times \text{nHM} + 0.24 \times \text{minddssS} + 0.17 \times \text{F06[O} - \text{O]} - 0.33 \times \text{Cl} - 089 - 0.30 \times \text{MaxsssN} \quad (1)$$

For the LP dataset:

$$Log(1/LC50) = 1.90 - 0.29 \times \text{F07[C} - \text{Cl]} + 1.12 \times \text{F03[F} - \text{Cl]} - 0.16 \times \text{F06[C} - \text{F]} + 0.55 \times \text{H} - 048 - 0.82 \times \text{ESOL} - 0.28 \times \text{nArX} - 0.62 \times \text{F09[N} - \text{O]} + 0.01 \times \text{TPSA(Tot)} \quad (2)$$

For the Miscellaneous dataset:

$$Log(1/LC50) = 2.17 - 0.52 \times \text{ESOL} + 0.72 \times \text{C} - 003 + 0.40 \times \text{F02[O} - \text{S]} + 0.88 \times \text{nCXr} = +0.38 \times \text{F05[N} - \text{O]} - 0.25 \times \text{F05[C} - \text{S]} - 0.46 \times \text{F09[O} - \text{Cl]} + 0.43 \times \text{B05[C} - \text{C]} \quad (3)$$

Strict internal and external validation has been carried out in this study following the OECD principle 4 to adjudge the models' performance; the computed validation metrics have been tabulated in the following **Table 4.4**. The computed R^2 and/or R^2_{Adj} values of the studied datasets couldn't cross the 0.6 threshold mark except for LP. Although R^2 and R^2_{Adj} are not the major criteria for models' goodness of fit. We have computed the cross-validated correlation coefficient (Q^2_{LOO}) and it crossed the 0.5 threshold mark for the three models. The external correlation coefficients' (Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}) values of these three PLS models are also in the acceptable range which suggests the good external predictability of the preliminary QSAR models.

Table 4.4. Computed Validation Metrics of developed PLS QSAR models for all the datasets

Datasets	Internal Validation Metrics				External Validation Metrics			
	R^2	R^2_{adj}	Q^2_{Loo}	MAE_{train}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	MAE_{test}
RT	0.53	0.51	0.51	0.79	0.55	0.54	0.52	0.75
LP	0.68	0.65	0.62	0.76	0.70	0.66	0.64	0.68
Miscellaneous	0.60	0.58	0.54	0.79	0.51	0.50	0.50	0.77

Table 4.5. Identified Structural Descriptors of developed PLS QSAR models for all the datasets:

Datasets	Descriptors	Class	Definition
RT dataset	LOGPcons	molecular properties	It indicates octanol-water partition coefficient (log P)
	SAdon	molecular properties	It indicates surface area of donor atoms from P_VSA like descriptors
	D/Dtr03	ring descriptors	It indicates distance/detour ring index of order 3
	nHM	constitutional indices	It indicates the number of heavy atoms
	minddssS	atom type E-state indices	It indicates minimum ddssS
	F06[O-O]	2D atom pairs	It indicates frequency of O-O at topological distance 6
	Cl-089	atom centered fragments	It indicates Cl attached to C1(sp2)
	MaxsssN	atom type E-state indices	It indicates maximum sssN
LP dataset	F07[C-Cl]	2D atom pairs	It indicates frequency of C-Cl at topological distance 7
	F03[F-Cl]	2D atom pairs	It indicates frequency of F-Cl at topological

			distance 3
	F06[C-F]	2D atom pairs	It indicates frequency of C-F at topological distance 6
	H-048	atom centered fragments	It indicates H attached to C2(sp3)/C1(sp2)/C0(sp)
	ESOL	molecular properties	It indicates estimated solubility (logS) for aqueous solubility using LOGPcons
	nArX	Functional group counts	It indicates number of X on aromatic ring
	F09[N-O]	2D atom pairs	It indicates frequency of N-O at topological distance 9
	TPSA(Tot)	molecular properties	It indicates topological polar surface area using N,O,S,P polar contributions
Miscellaneous dataset	ESOL	molecular properties	It indicates estimated solubility (logS) for aqueous solubility using LOGPcons
	C-003	atom centered fragments	It indicates CHR3 fragments
	F02[O-S]	2D atom pairs	It indicates frequency of O-S at topological

		distance 2
nCXr=	Functional group counts	It indicates number of X on ring C(sp2)
F05[N-O]	2D atom pairs	It indicates frequency of N-O at topological distance 5
F05[C-S]	2D atom pairs	It indicates frequency of C-S at topological distance 5
F09[O-Cl]	2D atom pairs	It indicates frequency of O-Cl at topological distance 5

4.2.2. Developed q-RASAR models of aquatic toxicity (Log 1/ LC50) endpoint

We aimed to develop the q-RASAR models for improving the external predictability of the associated QSAR models further in this study for all the datasets. Thus, we have computed the read-across-derived similarity, error, concordance, and prediction functions that are collectively called the RASAR descriptors from the identified structural features of each dataset. Final feature selections for RASAR model development have been carried out from the combined pool of RASAR descriptors and the corresponding 0D-2D descriptors using the best subset selection and the best combination of features have been identified based on the cross-validated correlation coefficient values (Q^2_{LOO}). Finally, the PLS regression has been employed with the optimum number of LVs (RT dataset = 4 LVs; LP dataset = 3 LVs; and Miscellaneous dataset = 1 LV) to get the desired q-RASAR models. The developed PLS q-RASAR models obtained from RT, LP and miscellaneous datasets have been represented mathematically in the following equations (4), (5) and (6) respectively.

For RT dataset:

$$Log(1/LC50) = 2.27 + 0.53 \times \text{RA function(ED)} - 0.61 \times \text{Pos. Avg. Sim} + 0.23 \times \text{LOGPcons} - 0.007 \times \text{SAdon} + 0.18 \times \text{nHM} + 0.15 \times \text{minddssS} - 0.23 \times \text{Cl} - 089 - 0.21 \times \text{MaxsssN} \quad (4)$$

For LP dataset:

$$Log(1/LC50) = 1.46 + 2.49 \times \text{SD similarity(GK)} - 0.39 \times \text{F07[C} - \text{Cl]} + 1.08 \times \text{F03[F} - \text{Cl]} - 0.20 \times \text{F06[C} - \text{F]} + 0.73 \times \text{H} - 048 - 0.80 \times \text{ESOL} - 0.61 \times \text{F09[N} - \text{O]} + 0.011 \times \text{TPSA(Tot)} \quad (5)$$

For other's dataset:

$$Log(1/LC50) = 0.32 + 0.49 \times \text{RA function(LK)} + 0.94 \times \text{SD Activity(LK)} - 0.30 \times \text{ESOL} + 0.12 \times \text{F02[O} - \text{S]} + 0.54 \times \text{nCXr} \quad (6)$$

The final q-RASAR models have been rigorously validated; the internal and external validation metrics have been computed and tabulated in the following **Table 4.6**. The Q^2_{LOO} values of all three PLS-based q-RASAR models (RT dataset = 0.50; LP dataset = 0.60; Miscellaneous = 0.52) have crossed the 0.5 threshold mark and therefore, the models are acceptable in terms of goodness of fit and robustness. The acceptable values of Q^2_{F1} also suggest sufficient external predictability of the developed q-RASAR models. Apart from that, we have performed the Y-randomization of each model in SIMCA-P software by 100 times random shuffling of response values (Log 1/LC50) (Rücker et al., 2007) and fitting the data again with the same descriptors of corresponding q-RASAR models. The original R^2 and Q^2_{LOO} values are plotted against the computed R^2 and Q^2 values of randomly developed models and finally, the intercepts (R^2_Y and Q^2_Y intercept) are noted. The Y-randomization plots of q-RASAR models have been given in Fig. S2 of Supplementary Material 1. The negative intercept values of the Y-randomization plots (R_Y^2intercept = -0.036, Q_Y^2intercept = -0.0839 for RT dataset; R_Y^2 intercept = 0.032, Q_Y^2intercept = -0.324 for LP dataset; and R_Y^2intercept = -0.005, Q_Y^2intercept = -0.05 for Miscellaneous dataset) confirm the models as not the result of any chance correlations. The observed vs. predicted scatter plots have also been prepared with the training and test set data and have been shown in **Figure 4.7A, 4.7C, 4.7E** of RT, LP, and Miscellaneous datasets respectively. A uniform scattering and equal

distribution of training and test data have been observed around the line passed through the origin in all the scatter plots that demonstrate the well fit of data and minimum residuals.

Table 4.6. Computed validation metrics of developed q-RASAR models for all the datasets

Datasets	Internal Validation Metrics				External Validation Metrics			
	R^2	R^2_{adj}	Q^2_{Loo}	MAE_{train}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	MAE_{test}
RT	0.52	0.51	0.50	0.79	0.58	0.57	0.52	0.71
LP	0.67	0.65	0.60	0.76	0.74	0.70	0.67	0.64
Miscellaneous	0.54	0.59	0.52	0.79	0.59	0.57	0.50	0.78

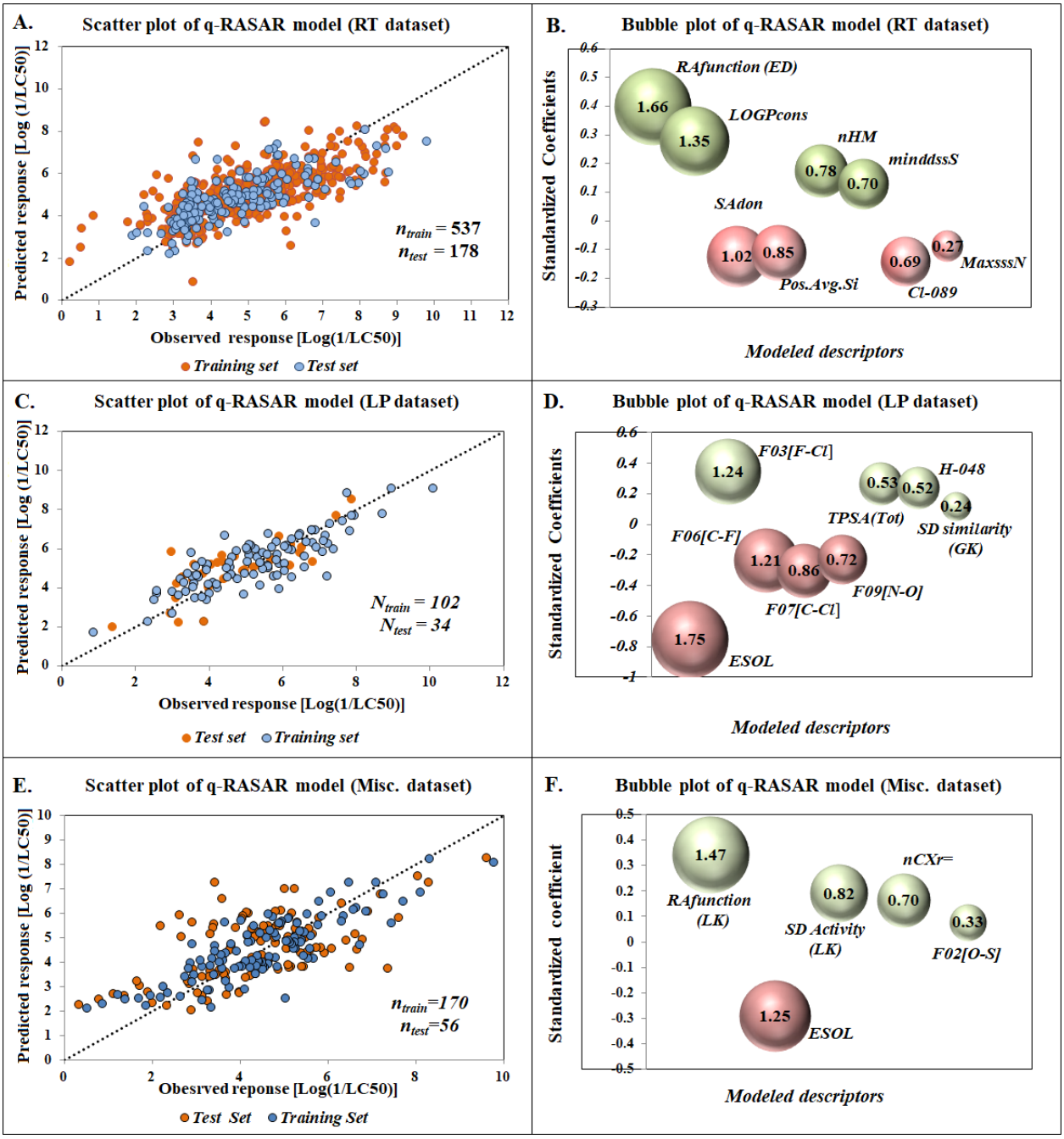


Fig 4.7. (A) Scatter plot of RT dataset**, (C)** Scatter plot of LP dataset**, (E)** Scatter plot of Miscellaneous dataset**, (B)** Bubble plot of q-RASAR model (RT dataset)**, (D)** Bubble plot of q-RASAR model of LP dataset**, (F)** Bubble plot of q-RASAR model of (Misc.: Miscellaneous dataset)

4.2.3. Insights on the Modeled Descriptors of the RT dataset

The final PLS-based q-RASAR model of RT dataset has been developed with two RASAR descriptors (***RAfunction (ED)***, ***and Pos. Avg. Sim***.) and six 0D-2D descriptors (LOGPcons,

SAdon, nHM, minddssS, Cl-089, and MaxsssN). To know the type of contribution of the descriptors and their relative importance, we made use of coefficient plots and variable importance plots in this study which is obtained from the SIMCA-P software. We have combined those two plots and represented them as a single bubble plot in **Figure 4.7B,** where the bubble radius indicates the VIP score, and the bubble colour indicates the type of contributions (green bubble = positive contribution, red bubble = negative contribution). Each bubble indicates a specific descriptor and the bubbles are arranged in descending order of importance from the left-hand side to the right-hand side of the X-axis. According to this plot, the order of importance of the modeled descriptors is – ***RAfunction (ED)***>LOGPcons>SAdon>Pos. Avg. Sim.>nHM>minddssS> Cl-089 >MaxsssN, and the first three descriptors namely ***RAfunction (ED)***, LOGPcons, SAdon are the most important descriptors towards toxicity with VIP score >1. An equal distribution between positive contributing descriptors and negative contributing descriptors has been observed in this model. RAfunction, LOGPcons, nHM, and minddssS have positive contributions whereas, the rest ***(Pos. Avg. Sim.,*** SAdon, Cl-089, and MaxsssN) have negative contributions towards the toxicity of Rainbow Trout (RT).

4.2.3.1. RA function (ED)

The *RAfunction (ED)* is the most important and positive contributing descriptor of the PLS-based q-RASAR model developed from RT dataset. It is a Read-across-derived prediction function obtained by the tool "RASAR-Desc-Calc-v2.0" employing the Euclidean distance-based similarity estimation algorithm. It contains information from all of the structural descriptors and process variables and functions similarly as a composite variable (Banerjee and Roy, 2022). The positive determination coefficient of it indicates a proportional relationship between this descriptor and toxic endpoint. The comparative assessment between different data points also suggests the same. For instance, a more toxic compound Cyphenothrin (**compound 4:** Log (1/LC50) = 9.04) has a greater RA function (ED) value (8.24) than a comparatively less toxic compound Famoxadone (**compound 37:** Log (1/LC50) = 7.53; RAfunction ED = 4.59).

4.2.3.2. LOGPcons

LOGPcons (computed octanol-water partition coefficient) is the second most important descriptor of the developed q-RASAR model of RT dataset as observed in the bubble plot (**Fig. 4.7B**). As a molecular property descriptor, it essentially provides the knowledge of the hydrophobicity of organic molecules. It seems to be a positive contributing descriptor of the model which suggests the linear correlation of it towards the toxic endpoint Log (1/LC50) in RT fish species. This linear relationship has been observed in the studied data points also. For example, bifenthrin (**Compound 2**), having a larger planar hydrophobic moiety in its structure possess higher LOGPcons value (LOGPcons=6.14) and thus more aquatic toxicity Log (1/LC50)=9.21 than the smaller hydrophilic compound cyphenothrin (**Compound 4:** Log (1/LC60) = 9.04) with less LOGPcons value (LOGPcons=5.08) The hydrophobicity of organic chemicals is a key factor for aquatic toxicity because it can drive the permeability of organic chemicals through lipophilic cell membrane into aquatic organisms and thus generation of toxic response (Mayer and Reichenberg, 2006; Trac et al., 2018)

4.2.3.3. SAdon

In the final q-RASAR model of RT dataset, the third most contributing descriptor towards the aquatic toxicity Log (1/LC50) endpoint is SAdon. As a molecular property descriptor, it indicates surface area of donor atoms(from P_VSA-like descriptor)(Todeschini and Consonni, 2000). It possesses a negative contribution towards the endpoint Log (1/LC50)) as observed in the bubble plot **(Figure: 4.7B)**. In this dataset, we have shown two different training compounds namely difenthiuron **(Compound 8)** and dinoterb **(Compound 26**) in **Figure: 4.8** for comparing the toxic character and toxicity. Due to the presence of the hydroxyl group (can act as a strong electron donor) (Li et al., 2020; Roberts and Marshall, 1995), dinoterb (**Compound 26**: SAdon = 42.68) has a greater surface area of doner atom than difenthiuron (**Compound 8**: SAdon = 36.02).Toxicity profiles of these two compounds have shown an exactly opposite trend from the descriptor Log (1/LC50): **Compound 8:** = 8.74; **Compound 26:** = 7.85) that approves the negative contribution of SAdon towards toxicity.

4.2.3.4. Pos.Avg.Sim.

It is another RASAR descriptor of the PLS-based q-RASAR model of RT dataset, which measures the average similarity of positive close source compounds (with respect to the mean of the source compounds) for a query compound. It shows a negative regression coefficient in the model equation, which indicates a decrease in the endpoint Log (1/LC50) value with an increase in the descriptor value and vice-versa. The same trend has been followed by many studied data points also. A comparatively more toxic compound Bifenthrin (**Compound 2**: Log (1/LC50) = 9.21) has less *Pos.Avg.Sim* (*Pos.Avg.Sim* = 0.86) value than a less toxic Cyfluthrin (**Compound 5**: Log (1/LC50)= 8.96; *Pos. Avg. Sim.* = 0.91).

4.2.3.5. nHM

The fifth most contributing descriptor of this model is the constitutional descriptor nHM, which indicates the number of heavy atoms present in a molecule. As observed in the bubble plot **(Figure 4.7B)**, it has a positive impact on the endpoint value that indicates a linear correlation between descriptors value and aquatic toxicity Log (1/LC50). For instance, more heavy atom containing Chlorpyrifos (**Compound 14**) (**nHM=5**) has higher toxicity (**Log (1/LC50) = 8.43**) than less heavy atom containing Dichlofenthion (**Compound 13:** nHM = 4; Log 1/LC50 = 8.40). We have structurally depicted the presence of heavy atoms in **Figure 4.8**.

4.2.3.6. minddssS

The minddssS is an atom-type E-state indices descriptor that identifies the presence of the sulfur atoms surrounded by two double bonds and two single bonds in a compound and also computes the minimum electronic state of the surrounding. It possesses a lower VIP score and thus shows a lower influence towards the end-point prediction. The positive regression coefficient as seen in the bubble plot (**Figure 4.7B**) indicates a linear correlation with the response values. If we compare two training compounds namelybensultap (**Compound 187**) and flusulfamide (**Compound 218**) as depicted in **Figure 4.8** , then it is observed that the toxicity increases from Bensultap to Flusulfamide Log (1/LC50):Compound **218 = 5.55; Compound 187 = 5.75)** with simultaneous increase in minddssS value (minddssS:

Compound 218= -3.49; **Compound 187**= -3.46). Therefore, these observations justify the linear relationship between minddssS and aquatic toxicity in this dataset.

4.2.3.7. Cl-089

It is an atom-centered fragment descriptor that denotes the number of Cl atoms attached to sp^2 hybridized carbon in the chemical structure of the compound as shown in **Figure: 4.8**. It shows a relatively lower VIP score and therefore, it possesses less importance towards the studied endpoint. It is a negative a negative contributing descriptor which indicates a decrease in the aquatic toxicity Log (1/LC50) with an increase in the descriptor value and vice versa. Compounds like pyridaben (**Compound 7**) having one Cl-atom attached to sp^2 hybridized carbon in its structure **(**Cl-089 **=** 1**)** show more aquatic toxicity (Log (1/LC50)=8.71 than Dichlofenthion (Compound **13:** Log (1/LC50) = 8.40) which contains two Cl attached to sp^2 hybridized carbon (Cl-089 = 2).

4.2.3.8. MaxsssN

MaxsssN is the least contributing descriptor of the q-RASAR model developed from the RT dataset. It is an atom-type E-state indices descriptor that computes the maximum E-state of a Nitrogen atom surrounded by three single bonds (-sssN) in a compound. It has a negative regression coefficient that indicates an inversely proportional relationship between toxicity and the mentioned descriptor. Studied compound like cyfluthrin (**Compound 5**) has no –sssN fragment in its structure (**MaxsssN=0**) but shows higher toxicity Log (1/LC50 = 8.96); on the contrary, a compound namely pyridaben (**Compound 7**) containing –sssN fragment (MaxsssN = 1.42) has less aquatic toxicity Log (1/LC50) = 8.71 to RT fish species as depicted in (**Figure 4.8**).

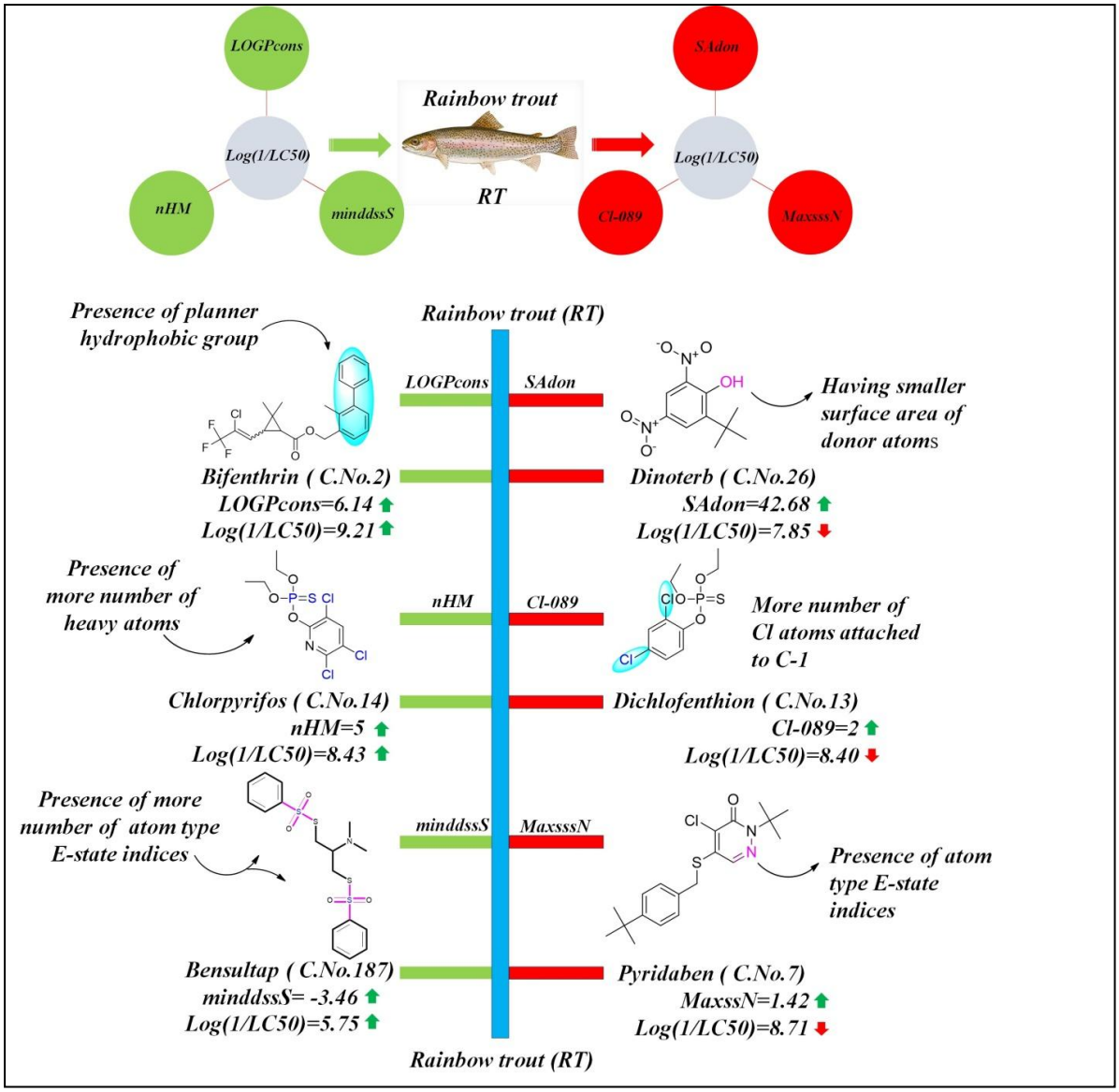


Fig 4.8. Insights of the modeled descriptors of RT (Rainbow trout) dataset

4.2.4. Insights on the Modeled Descriptors of the LP dataset

The final PLS-based q-RASAR model of LP dataset has been developed with one RASAR descriptor ***SD Similarity(GK)*** and seven 0D-2D descriptors (ESOL, F06[C-F],F03[F-Cl],F07[C-Cl],F09[N-O],TPSA(Tot),H-048). To know the type of contribution of the descriptors and their relative importance, we made use of coefficient plots and variable importance plots in this study which is obtained from the SIMCA-P software. We have combined those two plots and represented them as a single bubble plot in **Fig 4.7D**, where the bubble radius indicates the VIP score, and the bubble colour indicates the type of contributions (green bubble = positive contribution, red bubble = negative contribution). Each

bubble indicates a specific descriptor and the bubbles are arranged in descending order of importance from the left-hand side to the right-hand side of the X-axis. According to this plot, the order of importance of the modeled descriptors is – ESOL> F06[C-F]> F03[F-Cl]> F07[C-Cl]> F09[N-O]> TPSA(Tot),>H-048> ***SD Similarity(GK)***, and the first three descriptors namely ESOL> F06[C-F] > F03[F-Cl] are the most important descriptors towards toxicity with VIP score >1. F03[F-Cl], H-048, TPSA(Tot)have positive contribution whereas, the rest (ESOL, F06[C-F], F07[C-Cl], F09[N-O], ***SD Similarity(GK)***, have negative contribution towards toxicity of Lepomis (LP).

4.2.4.1. ESOL

According to the bubble plot (**Figure 4.7D**), it is the most contributing descriptor towards the response value (Log 1/LC50) of the resultant q-RASAR model of LP dataset. As a molecular property descriptor, it essentially transfers the knowledge of estimated solubility (logS) for aqueous solubility assessment using LOGPcons descriptor as well as indicating the hydrophobicity of organic molecules for determining toxic potency to aquatic species(Cronin, 2006). It seems to be a negative contributing descriptor of the model which suggests the inverse correlation of it towards the toxic endpoint Log (1/LC50) in LP fish species. This inverse relationship has been observed in other studied data points also of this dataset. For example, Cyhalothrin (**Compound 1**) has a larger hydrophobic moiety with less ESOL value (ESOL=-6.05) and thus causes more aquatic toxicity Log (1/LC50) = 8.99 than a smaller hydrophilic compound Fluazinam **(Compound 16:** Log (1/LC50) = 6.92 shows more ESOL value (ESOL=-6.60) as depicted in (**Figure 4.9)**. Therefore, it can be justified that hydrophobicity of organic molecules may be linearly correlated with aquatic toxicity in the form of Log (1/LC50) values (Stibany et al., 2017).

4.2.4.2. F06[C-F]

F06[C-F] is the second most important descriptor towards the end point prediction in the developed final q-RASAR model of LP dataset. It is a 2D atom pair descriptor indicating the frequency of C-F atom pairs at topological distance 6. As per the bubble plot **(Figure: 4.7D)**, it shows negative impact on the endpoint value for aquatic toxicity Log (1/LC50) prediction. Therefore, this 2D atom pair descriptor is inversely correlated with the endpoint (Log 1/LC50) of aquatic toxicity in this dataset, i.e., Triflubenzuron **(Compound 5)** with a less number of [C-F] atom pairs (F06[C-F]=2) at topological distance 6 shows comparatively

more aquatic toxicity Log (1/LC50)=7.76 than flumetralin (**Compound 10:** Log (1/LC50)=7.26) with descriptor value (F06[C-F]=3) and vice-versa. Therefore, we can justify that descriptor value is also related to molecular size of a chemical, that influences hydrophobicity of a molecule in the present dataset for aquatic toxicity in terms of Log (1/LC50) values (Monserud and Schwartz, 2012).

4.2.4.3. F03[F-Cl]

In the resultant q-RASAR model of LP dataset, F03[F-Cl] is the third most important descriptor towards the aquatic toxicity Log (1/LC50) prediction as observed in bubble plot (**Figure: 4.7D**). As a 2D atom pair descriptor, it indicates the frequency of F-Cl atom pairs at topological distance 3, as shown in **(Figure: 4.9)**. It has positive contribution towards the endpoint, which indicates linear correlation between the descriptor value and aquatic toxicity Log (1/LC50). Therefore, we can justify that (**compound 1**:[cyano-(3-phenoxyphenyl)methyl] 3-[(*Z*)-2-chloro-3,3,3-trifluoroprop-1-enyl]-2,2-dimethylcyclopropane-1-carboxylate) containing more number of [F-Cl] fragments (F03[F-Cl]=3) at topological distance 3, shows relatively higher aquatic toxicity Log(1/LC50)=7.76 than (**compound 6:**(1*S*,4*S*)-1,2,2,3,3,4,7-heptachloro-5,5-dimethyl-6-methylidenebicyclo[2.2.1]heptane and Log(1/LC50)=7.65 with no descriptor value (F03[F-Cl]=0**)** in this dataset. This implies that the presence of more [F-Cl] fragments in a molecule at topological distance 3, may decrease the polar surface area but increase the hydrophobicity of a molecule, as polar surface area is an important factor to control hydrophobicity for toxicity determination on aquatic species in terms of Log (1/LC50) values (Smith et al., 2010).

4.2.4.4. F07[C-Cl]

It is the fourth contributing 2D atom pair descriptor as shown in the bubble plot, which indicates the frequency of C-Cl atom pairs at topological distance 7. As seen in the bubble plot (**Figure: 4.7D**), it shows negative influence towards the response value, which indicates an inverse correlation with the endpoint Log (1/LC50) value for aquatic toxicity prediction to LP fish species. For example, Aldrin **(Compound 4)** with the devoid of[C-Cl] fragments (F07[C-Cl] = 0) at topological distance 7 causes more aquatic toxicity Log(1/LC50)=7.89 than Triflubenzuron **(Compound 5**: Log(1/LC50)=7.76) with this kind of fragments (F07[C-Cl]=2) at topological distance 7. In this dataset, descriptor value is also related to the

molecular size of a chemical, which may influence polar surface area as well as hydrophobicity of organic molecules for toxicity estimation on aquatic species (Mannhold, 2007).

4.2.4.5. F09[N-O]

F09[N-O] is a 2D atom pair descriptor, indicating the frequency of N-O atom pairs at topological distance 9, as illustrated in (**Figure: 4.7D**). It shows a moderate influence on the endpoint value for aquatic toxicity Log (1/LC50) prediction as seen in the bubble plot. It shows a negative influence on the endpoint prediction, which indicates a decrease in the aquatic toxicity Log (1/LC50) value with an increase in the descriptor value and vice versa. For instance, (**compound 89:**1-(2-fluorophenyl)-1-(4-fluorophenyl)-2-(1,2,4-triazol-1-yl)ethanol) having no atom pairs (F09[N-O]=0) at topological distance 9 causes more aquatic toxicity Log(1/LC50)=3.61 than (**Compound 95**:2-[(*Z*)-*N*-[(*E*)-3-chloroprop-2-enoxy]-*C*-ethylcarbonimidoyl]-3-hydroxy-5-(oxan-4-yl)cyclohex-2-en-1-one with Log(1/LC50)=3.58 due to the presence of [N-O] atom pairs at topological distance 9(F09[N-O]=2) as seen in (**Figure: 4.9**).

4.2.4.6. TPSA(Tot)

As a molecular property descriptor, it indicates topological polar surface area by using polar contributions of (N, O, S, P) atoms for prediction of aquatic toxicity in the form of Log (1/LC50) values. As observed in the bubble plot, it shows relatively lower VIP score and thus, possesses the least influence on the endpoint value. The positive regression coefficient indicates that this descriptor is linearly correlated with the response values as exemplified by Carbosulfun (**Compound 8)** having higher topological polar surface area (TPSA(Tot)=67.31) causes more aquatic toxicity Log (1/LC50)=7.40 than Allothrin (**Compound 12:** Log (1/LC50)=7.10) with small topological polar surface area (TPSA(Tot)=43.37) as depicted in (**Figure: 4.9**). In the present study, it can be justified that larger topological polar surface area of a molecule shows more hydrophobicity as well as more toxicity toward aquatic species because topological polar surface area is a key factor to control hydrophobicity in toxicity studies toward aquatic species in terms of Log (1/LC50) values (Prasanna and Doerksen, 2009).

4.2.4.7. H-048

It is an atom-centered fragment type of descriptor, indicating the number of H atoms attached to C2(sp3)/C1(sp2)/C0(sp) atoms of a molecule for aquatic toxicity prediction in the form of Log (1/LC50) values. It shows least influence for the model endpoint prediction as per the bubble plot. It shows positive influence on the end point value, which indicates if a (compound **36:***N*-(butoxymethyl)-2-chloro-*N*-(2,6-diethylphenyl)acetamide) having higher number of H atoms attached with C2(sp3)/C1(sp2)/C0(sp) atoms of a molecule (H-048=2), occurs comparatively more aquatic toxicity Log(1/LC50)=5.85 to LP fish species than the (**Compound 64:**(4*S*)-5-(4-chlorophenyl)-*N*-cyclohexyl-4-methyl-2-oxo-1,3-thiazolidine-3-carboxamide and Log (1/LC50)=4.60 with relatively less descriptor value (H-048=1) as seen in (**Figure: 4.9**). Therefore, we can justify that the compound containing large number of H atoms attached with C2(sp3)/C1(sp2)/C0(sp) atoms of a molecule represents more hydrophobicity as well as more toxicity toward aquatic species (Lambert et al., 2022).

4.2.4.8. SD Similarity (GK)

In the resultant q-RASAR model of LP dataset, it is the first RASAR descriptor indicating standard deviation of the similarity values of the selected close source compounds for each query compound. It shows a positive regression coefficient in the model equation, which indicates a linear relationship between *SD Similarity (GK)* and the modelled response and this can be justified in other studied data points also. For instance, (**Compound 1:** [cyano-(3-phenoxyphenyl)methyl] 3-[(*Z*)-2-chloro-3,3,3-trifluoroprop-1-enyl]-2,2-dimethylcyclopropane-1-carboxylate) has more aquatic toxicity Log (1/LC50)=8.99) as well as higher SD Similarity(GK) values (0.31), whereas., (**Compound 2**:[(*R*)-cyano-(3-phenoxyphenyl)methyl] (2*S*)-2-[2-chloro-4-(trifluoromethyl)anilino]-3-methylbutanoate**)**comparatively shows less aquatic toxicity Log (1/LC50)=8.74) with smaller SD Similarity(GK) values (**0.07**) in this dataset.

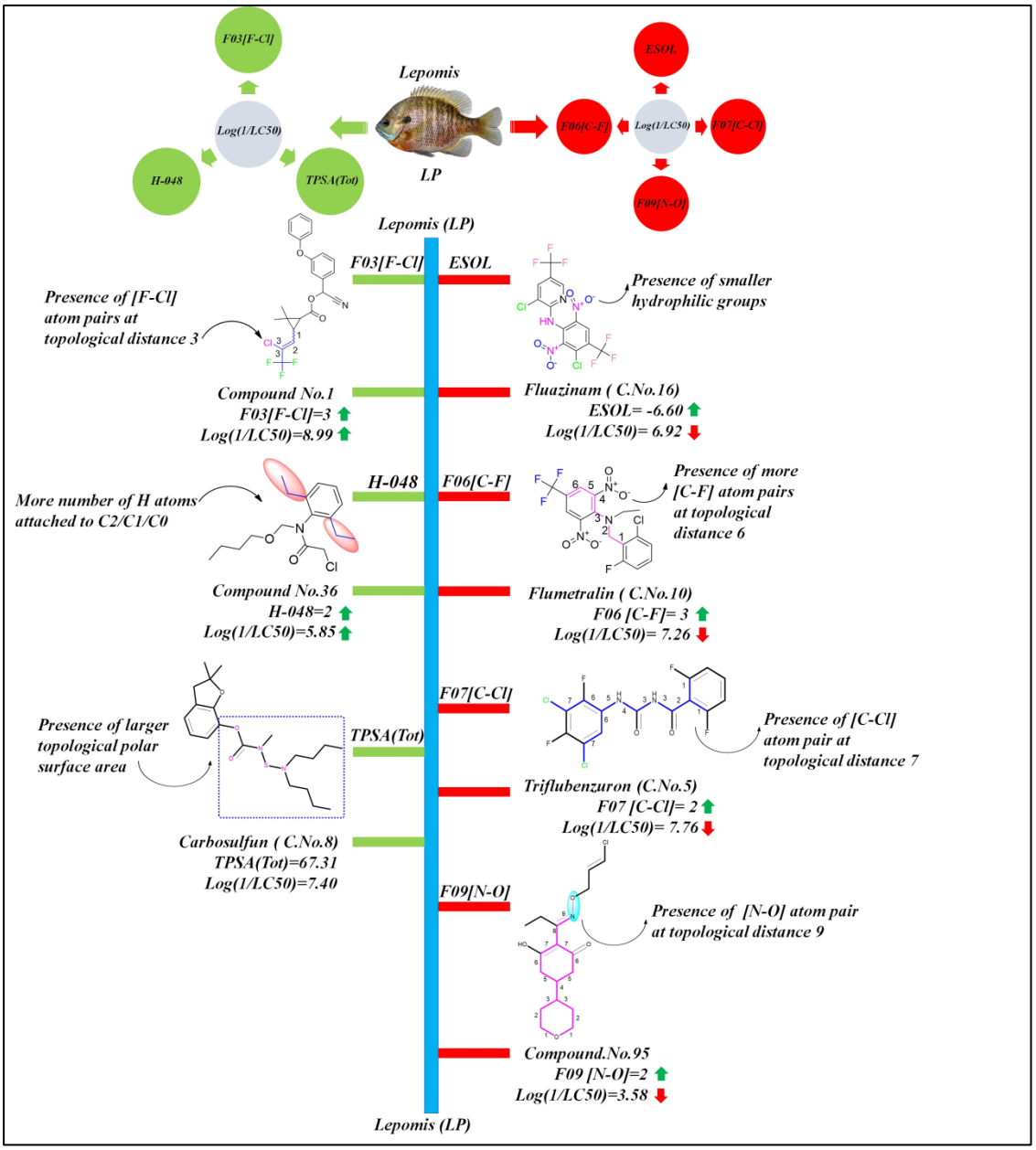


Fig 4.9. Insights of the modeled descriptors of LP (Lepomis) dataset

4.2.5. Insights on the Modeled Descriptors of the Miscellaneous dataset

The final PLS-based q-RASAR model of the Miscellaneous dataset has been developed with two RASAR descriptors (*RA function (LK)* and *SD Activity (LK)*) and three 0D-2D descriptors (ESOL, nCXr=, F02[O-S]). To know the type of contribution of the descriptors and their relative importance, we made use of coefficient plots and variable importance plots

in this study which is obtained from the SIMCA-P software. We have combined those two plots and represented them as a single bubble plot in Figure 4.7F, where the bubble radius indicates the VIP score, and the bubble colour indicates the type of contributions (green bubble = positive contribution, red bubble = negative contribution). Each bubble indicates a specific descriptor and the bubbles are arranged in descending order of importance from the left-hand side to the right-hand side of the X-axis. According to this plot, the order of importance of the modeled descriptors is – *RA function (LK)*> ESOL>*SD Activity (LK)*>nCXr=> F02[O-S], and the first three descriptors namely *RA function(LK)*> ESOL>*SD Activity(LK)* are the most important descriptors towards toxicity with VIP score >1. *RA function (LK), SD Activity (LK),*nCXr=,F02[O-S] have positive contributions whereas, the only rest descriptor ESOL has a negative contribution towards toxicity of Miscellaneous species(*Pimephales promelas* and *Brachydanio rerio*).

4.2.5.1. RA function (LK)

RA function (LK) is the most important and positive contributing descriptor of the PLS-based q-RASAR model developed from the Miscellaneous dataset. It is a read-across-derived prediction function obtained by the tool "RASAR-Desc-Calc-v2.0" employing the Laplacian kernel function similarity-based algorithm. It contains information from all the structural descriptors and process variables and functions similarly as a composite variable (Banerjee and Roy, 2022). The positive determination coefficient of it indicates a proportional relationship between this descriptor and the toxic endpoint. The comparative assessment between different data points also suggests the same. For instance, a more toxic compound (1R,3R,alphaR)-Cyfluthrin(**Compound 3**: Log (1/LC50) = 8.32) has a greater *RA function (LK)* value (8.29) than a comparatively less toxic compound Chloranil (**Compound 9**: Log (1/LC50) = 7.39;*RAfunction(LK)*= 4.68).

4.2.5.2. ESOL

According to the bubble plot **(Figure 4.7F)**, it is the most contributing descriptor towards the response value Log (1/LC50) of the resultant q-RASAR model of the Miscellaneous dataset. As a molecular property descriptor, it essentially transfers the knowledge of estimated solubility (logS) for aqueous solubility assessment using LOGPcons descriptor as well as indicating the hydrophobicity of organic molecules for determining toxic potency to aquatic species(Cronin, 2006). It seems to be a negative contributing descriptor of the model which

suggests the inverse correlation of it towards the toxic endpoint Log (1/LC50) in Miscellaneous fish species. This inverse relationship has been observed in other studied data points also of this dataset. For example, Oxychlordane (**Compound 3**) having larger hydrophobic moiety with less ESOL value (ESOL==-5.71) and thus causes more aquatic toxicity Log(1/LC50)=8.32 than a smaller hydrophilic compound Halfenprox **(Compound 16:**Log(1/LC50)=8.13) shows more ESOL value (ESOL=-7.11) as depicted in (**Figure 4.10)**. Therefore, it can be justified that the hydrophobicity of organic molecules may be linearly correlated with aquatic toxicity in the form of Log (1/LC50) values (Stibany et al., 2017).

4.2.5.3. SD Activity (LK)

SD Activity (LK) is the third most contributing RASAR descriptor in the resultant model of LP dataset as seen in the bubble plot. It indicates the standard deviation of the activity (observed) values of the selected close source compounds for each query compound. It is a positive regression coefficient, which seems positive influence on the response value for aquatic toxicity prediction. Therefore, compound (2,3,5,6-tetrafluoro-4-methylphenyl)methyl (1*S*,3*S*)-2,2-dimethyl-3-[(*E*)-prop-1-enyl]cyclopropane-1-carboxylate) (**Compound 5**) has higher *SD Activity(LK)*value (1.94) causes higher toxicity Log(1/LC50)=8.05 than (**Compound 7:** (1,9,10,11,12,12-hexachloro-4,6-dioxa-5λ^{6}-thiatricyclo[7.2.1.0^{2,8}]dodec-10-ene 5,5-dioxide and Log(1/LC50)=7.62) with smaller *SD Activity(LK)*values (0.48).

4.2.5.4. nCXr=

It is a "functional group count" type descriptor, indicating the number of X on ring C (sp2) in a molecule. It is the second least contributing descriptor towards the endpoint as observed from its lower VIP scores. As seen in the bubble plot **(Figure 4.7F)**, it shows a positive impact on the endpoint value, which indicates the presence of this kind of fragment increases the aquatic toxicity. For example, if a compound Endosulfan sulfate (**Compound 7**) containing this kind of fragment shows higher nCXr= values (nCXr==2) due to the presence of more unsaturation and thus more aquatic toxicity Log (1/LC50)=7.62 toward miscellaneous aquatic species than a compound Cyclohexanone (**Compound 195:** Log (1/LC50)=6.65**)** having less descriptor value (nCXr==1). Therefore, it can be justified that the presence of unsaturation in a molecule results in relatively higher toxicity on aquatic organisms in the form of Log (1/LC50) values (Nath et al., 2022).

4.2.5.5. F02[O-S]

It is the least contributing 2D atom pair descriptor as shown in VIP, which indicates the frequency of O-S atom pairs at topological distance 2. According to the bubble plot (**Figure 4.7F**), it has a positive influence towards the endpoint value for aquatic toxicity prediction. Then, we can justify that if a compound Bromophos-ethyl (**compound 23**) having this kind of atom pairs at topological distance 2 possess higher descriptor value **(**F02[O-S]=4) as well as more aquatic toxicity Log (1/LC50)=5.80 than a compound Cafenstrole **(compound 51:**Log(1/LC50)=5.26) with smaller descriptor value (F02[O-S]=3).This implies that the presence of more [O-S] atom pairs in a molecule shows relatively higher toxicity toward others aquatic species in terms of Log (1/LC50)values.

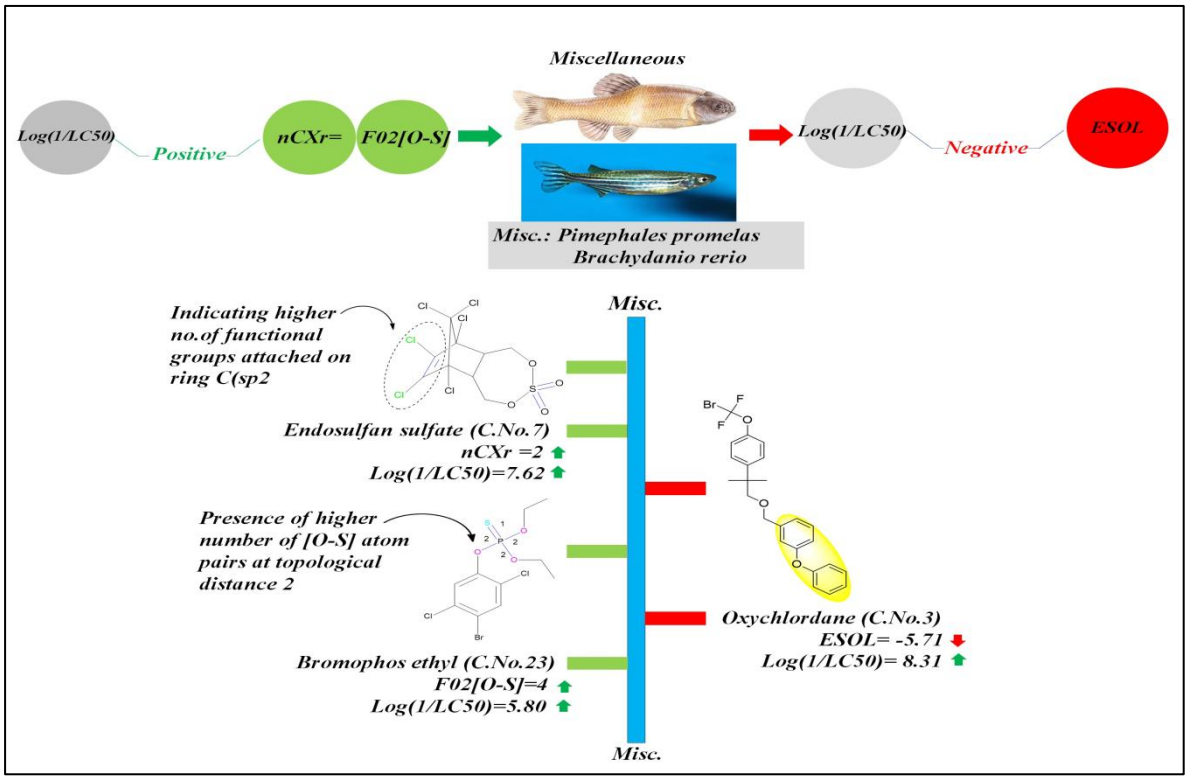


Fig. 4.10. Insights on the Modeled Descriptors of the Miscellaneous dataset

4.2.6. Importance of q-RASAR algorithm over conventional QSARs in the perspective of aquatic toxicity prediction

In this present study, both QSAR and q-RASAR models have been reported for the prediction of aquatic toxicity (Log1/LC50) of organic pesticides to different fish species (RT: *Oncorhynchus mykiss*; LP: *Lepomismacrochirus*; and (Miscellaneous: *Pimephales promelas* and *Brachydanio rerio*). Both the algorithms (QSAR and q-RASAR) can provide good

predictions as evidenced by the acceptable range of the computed validation metrics. We have comparatively analysed those computed validation metrics of QSAR and q-RASAR models of all studied datasets and to do so, we have tabulated the internal and external validation metrics of the developed models in the following **Table 4.7**. The q-RASAR models of all the datasets have clearly outperformed the QSAR models in terms of external predictivity as evidenced by the higher Q^2_{F1} (RT dataset: QSAR = 0.55, q-RASAR = 0.58; LP dataset: QSAR = 0.70, q-RASAR = 0.74; Micellaneous: QSAR = 0.51, q-RASAR = 0.59), Q^2_{F2} (RT dataset: QSAR = 0.54, q-RASAR = 0.57; LP dataset: QSAR = 0.66, q-RASAR = 0.70; Micellaneous: QSAR = 0.50, q-RASAR = 0.57) and lower MAE_{Test} (RT dataset: QSAR = 0.75, q-RASAR = 0.71; LP dataset: QSAR = 0.68, q-RASAR = 0.64; Micellaneous: QSAR = 0.77, q-RASAR = 0.78) values. The adjacency of R^2, Q^2_{LOO}, and MAE_{Train} between QSAR and q-RASAR models also proves the similar goodness of fit and robustness of the two algorithms. Therefore, it can be stated here that better external predictivity has been achieved in all three datasets using q-RASAR modeling approach without compromising the goodness of fit and robustness of QSARs. Previously, Li et al. reported three classification-based toxicity models developed from the PPDB datasets (*Oncorhynchus mykiss* = 829 data points, *Lepomis macrochirus* = 151 data points, and Miscellaneous (*Pimephales promelas* and *Brachydanio rerio*) = 278 data points) (Li et al., 2017). Although, there is no regression-based predictive model available in the literature. Therefore, the newly developed regression-based q-RASAR models can be a good source of reliable continuous predicted toxicity data which will be helpful for risk assessment and data-gap filling.

Table 4.7. Validation metrics of developed QSAR, q-RASAR models for RT, LP and Miscellaneous datasets

Datasets	Number of data points	Number of descriptors	No. of LVs	Type of models	Training set statistics			Test set statistics		
					R^2	Q^2_{Loo}	MAE_{Train}	Q^2_{F1}	Q^2_{F2}	MAE_{Test}
RT	n_{Train}=537 n_{Test}=178	8	5	QSAR	0.53	0.51	0.79	0.55	0.54	0.75
		8	4	RASAR	0.52	0.50	0.79	0.58	0.57	0.71
LP	n_{Train}=102 n_{Test}=34	8	3	QSAR	0.68	0.62	0.76	0.70	0.66	0.68
		8	3	RASAR	0.67	0.60	0.76	0.74	0.70	0.64
Miscellaneous	n_{Train}=170 n_{Test}=56	8	4	QSAR	0.60	0.54	0.79	0.51	0.50	0.77
		5	1	RASAR	0.54	0.52	0.79	0.59	0.57	0.78

CHAPTER - 5

CONCLUSION

"I have not failed. I've just found 10,000 ways that won't work."

- **Thomas Alva Edison (American inventor)**

5. CONCLUSION

The effectiveness of every research study depends on the findings and conclusions reached, which may emphasize previously known or undiscovered scientific insights. These results might also contribute to greater comprehension and expertise in the field in which the investigations were conducted. Results and explanations are always illustrative of previously established concepts. The present q-RASAR study might be beneficial for obtaining better external predictivity, interpretability, and transferability of the developed models, also it is an efficient technique that has the potential to be employed as a good alternative approach for retention time prediction and ecotoxicity potential identification. Here we have executed a straightforward and simple workflow in calculating descriptors, model development and judging their prediction reliability in a defined chemical space, and the diagnosed chemical information considering the OECD guidelines. Statistically validated models were derived employing different novel chemometric tools. All the developed models were validated through internal and external validation principles. Finally, the developed models were also subjected to a randomization analysis for validation purpose. Validation metrics indicate the goodness of fit, robustness, soundness and predictive ability of the developed models. Statistically reliable q-RASAR models obtained from these studies suggest that this approach is a purposefully added new method of risk assessment of pesticides that will help us to give more efficient aquatic toxicity predictions and to understand the toxicity mechanisms without harming animals.

5.1. Predictive q-RASPR modeling of the retention time (log t_R) of pesticide residues present in foods and vegetables

In this present study, a q-RASPR model has been developed with the experimental retention time data (log t_R) of 823 environmentally relevant pesticide residues (identified in foods and vegetable products), collected from the compound database (CDB). The retention time obtained from reverse-phase HPLC analysis has been used as the endpoint of this study due to its potential to be a toxicity indicator. This study was based on simple 2D descriptors to avoid complexity related to conformational analysis and energy minimization. The RASAR descriptors have been calculated from the 2D descriptors identified by the developed QSPR model in this study. The model was rigorously validated by various internal and external

validation metrics as recommended by the OECD. Additionally, the domain of applicability of the developed q-RASPR model has also been identified and reported. In a comparative analysis of the q-RASPR with the previously published work, the enhancement of external predictivity and reduction of prediction errors have been marked; the external predictivity in terms of Q^2_{F1} value of the q-RASPR model superseded the quality of the previously reported MLR model. From the descriptor interpretation, it has been observed that lipophilicity is the most important chemical property which positively affects the retention time, and as a result, to cellular toxicity. Besides this, several other features like the number of multiple bonds (nBM), graph density (GD) etc. have significant and inversely proportional relationships with the retention time endpoint. The software tools used in this study are easy to use and fast, and the majority of the tools are freely available which makes our workflow quite economical as compared to experimentation. The q-RASPR is a more efficient technique due to its better external predictivity, interpretability, and transferability; therefore, it has the potential to be used as a good alternative approach of retention time prediction and toxicity identification.

5.2. Quantitative Read-Across Structure-Activity Relationship (q-RASAR): A New Approach Methodology to Model Aquatic Toxicity of Organic Pesticides Against Three Different Fish Species

The prime objective of the present study was to develop regression-based in silico models for the aquatic prediction of various organic pesticides. In this direction, we have developed PLS-based q-RASAR models using experimental data [Log (1/LC50)] of organic pesticides to various fish species, including Rainbow trout (*Oncorhynchus mykiss*), Lepomis (*Lepomis macrochirus*), and Miscellaneous (*Pimephales promelas* and *Brachydanio rerio*). To encode the structural features of pesticide molecules, preliminary 0D-2D descriptors have been successfully used and therefore, the additional complexities associated with higher-order descriptor (3D-7D) calculation have been avoided. We have developed the preliminary QSAR models from the respective datasets, have strictly validated those models, and then applied the conjoined q-RASAR algorithm to further enhance the external predictability of those models. The OECD protocols have been followed during both QSAR and q-RASAR development. A marked increase in external predictability in terms of Q^2_{F1} and Q^2_{F2} has been observed after applying the RASAR algorithm. The errors of test set predictions (in terms of MAE_{Test}) have also been reduced from the preliminary QSAR models which indicate better predictability of q-RASAR models. During the QSAR and q-RASAR modeling,

hydrophobicity of the organic pesticides has been identified as a toxicity mediating factor which positively affects the overall aquatic toxicity. We have employed the freely available software tools for the model development to cut the additional cost and also didn't harm any animals during this study to make it a greener and ethical approach. As there were no previous predictive models for continuous aquatic toxicity data prediction of the PPDB pesticides, this study will be an effective alternative approach for aquatic toxicity prediction of different existing and untested organic pesticides without experimentations.

CHAPTER - 6

REFERENCES

"It's kind of fun to do the impossible." –

Walt Disney (American entrepreneur, animator)

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. DOI: 10.12732/ijpam.v94i3.2Adeniji, S.E., Uba, S., Uzairu, A., Arthur, D.E., 2019. A derived QSAR model for predicting some compounds as potent antagonist against mycobacterium tuberculosis : A theoretical approach. Adv. Prev. Med. 2019, 5173786. https://doi.org/10.1155/2019/5173786

Al-Saleh, I., 1994. Pesticides: a review article. J. Environ. Pathol. Toxicol. Oncol. 13, 151-161.

Ambure, P., Gajewicz-Skretna, A., Cordeiro, M.N.D., Roy, K., 2019. New workflow for QSAR model development from small data sets: Small dataset curator and small dataset modeller. Integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. J. Chem. Inf. Model. 59(10), 4070-4076. https://doi.org/10.1021/acs.jcim.9b00476

Asadollahi-Baboli, M., 2012. In silico prediction of the aniline derivatives toxicities to Tetrahymena pyriformis using chemometrics tools. Toxicol. Environ. Chem. 94, 2019–2034. https://doi.org/10.1080/02772248.2012.740046

Atkinson, A.C., 1985. Plots, Transformations and Regression, Oxford, Clarendon Press.

Auer, C.M., Nabholz, J.V., Baetcke, K.P., 1990. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environ. Health Perspect. 87, 183–197.

Balaban, A.T., 1982. Highly discriminating distance-based topological index. Chemical physics letters, 89(5), 399-404.

Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol. Divers. 26, 2847–2862. https://doi.org/10.1007/s11030-022-10478-6

Belanger, S.E., Rawlings, J.M., Carr, G.J., 2013. Use of fish embryo toxicity tests for the prediction of acute fish toxicity to chemicals. Environ. Toxicol. Chem. 32(8), 1768-1783. https://doi.org/10.1002/etc.2244

Benfenati, E. (Ed.), 2012. The e-Book on QSAR and REACH: Theory, Guidance and Applications.

Benfenati, E., Chaudhry, Q., Gini, G. and Dorne, J.L., 2019. Integrating in silico models and read-across methods for predicting toxicity of chemicals: A step-wise strategy. *Environment international*, *131*, p.105060.Abe, T., Saito, H., Niikura, Y., Shigeoka, T., Nakano, Y., 2001. Embryonic development assay with *Daphnia magna*: Application to toxicity of aniline derivatives. Chemosphere 45(4-5), 487-495. https://doi.org/10.1016/S0045-6535(01)00049-2

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K. and Wiswedel, B., 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11(1), 26-31.

Borman, S., 1990. New QSAR techniques eyed for environmental assessments. Chemical & Engineering News, 68(8), 20-23

Braunbeck, T., Böttcher, M., Hollert, H., Kosmehl, T., Lammer, E., Leist, E., Rudolf, M., Seitz, N., 2005. Towards an alternative for the acute fish LC50 test in chemical assessment: The fish embryo toxicity test goes multi-species - an update. ALTEX - Altern. to Anim. Exp. 22, 87–102

Brown, A.C. and Fraser, T.R., 1868. V.—on the connection between chemical constitution and physiological action. Part. I.—on the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. Earth and Environmental Science Transactions of The Royal Society of Edinburgh, 25(1), 151-203.

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

Cars, O., Mölstad, S. and Melander, A., 2001. Variation in antibiotic use in the European Union. The Lancet, 357(9271), 1851-1853.

Ccanccapa, A., Masiá, A., Navarro-Ortega, A., Picó, Y., Barceló, D., 2016. Pesticides in the Ebro River basin: occurrence and risk assessment. Environ. Pollut. 211, 414-424.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.

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, 189–203. https://doi.org/10.1039/d1en00725d

Chatterjee, M., Banerjee, A., Tosi, S., Carnesecchi, E., Benfenati, E., Roy, K., 2023. Machine learning - based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees. J. Hazard. Mater. 460, 132358. https://doi.org/10.1016/j.jhazmat.2023.132358

Chatterjee, M., Roy, K., 2021. Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors. J. Hazard. Mater. 408, 124936. https://doi.org/10.1016/J.JHAZMAT.2020.124936

Chatterjee, M., Roy, K., 2022. Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures. SAR QSAR Environ. Res. 33, 463–484. https://doi.org/10.1080/1062936X.2022.2081255

Chatterjee, Mainak, Roy, K., 2022. Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against Aliivibrio fischeri. Chemosphere 308, 136463. https://doi.org/10.1016/j.chemosphere.2022.136463

ChemSpider [WWW Document], n.d. URL http://www.chemspider.com/ (accessed 2.27.23).

Chen, H., Carlsson, L., Eriksson, M., Varkonyi, P., Norinder, U. and Nilsson, I., 2013. Beyond the scope of FreeWilson analysis: building interpretable QSAR models with machine learning algorithms. Journal of chemical information and modeling, 53(6), 1324-1336.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R. and Consonni, V., 2014. QSAR modeling:

where have you been? Where are you going to?. Journal of medicinal chemistry, 57(12), 4977-5010.

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), 2320-2335.

Chirico, N. and Gramatica, P., 2012. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. Journal of Chemical Information and Modeling, 52(8), 2044-2058.

CIOMS - Council for international organizations of medical sciences, 1985. International Guiding Principles for Biomedical Research Involving Animals, Available at http://cioms.ch/publications/guidelines/1985_texts_of_guidelines.htm (accessed 2 May 2016)

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q2 parameter for QSAR validation. J. Chem. Inf. Model. 49, 1669–1678.

Consonni, V., Ballabio, D., Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques. J. Chemometrics 24, 194–201.

Cortés-Ciriano, I., Ain, Q.U., Subramanian, V., Lenselink, E.B., Méndez-Lucio, O., IJzerman, A.P., Wohlfahrt, G., Prusis, P., Malliavin, T.E., van Westen, G.J. and Bender, A., 2015. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. MedChemComm, 6(1), 24-50.

Cronin, M.T.D., 2006. The Role of Hydrophobicity in Toxicity Prediction. Curr. Comput. Aided. Drug Des. 2, 405–413. https://doi.org/10.2174/157340906778992346

Darlington, R.B., 1990. Regression and Linear Models, McGraw Hill, New York.

Das, S., Ojha, P.K., Roy, K., 2017. Development of a temperature dependent 2D-QSPR model for viscosity of diverse functional ionic liquids. J. Mol. Liq. 240, 454-467. https://doi.org/10.1016/j.molliq.2017.05.113

Data Analytics Platform: Open Source Software Tools | KNIME [WWW Document], n.d.

URL https://www.knime.com/knime-analytics-platform (accessed 2.27.23).

De, P., Aher, R.B., Roy, K., 2018. Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector Aedes aegypti: application of ETA indices. RSC A 8, 4662–4670. https://doi.org/10.1039/c7ra13159c

Debnath, A.K. 2001. In: Ghose, A.K., Viswanadhan V.N., (Eds.), Combinatorial Library Design and Evaluation. Marcel Dekker, New York.

Devillers, J., 1996. Genetic algorithms in molecular modelling. Academic Press, San Diego, CA.

dos Santos, G.L.A.A., Reis, A.S., Besen, M.R., Furlanetto, R.H., Rodrigues, M., Crusiol, L.G.T., de Oliveira, K.M., Falcioni, R., de Oliveira, R.B., Batista, M.A., Nanni, M.R., 2023. Spectral method for macro and micronutrient prediction in soybean leaves using interval partial least squares regression. Eur. J. Agron. 143, 126717. https://doi.org/10.1016/J.EJA.2022.126717

Edwards, C.A. (Eds.), 2013. Environmental Pollution by Pesticides. Environmental Science Research, Vol. 3. Springer, Boston, MA. Springer, Boston, MA.

Embry, M.R., Belanger, S.E., Braunbeck, T.A., Galay-Burgos, M., Halder, M., Hinton, D.E., Léonard, M.A., Lillicrap, A., Norberg-King, T., Whale, G., 2010. The fish embryo toxicity test as an animal alternative method in hazard and risk assessment and scientific research. Aquat. Toxicol. 97(2), 79-87. https://doi.org/10.1016/j.aquatox.2009.12.008

Eriksson, M., Chen, H., Carlsson, L., Nissink, J.W.M., Cumming, J.G. and Nilsson, I., 2014. Beyond the scope of free-Wilson analysis. 2: Can distance encoded R-group fingerprints provide interpretable nonlinear models?. Journal of chemical information and modeling, 54(4), 1117-1128

European Commission, 2006. Regulation (EC) No 1907/2006 of The European Parliament and of The Council of 18December 2006 concerning the Registration, Evaluation. Authorization and Restriction of Chemicals (REACH),establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation(EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Off. J. Eur. Union Lett. 396, 1–849.

Fourches, D., Muratov, E. and Tropsha, A., 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. Journal of chemical information and modeling, 50(7), p.1189.

Fourches, D., Muratov, E. and Tropsha, A., 2016. Trust, but verify II: a practical guide to chemogenomics data curation. Journal of chemical information and modeling, 56(7), 1243-1252.

Free, S.M. and Wilson, J.W., 1964. A mathematical contribution to structure-activity studies. Journal of medicinal chemistry, 7(4), 395-399.

Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSARs models: Where theory meets reality. Int. J. Quant. Struct. Prop. Relationships 1, 45–63. https://doi.org/10.4018/IJQSPR.2016010102

Gallup, G.A., Gilkerson, W.R. and Jones, M.M., 1952. A Theoretical Formulation of the Hammett Equation. Transactions of the Kansas Academy of Science, 55(2), 232-236.

Gao, H., Katzenellenbogen, J.A., Garg, R. and Hansch, C., 1999. Comparative QSAR analysis of estrogen receptor ligands. Chemical Reviews, 99(3), 723-744.

Garg, R., Gupta, S.P., Gao, H., Babu, M.S., Debnath, A.K. and Hansch, C., 1999. Comparative quantitative structure− activity relationship studies on anti-HIV drugs. Chemical reviews, 99(12), 3525-3602.

Garg, R., Kurup, A., Mekapati, S.B. and Hansch, C., 2003. Cyclooxygenase (COX) inhibitors: a comparative QSAR study. Chemical reviews, 103(3), 703-732

Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J., 1998. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. The Journal of Physical Chemistry A, 102(21), 3762-3772.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Mod. 20, 269–276.

Gorzalczany, S.B., Rodriguez Basso, A.G., 2021. Strategies to apply 3Rs in preclinical

testing. Pharmacol. Res. Perspect. 9, e00863. https://doi.org/10.1002/PRP2.863

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.

Gutman, I. and Trinajstić, N., 1972. Graph theory and molecular orbitals. Total φ-electron energy of alternant hydrocarbons. Chemical physics letters, 17(4), 535-538.Glinski, D.A., Purucker, S.T., Van Meter, R.J., Black, M.C., Henderson, W.M., 2018. Analysis of pesticides in surface water, stemflow, and throughfall in an agricultural area in South Georgia, USA. Chemosphere 209, 496-507.

Hadjipavlou-Litina, D., Garg, R. and Hansch, C., 2004. Comparative Quantitative Structure− Activity Relationship Studies (QSAR) on Non-Benzodiazepine Compounds Binding to Benzodiazepine Receptor (BzR). Chemical reviews, 104(9), 3751-3794.

Håkanson, L. and Bryhn, A., 1999. Water pollution. Backhuys Publ, Leiden.

Hammett, L.P., 1935. Some relations between reaction rates and equilibrium constants. Chemical reviews, 17(1), 125-136.

Hengstler, J.G., Foth, H., Kahl, R., Kramer, P.-J., Lilienblum, W., Schulz, T, Schweinfurth, H., 2006. The REACH concept and its impact on toxicological sciences. Toxicol. 220, 232–239.

Hansch, C. and Leo, A., 1979. Substituent constants for correlation analysis in chemistry and biology. Wiley, New York.

Hansch, C., 1969. Quantitative approach to biochemical structure-activity relationships. Accounts of chemical research, 2(8), 232-239.

Hansch, C., 2011. The advent and evolution of QSAR at Pomona College. Journal of computer-aided molecular design, 25(6), 495-507.

Hansch, C., Gao, H., 1997. Comparative QSAR: Radical reactions of benzene derivatives in chemistry and biology. Chemical reviews, 97(8), 2995-3060.

Hansch, C., Hoekman, D., Leo, A., Weininger, D. and Selassie, C.D., 2002. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. Chemical reviews, 102(3), 783-812.

Hansch, C., Leo, A. and Taft, R.W., 1991. A survey of Hammett substituent constants and resonance and field parameters. Chemical reviews, 91(2), 165-195.

Hansch, C., Leo, A., Hoekman, D., 1995, Exploring QSAR: Hydrophobic, electronic, steric constants. ACS, Washington, DC.

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

Hill, A.J., Teraoka, H., Heideman, W., Peterson, R.E., 2005. Zebrafish as a model vertebrate for investigating chemical toxicity. Toxicol. Sci. 86(1), 6-19. https://doi.org/10.1093/toxsci/kfi110

Holland, J., 1975. Adaptation in natural and artificial systems: an introductory analysis with application to biology. Control and artificial intelligence. MIT Press, Cambridge, MA & London, England.

Hongmao, S., 2016. Quantitative Structure–Property Relationships Models for Lipophilicity and Aqueous Solubility, in: A Practical Guide to Rational Drug Design. Woodhead Publishing, pp. 193–223. https://doi.org/10.1016/B978-0-08-100098-4.00006-5

Hopfinger, A.J., Esposito, E.X., Llinàs, A., Glen, R.C., Goodman, J.M., 2009. Findings of the challenge to predict aqueous solubility. J. Chem. Inf. Model. 49, 1–5. https://doi.org/10.1021/ci800436c

Horie, Y., Yamagishi, T., Koshio, M., Iguchi, T., Tatarazako, N., 2017. Lethal and sublethal effects of aniline and chlorinated anilines on zebrafish embryos and larvae. J. Appl. Toxicol. 37(7), 836-841. https://doi.org/10.1002/jat.3431

Hotelling, H., 1931. The generalization of student's ratio. Ann. Math. Stat. 2, 360–378. https://doi.org/10.1214/aoms/1177732979

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496(7446), 498-503.

Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T., 2005. QSAR applicability domain estimation by projection of the training set descriptor space: a review. ATLA (Altern. Lab. Anim.) 33, 445–459

Jia, Q., Wang, S., Yu, M., Wang, Q., Yan, F., 2023. Two QSAR models for predicting the toxicity of chemicals towards Tetrahymena pyriformis based on topological-norm descriptors and spatial-norm descriptors. SAR QSAR Environ. Res. 34, 147–161. https://doi.org/10.1080/1062936X.2023.2171478

Jones, O.A.H., Voulvoulis, N. and Lester, J.N., 2001. Human pharmaceuticals in the aquatic environment a review. Environmental technology, 22(12), 1383-1394.

Jouan-Rimbaud, D., Massart, D.L., Leardi, R. and De Noord, O.E., 1995. Genetic algorithms as a tool for wavelength selection in multivariate calibration. Analytical Chemistry, 67(23), 4295-4301.

Kahn, I., Maran, U., Benfenati, E., Netzeva, T.I., Schultz, T.W. and Cronin, M.T., 2007. Comparative quantitative structure–activity–activity relationships for toxicity to Tetrahymena pyriformis and Pimephales promelas. Alternatives to Laboratory Animals, 35(1), 15-24.

Kaliszan, R., 2007. QSRR: quantitative structure-(chromatographic) retention relationships. Chemical reviews, 107(7), 3212-3246.

Kar, S., Leszczynski, J. (Eds.), 2023. Current trends in computational modeling in drug discovery. Springer Cham.

Khan, K., Khan, P.M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., Benfenati, E., 2019. QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors. Chemosphere 229, 8–17. https://doi.org/10.1016/j.chemosphere.2019.04.204

Kier, L.B., Hall, L.H., 1977. The nature of structure-activity relationships and their relation to molecular connectivity. Eur. J. Med. Chem. 12, 307–312.

Kier, L.B., Hall, L.H., 1986. Molecular connectivity in structure-activity analysis, Research Studies Press, John Wiley & Sons Ltd., Chichester, UK.

Kim, K.H., 2001. Thermodynamic aspects of hydrophobicity and biological QSAR. Journal of computer-aided molecular design, 15(4), 367-380.

Kirkpatrick, P., 2004. Gliding to success. Nature Reviews Drug Discovery, 3(4), 299.

Knöbel, M., Busser, F.J., Rico-Rico, A., Kramer, N.I., Hermens, J.L., Hafner, C., Tanneberger, K., Schirmer, K., Scholz, S., 2012. Predicting adult fish acute lethality with the zebrafish embryo: relevance of test duration, endpoints, compound properties, and exposure concentration analysis. Environ. Sci. Technol. 46(17), 9690-9700. https://doi.org/10.1021/es301729q

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

Kubinyi, H., 1988. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. Quantitative Structure- Activity Relationships, 7(3), 121-133.

Kumar, P., Kumar, A., 2018. Monte Carlo Method Based QSAR Studies of Mer Kinase Inhibitors in Compliance with OECD Principles. Drug Res. (Stuttg). 68, 189–195. https://doi.org/10.1055/s-0043-119288

Kurup, A., Garg, R. and Hansch, C., 2000. Comparative QSAR analysis of 5α-reductase inhibitors. Chemical reviews, 100(3), 909-924.

Kurup, A., Garg, R. and Hansch, C., 2001. Comparative QSAR study of tyrosine kinase inhibitors. Chemical reviews, 101(8), 2573-2600.

Lambert, F.N., Vivian, D.N., Raimondo, S., Tebes-Stevens, C.T., Barron, M.G., 2022. Relationships between aquatic toxicity, chemical hydrophobicity, and mode of action: Log Kow revisited. Arch. Environ. Contam. Toxicol. 83, 326–338. https://doi.org/10.1007/s00244-022-00944-5

Leclercq, M., Mathieu, O., Gomez, E., Casellas, C., Fenet, H. and Hillaire-Buys, D., 2009. Presence and fate of carbamazepine, oxcarbazepine, and seven of their metabolites at

wastewater treatment plants. Archives of environmental contamination and toxicology, 56(3), 408-415.

Leonard, J.T. and Roy, K., 2006. On selection of training and test sets for the development of predictive QSAR models. QSAR & Combinatorial Science, 25(3), 235-251.

Li, F., Fan, D., Wang, H., Yang, H., Li, W., Tang, Y., Liu, G., 2017. In silico prediction of pesticide aquatic toxicity with chemical category approaches. Toxicol. Res. (Camb). 6, 831–842. https://doi.org/10.1039/c7tx00144d

Li, T., Huang, Y., Wei, G., Zhang, Y. nan, Zhao, Y., Crittenden, J.C., Li, C., 2020. Quantitative structure-activity relationship models for predicting singlet oxygen reaction rate constants of dissociating organic compounds. Sci. Total Environ. 735, 139498. https://doi.org/10.1016/j.scitotenv.2020.139498

Lipnick, R. L., 1991. Studies of narcosis. Springer, Dordrecht, Netherlands.

Liu, T., Yan, F., Jia, Q. and Wang, Q., 2020. Norm index-based QSAR models for acute toxicity of organic compounds toward zebrafish embryo. Ecotoxicology and Environmental Safety, 203, p.110946.

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

Luch A. (ed.), 2010. Molecular, Clinical and Environmental Toxicology. Experientia Supplementum, vol. 100. Birkhäuser, Basel. https://link.springer.com/book/10.1007%2F978-3-7643-8338-1 (accessed 25 March 2021).

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, 198–212. https://doi.org/10.1093/toxsci/kfy152

Maltby,T., 2013. European Union energy policy integration: A case of European Commission policy entrepreneurship and increasing supranationalism. *Energy policy*, *55*, pp.435-444.

Mannhold, R. (Ed.), 2007. Molecular Drug Properties: Measurement and Prediction: Methods and Principles in Medicinal Chemistry. Wiley-VCH Verlag GmbH & Co. KGaA

Marvin, ChemAxon [WWW Document], n.d. URL https://chemaxon.com/marvin (accessed 2.27.23).

Mauri, A., 2020. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints, in: Roy, K. (Ed.), Methods in Pharmacology and Toxicology. Humana Press Inc., New York, pp. 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32

Mayer, P., Reichenberg, F., 2006. Can highly hydrophobic organic substances cause aquatic baseline toxicity and can they contribute to mixture toxicity? Environ. Toxicol. Chem. 25, 2639–2644. https://doi.org/10.1897/06-142r.1

McLeish, J.A., Chico, T.J., Taylor, H.B., Tucker, C., Donaldson, K., Brown, S.B., 2010. Skin exposure to microand nano-p

Melnikov, F., Costal, J., Voutchkova-Kostal, A., Zimmerman, J.B., Anastas, P.T., 2016. Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals. Green Chem 18, 4432–4445. https://doi.org/10.1039/C6GC00720A.

Milan, D.J., Peterson, T.A., Ruskin, J.N., Peterson, R.T., MacRae, C.A., 2003. Drugs that induce repolarization abnormalities cause bradycardia in zebrafish. Circulation 107(10), 1355-1358. https://doi.org/10.1161/01.CIR.0000061912.88753.87

Mitchell M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.

Mitra, I., Roy, K., 2012. On the Use of the Metric rm2 as an Effective Tool for Validation of QSAR Models in Computational Drug Design and Predictive Toxicology. Mini Rev. Med. Chem. 12, 491–504.

Monserud, J.H., Schwartz, D.K., 2012. Effects of molecular size and surface hydrophobicity on oligonucleotide interfacial dynamics. Biomacromolecules 13, 4002–4011. https://doi.org/10.1021/bm301289n

Nantasenamat, C., Isarankura-Na-Ayudhya, C. and Prachayasittikul, V., 2010. Advances in computational methods to predict the biological activity of compounds. Expert opinion on drug discovery, 5(7), 633-654.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T. and Prachayasittikul, V., 2009. A practical overview of quantitative structure-activity relationship.

Nath, A., De, P., Roy, K., 2022. QSAR modelling of inhalation toxicity of diverse volatile organic molecules using no observed adverse effect concentration (NOAEC) as the endpoint. Chemosphere 287, 131954. https://doi.org/10.1016/j.chemosphere.2021.131954

Nudds, R.L., John, E.L., Keen, A.N. and Shiels, H.A., 2014. Rainbow trout provide the first experimental evidence for adherence to a distinct Strouhal number during animal oscillatory propulsion. Journal of Experimental Biology, 217(13), 2244-2249.

OECD 236, 2013. Guidelines for the Testing of Chemicals, Section 2. Fish Embryo Acute Toxicity (FET). OECD Publishing, Paris.

OECD, 2004. OECD principles for the validation, for regulatory purposes, of (Quantitative) structure-activity relationship models [WWW Document]. URL https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

OECD, 2014. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264085442-en (accessed 7 March 2021).

Ojha, P.K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring rm2 metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 107, 194–205.

Panter, G.H., Thompson, R.S. and Sumpter, J.P., 1998. Adverse reproductive effects in male fathead minnows (Pimephales promelas) exposed to environmentally relevant concentrations of the natural oestrogens, oestradiol and oestrone. Aquatic Toxicology, 42(4), 243-253.

Prasanna, S., Doerksen, R.J., 2009. Topological polar surface area: a useful descriptor in 2D-QSAR. Curr. Med. Chem. 16, 21–41. https://doi.org/10.2174/092986709787002817

PubChem [WWW Document], n.d. URL https://pubchem.ncbi.nlm.nih.gov/ (accessed 2.27.23).

Qiao, K., Fu, W., Jiang, Y., Chen, L., Li, S., Ye, Q., Gui, W., 2020. QSAR models for the acute toxicity of 1, 2, 4-triazole fungicides to zebrafish (*Danio rerio*) embryos. Environ. Pollut. 265, 114837. https://doi.org/10.1016/j.envpol.2020.114837

Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., Cao, Z. and Zhu, R., 2017. The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope. Briefings in bioinformatics, 18(1), 125-136

Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment; Roy, K., Ed.; IGI Global: Hershey, PA, 2015.

Radoux, C.J., Olsson, T.S., Pitt, W.R., Groom, C.R. and Blundell, T.L., 2016. Identifying interactions that determine fragment binding at protein hotspots. Journal of medicinal chemistry, 59(9), 4314-4325.

Rekker, R., 1977. In The Hydrophobic Fragmental Constant. Elsevier, Amsterdam.

Roberts, D.W., Marshall, S.J., 1995. Application of hydrophobicity parameters to prediction of the acute aquatic toxicity of commercial surfactant mixtures. SAR QSAR Environ. Res. 4, 167–176. https://doi.org/10.1080/10629369508029914

Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G., 1999. Self-organizing molecular field analysis: A tool for structure− activity studies. Journal of medicinal chemistry, 42(4), 573-583.

Rogers, D., Hopfinger, A.J., 1994. Application of genetic function approximation to quantitative structure_-activity relationships and quantitative structure-property relationships. J. Chem. Inf. Comput. Sci. 34, 854–866.

Rojas, C., Aranda, J.F., Pacheco Jaramillo, E., Losilla, I., Tripaldi, P., Duchowicz, P.R., Castro, E.A., 2021. Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap. Food Chem. 342, 128354. https://doi.org/10.1016/j.foodchem.2020.128354

Roy, K. (Ed.), 2015. Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. IGI Global, Hershey, PA.

Roy, K. and 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. Journal of hazardous materials, 254, 166-178.

Roy, K. and Das, R.N., 2014. A review on principles, theory and practices of 2D-QSAR. Current drug metabolism, 15(4), 346-379.

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

Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opinion on Drug Discovery, 2(12), 1567-1577.

Roy, K., Ambure, P., Kar, S., 2018b. 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., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr. Intell. Lab. Syst. 152, 18-33. https://doi.org/10.1016/j.chemolab.2016.01.008

Roy, K., Kabir, H., 2012b. QSPR with extended topochemical atom (ETA) indices. 3. Modeling of critical micelle concentration of cationic surfactants. Chem. Eng. Sci. 81, 169–178.

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

Roy, K., Kar, S., 2015. Importance of Applicability Domain of QSAR Models, in: Roy. K. (Ed), Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. IGI Global, 180–211.

Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145, 22–29.

Roy, K., Kar, S., Das, R.N., 2015. A primer on QSAR/QSPR Modelling: Fundamental concepts. Springer briefs in molecular science, Springer Cham, Heidelberg, New York. https://doi.org/10.1007/978-3-319-17281-1

Roy, K., Kar, S., Das, R.N., 2015. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press.

https://doi.org/10.1016/C2014-0-00286-9

Roy, K., Mitra, I., 2011. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb. Chem. High Throughput Screen. 14, 450–474. https://doi.org/10.2174/138620711795767893

Roy, K., Mitra, I., 2011. On various metrics used for validation of predictive QSAR modelwith applications in virtual screening and focused library design. Comb. Chem. High Throughput Screen. 14, 450–474. https://doi.org/10.2174/138620711795767893

Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model. 52(2), 396-408. https://doi.org/10.1021/ci200520g

Roy, K., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., Kabir, H., 2012. Introduction of rm2(rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. Chemom. Intell. Lab. Syst. 118, 200–210. https://doi.org/10.1016/J.CHEMOLAB.2012.06.004

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

Sangodoyin, A.Y., 1996. Nutrient benefits and environmental aspects of land disposal of livestock waste. *Environmental Management and Health*, *7*(1), pp.33-38.

Saxena, A.K., Prathipati, P., 2003. Comparison of MLR, PLS and GA-MLR in QSAR analysis. SAR QSAR Environ. Res. 14, 433-445. https://doi.org/10.1080/10629360310001624015

Segall, M., Champness, E., Obrezanova, O., Leeding, C., 2009. Beyond profiling: using ADMET models to guide decisions. Chemistry & biodiversity, 6(11), 2144-2151.

Smith, K.E.C., Dom, N., Blust, R., Mayer, P., 2010. Controlling and maintaining exposure of hydrophobic organic compounds in aquatic toxicity tests by passive dosing. Aquat. Toxicol. 98, 15–24. https://doi.org/10.1016/j.aquatox.2010.01.007

Soares, J.X., Santos, Á., Fernandes, C., Pinto, M.M.M., 2022. Liquid Chromatography on the Different Methods for the Determination of Lipophilicity: An Essential Analytical Tool in Medicinal Chemistry. Chemosensors 10, 340. https://doi.org/10.3390/chemosensors10080340

Stibany, F., Schmidt, S.N., Schäffer, A., Mayer, P., 2017. Aquatic toxicity testing of liquid hydrophobic chemicals – Passive dosing exactly at the saturation limit. Chemosphere 167, 551–558. https://doi.org/10.1016/j.chemosphere.2016.10.014

Taylor P.J., 1991, Quantitative drug design. the rational design, mechanistic study and therapeutic applications of chemical compounds. Comprehensive medicinal chemistry. 4, 241-94.

Todeschini, R. and Consonni, V., 2008. Handbook of molecular descriptors (Vol. 11). John Wiley & Sons.

Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany. https://doi.org/10.1002/9783527613106

Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H., Perkins, R., 2005. Assessing QSAR limitations-A regulatory perspective. Current Computer-Aided Drug Design, 1(2), 195-205.

Toropov, A.A., Toropova, A.P., Selvestrel, G., Baderna, D. and Benfenati, E., 2020. Prediction of No Observed Adverse Effect Concentration for inhalation toxicity using Monte Carlo approach. SAR QSAR Environ. Res. 31(12), 1-12. https://doi.org/10.1080/1062936X.2020.1841827

Tunkel, J., Mayo, K., Austin, C., Hickerson, A., Howard, P., 2005. Practical considerations on the use of predictive models for regulatory purposes. Environ. Sci. Technol. 39, 2188–2199. https://doi.org/10.1021/es049220t

Umetri, A.B., 1996. SIMCA-P for Windows. Graphical Software for Multi-variate Process Modelling. Umeå, Sweden.

United States Environmental Protection Agency, n.d., Acute Toxicity WET Methods: Freshwater and Marine Organisms. https://www.epa.gov/cwa-methods/acute-toxicity-wet-methods (accessed 10 June 2021)

Venkatasubramanian, V., Sundaram, A., 2002. Genetic algorithms: introduction and applications. Encycl. Comput. Chem. https://doi.org/10.1002/0470845015.cga003

Verma, J., Khedkar, V.M. and Coutinho, E.C., 2010. 3D-QSAR in drug design-a review. Current topics in medicinal chemistry, 10(1), 95-115.

Verma, R.P. and Hansch, C., 2009. Camptothecins: a SAR/QSAR study. Chemical reviews, 109(1), 213-235

Vesterkvist, P. S. M.; Misiorek, J. O.; Spoof, L. E. M.; Toivola, D. M.; Meriluoto, J. A. O. Comparative Cellular Toxicity of Hydrophilic and Hydrophobic Microcystins on Caco-2 Cells. *Toxins (Basel).* **2012**, *4*, 1008–1023. https://doi.org/10.3390/toxins4111008.

Wang, J., Chow, W., Wong, J.W., Leung, D., Chang, J., Li, M., 2019. Non-target data acquisition for target analysis (nDATA) of 845 pesticide residues in fruits and vegetables using UHPLC/ESI Q-Orbitrap. Anal. Bioanal. Chem. 411, 1421–1431. https://doi.org/10.1007/s00216-019-01581-z

Weaver, S., Gleeson M.P., 2008. The importance of the domain of applicability in QSAR modeling. J. Mol. Graph. Model. 26, 1315–1326.

Wei, D.B., Zhang, A.Q., Han, S.K. and Wang, L.S., 2001. Joint QSAR analysis using the Free-Wilson approach and quantum chemical parameters. SAR and QSAR in Environmental Research, 12(5), 471-479.

Wiener, H., 1947. Structural determination of paraffin boiling points. Journal of the American chemical society, 69(1), 17-20.

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

Wollenberger, L., Halling-Sørensen, B. and Kusk, K.O., 2000. Acute and chronic toxicity of veterinary antibiotics to Daphnia magna. Chemosphere, 40(7), 723-730

Wu, Z., Li, D., Meng, J., Wang, H., 2010. Introduction to SIMCA-P and Its application, in: Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares. Springer, Berlin, Heidelberg, pp. 757–774. https://doi.org/10.1007/978-3-540-32827-8_33

Yu, X., Zeng, Q., 2022. Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes. Aquat. Toxicol. 251, 106265. https://doi.org/10.1016/j.aquatox.2022.106265

Yukawa, T.; Naven, R. Utility of Physicochemical Properties for the Prediction of Toxicological Outcomes: Takeda Perspective. *ACS Med. Chem. Lett.* **2020**, *11*, 203–209. https://doi.org/10.1021/acsmedchemlett.9b00536.

APPENDIX

REPRINTS

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

pubs.acs.org/JAFC

Article

Predictive Quantitative Read-Across Structure−Property Relationship Modeling of the Retention Time (Log t_R) of Pesticide Residues Present in Foods and Vegetables

Shilpayan Ghosh, Mainak Chatterjee, and Kunal Roy*

Cite This: *J. Agric. Food Chem.* 2023, 71, 9538−9548

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The retention time (log t_R) of pesticidal compounds in a reverse-phase high-performance liquid chromatography (HPLC) analysis has a direct relationship with lipophilicity, which could be related to the ecotoxicity potential of the compounds. The novel quantitative read-across structure−property relationship (q-RASPR) modeling approach uses similarity-based descriptors for predictive model generation. These models have been shown to enhance external predictivity in previous studies for several end points. The current study describes the development of a q-RASPR model using experimental retention time data (log t_R) in the HPLC experiments of 823 environmentally significant pesticide residues collected from a large compound database. To model the retention time (log t_R) end point, 0D−2D descriptors have been used along with the read-across-derived similarity descriptors. The developed partial least squares (PLS) model was rigorously validated by various internal and external validation metrics as recommended by the Organization for Economic Co-operation and Development (OECD). The final q-RASPR model is proven to be a good fit, robust, and externally predictive (n_{train} = 618, R^2 = 0.82, Q^2_{LOO} = 0.81, n_{test} = 205, and Q^2_{F1} = 0.84) that literally outperforms the external predictivity of the previously reported quantitative structure−property relationship (QSPR) model. From the insights of modeled descriptors, lipophilicity is found to be the most important chemical property, which positively correlates with the retention time (log t_R). Various other characteristics, such as the number of multiple bonds (nBM), graph density (GD), etc., have a substantial and inversely proportionate relationship with the retention time end point. The software tools utilized in this study are user-friendly, and most of them are free, which makes our methodology quite cost-effective when compared to experimentation. In a nutshell, to obtain better external predictivity, interpretability, and transferability, q-RASPR is an efficient technique that has the potential to be employed as a good alternative approach for retention time prediction and ecotoxicity potential identification.

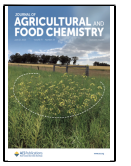
KEYWORDS: *q-RASPR, RASAR descriptors, retention time, pesticide residues, ecotoxicity*

1. INTRODUCTION

A pesticide is any compound or combination of substances intended to prevent, eliminate, repel, or control pests.[1] When pesticides are used on crops, residues might remain in food and enter the food chain. Pesticide residues have detrimental impacts on both humans and pollinators as a result of their toxicity and deleterious consequences. Pesticide mishandling has historically resulted in major difficulties for both human and animal health as well as the environment. Pesticide residues in food and the environment must be reduced, and more information about the risks of pesticides and their residues must be provided by regulators. In this regard, national and international authorities as well as environmental organizations have advocated for strict regulation of public health and environmental protection.[2] For the proper risk assessment of pesticides and their residues, experimental approaches, such as solvent extraction, high-performance liquid chromatography (HPLC), mass spectrometry, etc., are widely used. The pesticide extraction from food products (e.g., fruits, vegetables, grains, etc.) is regarded as an important stage in pesticide residue analysis because it serves as the foundation for pesticide detection at the trace level.[2] As a result of the intricacy of the process, various aspects must be optimized during pesticide extraction, which is time-consuming and labor-intensive. To avoid the associated complexities of experimental approaches, *in silico* new approach methodologies can be used.[3] The use of computational modeling, like quantitative structure−activity relationships (QSARs), read-across, machine learning models, etc., is recommended for the toxicity predictions of different chemicals against eco-toxicological end points by different chemical regulatory agencies.[4]

The quantitative structure−property relationship (QSPR) is an efficient mathematical tool that constructs a predictive quantitative relationship between a property (for example, retention time) for a group of molecules (pesticides) and the structural features of chemicals encoded by the molecular descriptors.[5] There are several ecotoxicological studies where

Received: March 7, 2023
Revised: April 29, 2023
Accepted: May 25, 2023
Published: June 9, 2023



QSAR/QSTRs have been successfully used for toxicity prediction.[6−9] The QSPR theory is used to supplement experimental findings from chemicals as well as to provide reliable predictions when experimental data are not available (data gap filling).[5] Read-across is another similarity-based predictive approach that uses simple algebraic calculations (unlike the QSPR, which uses different statistical algorithms for fitting the structural information of chemicals with the exerted properties by them) for obtaining the predictions of query chemicals from structural or functional analogues. It is also a widely used method in environmental toxicology and regulatory decision making. Read-across may be used to provide both quantitative and qualitative predictions, even when there is a dearth of experimental data. Scientists are also working on combinatorial modeling approaches. In 2018, Luechtefeld et al. reported a novel combinatorial modeling, namely, the read-across structure−activity relationship (RASAR), which uses similarity parameters obtained from fingerprints of chemicals by the conventional read-across algorithm to build supervised QSAR for the prediction of exerted activity/property/toxicity.[10] Enhanced accuracy and precision of predicted responses have been observed in the study, which indicates a significant enhancement of the predictability of RASAR over the previous modeling approaches. Recently, Banerjee and Roy have enriched the fundamental RASAR to a certain extent by proposing a quantitative RASAR method.[11] They combined the two-dimensional (2D) structural aspects of the chemicals with similarity and associated parameters obtained from their read-across prediction tool, for better predictability of the developed RASAR model. It is a simple, interpretable, and easily transferable technique. Unlike read-across, RASAR provides supervised mathematical models by which one can extract knowledge about the importance of modeled features and their contribution toward property prediction. RASAR can also be used for property predictions (read-across structure−property relationship or RASPR) from a comparatively small experimental data set (unlike QSPR, which requires sufficient experimental data for optimum performance) because it uses read-across-generated parameters (generally obtained from a small number of structural analogues). This is because RASAR uses composite similarity functions derived from multiple physicochemical features, which may behave like latent variables. Therefore, the advantages and efficiency of combinatorial modeling make them a promising new approach methodology in the field of chemoinformatics.

In the present paper, a q-RASPR modeling of retention time (log t_R) of pesticide residues has been reported. The retention time detected from reverse-phase HPLC can be considered a potential indicator of cellular toxicity. The retention time of chemicals depends upon their own lipophilicity, where the polarity of the mobile phase is constant.[12] Lipophilicity bears a directly proportional relationship with toxicity (more lipophilic chemicals can easily pass through the lipophilic plasma membrane and also stay in the lipid cells for a long time) as well.[13] Therefore, we can extrapolate the proportional relationship between toxicity and retention time also.[14] This assumption can be exemplified by the following observations: silafluofen, pyridalyl, and ivermectin (with t_R values of 11.05, 10.74, and 10.71 h, respectively, and X log P values of 8.2, 7.59, and 5.83, respectively). These compounds are reported to be very toxic to aquatic life, causing danger in specific target organ toxicity as well as reproductive toxicity. Again, compounds, like cyromazine, ethylene thiourea, and maleic hydrazide, have low t_R values (1.14, 1.08, and 1.06 h, respectively) and lower X log P values (−0.06, −0.66, and −0.84, respectively), and they are relatively less toxic than the previous compounds. We have attempted q-RASPR modeling in the present work for pesticide residues to detect toxicity levels in terms of retention time (log t_R) as a result of its simple, cost-effective, fast, and reproducible nature. No RASPR model has been reported in the literature for the prediction of the retention time of pesticides to the best of our knowledge. The entire computational research is devoid of animal experimentation, which complies with the 3R strategy (replacement, reduction, and refinement of animal experimentation) of the REACH legislation.[15]

2. MATERIALS AND METHODS

2.1. Data Set. The present work deals with the q-RASPR modeling of diverse classes of pesticides with a defined end point log t_R of 823 pesticides taken from the previous literature.[16] There are fair representations of cyclic (alicyclic and aromatic; homocyclic and heterocyclic structures) and acyclic (straight- and branched-chain structures) chemicals in the collected data set. The 2D structures of the aforesaid chemicals were retrieved (.sdf format) from the PubChem database[17] and were cross-verified with another popular chemical database ChemSpider.[18] After manual checking of all of the structures, chemical curation was performed using a chemical curation workflow in the KNIME platform[19] to remove inorganic salts, ions, radicals, repetitive compounds, and mixtures from the data set. However, all 823 compounds passed the curation, and the curated structures were selected for modeling in this study; the reported retention time (log t_R) data of curated 823 compounds has been represented in Table S1 of the Supporting Information.

2.1.1. Molecular Descriptor Calculation. Molecular descriptors are the mathematically encoded numerical representations of various structural and physicochemical features of chemical entities, which are used to correlate the structural aspects of chemicals with the exerted molecular properties (activity/property/toxicity) in QSAR studies.[20] There are various types of descriptors that are used to develop QSAR models. The 0D−7D descriptors can be used for modeling, and the level of complexity increases from 0D to 7D.[21] In this work, we have developed our models utilizing 0D−2D descriptors; they are extremely effective for avoiding the intricacy of alignment issues as well as the computational complexity of conformational analysis and energy minimization.[22] Apart from the exhaustive computation, the reproducibility and interpretability of higher order descriptors (3D and higher) are also limited, unlike 0D−2D descriptors. The computation of 0D−2D descriptors is fast and cost-effective. The 2D descriptors can effectively mark the structural features, and thus, mechanistic interpretation becomes easier. There are many previous reports, where the toxicity/activity/property have been efficiently correlated with the structural features by 0D−2D descriptors only.[8,9,23] The 2D structures of the molecules were drawn using Marvin Sketch software (version 14.10.27)[24] with proper aromatization and explicit hydrogen addition. In this study, nine classes of 2D descriptors, namely, electro-topochemical atom descriptors, molecular properties, constitutional index, functional group count, ring descriptors, atom-centered E-state indices, 2D atom pairs, connectivity index, and atom-centered fragments, have been computed by employing AlvaDesc software.[25] These descriptor classes were selected on the basis of their easy interpretability and usefulness in previous studies. After descriptor calculation, the intercorrelated descriptors (intercorrelation cutoff of >0.95) and the descriptors with the least variability in values (variance cutoff of <0.0001) were removed by the built-in pretreatment tool of AlvaDesc software.[25] As an additional measure, the flawed and extraneous data were eliminated by employing an in-house tool, namely, "DataPreTreatmentGUI1.2_3March2016" before using the data in the model

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

development, and a pool of 931 2D descriptors have been used in the preliminary modeling.

2.2. Data Set Division. Data set division is an essential part of QSAR model development. As a standard procedure, the data set is divided into training and test sets; the training set is solely employed for model development, whereas the test set is exclusively used for external validation of the developed model. In the present work, the sorted activity-based division technique has been used by employing the "datasetDivisionGUI1.2_19Feb2019" tool (accessible from http://dtclab.webs.com/software-tools) for 2D q-RASPR modeling. It is a response-based division technique, which divides a complete data set into clusters; the data points inside a cluster are similar to one another but different from those found in other clusters.[4] Although there is no stringent ratio for data set division, 70:30 to 80:20 proportions of training and test data points have been found in many previously published QSAR studies.[4] We chose around 25% of the data points from each cluster to serve as test set compounds (N_{test} = 205), with the remaining 75% to serve as the training set compounds (N_{train} = 618) after organizing the entire data set according to the cluster number and the corresponding activity levels. This method ensures a uniform representation from all of the clusters into training and test sets for feature selection and QSPR model development.

2.3. Feature Selection and Development of the QSPR Model. In this present study, we have developed a QSPR model using the retention time (log t_R) of pesticide residues as the response variable for the model development. As mentioned in the Molecular Descriptor Calculation section, the feature selection was started with 931 2D descriptors and the response variable log t_R. We have employed the genetic algorithm (GA) approach for feature selection. The GA technique was primarily applied by Rogers and Hopfinger[26] in QSAR investigations as a very effective tool with numerous advantages over other variable selection methods. It examines a variety of potential answers concurrently, each of which probes a unique area of the vector space delineated by computed descriptors. When QSPR modeling is performed, GA uses a fitness function based on mean absolute error (MAE)-based criteria to pick significant descriptors (variables). We have employed an in-house tool, namely, "GeneticAlgorithm_v4.1_Train" (available from https://dtclab.webs.com/software-tools) to identify the most correlated descriptors with the response variable. After identification of the important descriptors, we used the "partial least squares (PLS) regression" to build the preliminary QSPR models. The PLS regression approach is an extended version of the "multiple linear regression (MLR)" method that allows us to study strongly correlated, noisy, and collinear data as well as a large number of X variables.[27] The PLS regression has been carried out with a Java-based software tool "PLS_SingleY_version 1.0" (available from https://dtclab.webs.com/software-tools).

2.3.1. q-RASPR Model Development. The q-RASPR is a conjuncture of read-across and QSPR modeling where the read-across tool derived similarity, error, and concordance measures (RASAR descriptors) are used along with the previously identified descriptors to develop the QSAR-based final predictive models.[11,28] The q-RASPR model can be developed from both discrete and continuous data to form classification and regression-based models, respectively; however, only regression-based q-RASPR has been developed in this study. To calculate the RASAR descriptors, optimization of the read-across algorithm is a prior step. We have used the previous training and test sets with the identified 2D descriptors for determining the optimum read-across method. The training and test sets have been used as the source and target compounds, respectively, for read-across predictions. The "Read-across v4.1" tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) has been employed for the optimization, where Euclidean distance (ED)-based similarity, Gaussian kernel (GK) function similarity, and Laplacian kernel (LK) function similarity have been used for identifying the most similar compounds of the target chemical for further weighted average predictions.[21] The ED-, GK-, and LK-based similarity have been mathematically derived by following eqs 1, 2, and 3, respectively:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad (1)$$

where $d(X, Y)$ is the Euclidean distance between two compounds X and Y, X_i is the ith descriptor of compound X, and Y_i is the ith descriptor of compound Y, where i = 1, 2, 3, ..., n. The ED-based similarity is obtained by subtracting the scaled ED $[d(X, Y)]$ from 1

$$\mathrm{GK}(X, Y) = f = e^{-\|X_i - Y_i\|^2 / 2\sigma^2} \quad (2)$$

where GK(X, Y) is the Gaussian kernel function similarity (f) between two compounds X and Y, $\|X_i - Y_i\|^2$ is the Euclidean norm or L^2 norm, which can be measured by squaring the Euclidean distance, and σ determines the width of the Gaussian kernel function and can never become zero

$$\mathrm{LK}(X, Y) = \kappa = e^{(-\gamma \|X - Y\|_1)} \quad (3)$$

where LK(X, Y) is the Laplacian kernel function similarity (κ) between two compounds X and Y and γ is a non-zero and positive number, which determines the performance of the Laplacian kernel function. We have used the default setting, i.e., σ = 1, γ = 1, number of close training compounds = 10, distance threshold = 1, and similarity threshold = 0, for method optimization with subtraining and subtest sets (obtained after dividing the training set in a 75:25 ratio). As per the result of optimization (provided in Table S2 of the Supporting Information), Laplacian kernel function similarity read-across is the least error-prone method (as suggested by the least MAE value of 0.691) for this data set and has been used for RASAR descriptor calculation. Finally, the RASAR descriptors were calculated using the "RASAR descriptor v2.1" tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) after employing Laplacian kernel function similarity estimation as the optimized algorithm with γ = 1, number of closed training compounds = 10, and similarity threshold = 0 setting. The training and test sets have been used as the source and target compounds for the RASAR descriptor calculation of test set chemicals. However, the RASAR descriptor calculation of training set chemicals is a little bit tricky; the training set has been simultaneously used as the source and target compounds here. A "leave-same-out algorithm"[22] has been applied in the training RASAR descriptor calculation, where the same training chemical with a query compound is identified and removed before further procedures.

Finally, the pooled combination of RASAR descriptors and previously identified structural and physicochemical descriptors has been subjected to the best subset selection. On the basis of the cross-validated correlation coefficient (Q^2_{LOO}) of combinations, eight descriptors were selected and modeled using PLS regression with optimum latent variables. The final PLS q-RASPR model was thoroughly assessed by both internal and external validations;[21] the applicability domain[29] of the developed model has been analyzed; and all of the results have been discussed in the following Results and Discussion section. For a better understanding of the modeling, a schematic representation has been given in the following Figure 1.

3. RESULTS AND DISCUSSION

3.1. Developed QSPR Model for the Retention Time (Log t_R). A QSPR model has been developed in this study by employing PLS regression with five latent variables (LVs) from the identified eight different features by the best subset selection tool. The developed model has been represented in the following eq 4:

$$\begin{aligned} \log t_R = {} & 14.36 + 0.13\mathrm{O\text{-}060} + 0.43\mathrm{F05[N{-}S]} \\ & + 1.04\mathrm{LOGPcons} - 15.04\mathrm{GD} - 1.57\mathrm{nRNR2} \\ & - 0.089\mathrm{nBM} + 0.50\mathrm{B02[O{-}S]} - 7.47\mathrm{Me} \end{aligned} \quad (4)$$

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

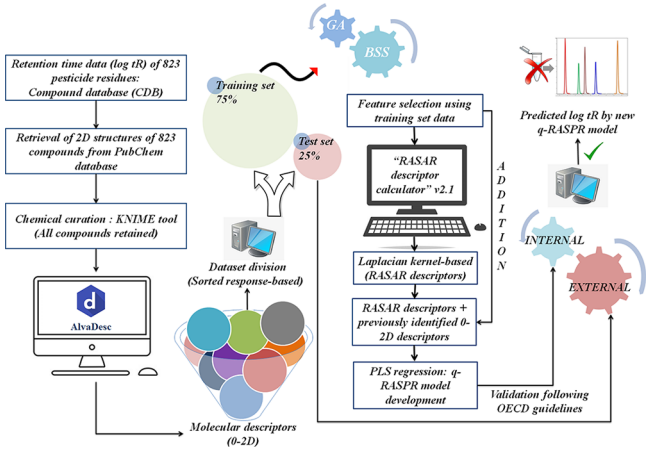


Figure 1. Flow diagram of q-RASPR model development (GA, genetic algorithm; BSS, best subset selection).

The performance of the model has been assessed using strict internal and external validation following the Organization for Economic Co-operation and Development (OECD) guidelines. Computed values of internal validation metrics, such as the determination coefficient (R^2 = 0.804), adjusted R^2 (R^2_{adj} = 0.789), leave-one-out cross-validated correlation coefficient (Q^2_{LOO} = 0.795), and r^2_m metrics of training set [$\overline{r^2}_m$(train) = 0.708 and Δr^2_m(train) = 0.162], signify the goodness of fit and robustness of the model, whereas the mean absolute error of training set predictions (MAE_{train} = 0.611) points out the error of predictions. Along with these, the external validation metrics, i.e., external predicted variance (Q^2_{F1} = 0.82 and Q^2_{F2} = 0.82), mean absolute error of test set predictions (MAE_{test} = 0.67), $\overline{r^2}_m$(test) = 0.72, Δr^2_m(test) = 0.15, concordance correlation coefficient (CCC = 0.90), etc., have also been calculated, which are the classic markers of good external predictability. The identified structural descriptors have been tabulated in the following Table 1 along with the definition and class.

3.2. Developed q-RASPR Model. Our objective was to develop a q-RASPR model and enhance the external predictability of the corresponding QSPR model further in this study. Thus, we combined the pool of eight structural descriptors along with the computed read-across-based RASAR descriptors and finally obtained the new combination of descriptors using the best subset selection method. The finally selected combined descriptors have been given in Table S3 of the Supporting Information. PLS regression has been utilized to model the final descriptors with the optimum number of latent variables (5 LVs), and the developed PLS q-RASPR model has been shown in the following eq 5:

$$\begin{aligned}\log t_R = {} & 6.90771 - 3.91482\mathrm{CVact(LK)} \\ & + 0.73673\mathrm{Pos.Avg.Sim} + 0.30878\mathrm{F05[N{-}S]} \\ & + 0.88941\mathrm{LOGPcons} - 10.04724\mathrm{GD} \\ & - 0.86658\mathrm{nRNR2} - 0.07316\mathrm{nBM} \\ & + 0.53026\mathrm{B02[O{-}S]} \end{aligned} \quad (5)$$

This final model has been rigorously validated employing several internal and external validation metrics[21] following the OECD guidelines, and the computed metrics have been given in the following Table 2.

Table 2. Computed Validation Metrics of the PLS q-RASPR Model

Internal Validation Metrics	
R^2	0.81
R^2_{adj}	0.81
Q^2_{LOO}	0.81
MAE_{train}	0.59
$\overline{r^2}_m$(train)	0.73
Δr^2_m(train)	0.15
External Validation Metrics	
Q^2_{F1}	0.84
Q^2_{F2}	0.84
$\overline{r^2}_m$(test)	0.75
Δr^2_m(test)	0.14
MAE_{test}	0.63

The acceptable range of internal validation metrics, like R^2, R^2_{adj}, Q^2_{LOO}, and r^2_m metrics [$\overline{r^2}_m$(train) and Δr^2_m(train)][30] justify the model as a robust and good fit, whereas the performance of the model on external compounds has been justified by the acceptable range of computed external validation metric values (Q^2_{F1} and Q^2_{F2}). Apart from the conventional validation procedure, we have performed the *Y*-randomization of the model with 100 times random shuffling of response values ($\log t_R$).[31,32] The *Y*-randomization plot has been shown in the following Figure 2A, where the Q^2 and R^2 values of the original model have been plotted against 100 new Q^2 and R^2 values of the randomly developed model. The intercepts of trend lines obtained by R^2 and Q^2 plots (R^2_Y and Q^2_Y) have been calculated, and these are −0.012 and 0.0946, respectively. Both of these values are less than the threshold limits (R^2_Y = 0.3 and Q^2_Y = 0.05), which confirm that the model was not the result of any chance correlation. An

Table 1. Identified Structural Descriptors with Definition and Class

descriptor	class	definition
O-060	atom-centered fragments	it indicates Al−O−Ar/Ar−O−Ar/R···O···R/R−O−C═X types of atom-centered fragments
F05[N−S]	2D atom pairs	it indicates the frequency of the N−S atom pair at topological distance 5
LOGPcons	molecular properties	it indicates the octanol−water partition coefficient (log *P*)
GD	constitutional indices	it is termed as the graph density, expressed by GD = 2nBo/nSK(nSK − 1), where nBo is the number of edges in the H-suppressed graph and nSK is the number of vertices in the H-suppressed graph
nRNR2	functional group counts	it indicates the number of tertiary amines (aliphatic)
nBM	constitutional indices	it indicates the number of multiple bonds present in a molecule

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

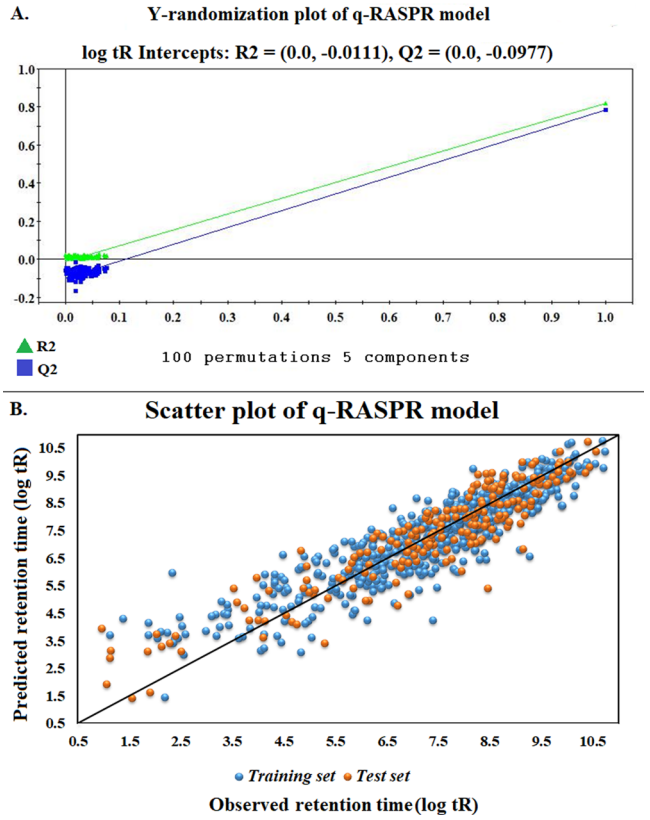


Figure 2. (A) *Y*-randomization plot of the q-RASPR model and (B) scatter plot of the q-RASPR model.

observed versus predicted scatter plot has also been produced with training and test set data and has been given in Figure 2B. The uniform scattering and equal distribution of training and test data points surrounded by the trend line indicate the goodness of fit of the generated model as well as how well the observed values correspond with the predicted values.

3.3. PLS Plots. *3.3.1. Coefficient Plot.* The descriptors appearing in the final q-RASPR model can be grouped into two classes depending upon their regression coefficient values. The modeled descriptors are plotted against their corresponding regression coefficient in the *Y* axis to obtain the coefficient plot.[33] Descriptors with positive regression coefficients are represented as upward columns, whereas the descriptors with negative regression coefficients are drawn as downward columns from the *X* axis. Here, in this study, the coefficient plot has been obtained using SIMCA-P software and has been presented in the following Figure 3A. According to the coefficient plot, LOGPcons, F05[N−S], B02[O−S], and Pos.Avg.Sim are positively contributing descriptors, whereas GD, nBM, nRNR2, and CVactivity(LK) are negatively contributing features of the developed model.

3.3.2. Variable Importance Plot (VIP). The VIP defines the relative importance of the modeled descriptors toward the end point (log t_R) predictions by a particular regression model.[34] It is essentially a column plot where the VIP scores of each descriptor are plotted on the *Y* axis and the modeled descriptors are plotted on the *X* axis in descending order of importance from left to right. As per the standard practice, the descriptors with a VIP score of >1 are considered to be the most important for end point prediction. We have generated

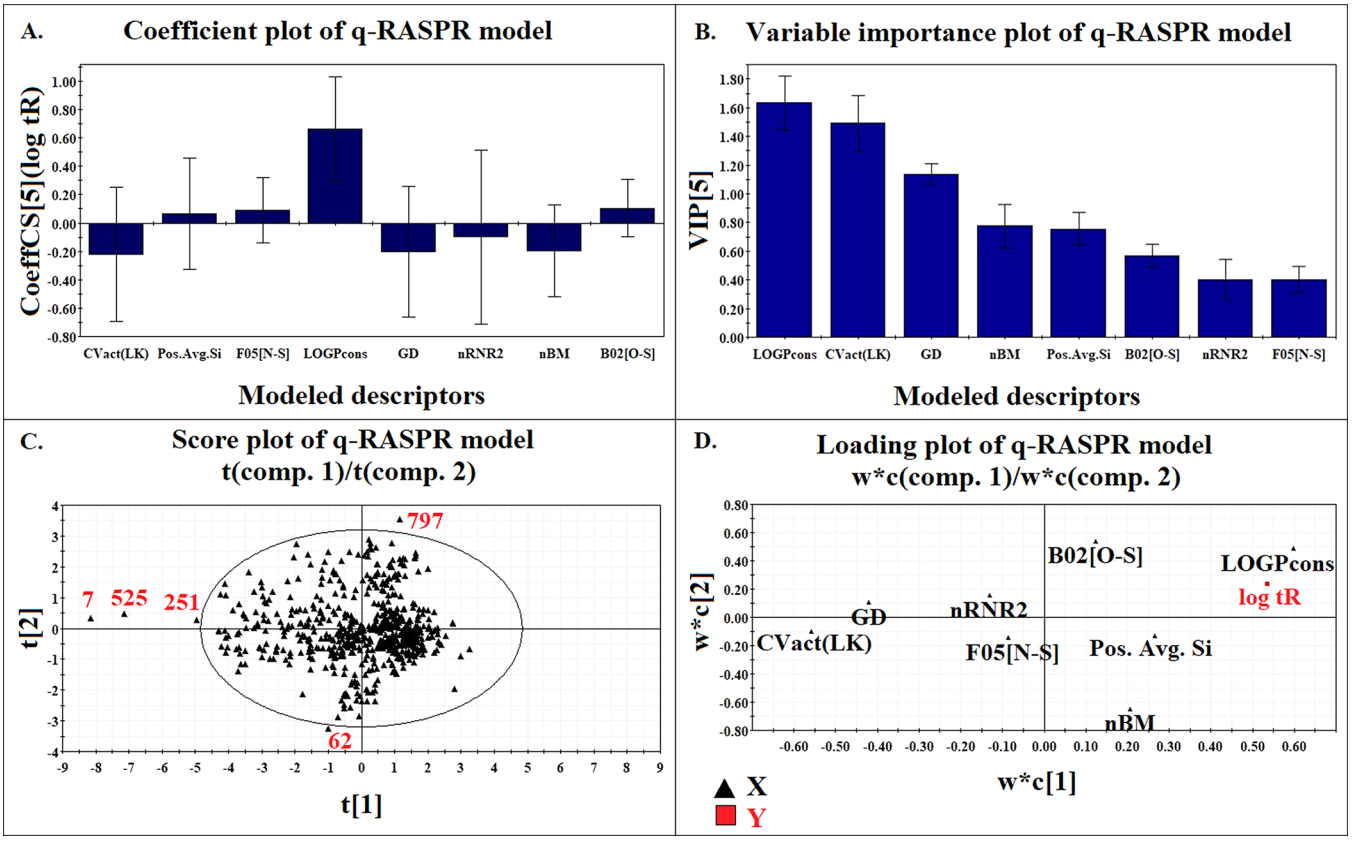


Figure 3. (A) Coefficient plot of the q-RASPR model, (B) variable importance plot (VIP) of the q-RASPR model, (C) score plot of the q-RASPR model, and (D) loading plot of the q-RASPR model.

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

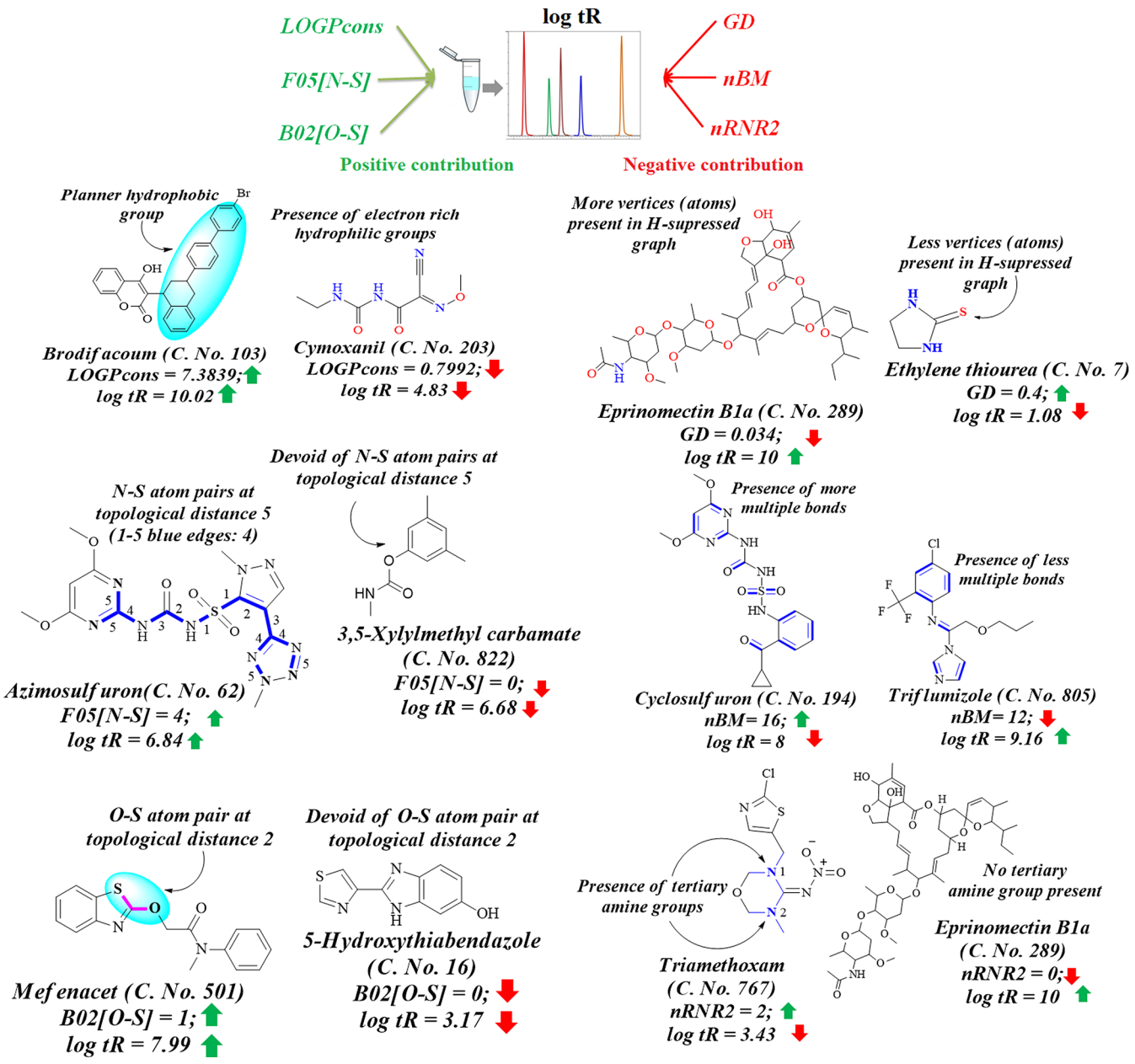


Figure 4. Insights of identified 0D−2D descriptors used in the q-RASPR model.

the VIP of developed q-RASPR by utilizing SIMCA-P software, which has been showcased in the following Figure 3B. According to the VIP, the order of the relative importance of contributing descriptors toward the response values appearing in the model is LOGPcons > CVactivity(LK) > GD > nBM > Pos.Avg.Sim > B02[O−S] > nRNR2 > F05[N−S]. LOGPcons, CVactivity(LK), and GD are the most important descriptors, which impact the end point prediction to a great extent. Apart from these, nBM and Pos.Avg.Sim are the descriptors having moderate importance toward the response (log t_R) prediction because their VIP scores are close to 1, and they are located in the middle region of the plot. B02[O−S], nRNR2, and F05[N−S] have the least importance because they have very low VIP scores (<1) and are located far right in the plot.

3.3.3. Score Plot. A score plot depicts the distribution of compounds in the latent variable space (up to 2 LVs) as determined by the scores.[35] The score plot of the first two components $t[1]$ and $t[2]$ has been shown in Figure 3C. The ellipse represents the applicability domain of the model, as described by Hotelling's t^2. Hotelling's t^2 is a multivariate version of the t test developed by Harold Hotelling and serves as a screening for compounds that adhere to multivariate normality.[36] Hotelling's t^2 for observation "i" based on "a" components is $t_i^2 = \sum((t_{ia} - t_{avg})^2/S_{ta}^2)$. Here, the summation is done over the range of the selected components, S_{ta}^2 = variance of t scores for ath component according to the class model. Hotelling's t^2 should be less than the critical value $F(p = 0.05, df = a, n - a)\ a(n - 1)/(n - a)$ (n = number of observations in the data set).[36] In this plot, compounds that are located near each other have similar kind of characteristics or properties, whereas compounds that are far from each other have dissimilar properties with respect to the retention time (log t_R) of pesticide residues. Compounds that are close to the center possess average properties, and we can conclude that compounds **7**, **62**, **251**, **525**, and **797** are identified as outliers

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

according to the plot because these compounds are situated at the outside of the ellipse. These compounds have Hotelling t^2 values of 26.2, 10.0, 9.76, 20.2, and 11.8 (on the basis of two components), all of which are higher than the critical value of 6.0, considering the F value (p = 0.05) of 3.0 at df of 2 and 616.

3.3.4. Loading Plot. The loading plot of this q-RASPR has also been generated with SIMCA-P software. This plot analyses the loading of both descriptors and response of a regression model considering up to the first two latent variable spaces; therefore, the knowledge of the interrelationship between X (descriptor) and Y (response) variables can be obtained from the loading plot.[37] The distance of the X variables from the origin correlates with the relevance of the descriptor. As per the loading plot given in Figure 3D, we can see that LOGPcons, Pos.Avg.Sim, CVactivity(LK), and GD descriptors provided the greatest impact to predict the end point (log t_R) because these are located far away from the origin, and nBM, B02[O−S], nRNR2, and F05[N−S] descriptors displayed less impact on the prediction of the end point (log t_R) because these are placed near the origin. It also can be inferred from the loading plot that LOGPcons is the most correlated descriptor with the response variable log t_R, and this can be justified by the VIP also.

3.4. Insights on the Modeled Descriptors. *3.4.1. LOGPcons Descriptor.* LOGPcons (computed octanol−water partition coefficient) is the most important descriptor of the final q-RASPR model as seen in the VIP, as a molecular property descriptor, which essentially transfers the knowledge of the hydrophobicity of molecules. As observed in the coefficient plot, LOGPcons is a positively contributing descriptor, which has a linear correlation with the retention time end point (log t_R). In the data set, a compound like brodifacoum (compound **103**) with a large planer hydrophobic moiety in its structure has larger LOGPcons (LOGPcons = 7.3839) and higher retention time (log t_R = 10.02) than a small hydrophilic compound, like cymoxanil (compound **203**, LOGPcons = −0.7992, and log t_R = 4.83), as observed in Figure 4. Hydrophobicity of chemicals potentiates cellular toxicity as a result of the ease of passing a lipid cell membrane.[38] Therefore, these observations justify the linear correlation between the retention time (log t_R) and hydrophobicity and, thus, cellular toxicity also.

3.4.2. CVactivity(LK) Descriptor. CVactivity(LK) is the second most important and the first RASAR descriptor of this q-RASPR model that describes the coefficient of variation of the observed responses of selected close source compounds for each query compound. It is a negatively contributing descriptor of the developed q-RASPR model, as shown in the coefficient plot (Figure 3A). It has been observed that the compound with a lower retention time value has more CVactivity(LK) [compound **525**, log t_R = 2.19, and CVactivity(LK) = 0.6669], whereas a compound with more retention time has a lower CVactivity(LK) value [compound **602**, log t_R = 10.36, and CVactivity(LK) = 0.02569] value, which justifies the negative contribution of CVactivity(LK) toward model end point prediction.

3.4.3. GD Descriptor. In the final q-RASPR model, the third most contributing descriptor toward retention time (log t_R) is GD. It is a constitutional descriptor that is termed as graph density. It is mathematically derived by the following equation 6:

$$GD = 2nBo/nSK(nSK - 1) \quad (6)$$

where nBo is the number of edges in the H-suppressed graph and nSK is the number of vertices in the H-suppressed graph.[39] We have shown two different compounds, namely, ethylene thiourea (compound **7**) and eprinomectin B1a (compound **289**), in Figure 4. Compound **289** contains a higher number of vertices than compound **7**, and thus, the nSK(nSK − 1) portion of the equation has a greater value, which decreases the resultant value of GD in compound **289**. As seen in the regression coefficient plot, GD has a negative contribution toward end point prediction. It is observed in the depicted compounds also; the retention time of compound **289** (GD = 0.034) is significantly higher (log tR = 10) than that of compound **7** (log t_R = 1.08) with less GD value.

3.4.4. nBM Descriptor. In the resultant model, the fourth most contributing descriptor toward retention time (log t_R) is the constitutional descriptor nBM, which indicates the number of multiple bonds present in a molecule. As shown by the regression coefficient plot, it has a negative impact on the end point value, which indicates a decrease in the retention time (log t_R) value with an increase in the descriptor value and vice versa. For instance, compound **194** in Figure 4 has 16 multiple bonds and less retention time (log t_R = 8.00) than compound **805** with 12 multiple bonds (log t_R = 9.16).

3.4.5. Pos.Avg.Sim Descriptor. It is the second RASAR descriptor of the q-RASPR model, which measures the average similarity of positive close source compounds (with respect to the mean of the source compounds) for a query compound. It shows a positive regression coefficient in the model equation, which indicates a linear relationship between Pos.Avg.Sim and the modeled response and can be justified by the observations in our data set. Compound **802** has more retention time (log t_R = 8.17) as well as Pos.Avg.Sim (0.7190) values, whereas compound **702** has less retention time (log t_R = 4.48) and Pos.Avg.Sim (0.0) values.

3.4.6. B02[O−S] Descriptor. B02[O−S] is a 2D atom pair descriptor, which indicates the presence or absence of the N−S atom pair at the topological distance 2. As shown in the regression coefficient plot, B02[O−S] has a positive regression coefficient; i.e., it linearly correlates with the retention time (log t_R). For instance, compound **501** possesses a O−S atom pair at the topological distance 2 in its structure, as depicted in Figure 4, and has a greater retention time (log t_R = 7.99) than that of compound **16** (log t_R = 3.17), which does not contain the O−S atom pair at the topological distance 2.

3.4.7. nRNR2 Descriptor. nRNR2 is a functional group count descriptor, which indicates the number of tertiary aliphatic amine groups in a compound. It shows a lower VIP score and, thus, possesses a lower influence toward the end point prediction. The negative regression coefficient indicates that this descriptor is inversely correlated with the response values. If we compare two training compounds, namely, triamethoxam (compound **767**) and eprinomectin B1a (compound **289**) (Figure 4), it can be observed that a higher number of tertiary amine group containing compound triamethoxam (nRNR2 = 2) has a lower retention time (log t_R = 3.43) than eprinomectin B1a, which is devoid of tertiary amine (nRNR2 = 0 and log t_R = 10). Therefore, these observations justify the inverse relationship between nRNR2 and the end point.

3.4.8. F05[N−S] Descriptor. This is a 2D atom pair descriptor, indicating the frequency of N−S atom pairs at

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

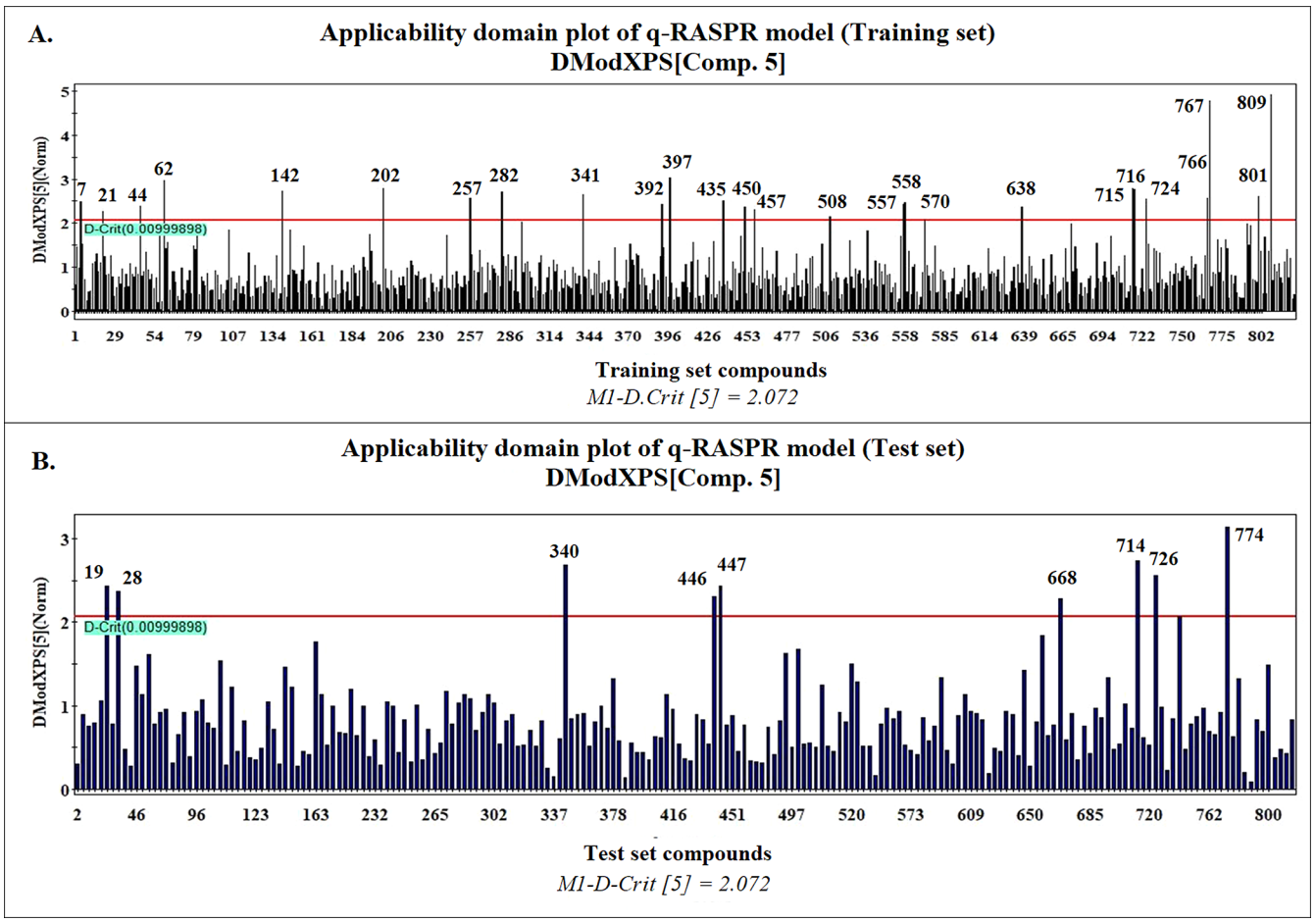


Figure 5. (A) AD plot for the test set and (B) AD plot for the training set.

Table 3. Validation Metrics of the Developed QSPR, q-RASPR, and Previously Reported MLR Models

			training set statistics		test set statistics		
work done by	type of models	number of data points	R^2	Q^2_{LOO}	Q^2_{F1}	MAE_{test}	number of descriptors
Rojas et al.	MLR	nTrain = 275, nVal = 275, and nTest = 273	0.87	0.86	0.75	0.62	5
present study	QSPR model (PLS)	nTrain = 618 and nTest = 205	0.80	0.80	0.82	0.67	6 (5 LVs)
	q-RASPR model (PLS)	nTrain = 618 and nTest = 205	0.82	0.81	0.84	0.63	8 (5 LVs)
	q-RASPR model (PLS)[a]	nTrain = (275 + 275) = 550 and nTest = 273	0.82	0.81	0.82	0.60	8 (5 LVs)

[a]q-RASPR developed with the training set (training + validation compounds of the previous work by Rojas et al.41). LVs = latent variables.

the topological distance 5, as illustrated in Figure 4. According to the regression coefficient plot, F05[N−S] linearly correlates with the retention time (log t_R); i.e., if the number or frequency of N−S at the topological distance 5 in a molecule increases, the response value (log t_R) will also increase for that molecule and vice versa. For instance, compound **62** with a higher number of N−S atom pairs at the topological distance 5 (F05[N−S] = 4) has a comparatively higher retention time (log t_R = 6.84) than compound **822** with no N−S atom pair (F05[N−S] = 0) at the topological distance 5, having a low retention time (log t_R = 6.68) (Figure 4). This implies that the presence of more fragments containing N−S atom pairs at the topological distance 5 increases the polar surface area, as observed in Figure 4. Thus, the polar surface area is an important factor to control lipophilicity as well as toxicity in terms of the retention time (log t_R) as seen in the work of Yukawa and Naven,[40] which justifies our observations.

3.5. Applicability Domain Analysis of the PLS q-RASPR Model. The applicability domain is the theoretical region in descriptor space in which the prediction obtained by the model is reliable.[29] The model's applicability domain (AD) was evaluated using the DModX (distance to model in *X* space) approach in the SIMCA-P software at the 99% confidence level with a *D* critical value of 0.009999. The AD plots of training and test sets and the corresponding outliers have been shown in panels A and B of Figure 5, respectively. We found 2.59% of training data (26 outliers; compounds **7**, **21**, **44**, **62**, **142**, **202**, **257**, **282**, **341**, **392**, **397**, **435**, **450**, **457**, **508**, **557**, **558**, **570**, **638**, **715**, **716**, **724**, **766**, **767**, **801**, and **809**) and 4.39% of test data (9 outliers; compounds **19**, **28**, **340**, **446**, **447**, **668**, **714**, **726**, and **774**) as outliers for the developed q-RASPR model. Considering the variety of molecules, the identification of any common structural traits for the outliers was difficult. However, from a statistical

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

standpoint, we can approximate that the standard deviation of X residuals is higher for outliers than for molecules lying within the applicability domain of the model.

3.6. Comparison of the Developed QSPR and q-RASPR Models to a Previously Published Work. In this present study, both QSPR and q-RASPR models have been developed for the prediction of the retention time (log t_R) of pesticide residues in the HPLC analysis. Here, we have extrapolated the knowledge of retention time toward the ecotoxicity potential of pesticide residues because of their proportional relationships with lipophilicity. The quality of the q-RASPR model in terms of goodness of fit, robustness, and external predictivity is better than the QSPR model as suggested by the tabulated validation metrics (in Table 3) of both of the models.

Previously, Rojas et al.[41] reported a food-informatic model, where they took 275 data points as the training and validation sets each and 273 data points as the test set. They employed five conformation-independent descriptors to correlate with the retention time end point using the MLR algorithm. The reported internal and external validation metrics of the previous model have been tabulated in Table 3. The previous model was developed with a remarkably lower number of training set compounds (nTrain = 275) in comparison to the present q-RASPR model (nTrain = 618), and thus, the previous model had somewhat better R^2 (0.87) and Q^2_{LOO} (0.86) values. However, both the QSPR and q-RASPR models have better external predictivity in terms of Q^2_{F1} (QSPR = 0.82 and q-RASPR = 0.84) than the previous MLR (Q^2_{F1} of previous MLR = 0.75). For a proper comparison of external predictivity between the two works, we have repeated the q-RASPR by taking the combination of the reported training and validation set compounds (from the work by Rojas et al.[41]) as the training set. The same set of descriptors and the same number of LVs (as used in the developed q-RASPR model) have been used for the prediction of the retention time (log t_R) of test set compounds (same as the previous work) by the q-RASPR model. It has been observed from the computed Q^2_{F1} (q-RASPR* = 0.82 and previous MLR = 0.75) and MAE_{test} (q-RASPR* = 0.60 and previous MLR = 0.62) values in Table 3 that the external predictivity and error of predictions of the q-RASPR model have held a better place than the previous model.

3.7. Overview. In this present study, a q-RASPR model has been developed with the experimental retention time data (log t_R) of 823 environmentally relevant pesticide residues (identified in foods and vegetable products), collected from the compound database (CDB). The retention time obtained from reverse-phase HPLC analysis has been used as the end point of this study as a result of its potential to be a toxicity indicator. This study was based on simple 2D descriptors to avoid complexity related to conformational analysis and energy minimization. The RASAR descriptors have been calculated from the 2D descriptors identified by the developed QSPR model in this study. The model was rigorously validated by various internal and external validation metrics, as recommended by the OECD. Additionally, the domain of applicability of the developed q-RASPR model has also been identified and reported. In a comparative analysis of the q-RASPR with the previously published work, the enhancement of external predictivity and reduction of prediction errors have been marked [Q^2_{F1}: QSPR = 0.82, q-RASPR = 0.84, q-RASPR* = 0.82 (same composition of training and test sets as used previously), and previous MLR = 0.75; MAE_{test}: QSPR = 0.67, q-RASPR = 0.63, q-RASPR* = 0.60, and previous MLR = 0.62]; the external predictivity in terms of the Q^2_{F1} value of the q-RASPR model superseded the quality of the previously reported MLR model. From the descriptor interpretation, it has been observed that lipophilicity is the most important chemical property, which positively affects the retention time and, as a result, cellular toxicity. Besides this, several other features, like the number of multiple bonds (nBM), graph density (GD), etc., have significant and inversely proportional relationships with the retention time end point. The software tools used in this study are easy to use and fast, and the majority of the tools are freely available, which makes our workflow quite economical compared to experimentation. The q-RASPR is a more efficient technique as a result of its better external predictivity, interpretability, and transferability;[42] therefore, it has the potential to be used as a good alternative approach of retention time prediction and toxicity identification.

■ ASSOCIATED CONTENT

Data Availability Statement

The DTC Lab software tools are available from http://teqip.jdvu.ac.in/QSAR_Tools/ (MLR BestSubsetSelection and MLR plus Validation) and https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home (Quantitative Read-Across, version 4.1, and RASAR Descriptor Calculator, version 2.0).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jafc.3c01438.

Raw data in Excel format for the list of compounds (Table S1), optimization of RASAR descriptors (Table S2), and descriptor and end point values (Table S3) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Kunal Roy − *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India;* orcid.org/0000-0003-4486-8074; Email: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in

Authors

Shilpayan Ghosh − *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India*

Mainak Chatterjee − *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jafc.3c01438

Author Contributions

Shilpayan Ghosh, computation, validation, and initial draft; Mainak Chatterjee, validation and editing; and Kunal Roy, conceptualization, supervision, and editing. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

Funding

Shilpayan Ghosh and Mainak Chatterjee sincerely acknowledge All India Council for Technical Education (AICTE), New Delhi, India, for their financial support as a form of Post-Graduation (PG) Fellowship and National Doctoral Fellowship (NDF), respectively.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Al-Saleh, I. Pesticides: A Review Article. *J. Environ. Pathol. Toxicol. Oncol.* **1994**, *13*, 151−161.
(2) Villaverde, J. J.; López-Goti, C.; Alcamí, M.; Lamsabhi, A. M.; Alonso-Prados, J. L.; Sandín-España, P. Quantum Chemistry in Environmental Pesticide Risk Assessment. *Pest Manage. Sci.* **2017**, *73*, 2199−2202.
(3) *Current Trends in Computational Modeling in Drug Discovery*; Kar, S., Leszczynski, J., Eds.; Springer: Cham, Switzerland, 2023.
(4) *Quantitative Structure−Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment*; Roy, K., Ed.; IGI Global: Hershey, PA, 2015; DOI: 10.4018/978-1-4666-8136-1.
(5) Kumar, P.; Kumar, A. Monte Carlo Method Based QSAR Studies of Mer Kinase Inhibitors in Compliance with OECD Principles. *Drug Res.* **2018**, *68*, 189−195.
(6) Gbeddy, G.; Egodawatta, P.; Goonetilleke, A.; Ayoko, G.; Chen, L. Application of Quantitative Structure-Activity Relationship (QSAR) Model in Comprehensive Human Health Risk Assessment of PAHs, and Alkyl-, Nitro-, Carbonyl-, and Hydroxyl-PAHs Laden in Urban Road Dust. *J. Hazard. Mater.* **2020**, *383*, 121154.
(7) Jia, Q.; Wang, S.; Yu, M.; Wang, Q.; Yan, F. Two QSAR Models for Predicting the Toxicity of Chemicals towards Tetrahymena Pyriformis Based on Topological-Norm Descriptors and Spatial-Norm Descriptors. *SAR QSAR Environ. Res.* **2023**, *34*, 147−161.
(8) Seth, A.; Ojha, P. K.; Roy, K. QSAR Modeling with ETA Indices for Cytotoxicity and Enzymatic Activity of Diverse Chemicals. *J. Hazard. Mater.* **2020**, *394*, 122498.
(9) Chatterjee, M.; Roy, K. Prediction of Aquatic Toxicity of Chemical Mixtures by the QSAR Approach Using 2D Structural Descriptors. *J. Hazard. Mater.* **2021**, *408*, 124936.
(10) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine Learning of Toxicological Big Data Enables Read-across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicol. Sci.* **2018**, *165*, 198−212.
(11) Banerjee, A.; Roy, K. First Report of Q-RASAR Modeling toward an Approach of Easy Interpretability and Efficient Transferability. *Mol. Diversity* **2022**, *26*, 2847−2862.
(12) Soares, J. X.; Santos, Á.; Fernandes, C.; Pinto, M. M. M. Liquid Chromatography on the Different Methods for the Determination of Lipophilicity: An Essential Analytical Tool in Medicinal Chemistry. *Chemosensors* **2022**, *10*, 340.
(13) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2009**, *49*, 1−5.
(14) Hongmao, S. Quantitative Structure−Property Relationships Models for Lipophilicity and Aqueous Solubility. In *A Practical Guide to Rational Drug Design*; Woodhead Publishing: Cambridge, U.K., 2016; Chapter 6, pp 193−223, DOI: 10.1016/B978-0-08-100098-4.00006-5.
(15) Gorzalczany, S. B.; Rodriguez Basso, A. G. Strategies to Apply 3Rs in Preclinical Testing. *Pharmacol. Res. Perspect.* **2021**, *9* (5), No. e00863.
(16) Wang, J.; Chow, W.; Wong, J. W.; Leung, D.; Chang, J.; Li, M. Non-Target Data Acquisition for Target Analysis (NDATA) of 845 Pesticide Residues in Fruits and Vegetables Using UHPLC/ESI Q-Orbitrap. *Anal. Bioanal. Chem.* **2019**, *411*, 1421−1431.
(17) *PubChem*; https://pubchem.ncbi.nlm.nih.gov/ (accessed Feb 27, 2023).
(18) *ChemSpider*; http://www.chemspider.com/ (accessed Feb 27, 2023).
(19) *Data Analytics Platform: Open Source Software Tools|KNIME*; https://www.knime.com/knime-analytics-platform (accessed Feb 27, 2023).
(20) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000; DOI: 10.1002/9783527613106.
(21) Roy, K.; Kar, S.; Das, R. N. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; Academic Press: Cambridge, MA, 2015; DOI: 10.1016/C2014-0-00286-9.
(22) Chatterjee, M.; Roy, K. Chemical Similarity and Machine Learning-Based Approaches for the Prediction of Aquatic Toxicity of Binary and Multicomponent Pharmaceutical and Pesticide Mixtures against Aliivibrio Fischeri. *Chemosphere* **2022**, *308*, 136463.
(23) Hoover, G.; Kar, S.; Guffey, S.; Leszczynski, J.; Sepúlveda, M. S. In Vitro and in Silico Modeling of per Fluoroalkyl Substances Mixture Toxicity in an Amphibian Fi Broblast Cell Line. *Chemosphere* **2019**, *233*, 25−33.
(24) Chemaxon. *Marvin*; https://chemaxon.com/marvin (accessed Feb 27, 2023).
(25) Mauri, A. AlvaDesc: A Tool To Calculate and Analyze Molecular Descriptors and Fingerprints. In *Methods in Pharmacology and Toxicology*; Roy, K., Ed.; Humana Press, Inc.: New York, 2020; pp 801−820, DOI: 10.1007/978-1-0716-0150-1_32.
(26) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(27) dos Santos, G. L. A. A.; Reis, A. S.; Besen, M. R.; Furlanetto, R. H.; Rodrigues, M.; Crusiol, L. G. T.; de Oliveira, K. M.; Falcioni, R.; de Oliveira, R. B.; Batista, M. A.; Nanni, M. R. Spectral Method for Macro and Micronutrient Prediction in Soybean Leaves Using Interval Partial Least Squares Regression. *Eur. J. Agron.* **2023**, *143*, 126717.
(28) Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. A Novel Quantitative Read-across Tool Designed Purposefully to Fill the Existing Gaps in Nanosafety Data. *Environ. Sci. Nano* **2022**, *9*, 189−203.
(29) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSARs Models: Where Theory Meets Reality. *Int. J. Quant. Struct. Prop. Relat.* **2016**, *1*, 45−63.
(30) Roy, K.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N.; Kabir, H. Introduction of Rm2(Rank) Metric Incorporating Rank-Order Predictions as an Additional Tool for Validation of QSAR/QSPR Models. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 200−210.
(31) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(32) Chatterjee, M.; Roy, K. Application of Cross-Validation Strategies to Avoid Overestimation of Performance of 2D-QSAR Models for the Prediction of Aquatic Toxicity of Chemical Mixtures. *SAR QSAR Environ. Res.* **2022**, *33*, 463−484.
(33) Davrieux, F.; Dufour, D.; Dardenne, P.; Belalcazar, J.; Pizarro, M.; Luna, J.; Londono, L.; Jaramillo, A.; Sanchez, T.; Morante, N.; Calle, F.; Lopez-Lavalle, L. A. B.; Ceballos, H. LOCAL Regression Algorithm Improves near Infrared Spectroscopy Predictions When the Target Constituent Evolves in Breeding Populations. *J. Near Infrared Spectrosc.* **2016**, *24*, 109−117.
(34) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(35) De, P.; Aher, R. B.; Roy, K. Chemometric Modeling of Larvicidal Activity of Plant Derived Compounds against Zika Virus Vector Aedes Aegypti: Application of ETA Indices. *RSC A* **2018**, *8*, 4662−4670.
(36) Hotelling, H. The Generalization of Student's Ratio. *Ann. Math. Stat.* **1931**, *2*, 360−378.

(37) Trygg, J.; Wold, S. Orthogonal Projections to Latent Structures (O-PLS). *J. Chemom.* **2002**, *16*, 119−128.
(38) Vesterkvist, P. S. M.; Misiorek, J. O.; Spoof, L. E. M.; Toivola, D. M.; Meriluoto, J. A. O. Comparative Cellular Toxicity of Hydrophilic and Hydrophobic Microcystins on Caco-2 Cells. *Toxins* **2012**, *4*, 1008−1023.
(39) Krishna, J. G.; Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. Chemometric Modeling of Power Conversion Efficiency of Organic Dyes in Dye Sensitized Solar Cells for the Future Renewable Energy. *Nano Energy* **2020**, *70*, 104537.
(40) Yukawa, T.; Naven, R. Utility of Physicochemical Properties for the Prediction of Toxicological Outcomes: Takeda Perspective. *ACS Med. Chem. Lett.* **2020**, *11*, 203−209.
(41) Rojas, C.; Aranda, J. F.; Pacheco Jaramillo, E.; Losilla, I.; Tripaldi, P.; Duchowicz, P. R.; Castro, E. A. Foodinformatic Prediction of the Retention Time of Pesticide Residues Detected in Fruits and Vegetables Using UHPLC/ESI Q-Orbitrap. *Food Chem.* **2021**, *342*, 128354.
(42) Banerjee, A.; Roy, K. On Some Novel Similarity-Based Functions Used in the ML-Based q-RASAR Approach for Efficient Quantitative Predictions of Selected Toxicity End Points. *Chem. Res. Toxicol.* **2023**, *36*, 446−464.

https://doi.org/10.1021/acs.jafc.3c01438
J. Agric. Food Chem. 2023, 71, 9538−9548

Reg id: 1441
Poster id ADD5

Predictive q-RASPR modelling of the retention time of pesticide residues in foods and vegetables measured by the Hypersil Gold stationary phase

A) Shilpayan Ghosh
B) Arkaprava Banerjee
C) Mainak Chatterjee
D) Kunal Roy*
A) Drug Theoretics & Cheminformatics Laboratory,Department of Pharmaceutical Technology,Jadavpur University ,Kolkata 700032, India
B) Drug Theoretics & Cheminformatics Laboratory,Department of Pharmaceutical Technology,Jadavpur University ,Kolkata 700032, India
C) Drug Theoretics & Cheminformatics Laboratory,Department of Pharmaceutical Technology,Jadavpur University ,Kolkata 700032, India
D) Drug Theoretics & Cheminformatics Laboratory,Department of Pharmaceutical Technology,Jadavpur University ,Kolkata 700032, India

This study describes the development of quantitative read-across structure-property relationship (q-RASPR) models to predict the retention time (tR) of a large compound database (CDB) of pesticides detected in fruits & vegetables [1] measured by the Hypersil Gold stationary phase. A KNIME tool has been used for chemical curation workflow of pesticides which were then represented by conformation independent molecular descriptors. The feature selection was done by a genetic algorithm approach in the initial quantitative structure-property relationship (QSPR) analysis. The selected features were then subjected to the Read across tool https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home for a similarity analysis using different measures like Euclidian distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity with structural analogues for the query compounds [2], and corresponding weighted average predictions were made. Based on the read-across optimized setting, RASPR descriptors were calculated using the RASAR descriptor Calculator v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home), and these were then clubbed with the previously selected structural descriptors, and finally partial least squares (PLS) models were developed from the features selected by the best subset selection approach. The models were rigorously validated by various internal and external validation metrics as recommended by the Organization for Economic Co-operation & Development. Finally, the obtained predictions were further subjected to an intelligent consensus analysis for enhancement of external predictivity. The quality of predictions of q-RASPR models superseded the quality of previously reported QSPR models. The adopted approach of q-RASPR may as well be applied for modeling of other endpoints relevant to drug discovery, materials science and predictive toxicology.

Keywords:- RASPR, RASAR KNIME, Genetic Algorithm, Read-across, PLS

[1] Rojas, C., Aranda J. F., Jaramillo, E. P., Losilla, I., Tripaldi, P., Duchowicz, P.R., Castro, E. A., 2021. Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap. Food Chem. 342, 128354, https://doi.org/10.1016/j.foodchem.2020.128354

[2] Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling towards an approach of easy interpretability and efficient transferability. Mol. Divers. https://doi.org/10.1007/s11030-022-10478-6

KUNAL ROY
kunalroy_in@yahoo.com