QSAR modeling of toxicity of chemical mixtures

Thesis submitted
by

MAINAK CHATTERJEE

Doctor of Philosophy (Pharmacy)

Department of Pharmaceutical Technology
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India
2023

JADAVPUR UNIVERSITY
KOLKATA – 700032
INDIA

INDEX NO. - 01/20/Ph

1. Title of the thesis:

"QSAR modeling of toxicity of chemical mixtures"

2. Name, Designation & Institution of the Supervisor/s:

Dr. Kunal Roy
Professor
Drug Theoretics and Cheminformatics Laboratory
Department of Pharmaceutical Technology
Jadavpur University
Kolkata – 700032
West Bengal, India

3. List of publications:

Papers related to the dissertation

Research Article(s)

(6) **Mainak Chatterjee**, Kunal Roy*. Predictive binary mixture toxicity modeling of fluoroquinolones (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A combination of 2D-QSAR and machine-learning approaches. *Environmental Science: Processes & Impacts*, **2024**, *26*, 105-118

(5) **Mainak Chatterjee**, Kunal Roy*. "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary antibiotic mixtures toward three bacterial species. *Journal of Hazardous Materials***, 2023,** *459***,** 132129

(4) **Mainak Chatterjee**, Arkaprava Banerjee, Simone Tosi, Edoardo Carnesecchi, Emilio Benfenati, Kunal Roy*. Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees. *Journal of Hazardous Materials*, **2023**, *460*, 132358

(3) **Mainak Chatterjee**, Kunal Roy*. 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

(2) **Mainak Chatterjee**, Kunal Roy*. Application of Cross-Validation Strategies to Avoid Overestimation of Performance of 2D-QSAR Models for the Prediction of Aquatic Toxicity of Chemical Mixtures. *SAR and QSAR in Environmental Research*, **2022**, *33*, 463–484

(1) **Mainak Chatterjee**, Kunal Roy*. Prediction of Aquatic Toxicity of Chemical Mixtures by the QSAR Approach Using 2D Structural Descriptors. *Journal of Hazardous Materials*, **2021**, *408*, 124936

Review Article(s)

(1) **Mainak Chatterjee**, Kunal Roy*. Recent advances on modelling the toxicity of environmental pollutants for risk assessment: From single pollutants to mixtures. *Current Pollution Reports*, **2022**, *8*, 81–97

Papers not related to the dissertation

(4) Shilpayan Ghosh, **Mainak Chatterjee**, Kunal Roy*. Predictive quantitative read-across structure–property relationship modeling of the retention time ($\log t_R$) of pesticide residues present in foods and vegetables. *Journal of Agricultural and Food Chemistry*, **2023**, *71*, 9538-9548

(3) Rahul Paul, **Mainak Chatterjee**, Kunal Roy*. First report on soil ecotoxicity prediction against *Folsomia candida* using intelligent consensus predictions and chemical read-across. *Environmental Science and Pollution Research*, **2022**, *29*, 88302-88317

(2) Arkaprava Banerjee, **Mainak Chatterjee**, Priyanka De, Kunal Roy*. Quantitative predictions from chemical read-across and their confidence measures. *Chemometrics and Intelligent Laboratory Systems*, **2022**, *227*, 104613

(1) **Mainak Chatterjee**, Arkaprava Banerjee, Priyanka De, Agnieszka Gajewicz-Skretna, Kunal Roy*. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environmental Science: Nano*, **2022**, *9*, 189-203

Book chapter related to the dissertation

(1) **Mainak Chatterjee**, Kunal Roy*. Computational Modeling of Mixture Toxicity. In *In Silico Methods for Predicting Drug Toxicity*; Benfenati, E., Ed.; Humana Press, Springer, **2022**; pp 561–587

Book chapter not related to the dissertation

(1) **Mainak Chatterjee**, Kunal Roy*. Quantitative Structure-Activity Relationships (QSARs) in Medicinal Chemistry. In *Cheminformatics, QSAR and MachineLearning*

Applications for Novel Drug Development; Roy, K.; Ed.; Academic Press, **2023**, pp 3-28

4. List of Patent(s): **Nil**

5. List of presentation/(s) in National/International/Conferences/Workshops:

International Conference/Workshop(s)

(3) **Mainak Chatterjee**, Kunal Roy. **Poster presentation- In-person**: entitled "Chemical similarity-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical mixtures against *Aliivibrio fischeri*" at the **23rd EuroQSAR**- 23rd Symposium on Quantitative Structure-Activity Relationship, Heidelberg, Germany, **September 26-30, 2022**

(2) **Mainak Chatterjee**, Kunal Roy. **Poster presentation- Virtual:** entitled "Application of 2D-QSAR and read-across techniques to predict aquatic toxicity of binary and multicomponent pharmaceutical mixtures against *Aliivibrio fischeri*" at **ACS Fall 2022**, Chicago-IL and Hybrid, **August 21-25, 2022**

(1) **Mainak Chatterjee**, Arkaprava Banerjee, Priyanka De, Agnieszka Gajewicz-Skretna, Kunal Roy. **Poster presentation- Virtual**: entitled "Development and application of a new similarity-based read-across algorithm for the prediction of nanotoxicity of small datasets" at **OpenTox Euro 2021** Virtual Conference, **September 20-24, 2021**

National conference/Workshop(s)

(4) **Mainak Chatterjee**. Delivered an expert talk on "QSAR Practical: Hands-on Session on the DTC Lab Tools" in SERB sponsored workshop on "Hands-on training on Computer Aided Drug Design and Discovery Tools" organised by Ganpat University on **11-17 July, 2022**.

(3) **Mainak Chatterjee**. Attended webinar on "Gaining acceptance in next-generation PBK modelling approaches for regulatory assessments: Case studies", OECD Webinar, Organized by European Society of Toxicology In Vitro (ESTIV) held on **May 10, 2022**.

(2) **Mainak Chatterjee**. Attended webinar on "Biodiversity and Environmental Protection during Pandemic Outbreak of COVID 19" organized by International Society of Waste Management Air and Water (ISWMAW) jointly by Chairman Prof. Dr. Sadhan Kumar Ghosh of India and Co-Chairman Dr. Abas Basir of Sri Lanka on **5th June 2020**.

(1) **Mainak Chatterjee**. Attended a One-week Workshop on "Chemometric data processing for express analytical systems in agricultural and environmental studies" conducted by Prof. D. Kirsanov of St. Petersburg State University, Russia, at Dept. of Instrumentation and Electronics Engineering, Jadavpur University on **2-7 December 2019**.

"Statement of Originality"

I ***Mainak Chatterjee*** registered on ***20th July 2020*** do hereby declare that this thesis entitled "***QSAR modeling of toxicity of chemical mixtures***" contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies. All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work. I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is **8%***.

Mainak Chatterjee. 08.11.2023

Signature of Candidate with Date

Certified by Supervisor(s):

08.11.2023

DR. KUNAL ROY
PROFESSOR & EX-HEAD
DEPT. PHARMACEUTICAL TECHNOLOGY
JADAVPUR UNIVERSITY
KOLKATA - 700 032 INDIA

(Signature with date & seal)

*Ignoring smaller match of up to 14 words (as per the UGC) and ignoring own publications as the source.

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled "**QSAR modeling of toxicity of chemical mixtures**" submitted by Shri **Mainak Chatterjee**, who got his name registered on **20th July 2020** for the award of Ph.D. (Pharmacy) degree of Jadavpur University is absolutely based upon his own work under the supervision of **Dr. Kunal Roy**, and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

08-11-2023

(Signature with Date and Seal)

DR. KUNAL ROY
PROFESSOR & EX-HEAD
DEPT. PHARMACEUTICAL TECHNOLOGY
JADAVPUR UNIVERSITY
KOLKATA - 700 032, INDIA

Acknowledgements

I deem it a pleasure and privilege to work under the guidance of ***Dr. Kunal Roy***, Professor, Drug Theoretics & Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata. 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 also greatly indebted to him for his valuable guidance throughout my PhD tenure that enabled me to complete the work.

I am thankful to the Research Advisory Committee members ***Prof. Tarun Jha***, and ***Prof. Tapan Kr. Maity*** for timely evaluating my progress and also for their valuable feedback on my work. I also express my gratitude to the Head, Professors, non-teaching staff of the Department of Pharmaceutical Technology, and the entire University Authority. I got their kind help when needed. My sincere thanks to the ***All India Council for Technical Education (AICTE), New Delhi*** for providing financial assistance in the form of a ***National Doctoral Fellowship (NDF)*** in the past four years.

With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, perpetual assistance, and cooperation from my senior Dr. Probir Kr. Ojha, Dr. Vinay Kumar, Dr. Kabiruddin I. Khan, Dr. Pathan Mohsin Khan, and Dr. Gopala Krishna Jillela. 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 friends Dr. Priyanka De, Mrs. Joyita Roy and my juniors Arnab Seth, Sapna Kumari Pandey, Arkaprava Banerjee, Souvik Pore and Shilpayan Ghosh, Rahul Paul, Ankur Kumar, Subham Kr. Pandey, Doelima Bera, and MD. Mobarak Hossain who all have extended their helping hands and friendly cooperation all through my 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. Malay Chatterjee*** and ***Mrs. Bandana Chatterjee*** for their unconditional love and affection; their support keep me calm and strong during many undulations in the path of life.

Mainak Chatterjee.

MAINAK CHATTERJEE

Date: 08-11-2023
Place: KOLKATA

Preface

The research work presented in this dissertation is the result of constant effort over the last four and a half years. This dissertation is mainly focused on different computational tools and techniques that have been employed for the prediction of toxicity of chemical mixtures against different environmentally relevant endpoints. The prime objective of this dissertation was to develop *in silico* modeling of chemical mixtures and using those models, further risk assessment of chemicals. With industrial advancement, the chemical space has grown. Thus, the existence of hazardous chemicals in the environment is not an extended reality anymore. These chemicals are mostly present in the environment as multicomponent mixtures but we have very little mixture toxicity and safety data. Experimental toxicity determination of isolated chemicals and/or mixtures is a complicated process and it is unfeasible to use the experimental approach to fill the existing mixtures' toxicity data gap. The *in silico* modeling techniques have been proven advantageous over experimentation; these are fast, cheap, and environmentally safe alternatives that can provide reliable predictions. These computational models can be developed with limited toxicity data in hand and using those developed models, one can get predictions of toxicity of new or untested mixtures/ solo chemicals without even preparing it.

Over the past few decades, in silico modeling has evolved a lot and grabbed sufficient attention of the scientific community. The Quantitative structure-activity relationships (QSARs), Read-across predictions, Machine learning (ML) models, and the new quantitative Read-across structure-activity relationships (q-RASARs) are becoming more popular now in predictive toxicology. These methods are similarity-based approaches where the prediction of endpoints can be obtained from structurally similar chemical analogues. The majority of work in this dissertation deals with the development of predictive and statistically sound mixture-QSAR models; although different other algorithms like read-across predictions, q-RASAR, and ML have also been used. A standard QSAR methodology comprises the steps i.e., data collection and curation, descriptor calculation and dataset division, feature selection and model development, validation of the developed model, and mechanistic interpretation of the response in light of the identified structural features. Various statistical tools are used for developing the model whereas, the developed model must be validated using strict internal and external validation protocols. Organization for Economic Cooperation and Development (OECD) prescribes the guidelines of QSAR modeling that have been followed in all presented works. During this study, many software tools have been employed that are freely available at http://teqip.jdvu.ac.in/QSAR_Tools/. The 3R strategy (Reduction, Refinement, and Replacement) of animal experimentation has been followed throughout the course of this study; therefore, no animal has been harmed.

The works presented in this dissertation are as follows:

Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors

Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against *Aliivibrio fischeri*

Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species

Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach

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

Chapter 1: Introduction

Chapter 2: Present Work

Chapter 3: Materials and Methods

Chapter 4: Result and Discussion

Chapter 5: Conclusion

Reference

Appendix: Reprints

The overall background knowledge of the mixture toxicity assessment, the computational approaches, and the previous works done in this area of research have been discussed under the heading "***Introduction***". Additionally, information on a brief history of QSAR, the basic principles of in-silico approaches that have been employed, and the validation protocol of in-silico models have also been provided. Chapter "***Present Work***" briefly discusses the performed research works along with the collected datasets. In the "***Materials and Methods***" section, the methodology of the entire work done has been thoroughly discussed. Schematic representations of all studies and the important numerical formula used for calculations have been provided in this section to avoid ambiguities. The results of all studies performed have been discussed in the "***Results and Discussions***" in detail. An overall "***Conclusion***" has been provided in this dissertation followed by the "***References***". Most of the studies from this dissertation work have been published in ***reputed international journals*** and also presented at different national and international conferences. Those published works are thus also included in the ultimate section of this dissertation under section "***Reprints***". However, the research work performed and presented in the limited scope of this dissertation constitutes only a small portion in the broad spectram of work in this area of research. Various interesting aspects of research that have arisen in this dissertation, can be further investigated more appropriately in the future study.

Abbreviation

2D QSAR	Two dimensional QSAR
3D QSAR	Three dimensional QSAR
4D QSAR	Four dimensional QSAR
5D QSAR	Five dimensional QSAR
6D QSAR	Six dimensional QSAR
7D QSAR	Seven dimensional QSAR
AD	Applicability domain
ANN	Artificial neural network
ATDSR	Agency for the Toxic Substance and Disease Registry
BSS	Best subset selection
CA	Concentration addition
CCC	Concordance correlation coefficient
CFC	chlorofluorocarbon
CTC	Close training compound
DDT	dichlorodiphenyltrichloroethane
EC	Effective concentration
ECHA	European Chemical Agency
ED	Euclidean distance
EFSA	European Food Safety Authority
ETA	Extended topochemical atom
EU	European Union
GA	Genetic algorithm
GB	Gradient boost
GCA	Generalized concentration addition
GK	Gaussian kernel
IA	Independent action
IAI	Integrated addition-interaction
IC	Inhibitory concentration
IPPC	Integrated Pollution and Prevention Control Directive
KNIME	Konstanz Information Miner
LD	Lethal dose
LDA	Linear discriminant analysis
LK	Laplacian Kernel
MAE	Mean absolute error
MCC	Matthews correlation coefficient
ML	Machine learning
MLP	Multilayer perception
NOEC	No observed effective concentration
OECD	Organization for Economic Cooperation and Development
PCB	polychlorinated biphenyls
PCRA	Principal component regression analysis
PLS	Partial least squares
POP	Persistent organic pollutants
PPP	Plant Protection Products
PRI	Prediction reliability indicator
QSAAR	Quantitative structure-activity-activity relationship
QSAR	Quantitative structure-activity relationship
RA	Read-across
RASAAR	Read-across structure-activity-activity relationship
RASAR	Read-across structure-activity relationship
REACH	Registration, evaluation, authorisation and restriction of chemicals
RF	Random forest
RMSEC	Root mean square error of calibration
RR	Ridge regression
SETAC	Society of Environmental Toxicology and Chemistry
SMILES	Simplified Molecular Input Line Entry Systems
S-MLR	Stepwise multiple linear regression
SVM	Support vector machine
US EPA	United States Environment Protection Agency
WAP	Weighted average prediction
WFD	Water Framework Directive
WHO	World Health Organization
XGB	Extreme gradient boost

Contents

Chapter 1: Introduction……...1

1.1. Environmental pollution: A threat to environmental safety and human health……1

1.2. Environmental pollutants and mixtures…….2

1.3. The role of chemical regulatory agencies in mixture risk assessment……..2

1.4. Toxicity assessment of environmental pollutants…….3

1.5. Toxicity assessment of chemical mixtures……4

1.5.1. Conventional techniques for mixture toxicity assessment……..4

1.5.2. Computational techniques in mixture toxicity assessment……8

1.6. In silico predictive modeling approaches……..9

1.6.1. Quantitative structure-activity relationship (QSAR)……...9

1.6.2. Read-across predictions……26

1.6.3. Read-across structure-activity relationship (RASAR)……..28

1.6.4. Machine-learning (ML) algorithms……...29

1.7. Application of in silico modeling approaches for mixture toxicity assessment……..30

Chapter 2: Present Work……..33

2.1. Objectives of the dissertations…….34

2.2. Datasets employed for the development of Mixture-QSAR/q-RASAR/ML models..35

2.1.1. Dataset I (Study 1)……36

2.1.2. Dataset II (Study 2)……...36

2.1.3. Dataset III (Study 3)……..37

2.1.4. Dataset IV (Study 4)……..37

2.1.5. Dataset V (Study 5)……...38

2.1.6. Dataset VI (Study 6)……..39

Chapter 3: Materials and Methods ……..40

3.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors……...40

3.1.1. Toxicity data……..40

3.1.2. Molecular Descriptors……...45

3.1.3. Dataset Division……46

3.1.4. Model Development……..46

3.1.5. Statistical Validation Metrics……....48

3.1.6. Applicability Domain……49

3.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures……..49

3.2.1. Toxicity data……49

3.2.2. Molecular descriptors……55

3.2.3. Dataset division……56

3.2.4. Feature selection and model development……56

3.2.5. Internal and external validation……57

3.2.6. Cross-validation……59

3.2.7. Chemical read-across……61

3.2.8. Prediction reliability indicator and designing of empirical dataset……..61

3.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutica and pesticide mixtures against Aliivibrio fischeri……63

3.3.1. Toxicity data……..63

3.3.2. Mixture descriptor calculation……...71

3.3.3. Dataset division……72

3.3.4. Feature selection and model development……73

3.3.5. Chemical read-across……75

3.3.6. Statistical validation and Applicability domain analysis……..76

3.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees……..77

3.4.1. Toxicity data……..77

3.4.2. Mixture descriptors……...82

3.4.3. Dataset division and feature selection for QSAR and subsequent modeling……83

3.4.4. Read-across and RASAR descriptors calculation……86

3.4.5. q-RASAR modeling……..87

3.4.6. Validation protocols……..88

3.4.7. Prediction of unknown mixtures with PLS and ML-based q-RASAR models and reliability analysis with PRI tool……89

3.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species……100

3.5.1. Mixture toxicity data collection and dataset preparation ……100

3.5.2. Mixture descriptor calculation considering three different mixing rules……110

3.5.3. Dataset division and proper mixing rules identification……110

3.5.4. Multilayered descriptor selection and QSAR/QSAAR model development…..115

3.5.5. Application of q-RASAR in QSAAR development……117

3.5.6. Statistical model validation……120

3.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach……...121

3.6.1. Toxicity data collection……...121

3.6.2. Mixture descriptor calculation……125

3.6.3. Dataset division and optimization of mixing rule……...126

3.6.4. Feature selection and model development……..127

3.6.5. Machine-learning models……127

3.6.6. Statistical validation……129

3.6.7. Toxicity prediction of binary mixture dataset designed from parent FQs and their degradates: Use of the PRI tool…….129

Chapter 4: Results and Discussions……145

4.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors…….145

4.1.1. QSAR study of toxicity of individual chemicals (Series 1) to P. phosphoreum.147

4.1.2. QSAR study of toxicity of individual chemicals (Series 2) to S. capricornutum……150

4.1.3. QSAR study of toxicity of mixtures (Type-1) to P. phosphoreum……..154

4.1.4. QSAR study of toxicity of mixtutes (Type-2) to S. capricornutum……160

4.1.5. Applicability domain and Y-randomization of the developed PLS models……163

4.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures……164

4.2.1. The developed model and its validation……..164

4.2.2. Result of cross-validation and prediction reliability indicator……167

4.2.3. Mechanistic interpretation of modeled descriptors…….167

4.2.4. Result of chemical read-across……172

4.2.5. Comparison with a previously published model…….174

4.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against Aliivibrio fischeri……..174

4.3.1. The developed model and its validation……..174

4.3.2. Results of cross-validation……..177

4.3.3. Study of modeled descriptors……..179

4.3.4. Result of chemical read-across……187

4.3.5. A comparative assessment of the model with previous reports……..188

4.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees……190

4.4.1. Developed q-RASAR models and best mixing rule identification…….190

4.4.2. Best model and result of the validation……...193

4.4.3. Applicability domain analysis of the PLS q-RASAR model……..196

4.4.4. Insights on modeled descriptors……..197

4.4.5. Prediction of toxicity of unknown mixtures and reliability analysis……...204

4.4.6. Comparative assessment of the quality of q-RASAR model with previous work……...215

4.4.7. Limitations of the developed q-RASAR models…….....218

4.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species……218

4.5.1. Developed QSAR models and their validations…….218

4.5.2. Insights of modeled descriptors identified in the QSAR analysis……...228

4.5.3. Insights of modeled descriptors identified in the QSAAR analysis……239

4.5.4. Importance of q-RASAR algorithm in interspecies mixture toxicity modeling.242

4.5.5. Comparative analysis with the previous report……..245

4.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach……...248

4.6.1. Optimized mixing rule and best division……248

4.6.2. Partial least squares (PLS) regression model and validation……..249

4.6.3. Machine learning models and the result of validation……251

4.6.4. Insights of modeled descriptors……...253

4.6.5. Applicability domain analysis…….257

4.6.6. Predicted responses of untested mixtures and the result of reliability analysis..258

4.6.7. Comparison of models……271

Chapter 5: Conclusions……273

5.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors……...273

5.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures……..274

5.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against Aliivibrio fischeri……...274

5.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees……275

5.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species……..276

5.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach...277

References…….278

Appendix: Reprints

CHAPTER - 1

INTRODUCTION

Chapter 1: Introduction

1.1. Environmental pollution: A threat to environmental safety and human health

Risk assessment of environmental pollutants is an important concern of modern science to look after due to huge chemical expansion and environmental pollution. Nowadays, pollution is one of the leading environmental causes of disease and premature death worldwide. As per a report of the Lancet Commission published in 2017, almost 16% of total premature deaths throughout the world (9 million) in 2015 were caused due to environmental pollution- which is three times higher than the cumulative death caused by Malaria, Tuberculosis, and AIDS (Landrigan et al., 2018). The maximum number of pollution-related mortality were reported from low to middle-income countries of Southeast Asia (3.2 million deaths) and the Western Pacific region (2.2 million) (Forouzanfar et al., 2016). Highly populated countries such as India (located in Southeast Asia) and China (located in the Western Pacific region) are heavily affected by environmental pollution (Forouzanfar et al., 2016). However, various natural events can be the reason for pollution; the word pollutants generally implies contaminants from different sources including anthropogenic origin. The worldwide growth of chemical and petroleum industries (E. Ite et al., 2018), excessive use of agrochemicals (Baek et al., 2019; Singh et al., 2016), pharmaceuticals (Adhikari and Mishra, 2018; Cizmas et al., 2015), and other organic substances for various needs (Horton and Barnes, 2020; Malik et al., 2019; Yi et al., 2019; Zhang et al., 2018), and improper disposal of these chemicals boost environmental pollution to a large extent. Among various kinds of pollution, water pollution (Cizmas et al., 2015; Hasan et al., 2019; Ighalo et al., 2021), air pollution (Manisalidis et al., 2020), and soil pollution (Rodríguez-Eugenio et al., 2018) are the most common and major causes of toxicity towards environmental ecosystems. Pollution also affects human beings besides the environment. However, the forms of pollution associated with profound poverty and traditional lifestyles are slowly declining; the forms of pollution produced by industry, mining, electricity generation, mechanised agriculture, and petroleum-powered vehicles are all on the rise, with the most marked increases in the developing 3rd world countries.

Unpredictable chemical exposure is another global issue associated with pollution. The environmental ecosystems and human beings are exposed to various chemicals and multicomponent mixtures but their effects and disease-forming ability on living organisms and human beings are mostly underestimated. Since 1950, beyond 140000 new chemical entities including pesticides and pharmaceuticals have been synthesized (Landrigan et al., 2018). Among these materials, more than 20000 pesticides (rodenticides, herbicides, insecticides, fungicides etc.) are available in the global market and almost 5.2 billion pounds are produced each year globally (Grube et al., 2011). The majority of these highly produced chemicals have not even passed through a sufficient number of stringent toxicity assessments or pre-marketing safety evaluations, because it has been mandatory since last decade only in a very few high-income countries. As a result of this negligence, human beings and the environment are adversely affected repeatedly. There are some historically proven toxic chemicals such as asbestos, lead, polychlorinated biphenyls (PCBs), chlorofluorocarbon (CFCs), dichlorodiphenyltrichloroethane (DDT), etc. which came upfront in several instances for their adverse effects towards the living environment (Jacobs et al., 2019; Kovshov, 2019; Onuzulu et al., 2019; Torrens and Castellano, 2019). The new chemicals that have come into the market in the past 2-3 decades have undergone few pre-marketing safety assessments. These tests prove the toxic characteristics of various new-generation chemicals such as endocrine disruptors, developmental neurotoxicants, novel insecticides, herbicides, pharmaceutical wastes, and various nanomaterials. If these chemicals are also not properly regulated as the

previous generations during the marketing and disposal, then these observations can be the marker of threat to human health and environmental safety in the near future.

1.2. Environmental pollutants and mixtures

The pollutants are never found alone in the environment. Multicomponent chemical mixtures are formed continuously with variable combinations, and the number of such combinations is increasing. Chemical mixtures can be defined as a combination of two or more pure chemicals which are present together but not chemically bonded. As per an article published in 1998 by Yang *et al.*, "there is no such object as single chemical exposure" (Yang et al., 1998). Although most of the environmental pollutants are found at very low concentrations, living organisms and human beings are permanently exposed to these mixtures of pollutants which can be the reason for adverse response (Altenburger et al., 2003). The probable interactions (both additive response and synergism) among mixture components can cause behavioural changes of toxic responses. Additive response and synergism usually enhance the toxic response of a pollutant in the presence of others and these are not at all desirable. The aquatic environment is mostly affected by chemical pollutants and their mixtures due to the disposal of industrial, agricultural, and pharmaceutical wastes, heavy metals, and nanomaterials. Industrial wastes are usually passed through wastewater treatment plants that contain activated sludge systems. Conventional plants cannot screen all the pollutants efficiently and the organic structures cannot be decomposed completely. Therefore the resultant wastewater can be toxic to the activated sludge and the aquatic species as well (Dsikowitzky and Schwarzbauer, 2014). The presence of such mixtures of toxic chemicals in fresh and marine water can cause serious damage to the environmental ecosystems and due to bioaccumulation, human beings can also be affected (Bilal et al., 2019b, 2019a; Rasheed et al., 2018; Singh et al., 2018). In the recent past, the European Union (EU) funded project- 'SOLUTION (2013-2018)' (Brack et al., 2015), and the Society of Environmental Toxicology and Chemistry (SETAC) funded 'Horizon Scanning Project (2013)' (Van den Brink et al., 2018) were launched to identify potential chemical species and mixtures in the environment. After assessing the report of 'SOLUTION', Kortenkamp *et al.* claimed that the mixture risks threaten the water quality and that the Water Framework Directive (WFD) of the EU should be revised for better coordination across all chemical legislations to improve the protection of the aquatic environment from the multicomponent mixtures (Kortenkamp et al., 2019). There are a handful of studies that report the presence of multicomponent mixtures in the environment (mostly in water bodies) that again suggest the importance of the mixture's toxicity estimation for better risk assessment of environmental pollutants (Altenburger et al., 2015; Brack et al., 1999; Nováková et al., 2020; Pereira et al., 2018; Posthuma and De Zwart, 2006; Silva et al., 2012).

1.3. The role of chemical regulatory agencies in mixture risk assessment

Regulatory agencies like the United States Environment Protection Agency (US EPA), European Chemical Agency (ECHA) etc. have taken the issue of human safety and environmental protection seriously and they further reinforced the safety protocols for global outreach. The European Union (EU) has abruptly bifurcated the European chemical regulations into two gross legislation for environmental safety and human health, i.e., substance and product-based legislation and process and media-based legislation. 'Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH)', 'Classification, Labeling, and Packaging (CLP)', and the 'Placing of Plant Protection Products (PPP)' regulations are the parts of the substance and product-based legislation whereas, 'Integrated Pollution and Prevention Control Directive (IPPC)' and the 'Water Framework Directive (WFD)' are the parts of the process and media-based legislation (European Commision, 2003). There are

various other regulatory agencies such as 'World Health Organization (WHO)', the 'Agency for the Toxic Substance and Disease Registry (ATDSR)', 'European Food Safety Authority (EFSA)' etc. which are also coming upfront for the human safety assessment and environmental risk assessment of environmental pollutants (Altenburger et al., 2003; Altenburger and Greco, 2009; Li et al., 2008; Syberg et al., 2009). However, sufficient attention is still not given to the risk assessment of chemical mixtures even after having such stringent regulatory guidelines.

1.4. Toxicity assessment of environmental pollutants

Ecotoxicity assessment is an integral part of applied science ecotoxicology which basically deals with the toxic effects on environmental ecosystems and all other higher animals, caused by synthetic or naturally produced environmental pollutants (Truhaut, 1977). Various environmental species can be the classic markers of environmental pollution because they get affected easily by the presence of pollutants. Ecotoxicity is assessed by direct experimentations or mathematical modeling studies using the toxicity data obtained from various environmental species such as plants (*Oriza sativa*, *Liliopsida sp.*, *Vigna radiata*) (Kim et al., 2019), bacteria (*Photobacteriaum phosphoreum*) (Khan and Roy, 2021), algae (*Chlorella vulgaris*, *Selenastrum capricornutum*) (Seth and Roy, 2020; Yao et al., 2013), invertebrates (*Daphnia magna, Folsomia candida*) (Khan et al., 2019; Ogungbemi and van Gestel, 2018), earthworm (*Eisenia foetida*) (Ghosh et al., 2020), fish (*Danio rerio, Oncorhynchus mukiss, Lepomis macrochirus, Pimephales promelas*, etc.) (Pandey et al., 2020), birds (*Colinus virginianus, Anas platyrhynchos, Phasianus colchicus etc.*) (Mukherjee et al., 2022) etc. Toxicity is usually expressed by different forms such as EC_x (effective concentration for producing maximum response against x% population), LC_{50} (median lethal concentration), NOEC (no observed effective concentration), LOEC (lowest observed effective concentration), IC_{50} (inhibitory concentration of chemical for 50% population) etc. (Sigurnjak Bureš et al., 2021). The EC_{50} (concentration of a toxic chemical which produces a maximum response on 50% population) is the most used endpoint by toxicologists for the ecotoxicity assessment using mathematical modeling studies (Sigurnjak Bureš et al., 2021). Endpoints can also be divided into two broad categories (e.g., acute toxicity and chronic toxicity) based on the duration of toxic effects. Acute toxicity is used mostly in ecotoxicity assessment, and it represents the toxic effect of a chemical after a short-term exposure to a biological species. On the other hand, chronic toxicity represents the long-term toxic effect on an organism by a chemical (Elisabeth et al., 2014). As per the United Nation's Globally Harmonized System of Classification and Labelling of Chemicals (Nations, 2017), the pollutants can be categorised into different toxic classes-

i. Highly toxic chemical ($EC50 \leq 1$ mg/L)
ii. Moderately Toxic Chemical (1 mg/L $< EC50 \leq 10$ md/L)
iii. Poorly toxic chemical ($EC50 \leq 100$ mg/L)
iv. Non-toxic chemical ($EC50 \geq 100$ mg/L)

Besides the knowledge of environmental species, endpoints, and toxicity profile of chemicals, it is also important to know the mechanism of action to properly explain the toxic response among various classes of chemicals on certain organisms of interest. As per a literature report published by Verhaar *et al.*, there are four distinct categories of the mode of action (Verhaar et al., 1992):

i. Inert chemicals with baseline toxicity, which do not have sufficient reactivity to show acute response and also do not bind with a specific receptor of an organism. Nonpolar narcosis

is an example of this and the response is driven by physical factors such as the hydrophobicity of compounds.

ii. Less inert chemicals show polar narcosis by the hydrogen bond donor activity.

iii. Reactive chemicals show non-specific but relatively stronger reactive responses. Chemical species such as aldehydes, epoxides etc. show this response without affecting the specific working site of the test organisms.

iv. Specifically acting chemicals: the pharmaceuticals, pesticides, some polychlorinated hydrocarbons etc. have a specific affinity towards selected receptors of the organism. Their mode of action is non-superficial but more accurately characterized.

As a part of the risk assessment of environmental pollutants, the single chemical toxicity assessment and toxicity data generation are of the primary concern. The environmental toxicity of chemicals such as pharmaceuticals, pesticides, herbicides, plant hormones, persistent organic pollutants (POPs), ionic liquids, nanomaterials, microplastics, organic dyes, heavy metals etc. is evident from various scientific reports and hence the ecotoxicity assessments of these classes of chemicals are also observed in the literature. Although there are sufficient experimental methods (in vitro and in vivo) available for the toxicity study of chemicals, a data gap still exists. The experimental approaches are associated with ethical concerns in addition to the involvement of cost and resources. On the contrary, in silico predictive approaches such as quantitative structure-activity relationships (QSARs), read-across etc. are fast, less expensive, and reliable.

1.5. Toxicity assessment of chemical mixtures

As mentioned earlier, common risk assessment of environmental pollutants is majorly based on the toxicity study of individual pollutants instead of multicomponent mixtures. However, single chemical exposure in the environment is a myth. The combined effect of the mixture's components changes the overall response radically from the response of a single component. Mixtures can produce a higher response than the corresponding components (synergistic effect/additive effect) or a lower response as well (antagonistic effect). The combinations of the same components with different mixing ratios can produce different responses to test organisms. Due to the above-mentioned interactive nature of compounds and other associated complexities, the mixture effect cannot be estimated by the simple average of the responses of corresponding components or by using any particular equation (Logan and Wilson, 1995; Nowell et al., 2014). The six important features and processes such as bioavailability, internal transport, uptake, metabolization, binding at the target site and excretion are important determinants of the toxic effect of chemicals on an organism which can be altered due to the interaction between mixture components (Cedergreen et al., 2012).

1.5.1. Conventional techniques for mixture toxicity assessment

Conventional toxicity assessment techniques are mostly used for determining the mixture toxicity. Technically two major concepts are applied for the risk assessment of chemical mixtures: (i) a whole mixture-based approach to determine the toxicity of mixture samples, (ii) a component-based approach to calculate mixture toxicity from the available toxicity data of mixture components (Kar and Leszczynski, 2019). The conventional techniques of mixture risk assessment have been summarized as a flowchart in **Figure 1.1.**

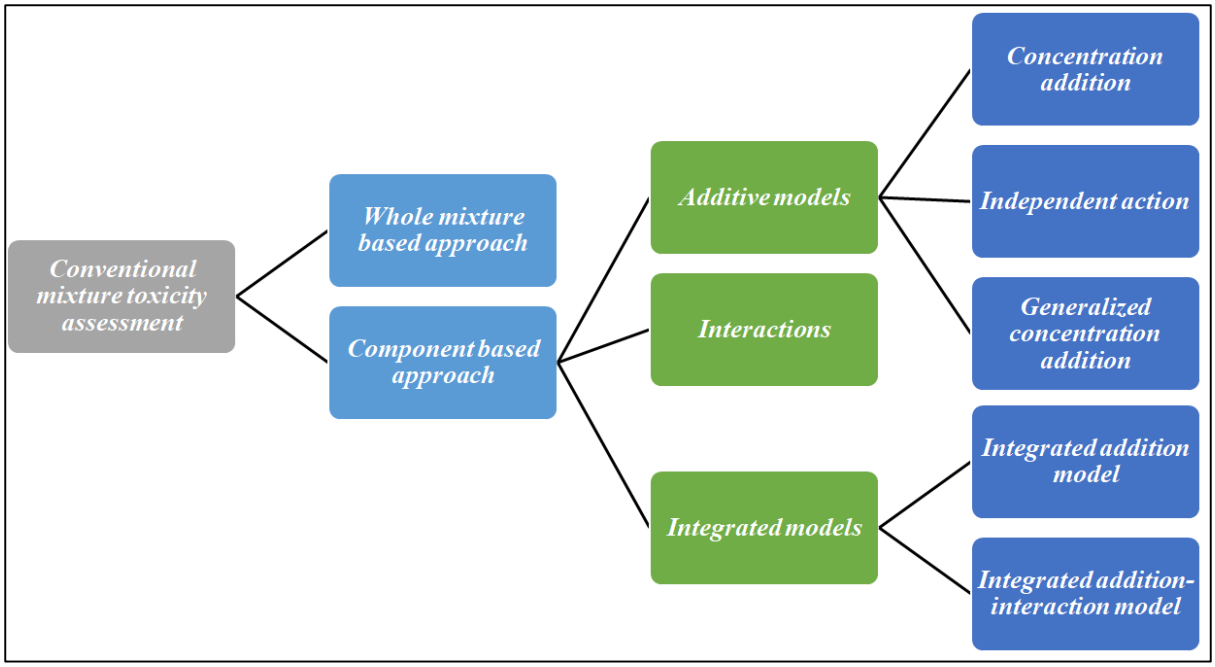


Figure 1.1. Different classes of conventional mixture toxicity assessment

1.5.1.1. Whole mixture-based approach

It is essentially the direct toxicity experimentation using the whole mixture as a sample. This method is more accurate and can be executed for both the intentional mixtures (prepared in the laboratory) and unintentional mixtures (collected from the environment) in the laboratory setup. Accurate results can be obtained from the whole mixture-based approach even after the presence of any unidentified components and/or any kind of possible interactions among mixture components. In vivo or in vitro experimentation is an integrated part of the whole mixture-based toxicity assessment (Kar and Leszczynski, 2019)

The major limitations of this approach are:

i. It is a troublesome, costly, time-consuming and ethically complicated process.

ii. The environmental mixtures can exist in an infinite number of combinations and can affect various environmental species at a time; hence ecotoxicological studies are practically unfeasible in the whole mixture-based approach.

iii. This approach is limited to very small-scale toxicity assessment under laboratory conditions like dietary and consumer product mixtures.

iv. Without the physical presence of the mixture, prior predictions of toxicity cannot be made in a whole mixture-based approach.

v. The knowledge of the mixture components and their interactions cannot be acquired in the whole mixture-based approach.

1.5.1.2. Component-based approaches

Component-based approaches typically depend upon the available knowledge of toxicity and exposure of target mixture components. These methods are a more practical and popular way to predict the toxicity of mixtures than the whole mixture-based approach. The basic concepts of component-based approaches can be categorized according to the complexity of the process (how the mixture toxicity is being computed from the component chemicals) into two parts: (i) simple additive models and (ii) integrated models. The simple additive models are based on

the assumption that there is no interaction occurring among the mixture components and hence these are simple combined effects of the mixture components. Under the simple additive effect group, there are 3 types of modeling: Concentration addition model (CA), Independent action model (IA), and Generalized concentration addition model (GCA). The Integrated addition model and Integrated addition-interaction (IAI) model come under the category of integrated models. Till date, there are no precise modeling techniques to predict the interaction effects.

Concentration addition (CA): It is the most popular additive toxicity modelling which is based on a similar mechanism of action hypothesis. The basic assumptions of CA are: (1) the mixture components act via the same mechanism of action after binding into the same target site of an organism and that is why it is also called concentration dilution; (2) the component chemicals of a mixture are not interactive among each other (Heys et al., 2016). This concept was first proposed by scientists Loewe and Muischnek in 1926 (Loewe and Muischnek, 1926). It has been paid more attention by the US EPA and European Union Commission and is considered as the default toxicity model for a mixture's risk assessment (Kortenkamp et al., 2009; US-EPA, 1986). In this model, the components with negligible concentrations (even below NOEC) can contribute to the overall toxicity of the mixture. Concentration addition can also be expressed mathematically by the following equation 1.1-

$$ECx_{Mix} = \sum_{i=0}^{n} \left({}^{C_i}/_{EC_i} \right) \qquad (1.1)$$

where, ECx_{Mix} represents the mixture's effective concentration, C_i represents the concentration of component *i* present in the mixture, EC_i represents the effective concentration of component *i*.

Independent action (IA): This model is another additive model and is also termed as response addition (Altenburger et al., 2000; Bliss, 1939). It assumes that the mixture components act by completely different mechanisms of action on different binding sites. It also does not consider any interaction among the components. IA model simply adds the response obtained by the single substances of the mixture, and hence the components at NOEC do not have any effect on overall toxicity estimation (Cassee et al., 1998; Olmstead and LeBlanc, 2005). When $E(C_{mix})$ is the overall mixture's toxic response and $E(C_i)$ represents the effective concentration of mixture component *i*, then the IA model can be mathematically shown by the equation 1.2-

$$E(C_{mix}) = 1 - \prod_{i=0}^{n} E(C_i) \qquad (1.2)$$

Generalized concentration addition (GCA): This modeling technique was derived to overcome the major gaps of the previously discussed models (CA & IA). lies behind the strict assumptions of the concepts. Mixtures containing only similarly acting or dissimilarly acting components are hardly found in the real world. To fill this gap in concept, the generalized concentration addition (GCA) model was proposed in 2009 by Scientists Howard and Webster (Howard and Webster, 2009). GCA is the combined effect of the efficacy and potency of each component of a mixture. The model can be mathematically expressed as the following equation 1.3-

$$E = \frac{Max\ effect\ level_A [A]/EC_{50A} + Max\ effect\ level_B [B]/EC_{50B} + \cdots}{1 + \frac{[A]}{EC_{50A}} + \frac{[B]}{EC_{50B}} + \cdots} \qquad (1.3)$$

where E stands for the effect of a mixture at a specific concentration; max effect level A and max effect level B are the maximum effective level of mixture components A and B respectively; [A] and [B] are the concentrations of components A and B, respectively in the

mixture at an explicit mixture concentration; EC_{50A} is the median effective concentration of component A; EC_{50A} is the median effective concentration of mixture component B.

Integrated addition: In 2005, scientists Olmstead and LeBlanc proposed the integrated addition model simply by merging the concentration addition and independent action (Olmstead and LeBlanc, 2005). It is a two-stage prediction (TSP) where CA and IA are utilised separately (Altenburger et al., 2004). During the construction of this model, a preliminary chemical clustering is required based on their modes of action. The toxicity of all clusters containing only similar-acting components is calculated by applying the concentration addition model in the 1st stage. The independent action is the 2nd stage process where the overall toxicity is determined from the calculated toxicities of different clusters (Altenburger et al., 2004)(Altenburger et al., 2005). This model can also be expressed mathematically by the equation 1.4-

$$EC_{50,mix} = 1 - \prod_{x=1}^{n} \left\{ 1 - \frac{100}{1+\left[\sum_{x=1}^{n}\left(C_x/EC_{50,x}\right)\right]^{P'}} \right\} \qquad (1.4)$$

where, $EC_{50,mix}$ is the toxicity of mixture; $EC_{50,x}$ is the required concentration of mixture component x to produce a maximum effect on 50% of the population; C_x = concentration of *mixture component x*; P' = average power of a single substance within a cluster. There are certain limitations of the mentioned model such as non-consideration of interactions, heavy data demanding approach, and biological knowledge-based (mechanisms of action) approach.

Integrated addition-interaction: Scientists Rider and LeBlanc first considered the effect of interactions in mixture toxicity and proposed an integrated addition-interaction (IAI) model in 2005 (Rider and LeBlanc, 2005). An experimentally derived K function is used to incorporate the effect of interaction. This is again a two-step process like the previous where the K function is determined 1st and then using that data, the overall toxic response of different clusters are calculated. The IAI model can be represented mathematically by the following equation 1.5-

$$R = 1 - \prod_{I=1}^{N} \left\{ 1 - \frac{1}{1+\frac{1}{\left(\sum_{i=1}^{n}\frac{K_X,i(C_X)\times C_i}{EC_{50i}}\right)^{P'}}} \right\} \qquad (1.5)$$

where, K_x,i is a function that describes the extent of a chemical (x) present after interaction among x and i. Cx and Ci are the concentration of components x and i in the mixture; EC_{50i} is the effective concentration of component i on 50% population. P' is the average power of chemicals within a group. For being an experiment-dependent and highly data-demanding procedure (that needs the dose-response curves of the mixture components to identify the interactions among each other), its application is very limited.

1.5.1.3. Major advantages and disadvantages of component-based approaches

The component-based approaches for mixtures toxicity assessment are age-old and time-tested processes; therefore, they are more frequently used. The techniques are associated with various advantages and limitations. One of the major advantages of component-based conventional modeling techniques is the ease of calculation. These models are not associated with any regression or classification-based complex statistical calculations. Apart from that, the component-based approaches are fairly efficient for predicting the mixture toxicity. One of the

the integrated modeling approaches considers the effect of interactions also and hence more reliable predictions can be obtained. Using component-based approaches, the frequency of the in vitro or in vivo experiments can be reduced as compared to the whole mixture-based approach.

The limitations are also very prominent in conventional component-based approaches. The major disadvantage is the requirement of experimentation. Although the frequency for experimentation is less than in the whole mixture-based approach. It needs experimental toxicity data of the mixture components for further calculations. Without complete knowledge of the mixture component and the composition, one cannot predict the mixture toxicity using the component-based approaches. Thus, the calculation of toxicity of untested/new mixture becomes costly, time-consuming, manpower and resource-driven in component-based approaches. The need of in vivo experiments also makes these approaches ethically complicated. The most used methods like CA and IA do not consider the effect of interactions and hence are not always reliable. On the other hand, more reliable approaches like integrated addition-interaction models are more data demanding and to determine the effect of interactions, it needs additional experimentations which further limit the application of the method.

1.5.2. Computational techniques in mixture toxicity assessment

Computational toxicology is a developing field of science which combines developments in molecular biology and chemistry with computational science for the prediction of various toxicity endpoints for single chemicals as well as mixtures (Raies and Bajic, 2016). There are still existing data gaps for the toxicity of single chemicals which are present in the environment. The case of mixture toxicity data is even worse. As we have discussed earlier, conventional toxicity assessments need in vivo or in vitro models for both the individual chemicals and mixtures as well. Animal experimentation is an ethically complicated and troublesome process which needs a lot of money and manpower. Computational methods are one of the promising techniques of toxicity prediction by using multiple chemometric tools and expert systems. These approaches are not at all the exact alternatives of animal testing; rather they are usually used to complement the experimentations by reducing the animal tastings. These cost-effective alternatives are green, time efficient, and can improve toxicity predictions and risk assessment. Additionally, in the case of mixture risk assessment, computational methods can predict the toxicity of any number of combinations in advance. Various regulatory authorities followed by the industries are now using the in silico techniques for the chemical risk assessment. For example, the use of QSAR is already accepted for hazard assessment by the European Union's 'REACH legislation'. Various commercial software systems like DEREK expert systems (DEREK), TOPKAT (Accelrys), MultiCase etc. are available for toxicity prediction of single chemicals. Organization for Economic Cooperation and Development (OECD) has developed a noncommercial software called OECD Toolbox for the same purpose (OECD, 2021). These expert systems are based on some statistical techniques and machine learning (ML) algorithms to relate various biological (LC_{50}, EC_{50}, IC_{50}, and so on), and physicochemical endpoints with molecular structures in the form of descriptors. The major advantages of in silico methods in mixture toxicity predictions are as follows (Kar and Leszczynski, 2019):

i. To efficiently predict the hazards posed by single chemicals and multi-component mixtures in the environment.
ii. The unethical practice of animal sacrifice can be minimized by the successful application of computational tools in the toxicity studies of chemicals and mixtures.

iii. To predict the toxicity of combinations where the component chemicals are present in very low concentrations (less than their EC_{50} of NOEC).
iv. Using the chemometric models of existing mixtures, toxicity prediction of new and unknown combinations against the specific species is possible if they fall under the applicability domain of the model.
v. In silico tools are fast, less costly, and reproducible as compared to in vivo or in vitro experiments.
vi. It is a reliable method to generate more toxicity data of multi-component chemical mixtures, and hence it leads to easier risk assessment of mixtures.

1.6. In silico predictive modeling approaches

After the introduction of computational technology in the mixture toxicity assessment, various in silico algorithms have been utilized. These algorithms generate predictive models by correlating the observed query responses with structural/biological aspects of the corresponding chemicals. Among the in silico approaches, Quantitative structure-activity relationship (QSAR), read-across (RA) predictions, Quantitative read-across structure-activity relationships (q-RASAR), and Machine learning (ML) algorithms are mostly used. During the development of in silico models, i) data gathering, (ii) descriptor generation, (iii) modeling, (iv) model evaluation, and (v) model interpretation are the general steps that are followed by all algorithms. The general steps are graphically represented in **Figure 1.2**.

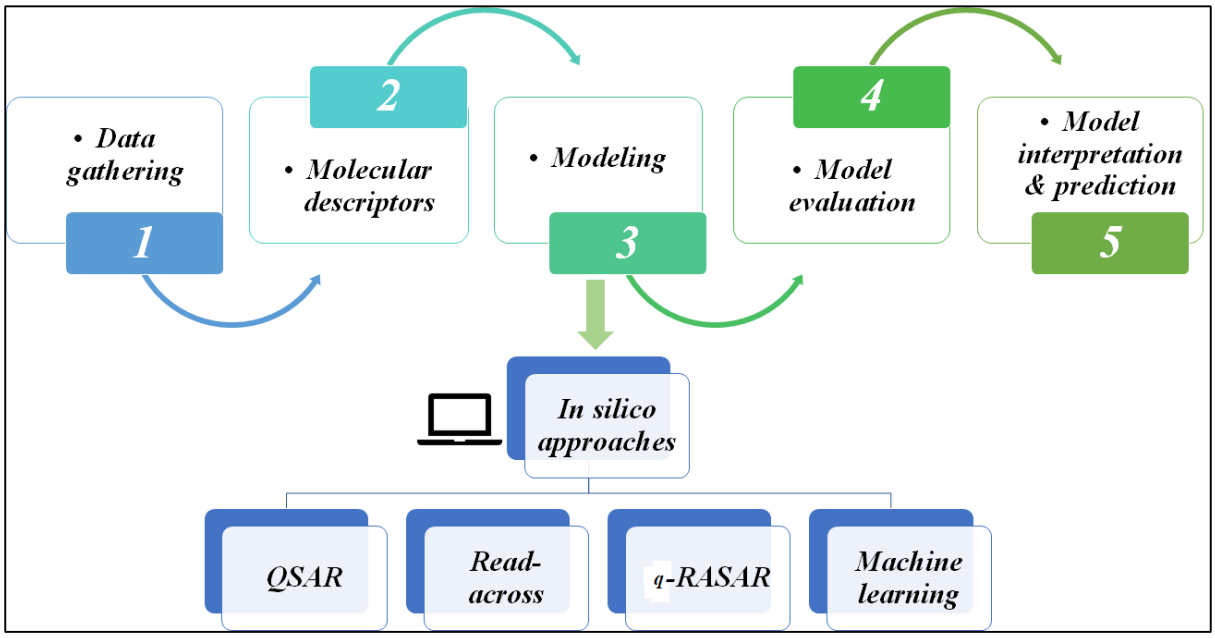


Figure 1.2. General steps of in silico modeling

1.6.1. Quantitative structure-activity relationship (QSAR)

Quantitative structure-activity relationship (QSAR) is the most commonly used statistical approach for the mixture toxicity assessment. It is a simple mathematical modeling technique that can correlate chemistry with specific molecular properties (biological/physicochemical/toxicological) using various computationally or experimentally derived quantitative parameters termed descriptors (Roy et al., 2015). A QSAR model is developed from the dataset of a series of compounds having a specific response. In the absence of experimental data, a compound's characteristics (having an acceptable relationship between

structural properties and activity) can be predicted by QSAR models. In simple terms, molecular responses are the function of a set of structural features which can be numerically encoded by molecular descriptors. Thus, QSAR is most commonly expressed by the following equation 1.6 -

Molecular response = (activity/toxicity/property) = *f*(quantitative structural information or properties) = *f*(Descriptors) (1.6)

1.6.1.1. Background and its brief history

Finding a definite starting point for QSAR is not an easy task because the basic idea has been transformed through various changes over a century from 1860 to the period of Hansch (Prucell et al., 1973; Rekker, 1992; Tute, 1990; van De Waterbeemd, 1992). Very few works were going on sparsely around the world at a very preliminary stage during the initial days. In 1863, the very first observation of the relationship between physicochemical and biological properties was reported in a thesis by A. F. A. Cros in University of Strasbourg (Borman, 1990). Cros found an increase in the toxicity of alcohols to mammals with decreasing aqueous solubility, up to a certain extent. In a separate study in 1868, A. Crum Brown and T. Fraser observed pronounced differences between the basic and the permanently charged quaternary nitrogen compounds during the study of the biological effects of certain alkaloids (including morphine, codeine, strychnine, atropine, etc) before and after methylation of a basic nitrogen atom. From this observation, they concluded that the paralytic property (X) of these compounds should be a function of the chemical constitution (Brown and Fraser, 1868). Although there was no way to encode the chemical structure by quantitative terms at that time, the researchers were unable to validate this relationship using other examples, but the general mathematical expression of the biological property proposed by them was of utmost importance. In the consecutive year, B.J. Richardson reported a proportional relationship between the narcotic property of alcohol and their molecular weights (Richardson, 1869). In 1893, M.C. Richet claimed that the toxicity of ethers, alcohols, aldehydes, ketones, etc., had an inverse relationship with their water solubility (Richet, 1893). During the juncture of the 19th and 20th centuries, many theories of narcosis and the effects of narcotic chemicals had come up. H.H. Mayer and C.E. Overton independently reported the relationship of toxicity with the partition coefficient of chemical compounds in this period (Lifnick, 1989; Lipnick, 1991). After these, there is hardly any document that states the significant development of QSARs in the first quarter of the 20th century.

A major contribution came in 1930 by L.P. Hammett in the domain of theoretical organic chemistry. From the ionization constants of benzoic acids and derivatives as standard, Hammett proposed two equations for the estimation of equilibrium constant *K* and reaction rate constant *k* of aromatic compounds as shown in the following equations 1.7 and 1.8 (Hammett, 1970)-

$$\log K_{R-X} - \log K_{R-H} = \rho\sigma \qquad (1.7)$$
$$\log k_{R-X} - \log k_{R-H} = \rho\sigma \qquad (1.8)$$

where ρ represents the reaction constant which is specific for the reaction type and σ represents an electronic constant which is a variable number and depends on the electron donating or withdrawing properties of substituents. This σ was also termed as Hammett electronic constant. In the subsequent years, some scientists made several attempts to apply the Hammett equation in biology to correlate biological activity (Hansen, 1962) with the Hammett electronic constant (σ) or by demonstrating a separate "biological Hammett equation" (equation 1.9) (Hansen, 1962; Zahradnik, 1962; Zahradník, 1962).

$$\log \tau_i - \log \tau_{Et} = \alpha\beta_i \qquad (1.9)$$

were, τ_i and τ_{Et} represent biological activities, α represents a constant term specific to the biological activity, and it resembles the ρ of the Hammett equation, and β_i represents the constant property contribution of substituent group X_i. However, due to the presence of unspecified biological activity terms in equation 1.9, it could not be applied in a generalized manner. Two other important contributions in this domain should be mentioned: (i) demonstration of group contribution of logarithmic response values by scientists T. C. Bruice, N. Kharasch, and R. J. Winzler through a series of thyroid hormone experiments in 1956 (Bruice et al., 1956), (ii) the demonstration of the derivation of substitution "enhancement factor" by scientists J. Fried and A. Borman in 1958 from the biological response data of several gluco- and mineralocorticoids (Fried and Borman, 1958).

1.6.1.2. Origin of classical QSAR: Contributions of Hansch, Fujita, Free, and Wilson

The modern QSAR was developed based on the linear free energy relationship in general, and the Hammett equation in specific. As mentioned earlier, many academicians mark 1962 as the foundation year of modern QSAR after the publication of a pioneering research paper by Hansch et al. (Hansch et al., 1962). This work was a general representation of the previous 15 years of research based on the Hammett electronic constant where we can observe several attempts to relate the substituent's effects on biological activity. In 1964, Hansch and Fujita published a research paper entitled "*ρ-σ-π* analysis. A Method for the correlation of biological activity and chemical structure", a preliminary attempt at classical QSAR which is used even today (Hansch and Fujita, 1964). Scientists Spencer M. Free and James W. Wilson published another remarkable research entitled "A mathematical contribution to structure-activity studies" in 1964, which also contributed to the development of classical QSAR (Free and Wilson, 1964). Hansch contributed especially to understanding the effect of lipophilicity of the substituent chemical species toward biological activity. He demonstrated lipophilicity as an additive, constitutive-physicochemical property of a molecule and derived a Hammett-type equation to estimate the lipophilicity of aromatic hydrocarbons as shown in equation 1.10:

$$\log P_{R-X} - \log P_{R-H} = \pi_X \qquad (1.10)$$

where π_X represents the lipophilicity contribution of the substituent X of a substituted aromatic hydrocarbon. Another notable input of Hansch was to use several parameters (which encode different properties of a chemical compound) in an equation instead of the use of a single parameter, after the proposal of Fujita (Hansch, 1969; Hansch and Fujita, 1964). The proposed multiparametric model has been given in the following equation 1.11 –

$$\log 1/C = a\,\pi + b\,\sigma + \cdots + constant \qquad (1.11)$$

This expression is popularly known as Hansch analysis and it is advantageous over its uni-parametric counterpart due to its flexibility and also the ability to interpret many biological activities. Scientists Free and Wilson also proposed another mathematical model during this time in an independent publication, and this model is popularly known as Free-Wilson formalism (Free and Wilson, 1964). According to this formalism, biological activity on a logarithmic scale is the summation of the biological activities of the base chemical and the attached substituent groups on it at different positions. The current form of that model is shown in the following equation 1.12, where a_i represents the contribution of substituents while μ represents the contribution of the base chemical to biological activity (Fujita and Ban, 1971; Kubinyi, 1988).

$$\log 1/C = \sum a_i + \mu \qquad (1.12)$$

The Free-Wilson formalism is an easy approach because it does not consider any physicochemical parameters into account. Binary values (0 and 1) are used to indicate the presence and absence of certain substituents on the base chemical moiety which indirectly correlate with the biological activity. However, there is a limitation of this model due to the low degree of freedom in the case of statistical analysis. It is also a more restricted approach than the Hansch analysis due to the use of many parameters which represent the substituent groups. Hansch model and Free-Wilson formalism can be combined to obtain a mixed approach. One such attempt was reported by Hugo Kubinyi in 1976 where the Hansch analysis has been used for the quantification of substituent's effects after a major structural change and Free-Wilson formalism has been used for a minor structural variation (Kubinyi, 1976). The mixed approach has been represented in the following equation 1.13 which improves the extent of applicability of QSAR models up to a certain limit.

$$\log 1/C = a\,\pi + b\,\sigma + \cdots + \sum a_i + constant \qquad (1.13)$$

1.6.1.3. Evolution of QSAR

As time passed, the importance of Hansch analysis to QSAR was getting established. In the meantime, the availability of resources for basic quantum chemical calculations started increasing which opened up an alternative arena for chemists for correlating the electronic and steric functionality of biological activity of closely related chemical congeners. However, the quantum chemical approach was not affordable in most parts of the world at that time due to the limitations of computations and expertise. In one of the first attempts, Pullman and Pullman demonstrated the relationships between the carcinogenic potential of aromatic hydrocarbons with their electronic structural features in the experimentation of DNA-reactive diol-epoxide intermediates for predicting the potential for bay region metabolic activation (Pullman and Pullman, 1955). Apart from the physicochemical properties and molecular electronic-steric properties, the concept of using small parts of molecules as structural features was initiated in the early 1980s. The idea of 2D fragment-based descriptors came after the publication of breakthrough research work by G. Klopman in 1984 where he produced a large number of computer-generated 2D constitutive fragments from a series of training set molecules (Klopman, 1984). Then he tried to correlate the number of each fragment separately with the biological activity to find suitable relationships between biology and chemical structure. This was a successful approach of that time because of its simplicity and ability to interpret the effect of structural features on the biological activity of a large set of compounds. From then onwards, scientists started working on highly diverse and all-inclusive data sets for developing so-called global QSARs. Accounting for the steric effect of substituent fragments was a major concern of preliminary QSAR scientists. In the early QSARs, Taft's steric constant (E_s) was used. Scientists like Hansch, Kutter, and Charton demonstrated, how Taft's *Es* constants are related to the radius of symmetric substituents (Charton, 1969; Kutter and Hansch, 1969). The use of molecular refractivity as a measure of molecular bulk was also shown by Hansch. Verloop and co-workers stretched this domain further and developed five different parameters to consider the bulk of substituent atoms or groups into account (Verloop et al., 1976). Comparative molecular field analysis (CoMFA) was a breakthrough in this domain where the electrostatic potential interaction energies of closely related congeners are taken into account (Cramer et al., 1988). This technique was first commercially used successfully during a 3D-QSAR analysis where the scientists attempted to correlate the molecular field interactions with biological activity by a partial least squares (PLS) regression (Wold et al., 1984).

With time, the available datasets were becoming structurally diverse, large in number, and more complex, and that is why the preliminary designed physicochemical descriptors were becoming insufficient to demonstrate the correlation between structural aspects of molecules and biological activity. Scientists were trying to use the indicator variable to make a distinction between two series of molecules (Martin and Lynn, 1971), but the general solution was to generate more descriptors that could efficiently represent various structural features of molecules. Besides the traditional QSARs with non-zero measurable biological activity values (quantitative QSARs), there are other datasets where only the binary qualitative data (active/inactive, present/absent, etc.) are available and the quantitative QSAR is an improper way to model such data. Discriminant analysis, logistic regression, and some other classification-based techniques were introduced to tackle such qualitative datasets (Martin et al., 1974). Several newer classification-based methods such as random forests (Hawkins et al., 1997), support vector machines (Byvatov et al., 2003), Bayesian classifiers (Ajay et al., 1998), etc. also came up with time and successfully applied to categorical datasets. Another aspect of QSARs is the statistical part which helps the researchers to assess the model quality, reliability, and external predictivity. Various preliminary works adopted different statistical metrics to justify the model quality and reliability of predictions. We can see the pioneering work of Topliss and Edward where they introduced the randomization technique to nullify the chance correlation (Topliss and Edwards, 1979). In more than 50 years of active development, the field of QSAR modeling has grown tremendously concerning the diversity of both methodologies and applications. The subsequent sections of this chapter provide more details about the stages of QSAR model development in the context of chemical mixtures.

1.6.1.4. Chemical information and descriptors

The QSARs are directly related to the molecular structure of chemicals, and they find the relationship between the structural features (encoded by molecular descriptors) and exerted molecular responses (biological/physic-chemical/toxicological). The structural aspects of molecules are represented as various structural features with specific numerical values and are termed molecular descriptors. These molecular descriptors are one of the core components of model development that serve a critical role in QSAR analysis. Todeschini and Consonni define molecular descriptor as "*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*" (Consonni and Todeschini, 2010). In simple words, the response variable of a QSAR is the function of structural features or descriptors. The quality of any QSAR model depends on the extent of encoding of structural features into descriptors values that correlate the exerted properties with structures. For successful encoding of molecular structures and good model development, the descriptors should ideally have the following properties (Randić, 1991)- (i) easy to calculate with the unambiguous algorithm, (ii) minimal degeneracy, (iii) easily interpretable, (iv) applicable to a broad range of compounds, (v) independent of experimentation, (vi) significantly correlated with the endpoint but very less correlated with other descriptors, (vii) ability to map the value into the molecular structure for better visualization, and (viii) reproducible.

In the very early days of QSARs, the physicochemical descriptors (electronic/ steric/hydrophobic) were mostly used. But the modeling technique has crossed a long way since the beginning, and nowadays various kinds of descriptors are used based on structural complexity (structural and topological descriptors), molecular orbital calculations (electronic), molecular surface area calculations (geometrical descriptors), Randic's brunching descriptors, Kier and Hall's flexibility descriptors etc (Roy et al., 2015). The dimensionality of chemical

structure is another important aspect of descriptor calculation. Dimensions generally signify the degree of descriptors used and thus the complexity of the modeling technique. Dimension can be defined as the minimum number of coordinates required to specify a point in chemical space (Katritzky et al., 2002). It is an inherent property of individual chemicals and independent of the chemical space or any other external fields. The use of added dimensions in descriptor calculation helps to extract a higher level of chemical information which further assists in pursuing a higher degree of QSAR analysis (Katritzky et al., 2002). Based on the dimensionality, 8 classes of descriptors (0D – 7D) can be calculated by different algorithms. The 0D – 1D descriptors are obtained from the molecular formula only and termed the constitutional index. These are the simplest among all descriptors and compute very limited information such as the number and type of atoms used, the number and type of bonds, molecular weight, etc. 2D- and 3D- descriptors are the most popular and widely used for QSAR model development. 2D-descriptors are generally obtained from simple 2D- molecular graphs, thus devoid of 3D- energy minimization and conformational analysis. Although the 2D descriptors can efficiently encode the structural features, these are easy to calculate and interpret; thus these are most commonly used in descriptive QSARs. On the contrary, the 3D-descriptors are more complex than 2D ones and need energy-minimized 3D structures as input for descriptor calculation. Due to topological similarities among compounds, sometimes 2D-descriptors cannot distinguish different compounds separately; 3D-descriptors come in handy in those situations and are used mostly in predictive QSARs nowadays. The three-dimensional descriptors are the computed atom-based numerical parameters that are obtained from the spatial arrangements (3-dimensional representation) of molecular structures. The 3D-QSARs are the extension of classical QSARs proposed by pioneer scientists Hansch, Fujita, Free, and Willson, where the 3D-descriptors are employed to correlate the structural aspects of the molecules with exerted molecular properties. Different types of descriptors have been shown based on their dimensionality in the following **Table 1.1**.

Table 1.1. Classification of descriptors based on their dimensionality

Dimensionality of descriptors	Description	Representative Example	References
0D – 1D	These descriptors are derived from molecular formulas and different sub-structural fragments	Constitutional indices, bond counts, atom counts, molecular weight, fragment counts, fingerprints	(Todeschini and Consonni, 2009)
2D	These descriptors are obtained mainly from 2D-molecular graph theoretical representation along with some physiochemical indices.	Structural descriptors, topological descriptors, physicochemical descriptors, thermodynamic descriptors	(Todeschini and Consonni, 2009)
3D	These descriptors are obtained after giving static 3D structures of the molecules as input. Various geometrical and spatial	Quantum chemical descriptors, molecular surface area (MSA) parameters,	(Guner, 2005; Todeschini and

	features are taken into account.	molecular field analysis (MFA) parameters, WHIM descriptors, Jurs parameters, GATEWAY descriptors, CoMFA parameters, CoMSIA parameters, MoRSE descriptors, geometric descriptors, electronic parameters, etc.	Consonni, 2009)
4D	Here, multiple orientations of 3D structures of ligand molecules are used to generate the descriptors. The configuration, conformation, and protonation states are also varied.	GRID, Raptor, and Volsurf derived descriptors.	(Hopfinger et al., 1997)
5D	Here, the receptor characteristics are also taken into account along with the ligand to establish a ligand-based virtual receptor model. The induced-fit parameters are considered.	Flexible protein docking.	(Vedani and Dobler, 2002)
6D	Various solvation states are considered with the 5D information.	Quasar	(Vedani et al., 2005)
7D	In this case, real receptor binding scenarios or target-based modeling data are considered.	-	(Polanski, 2009)

1.6.1.5. The standard algorithm of QSAR modeling

QSAR is a process of finding a mathematical relationship between the structural or chemical aspects of a molecule with the exerted biological/toxicological/physicochemical properties. Previously, there was no uniform protocol for developing QSAR models, and that is why it was very difficult to find the merits and predictivity of developed models. Validation of a QSAR model was a key element for being recognized as an acceptable and reliable model (Balls et al., 1995). In recent times, model validation has been accepted as the most holistic approach for assessing the quality, reliability, and applicability of a developed model (Aptula et al., 2005). But some basic questions may arise such as

(i) What should the validation approach be followed to justify the quality, predictivity, and reliability of developed models?
(ii) What are the criteria for acceptance of the predicted value from a QSAR model?
(iii) What are the eligibility criteria of a QSAR model to be used for regulatory purposes?
(iv) Can a QSAR model predict the activity/toxicity value of new or untested chemicals?

Experts from the Organization of Economic Cooperation and Development (OECD) member countries have formulated five generic principles related to QSAR model development which provide an unambiguous idea about the workflow of QSAR model development and thus, suggest the answers to the above questions as well (OECD, 2014). The principles are as follows-

OECD principle 1 (A defined endpoint)- The endpoint of each QSAR model must be properly defined along with the unit and experimental condition used to obtain the data. Any ambiguity of the endpoint value during data collection may produce a prediction error.

OECD principle 2 (An unambiguous algorithm)- The unambiguous algorithm is the second important criterion that has to be documented properly (each step should be reported clearly) to guarantee transparency and reproducibility of the results.

OECD principle 3 (A defined domain of applicability). Applicability domain analysis is performed to specify a theoretical space in which the predicted response from a developed model is reliable.

OECD principle 4 (Appropriate measure of goodness-of-fit, robustness, and predictivity)- A proper validation of a developed QSAR model is necessary to clarify the goodness-of-fit, robustness, and predictivity. Internal validation with training set molecules is carried out to evaluate the goodness-of-fit and robustness of the model, and external validation with test set molecules is carried out to evaluate the predictivity of the model.

OECD principle 5 (A mechanistic interpretation, if possible)- The mechanistic interpretation can be given to provide a probable idea about the working principle of the chemical.

A standard QSAR modeling approach following the OECD protocol has been explained with some of the key points in the following section and a graphical workflow has also been given in **Figure 1.3**.

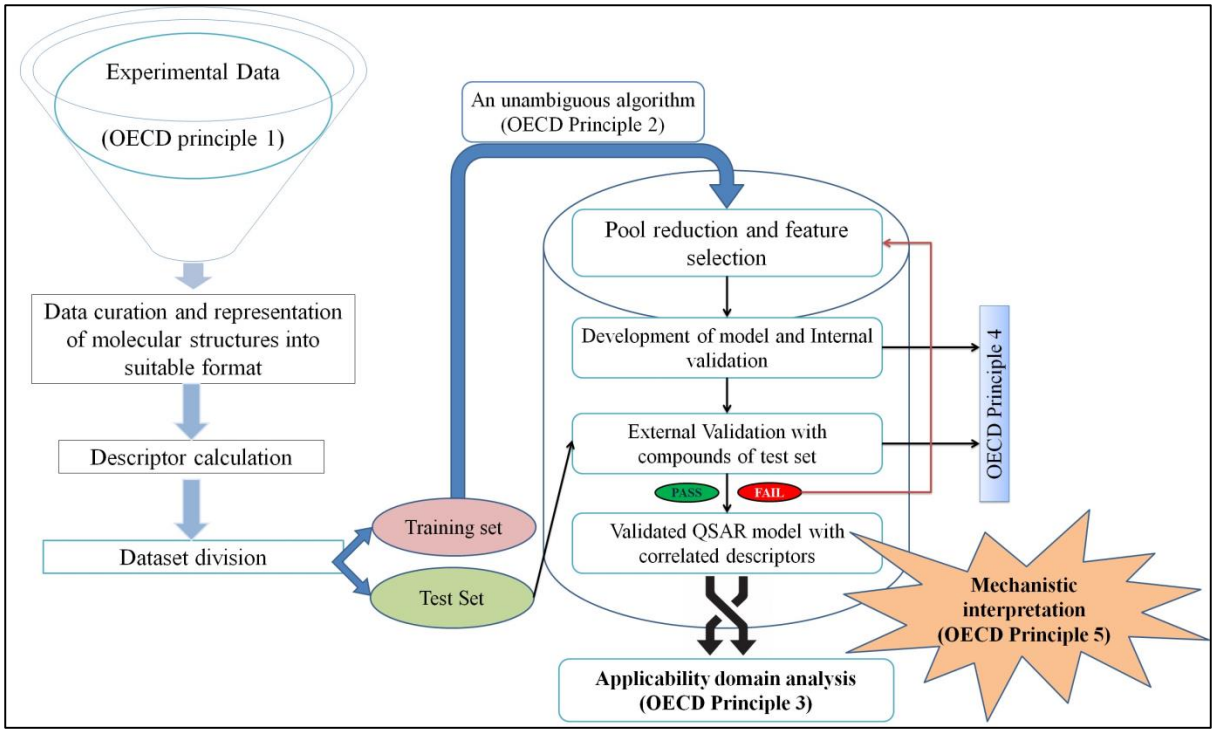


Figure 1.3. Standard protocol of QSAR development following the OECD guidelines

Data curation and descriptor calculation

One of the basic assumptions of QSAR or any cheminformatics study is the correctness of input data. To develop QSAR models, chemical structures and corresponding activity/property data are the fundamental inputs that are collected from databases or scientific literature. The values of the response variable (activity/property data) are primarily obtained from the in vitro/in vivo experimentations whereas the chemical structures, CAS numbers, names, etc. of molecules are primarily collected from the databases. The error may happen during the experiment itself which cannot be corrected by the QSAR modelers. However, the simple human errors that occur during the data collection can be avoided to a great extent. The incorrect and/or imprecise input data in the modeling sets is increasingly considered a major concern for building in silico models. In 2005, Olah et al. (Olah et al., 2005) demonstrated the presence of at least two chemical annotation errors in each medicinal chemistry publication. They also reported an overall error rate as high as 8% for the chemicals indexed in the commercial database WOMBAT (Olah et al., 2008). Apart from these findings, the importance of data curation has also been explained in some published articles (Fourches et al., 2010; Olah et al., 2004; Young et al., 2008). The descriptor values calculated from the incorrect structural data are erroneous and cause detrimental effects on model performance. Thus, rigorous data curation and removal of questionable data points often lead to an improvement in model predictivity. Fourches and co-workers (Fourches et al., 2010) developed a guideline of the best practice of data curation where they suggested removing the inorganics, organometallics, counterions, and mixtures from the input data. Most descriptor-calculating software cannot handle organometallics. Mixtures and salts may lead to confounding duplicates when simplified (e.g., desalted). Additional curation elements include structural cleaning (e.g., detection of valence violations), ring aromatization, normalization of specific chemotypes, and standardization of tautomeric forms. Postprocessing entails the deletion of duplicates resulting from curation, standardization and normalization, and manual checking of complex cases.

Descriptor calculation for the conventional single molecular QSAR is a straightforward approach. Various software tools like Dragon, PaDEL Descriptor, SiRMS, and Alvadesc etc. are employed to compute molecular descriptors. However, the calculation of mixture descriptors is a tricky process. The weighted descriptor approach is mostly applied to obtain the mixture descriptor values from the descriptors of component chemicals of a mixture (Kar and Leszczynski, 2019). According to this strategy, the mixture descriptors can be obtained by multiplying the original descriptor values of a mixture component with the percentage ratio of that component in the mixture and finally summing up for all components present in the mixture. The approach can be simplified as Equation 1.14 where $D_{mix}$is the estimated value of the mixture descriptor, x_i is the percentage ratio of a specific component, D_i is the original descriptor value of that component:

$$D_{mix} = \sum(D_i \times X_i) \qquad (1.14)$$

The sum of the percentage ratios of all components should be 100% for a mixture. Apart from weighted descriptors, some other approaches are reported in the literature such as (i) integral additive descriptors (weighted sum of descriptors of pure components), (ii) integral non-additive descriptors of mixtures, (iii) fragment non-additive descriptors (various structural fragments are taken into consideration) (iv) descriptors based on the partition coefficient of mixtures, and miscellaneous descriptors for mixtures (Kar and Leszczynski, 2019). Qin *et al.*utilized various mixing rules to develop a QSAR model for the prediction of the toxicity of binary and multi-component mixtures composed of two antibiotics (tetracycline hydrochloride and chloramphenicol) and four pesticides (metribuzine, trichlorfon, dichlorvos, and linuron) against *Aliivibrio fischeri* (Qin et al., 2018). The authors considered the additive and non-additive (synergism and antagonism) effects of mixtures during model development and calculated the mixture descriptors by following seven different mixing rules besides the weighted descriptors published in various other literature (equation 1.15 to 1.21) (Altenburger et al., 2003; Chatterjee and Roy, 2022; Gaudin et al., 2015; Hoover et al., 2019; Qin et al., 2018; Sobati et al., 2016) to encode the nonlinear additive effects of mixtures -

$$D_{mix,i} = (\sum_{i=1}^{n} p_i x_i)^2, \qquad (1.15)$$

$$D_{mix,i} = \sqrt{\sum_{i=1}^{n} (p_i x_i)^2}, \qquad (1.16)$$

$$D_{mix,i} = \sum_{i=1}^{n} \sqrt{p_i x_i}, \qquad (1.17)$$

$$D_{mix,i} = \sum_{i=1}^{n} \sqrt{|x_i|}, \qquad (1.18)$$

$$D_{mix,i} = \sum_{i=1}^{n} p_i (x_i)^2, \qquad (1.19)$$

$$D_{mix,i} = \sum_{i=1}^{n} (p_i)^2 x_i, \qquad (1.20)$$

$$D_{mix,i} = \sqrt[3]{\sum_{i=1}^{n} (p_i)^3 x_i}, \qquad (1.21)$$

where $D_{mix,i}$ represents the overall mixture descriptor, i represents the ith component in a mixture, and x_i represents the descriptor of ith component.

Dataset division

Dataset division is important in QSAR from the perspective of model validation. For performing the external validation, sufficient new experimental data is not always available.

Thus, the dataset is divided into training and test sets; the training set is used for developing the model and internal validation, whereas, the test set is exclusively used for external validation. The dataset should be divided in such a way that the training and test set compounds are scattered well in the whole multi-dimensional descriptor space (Golbraikh and Tropsha, 2000; Guha and Jurs, 2005). Each test set compound should be in the vicinity of at least one training set compound so that the theory of similarity can be employed for predicting the properties of test set compounds (Golbraikh and Tropsha, 2000). There is no fixed ratio of the number of training and test set compounds. For a dataset with sufficient data points (n>50), 70-80% of data is kept in the training set whereas 20-30% of data is kept in the test set (Roy et al., 2008). The ratio of numbers of training and test sets is usually selected considering the "*Bias-variance dilemma*" (Almaliki, 2018; Kononenko and Kukar, 2007). Various approaches are used by the QSAR modelers for the selection of training and test set compounds. Those data division approaches can be categorized into three broad domains- (i) Random division, (ii) Y-response-based division, and (iii) X-response-based division. In the random selection, the training and test set compounds are selected randomly without following any specific algorithm. It is the simplest and least biased approach. The Y-response-based selection is performed after sorting the activity/property/toxicity values; thus it is also called "Sorted activity-based division" (Carlson and Carlson, 1992). The data points are classified into some clusters based on sorted activity and then the dataset is divided by random sampling from different clusters. It is a popular and widely used strategy among others. In the X-response-based approaches, the training and test set compounds are selected by considering the descriptor values only. There are various X-response-based approaches such as, "Kennard-Stone selection" (Snarey et al., 1997), "Euclidean distance-based selection", "Kohonen's self-organizing map selection" (Gasteiger and Zupan, 1993), "K-means clustering selection" (Roy et al., 2008), etc. which are often used by QSAR modelers.

Feature selection

Feature selection is one of the important steps in QSAR modeling where the most correlated features with response variables are identified from a large number of computed descriptors. Several approaches are used for selecting the most important descriptors among which stepwise selection, all possible subset selection, genetic algorithm, etc. are the most popular. Apart from these, factor analysis (Franke, 1984), particle swarm optimization (Kennedy and Eberhart, 1994), ant colony optimization (Dorigo and Stutzle, 2004), K-nearest neighbourhood method (Zheng and Tropsha, 2000), etc. are also used.

Stepwise selection (Darlington, 1990)*:* It is one of the popular approaches for descriptor selection. In this method, the descriptors are selected based on some stepping criteria. (1) The descriptor from the pool is initially selected based on the maximum correlation with the response variable. (ii) In the next step, one additional descriptor is added to the 1st descriptor in such a way that the contribution of the incoming descriptor is maximum. Here, the objective criterion is the *F*-for inclusion where *F* is the square of the *t-value* of the new descriptor added. (iii) The same process is repeated until no suitable descriptor is selected with the stepping criteria. During descriptor selection, some of the preliminary set of descriptors may be insignificant. Those descriptors are then eliminated with F-for exclusion criteria.

All possible subset selection: All possible subset selection is a computationally exhaustive method of variable selection. It is used when the number of descriptors in the pool is not very high. Otherwise, the number of descriptors is reduced before all possible subset selections.

Various screening filters such as determination coefficient cut-off and cross-validated variance cut-off can be used for the selection of the best combination.

Genetic algorithm (Fan et al., 2001; Rogers and Hopfinger, 1994): Genetic algorithm was first used by scientists Rogers and Hopfinger in 1994. It is a more efficient technique with many merits compared to other feature selection strategies. It simultaneously explores the whole region of the vector space defined by the training set descriptors. As a preliminary step, a population of *N* descriptors is created and the fitness score of all descriptors is determined. In the next step, a part of a new generation of descriptors (children) is produced by cross-over, and the rest are produced by mutation from the parent descriptors based on their scaled fitness scores. The new generation of descriptors must contain the features of two or one of its parent descriptors. These selections, cross-over, and mutation processes are repeated until all *N* parent descriptors are replaced by new generation *f* descriptors. The fitness score of each descriptor of the new generation is calculated, and the cycles continue until 80% of new descriptors achieve the target fitness score.

Statistical modeling approaches

Quantitative structure-activity relationship models can be developed using several statistical approaches. As a whole, these statistical modeling techniques are classified into two broad categories namely- (i) Regression-based modeling, and (ii) Classification-based modeling. Apart from these classical approaches, artificial intelligence-based machine-learning algorithms are the most popular and widely used nowadays. The regression-based models are mostly used among the classical modeling approaches and use continuous data as input. Multiple linear regression (MLR) (Snedecor and Cochran, 1967), Partial least squares (PLS) regression (Wold et al., 2001), Principal component regression analysis (PCRA) (Frank and Friedman, 1993), and Ridge regression (Horel, 1962), etc. are the common regression-based modeling approaches. The classification-based models are usually developed with discrete data when only binary responses are available. Linear discriminant analysis (LDA) (Agresti, 1996), Logistic regression (Harell, 2001), etc. are the common types of classification-based modeling approaches. Here, in the limited space of this chapter, we have briefly discussed the most popular and widely used modeling approaches. The readers may consult the cited references for more detailed knowledge.

Multiple linear regression (MLR): Multiple linear regression is a simple and reproducible regression technique that can find the linear relationship between a dependent variable and multiple independent variables. This method is widely popular due to its simplicity and ability to handle multiple descriptors (independent variables) simultaneously. The general mathematical expression of an MLR equation is as follows (equation 1.22)-

$$Y = c_0 + c_1 \times D_1 + c_2 \times D_2 + \cdots + c_n \times D_n \qquad (1.22)$$

where *Y* is the predicted response, D_1, D_2, …, D_n are the modeled descriptors with c_1, c_2, …, c_n regression coefficients respectively. There are various advantages of MLR such as easy implementation, the ability of interpretation, and reproducibility. There are a few limitations of an MLR model such as the inability to handle noisy and colinear data, and more chances of overfitting. To overcome all these limitations, the partial least squares regression method was devised.

Partial least squares (PLS): PLS regression is the generalization of MLR. The fundamental concept of PLS lies in the extraction of latent variables from the sampled factors and the

response. The extracted variables are accounted for the manifestation of factor variation and then are utilized for modeling the response. From the X space (descriptors), X-scores (latent variable - T) are extracted which are further used to calculate the Y-scores (latent variable - U). Finally, the Y-scores (U) are used to derive the predictions from the response variables. The number of latent variables used for modeling the response should always be less than the number of descriptors used. A PLS model can handle noisy, collinear, and correlated data well, and it is reproducible also. Besides these, simultaneous regression of multiple Y- variables (response) is possible with PLS. Chances of overfitting can also be avoided using this approach.

Principal component regression analysis (PCRA): Principal component regression analysis is a data reduction strategy followed by regression where instead of regressing the descriptors directly to the response, the principal components of descriptors are used as regressors. This is mostly used instead of MLR to overcome the multicollinearity problem. The principal components are the identifiers of the direction in which data varies. Mathematically, the principal components are identified by computing the eigenvalue and eigenvectors of the data covariance matrix. To perform PCRA, both the independent and dependent variables need to be standardized by the following equation 1.23-

$$X_{std} = {x_i - \bar{x}}/{x_{stdev}} \qquad (1.23)$$

where X_{std} is the standardized value of x, x_i is the original value of the ith component, $\bar{x}$ is the mean of x variable values, and x_{stdev} is the standard deviation of x variable values.

Linear discriminant analysis (LDA): Linear discriminant analysis is a form of classification-based QSAR modeling that classifies the data into two or more categories based on their response, and then fits the classified response data with the continuous values of descriptors in an equation. The LDA does the same task as MLR but with a classified response. In the following equation 1.24, the expression of the general LDA equation has been represented.

$$DF = c_1 \times D_1 + c_2 \times D_2 + c_3 \times D_3 + \cdots + c_m \times D_m + a \qquad (1.24)$$

where DF represents the discriminating function which is the sum of discriminating variables, c_1, c_2, c_3, …, c_m represent the discriminant coefficient of the predictor variables, D_1, D_2, D_3, …, D_m are the respondent's score of the descriptors, m is the number of descriptors.

Validation protocols

With the improvement of computing power and modern technology, predictive in silico models are now widely available. Although the chances of over-fitting, chance correlation, and less predictability of the models cannot be avoided. According to the OECD principle 4, each QSAR model should have a measure of goodness-of-fit, robustness, and predictivity for being acceptable (OECD, 2014). These criteria are evaluated by model validation which is a fundamental requirement of any QSAR model. Fitting of data and robustness are not the sole criteria, rather predictability is the key feature that judges the quality of a QSAR model. The predictability represents how well a model can predict the properties of a chemical which does not belong to the training set.

For validating a QSAR model, two major strategies are followed namely- internal validation, and external validation. Internal validation is performed with the training set compounds and evaluates the goodness-of-fit and robustness of the model. External validation is performed for evaluating the predictability of the model with the test set compounds only. There is another validation namely true external validation which is performed with an external data set against the same endpoint. However the availability of a new experimental dataset against the same endpoint is not always possible and in that case, external validation with the test set is considered as the true external validation. Various statistical metrics are used for model validation. Different sets of validation metrics are used for regression-based models and classification-based models. Here in **Table 1.2**, the description and mathematical formula of regression and classification-based metrics (Roy and Mitra, 2011) are discussed.

Table 1.2. Mathematical formulae and description of validation metrics used in the QSAR modeling study

Regression-based validation metrics		
Metrics	**Formula**	**Description**
$\boldsymbol{R^2}$	$R^2 = 1 - \frac{\sum(Y_{obs} - Y_{cals})^2}{\sum(Y_{obs} - \overline{Y_{training}})^2}$	The determination coefficient is a measure of goodness of fit. For a good model, the R^2 value should be 1 or near 1. Deviation from the ideal value represents the deterioration of model fitness. Y_{obs}= observed response, Y_{calc} = fitted response, $\overline{Y_{training}}$ = mean of the observed responses of the training set
$\boldsymbol{R^2_{Adj}}$	$R^2_{Adj} = \frac{\{(n-1) \times R^2\} - p}{n - p - 1}$	The adjusted determination coefficient is a measure of goodness of fit. The R^2 value increases with the increasing descriptor count in the model, thus a high R^2 can be a result of the overfitted model also. To penalize the overfitted models, the number of descriptors is considered along with the number of data points in the calculation of adjusted R^2. The ideal value of R^2_{Adj} is 1 or near 1. n = number of data points p= number of descriptor R^2 =determination coefficient
$\boldsymbol{Q^2 (Q^2_{LOO})}$	$Q^2 = 1 - \frac{\sum(Y_{obs(Train)} - Y_{calc(Train)})^2}{\sum(Y_{obs(Train)} - \overline{Y_{training}})^2}$	The cross-validated correlation coefficient is a measure of robustness. Leave-one-out (*LOO*) cross-validation is mostly

		performed where one data point is removed from the training set and predicted by the model developed from the rest of the molecules. Leave-many-out (*LMO*) cross-validation can also be possible where a set of molecules are removed from the training set. The ideal value of Q^2_{LOO} should be 1 or near 1 and any deviation from the ideal value suggests the deterioration of the model quality. $Y_{obs(train)}$ = training set observed response $Y_{calc(train)}$ = training set fitted response $\overline{Y_{training}}$ = mean of the observed responses of the training set
$RMSE_c$	$RMSE_c = \sqrt{\frac{\sum(Y_{obs(train)} - Y_{calc(train)})^2}{n_{train}}}$	Root mean square error of calibration is an internal validation metric which is a measure of prediction error in the training set by the model. Ideally, the computed value should be as low as possible. $Y_{obs(train)}$ = training set observed response $Y_{calc(train)}$ = training set fitted response N_{train} = number of training data points
$\boldsymbol{Q^2_{F1}}$	$Q^2_{F1} = 1 - \frac{\sum(Y_{obs(test)} - Y_{calc(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{training}})^2}$	Q^2_{F1} is the external correlation coefficient that evaluates the predictability of the model. The ideal value of Q^2_{F1} should be 1 or near 1 and any deviation from the ideal value suggests a less predictive model. $Y_{obs(test)}$ = test set observed response $Y_{calc(test)}$ = test set fitted response $\overline{Y_{training}}$ = mean of the observed responses of the training set
$\boldsymbol{Q^2_{F2}}$	$Q^2_{F2} = 1 - \frac{\sum(Y_{obs(test)} - Y_{calc(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{test}})^2}$	Q^2_{F2} is another external correlation coefficient that evaluates the predictability of the model. The ideal value of Q^2_{F1} should be 1 or near 1 and any deviation from the ideal value suggests a less predictive model. It is dependent on the number of the training set. $Y_{obs(test)}$ = test set observed response

		$Y_{calc(test)}$ = test set fitted response $\overline{Y_{test}}$ = mean of the observed responses of the test set
CCC	$CCC = \frac{2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - \bar{y})}$	CCC is the concordance correlation coefficient that is a measure of the reliability of external set predictions obtained from the model. CCC also measures precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. In ideal cases, both lines should coincide and thus CCC value should be 1. A lower value of CCC suggests a deviation of the regression line from the concordance line. x_i = observed response of i^{th} test compound y_i = fitted response of i^{th} test compound $\bar{x}$ = mean of observed response value of test compounds $\bar{y}$ = mean of predicted response data of test compounds
$RMSE_p$	$RMSE_p = \sqrt{\frac{\sum(Y_{obs(test)} - Y_{calc(test)})^2}{n_{test}}}$	Root mean square error of prediction is an external validation metric which is a measure of prediction error in the test set by the model. Ideally, the computed value should be as low as possible. $Y_{obs(test)}$ = test set observed response $Y_{calc(test)}$ = test set fitted response N_{test} = number of training data points

Classification-based validation metrics

Metrics	Formula	Description
Sensitivity (%)	$Sensitivity = \frac{TP}{TP + FN}$	A greater value of sensitivity (near 100%) suggests a good classification
Specificity (%)	$Specificity = \frac{TN}{TN + FP}$	A greater value of sensitivity (near 100%) suggests a good classification

Precision (%)	$Precision = \frac{TP}{TP + FP}$	A greater value of sensitivity (near 100%) suggests a good classification
Accuracy (%)	$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$	A greater value of sensitivity (near 100%) suggests a good classification
F-measure (%) (harmonic mean of recall)	$F - measure(\%) = \frac{2}{\frac{1}{Precision} + \frac{1}{Sensitivity}}$	*F*- measure usually evaluates the model's ability to classify the chemicals into active and inactive classes. It is also termed as the harmonic mean of recall and precision. A higher value of the F-measure suggests greater ability for classification.
G-means (geometric mean)	$G - means = \sqrt{Specificity \times Sensitivity}$	*G*-means combines sensitivity and specificity by geometric mean which is a measure of model's ability to classify positive and negative compounds.
Cohen's kappa (k)	$P_r(a) = \frac{(TP + TN)}{(TP + FP + TN + FN)}$ $P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2}$ $Cohen's\ K = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$	Cohen's *k* is a measure of agreement between known classification and predicted classification. It may have a value ranging from -1 to 1. The resultant Cohen's *k* values between -1.0 and 0.4 suggest a poor model; values between 0.4 and 0.6 suggest an average model; values between 0.6 and 0.8 indicates an acceptable model; and values between 0.8 and 1.0 imply a highly predictive model.
Matthews correlation coefficient (MCC)	$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$	MCC is the correlation coefficient between observed and predicted responses in binary classification. It may have values from -1 to 1 where 1 represents perfect classification, 0 represents the average and random classification, and -1 represents an inverse classification.

**TP = true positive, TN = true negative, FP = false positive, FN = false negative.

1.6.2. Read-across predictions

Read-across (RA) predictions are derived adopting a similarity-based mathematical approach where the activity/property/toxicity of unknown chemicals can be predicted from the known activity/property/toxicity data of structural/biological/functional analogues (Dimitrov and Mekenyan, 2010; Jeliazkova et al., 2010; Modi et al., 2012; Venkatapathy and Wang, 2013). It is accepted by REACH and the US EPA and used for filing the data gaps of thousands of compounds (Vink et al., 2010). Before accepting the RA predictions, the knowledge of the chemical category is important. A chemical category can be defined as a class of structurally similar chemicals which follow a similar pattern of biological action and show similar toxicity or properties. The chemicals in a specific category are called source chemicals which are used for the toxicity estimation of unknown chemicals (target compounds) of similar category. There are two basic ways to develop a read-across model i.e., (i) analogue approach and (ii) category approach (Jeliazkova et al., 2010; Venkatapathy and Wang, 2013). The analogue approach is the one-to-one prediction method where the unknown toxicity of a compound is predicted from a known analogue. This method is not reliable because of its sensitivity to outliers. Two different analogues may have different toxicity patterns, and thus, the chance of false prediction is higher. The category approach is a more confident prediction where many analogues are used to determine a trend among categories and predict the toxicity of unknown analogues. Read-across can be used with both the qualitative and quantitative data. The quantitative read across is mostly used (Gajewicz, 2017; Gajewicz et al., 2017). The read-across prediction approach has been graphically represented in **Figure 1.4**.

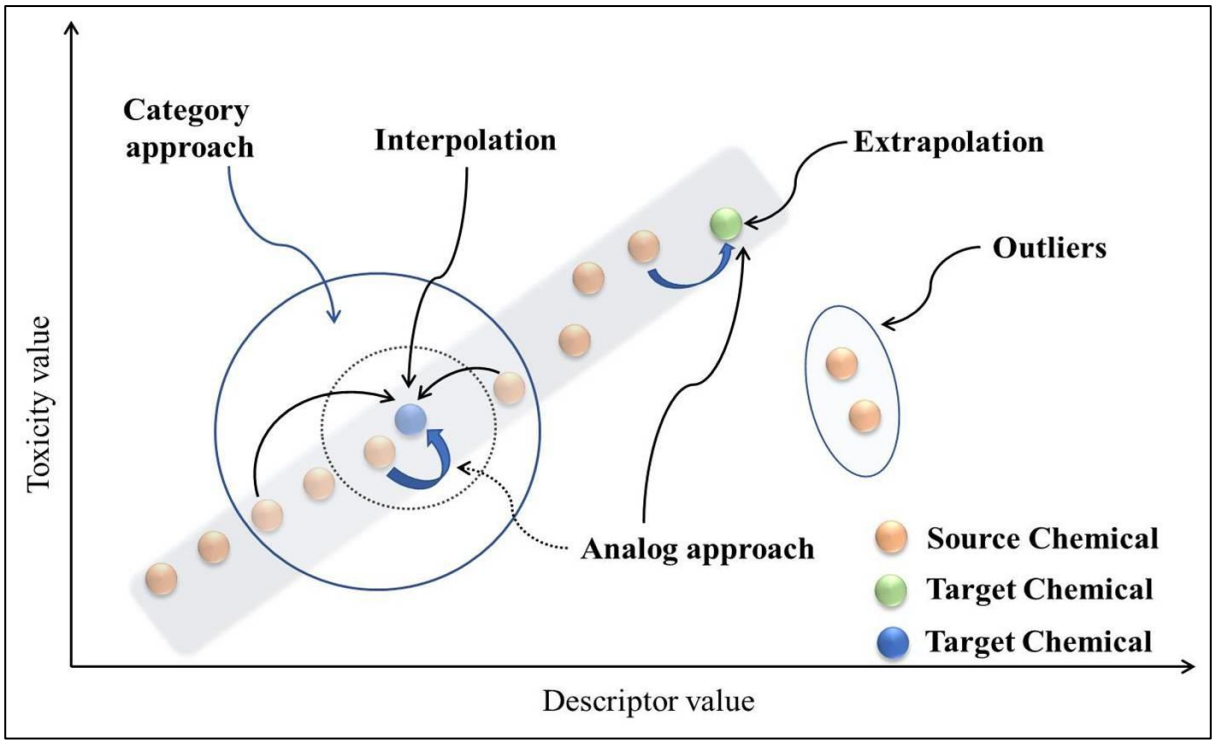


Figure 1.4. Graphical representation of Read-across prediction

1.6.2.1. Similarity estimation techniques used in read-across predictions

The similarity among compounds of the same chemical category must be established for successful read-across predictions. According to the Read-across assessment framework (RAAF) of the European Chemical Agency (ECHA), structural/biological/functional

similarities must be proven before the read-across argument. However, for quantitative read-across predictions, the estimation of structural similarity between source and target chemicals is mostly performed. As per the report published by Chatterjee et al., the Euclidean distance (ED)-based, the Gaussian kernel (GK) function, and the Laplacian kernel (LK) function similarities are mostly used for chemical category identification (Chatterjee et al., 2022).

Euclidean distance-based similarity: Euclidean distance is the perpendicular distance between two points in chemical space. This distance can be measured by the Pythagorean theorem and the formula is expressed in the following equation 1.25-

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad (1.25)$$

where, Euclidean distance between two points X and Y has been represented by *d(X,Y)*, whereas, X_i is the i^{th} descriptor of compound X, and Y_i is the i^{th} descriptor of compound Y (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.

Gaussian kernel function similarity: Gaussian kernel function is another approach which is employed to find out the most similar training compounds for each test set compound. Gaussian kernel function is a radial basis kernel function (Roy et al., 2013) which is defined by the following equation 1.26-

$$Similarity\ (X,Y) = f = e^{-\frac{\|X_i - Y_i\|^2}{2\sigma^2}} \qquad (1.26)$$

where, *Similarity (X, Y)* is the similarity 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; σdetermines the width of the Gaussian and can never become zero. We have estimated the similarity using various positive values of σ (0.25, 0.5, 0.75, 1.0, 1.5, 2.0).

Laplacian kernel function similarity: Laplacian kernel function is another radial basis function which is used for the similarity estimation is this study. The Laplacian kernel function (Chen et al., 2011) is defined as following equation 1.27-

$$\kappa(X,Y) = e^{(-\gamma\|X-Y\|_1)} \qquad (1.27)$$

where, $\kappa(X,Y)$is the similarity between two compounds X and Y; γ is a non-zero and positive number which determines the performance of the kernel. In this study, we have estimated the similarity using various values of γ (0.25, 0.5, 0.75, 1.0, 1.5, 2.0). $\|X\text{-}Y\|_1$ is the L^1 norm or Manhattan distance which can be defined as the distance between two points (here two compounds) measured along axes at 90º angle. If X and Y are two points with (x_1, x_2) and (y_1, y_2) descriptors respectively, the $\|X\text{-}Y\|_1 = (x_1 - y_1) + (x_2 - y_2)$.

1.6.2.2. Advantages and limitations of Read-across predictions

The read-across prediction has certain advantages like

i. Read-across predictions can be used for both qualitative and quantitative data.
ii. It is a simple and transparent technique.
iii. Implementation is very easy.

The limitations of the method are as follows-

i. RA can be used for small datasets in practice because the number of analogues is limited for a target chemical.
ii. No mathematical model is generated in read-across prediction and thus, the importance of structural features cannot be understood properly.
iii. Inaccurate predictions may be possible in the case of analogues having conflicting toxicity profiles.

1.6.3. Read-across structure-activity relationship (RASAR)

Read-across structure-activity relationship (RASAR) is a thoughtful combination of QSAR and read-across, purposefully developed to improve the external predictability of the model further. For being a statistical approach, QSAR cannot be used for modeling small data sets; on the contrary, read-across cannot provide any mathematical equation for further studying the impact of structural features. To avoid these individual drawbacks, scientists have tried to combine these two approaches (Luechtefeld et al., 2018). Recently, Banerjee and Roy (Banerjee and Roy, 2022) reported a novel quantitative RASAR algorithm for quantitative predictions for the 1st time where they modeled androgen receptor binding affinity with the physicochemical descriptors and the reliability parameters (RASAR descriptors) obtained from a quantitative read-across prediction tool (Banerjee et al., 2022). They reported a significant improvement in model quality due to the improvement of the feature space information content after introducing RASAR descriptors. RASAR descriptors are the read-across derived similarity, error, concordance, and prediction functions calculated from the important structural features of the training set. This method has shown promising quality, easy interpretability, and transferability. It is a relative prediction method where the response of query compound is predicted from the structural and biological data of source analogues; thus the RASAR descriptor generation is also relative and obtained after optimizing the read-across algorithm from the source compounds. In **Table 1.3**, different RASAR descriptors have been defined.

Table 1.3. Definition of the RASAR descriptors

Reliability parameters	Definition
Dispersion measures	
SD_Activity	Standard deviation of the observed response values of selected close source compounds (structural analogues) for a query compound
CV_Activity	Coefficient of variation of observed responses of selected source compounds (structural analogues) for a query compound
Similarity measures	
Average similarity	Average of calculated similarity values of the selected source compounds (structural analogues) against a specific query compound.
SD_Similarity	Standard deviation of the calculated similarity values of the selected source compounds (structural analogues) against a specific query compound.
MaxPos	Maximum similarity level of positive (based on source responses mean) source compound (structural analogues) against a specific query compound.

MaxNeg	Maximum similarity level of negative (based on source responses mean) source compound (structural analogues) against a specific query compound.
AbsDiff	Absolute difference between MaxPos and MaxNeg values
Concordance measures	
g	$g = 1 - 2 \times \lvert PosFrac - 0.5 \rvert$; (PosFrac is the fraction of positive source compounds amound alelected source compounds).
Banerjee-Roy coefficient (g_m)	$g_m = (-1)^n \times 2\lvert PosFrac - 0.5 \rvert$ Concordance measure that suggests the probability of the query compound to be positive or negative, based on the MaxPos, MaxNeg, and PosFrac values of the corresponding query compound. n is 1 when MaxNeg > MaxPos and vice versa. The value of g_m ranges from -1 to +1.
Predictive measure	
RA-Function	Read-across derived function which encodes the most important physicochemical and structural features responsible for generating a response

The RASAR algorithms have several advantages over QSAR and read-across such as-

i. Improve the external predictability from both QSAR and read-across.
ii. One can get the mathematical equation which further helps to study the impact of structural features.
iii. Can be used for small datasets also.

1.6.4. Machine-learning (ML) algorithms

Machine learning is an artificial intelligence-based approach that uses the knowledge of computer science and statistics with great proficiency and is applied for predictive and descriptive modeling in nearly all domains nowadays. Fundamentally, ML algorithms learn from the training data. Based on the learning algorithms, ML approaches can be classified into two broad categories namely (i) Unsupervised learning, and (ii) Supervised learning. A set of features (here descriptors in QSARs) is used to train the unsupervised learning algorithm and extract insights from them. Unsupervised learning is used in descriptive modeling approaches due to its ability to identify the important features of the data without considering the response or dependent variables (Y variables). On the other hand, supervised learning approaches consider both the features and the response for learning the algorithm and try to find the relationship between them; thus they are specifically used in predictive modeling. According to Kuhn and Johnson (Kuhn and Johnson, 2013), predictive modeling is "the process of developing a mathematical tool or model that generates an accurate prediction". For the regression and classification modeling of QSAR, supervised machine learning is mostly used. Various machine learning approaches are used for regression and classification modeling such as Artificial neural network (ANN) (Andrea and Kalayeh, 1991; So and Richards, 1992), Random forest (Amit and Geman, 1997; Andres and Hutter, 2006), Support vector machine (SVM) (Vapnik, 1998), etc.

The foundation of the artificial neural network is inspired by the natural neuron. ANN connects the computationally built artificial neurons in layers to process the information (Andrea and Kalayeh, 1991; So and Richards, 1992). Each artificial neuron performs four operations namely- input and output functions, summation function, activation function, and transfer function. The input function receives the output signal from the previous layer of neurons, determine the strength of the otput signal and accordingly passes the output signal to the next layers. The summation function computes the total signal received in a particular neuron from the previous layer. The activation function allows the output signal to vary with respect to time. Finally, the transfer function prepares a suitable output value and transfers it to the next layer as the output signal. During learning or training of the algorithm, one can modify the weight of the artificial neuron, and it modifies the output accordingly.

Random forest is an ensemble learning-based approach that uses consensus classification to reduce overfitting and improve accuracy (Amit and Geman, 1997). This method grows numerous decision trees (Andres and Hutter, 2006) that make the ultimate predictions based on the majority voting from the single tree's predictions. A randomly selected reduced number of data is taken for developing a decision tree and the randomly selected descriptors are chosen without data pruning. The user can optimize the number of trees and the number of descriptors used in each tree. The number of descriptors taken should be the square root of the total number of descriptors in the original data set. Random forest is an easy-to-use approach among others and can handle a large number of training data and descriptors. Random forests can also be used for quantitative predictions.

Support vector machine (SVM) algorithms are the supervised learning models that are used for both regression and classification analysis (Vapnik, 1998). SVM analyses the data and recognizes the pattern during learning. The foundation of SVM lies in the construction of hyperplanes in multidimensional descriptor space that separates the cases into several class levels. SVM can handle the data with high dimensionality even when the number of dimensions is more than the number of samples. Among other advantages, SVM can avoid the difficulties of using the linear function in high-dimensional feature space. The disadvantage of SVM is its inability to provide direct probability estimates; rather it uses the computationally exhaustive 5-fold cross-validation for this.

1.7. Application of in silico modeling approaches for mixture toxicity assessment

The scientific communities around the globe are continuously working on the mixture risk assessment. Various new ideas and techniques are utilized to develop a more reliable and precise prediction model and to construct reliable guidelines for mixture risk assessment. Popular in silico techniques, QSAR, and machine learning approaches, are successfully used in various computational studies of mixture toxicity prediction. Some representative case studies are mentioned below.

Qin et al. developed a QSAR model for the prediction of the toxicity (median effective concentration- EC_{50}) of binary and multi-component mixtures composed of two antibiotics (tetracycline hydrochloride and chloramphenicol) and four pesticides (metribuzine, trichlorfon, dichlorvos, and linuron) against *Aliivibrio fischeri* (Qin et al., 2018). The authors considered the additive and non-additive effects of mixtures during model development and calculated the mixture descriptors by different mixing rules. They used 25 binary, 15 ternary, and 5 quaternary mixtures data to construct an effective model for multi-component mixtures. The study described the conventional toxicity prediction by concentration addition and independent action and a comparison of results among additive toxicity models with QSAR. Based on the

CA and IA predictions, 18 mixtures presented additive effects whereas 27 mixtures showed non-additive effects. The authors developed a genetic algorithm-based MLR (multiple linear regression) model with three descriptors. An agreement between predicted and observed toxicity data can be expressed with a high coefficient of determination value (R^2 = 0.9366). Out of 45 mixture predictions, 12 mixtures are out of the lower and upper range of 95% confidence interval which is a better result than CA and IA models (27 mixtures were out of range). The authors have measured root mean square error (RMSE) to evaluate the deviation between observed and predicted results; the value for the QSAR model based on the whole data set (0.1663) was less than both CA (0.2045) and IA predictions (0.2378). Therefore the authors claimed better predictive capability of the QSAR models from CA and IA for prediction of non-additive mixture's toxicity in the case of binary and multi-component mixtures as well.

Mwense et. al. proposed a new methodology called integrated fuzzy concentration addition-independent action model (INFCIM) which utilizes molecular descriptors and a fuzzy set theory to characterize the similarity and dissimilarity of mixture components (Mwense et al., 2004). Based on that, the model also integrates the CA and IA models for better prediction of mixture toxicity. The authors tested their model in two cases. In the 1st case, they took the toxicity data of two types of mixtures (Mixture 1 and Mixture 2) containing 18 s-triazine and 16 acting chemical species, respectively tested on a freshwater algae *Scenedemus vacuolatus*. The inhibition of reproductive ability was tested with both mixtures. In the 2nd case, the toxicity data contains mixture 3 (consisting of 10 quinolones) and mixture 4 (consisting of 16 phenol derivatives) causing long-term inhibition of bioluminescence of the marine bacterium *Vibrio fischeri*. In the proposed INFCIM approach, molecular descriptors of individual components are calculated for the assessment of similarity among the components. The fewer prediction errors between the experimental data and INFCIM for all the datasets (less than 10% for mixture 1, less than 16% for mixture 2, and 11% for mixture 3 and 4) justify the good predictive ability of the method. The authors claimed that the INFCIM approach can be used for the toxicity prediction of the same mixture with different compositions or different mixtures as well until the endpoint remains the same.

Cipullo et. al. employed two machine learning models, i.e., Artificial Neural Network (ANN) and Random Forest (RF) to predict temporal bioavailability which was further used to predict toxicity of complex chemical mixtures (Cipullo et al., 2019). The authors developed a two-step process; temporal bioavailability of heavy metal/metalloids and petroleum hydrocarbon were predicted in the 1st step, and further in the 2nd step, the soil toxicity was predicted from the bioavailability data. Six months of mesocosms experimental data were used for the study. Parameters like amendment (biochar, compost), soil type, incubation time, and initial concentration of specific contaminants were employed as the input data for the modeling. The ANN was used to predict the bioavailability whereas the RF was employed for the identification of important input variables for temporal bioavailability and toxicity prediction. A series of biological indicators including soil respiration, bacteria count, seed germination, earthworms' lethality, microbial community fingerprint, and bioluminescent bacteria were evaluated for the environmental risk assessment. The ML model showed the importance of multiple features (combined effect) for the change in toxicity, and this information may not be accounted for in classical linear regression models. It has been found that the prediction of earthworm acute toxicity index was mainly affected by phenanthrene, anthracene, acenaptene, fluorine, pyrene, and the aliphatic fraction 17 to 19 of petroleum hydrocarbons. The aliphatic hydrocarbons with medium chain length were identified as important indicators for the acute toxicity of soil organisms like earthworms. The authors claimed that the engagement of an ML

model could improve our understanding of the rate-determining processes affecting the bioavailable fraction of soil contaminants. Therefore, the potential risk assessment of environmental mixtures becomes easier and the appropriate safety measures can be adopted.

Zeng et. al. have applied QSAR for the risk assessment of heavy metal-containing mixtures (Zeng et al., 2019). The authors have evaluated the toxicity of persistent heavy metals such as Cu^{2+}, Co^{2+}, Fe^{3+}, Zn^{2+}, and Cr^{3+} by experimentation. They also determined the toxicity of corresponding binary mixtures against *Photobacterium phosphoreum*. The authors clearly pointed out two distinct observations. Firstly, they showed the toxicity order of individual heavy metals against marine bacterium (*P. phosphoreum*): Zn^{2+}> Cu^{2+}> Co^{2+}> Cr^{3+}> Fe^{3+}. Zinc (II) ion was the most toxic one (-log EC_{50} = 4.75) among the five whereas, the ferric (III) ion showed the least toxicity (-log EC_{50} = 3.64). Secondly, various interactions (synergism and antagonism) have been observed in addition to the additive effects of the binary mixtures of the mentioned metals. The authors developed a partial least squares (PLS) regression model for the prediction of mixture toxicity of heavy metals by employing the ion characteristic parameters as dependent variables. Acceptable R^2 and Q^2 values (R2 = 0.75; Q^2 = 0.649) prove a good correlation between the toxicity endpoint (-log EC_{50}) and the ion characteristic parameters. From the developed model, it has been found that the toxicity of mixtures of heavy metals is affected by the $(\Delta IP)^{mix}$ (change in ionization potential), $\log(KOH)^{mix}$ (1st hydrolysis constant), and $\log K_f^{mix}$ (formation constant value). Finally, the authors claimed the ion characteristics-based QSAR as an efficient approach for toxicity prediction and proposed it as a supplementary tool for mixture risk assessment of heavy metals.

Halder, A.K. et. al. employed machine learning-based QSTR modeling for the risk assessment of deep eutectic solvents (DES) against *Aliivibrio fischeri* (Halder and Cordeiro, 2019). The authors have employed the toxicity data of 72 DESs and their constituents to find out the structural features associated with DES-mediated toxicity. Mixture descriptors were obtained by the 'sum and absolute difference of weighted (by molar fraction) descriptors' strategy. According to this strategy, the composition of mixture components along with their molar fractions affects the properties of mixtures. The authors have employed 'points out', 'mixtures out', 'compounds out', and 'everything out' strategies for the dataset division and model validation. The non-linear QSTR models were developed with the help of three machine learning techniques called: 'support vector machine (SVM)', 'random forest' (RF), and 'multilayer-perception artificial neural network (MLP-ANN)'. The authors also developed several linear QSTR models for the identification of the important structural features of the associated toxicity of DESs. The authors suggested that the 'number of nitrogen atoms', 'topological distance between oxygen atoms', 'van der Waals surface area related topological descriptors', 'dipole moment', 'molar refractivity', and 'molecular mass' components of DESs may affect the toxicity of DESs.

CHAPTER - 2

PRESENT WORK

Chapter 2: Present Work

The continued population growth, rapid urbanization, and simultaneous chemical advancement greatly increase chemical pollution worldwide (Carley and Christie, 2019). The chemical industries are flourishing in different parts of the world (mostly in developing countries) to fulfill the growing demand of mankind (Landrigan et al., 2018), and these socio-economical and industrial growths are creating a serious threat to environmental safety and ground-water prevention (Daughton, 2003). The chemicals are mostly discharged through different sewage systems such as hospitals, cattle fields, aquacultures, etc. (Álvarez-Muñoz et al., 2015; Xiong et al., 2015); and thus, a majority of chemical pollutants i.e., organic chemicals, plant protection products, pharmaceuticals, industrial, and household chemicals have been found in soil and water resources in ng/L to mg/L ranges (Lamastra et al., 2016; Luo et al., 2014). The bioactive chemicals i.e., pharmaceuticals, pesticides, insecticides, acaricides etc. are of major concern due to their bioactive nature. Antibiotics inevitably impact the lower environmental organisms (mostly aquatic and soil species) (Mendoza et al., 2015). They mostly target the microorganism, especially the bacteria in the environment. Moreover, the bacterial species come in contact with multiple chemicals simultaneously, which may further potentiate the toxic effects (Jiang et al., 2011; Li et al., 2015). Different other species i.e., aquatic species (fish, algae, etc.), soil species (earthworm, termites, fungi, etc.), insets (honey bees etc.), etc. are also affected. Besides the environmental damage, the presence of such mixtures of toxic chemicals can cause serious damage to human beings due to bioaccumulation and biomagnification (Bilal et al., 2019b, 2019a; Rasheed et al., 2018; Singh et al., 2018).

The pollutants are never found alone in the environment. Multicomponent chemical mixtures are formed continuously with variable combinations, and the number of such combinations is increasing. In some previous research articles, deviations of toxicity from the standard were reported due to significant interactions (additive response, synergism, antagonism) among chemicals (Kar and Leszczynski, 2020; Zou et al., 2012). Besides this, the degradation products of chemicals also play an essential role in mixture toxicity as reported by Wang and co-workers (Wang et al., 2020). Therefore, a mixture toxicity assessment is a significant and necessary task which should be carried out for a proper risk assessment of chemicals against different ecological species. In the past decades, contamination of the aquatic environment with various pharmaceuticals and pesticides has grabbed the attention of global scientific communities (Gao et al., 2012; Kemper, 2008; Topp et al., 2016) and it is becoming an important topic of discussion after finding their simultaneous presence in water bodies (Elliott et al., 2018; Kolpin et al., 2002). Scientists are now working on the risk assessment of chemical mixtures; however, mixture risk assessment by physical experimentation is not a good idea due to the unending possibilities of combinations. Besides this, the physical experiments are associated with numerous complications including ethical issues. Animal models for toxicity testing are time-consuming and costly as well as unethical. In 1959, Russell and Burch introduced the “3R” concept (Replacement, Reduction, and Refinement), to reduce the use of animals in scientific procedures (Russell and Burch, 1959).

There are two largely accepted and widely used mixture toxicity models, namely concentration addition (CA) (Loewe and Muischnek, 1926) and independent action (IA) (Bliss, 1939) which assume the additive nature of mixture components, but they never consider the probable interactions (especially for the biologically active chemicals like pharmaceuticals and pesticides) among the components (Kim et al., 2013). Apart from this, the CA and IA models need preliminary experimentation, which limits the application of

these modeling to some extent. To get rid of these above-mentioned issues, scientists and regulatory agencies are working together; however, a universally acceptable toxicity protocol for mixture toxicity study is still rare (Iwasaki and Gauthier, 2016). Therefore, an efficient method for the identification and estimation of interactive effects (synergism and/or antagonism) along with the additive effect of mixture components is necessary for better risk assessment of chemicals in a real-world scenario.

The concept of similarity stands upon a relative analysis of different points with respect to others. In the context of ecotoxicology and drug design, it is necessary to have an objective measure of similarity to compare various chemicals concerning their toxicity or activity (Gallegos Saliner A., 2006). Machine learning algorithms are now gaining sufficient attention among scientific communities for predictive modeling of toxicity and biological activity based on chemical similarity (Nazemzadeh et al., 2021). The hybrid modeling approaches can be very useful in this regard due to their ability to combine the detailed process understanding with machine learning; thus they can provide an opportunity to integrate prior process knowledge with various measurement data for efficient modeling of toxicity of hazardous mixtures (Nazemzadeh et al., 2021). Quantitative structure-activity relationships (QSARs), read-across predictions, q-RASAR, and ML algorithms are the most widely used similarity-based computational approaches which can efficiently predict the toxicity of individual chemicals by relative similarity analysis with respect to the chemical structures and properties (Hao et al., 2020; Li et al., 2022); hence the use of these *in silico* methods is widely accepted and encouraged by the concerning authorities like the Organization of Economic Co-operation and Development (OECD), European Chemical Agency (ECHA), United States Environmental Protection Agency (US-EPA), and regulations like Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) etc. (Kar et al., 2020; Klatte et al., 2017; Mansour et al., 2016; Raies and Bajic, 2016). Therefore these methods can be a good alternative for mixture toxicity assessment, which can give predictions based on various structural features of the mixture components. Various research groups have applied QSAR for the mixture toxicity assessment in the recent past and successfully proposed numerous mathematical models for different environmental endpoints (Carnesecchi et al., 2020; Chatterjee and Roy, 2021; Hoover et al., 2019; Qin et al., 2021). QSAR can also help to identify the interactions among structural features which may be essential for the toxicological effect (biological properties in general). It should be noted here that the QSAR modeling of mixtures is not as simple as the individual toxicants; thus, various additional steps should be followed for successful modeling. Among the key steps, mixture descriptor calculation, and validation of the developed model are the most important. There are very few approaches in the literature, where the cross-validation strategies have been properly used for the mixture toxicity predictions, and hence the chance of overestimation of quality metrics is persisting. Therefore, to study the mixture toxicity of chemical mixtures by in silico models rigorously, more attempts should be made by following the appropriate procedures and without compromising the guidelines of QSAR model development for single chemicals given by OECD.

2.1. Objectives of the dissertations

The objectives of the dissertation were the (i) collection of experimental toxicity data of various chemical mixtures against different environmental endpoints, (ii) the development of QSAR models for predicting toxicity of different classes of chemical mixtures (pharmaceuticals, industrial chemicals, plant protection products, etc.), (iii) rigorous validation of the developed models in accordance with the OECD guidelines, (iv)

identification of various structural features, responsible for the toxicity of chemical mixtures, (iv) make use of different other machine learning-based new approach methodologies such as read-across and q-RASAR for enhancing the performance of predictive model further.

2.2. Datasets employed for the development of Mixture-QSAR/q-RASAR/ML models

In this dissertation, six different in silico prediction modeling studies have been carried out and presented. Aquatic mixture toxicity data against *Photobacterium phosphoreum, Selenastrum capricornutum, and Aliivibrio fischeri* have been collected and employed for modeling study. Mixture toxicity data against soil and intestinal bacterium such as *Bacillus subtilis* and *Escherichia coli* have also been collected. Besides the bacterium and algae, acute contact mixture toxicity data of honey bees (*Apis millifera*) have also been collected for mixture toxicity modeling. In the following **Table 2.1**, all the collected mixture datasets along with their respective sources have been mentioned.

Table 2.1. Collected mixture datasets employed in this dissertation

Dataset	Endpoint	No. of mixtures	Class of mixtures	Sources
I-A	pEC50 (Mol/L) to *Photobacterium phosphoreum*	79 mixtures	Binary mixtures obtained from different organic chemicals [aldehydes (AHs), cyanogenics (CGs), sulfonamides (SAs), and Trimethoprim (TMP)]	Literature sources(Chung-Yuan and Ch'i-Feng, 1996; Jiang et al., 2010; Lin et al., 2003; Tian et al., 2012; Zou et al., 2012)
I-B	pEC50 (Mol/L) to *Selenastrum capricornutum*	20 mixtures	Binary mixtures obtained from different organic chemicals [triazines (TAs) and urea herbicides (UEs)]	Literature sources (Fairchild et al., 1997; Gaggi et al., 1995; Ma et al., 2006)
II	Log 1/EC50 (Mol/L) to *Photobacterium phosphoreum*	84 mixtures	Binary mixtures obtained from non-polar narcotic toxicants	Literature source (Lin et al., 2003)
III	pEC50 (Mol/L) to *Aliivibrio fischeri*	198 mixtures	178 binary mixtures, 15 ternary mixtures, and 5 quaternary mixtures	Literature sources (Bures et al., 2021; Qin et al., 2018)
IV	Acute contact toxicity data: pLD50 values (in μg active substance/bee unit) to honey bees	123 mixtures	Binary and multicomponent mixtures of organic plant protection products	Literature source (Carnesecchi et al., 2020)

V-A	pEC50 (Mol/L) to *Vibrio fischeri*	76 mixtures	51 binary and 25 ternary mixtures obtained from sulfonamides, sulfonamide potentiators, and tetracycline antibiotics	Liturature source (D. Wang et al., 2018)
V-B	pEC50 (Mol/L) to *Escherichia coli*	67 mixtures	43 binary mixtures and 24 ternary mixtures obtained from sulfonamides, sulfonamide potentiators, and tetracycline antibiotics	Liturature source (D. Wang et al., 2018)
V-C	pEC50 (Mol/L) to *Bacillus subtilis*	43 mixtures	29 binary mixtures and 14 ternary mixtures obtained from sulfonamides, sulfonamide potentiators, and tetracycline antibiotics	Liturature source (D. Wang et al., 2018)
VI	Log (1/IC50) in molar scale to *Escherichia coli*	78 mixtures	Binary mixtures obtained from fluoroquinolones antibiotics	Literature source (Wang et al., 2020)

2.1.1. Dataset I (Study 1)

In the current study, various QSAR models were developed for the future prediction of aquatic toxicity and risk assessment of various pharmaceuticals, organic chemicals and their corresponding binary mixtures. Two aquatic organisms, *Photobacterium phosphoreum* and *Selenastrum capricornutum* were chosen against which the toxicity data are collected. *P. phosphoreum* is a gram negative and bioluminescent bacterium living in symbiosis with various marine organisms and a classic marker of the presence of marine toxicants (De Zwart and Slooff, 1983; Ribo and Kaiser, 1987), whereas *S. capricornutum* is a type of microalgae of freshwater environment and is commonly used as a sensitive biomarker of the presence freshwater toxicants (Hübner et al., 2003; Ward et al., 2002). Negative logarithms of median effective concentration (pEC50) of the chemicals species and mixtures were used as the response variable (Y-variable in the data matrix) of the data sets. The weighted descriptors generation approach of binary mixtures has been used to calculate the mixture descriptors in this work. Here, the objectives of the current work are- to identify the maximum possible structural features of the collected data sets of individual chemicals and mixtures responsible for the toxicity, to propose general QSAR models based on the structural features of the individual compounds and the corresponding mixtures, to show the significant structural features in the individual compounds as well as in the mixture components and their effects towards modeled responses (pEC50).

2.1.2. Dataset II (Study 2)

In this current study, we have developed a properly validated mixture QSAR model for the prediction of aquatic toxicity from the dataset of 84 binary mixtures of polar and non-polar

narcotic toxicants (Lin et al., 2003). We have taken the negative logarithm of median effective concentration (*Log 1/EC50*) of binary mixtures against marine bacteria *Photobacterium phosphoreum* as the endpoint (Y-variable) of this study. *Photobacterium phosphoreum* is a type of gram-negative bioluminescence bacteria that lives in symbiosis with marine organisms and is a classical marker of marine toxicants (De Zwart and Slooff, 1983; Ribo and Kaiser, 1987). The objectives of the current research are- to identify the important structural features of the mixtures of polar and non-polar narcotic toxicants responsible for aquatic toxicity, to develop and propose a new mixture QSAR model following the OECD protocols for the prediction of aquatic toxicity, to cross-validate the model for more rigorous validation and for avoiding the development bias of mixture QSAR (Muratov et al., 2014; Oprisiu et al., 2012), to interpret the modeled descriptors mechanistically for better understanding, to improve the quality of predictions of binary mixtures using a machine learning-based chemical read-across algorithm.

2.1.3. Dataset III (Study 3)

The objective of this study was to apply the in-silico approaches for the prediction of aquatic toxicity of binary and multicomponent mixtures effectively. Keeping the objective in mind, we developed a QSAR model for the prediction of aquatic toxicity of binary and multi-component chemical mixtures in this communication. To build the model, a toxicity dataset of 198 hazardous mixtures against *Aliivibrio fischeri* was taken. The mixtures were composed of 12 pharmaceuticals and 8 pesticides which are abundantly used either for the treatment of various diseases, or the agricultural purposes and hence can be present in wastewater streams and can flow through the river and freshwater resources (Gao et al., 2012; Kemper, 2008; Topp et al., 2016). 2D structural descriptors have been used in this study for the ease of calculation and interpretation (Chatterjee and Roy, 2021; Mukherjee et al., 2022; Sun et al., 2021). Mixture descriptor calculation is one of the tricky steps of mixture-QSAR. Qin *et al.* reported various mixing rules for the computation of mixture descriptors in 2018 among which the theory of additivity was an important one (Qin et al., 2018). In 2019, Hoover *et al.*, and Kar *et al.* have separately mentioned about the theory of additivity as one of the promising approaches for mixture descriptor calculation (Hoover et al., 2019; Kar and Leszczynski, 2019). Thus, we have also applied the theory of additivity for obtaining the mixture descriptors in this research. Partial least squares (PLS) regression has been applied to develop the model, which has been cross-validated using mixtures out, compounds out, and everything out cross-validation strategies afterwards (Muratov et al., 2014; Oprisiu et al., 2012). The selected descriptors have been thoroughly studied in the light of this model. We have also reported the applicability domain (Gadaleta et al., 2016) of the developed model which is essential to claim the reliability of prediction with the new model. Besides all these, a machine learning-based chemical read-across prediction tool has been used to justify the relevance of the selected structural features in this study, and it is the 1st instance where a chemical read-across algorithm has been applied for mixture toxicity prediction. The complete study has been performed without any animal experimentation and in the environmental friendly condition to comply with the 3R concept (Replacement, Reduction, and Refinement) of REACH legislation (Ritskes-Hoitinga and Luijk, 2019).

2.1.4. Dataset IV (Study 4)

We have developed a quantitative read-across structure-activity relationship model (q-RASAR) for the prediction of binary mixture toxicity in honey bees in this research. RASAR is the combination of read-across and quantitative structure-activity relationship (QSAR), which is developed for the improvement of predictability from both read-across and QSAR

approaches (Luechtefeld et al., 2018). It incorporates the similarity and error-based parameters obtained from the read-across prediction tool with the pool of structural and physicochemical features of the dataset and then finalizes the combination of descriptors (Banerjee and Roy, 2022). Ridge regression, linear and non-linear support vector machine algorithms are also used in this study to enhance the predictive ability of the q-RASAR model further. Apart from the modeling, the best mixing rule to calculate the mixture descriptors has been identified in this research. The calculated validation metrics indicate the goodness-of-fit, robustness, and good predictability of the model. Here, the toxicity values of several new untested binary mixtures (obtained from the real-world data recently published by Tosi et al., 2022) have also been predicted with the new model, and the reliability of the predictions has been checked by the prediction reliability indicator (PRI) tool. This computational research is in consonance with the 3R strategy (Replacement, Reduction, and Refinement of animal experimentation) of the REACH legislation (Ritskes-Hoitinga and Luijk, 2019).

2.1.5. Dataset V (Study 5)

In the current study, we have developed three QSAR models from three different mixture datasets against three common bacterial endpoints i.e., *Vibrio fischeri*, *Escherichia coli*, and *Bacillus subtilis*. The data comprised binary and ternary mixtures of commonly used antibiotics such as sulfonamides (SAs), sulfonamide potentiators (SAPs), and tetracyclines (TCs). Simple 0D-2D descriptors have been employed in this study to model the mixture toxicity for avoiding more complex conformational analysis and energy minimization (important prerequisites for 3D and higher-order descriptor calculations) without compromising the predictability of developed models. For mixture descriptor calculation, three different mixing rules (linear combination of molecular contributions, the squared molecular contributions, and norm of molecular contributions) (Chatterjee and Roy, 2022a) have been used and the best one has been identified. We have also developed three interspecies QSAAR models using the predicted responses as X-variable with additional descriptors. The quantitative Read-across Structure-Activity Relationship (q-RASAR) (Banerjee and Roy, 2022) and partial least squares (PLS) (Wold et al., 2001) algorithms have been employed for developing the QSAARs. The RASAR is a thoughtful merging of read-across (RA) predictions and QSAR for using the non-linearly derived prediction functions obtained by the read-across algorithm in a linear regression framework of QSAR (Banerjee and Roy, 2022). There are several instances in the literature where RASAR has been applied and proved to be effective for obtaining better external predictability than both QSAR and read-across predictions (Banerjee and Roy, 2022; Luechtefeld et al., 2018). In mixture toxicity predictions, RASAR has not yet been used much; hence, we have incorporated the RASAR algorithm for developing more efficient interspecies mixture-QSAARs in this work, and the final models may thus more appropriately be termed as quantitative Read-across Structure-Activity-Acivity Relationships (q-RASAARs). In these models, the predicted toxicity to other bacterial species are used as descriptor(s) after an appropriate “data fusion”. The important structural features, responsible for mixture toxicity against each organism have been identified and discussed here in this study. We have also emphasized the consideration of intercorrelated toxic endpoints for better toxicity predictions from the insights of developed QSAAR models. The toxicity data was obtained by simple assay where the concentration of sample broth is the determinant of lethality. Thus, the exact toxicity mechanism is uncertain and thus, the use of specific receptor–based descriptors is inappropriate. Apart from that, the targets for the toxicity may vary depending on the chemical classes and the organism being considered; therefore, the use of a fixed target

receptor or protein for docking-based descriptor calculation potentially limits the efficiency of predictions. To avoid these issues, 0D-2D descriptors have been used in this study. The predicted toxicity values of other bacteria have also been used as descriptors for improving the predictive performance further. To the best of our knowledge, this is the first attempt where interspecies mixture toxicity models have been developed for the prediction of cellular toxicity of binary and ternary mixtures of SAs, SAPs, and TCs towards any of the three above-mentioned organisms. In a nutshell, this study helps the readers to understand the basics of mixture toxicity predictions, the use of predicted toxicity of certain species for improving the predictability, and the importance of q-RASAAR for developing statistically sound models for mixture toxicity predictions.

2.1.6. Dataset VI (Study 6)

In the present work, we have developed well-validated mixture-QSAR models from the existing toxicity data of fluoroquinolone mixtures in *Escherichia coli.*(Wang et al., 2020) *E. coli* is one of the resident bacteria present in the lower intestine of warm-blooded animals including human beings and it has several beneficial effects.(Martinson et al., 2019; Martinson and Walk, 2020) Therefore, it is a good marker to assess the detrimental effects of fluoroquinolone mixtures towards gut microbiota. Different machine-learning algorithms have been employed with the conventional partial least squares (PLS) regression to optimize the modeling method. We have also optimized the mixing rule for mixture descriptor calculation in this study. To avoid exhaustive calculations, simple, interpretable, and reproducible 0D-2D descriptors have been used. The models have been validated employing the strict validation protocol suggested by the Organization for Economic Co-operation and Development (OECD).(OECD, 2004) The probable degradation products of the fluoroquinolones have been identified from the literature. Based on that, the hypothetical binary mixtures have been generated out of the degradation products, and their parent compounds and their combined toxicities have been predicted using the newly developed QSAR model. Mostly the freely available software tools have been used to model the toxicity which makes this work cost-effective.

CHAPTER - 3

MATERIALS AND METHODS

Chapter 3: Materials and Methods

3.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors

3.1.1. Toxicity data

3.1.1.1. Series 1 Compounds and Type 1 Mixture Data sets

The toxicity endpoint values of Series 1 compounds [13 aldehydes (AHs), 11 cyanogenics (CGs), 10 sulfonamides (SAs), and Trimethoprim (TMP)], measured through the bioluminescence assay of a bioluminescent marine bacterium *P. phosphoreum*, were collected from the literature (Chung-Yuan and Ch'i-Feng, 1996; Jiang et al., 2010; Lin et al., 2003; Tian et al., 2012; Zou et al., 2012). The bioluminescence assay was performed using the diluted bacteria (maintained at agar slants at 4ºC) that had been cultured at 20ºC in yeast-tryptone-salts-glycerol broth for 12-14 hours. Toxicity was measured with the Microtox test instrument (Model toxicity analyzer DXY-2) by quantifying the decrease in light emission from the bacteria after exposure to 3% NaCl solution of the test chemical for 15 min. Based on the decrease in light emission, EC_{50} (median effective concentration) was calculated by using the Probit model (Finney, 1971) and reported as negative logarithm of EC_{50} (pEC50). Toxicity data of series 1 compounds are enlisted in **Table 3.1**. The toxicity test of type 1 mixtures (obtained from the Series 1 compounds), performed in same manner as the individual chemical, is discussed in the equation 3.1 as follows (Preston et al., 2000)-

$$EC_{50M} = \frac{C_M}{\frac{C_A}{EC_{50A}} + \frac{C_B}{EC_{50B}}} \qquad (3.1)$$

where, EC_{50M} is the median effective concentration of sample mixture; C_A, C_B, and C_M are the concentrations of mixture component A, mixture component B, and the whole mixture respectively; EC_{50A} and EC_{50B} are the median effective concentration of component A and B respectively. Toxicity data of type 1 mixtures were obtained from the literature by (Yao et al., 2013) and presented in **Table 3.2**.

Table 3.1. Toxicity data of Series 1 compounds

Sl. No.	Individual Compounds	Observed Toxicity [pEC50 (mol/L)]
1	Acetaldehyde	2.36
2	Propionaldehyde	2.72
3	Butyraldehyde*	3.25
4	Valeraldehyde	3.27
5	Benzaldehyde	3.43
6	*p*-Nitrobenzaldehyde*	4.28
7	*p*-Terephthaldehyde	4.31
8	*p*-Chlorobenzaldehyde	4.25
9	*p*-Bromobenzaldehyde	4.30
10	*p*-Hydrobenzaldehyde	4.54
11	*p*-Methylbenzaldehyde	3.82

12	*p*-Methoxybenzaldehyde*	4.03
13	*p*-Dimethylaminobenzaldehyde	5.40
14	Malononitrile*	2.55
15	Glycolonitrile	2.98
16	α-Hydroxyisobutyronitrile	3.61
17	Allyl cyanide*	1.45
18	Benzonitrile	3.48
19	Benzyl cyanide	4.23
20	Acetonitrile	0.75
21	Acrylonitrile	1.51
22	Succinonitrile	0.36
23	Phthalonitrile*	3.51
24	Lactonitrile	2.01
25	Sulfamethazine	4.08
26	Sulfapyridine	3.84
27	Sulfamethoxazole	4.45
28	Sulfadiazine	4.50
29	Sulfisoxazole	4.43
30	Sulfamonomethoxine	5.05
31	Sulfachloropyridazine	4.78
32	Sulfachinoxalin*	4.53
33	Sulfamethoxydiazine*	4.41
34	Sulfamethoxypyridazine	4.36
35	Trimethoprim	3.22

* : Test Set compounds

Table 3.2. Chemical components, mass ratio, and observed toxicity data of Type-1 mixtures

Mixture No.	Components in the mixture	The ratio of toxic unit	Observed pEC50values
1*	1\14	1\1	2.44
2	2\14	1\1	2.63
3	3\14	1\1	2.77
4	4\14	1\1	2.78
5	5\14	1\1	2.80
6*	6\14	1\1	2.84
7	7\14	1\1	2.84
8	8\14	1\1	2.83
9*	9\14	1\1	2.84
10	10\14	1\1	2.85

11	11\14	1\1	2.83
12	12\14	1\1	2.84
13*	13\14	1\1	2.85
14	5\15	1\1	3.15
15	6\15	1\1	3.26
16*	7\15	1\1	3.25
17	8\15	1\1	3.24
18	9\15	1\1	3.26
19	10\15	1\1	3.27
20*	11\15	1\1	3.22
21	13\15	1\1	3.28
22	1\16	1\1	2.64
23*	2\16	1\1	2.97
24	3\16	1\1	3.39
25*	5\16	1\1	3.51
26	6\16	1\1	3.83
27	7\16	1\1	3.78
28	8\16	1\1	3.75
29*	9\16	1\1	3.83
30	10\16	1\1	3.86
31	11\16	1\1	3.70
32	12\16	1\1	3.77
33*	13\16	1\1	3.90
34	1\17	1\1	2.18
35	3\17	1\1	2.33
36	4\17	1\1	2.34
37	5\17	1\1	2.34
38	6\17	1\1	2.36
39*	7\17	1\1	2.36
40	8\17	1\1	2.36
41	10\17	1\1	2.36
42*	11\17	1\1	2.35
43	12\17	1\1	2.35
44*	13\17	1\1	2.36
45	5\18	1\1	3.45
46*	6\18	1\1	3.72
47	7\18	1\1	3.68
48	8\18	1\1	3.66
49	10\18	1\1	3.74
50	11\18	1\1	3.62

51*	12\18	1\1	3.67
52	13\18	1\1	3.78
53	5\19	1\1	3.67
54	6\19	1\1	4.25
55	7\19	1\1	4.14
56	8\19	1\1	4.08
57	9\19	1\1	4.26
58	10\19	1\1	4.36
59	13\19	1\1	4.50
60*	25\35	1\1	5.08
61	26\35	1\1	4.85
62	27\35	1\1	5.50
63	28\35	1\1	5.42
64	29\35	1\1	5.45
65*	30\35	1\1	6.01
66	31\35	1\1	5.73
67	27\35	13396\1	3.49
68	27\35	8587\1	3.49
69	27\35	2747\1	3.49
70	27\35	858\1	3.51
71	27\35	274\1	3.55
72	27\35	85\1	3.67
73	27\35	27\1	3.92
74*	27\35	15\1	4.08
75	27\35	4\1	4.52
76	27\35	1\6	5.34
77	27\35	1\21	5.43
78	27\35	1\37	5.45
79*	27\35	1\116	5.46

*Test set Mixtures; [The toxicity values of mixtures are against *Photobacterium phosphoreum*]

3.1.1.2. Series 2 Compounds and Type 2 Mixture Data sets

The toxicity endpoint values of Series 2 compounds [10 triazines (TAs) and 10 urea herbicides (UEs)), measured through the toxicity assay of the freshwater microalgae *S. capricornutum*, were collected from the literature (Fairchild et al., 1997; Gaggi et al., 1995; Ma et al., 2006). The toxicity assay was performed using the diluted bacteria (maintained at agar slants at 4ºC) that had been photo-autotrophically incubated in HB-4 medium (Li et al., 1959) in a rotator shaker (100 rpm) at 25ºC with illumination by cool white fluorescent light (light intensity = 450 E $m^{-1}s^{-1}$) in continuous mode (Ma et al., 2003; Verdisson et al., 2001). For toxicity test, 15-mL of aliquots of HB-4 medium containing 5×10^4 mL^{-1} algal cells was inoculated in 50-mL Erlenmeyer flask. Different concentrations of the test samples were

added in different flasks and incubated for 96 hours at the previously mentioned condition. Finally, the optical density of the culture was measured with the spectrophotometer at 680 nm and compared with control sample (without any test sample). The growth inhibition was estimated from this comparative study and reported. Toxicity data of series 1 compounds are enlisted in **Table 3.3**. The toxicity data of type 2 mixtures (obtained from the Series 2 compounds), estimated in the same manner as the type 1 mixtures, was obtained from the literature by (Yao et al., 2013) and presented in **Table 3.4**.

Table 3.3. Toxicity data of Series 2 Compounds

Sl. No.	Individual Compounds	Observed Toxicity [pEC50 (mol/L)]
36	Atrazine	6.68
37	Prometryn	8.07
38	Simetryn	6.29
39	Prometone	8.99
40	Simazine	5.43
41	Metribuzin	5.7
42	Cyanazine*	6.61
43	Terbutryn*	7.95
44	Terbutylazine	6.93
45	Ametryn	6.56
46	Diuron	7.72
47	Chlorotoluron	8.4
48	Monolinuron	7.33
49	Monuron	6.3
50	Methabenzthiazuron	7.02
51	Isoproturon*	7.18
52	Isoproturon	7.33
53	Ethametsulfuron*	4.13
54	Chlorsulfuron	6.42
55	Metsulfuron*	6.29

*: Test Set compounds

Table 3.4. Chemical components, mass ratio, and observed toxicity data of Type-2 mixtures

Mixture No.	Components in the mixture	The ratio of toxic unit	Observed pEC50 values

80	36\46	1\1	7.78
81	36\47	1\1	6.97
82	36\48	1\1	6.89
83*	36\49	1\1	6.45
84	37\46	1\1	7.86
85*	37\47	1\1	8.20
86	37\48	1\1	7.56
87	37\49	1\1	6.59
88	38\46	1\1	6.58
89*	38\47	1\1	6.59
90*	38\48	1\1	6.55
91	38\49	1\1	6.29
92	39\46	1\1	8.00
93*	39\47	1\1	8.60
94	39\48	1\1	7.62
95	39\49	1\1	6.60
96	40\46	1\1	5.73
97	40\47	1\1	5.73
98	40\48	1\1	5.73
99	40\49	1\1	5.68

*: Test set Mixtures; [The toxicity values of mixtures are against *Selenastrum capricornutum*]

3.1.2. Molecular Descriptors

Molecular descriptors are the quantitative values of structural features of molecules which are computed and used to correlate the molecular structure with the biological response (activity or toxicity) or chemical properties in a QSAR study. Various classes of descriptors based on dimensionality (e.g., 0 Dimensional- 7 Dimensional descriptors) can be used for the QSAR analysis. 2D descriptors are easy to calculate, interpretable, can correlate sufficient structural features with the response and were extensively used in many published QSAR literature (Hoover et al., 2019; Seth et al., 2020). Hence, in this work, the 2D descriptors were used for the individual and mixture datasets. The 2D structures of 55 individual compounds (Series 1: 35 compounds; Series 2: 20 compounds) have been drawn using "MarvinSketch" (version 14.10.27) software (ChemAxon, 2013) with explicit hydrogen addition and proper aromatization for descriptor calculation. These structures have been used without energy minimization and conformational analysis, as we have used only 2D descriptors in our study. Total nine classes of 2D descriptors namely Constitutional indices, Ring descriptors, Functional group counts, Atom centered fragments, Atom type E-state indices, 2D atom pairs, Molecular properties, and Extended Topochemical Atom (ETA) indices have been

calculated. All the classes of descriptors are easy to calculate and easily interpretable. The ETA descriptors (Roy and Ghosh, 2010) have been calculated using the "PaDel-Descriptor" software tool (Yap, 2011) whereas the other eight classes of descriptors have been computed using "DRAGON 6" (2011) software (Mauri et al., 2006). Descriptor generation of mixtures is not as simple as the individual chemicals. One way to calculate the descriptors for mixtures is to estimate the values of variables from individual components, and then, the obtained descriptor values from individual components are multiplied by their concentration ratio present in the mixture and finally added together to get the final descriptor value for mixtures. This strategy is known as weighted descriptors generation approach and has been used in many published QSAR analysis of mixtures (Khan et al., 2018; Rasulev et al., 2017). In the present work, the descriptors of mixtures have been calculated as following-

Descriptors of molecule M: D_1^M, D_2^M, D_3^M, ……..,D_n^M

Descriptors of molecule N: D_1^N, D_2^N, D_3^N, ……..,D_n^N

If the molecule M and N are mixed in a toxic ratio X:Y then i^{th} descriptor values of the mixture MN will be calculated as per the following equation 3.2-

$$Di^{MN} = Di^M \times X + D_i^N \times Y \qquad (3.2)$$

where, Di^{MN} is the i^{th} descriptor of mixture MN, D_i^M is the i^{th} descriptor of component M, D_i^N is the i^{th} descriptor of component N, and X+Y = 1.

3.1.3. Dataset Division

For the QSAR model development and external validation, each of the 4 datasets was divided into training and test sets. The training set compounds were used for model building whereas the test set compounds (marked with *in the datasets) were used for the rigorous external validation of the developed QSAR models. Four division methods and two statistical tools have been employed to improve the precision of QSAR models and to reduce the selection bias of the division techniques. Each datasets were divided using 1) Sorted activity based division; 2) Euclidean distance based division; 3) Kennard-Stone selection; and 4) Modified k-Medoids clustering technique (Roy et al., 2015). The Sorted activity based division method is a Y-response based strategy (activity based) whereas the other three are X-response based (structural features) strategies. From the 4 division methods, total 13 training-test set pairs (1 seed from Kennard-Stone selection, 4 seeds from Activity sorted response based divisions, 4 seeds from Euclidian distance based divisions, 4 seeds from k-Medoids clustering techniques) were obtained in each dataset. The first three approaches were executed using **'Dataset Division GUI 1.2'** tool and the last one was performed using 'Modified k-Medoids GUI 1.3' tool available at http://teqip.jdvu.ac.in/QSAR_Tools/. Preliminary models were developed from each division by using stepwise multiple linear regressions (S-MLR) and examined through rigorous validation approaches. The sorted activity based division and Euclidian distance based division were chosen for the individual data sets and the mixture data sets, respectively based on the validation results from the 4 datasets and were taken forward for the further improvement of the quality of models.

3.1.4. Model Development

Initially we reduced the descriptor pool from the best division of each dataset by using Stepwise Multiple Linear Regression (S-MLR) method in 'MINITAB' software in multiple

iterations. After the size reduction of the descriptor matrix, two statistical tools i.e., Genetic Algorithm (GA) (Leardi, 2001), and Best Subset Selection (BSS) (Roy et al., 2015) were employed to select the best descriptor combinations. GA and BSS were carried out using 'DCVToolv2.0' and 'BestSubsetSelection_v2.1' tool respectively. Various combinations of descriptors were obtained from these tools and the final combinations were chosen on the basis of internal and external validation metrics values (see section ***3.1.5. Statistical Validation Metrics*** for details). The best combinations of descriptors from the 4 datasets were obtained from BSS which show acceptable values of validation parameters. Finally four partial least square (PLS) models were constructed from the best descriptor combinations by using 'PLS_SingleY_1.0' software tool (available at http://teqip.jdvu.ac.in/QSAR_Tools/). After development of the models, the predictivity of the models was checked by rigorous external validation using 'XternalValidationPlus_1.2' tool (available at http://teqip.jdvu.ac.in/QSAR_Tools/). The whole model development procedure is schematically represented in **Figure 3.1** and the size reduction of descriptor matrices of four datasets in each step of model development has been shown in **Figure 3.2**.

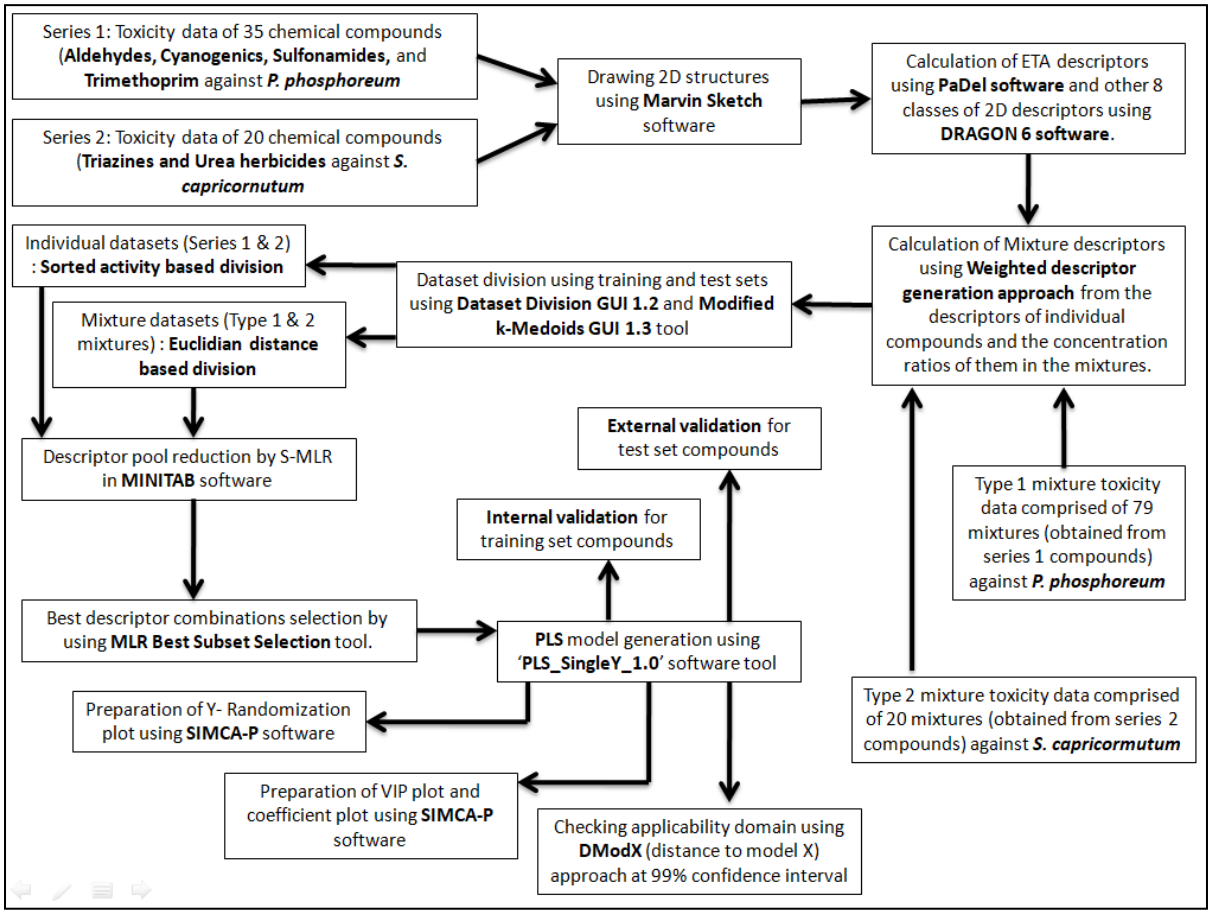


Figure 3.1. Schematic representation of the model development process

The best subset selection is a strategy to optimize the best descriptor combinations in order to obtain the important structural information responsible for the toxicity (pEC50). This method can be used for comparatively smaller descriptor pools. For larger descriptor pools, a pool reduction is required as a prerequisite of BSS to save the computation time. BSS usually explores all possible descriptor combinations among which one is taken as the best on the basis of the statistical parameters. We performed BSS by using 'MLR Best Subset Selection' tool available at http://teqip.jdvu.ac.in/QSAR_Tools/.

Partial Least Squares regression is a generalization of MLR (Wold et al., 2001). It can be employed for strongly collinear, correlated, noisy, and numerous descriptors (X-variables). It can also model several response variables (Y-variables) simultaneously. The basic idea behind PLS regression is to extract the latent variables (LVs) considering for maximum of the manifest factor variation during modeling the response. This can be achieved by extracting the latent variables T (X-scores) and U (Y-scores) from descriptors and response, respectively. The X-scores are employed to predict the Y-scores which are used for constructing the response predictions.

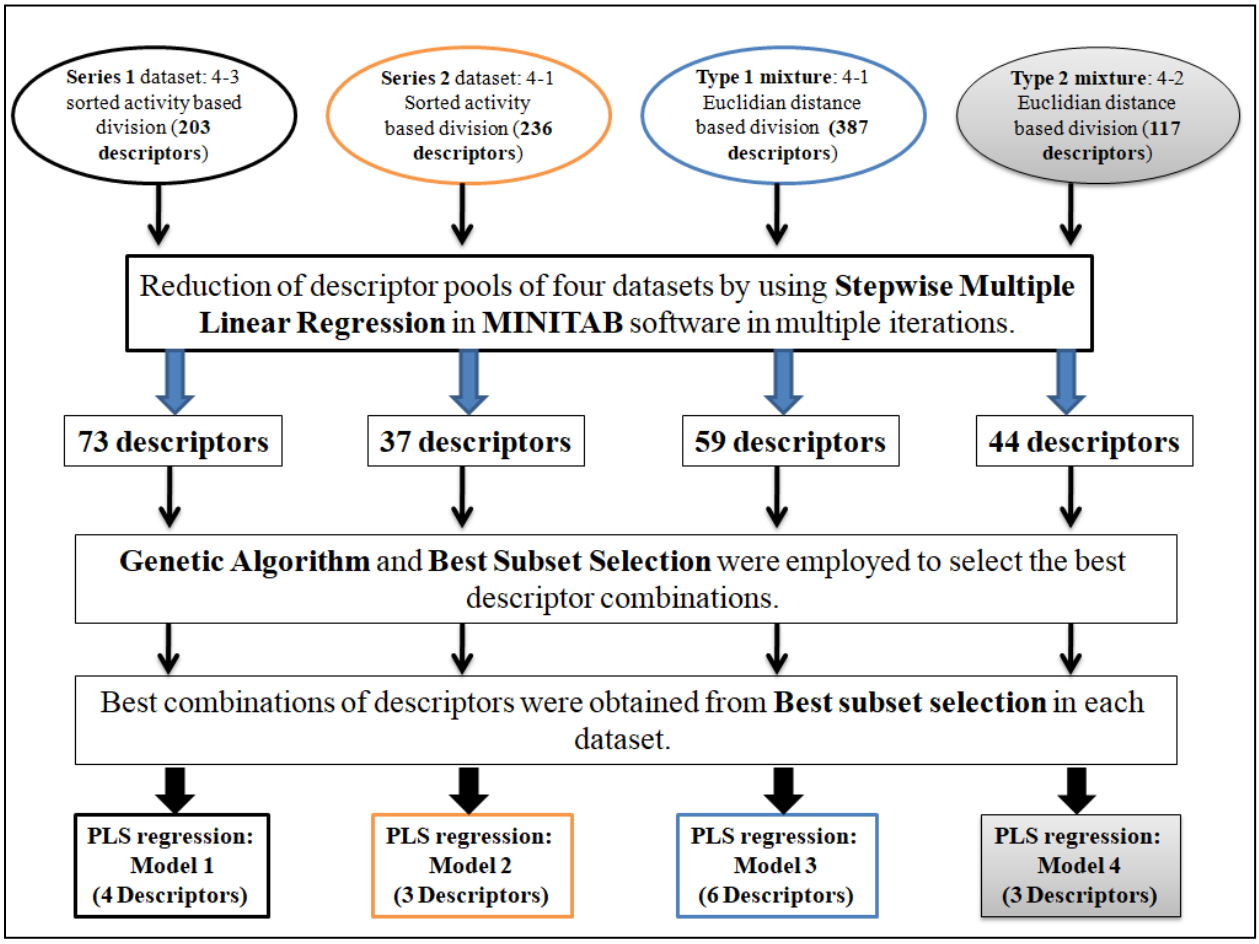


Figure 3.2. Schematic representation of the size reduction of descriptor matrices of all data set in each step of model development

3.1.5. Statistical Validation Metrics

Statistical validation metrics are employed for the evaluation of model quality, measurement of the fitness, robustness, stability and predictivity of the constructed models. The determination coefficient (R^2) and adjusted R^2 are the measures of goodness-of-fit (Roy et al., 2016). For an ideal model, the R^2 value should be equal to 1, and as it deviates from the ideal value, the quality of the model gets deteriorated. If we go on increasing the number of descriptors in a model with a constant number of observations, then R^2 will keep increasing as well but this will lead to a low statistical reliability and a decrease of degree of freedom. Calculation of adjusted R^2 considers the number of descriptors into account and it penalizes the models with large descriptor numbers. Internal validation parameters (based on the training set compounds) are computed to check the robustness of constructed models. It is

determined by a cross-validated correlation coefficient Q^2_{LOO} (Leave-one-out). But the internal validation parameters are not sufficient to claim on the performance of a model for a new set of compounds. The external validation deals with the predictive ability of the model based on test set compounds. It can be determined by checking parameters like predictive R^2 (R^2_{Pred}) or Q^2_{F1} and Q^2_{F2} (Chirico and Gramatica, 2012). Additionally, the r_m^2 rank metrics like average $r_m^2{}_{(LOO)}$, $\Delta r_m^2{}_{(LOO)}$, average $r_m^2{}_{(test)}$, and $\Delta r_m^2{}_{(test)}$ were evaluated to check the internal, external, and overall predictivity (Roy et al., 2015).

3.1.6. Applicability Domain

The applicability domain (AD) is the theoretical region in chemical space, defined by the model descriptors and the modeled response where the predictions are reliable (Gadaleta et al., 2016). According to the OECD (Organization of Economic Cooperation and Development) principle 3, the applicability domains of all QSAR models should be identified so that one can predict the untested compounds reliably with the developed model if they fall in the AD space of training set compounds. To comply with the OECD principle 3, we performed the AD assessment of all the studied models in this work. DModX approach is usually used to predict the applicability domain of PLS models and is defined in the following equation 3.3 (Wold et al., 2001).

$$DModX = \frac{\sqrt{\frac{SSE_i}{K-A}}}{\sqrt{\frac{SSE}{(N-A-A_0)(K-A)}}} \qquad (3.3)$$

For observation *i*, in a model with *A* components, *K* descriptors, and *N* observations, SSE is the squared sum of the residual of X space. A_0 is 1 if the model is centered and 0 otherwise. If the DModX value larger than around 2.5 times of the standard deviation (SD) of the X residuals (correspondent to an F value of 6.25) signifies that the observation is outside the AD of the model considering a normally distributed data. The DModX has been performed at a confidence level of 99% using SIMCA-P software ("U. Simca-P, 10.0," 2002) to get the applicability domain analysis. A threshold line is attributed as D-critical line and in this study, a confidence level of 99% has been used as standard (D-Crit = 0.009999).

3.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

3.2.1. Toxicity data

The toxicity data were collected from published literature (Lin et al., 2003), where Log 1/EC50 (negative logarithm of median effective concentration) of 84 binary mixtures were reported in mol/l scale against the bioluminescent marine bacterium *Photobacterium phosphoreum*. The binary mixtures were constructed from four environmental non-polar narcotic toxicants (chlorobenzene: #1; 1,2,3-trichlorobenzene: #2; bromobenzene: #3; 1-bromo-2,3-dichlorobenzene: #4) and 4 environmental polar narcotic toxicants (phenol: #5; 2,4-dichlorophenol: #6; aniline: #7; 3,4-dichloroaniline: #8) as classified by (Veith and Broderius, 1990). Each combination of mixtures was reported thrice in three specific ratios (0.5:1; 1:1; 1:0.5). Among the 84 binary mixtures, 18 mixtures (6 combinations) were constructed from only non-polar narcotic toxicants (#1 to #4), 18 mixtures (6 combinations) were constructed from only polar narcotic toxicants (#5 to #8), and the rest 48 mixtures (16 combinations) were constructed from both the non-polar and polar narcotic toxicants (#1 to #8). The toxicity data of all mixtures have been enlisted in **Table 3.5**. The toxicity study was

reported to be performed (Lin et al., 2003) in Microtox test instrument (Model toxicity analyzer DXY-2) using the freeze-dried *Photobacterium phosphoreum* (T_3 mutation). The organism was reconstituted in agar slants at 4ºC and used in the bioluminescence assays. The assays were performed in yeast-tryptone-salts-glycerol broth culture, maintained at 20ºC for 12-14 hours. 3% NaCl solutions of test samples were applied to the bacteria culture and the intensity of the luminance was measured after 15 min of exposure. Each mixture was tested thrice to check its reproducibility (Lin et al., 2003). The EC_{50} was computed using the Probit model (Finney, 1971) from the measured luminance and reported as Log 1/EC50 (mol/l).

Table 3.5. Toxicity data, Concentration ratio, and components of studied mixtures

Mixture No.	Components in Mixture	Mass Ratio	Log 1/EC50 (mol/l)
51	Chlorobenzene (#1) + Bromobenzene(#3)	2:1	3.73
52*	Chlorobenzene (#1) + Bromobenzene(#3)	1:1	3.67
53	Chlorobenzene (#1) + Bromobenzene(#3)	1:2	3.76
54*	Chlorobenzene (#1) + 1,2,3-Trichlorobenzene (#2)	2:1	3.89
55	Chlorobenzene (#1) + 1,2,3-Trichlorobenzene (#2)	1:1	3.85
56	Chlorobenzene (#1) + 1,2,3-Trichlorobenzene (#2)	1:2	4.21
57	Chlorobenzene (#1) + 1-Bromo-2,3-dichlorobenzene (#4)	2:1	3.9
58*	Chlorobenzene (#1) + 1-Bromo-2,3-dichlorobenzene (#4)	1:1	3.74
59	Chlorobenzene (#1) + 1-Bromo-2,3-dichlorobenzene (#4)	1:2	4.07
60	Bromobenzene (#3) + 1,2,3-Trichlorobenzene (#2)	2:1	4.06
61	Bromobenzene (#3) + 1,2,3-Trichlorobenzene (#2)	1:1	3.98
62*	Bromobenzene (#3) + 1,2,3-Trichlorobenzene (#2)	1:2	4.19
63*	Bromobenzene (#3) + 1-Bromo-2,3-dichlorobenzene (#4)	2:1	4.04
64	Bromobenzene (#3) + 1-Bromo-2,3-dichlorobenzene (#4)	1:1	3.95

65	Bromobenzene (#3) + 1-Bromo-2,3-dichlorobenzene (#4)	1:2	4.18
66	1,2,3-Trichlorobenzene (#2) + 1-Bromo-2,3-dichlorobenzene (#4)	2:1	4.7
67*	1,2,3-Trichlorobenzene (#2) + 1-Bromo-2,3-dichlorobenzene (#4)	1:1	4.76
68	1,2,3-Trichlorobenzene (#2) + 1-Bromo-2,3-dichlorobenzene (#4)	1:2	4.7
69*	Chlorobenzene (#1) + Phenol (#5)	2:1	3.2
70	Chlorobenzene (#1) + Phenol (#5)	1:1	3.03
71	Chlorobenzene (#1) + Phenol (#5)	1:2	2.91
72	Chlorobenzene (#1) + 2,4-Dichlorophenol (#6)	2:1	3.57
73	Chlorobenzene (#1) + 2,4-Dichlorophenol (#6)	1:1	3.42
74	Chlorobenzene (#1) + 2,4-Dichlorophenol (#6)	1:2	3.68
75	Chlorobenzene (#1) + Aniline (#7)	2:1	2.45
76*	Chlorobenzene (#1) + Aniline (#7)	1:1	2.45
77	Chlorobenzene (#1) + Aniline (#7)	1:2	2.22
78	Chlorobenzene (#1) + 3,4-Dichloroaniline (#8)	2:1	3.52
79	Chlorobenzene (#1) + 3,4-Dichloroaniline (#8)	1:1	3.67
80*	Chlorobenzene (#1) + 3,4-Dichloroaniline (#8)	1:2	3.81
81	Bromobenzene (#3) + Phenol (#5)	2:1	3.17
82	Bromobenzene (#3) + Phenol (#5)	1:1	3.28
83	Bromobenzene (#3) + Phenol (#5)	1:2	3.07
84	Bromobenzene (#3) + 2,4-Dichlorophenol (#6)	2:1	3.8

85	Bromobenzene (#3) + 2,4-Dichlorophenol (#6)	1:1	3.77
86	Bromobenzene (#3) + 2,4-Dichlorophenol (#6)	1:2	3.82
87	Bromobenzene (#3) + Aniline (#7)	2:1	2.63
88	Bromobenzene (#3) + Aniline (#7)	1:1	2.68
89*	Bromobenzene (#3) + Aniline (#7)	1:2	2.38
90	Bromobenzene (#3) + 3,4-Dichloroaniline (#8)	2:1	2.91
91	Bromobenzene (#3) + 3,4-Dichloroaniline (#8)	1:1	3.87
92	Bromobenzene (#3) + 3,4-Dichloroaniline (#8)	1:2	4.04
93	1,2,3-Trichlorobenzene (#2) + Phenol (#5)	2:1	3.05
94	1,2,3-Trichlorobenzene (#2) + Phenol (#5)	1:1	3.39
95*	1,2,3-Trichlorobenzene (#2) + Phenol (#5)	1:2	2.98
96*	1,2,3-Trichlorobenzene (#2) + 2,4-Dichlorophenol (#6)	2:1	4.41
97*	1,2,3-Trichlorobenzene (#2) + 2,4-Dichlorophenol (#6)	1:1	4.27
98	1,2,3-Trichlorobenzene (#2) + 2,4-Dichlorophenol (#6)	1:2	4.21
99	1,2,3-Trichlorobenzene (#2) + Aniline (#7)	2:1	2.39
100	1,2,3-Trichlorobenzene (#2) + Aniline (#7)	1:1	2.63
101	1,2,3-Trichlorobenzene (#2) + Aniline (#7)	1:2	2.45
102*	1,2,3-Trichlorobenzene (#2) + 3,4-Dichloroaniline (#8)	2:1	4.1
103	1,2,3-Trichlorobenzene (#2) + 3,4-Dichloroaniline (#8)	1:1	4.29
104	1,2,3-Trichlorobenzene (#2) + 3,4-Dichloroaniline (#8)	1:2	4.15

105*	1-Bromo-2,3-dichlorobenzene (#4) + Phenol (#5)	2:1	3.32
106*	1-Bromo-2,3-dichlorobenzene (#4) + Phenol (#5)	1:1	3.42
107	1-Bromo-2,3-dichlorobenzene (#4) + Phenol (#5)	1:2	3.12
108*	1-Bromo-2,3-dichlorobenzene (#4) + 2,4-Dichlorophenol (#6)	2:1	4.61
109	1-Bromo-2,3-dichlorobenzene (#4) + 2,4-Dichlorophenol (#6)	1:1	4.66
110	1-Bromo-2,3-dichlorobenzene (#4) + 2,4-Dichlorophenol (#6)	1:2	4.53
111	1-Bromo-2,3-dichlorobenzene (#4) + Aniline (#7)	2:1	2.56
112	1-Bromo-2,3-dichlorobenzene (#4) + Aniline (#7)	1:1	2.91
113	1-Bromo-2,3-dichlorobenzene (#4) + Aniline (#7)	1:2	2.57
114	1-Bromo-2,3-dichlorobenzene (#4) + 3,4-Dichloroaniline (#8)	2:1	4.29
115	1-Bromo-2,3-dichlorobenzene (#4) + 3,4-Dichloroaniline (#8)	1:1	4.52
116	1-Bromo-2,3-dichlorobenzene (#4) + 3,4-Dichloroaniline (#8)	1:2	4.32
117	Phenol (#5) + 2,4-Dichlorophenol (#6)	2:1	3.22
118*	Phenol (#5) + 2,4-Dichlorophenol (#6)	1:1	3.11
119	Phenol (#5) + 2,4-Dichlorophenol (#6)	1:2	3.29
120	Phenol (#5) + Aniline (#7)	2:1	2.18
121	Phenol (#5) + Aniline (#7)	1:1	2.5
122	Phenol (#5) + Aniline (#7)	1:2	2.33
123	Phenol (#5) + 3,4-Dichloroaniline (#8)	2:1	3.37
124	Phenol (#5) + 3,4-Dichloroaniline (#8)	1:1	3.16

125	Phenol (#5) + 3,4-Dichloroaniline (#8)	1:2	3.54
126	2,4-Dichlorophenol (#6) + Aniline (#7)	2:1	2.43
127	2,4-Dichlorophenol (#6) + Aniline (#7)	1:1	2.6
128*	2,4-Dichlorophenol (#6) + Aniline (#7)	1:2	2.61
129	2,4-Dichlorophenol (#6) + 3,4-Dichloroaniline (#8)	2:1	4.17
130	2,4-Dichlorophenol (#6) + 3,4-Dichloroaniline (#8)	1:1	4.44
131	2,4-Dichlorophenol (#6) + 3,4-Dichloroaniline (#8)	1:2	4.23
132*	Aniline (#7) + 3,4-Dichloroaniline (#8)	2:1	2.55
133	Aniline (#7) + 3,4-Dichloroaniline (#8)	1:1	2.5
134*	Aniline (#7) + 3,4-Dichloroaniline (#8)	1:2	2.71

*Test set mixtures; #Serial number of component

3.2.2. Molecular descriptors

Molecular descriptors are defined as the numerical values of the structural features of molecules that are essentially used for correlating the molecular structure with different molecular properties (activity/toxicity/property) in the QSAR study (Gramatica, 2020; Roy et al., 2015). There are various classes of descriptors available which can be used for QSAR analysis. Based on dimensionality, descriptors can be classified into eight classes (0 dimensional to 7 dimensional) (Roy et al., 2015). The lower-dimensional descriptors (0D and 1D) represent various constitutional features of molecules such as molecular weight, number of atoms, number of bonds etc. Two-dimensional descriptors need 2D molecular graph representation for calculations, and they include topological descriptors, 2D atom pairs, connectivity index etc. Three-dimensional descriptors are of higher-order and need the pretreatment of 3D structures (energy minimization and conformational analysis) for calculations. Electronic parameters, spatial parameters, WHIM descriptors, molecular surface area parameters etc. are examples of 3D descriptors. Four-dimensional to seven-dimensional descriptors usually represent various interactive parameters of ligand and receptors, solvation circumstances, induced-fit parameters etc. These are computationally more exhaustive, tough to interpret, and time-consuming for computation, thus they are used in much less extent in conventional QSAR. Among all classes, 0D to 2D descriptors are the most simple, easy to calculate, and easily interpretable descriptors which can efficiently correlate the structural properties with the molecular properties. These descriptors have been used in various previously published QSAR studies also (Chatterjee and Roy, 2021; Hoover et al., 2019). Therefore, for the convenience of work, different sub-classes of 0D to 2D descriptors (constitutional indices, ring descriptors, molecular properties, connectivity index, 2D atom pairs, functional group counts, atom type-E-state indices, atom-centred fragments, and extended topochemical atom indices) have been utilized in this communication.

Computation of mixture descriptors is not an easy task and cannot be obtained directly as software output. Mixture descriptors were calculated in two steps in this work. Firstly, the previously mentioned descriptors of individual mixture components were calculated using alvaDesc software (Mauri, 2020). Before descriptor calculation, 2D structures (including explicit hydrogen) of mixture components were drawn and properly aromatized using Marvin Sketch software ("Marvin|ChemAxon," 2013). The 2D structures were taken into the descriptor calculation step without prior energy minimization or conformational analysis (which is required only for higher dimensional descriptors). After getting the descriptors of individual components, mixture descriptors have been obtained by using the weighted descriptor generation approach using a Python code available in laboratory (Paul, 2021). According to this principle, one can get the mixture descriptors by summing up the multiplications of descriptor value of mixture components and the mass ratio of the corresponding components in the mixture. For simplification, the weighted descriptor generation approach has been mathematically defined in the following equation 3.4 (Hoover et al., 2019; Kar and Leszczynski, 2019)-

$$D_iMix = \sum D_iComponent \times X_{Component} \qquad (3.4)$$

where D_iMix is i^{th} the descriptor of a mixture; $D_iComponent$ is the i^{th} descriptor of a specific component, $X_{Component}$ is the mass ratio of the corresponding component, and $\sum X_{Component}$ is always equal to 1 for a mixture.

3.2.3. Dataset division

Dataset division is an integral part of QSAR modeling. A dataset is usually divided into training and test sets for model development and validation. The training set is used for model building, whereas the test set is used for rigorous external validation. The compositions of both sets are kept mutually exclusive in classical QSAR analysis so that the developed model can be properly validated by using a set of molecules (test set) completely different from the training set. In mixture QSARs, the values of mixture descriptors are unique and that is why the test and training sets of mixture data attain mutual exclusiveness in quantitative terms. In the present communication, the mixture data have been divided into training and test sets by utilizing four statistical data division tools namely, *(i)* Euclidean distance-based division, *(ii)* Kennard-Stone division, *(iii)* Sorted response-based division, *(iv)* Modified k-Medoids clustering. Among these four techniques, the sorted response-based division is a Y-variable-based strategy, whereas the rest are X-variable-based strategies. The modified k-medoids clustering was performed in the "Modified k-Medoid GUI 1.3" tool and the rest techniques were executed by the "Dataset Division GUI 1.2" tool available at http://teqip.jdvu.ac.in/QSAR_Tools/. From these four distinct methods, 13 types of division (Kennard-Stone division: 70% training set and 30% test set; 4 seeds of Euclidean distance-based divisions: 75% training set and 25% test set; 4 seeds of sorted response-based divisions: 75% training set and 25% test set; and 4 seeds of modified k-medoids divisions: 75% of training set and 25% test set) were obtained. Based on the validation metrics (Q^2_{LOO}, Q^2_{F1}) obtained from the preliminary stepwise multiple linear regression (S-MLR) models, the 4^{th} seed- 4 clustered (every 4^{th} data point of each cluster is taken into the test set) sorted response-based division was chosen for further analysis.

3.2.4. Feature selection and model development

After getting the best division among all, the descriptor pool was initially reduced by a multilayered (multiple iterations of) stepwise regression algorithm in the MINITAB software. Two feature selection techniques namely, Genetic algorithm (Leardi, 2001) and Best subset selection (Roy et al., 2015) were used afterwards for identifying the most important structural features of the descriptor matrix responsible for mixture toxicity. The Genetic algorithm and Best subset selection were executed by "DCVTool_v2.0" and "BestSubsetSelection_v2.1" respectively and various descriptor combinations were obtained. The best combination was chosen based on the statistical metrics values (section "Internal and external validation"). The best combination of descriptors was obtained from Best subset selection and at the end, a partial least square (PLS) regression model was developed by "PLSSingleY_v1.0" tool from the selected descriptors. Another tool "XternalValidationPlus_v1.2" was used for the rigorous external validation of the constructed model with the test set of mixtures. Various PLS plots such as Variable importance plot, Coefficient plot, Applicability domain plot, Y-randomization plot etc. were developed in "SIMCA-P" software for visual interpretation of the developed model. The complete model development process has been schematically represented in the **Figure 3.3**.

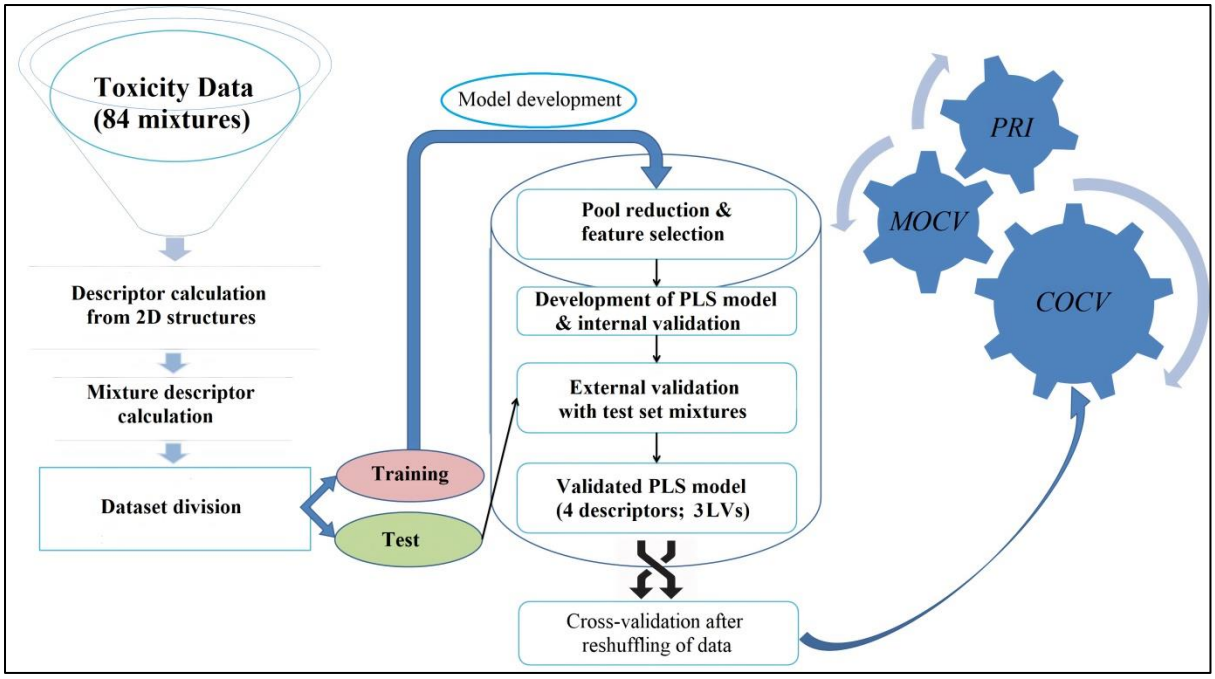


Figure 3.3. Schematic representation of the model development process and validation

The best subset selection is one of the feature selection strategies available for the QSAR study. It finds out the best descriptor combination to identify the important structural features responsible for toxicity (biological activity in general). It explores all possible descriptor combinations and that is why it is a computationally exhaustive process. Therefore, to minimize the computation time, a reduced pool of descriptors should be given as input instead of a whole descriptor matrix. The best combination of descriptors is selected based on the computed values of internal and external validation metrics provided by the tool itself. We performed this process in the "BestSubsetSelection_v2.1" tool available at http://teqip.jdvu.ac.in/QSAR_Tools/.

Partial least squares (PLS) regression is the generalized form of the multiple linear regression method (Wold et al., 2001). It is more advanced than MLR and can be used for correlated, strongly collinear, and noisy data. PLS can handle numerous X variables (descriptors) and various Y variables (responses) simultaneously. As an algorithm, PLS extracts the latent variables (LVs) considering maximum variation in the descriptor matrix during modeling the response. It can be achieved by calculating the X-scores (latent variable T) and Y-score (latent variable U) from the X and Y variables respectively. The T vectors ate utilized to get the U vectors whereas the U vectors are further employed to get the response prediction. During reporting a PLS model, the number of LVs is very important and it should not be equal to or more than the number of descriptors in the model.

3.2.5. Internal and external validation

Internal and external validation tests are executed for the judgment of model quality, measurement of the fitness, robustness, reproducibility, and predictive ability of the developed models. The internal validation metrics (R^2, R^2_{adj}, Q^2_{LOO} etc.) are employed to check the goodness-of-fit and robustness of the model, whereas the external validation metrics (Q^2_{F1}, Q^2_{F2}, $RMSE_P$ etc.) are employed to judge the overall predictivity. The fitness of the model is judged by measuring the determination coefficient (R^2) and adjusted R^2 (Roy et al., 2015). The quality of a model gets deteriorated with the deviation of R^2 from its ideal

value 1. The value of R^2 depends upon the number of descriptors used in the model; hence it increases with increasing the number of descriptors. This creates the problem of overfitting and low statistical reliability. Therefore, to avoid overfitting, the metric R^2_{adj} is employed. The computation of R^2_{adj} considers the number of X-variables, and hence it penalizes the model with large number of descriptors. The robustness of a model is adjudged by another internal validation metric namely, cross-validated correlation coefficient Q^2_{LOO} (Leave-one-out). The ideal value of Q^2_{LOO} is also 1 and any deviation from the ideal value suggests the deterioration of the model quality. Various error-based metrics such as, mean average error (MAE) and root mean square error of prediction ($RMSE_P$) are also used to check the error of predictions of a model and these metric values should be near 0. The internal validation metrics and the error-based metrics cannot suggest the predictive performance of the model on a new set of compounds; thus, external validation is used. External validation metrics are the measures of the predictive ability of a model and are computed from an external set of compounds called the test set. The external correlation coefficients like Q^2_{F1} (also termed as R^2_{Pred}) and Q^2_{F2} are employed to determine the predictivity of the model (Chirico and Gramatica, 2012). Some additional r_m^2 rank metrics (e.g. average $r_m^2{}_{(LOO)}$, $\Delta r_m^2{}_{(LOO)}$, average $r_m^2{}_{(test)}$, and $\Delta r_m^2{}_{(test)}$) are also calculated to evaluate the internal, external, and overall predictivity (Roy et al., 2015).

The Variable importance plot (VIP) is a "SIMCA-P" plot which signifies the order of importance of descriptors towards modeled response. The descriptors having a VIP score of more than 1 are statistically more significant and they are arranged in descending order of VIP score in the graph from left to right (Akarachantachote et al., 2014). Coefficient plots show the positive and negative relationships of the descriptors with the modeled response. This is essentially a bar diagram obtained from "SIMCA-P" where positive coefficients can be identified by the upward bars and negative coefficients can be identified by the downward bars. The positive and negative coefficients signify the proportional and inverse relationships of the descriptors respectively, with the response. Y-randomization is a process to check the developed model, whether it is a result of any chance correlation or not (Topliss and Edwards, 1979). For that, the Y variables of the training set are shuffled randomly and X-variables remain unaltered. The newly formed model is then fitted with the permuted random models and compared with the best fit. The resulting R^2 and Q^2 values of the random models are plotted (Y-axis) against the correlation coefficient among original Y values and permuted Y values (X-axis). Very poor R^2 and Q^2 values (<0.05) signify the statistical credibility of the developed model (Topliss and Edwards, 1979). This Y-randomization plot was obtained in this communication from "SIMCA-P" software and discussed in the Result and Discussion section.

The applicability domain is a theoretical portion of the chemical space defined by the modeled descriptors and response, in which the predictions are reliable (Gadaleta et al., 2016). The toxicity of a complete spectrum of chemical substances cannot be predicted with sufficient confidence with a single QSAR model due to their structural diversity, and that is why the QSAR modelers have to provide the applicability domain plot of the developed models so that the toxicity can be predicted with reliability. OECD (Organization of Economic Co-operation and Development) also suggests to analyze the applicability domain of QSAR models in principle 3 (OECD, 2004); thus to comply with the guideline, we have carried out the AD analysis in this research. Among various methods of applicability domain analysis, DModX is used for the PLS models, and we used "SIMCA-P" software in this communication to obtain the DModX plot. The DModX approach is defined in the previous equation 3.3 (Wold et al., 2001). If the DModX score is higher than 2.5 times of the standard

deviation of the X residuals (F value 6.25), it indicates that the observation is beyond the applicability domain of the particular model considering a normal data distribution. A threshold line is termed as a D-critical line and in this communication, a confidence level of 99% has been applied as the standard (D-Crit = 0.00999898).

3.2.6. Cross-validation

Mixture QSAR is different from the conventional ones in various aspects, and one of the important aspects is the dataset division. As seen in the 'Molecular descriptors' section, each mixture descriptor is composed of the descriptors of two different chemicals; therefore, there is structural information of more than one pure compound in each data point of the mixture. Now, this leads to a question on structural exclusiveness between training and test sets during model development, because a single mixture component may be present in both test and training sets simultaneously as different mixtures. It may lead us to compute overestimated validation metrics and hence the quality of the developed model may be misleading. To overcome this problem, Oprisiu et al. suggested 3 different cross-validation (CV) strategies namely, the Points-out cross-validation, Mixtures-out cross-validation, and the Compounds-out cross-validation (Oprisiu et al., 2012) whereas, Muratov et al suggested the Everything out cross-validation exclusively for mixture QSARs (Muratov et al., 2014).

To obtain the cross-validated metrics, the whole data set is again divided manually by following specific patterns of different cross-validations. These new training sets are used to develop new PLS models with the same combination of descriptors and the same number of latent variables as in the original PLS model; then these resultant models are validated with the new test sets and the metrics are calculated. In mixtures out CV, the mixtures (those specific combinations of component chemicals) are unique among themselves in training and test sets. In other words, the test mixtures must not be present in the training set. Although in this cross-validation, the individual components may be present in both sets as separate mixtures, the mixtures are unique among themselves with respect to the new training and test sets. The compounds-out CV suggests the partial structural exclusiveness between the training and test sets. The binary test mixtures must contain one component which is unique with respect to the training set. To perform this cross-validation, the data set was divided again. The compounds-out CV has been achieved after keeping out the common mixture component of the test set (#5: phenol) from the training set. The training and test sets of both the mixtures-out and compounds-out CV have been stored in an openly accessible online repository that is available at https://github.com/mchatt1012/PhD-Thesis_QSAR-Modeling-of-Toxicity-of-Chemical-Mixtures.git. The PLS regression has been carried out again twice with the same set of descriptors and same number of LVs with the new training-test pairs by "PLSSingleY_v1.0" tool (available at http://teqip.jdvu.ac.in/QSAR_Tools/.) to get the cross-validated metrics. These cross-validations have been pictorially represented in **Figure 3.4** for better understanding.

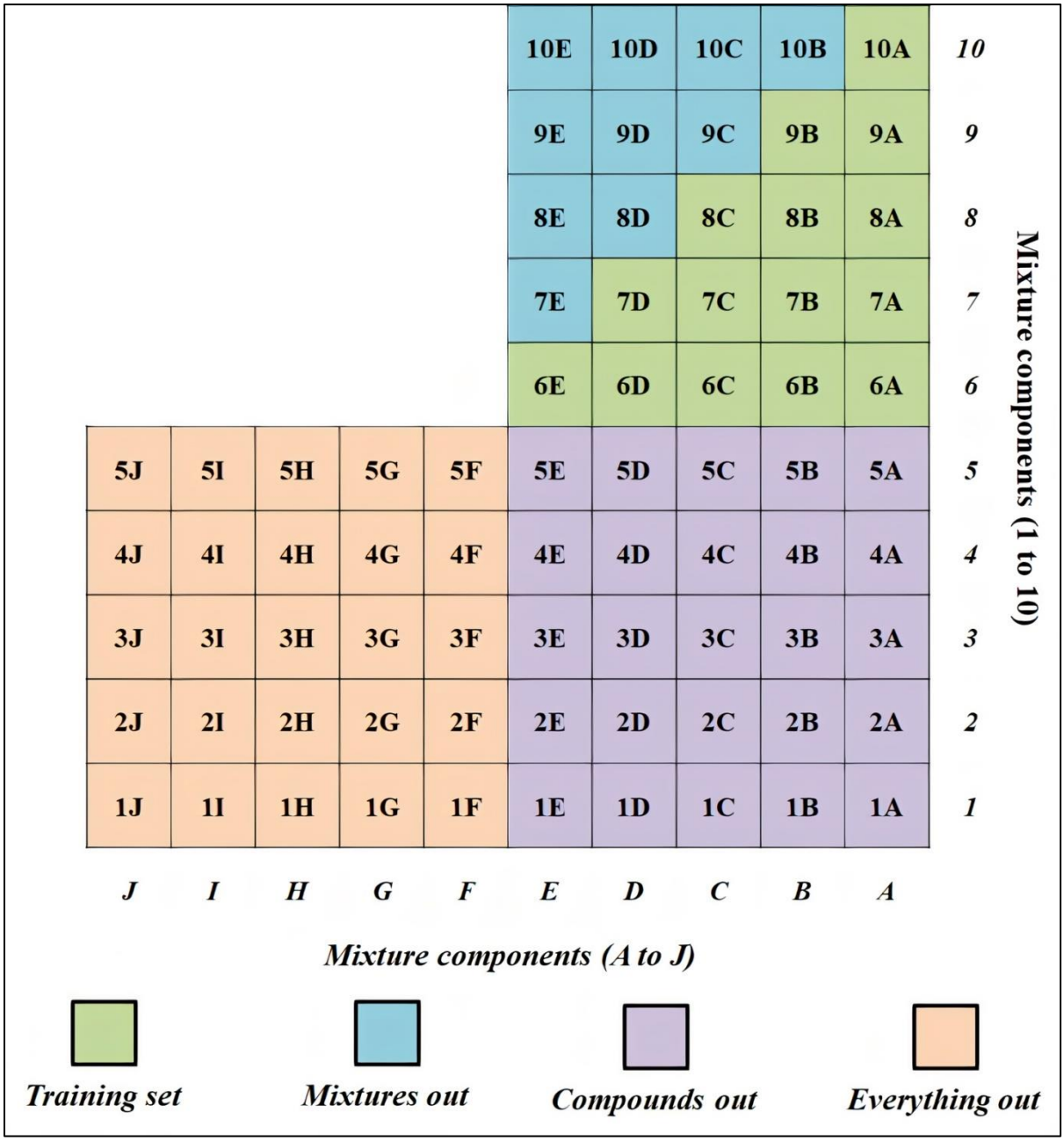


Figure 3.4. Division patterns of mixtures out, compounds out, and everything out cross-validations (The boxes represent the mixtures)

The Everything-out CV is the complete structural exclusiveness between the training and test sets which can be achieved by placing completely different mixture components in the test set from the training set. In other words, all the components of the mixtures of the test set should be structurally different from the training set mixtures. Due to the scarcity of toxicity data, everything-out CV could not be performed in this research; instead of this, a set of mixtures has been designed (containing unique mixture components from the original modeling set), and the quality of prediction of the developed model has been judged using this designed set by a Prediction reliability indicator (Roy et al., 2018a) (Section "***3.2.8. Prediction reliability indicator and designing of empirical dataset***")

3.2.7. Chemical read-across

Read-across prediction is a similarity-based *in silico* approach where the toxicity or biological response of untested chemicals (target chemical) is predicted from the known toxicity or biological response values of chemical analogue/s (source chemical/s) (Raies and Bajic, 2016). Two types of read-across algorithms are mostly seen, such as the analogue approach (the toxicity/response of an untested chemical is predicted from a single known analogue) and the category approach (the toxicity/response of an untested chemical is predicted from more than one known analogues). The category approach is more reliable in terms of quantitative prediction because the analogue approach can provide the toxicity value of itself as the prediction. Recently, the present authors' group has reported a machine learning-based chemical read-across tool- 'Read-across v4.1' which can predict the toxicity (biological activity in general) based on the structural similarity of the source and target chemicals (Banerjee and Roy, 2022). This tool calculates the similarity (structural similarity or the similarity of descriptors) by three distinct approaches namely Euclidean distance-based approach, Gaussian kernel function similarity, and Laplacian kernel function similarity, and gives the prediction by computing the weighted average of the similar source compounds based on their similarity (Chatterjee et al., 2022).

In this study, we have used the machine learning-based chemical read-across algorithm for the prediction of toxicity of the test mixtures with the modeled descriptors. For a successful prediction with Gaussian ad Laplacian kernel function similarity estimations, we have optimized the hyperparameter sigma (σ) and gamma (γ) respectively. For the optimization, the original training set (used for QSAR modeling) is randomly divided into subtrain and subtest sets in 3:1 proportion. The subtrain and subtest sets were then subjected to 'Read-across v4.1' with different σ and γ values (we have used 0.25 to 2 as σ and γ with 0.25 interval). The other determinants of the tool such as the number of close training compounds (10), distance threshold (1), and similarity threshold (0) were kept constant during optimization. The optimized setting has been chosen by checking the quality metrics Q^2_{F1} (calculated by the tool itself with the predictions). The optimized setting was used finally with the original training and test set to get the final prediction. We have gradually decreased the number of close training compounds (10 to 2) to get the best predictions.

3.2.8. Prediction reliability indicator and designing of empirical dataset

To develop a good QSAR model, proper validation is necessary, and to comply with the strict validation criteria, we performed external validation with the test set. Due to the scarcity of available mixture data, we could not perform the true external validation with an external dataset. Thus, we have designed an empirical mixture dataset (18 data points) to check the reliability of the prediction of the untested mixtures by the developed QSAR model. The designed dataset has been shown in the following **Table 3.6**. This dataset contains 18 empirical mixtures with equal mass (**0.5:0.5**) of two different components. Mixture No. 135 to 148 (8 combinations) are designed from a compound of the training set (#1 to #8) [11] and a completely new compound (#9 to #13) with respect to the original training set. On the contrary, mixture no. 143 to 152 (10 combinations) were designed from completely new compounds (#9 to #13) which were not present in the training set. The new compounds are the close congeners of the mixture components of this study and that is the reason behind the selection. The structures of the new pure compounds were collected from the PubChem chemical database maintained by National Centre for Biotechnology Information, USA (https://pubchem.ncbi.nlm.nih.gov/).

The mixture descriptors were calculated using the method discussed in the section "***3.2.2. Molecular Descriptor***". The reliability of the prediction of the query dataset was adjudged by the statistical tool called "Prediction Reliability Indicator" (Roy et al., 2018a) which is available from http://teqip.jdvu.ac.in/QSAR_Tools/. This tool categorizes the reliability of prediction of the query chemicals of the designed dataset into 'Good', 'Moderate', and 'Bad', depending upon the composite score. This tool essentially uses three sets of principles for the categorization of the quality of predictions and calculate the composite score; these are- (i) mean absolute error of prediction (leave-one-out) of a query compound from 10 most close training compounds; (ii) applicability domain in terms of similarity, by using the standardization approach; (iii) closeness of predicted values to the mean of the observed response of training set compounds.

Table 3.6. Designed query mixture dataset

Mixture No.	Component A	Component B	Mass ratio
135	Chlorobenzene (#1)	2-bromo-4-chlorophenol (#11)	0.5:0.5
136	1,2,3-Trichlorobenzene (#2)	3,4-dibromoaniline (#12)	0.5:0.5
137	Bromobenzene(#3)	p-Chloroaniline (#13)	0.5:0.5
138	1-Bromo-2,3-dichlorobenzene (#4)	2,4-Dibromo-1-chlorobenzene (#9)	0.5:0.5
139	Phenol (#5)	2,4-Dibromo-1-chlorobenzene (#9)	0.5:0.5
140	2,4-Dichlorophenol (#6)	Fluorobenzene (#10)	0.5:0.5
141	Aniline (#7)	p-Chloroaniline	0.5:0.5
142	3,4-Dichloroaniline (#8)	Fluorobenzene (#10)	0.5:0.5
143	2,4-Dibromo-1-chlorobenzene (#9)	2-bromo-4-chlorophenol (#11)	0.5:0.5
144	2,4-Dibromo-1-chlorobenzene (#9)	3,4-dibromoaniline (#12)	0.5:0.5
145	2,4-Dibromo-1-chlorobenzene (#9)	p-Chloroaniline (#13)	0.5:0.5
146	Fluorobenzene (#10)	2-bromo-4-chlorophenol (#11)	0.5:0.5
147	Fluorobenzene (#10)	3,4-dibromoaniline (#12)	0.5:0.5
148	Fluorobenzene (#10)	p-Chloroaniline (#13)	0.5:0.5
149	2,4-Dibromo-1-chlorobenzene (#9)	Fluorobenzene (#10)	0.5:0.5
150	2-bromo-4-chlorophenol (#11)	p-Chloroaniline (#13)	0.5:0.5
151	2-bromo-4-chlorophenol (#11)	3,4-dibromoaniline (#12)	0.5:0.5
152	p-Chloroaniline (#13)	3,4-dibromoaniline (#12)	0.5:0.5

#: number of mixture components; #1 to #8: mixture components present in the studied dataset; #9 to #13: components of designed mixtures

3.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutica and pesticide mixtures against *Aliivibrio fischeri*

3.3.1. Toxicity data

Good quality toxicity data (biological data) is one of the most important prerequisites of QSAR model development, and the quality of such dataset is an important factor along with the number of data points for the performance of the model. Therefore, the dataset should be chosen and curated carefully for obtaining the optimum prediction from the developed model. In this research, the toxicity dataset of 198 hazardous mixtures (178 binary mixtures, 15 ternary mixtures, and 5 quaternary mixtures) has been utilized for model development (**Table 3.7**). We have taken the toxicity data of 153 binary mixtures and 45 multi-component mixtures separately against *Aliivibrio fischeri* from two published articles by Bures et al. (Bures et al., 2021) and Qin et al. (Qin et al., 2018) respectively. The test organism (also called *Vibrio fischeri*) is a gram-negative, rod-shaped, bioluminescent marine bacterium that lives globally in symbiosis with various sea creatures such as Hawaiian bobtail squid, and it is a classic marker of aquatic toxicity (Abbas et al., 2018). The data reported by Bures et al. contains 51 binary combinations of 3 different molar concentration ratios (0.75:0.25; 0.5:0.5; 0.25:0.75) whereas, the data reported by Qin et al. contains 5 binary combinations, 3 ternary combinations, and 1 quaternary combination with 5 variable molar concentration ratios each. The studied mixtures were composed of 12 marketed pharmaceuticals and 8 marketed pesticides. All the mixture components along with their respective class have been displayed in **Table 3.8**.

Toxicity testing was performed by using freeze-dried *Aliivibrio fischeri* strain maintained in 1.5%–2% agar medium at 4°C. The liquid culture medium was prepared by adding 2 g potassium dihydrogen phosphate (KH_2PO_4), 9.4 g hydrated disodium hydrogen phosphate ($Na_2HPO_4 \cdot 12H_2O$), 0.6 g hydrated magnesium sulphate ($MgSO_4 \cdot 7H_2O$), 1 g di-ammonium hydrogen phosphate ($(NH_4)_2HPO_4$), 60 g sodium chloride (NaCl), 6.63 g tryptone, 8.78g yeast extract and 6.90 g glycerine, and dissolved in 1000 mL distilled water. The culture medium was packed in 100 mL conical flask with a volume of 50 mL, sterilized at 121 °C, and 15 psi pressure steam for 20 min, and the flasks were preserved in the 4 °C. The initial mixture solutions (0.1 mM) were firstly prepared by dissolving a specified amount of solid mixtures in deionised water. The original solutions were diluted with sodium chloride solution to obtain 2% solutions. After getting the mixture solutions, pH was adjusted to 7.0 ± 0.2 by adding hydrochloric acid and sodium hydroxide solutions, respectively. The inoculum of the luminescent bacteria was prepared in the liquid cultured broth and subsequently added to each sample. As mentioned by the standard, the test samples were incubated at 15 ± 1°C for 15 mins, and after incubation, the intensity of luminescence was measured by BioFix Lumi-10 toxicity analyzer (Macherey-Nagel GmbH & Co., Düren, Germany). The toxicity expressed as the inhibition value (E) was calculated in triplicate by the following equation 3.5-

$$E(\%) = \frac{I_0 - I}{I_0} \times 100 \qquad (3.5)$$

where I is the average relative light unit (RLU) of the test organism exposed to the test mixture sample and I_0 is an average of the RLUs of the controls (without any test sample). The inhibition values were then converted to EC_{50} (mM) by subsequent calculation. Finally, the negative logarithmic values of EC_{50} (*pEC50*) were calculated and taken for the endpoint of the study.

Table 3.7. Toxicity data, molar concentration ratio, and components of studied mixtures

Sl. No.	Composition	Molar concentration ratio	pEC50
21	Diclofenac (#6) + Oxytetracycline (#4)	0.75:0.25	1.55
22	Diclofenac (#6) + Oxytetracycline (#4)	0.5:0.5	1.44
23	Diclofenac (#6) + Oxytetracycline (#4)	0.25:0.75	1.32
24*	Diclofenac (#6) + Carbamazepine (#3)	0.75:0.25	1.4
25	Diclofenac (#6) + Carbamazepine (#3)	0.5:0.5	1.27
26	Diclofenac (#6) + Carbamazepine (#3)	0.25:0.75	1.13
27	Diclofenac (#6) + Dexamethasone (#5)	0.75:0.25	1.35
28	Diclofenac (#6) + Dexamethasone (#5)	0.5:0.5	1.24
29*	Diclofenac (#6) + Dexamethasone (#5)	0.25:0.75	1.17
30	Diclofenac (#6) + Azithromycin (#1)	0.75:0.25	1.23
31	Diclofenac (#6) + Azithromycin (#1)	0.5:0.5	1.08
32	Diclofenac (#6) + Azithromycin (#1)	0.25:0.75	0.87
33	Diclofenac (#6) + Erythromycin (#2)	0.75:0.25	1.26
34	Diclofenac (#6) + Erythromycin (#2)	0.5:0.5	1.06
35	Diclofenac (#6) + Erythromycin (#2)	0.25:0.75	0.9
36*	Oxytetracycline (#4) + Carbamazepine (#3)	0.75:0.25	1.06
37	Oxytetracycline (#4) + Carbamazepine (#3)	0.5:0.5	1.06
38	Oxytetracycline (#4) + Carbamazepine (#3)	0.25:0.75	0.89
39	Oxytetracycline (#4) + Dexamethasone (#5)	0.75:0.25	0.96
40*	Oxytetracycline (#4) + Dexamethasone (#5)	0.5:0.5	0.99
41	Oxytetracycline (#4) + Dexamethasone (#5)	0.25:0.75	0.97
42*	Oxytetracycline (#4) + Azithromycin (#1)	0.75:0.25	0.95
43	Oxytetracycline (#4) + Azithromycin (#1)	0.5:0.5	0.86
44	Oxytetracycline (#4) + Azithromycin (#1)	0.25:0.75	0.79
45	Oxytetracycline (#4) + Erythromycin (#2)	0.75:0.25	0.94
46*	Oxytetracycline (#4) + Erythromycin (#2)	0.5:0.5	0.84
47	Oxytetracycline (#4) + Erythromycin (#2)	0.25:0.75	0.76
48	Carbamazepine (#3) + Dexamethasone (#5)	0.75:0.25	0.8
49*	Carbamazepine (#3) + Dexamethasone (#5)	0.5:0.5	0.87

50	Carbamazepine (#3) + Dexamethasone (#5)	0.25:0.75	0.95
51*	Carbamazepine (#3) + Azithromycin (#1)	0.75:0.25	0.82
52	Carbamazepine (#3) + Azithromycin (#1)	0.5:0.5	0.8
53	Carbamazepine (#3) + Azithromycin (#1)	0.25:0.75	0.83
54*	Carbamazepine (#3) + Erythromycin (#2)	0.75:0.25	0.75
55	Carbamazepine (#3) + Erythromycin (#2)	0.5:0.5	0.79
56	Carbamazepine (#3) + Erythromycin (#2)	0.25:0.75	0.74
57*	Dexamethasone (#5) + Azithromycin (#1)	0.75:0.25	0.89
58	Dexamethasone (#5) + Azithromycin (#1)	0.5:0.5	0.83
59	Dexamethasone (#5) + Azithromycin (#1)	0.25:0.75	0.8
60	Dexamethasone (#5) + Erythromycin (#2)	0.75:0.25	0.82
61	Dexamethasone (#5) + Erythromycin (#2)	0.5:0.5	0.84
62	Dexamethasone (#5) + Erythromycin (#2)	0.25:0.75	0.76
63	Azithromycin (#1) + Erythromycin (#2)	0.75:0.25	0.8
64	Azithromycin (#1) + Erythromycin (#2)	0.5:0.5	0.76
65*	Azithromycin (#1) + Erythromycin (#2)	0.25:0.75	0.73
66	Isoproturon (#12) + Alachlor (#11)	0.75:0.25	0.9
67	Isoproturon (#12) + Alachlor (#11)	0.5:0.5	0.88
68*	Isoproturon (#12) + Alachlor (#11)	0.25:0.75	0.74
69	Isoproturon (#12) + Diuron (#13)	0.75:0.25	0.86
70	Isoproturon (#12) + Diuron (#13)	0.5:0.5	0.8
71*	Isoproturon (#12) + Diuron (#13)	0.25:0.75	0.64
72*	Isoproturon (#12) + Chlorfenvinphos (#14)	0.75:0.25	0.93
73	Isoproturon (#12) + Chlorfenvinphos (#14)	0.5:0.5	0.98
74	Isoproturon (#12) + Chlorfenvinphos (#14)	0.25:0.75	1
75	Alachlor (#11) + Diuron (#13)	0.75:0.25	0.69
76	Alachlor (#11) + Diuron (#13)	0.5:0.5	0.61
77	Alachlor (#11) + Diuron (#13)	0.25:0.75	0.64
78	Alachlor (#11) + Chlorfenvinphos (#14)	0.75:0.25	0.75
79	Alachlor (#11) + Chlorfenvinphos (#14)	0.5:0.5	0.88
80	Alachlor (#11) + Chlorfenvinphos (#14)	0.25:0.75	1.03
81*	Diuron (#13) + Chlorfenvinphos (#14)	0.75:0.25	0.74
82	Diuron (#13) + Chlorfenvinphos (#14)	0.5:0.5	0.9

83	Diuron (#13) + Chlorfenvinphos (#14)	0.25:0.75	0.97
84*	Diclofenac (#6) + Isoproturon (#12)	0.75:0.25	1.39
85	Diclofenac (#6) + Isoproturon (#12)	0.5:0.5	1.41
86	Diclofenac (#6) + Isoproturon (#12)	0.25:0.75	1.55
87	Diclofenac (#6) + Alachlor (#11)	0.75:0.25	1.51
88	Diclofenac (#6) + Alachlor (#11)	0.5:0.5	1.01
89	Diclofenac (#6) + Alachlor (#11)	0.25:0.75	0.82
90*	Diclofenac (#6) + Diuron (#13)	0.75:0.25	1.44
91	Diclofenac (#6) + Diuron (#13)	0.5:0.5	1.26
92	Diclofenac (#6) + Diuron (#13)	0.25:0.75	1.03
93	Diclofenac (#6) + Chlorfenvinphos (#14)	0.75:0.25	1.85
94	Diclofenac (#6) + Chlorfenvinphos (#14)	0.5:0.5	1.8
95	Diclofenac (#6) + Chlorfenvinphos (#14)	0.25:0.75	1.8
96	Oxytetracycline (#4) + Isoproturon (#12)	0.75:0.25	0.87
97	Oxytetracycline (#4) + Isoproturon (#12)	0.5:0.5	0.86
98	Oxytetracycline (#4) + Isoproturon (#12)	0.25:0.75	0.86
99	Oxytetracycline (#4) + Alachlor (#11)	0.75:0.25	1.29
100	Oxytetracycline (#4) + Alachlor (#11)	0.5:0.5	1.18
101	Oxytetracycline (#4) + Alachlor (#11)	0.25:0.75	1
102*	Oxytetracycline (#4) + Diuron (#13)	0.75:0.25	1.1
103	Oxytetracycline (#4) + Diuron (#13)	0.5:0.5	1.04
104	Oxytetracycline (#4) + Diuron (#13)	0.25:0.75	0.94
105	Oxytetracycline (#4) + Chlorfenvinphos (#14)	0.75:0.25	1.33
106	Oxytetracycline (#4) + Chlorfenvinphos (#14)	0.5:0.5	1.36
107	Oxytetracycline (#4) + Chlorfenvinphos (#14)	0.25:0.75	1.45
108	Carbamazepine (#3) + Isoproturon (#12)	0.75:0.25	0.84
109	Carbamazepine (#3) + Isoproturon (#12)	0.5:0.5	0.7
110	Carbamazepine (#3) + Isoproturon (#12)	0.25:0.75	0.71
111*	Carbamazepine (#3) + Alachlor (#11)	0.75:0.25	0.85
112	Carbamazepine (#3) + Alachlor (#11)	0.5:0.5	0.78
113	Carbamazepine (#3) + Alachlor (#11)	0.25:0.75	0.66
114	Carbamazepine (#3) + Diuron (#13)	0.75:0.25	0.66
115	Carbamazepine (#3) + Diuron (#13)	0.5:0.5	0.76

116	Carbamazepine (#3) + Diuron (#13)	0.25:0.75	0.61
117	Carbamazepine (#3) + Chlorfenvinphos (#14)	0.75:0.25	0.85
118	Carbamazepine (#3) + Chlorfenvinphos (#14)	0.5:0.5	1.02
119	Carbamazepine (#3) + Chlorfenvinphos (#14)	0.25:0.75	1.16
120*	Dexamethasone (#5) + Isoproturon (#12)	0.75:0.25	0.91
121	Dexamethasone (#5) + Isoproturon (#12)	0.5:0.5	0.75
122	Dexamethasone (#5) + Isoproturon (#12)	0.25:0.75	0.63
123	Dexamethasone (#5) + Alachlor (#11)	0.75:0.25	1.14
124	Dexamethasone (#5) + Alachlor (#11)	0.5:0.5	1.04
125	Dexamethasone (#5) + Alachlor (#11)	0.25:0.75	0.87
126*	Dexamethasone (#5) + Diuron (#13)	0.75:0.25	0.96
127	Dexamethasone (#5) + Diuron (#13)	0.5:0.5	0.91
128	Dexamethasone (#5) + Diuron (#13)	0.25:0.75	0.82
129	Dexamethasone (#5) + Chlorfenvinphos (#14)	0.75:0.25	1.1
130	Dexamethasone (#5) + Chlorfenvinphos (#14)	0.5:0.5	1.25
131	Dexamethasone (#5) + Chlorfenvinphos (#14)	0.25:0.75	1.36
132	Erythromycin (#2) + Isoproturon (#12)	0.75:0.25	0.66
133*	Erythromycin (#2) + Isoproturon (#12)	0.5:0.5	0.63
134	Erythromycin (#2) + Isoproturon (#12)	0.25:0.75	0.74
135	Erythromycin (#2) + Alachlor (#11)	0.75:0.25	1.55
136	Erythromycin (#2) + Alachlor (#11)	0.5:0.5	0.97
137*	Erythromycin (#2) + Alachlor (#11)	0.25:0.75	0.75
138	Erythromycin (#2) + Diuron (#13)	0.75:0.25	0.87
139*	Erythromycin (#2) + Diuron (#13)	0.5:0.5	0.81
140*	Erythromycin (#2) + Diuron (#13)	0.25:0.75	0.75
141	Erythromycin (#2) + Chlorfenvinphos (#14)	0.75:0.25	0.98
142	Erythromycin (#2) + Chlorfenvinphos (#14)	0.5:0.5	1.45
143	Erythromycin (#2) + Chlorfenvinphos (#14)	0.25:0.75	1.55
144	Azithromycin (#1) + Isoproturon (#12)	0.75:0.25	0.76
145	Azithromycin (#1) + Isoproturon (#12)	0.5:0.5	0.81
146	Azithromycin (#1) + Isoproturon (#12)	0.25:0.75	0.85

147*	Azithromycin (#1) + Alachlor (#11)	0.75:0.25	0.88
148	Azithromycin (#1) + Alachlor (#11)	0.5:0.5	1.06
149	Azithromycin (#1) + Alachlor (#11)	0.25:0.75	0.82
150*	Azithromycin (#1) + Diuron (#13)	0.75:0.25	1.02
151	Azithromycin (#1) + Diuron (#13)	0.5:0.5	0.96
152*	Azithromycin (#1) + Diuron (#13)	0.25:0.75	1.02
153*	Azithromycin (#1) + Chlorfenvinphos (#14)	0.75:0.25	1.08
154*	Azithromycin (#1) + Chlorfenvinphos (#14)	0.5:0.5	1.39
155	Azithromycin (#1) + Chlorfenvinphos (#14)	0.25:0.75	1.65
156	Omeprazole (#10) + Imatinib (#8)	0.75:0.25	2.98
157*	Omeprazole (#10) + Imatinib (#8)	0.5:0.5	2.7
158	Omeprazole (#10) + Imatinib (#8)	0.25:0.75	2.28
159	Omeprazole (#10) + Tobramycin (#9)	0.75:0.25	2.73
160	Omeprazole (#10) + Tobramycin (#9)	0.5:0.5	2.33
161	Omeprazole (#10) + Tobramycin (#9)	0.25:0.75	1.8
162	Omeprazole (#10) + Desloratadine (#7)	0.75:0.25	2.98
163	Omeprazole (#10) + Desloratadine (#7)	0.5:0.5	2.55
164*	Omeprazole (#10) + Desloratadine (#7)	0.25:0.75	1.93
165	Imatinib (#8) + Tobramycin (#9)	0.75:0.25	0.39
166	Imatinib (#8) + Tobramycin (#9)	0.5:0.5	0.26
167	Imatinib (#8) + Tobramycin (#9)	0.25:0.75	0.13
168*	Imatinib (#8) + Desloratadine (#7)	0.75:0.25	0.65
169	Imatinib (#8) + Desloratadine (#7)	0.5:0.5	0.49
170*	Imatinib (#8) + Desloratadine (#7)	0.25:0.75	0.29
171*	Tobramycin (#9) + Desloratadine (#7)	0.75:0.25	0.09
172*	Tobramycin (#9) + Desloratadine (#7)	0.5:0.5	0.04
173	Tobramycin (#9) + Desloratadine (#7)	0.25:0.75	0.02
174	Tetracycline hydrochloride (#15) + Dichlorvos (#19)	0.078:0.922	2.94
175*	Tetracycline hydrochloride (#15) + Dichlorvos (#19)	0.033:0.967	2.58
176	Tetracycline hydrochloride (#15) + Dichlorvos (#19)	0.017:0.983	2.37
177*	Tetracycline hydrochloride (#15) + Dichlorvos (#19)	0.008:0.992	2.04
178	Tetracycline hydrochloride (#15) +	0.003:0.997	1.86

	Dichlorvos (#19)		
179	Tetracycline hydrochloride (#15) + Trichlorfon (#18)	0.054:0.946	3.47
180	Tetracycline hydrochloride (#15) + Trichlorfon (#18)	0.022:0.978	3.35
181	Tetracycline hydrochloride (#15) + Trichlorfon (#18)	0.011:0.989	3.3
182	Tetracycline hydrochloride (#15) + Trichlorfon (#18)	0.006:0.994	3.26
183	Tetracycline hydrochloride (#15) + Trichlorfon (#18)	0.002:0.998	3.24
184	Tetracycline hydrochloride (#15) + Linuron (#20)	0.814:0.186	4.16
185*	Tetracycline hydrochloride (#15) + Linuron (#20)	0.636:0.364	4.26
186	Tetracycline hydrochloride (#15) + Linuron (#20)	0.467:0.533	4.25
187	Tetracycline hydrochloride (#15) + Linuron (#20)	0.304:0.696	4.24
188	Tetracycline hydrochloride (#15) + Linuron (#20)	0.149:0.851	4.23
189	Chloramphenicol (#16) + Dichlorvos (#19)	0.549:0.451	3.15
190*	Chloramphenicol (#16) + Dichlorvos (#19)	0.328:0.672	3.21
191	Chloramphenicol (#16) + Dichlorvos (#19)	0.196:0.804	3.25
192	Chloramphenicol (#16) + Dichlorvos (#19)	0.109:0.891	3.28
193	Chloramphenicol (#16) + Dichlorvos (#19)	0.046:0.954	3.28
194	Chloramphenicol (#16) + Linuron (#20)	0.984:0.016	3.01
195	Chloramphenicol (#16) + Linuron (#20)	0.962:0.038	3.11
196*	Chloramphenicol (#16) + Linuron (#20)	0.927:0.073	3.22
197	Chloramphenicol (#16) + Linuron (#20)	0.861:0.136	3.37
198	Chloramphenicol (#16) + Linuron (#20)	0.717:0.283	3.61
199	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17)	0.033:0.483:0.484	3.34
200	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17)	0.037:0.367:0.596	3.35
201	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17)	0.038:0.277:0.684	3.06
202	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine	0.041:0.166:0.793	3.12

	(#17)		
203	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17)	0.12:0.279:0.601	3.08
204*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Linuron (#20)	0.06:0.871:0.069	3.62
205*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Linuron (#20)	0.083:0.833:0.084	3.32
206	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Linuron (#20)	0.11:0.795:0.095	3.5
207	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Linuron (#20)	0.173:0.71:0.117	3.21
208*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Linuron (#20)	0.253:0.589:0.158	3.77
209	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Trichlorfon (#18)	0.01:0.141:0.849	2.95
210*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Trichlorfon (#18)	0.013:0.133:0.854	2.57
211	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Trichlorfon (#18)	0.017:0.124:0.858	2.59
212	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Trichlorfon (#18)	0.027:0.112:0.861	2.78
213	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Trichlorfon (#18)	0.043:0.1:0.857	2.89
214*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17) + Linuron (#20)	0.032:0.465:0.466:0.037	3.3
215*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17) + Linuron (#20)	0.035:0.354:0.575:0.036	3.28
216*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17) + Linuron (#20)	0.037:0.268:0.663:0.032	3.19
217*	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17) + Linuron (#20)	0.039:0.162:0.772:0.027	3.2
218	Tetracycline hydrochloride (#15) + Chloramphenicol (#16) + Metribuzine (#17) + Linuron (#20)	0.039:0.091:0.846:0.024	3.32

*Test set mixtures

Table 3.8. Mixture components along with their class (pharmaceuticals/pesticides)

Sl. No.	Name of Component	Class
#1	Azithromycin	Pharmaceutical (Antibiotic)
#2	Erythromycin	Pharmaceutical (Antibiotic)
#3	Carbamazepine	Pharmaceutical (Anti-convulsant)
#4	Oxytetracycline	Pharmaceutical (Antibiotic)
#5	Dexamethasone	Pharmaceutical (Corticosteroid)
#6	Diclofenac	Pharmaceutical (NSAID)
#7	Desloratadine	Pharmaceutical (Antihistamine)
#8	Imatinib	Pharmaceutical (Tyrosine kinase inhibitor)
#9	Tobramycin	Pharmaceutical (Antibiotic)
#10	Omeprazole	Pharmaceutical (Proton pump inhibitor)
#11	Alachlor	Pesticide
#12	Isoproturon	Pesticide
#13	Diuron	Pesticide
#14	Chlorfenvinphos	Pesticide
#15	Tetracycline hydrochloride	Pharmaceutical (Antibiotic)
#16	Chloramphenicol	Pharmaceutical (Antibiotic)
#17	Metribuzin	Pesticide
#18	Trichlorfon	Pesticide
#19	Dichlorvos	Pesticide
#20	Linuron	Pesticide

NSAID: Non-steroidal anti-inflammatory drug

3.3.2. Mixture descriptor calculation

Todeschini and Consonni define molecular descriptor as "*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*" (Consonni and Todeschini, 2010). 2D descriptors can efficiently model the structural features of chemical components with toxicity (biological activity in general) and have been used in classical QSARs as well as mixture QSARs. Thus, only 2D structural descriptors have been used among various other classes of descriptors (0D to 7D) in this study. The 2D structures of mixture components have been drawn in MarvinSketch software (version 14.10.24) with explicit hydrogen addition and proper aromatization and used for descriptor calculation without further energy minimization and conformational analysis

(exclusive for 3D and above order of descriptor calculation). Among various classes of 2D descriptors, we have computed constitutional indices, ring descriptors, functional group counts, connectivity indices, atom-centred fragments, atom type E-state indices, 2D atom pairs, molecular properties, and extended topochemical atom (ETA) indices with the alvaDesc software (Mauri, 2020). These particular classes of descriptors have been used for their easy interpretability.

It has been observed to use the software-derived descriptors of individual chemicals directly for the classical QSARs (Ghanem et al., 2017; Mukherjee et al., 2022). But in the case of mixture's toxicity modeling, one has to consider the influence of at least 2 different chemicals (binary mixture) simultaneously during descriptor calculation. The theory of additivity is mostly employed during the calculation of mixture descriptors from the descriptors of component chemicals, and it has been successfully used in various instances (Chatterjee and Roy, 2021; Hoover et al., 2019; T. Wang et al., 2018). Therefore, we have also applied the theory of additivity to get the corresponding mixture descriptors (weighted descriptors) in this study. The descriptors of mixture components were obtained from the alvaDesc software (251 descriptors of the individual component were obtained after internal pretreatment of alvaDesc software) (Mauri, 2020). The mixture descriptors have been obtained by employing an in-house tool 'Mixture Descriptor Calculator v1.0' (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) from the 251 descriptors of each mixture components. If a mixture is composed of *m* chemicals (1 to m) with molar concentration ratios M_1, M_2, …., M_m ($M_1+M_2+\ldots+M_m = 1$), and the descriptors of each chemical are D^1 (D^1_1, D^1_2, …., D^1_m), D^2 (D^2_1, D^2_2, …., D^2_m), D^n (D^n_1, D^n_2, …., D^n_m) respectively, then according to the theory of additivity, the corresponding weighted descriptors (D_{Mix}) can be calculated by the following equations 3.6 to 3.8-

$$D^1_{Mix} = (D^1_1 \times M_1) + (D^1_2 \times M_2) + \cdots + (D^1_m \times M_m) \qquad (3.6)$$

$$D^2_{Mix} = (D^2_1 \times M_1) + (D^2_2 \times M_2) + \cdots + (D^2_n \times M_m) \qquad (3.7)$$

$$D^n_{Mix} = (D^n_1 \times M_1) + (D^n_2 \times M_2) + \cdots + (D^n_m \times M_m) \qquad (3.8)$$

After calculating the descriptors, a pretreatment operation was carried out by using the 'DataPreTreatmentGUI' tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) where the software removes the intercorrelated descriptors (intercorrelation cutoff 0.95), descriptors with the least variability among compounds (variability cutoff 0.0001), descriptors with constant values or near-constant values, and descriptors with at least one missing value. Finally, a pool of 247 mixture descriptors was obtained for modeling.

3.3.3. Dataset division

Dataset division is an integral part of QSAR modeling, and it is necessary for developing the model and proper validation. As per the standard protocol, the whole dataset should be divided into training and test sets; the training set is used for model building and internal validation whereas the test set is exclusively used for rigorous external validation. There is no fixed ratio of the number of training and test compounds; however, 25-30% of data is kept in the test set as common practice. The algorithms of Kennard-Stone selection and Euclidean distance-based division check the structural variability (independent variables: X) among compounds/mixtures, whereas sorted activity-based division checks the variability of toxicity/activity/property (dependent variable: Y) among compounds/mixtures. From these 3

unique approaches, total nine training-test pairs were obtained (4 seeds of Euclidean distance-based divisions, 4 seeds of sorted activity-based divisions, and 1 Kennard-Stone selection) among which the Kennard-Stone selection was a 70:30 training-test division and the rest pairs were 75:25 training-test divisions.

The dataset was divided into training and test sets for model development and external validation respectively, by applying Kennard-Stone selection, Euclidean distance-based division, and sorted activity-based division with the freely available in-house tool 'Dataset Division GUI 1.2' (available at: http://teqip.jdvu.ac.in/QSAR_Tools) (Roy et al., 2015). Preliminary regression models were generated using stepwise selection in the Minitab software ("Minitab Statistical Software," 2022) and based on the internal correlation coefficient (Q^2_{LOO}) and external correlation coefficient Q^2_{F1} (Q^2_{LOO} and Q^2_{F1} values of 9 preliminary model obtained from 9 divisions have been incorporated in the **Table 3.9**, only a Euclidean distance-based division (4^{th} seed) was selected (149 training data, 49 test data with 247 descriptors) for further processing. The final 4-4 Euclidean distance-based division has been displayed in **Table 3.7.**

Table 3.9. Internal and External validation metric values of preliminary MLR models obtained from 9 classes of divisions

Division	Q^2_{LOO}	Q^2_{F1}	Number of descriptors
4-1 activity-based division	0.94	0.92	10
4-1 Euclidean distance-based division	0.91	0.69	10
4-2 activity-based division	0.93	0.90	8
4-2 Euclidean distance-based division	0.94	0.89	10
4-3 activity-based division	0.89	0.87	9
4-3 Euclidean distance-based division	0.90	0.91	11
4-4 activity-based division	0.86	0.93	8
4-4 Euclidean distance-based division	**0.93**	**0.95**	**10**
70:30 Kennard-Stone division	0.89	0.92	10

3.3.4. Feature selection and model development

In this communication, the data matrix of the final Euclidean distance-based division (4^{th} seed) was taken forward to further steps for model development which consists of 247 descriptors (X variables) and the pEC50 values (Y variable). The pool of 247 descriptors was reduced to 40 by using both stepwise selection (Wagner and Shimshak, 2007) and genetic algorithm (Leardi, 1996; Rogers and Hopfinger, 1994). Genetic algorithm is an efficient feature selection technique compared to other tools and was first used by Rogers and Hopfingers. The simultaneous exploration of many possible solutions in the various regions of descriptor space makes it more powerful. It is a two-steps process where N number of descriptor population (parents) are selected from the whole pool first, and the fitness of each combination is determined. In the second step, crossover and mutation are taken place to get

a new generation of descriptor combinations from the parents based on their scaled fitness scores. It is a process of genetic movement (movement of features) from parents to children; thus selection, crossover, and mutation are repeated until all the parent descriptors are replaced with the new one. The fitness scores of the new combinations are again evaluated and the process continues until 80% of the new combinations attain the target score. We have performed GA using a freely available tool 'GeneticAlgorithm_v4.1' (available at http://teqip.jdvu.ac.in/QSAR_Tools) with the training set of mixtures. The stepwise selections have been carried out with the F-to-enter 4 and F-to-remove 3.9 setting in multiple iterations using Minitab software ("Minitab Statistical Software," 2022). This F feature is the stepping criterion which is calculated by doing square of the t value of the incoming descriptors. For a more detailed explanation of the stepwise selection strategy, one can refer (Wagner and Shimshak, 2007). The train-test pair with the reduced pool of 40 descriptors along with the dependent variable pEC50 were then subjected to the Best subset selection algorithm (Roy et al., 2015) for obtaining various descriptor combinations (subsets), and the best combination of features was selected (7 descriptors) based on the internal and external validation parameters such as Q^2_{LOO} and Q^2_{F1} respectively (Roy et al., 2015). It is another statistical tool specifically used for variable selection during QSAR model development. It is simply based on the multiple linear regression (MLR) algorithm and explores all possible combinations of the descriptor from a provided pool. This is computationally more exhaustive and takes more time than a genetic algorithm; thus a descriptor pool reduction is carried out before BSS. In this study, the final selection of descriptors was performed with the best subset selection using a freely available in-house tool 'BestSubsetSelectionModified_v2.1' (available at http://teqip.jdvu.ac.in/QSAR_Tools). This tool provides various validation parameters such as regression coefficient (R^2), cross-validated correlation coefficient (Q^2_{LOO}), external correlation coefficient (Q^2_{F1}), and other error-based metrics (MAE) etc. along with the descriptors, which are used to identify the best combination of descriptors among all. In this tool, additional filters like intercorrelation cutoff, determination coefficient cutoff, or cross-validated variance cutoff can also be applied to save computational time. Finally, partial least squares (PLS) regression (Cramer, 1993; Wold et al., 2001) was carried out to obtain the desired model with 7 selected descriptors of the training set, and it was rigorously validated with the test set of mixtures. We have also constructed various PLS plots such as coefficient plot, variable importance plot, score plot, loading plot, Y-randomization plot, and applicability domain (DModX) plot using the SIMCA-P software (Wu et al., 2010). PLS regression is the generalization of multiple linear regression and is used to model strongly correlated, collinear, and noisy data with numerous descriptors. It can also handle numerous Y variables (biological responses/molecular properties) simultaneously. The problem of overfitting with a large number of descriptors (close to the number of data points) in MLR can be successfully resolved with PLS. In this method, latent variables (LVs) such as X scores (T) and Y scores (U) are calculated. Y scores are calculated from the X scores and response variables and then used for the final prediction. The LVs account for the maximum variation of the descriptor matrix to calculate the response. In this research, we have performed the PLS using the freely available tool 'PLS_Single Y_version 1.0' (available at: http://teqip.jdvu.ac.in/QSAR_Tools) and reported the resultant model with the number of LVs (number of LVs must be less than the number of descriptors).

To avoid the chances of overestimating validation metrics, we have performed 3 cross-validation tests namely mixtures out (Oprisiu et al., 2012), compounds out (Oprisiu et al., 2012) and everything out cross-validation (Muratov et al., 2014) in this study. To perform these three cross-validations, we shuffled the mixture data points among training and test

sets. These new training and test sets for the cross-validations have been stored in an openly accessible online repository that is available at https://github.com/mchatt1012/PhD-Thesis_QSAR-Modeling-of-Toxicity-of-Chemical-Mixtures.git. The new training-test pairs were then subjected to the 'PLS_Single Y_version 1.0' tool (available at http://teqip.jdvu.ac.in/QSAR_Tools) without changing the descriptor combination and the number of LVs. The internal and external validation metrics (R^2, Q^2_{LOO}, and Q^2_{F1}) were computed and reported in each case to check the quality. According to Muratov et al., everything out cross-validation penalizes all possible interferences among test and training set compounds and thus provides realistic metric values of the model (Muratov et al., 2014). The complete process of model development along with the data division and descriptor calculation have been showcased schematically in **Figure 3.5**.

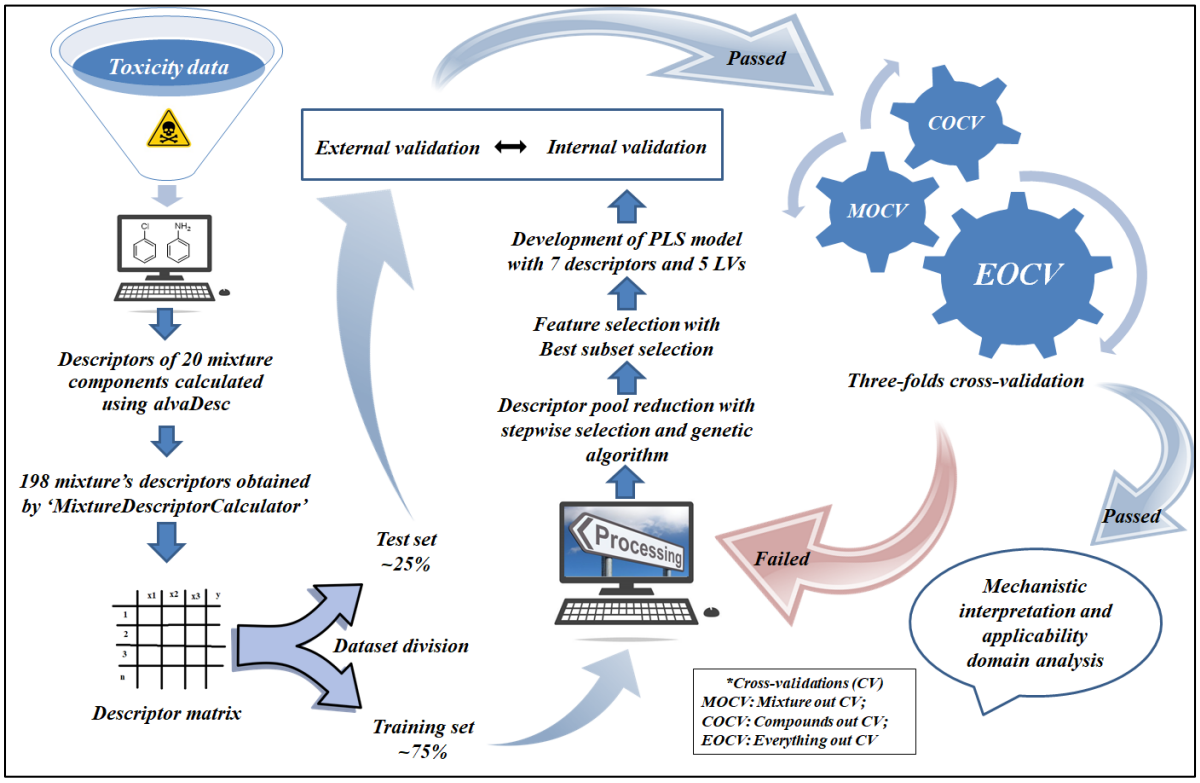


Figure 3.5. Workflow of QSAR model development

3.3.5. Chemical read-across

Read-across is a similarity-based *in silico* method which predicts the toxicity (biological activity in general) of the untested chemicals from the known toxicity data of chemical analogues from the same chemical category (Raies and Bajic, 2016). It is one of the most important alternative approaches in chemical hazard assessment aimed to fill data gaps. Recently, we have developed and reported a similarity-based machine learning algorithm 'Read-across v4.1' (available at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) that predicts the toxicity of unknown substances (target chemicals) from the structural information of known analogues (source chemicals) (Banerjee and Roy, 2022; Chatterjee et al., 2022). Three different similarity estimation strategies such as Euclidean distance-based similarity, Gaussian kernel function similarity (GK), and Laplacian kernel function similarity (LK) have been utilized in this machine learning algorithm. The success of this method depends on the relevance of provided structural features (descriptors) towards toxicity based on which the similarity is calculated (Schultz et al., 2015). The software selects

up to 10 most similar source compounds for each target chemical and predicts the responses by taking the weighted average of toxicity values of them. In the current study, we have applied chemical read-across for mixture toxicity prediction for the 1st time. It has been also used as an additional tool to cross-check the model quality as well as the relevance of the modeled descriptors.

There are two hyperparameters namely sigma (for GK) and gamma (for LK) used in this tool, which should be optimized first for obtaining the best prediction. To optimize the hyperparameters, we randomly divided the model training set into 70% sub-training set and 30 % calibration set and used them as source and target mixture respectively, by changing the values of sigma and gamma from 0.25 to 2 at 0.25 intervals. The optimized setting has been found based on the external validation metric Q^2_{F1} (provided by the tool itself). Finally, we subjected the training and test mixtures to 'Read-across v4.1' with gradual changes in the number of close training compounds from 10 to 2 to get the best prediction. One can also apply the distance (1 to 0) and similarity thresholds (0 to 1) instead of providing the number of close training compounds (10 to 2); however, both the determinants (thresholds of distance & similarity, and the number of close training compounds) cannot be used simultaneously. The workflow for read-across predictions has been presented in **Figure 3.6.**

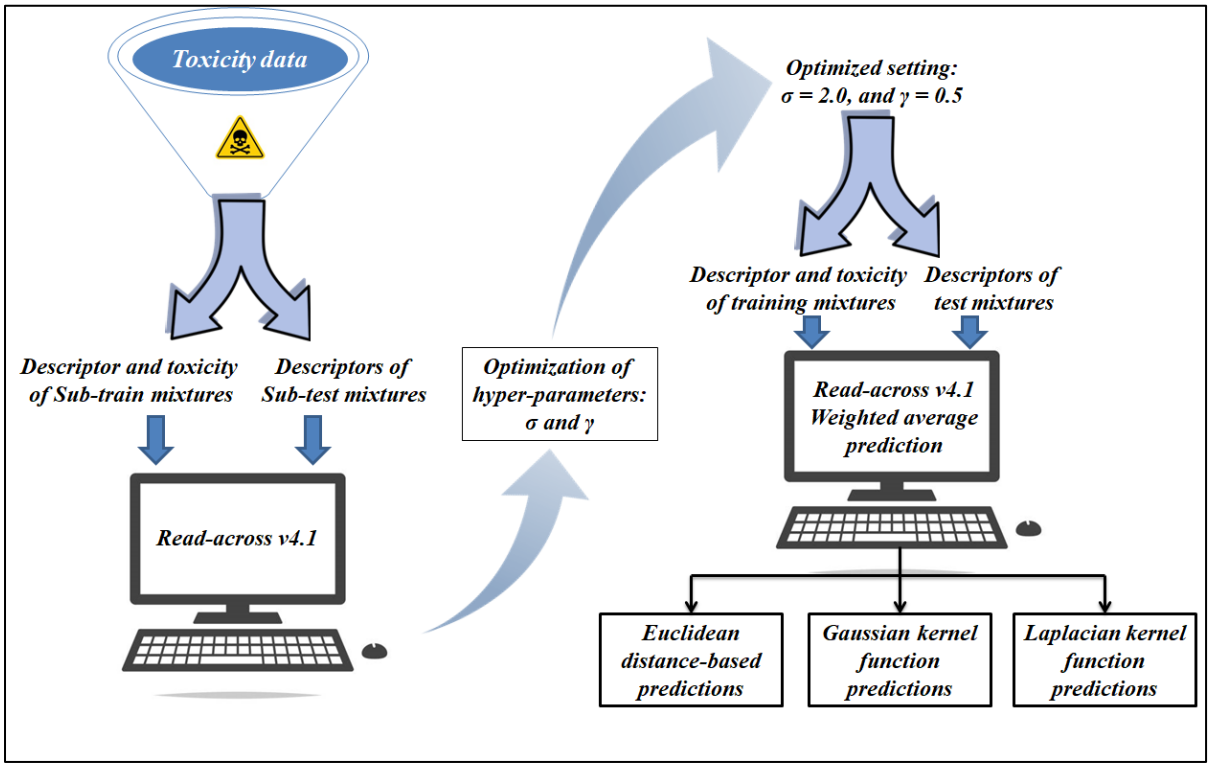


Figure 3.6. Workflow for chemical read-across predictions.

3.3.6. Statistical validation and Applicability domain analysis

Statistical validation metrics are extensively used to evaluate the quality and predictivity of the developed QSAR models, based on which the models are accepted or rejected. The QSAR modellers must perform proper statistical validation during modeling to comply with the OECD guideline 4. Two types of metrics are calculated two comply with the guideline. The internal validation metrics such as regression coefficient (R^2), cross-validated (*leave-one-out*) correlation coefficient (Q^2_{LOO}), root mean square error of calibration (RMSEc), mean absolute error (MAE) etc. are computed with the training data points. R^2, Q^2_{LOO},

RMSEc and MAE are the measure of goodness-of-fit, robustness, and errors of measurement respectively. There are high chances of overestimation of R^2 where the number of descriptors is close to the number of data points. To penalize the overestimated R^2, adjusted R^2 is calculated which considers the number of descriptors also along with the number of data points. The value of R^2, and R^2_{adj}, and Q^2_{LOO} should be near 1, and more deviation of the value from 1 suggests inferiority of the model. The optimum value of Q^2_{LOO} >0.5 is mostly used and suggested by OECD[49]. The error metrics (RMSEc, MAE) should be near zero. The external validation is performed to check the predictivity of the developed model and the test set compounds are exclusively used for this purpose. External correlation coefficient such as Q^2_{F1} and Q^2_{F2} are the classic markers of predictivity and their values should be near 1. In common practice, the optimum value of these two metrics for the selection of model should be > 0.6. Error metrics like root mean square error of prediction (RMSEp) is also calculated to check the error of predicted results. Besides these classical metrics, r_m^2 metrics (*average* $r_m^{2}{}_{(LOO)}$, $\Delta r_m^{2}{}_{(LOO)}$, *average* $r_m^{2}{}_{(Test)}$, $\Delta r_m^{2}{}_{(Test)}$) are also calculated for internal validation as well as external validation. The optimum value of average $r_m^2{}_{(LOO)}$ and average $r_m^2{}_{(Test)}$ is >0.5, whereas the expected value for $\Delta r_m^2{}_{(LOO)}$ and $\Delta r_m^2{}_{(Test)}$ should be <0.2.

3.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

3.4.1. Toxicity data

For developing any *in silico* model, the availability of good quality response (activity/toxicity/property) data is of utmost importance. The number of data points is also a major concern while developing a statistically sound model. In 2019, Carnesecchi et al. reported several ecotoxicologically important mixtures with corresponding acute contact/oral toxicity data against honey bees in a meta-analysis (Carnesecchi et al., 2019). The same research group further clarified the experimental data and reported in a separate modeling study in 2020 (Carnesecchi et al., 2020). We have collected the toxicity data of 123 organic chemical mixtures (SM 1 to SM 123) from the latter publication of Carnesecchi et al. (Carnesecchi et al., 2020) and have used it in our modeling (can be accessed from the open online repository available at https://github.com/mchatt1012/PhD-Thesis_QSAR-Modeling-of-Toxicity-of-Chemical-Mixtures.git). These mixtures were mostly composed of two (binary mixtures) or more (multicomponent mixtures on very few occurrences) plant protection products (pesticides, insecticides, acaricides, defoliants, plant growth retardants, etc.), that have been tabulated in **Table 3.10.**

Table 3.10. Mixture components (Carnesecchi et al., 2020)

Sl. No.	Compounds	Type
C1	Acetamiprid	Insecticide/Acaricide
C2	Amitraz	Insecticide/Acaricide
C3	Boscalid	Fungicide
C4	Chlorothalonil	Fungicide
C5	Coumaphos	Insecticide/Acaricide
C6	Cyfluthrin	Insecticide/Acaricide
C7	Cyhalothrin	Pesticide

C8	Diethyl but-2-enedioate	Pesticide
C9	Epoxiconazole	Fungicide
C10	Fenbuconazole	Fungicide
C11	Fenpyroximate	Insecticide/Acaricide
C12	Fluvalinate	Insecticide/Acaricide
C13	Imidacloprid	Insecticide/Acaricide
C14	Metconazole	Fungicide
C15	Myclobutanil	Fungicide
C16	Oxalic acid	Disinfectant/Pesticide
C17	Phenothrin	Insecticide/Acaricide
C18	Piperonyl butoxide	Pesticide synergist
C19	Prochloraz	Fungicide
C20	Propiconazole	Fungicide
C21	Pyraclostrobin	Fungicide
C22	Thiacloprid	Insecticide/Acaricide
C23	Thymol	Pesticide
C24	Triadimefon	Fungicide
C25	Tribufos/Butifos	Defoliant
C26	Triflumizole	Fungicide
C27	Uniconazole	Plant growth retardant

*C: Component

Carnesecchi and coworkers (Carnesecchi et al., 2020) reported the acute contact toxicity data (pLD50 values) as μg active substance/bee unit (the endpoint of this research), which represents the concentration of the mixtures that is enough to produce a lethal effect on 50% of the total population of honey bees. There is no data on mass or concentration ratio in the source article; therefore, we have used a mass ratio 50:50 as the default value for binary mixtures as an approximation (0.33:0.67 mass ratio has been taken for mixtures SM116, SM119, and SM 122 where 1 component is present twice than the other) in this research. Among 123 data points, 24 different sets of duplicates (69 duplicate entries) have been found. Among these 24 sets, 11 sets (24 entries) have response values in a very close range (variance <0.02; standard deviation <0.15). Thus, the response means of these 11 sets of duplicate data have been taken as the observed response and used in modeling. Therefore, a total of 65 mixtures (initial 54 mixtures + 11 duplicate mixtures) have been employed as the modeling data set. The rest 13 sets (44 entries) of mixtures have response values with higher ranges (variance >0.05 and/or standard deviation > 0.2) and thus have been used as true external set mixtures. These true external set mixtures have been employed for true external validation in this research. The modeling set mixtures and true external set mixtures have been tabulated in **Table 3.11** and **Table 3.12** respectively.

Table 3.11. Modeling set mixtures

Mix. No.	Source Sl. No.	Component A	Mass Ratio	Component B	Mass Ratio	Observed toxicity
1	SM1	C1	0.5	C18	0.5	-0.068
2	SM2	C1	0.5	C25	0.5	-0.378
3	SM3	C1	0.5	C8	0.5	-0.841
4*	SM4	C1	0.5	C26	0.5	1.523
5	SM5	C1	0.5	C20	0.5	1.155
6*	SM6	C1	0.5	C24	0.5	1.097
7	SM7	C1	0.5	C9	0.5	0.301
8	SM8	C1	0.5	C27	0.5	-0.049
9	SM15	C1	0.5	C10	0.5	-1.155
10	SM12	C22	0.5	C1	0.5	1.046
11	SM31	C12	0.5	C2	0.5	-0.573
12	SM68	C2	0.5	C16	0.5	-1.164
13	SM69	C2	0.5	C21	0.5	-0.215
14	SM70	C2	0.5	C3	0.5	-0.683
15	SM71	C2	0.5	C4	0.5	-0.524
16	SM72	C2	0.5	C19	0.5	-0.394
17	SM73	C2	0.5	C8	0.5	-0.362
18	SM74	C2	0.5	C25	0.5	-0.336
19*	SM75	C2	0.5	C18	0.5	-0.382
20	SM115	C17	0.5	C2	0.5	0.444
21*	SM35	C12	0.5	C3	0.5	-1.064
22	SM46	C5	0.5	C3	0.5	-1.354
23*	SM58	C11	0.5	C3	0.5	-0.751
24	SM82	C23	0.5	C3	0.5	-1.673
25*	SM36	C12	0.5	C4	0.5	-0.86
26*	SM47	C5	0.5	C4	0.5	-1.220
27	SM59	C11	0.5	C4	0.5	-0.807
28	SM83	C23	0.5	C4	0.5	-1.474
29	SM17	C5	0.5	C8	0.5	-1.299
30	SM18	C5	0.5	C25	0.5	-0.863
31	SM19	C5	0.5	C18	0.5	-0.702
32	SM44	C5	0.5	C23	0.5	-1.364
33	SM45	C5	0.5	C16	0.5	-1.167
34*	SM48	C5	0.5	C19	0.5	0.357
35	SM116	C17	0.33	C5*2	0.67	0.854
36	SM106	C6	0.5	C8	0.5	1.222
37	SM107	C6	0.5	C25	0.5	1.523

38*	SM108	C6	0.5	C18	0.5	2.699
39	SM109	C7	0.5	C8	0.5	1.398
40	SM110	C7	0.5	C25	0.5	1.523
41	SM111	C7	0.5	C18	0.5	2.886
42	SM38	C12	0.5	C8	0.5	-0.917
43*	SM61	C11	0.5	C8	0.5	-0.641
44	SM85	C23	0.5	C8	0.5	-1.806
45	SM16	C13	0.5	C10	0.5	0.523
46	SM56	C11	0.5	C16	0.5	-0.176
47*	SM57	C11	0.5	C21	0.5	-0.32
48	SM60	C11	0.5	C19	0.5	0.602
49*	SM62	C11	0.5	C25	0.5	-0.100
50*	SM63	C11	0.5	C18	0.5	0.569
51	SM32	C12	0.5	C23	0.5	-1.009
52	SM33	C12	0.5	C16	0.5	-0.848
53	SM39	C12	0.5	C25	0.5	-0.292
54	SM34	C12	0.5	C21	0.5	-0.646
55	SM9	C13	0.5	C18	0.5	1.979
56	SM10	C13	0.5	C26	0.5	2.013
57*	SM11	C13	0.5	C20	0.5	1.921
58	SM120	C17	0.5	C18	0.5	0.495
59	SM13	C22	0.5	C26	0.5	2.000
60	SM14	C22	0.5	C20	0.5	1.523
61	SM80	C23	0.5	C16	0.5	-1.487
62*	SM81	C23	0.5	C21	0.5	-1.450
63	SM84	C23	0.5	C19	0.5	-1.591
64*	SM86	C23	0.5	C25	0.5	-1.545
65	SM87	C23	0.5	C18	0.5	-1.511

*Test set mixtures of modeling division; [SM: Source mixture; C: Component; Mass ratio: 0.5 has been taken as the default mass ratio in most of the cases as there is no information of mass of components in the source article].

Table 3.12. Modeling external set mixtures

Sl. No.	Source Mix. No.	Component A	Mass Ratio	Component B	Mass Ratio	Observed toxicity
66	SM43	C5	0.5	**C2**	0.5	-0.964
66	SM 65	**C2**	0.5	C5	0.5	-0.436
67	SM 54	C11	0.5	**C2**	0.5	-0.255

67	SM 66	**C2**	0.5	C11	0.5	-0.66
68	SM 67	**C2**	0.5	C23	0.5	-0.592
68	SM 79	C23	0.5	**C2**	0.5	-1.635
69	SM 20	**C5**	0.5	C12	0.5	-1.322
69	SM 21	**C5**	0.5	C12	0.5	-1.106
69	SM 22	**C5**	0.5	C12	0.5	-0.782
69	SM 23	**C5**	0.5	C12	0.5	-0.782
69	SM 41	**C5**	0.5	C12	0.5	-0.813
69	SM 24	C12	0.5	**C5**	0.5	-0.788
69	SM 25	C12	0.5	**C5**	0.5	-0.517
69	SM 26	C12	0.5	**C5**	0.5	-0.428
69	SM 27	C12	0.5	**C5**	0.5	-0.185
69	SM 28	C12	0.5	**C5**	0.5	0.678
69	SM 29	C12	0.5	**C5**	0.5	0.108
70	SM 42	**C5**	0.5	C11	0.5	-0.615
70	SM 53	C11	0.5	**C5**	0.5	-0.307
71	SM 91	C12	0.5	**C10**	0.5	-1.021
71	SM 97	C12	0.5	**C10**	0.5	-0.793
71	SM 103	C12	0.5	**C10**	0.5	0.26
72	SM 52	**C11**	0.5	C12	0.5	-0.744
72	SM 30	C12	0.5	**C11**	0.5	-0.38
73	SM 55	**C11**	0.5	C23	0.5	-0.567
73	SM 78	C23	0.5	**C11**	0.5	-1.543

74	SM 40	**C12**	0.5	C18	0.5	2
74	SM 88	**C12**	0.5	C18	0.5	-1.041
74	SM 94	**C12**	0.5	C18	0.5	-0.928
74	SM 100	**C12**	0.5	C18	0.5	-0.188
74	SM 114	**C12**	0.5	C18	0.5	2
75	SM 37	**C12**	0.5	C19	0.5	2
75	SM 89	**C12**	0.5	C19	0.5	-1.045
75	SM 95	**C12**	0.5	C19	0.5	-0.667
75	SM 101	**C12**	0.5	C19	0.5	0.658
76	SM 90	**C12**	0.5	C20	0.5	-1.143
76	SM 96	**C12**	0.5	C20	0.5	-0.776
76	SM 102	**C12**	0.5	C20	0.5	0.131
77	SM 92	**C12**	0.5	C14	0.5	-1.104
77	SM 98	**C12**	0.5	C14	0.5	-0.717
77	SM 104	**C12**	0.5	C14	0.5	0.347
78	SM 93	**C12**	0.5	C15	0.5	-1.111
78	SM 99	**C12**	0.5	C15	0.5	-0.574
78	SM 105	**C12**	0.5	C15	0.5	0.921

*SM: Source mixture; C: Component; Mass ratio: 0.5 has been taken as the default mass ratio in most of the cases as there is no information of mass of components in the source article.

3.4.2. Mixture descriptors

The calculation of mixture descriptors is not as simple as the descriptors of individual chemicals. The theory of additivity is mostly used to calculate mixture descriptors where a linear combination of the molecular descriptors of mixture components weighted by their respective masses are employed to consider the effects of each component (Chatterjee and Roy, 2022b; Qin et al., 2018). But the bioactive mixture components may interact with each other and develop synergistic or antagonistic effects. If we thoroughly review the scientific literature, then several mixing rules apart from the theory of additivity can be found which are used to consider the interactive effects of mixture components (Altenburger et al., 2003; Chatterjee and Roy, 2022c; Gaudin et al., 2015; Hoover et al., 2019; Qin et al., 2018; Sobati

et al., 2016). In this study, we have employed three different mixing rules namely the linear combination of molecular contributions, the squared molecular contributions, and the norm of molecular contributions, to optimize the descriptor calculation. The mathematical forms of the three mixing rules have been given in the following equations 3.9, 3.10, and 3.11 respectively (Saldana et al., 2013)-

$$Mixture\ Descriptor\ 1 = \sum(p_i \times x_i) \qquad (3.9)$$

$$Mixure\ Descriptor\ 2 = (\sum(p_i \times x_i))^2 \qquad (3.10)$$

$$Mixture\ Descriptor\ 3 = \sqrt{\sum(p_i \times x_i)^2} \qquad (3.11)$$

where p_i represents the mass fraction i^{th} component, x_i represents the descriptor of i^{th} component.

Several classes of molecular descriptors can be obtained for modeling. The most common classification of descriptors is based on dimensionality; 0-dimensional to 7-dimensional descriptors can be computed among which 2D and 3D descriptors are mostly used. In the present study, up to 2D descriptors are employed for the ease of computation and interpretation of descriptors, and reproducibility and transferability of the developed models. 2D molecular structures, obtained from Marvin sketch software (ChemAxon, 2018), have been used for descriptor calculation without any energy minimization and conformational analysis (prerequisites of 3D descriptor calculation). "AlvaDesc" software (Mauri, 2020) has been utilized to calculate 9 types of 0D-2D descriptors (molecular properties, constitutional indices, ring descriptors, connectivity index, functional group count, 2D-atom pairs, atom type-E state indices, atom-centred fragments, and electro-topochemical atom indices) in this study. After obtaining the descriptors of individual components, the mixture descriptors have been computed with the above-mentioned equations 3.6 to 3.8 by using our in-house tool "Mixture_Desc_Calc v1.0".

3.4.3. Dataset division and feature selection for QSAR and subsequent modeling

Dataset division is one of the integral parts of any statistical modeling (especially QSARs) when the external data points are not available in sufficient numbers or are practically unavailable. In that case, the modeling set is divided into two parts namely training and test sets, that are used respectively for model development and external validation. There is no fixed ratio of the number of data points in training and test sets; although there must be a justification for the number of data points selected in both sets. By keeping in mind the possibility of the bias-variance dilemma (Almaliki, 2018), the QSAR experts suggest a 70:30 to 80:20 division of training-test data points which have been used in many previous studies (Chatterjee and Roy, 2021, 2022d, 2022a; T. Wang et al., 2018). In this research, we have used 75% of the data for model development and internal validation (training set), and 25% of data for external validation (test set). A dataset can be divided either randomly or following some protocols based on the descriptor values (X-based division) or exerted response values (Y-based division) (Roy et al., 2015). We have utilized both X-based (Euclidean distance-based division) and Y-based divisions (sorted activity-based division) for each class of mixture descriptors (Mixture Descriptor 1, Mixture Descriptor 2, and Mixture Descriptor 3) with the help of our in-house "DatasetDivision GUI 1.2" tool. Eight types of divisions have been obtained (4 seeds of Euclidean distance-based divisions and 4 seeds of sorted activity-based divisions) from each mixture descriptor class, that has been used for the selection of the proper division set. Preliminary stepwise multiple linear regression (SMLR)

models have been developed from all the training sets by using MINITAB software with F to enter = 4 and F to remove = 3.9 criteria in multiple iterations. For a standard MLR equation, each descriptor should be significant in 95% confidence level ($p < 0.05$) which corresponds to an approximate t-value more than 2 ($F = t^2 = (2)^2 = 4$). Thus, we used F to enter = 4 and F to remove = 3.9. The best divisions from each class of mixture descriptor have been selected based on the proper balancing between results of both internal and external validation results of the preliminary SMLR models in terms of cross-validated correlation coefficient (Q^2_{LOO}) and external correlation coefficient (Q^2_{F1}). The quality metrics R^2, Q^2_{LOO}, and Q^2_{F1} of all preliminary models have been tabulated in the **Table 3.13** as the indicators of goodness of fit, robustness, and external predictability respectively.

Table 3.13. Internal and external validation metrics of preliminary SMLR models

Mixture descriptor 1: Linear combination of molecular contribution			
Division	R^2	Q^2_{LOO}	Q^2_{F1}
4-1 ABD	0.87	0.81	0.69
4-1 DBD	0.91	0.86	0.60
4-2 ABD	0.90	0.87	0.72
4-2 DBD	0.89	0.84	0.56
4-3 ABD	0.92	0.89	0.43
4-3 DBD	0.88	0.84	0.67
4-4 ABD*	0.85	0.79	**0.82**
4-4 DBD	0.84	0.78	0.67
Mixture descriptor 2: Square molecular contribution			
Division	R^2	Q^2_{LOO}	Q^2_{F1}
4-1 ABD	0.87	0.83	0.53
4-1 DBD	0.89	0.82	0.44
4-2 ABD	0.89	0.86	0.68
4-2 DBD	0.87	0.82	0.64
4-3 ABD	0.90	0.88	0.18
4-3 DBD	0.89	0.85	0.59
4-4 ABD*	0.86	0.78	**0.91**
4-4 DBD	0.84	0.79	0.62
Mixture descriptor 3: Norm of molecular contribution			
Division	R^2	Q^2_{LOO}	Q^2_{F1}

4-1 ABD	0.90	0.88	0.71
4-1 DBD	0.86	0.81	0.45
4-2 ABD	0.87	0.82	0.68
4-2 DBD	0.92	0.88	0.60
4-3 ABD	0.94	0.91	0.04
4-3 DBD	0.90	0.88	0.78
4-4 ABD*	0.87	0.82	**0.79**
4-4 DBD	0.82	0.72	0.66

*Selected divisions for further studies; ABD: Sorted activity-based division; DBD: Euclidean distance-based division

The results of preliminary SMLR models suggest the 4th seed among four sorted activity-based divisions from each descriptor class as the best division. Thus, we went ahead with the 4-4 sorted activity-based divisions (4-4 ABD) for feature selection. The feature selection was performed parallelly for each class of mixture descriptors. We have used all possible subset selection strategy for identifying the most important descriptors responsible for toxicity in honey bees (using a freely available in-house tool "BestSubsetSelectionModified_v2.1"). It is simply based on the multiple linear regression (MLR) algorithm and explores all possible combinations of the descriptors from a provided pool (Roy et al., 2015). All possible subset selection is a computationally exhaustive process which needs greater time and computer proficiency than the other feature selection approaches. Therefore it is always accompanied by a descriptor pool reduction step. Here in this study, we have used Genetic algorithm (GA) for pool reduction. The GA is another efficient feature selection technique and was first used by Rogers and Hopfingers (Rogers and Hopfinger, 1994). The simultaneous exploration of many possible solutions in the various regions of descriptor space makes it more powerful (Leardi, 1996). It is a two-step process where N number of descriptor population (parents) are selected from the whole pool first, and the fitness of each combination is determined. In the second step, crossover and mutation are taken place to get a new generation of descriptor combinations from the parents based on their scaled fitness scores. It is a process of genetic movement (movement of features) from parents to children; thus selection, crossover, and mutation are repeated until all the parent descriptors are replaced with the new ones. The fitness scores of the new combinations are again evaluated and the process continues until 80% of the new combinations attain the target score. We have performed GA using a freely available tool "GeneticAlgorithm_v4.1" with the training set of mixtures.

Finally, three different combinations of six descriptors have been obtained from three types of mixture descriptors (linear combination of molecular contributions, squared molecular contributions, and norm of molecular contributions) by all possible subset selections. From these three sets of descriptors, we have developed three different QSAR models by partial least squares regression (PLS). PLS is the generalization of multiple linear regression and is used to model strongly correlated, collinear, and noisy data with numerous descriptors (Cramer, 1993). It can also handle numerous Y variables (biological responses/molecular properties) simultaneously. The problem of overfitting with a large number of descriptors (close to the number of data points) in MLR can be successfully resolved with PLS (Wold et al., 2001). In this method, latent variables (LVs) such as X scores (T) and Y scores (U) are

calculated. Y scores are calculated from the X scores and response variables and then used for the final prediction. The LVs account for the maximum variation of the descriptor matrix to calculate the response. In this research, we have performed the PLS using the freely available tool "PLS_Single Y_version 1.0" and reported the resultant model with the optimum number of LVs (the number of LVs must be less than the number of descriptors) in **Table 3.14**. These selected descriptors were taken forward for the next step i.e., RASAR descriptor calculation, which has been discussed afterwards.

Table 3.14. Selected descriptors from the pool of three different types of descriptors obtained by different mixing rules; quality metrics of developed PLS model with selected descriptors

Mixture descriptor type	R^2	Q^2_{LOO}	Q^2_{F1}	Selected descriptors	No. of LVs
1: (Linear combination of molecular contributions)	0.888	0.853	0.864	**Six descriptors:** B05[C-O], B06[C-Cl], Eta_betaS_A, T(O..O), F06[C-F], F06[O-Cl]	3
2: (Squared molecular contributions)	0.869	0.835	0.862	**Six descriptors:** B06[O-Cl], D/Dtr03, Eta_betaS_A, mindsN, nCsp2, RDCHI	3
3: (Norm of molecular contributions)	0.894	0.854	0.862	**Six descriptors:** B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, B01[C-S], NssssC	3

LVs: Latent variables

3.4.4. Read-across and RASAR descriptors calculation

The read-across structure-activity relationship (RASAR) approach combines two predictive modeling techniques namely read-across and QSAR, for better predictability. The RASAR descriptors include similarity, error, and concordance measures obtained from the read-across prediction tool (Banerjee and Roy, 2022). To develop RASAR models, these read-across-derived parameters are clubbed together with significant structural and physicochemical descriptors that are correlated with the exerted molecular property in general. The RASAR descriptor calculation is accompanied by the read-across predictions and subsequent hyperparameter optimization. Read-across is a similarity-based prediction method where the properties (biological/physicochemical) of new/unknown compounds are predicted from their structural analogues. These structural analogues are selected based on structural similarity between query and source compounds, which can be obtained by several similarity estimation techniques. In one of our previous works, we have explained the similarity estimation by the Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity in detail along with the hyperparameter optimization (sigma: Gaussian kernel similarity estimation; gamma: Laplacian kernel similarity estimation, and the number of close source analogues for optimum prediction: all the three similarity estimation techniques) which is necessary for obtaining the best quality of predictions (Chatterjee et al., 2022). In this study, different sigma values (0.25 to 2 with 0.25 interval), different gamma values (0.25 to 2 with 0.25 interval), and 2 to 10 number of close source analogues have been used for optimizing these hyperparameters (based on the quality of prediction in terms of Q^2_{F1} value for a validation set). The hyperparameters have been optimized by employing an in-house tool "Auto RA Optimize v1.0" and the optimized settings have been reported in **Table 3.15**. Finally, the RASAR descriptors were calculated with the optimized

hyperparameters by employing a freely available in-house tool "RASAR Descriptor Calculator v2.0". One point should be noted here: the training set is used as both source and query mixture set for the calculation of RASAR descriptors of the training set; therefore, there is a chance of selecting the query compound itself as the most similar compound along with the others for each training data point. To avoid this issue, we have followed the "leave-same-out" strategy where the same compound (query compound present in the training set) is removed from the training dataset based on the similarity and corresponding response value during RASAR descriptor calculation. However, the RASAR descriptors of the test set can be calculated conventionally by considering training and test set compounds as the source and query chemicals respectively.

Table 3.15. Optimized settings for RASAR-descriptor calculation of three different classes

Class of mixture descriptors	Similarity estimation technique of RA used	Optimum sigma (σ) value	Optimum gamma (γ) value	Optimum number of close-source mixtures	Distance threshold	Similarity threshold
Mix. Desc. 1	LK	-	2	2	-	0
Mix. Desc. 2	GK	1	-	6	-	0
Mix. Desc. 3	ED	-	-	2	1	0

Mix. Desc.: Mixture descriptors; LK: Laplacian kernel function similarity estimation; GK: Gaussian kernel function similarity estimation; ED: Euclidean distance-based similarity estimation

3.4.5. q-RASAR modeling

Three different q-RASAR models have been developed in this research from three different classes of descriptors. To do so, we have performed another round of feature selection with all possible subset selection strategy. In each descriptor class, a pool of 21 combined descriptors (physicochemical, structural, and RASAR) has been used. All possible subset selection has been performed using the 'BestSubsetSelectionModified_v2.1' tool. Although the models were developed solely based on training set information, we have shown the model with the highest external predictivity as evidenced from the external correlation coefficient values (Q^2_{F1}). Finally, three different PLS q-RASAR models have been developed with the selected descriptors by using the 'PLS_Single Y_version 1.0' tool. From the quality of predictions of the developed PLS q-RASAR models, the most important mixing rule has been identified and reported as the final model of this study. To enhance the predictability of the final model, we have used machine learning (ML)-based linear and non-linear support vector machines (SVM) algorithms (Hearst et al., 1998), and ridge regression (Hoerl and Kennard, 1970). We have used Python-based codes using the Jupyter Notebook web application (Kluyver et al., 2016) in the Anaconda Navigator version 2022.05 with Python version 3.10.4 for ML-based model development. The associated hyperparameters were optimized by the Grid search technique using 5-fold cross-validation mean squared error as the objective function in each of the ML approaches, and the best model was selected for the prediction of test set compounds for which the quality was judged in terms of Q^2_{F1} and $MAE_{(Test)}$ values. A detailed flow diagram of this study has been given in **Figure 3.7**.

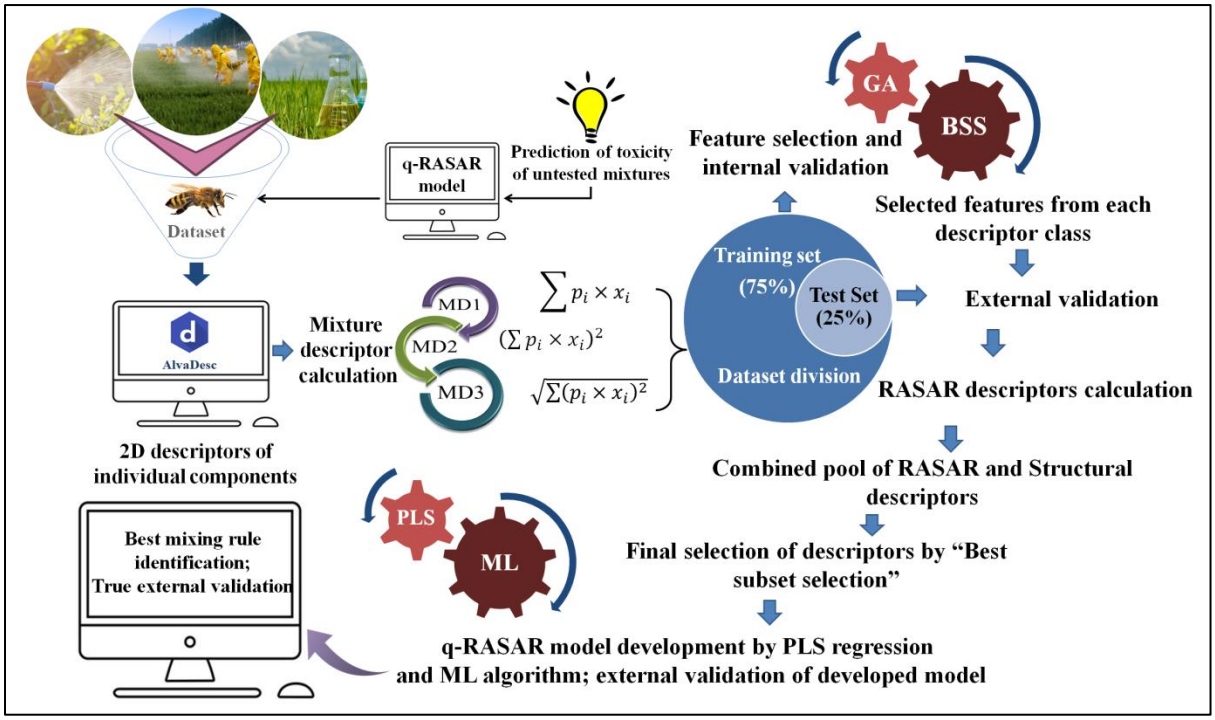


Figure 3.7. Flow diagram of the q-RASAR modeling approach (*BSS: Best subset selection; GA: Genetic algorithm; ML: Machine learning algorithm; PLS: Partial least squares)

3.4.6. Validation protocols

The acceptance of any predictive model depends on proper validation. The Organization of economic cooperation and development (OECD) proposed five different principles which are obeyed worldwide for the characterization of QSAR models for regulatory purposes (OECD, 2007). Statistical validation metrics are extensively used to evaluate the quality and predictability of the developed QSAR models, based on which the models are accepted or rejected. The QSAR modelers must perform proper statistical validation during modeling to comply with the OECD guideline 4 (OECD, 2004). Two types of metrics are calculated two comply with the guideline. The statistical quality and internal validation metrics such as regression coefficient (R^2), cross-validated (*leave-one-out*) correlation coefficient (Q^2_{LOO}), mean absolute error (MAE) etc. are computed with the training data points (Roy et al., 2015). R^2, Q^2_{LOO}, and MAE are the measures of goodness-of-fit, robustness, and errors of prediction respectively. There are high chances of overestimation of R^2 where the number of descriptors is close to the number of data points. To penalize the overestimated R^2, adjusted R^2 is calculated which considers the number of descriptors also along with the number of data points. The value of R^2, R^2_{adj}, and Q^2_{LOO} should be near 1, and a higher deviation from the ideal value suggests the inferiority of the model. An optimum value of $Q^2_{LOO} > 0.6$ is mostly used (Roy et al., 2015). The error metric (MAE) should be near zero. The Akaike information criteria (AIC) values were also computed for comparing the q-RASAR models obtained from three different mixture descriptors in this study. This metric estimates the prediction error of the models developed from a single dataset and thereby, judges the relative efficiency of the models. For a regression model, AIC is computed by employing the following equation 3.12 (Burnham and Anderson, 2002):

$$AIC = n \times log(\sigma^2) + 2 \times K \qquad (3.12)$$

where, n = number of training data, K = number of descriptors (LVs for PLS), and σ^2 = mean sum squared residual. The least AIC value of a model is the indicator of higher relative efficiency among various models developed from a dataset.

To check the possibility of any chance correlation during model development, Y-randomization analysis (Topliss and Edwards, 1979) has been performed using the "SIMCA-P" software (Wu et al., 2010) and the results are reported. The response values are randomly shuffled (100 times in this study) of the training set without altering the descriptors and then refitted with the original model for computing the R^2 and Q^2 (computed metrics of randomized models) values. Finally, the computed R^2 and Q^2 values are plotted (Y-axis) against the correlation coefficient between the original Y values and the permuted Y values (X-axis). The intercept of R^2 and Q^2 (randomized) plots are recorded. Poor values of these intercepts (R^2 intercept < 0.3; Q^2 intercept < 0.05) are the indicator of a statistically significant model without any chance correlation (Rücker et al., 2007). To comply with the OECD principle 3 (a defined domain of applicability), an applicability domain analysis (Gadaleta et al., 2016) has also been performed by DModX approach using "SIMCA-P" software.

The external validation is performed to check the predictability of the developed model and the test set compounds are exclusively used for this purpose. External correlation coefficients such as Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} are the classic markers of predictability and their values should be near 1 (Roy et al., 2015). In common practice, the optimum value of these three metrics for the selection of model should be > 0.6. Error metrics like mean absolute error (MAE_{Test}) is also calculated to check the error of predicted results. Besides these classical metrics, r_m^2 metrics (*average* $r_{m}^{2}{}_{(LOO)}$, $\Delta r_{m}^{2}{}_{(LOO)}$, *average* $r_{m}^{2}{}_{(Test)}$, $\Delta r_{m}^{2}{}_{(Test)}$) are calculated for internal validation as well as external validation (Ojha et al., 2011). The optimum value of average $r_m^2{}_{(LOO)}$ and average $r_m^2{}_{(Test)}$ is > 0.5, whereas the expected value for $\Delta r_m^2{}_{(LOO)}$ and $\Delta r_m^2{}_{(Test)}$ should be < 0.2. A true external validation has also been performed in this study with 13 true external set mixtures, which are showcased in **Table 3.12**. As discussed in the section "***3.4.1.Toxicity data***", these 13 sets of duplicate mixtures correspond to 44 data points and respective 44 observed toxicity values. The most predictive model (reported in the section "4.4.2. Best model and result of the validation") has been used for the prediction of toxicity of these 13 different true external set mixtures and classify them as toxic or non-toxic based on a threshold value (response mean of the training set = -0.127 μg active substance/bee). Among the observed toxicity values, the most similar one among duplicate data points with the predicted response is selected for the calculation of classification-based metrics. The classification-based metrics have been calculated and reported subsequently.

3.4.7. Prediction of unknown mixtures with PLS and ML-based q-RASAR models and reliability analysis with PRI tool

To validate the performance of the new q-RASAR model against unknown mixtures (different from the modeling set), we have used the relevant, reliable dataset of 361 binary mixtures composed of organic chemicals used as plant protection products published by Tosi et al., 2022. After manual checking of these new binary mixtures, 143 data points have been excluded (reasons: same as modeling mixtures, duplicates, ternary mixtures, or metallic mixtures). The remaining 218 data points have included in our analysis for toxicity prediction and further reliability analysis. These 218 binary mixtures are composed of 64 new components along with the previous 27 components (**Table 3.10**) that are present in the modeling set. The unique mixture components and the composition of untested mixtures have

been represented in **Tables 3.16** and **Table 3.17** respectively. The 2D structures of the new components have been retrieved from the PubChem database ("PubChem," 2023) and also cross-checked with the ChemSpider database (RSC, 2022). The 2D descriptors of new components have been calculated by "AlvaDesc" software; then the mixture descriptors have been computed with the most relevant mixing rule (identified by the PLS q-RASAR model) discussed in the section "3.1.Developed q-RASAR models and best mixing rule identification". The RASAR descriptors of true external mixtures have been computed by the methodology discussed in the section "Read-across and RASAR descriptors calculation". Finally, the predicted response from PLS q-RASAR and different ML-based q-RASAR models (linear SVM, non-linear SVM, ridge regression) have been reported.

Table 3.16. New mixture components present in true external mixture dataset (Tosi et al., 2022)

Sl. No.	Name
C28	Acephate
C29	Aldrin
C30	Alpha-cypermethrin
C31	Arsenic acid
C32	Atrazine
C33	Azoxystrobin
C34	Bifenthrin
C35	Carbaryl
C36	Carbendazim
C37	Chlorantraniliprole
C38	Chlorpyrifos
C39	Clothianidin
C40	Coroxon
C41	Cypermethrin
C41	Cyprodinil
C42	Deltamethrin
C43	Dieldrin
C44	Diethyl maleate
C45	Difenoconazole
C46	Dimethoate
C47	Dinotefuran

C48	DMPF
C49	Endosulfan
C50	Endosulfan sulfate
C51	Esfenvalerate
C52	Fenpropathrin
C53	Fipronil
C54	Flumethrin
C55	Flupyradifurone
C56	Flutriafol
C57	Fumagillin
C58	Glyphosate
C59	Imazalil
C60	Indole-3-carbinol
C61	Iprodione
C62	Ivermectin
C63	Methoxyfenozide
C64	Metolachlor
C65	Omethoate
C66	Oxamyl
C67	Oxyfluorfen
C68	Oxytetracycline
C69	Paradichlorobenzene
C70	Parathion-methyl
C71	Penconazole
C72	Pendimethalin
C73	Permethrin
C74	Phenobarbital
C75	Pristine
C76	Pyrethrum
C77	Pyrimethanil

C78	Pyriproxyfen
C79	Quercetin
C80	Salicylic acid
C81	Sulfoxaflor
C82	Tebuconazole
C83	Tetraconazole
C84	Thiabendazole
C85	Thiamethoxam
C86	Thiophanate-methyl
C87	Triadimenol
C88	Trifloxystrobin
C89	Tylosin
C90	Verapamil
C91	Xanthotoxin

*C: Components

Table 3.17. Composition of external mixtures (Tosi et al., 2022)

Sl. No.	Pesticide A + Pesticide B	Mass ratio
UM 1	C1 + C44	0.5:0.5
UM 2	C1 + C57	0.5:0.5
UM 3	C1 + C79	0.5:0.5
UM 4	C1 + C85	0.5:0.5
UM 5	C1 + C90	0.5:0.5
UM 6	C2 + C44	0.5:0.5
UM 7	C2 + C57	0.5:0.5
UM 8	C2 + C68	0.5:0.5
UM 9	C2 + C89	0.5:0.5
UM 10	C3 + C21	0.5:0.5
UM 11	C3 + C41	0.5:0.5
UM 12	C3 + C72	0.5:0.5
UM 13	C3 + C77	0.5:0.5

UM 14	C4 + C3	0.5:0.5
UM 15	C4 + C21	0.5:0.5
UM 16	C4 + C32	0.5:0.5
UM 17	C4 + C41	0.5:0.5
UM 18	C5 + C44	0.5:0.5
UM 19	C5 + C57	0.5:0.5
UM 20	C5 + C68	0.5:0.5
UM 21	C5 + C89	0.5:0.5
UM 22	C6 + C44	0.5:0.5
UM 23	C7 + C4	0.5:0.5
UM 24	C7 + C10	0.5:0.5
UM 25	C7 + C11	0.5:0.5
UM 26	C7 + C13	0.5:0.5
UM 27	C7 + C15	0.5:0.5
UM 28	C7 + C19	0.5:0.5
UM 29	C7 + C20	0.5:0.5
UM 30	C7 + C24	0.5:0.5
UM 31	C7 + C33	0.5:0.5
UM 32	C7 + C34	0.5:0.5
UM 33	C7 + C36	0.5:0.5
UM 34	C7 + C44	0.5:0.5
UM 35	C7 + C56	0.5:0.5
UM 36	C7 + C59	0.5:0.5
UM 37	C7 + C71	0.5:0.5
UM 38	C7 + C74	0.5:0.5
UM 39	C7 + C86	0.5:0.5
UM 40	C7 + C87	0.5:0.5
UM 41	C10 + C33	0.5:0.5
UM 42	C10 + C72	0.5:0.5
UM 43	C11 + C5	0.5:0.5

UM 44	C11 + C10	0.5:0.5
UM 45	C11 + C41	0.5:0.5
UM 46	C11 + C44	0.5:0.5
UM 47	C11 + C57	0.5:0.5
UM 48	C11 + C68	0.5:0.5
UM 49	C11 + C72	0.5:0.5
UM 50	C11 + C89	0.5:0.5
UM 51	C12 + C5	0.5:0.5
UM 52	C12 + C11	0.5:0.5
UM 53	C12 + C44	0.5:0.5
UM 54	C12 + C45	0.5:0.5
UM 55	C12 + C57	0.5:0.5
UM 56	C12 + C60	0.5:0.5
UM 57	C12 + C68	0.5:0.5
UM 58	C12 + C74	0.5:0.5
UM 59	C12 + C79	0.5:0.5
UM 60	C12 + C80	0.5:0.5
UM 61	C12 + C86	0.5:0.5
UM 62	C12 + C89	0.5:0.5
UM 63	C12 + C91	0.5:0.5
UM 64	C13 + C4	0.5:0.5
UM 65	C13 + C5	0.5:0.5
UM 66	C13 + C7	0.5:0.5
UM 67	C13 + C15	0.5:0.5
UM 68	C13 + C28	0.5:0.5
UM 69	C13 + C33	0.5:0.5
UM 70	C13 + C34	0.5:0.5
UM 71	C13 + C40	0.5:0.5
UM 72	C13 + C52	0.5:0.5
UM 73	C13 + C58	0.5:0.5

UM 74	C13 + C59	0.5:0.5
UM 75	C13 + C66	0.5:0.5
UM 76	C13 + C72	0.5:0.5
UM 77	C13 + C81	0.5:0.5
UM 78	C13 + C83	0.5:0.5
UM 79	C13 + C86	0.5:0.5
UM 80	C17 + C5	0.5:0.5
UM 81	C18 + C19	0.5:0.5
UM 82	C18 + C73	0.5:0.5
UM 83	C21 + C72	0.5:0.5
UM 84	C21 + C77	0.5:0.5
UM 85	C22 + C10	0.5:0.5
UM 86	C22 + C15	0.5:0.5
UM 87	C22 + C18	0.5:0.5
UM 88	C22 + C33	0.5:0.5
UM 89	C22 + C41	0.5:0.5
UM 90	C22 + C82	0.5:0.5
UM 91	C23 + C5	0.5:0.5
UM 92	C23 + C16	0.5:0.5
UM 93	C23 + C44	0.5:0.5
UM 94	C23 + C57	0.5:0.5
UM 95	C23 + C68	0.5:0.5
UM 96	C23 + C89	0.5:0.5
UM 97	C28 + C7	0.5:0.5
UM 98	C28 + C38	0.5:0.5
UM 99	C28 + C39	0.5:0.5
UM 100	C28 + C58	0.5:0.5
UM 101	C28 + C66	0.5:0.5
UM 102	C28 + C83	0.5:0.5
UM 103	C29 + C74	0.5:0.5

UM 104	C30 + C4	0.5:0.5
UM 105	C30 + C19	0.5:0.5
UM 106	C30 + C20	0.5:0.5
UM 107	C30 + C36	0.5:0.5
UM 108	C30 + C82	0.5:0.5
UM 109	C32 + C65	0.5:0.5
UM 110	C33 + C3	0.5:0.5
UM 111	C33 + C4	0.5:0.5
UM 112	C33 + C36	0.5:0.5
UM 113	C33 + C72	0.5:0.5
UM 114	C34 + C4	0.5:0.5
UM 115	C34 + C12	0.5:0.5
UM 116	C34 + C33	0.5:0.5
UM 117	C35 + C12	0.5:0.5
UM 118	C35 + C32	0.5:0.5
UM 119	C37 + C20	0.5:0.5
UM 120	C37 + C61	0.5:0.5
UM 121	C38 + C3	0.5:0.5
UM 122	C38 + C4	0.5:0.5
UM 123	C38 + C5	0.5:0.5
UM 124	C38 + C7	0.5:0.5
UM 125	C38 + C10	0.5:0.5
UM 126	C38 + C11	0.5:0.5
UM 127	C38 + C12	0.5:0.5
UM 128	C38 + C20	0.5:0.5
UM 129	C38 + C21	0.5:0.5
UM 130	C38 + C23	0.5:0.5
UM 131	C38 + C32	0.5:0.5
UM 132	C38 + C33	0.5:0.5
UM 133	C38 + C34	0.5:0.5

UM 134	C38 + C36	0.5:0.5
UM 135	C38 + C41	0.5:0.5
UM 136	C38 + C49	0.5:0.5
UM 137	C38 + C63	0.5:0.5
UM 138	C38 + C72	0.5:0.5
UM 139	C38 + C76	0.5:0.5
UM 140	C38 + C77	0.5:0.5
UM 141	C38 + C82	0.5:0.5
UM 142	C39 + C3	0.5:0.5
UM 143	C39 + C12	0.5:0.5
UM 144	C39 + C15	0.5:0.5
UM 145	C39 + C20	0.5:0.5
UM 146	C39 + C21	0.5:0.5
UM 147	C39 + C32	0.5:0.5
UM 148	C39 + C46	0.5:0.5
UM 149	C39 + C47	0.5:0.5
UM 150	C39 + C69	0.5:0.5
UM 151	C39 + C82	0.5:0.5
UM 152	C39 + C85	0.5:0.5
UM 153	C41 + C10	0.5:0.5
UM 154	C41 + C21	0.5:0.5
UM 155	C41 + C32	0.5:0.5
UM 156	C41 + C33	0.5:0.5
UM 157	C41 + C72	0.5:0.5
UM 158	C41 + C77	0.5:0.5
UM 159	C41 + C88	0.5:0.5
UM 160	C42 + C19	0.5:0.5
UM 161	C42 + C45	0.5:0.5
UM 162	C42 + C86	0.5:0.5
UM 163	C43 + C74	0.5:0.5

UM 164	C46 + C20	0.5:0.5
UM 165	C46 + C32	0.5:0.5
UM 166	C48 + C5	0.5:0.5
UM 167	C48 + C11	0.5:0.5
UM 168	C48 + C12	0.5:0.5
UM 169	C49 + C33	0.5:0.5
UM 170	C49 + C34	0.5:0.5
UM 171	C50 + C33	0.5:0.5
UM 172	C50 + C34	0.5:0.5
UM 173	C50 + C38	0.5:0.5
UM 174	C50 + C49	0.5:0.5
UM 175	C51 + C72	0.5:0.5
UM 176	C53 + C21	0.5:0.5
UM 177	C54 + C86	0.5:0.5
UM 178	C55 + C20	0.5:0.5
UM 179	C62 + C57	0.5:0.5
UM 180	C62 + C79	0.5:0.5
UM 181	C62 + C90	0.5:0.5
UM 182	C63 + C4	0.5:0.5
UM 183	C63 + C7	0.5:0.5
UM 184	C63 + C10	0.5:0.5
UM 185	C63 + C34	0.5:0.5
UM 186	C63 + C41	0.5:0.5
UM 187	C63 + C72	0.5:0.5
UM 188	C69 + C32	0.5:0.5
UM 189	C69 + C38	0.5:0.5
UM 190	C69 + C72	0.5:0.5
UM 191	C69 + C85	0.5:0.5
UM 192	C70 + C12	0.5:0.5
UM 193	C72 + C32	0.5:0.5

UM 194	C72 + C64	0.5:0.5
UM 195	C72 + C67	0.5:0.5
UM 196	C73 + C4	0.5:0.5
UM 197	C73 + C32	0.5:0.5
UM 198	C77 + C33	0.5:0.5
UM 199	C77 + C72	0.5:0.5
UM 200	C77 + C88	0.5:0.5
UM 201	C78 + C3	0.5:0.5
UM 202	C78 + C41	0.5:0.5
UM 203	C78 + C72	0.5:0.5
UM 204	C78 + C77	0.5:0.5
UM 205	C82 + C36	0.5:0.5
UM 206	C85 + C3	0.5:0.5
UM 207	C85 + C7	0.5:0.5
UM 208	C85 + C15	0.5:0.5
UM 209	C85 + C20	0.5:0.5
UM 210	C85 + C32	0.5:0.5
UM 211	C85 + C33	0.5:0.5
UM 212	C85 + C38	0.5:0.5
UM 213	C85 + C47	0.5:0.5
UM 214	C85 + C64	0.5:0.5
UM 215	C85 + C82	0.5:0.5
UM 216	C85 + C84	0.5:0.5
UM 217	C86 + C4	0.5:0.5
UM 218	C88 + C72	0.5:0.5

*UM: Untested mixtures; C: components

As an additive measure, the reliability of PLS q-RASAR predictions has been checked by an in-house java-based tool "Prediction reliability indicator" (PRI) (Roy et al., 2018a). This tool can categorize the reliability of the predicted value into 'Good', 'Moderate', and 'Unreliable', depending upon the composite score. PRI tool essentially uses three sets of principles for the reliability analysis of predicted responses and calculates the composite scores; these are- "*(i) mean absolute error of prediction (leave-one-out) of a query compound from 10 most close training compounds; (ii) applicability domain in terms of similarity, by using standardization*

approach; (iii) closeness of predicted values to the mean of the observed response of training set compounds".

3.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species

3.5.1. Mixture toxicity data collection and dataset preparation

The toxicity data for individual antibiotics (SAs, SAPs, TCs) and their binary and ternary mixtures were collected against *Vibrio fischeri*, *Escherichia coli*, and *Bacillus subtilis* separately from the published work by Wang et al. (D. Wang et al., 2018). We compiled all the data and represented them into three different mixture datasets (**MD1**, **MD2**, and **MD3**). The MD1 contains the toxicity data of 51 binary and 25 ternary mixtures against *V. fischeri*, the MD2 contains the toxicity data of 43 binary mixtures and 24 ternary mixtures against *E. coli*, and the MD3 contains the toxicity data of 29 binary mixtures and 14 ternary mixtures against *B. subtilis*. All the mixtures from these three mixture datasets (MD1, MD2, and MD3) were composed of different compositions of 8 SAs, 4 SAPs, and 4 TCs which have been showcased along with their individual toxicity values (pEC50 (Mol/L)) against the studied bacteria (*V. fischeri*, *E. coli*, *B. subtilis*) in **Table 3.18**. All the mixtures were composed of an equitoxic quantity of the component antibiotics (amount required to produce one toxic unit for individual components) based on their individual toxicity towards the corresponding bacterium. The mixtures of MD1, MD2, and MD3 were tabulated along with the composition and observed toxicity values in **Table 3.19**, **3.20**, and **3.21** respectively.

Table 3.18. Name, type, CAS No., and toxicity (pEC50 (Mol/L)) of individual antibiotics (here mixture components)

Sl. No.	Name	Type	CAS No.	Toxicity against *V. fischeri* [pEC50 (Mol/L)]	Toxicity against *E. coli* [pEC50 (Mol/L)]	Toxicity against *B. subtilis* [pEC50 (Mol/L)]
1	Sulfamethazine	SA	57-68-1	4.45	5.23	4.71
2	Sulfameter	SA	651-06-9	4.36	5.43	4.72
3	Sulfamethoxypyridazine	SA	80-35-3	4.86	5.3	4.75
4	Sulfisoxazole	SA	127-69-5	4.63	5.47	4.83
5	Sulfadoxine	SA	2447-57-6	4.04	5.34	5.07
6	Sulfamethoxazole	SA	723-46-6	4.87	5.79	5.08
7	Sulfadiazine	SA	68-35-9	4.28	5.56	5.11
8	Sulfamonomethoxine	SA	1220-83-3	4.89	5.73	5.49
9	Trimethoprim	SAP	738-70-5	6.49	7.45	5.78
10	Diaveridine	SAP	5355-16-8	6.49	6.4	6.15
11	Ormetoprim	SAP	6981-18-6	6.51	6.81	6.3
12	Pyrimethamine	SAP	58-14-0	5.04	NA	NA
13	Doxycycline hyclate	TC	24390-14-5	4.41	7.82	6.41
14	Tetracycline hydrochloride	TC	64-75-5	3.96	7.3	5.38
15	Chlortetracycline hydrochloride	TC	64-72-2	4.42	7.71	6.28
16	Oxytetracycline hydrochloride	TC	2058-46-0	3.71	7.18	4.64

Table 3.19. Compositions and toxicity data (pEC50 (Mol/L)) of 76 mixtures against *V. fischeri*

SL. NO.	Component 1	Fraction 1	Component 2	Fraction 2	Component 3	Fraction 3	pEC50M (Mol/L)
VF1	9	0.602599814	7	0.397400186			4.84
VF2	9	0.57029877	8	0.42970123			5.56
VF3*	9	0.571806167	3	0.428193833			5.43
VF4	9	0.571302817	6	0.428697183			5.43
VF5	9	0.593235832	1	0.406764168			5.02
VF6	9	0.598156682	2	0.401843318			4.75
VF7*	10	0.57029877	8	0.42970123			5.62
VF8	10	0.571806167	3	0.428193833			5.54
VF9	10	0.593235832	1	0.406764168			5.11
VF10*	10	0.583633094	4	0.416366906			5.28
VF11	10	0.598156682	2	0.401843318			5.05
VF12*	11	0.571052632	8	0.428947368			5.68
VF13	11	0.572559367	3	0.427440633			5.6
VF14	11	0.593978102	1	0.406021898			5.13
VF15	11	0.58438061	4	0.41561939			5.31
VF16	11	0.598896044	2	0.401103956			5.06
VF17	12	0.50755287	8	0.49244713			5.23
VF18	12	0.509090909	3	0.490909091			5.09
VF19	12	0.508577195	6	0.491422805			5.18
VF20	12	0.531085353	1	0.468914647			4.93
VF21	12	0.521199586	4	0.478800414			5.05
VF22	12	0.536170213	2	0.463829787			4.82
VF23	14	0.447457627	8	0.552542373			4.07
VF24*	14	0.448979592	3	0.551020408			4.05

VF25*	14	0.448471121	6	0.551528879			4.11
VF26	14	0.470868014	1	0.529131986			4.01
VF27	14	0.461001164	4	0.538998836			3.99
VF28	14	0.475961538	2	0.524038462			3.96
VF29*	15	0.522458629	5	0.477541371			4
VF30	15	0.508045977	7	0.491954023			4.07
VF31	15	0.503416856	2	0.496583144			4.04
VF32*	16	0.431395349	8	0.568604651			3.75
VF33	16	0.432905484	3	0.567094516			3.71
VF34	16	0.432400932	6	0.567599068			3.69
VF35	16	0.454656863	1	0.545343137			3.81
VF36*	16	0.444844125	4	0.555155875			3.8
VF37	16	0.459727385	2	0.540272615			3.79
VF38	13	0.521893491	5	0.478106509			3.92
VF39*	13	0.507479862	7	0.492520138			3.93
VF40	13	0.497742664	1	0.502257336			4.08
VF41*	13	0.502850627	2	0.497149373			4.1
VF42*	14	0.378947368	9	0.621052632			4.1
VF43	14	0.378947368	10	0.621052632			4.06
VF44	14	0.378223496	11	0.621776504			4.08
VF45	14	0.44	12	0.56			4.01
VF46	15	0.467230444	12	0.532769556			4.39
VF47	16	0.36372549	9	0.63627451			3.83
VF48	16	0.36372549	10	0.63627451			3.88
VF49	16	0.363013699	11	0.636986301			3.84
VF50	16	0.424	12	0.576			3.76
VF51	13	0.466666667	12	0.533333333			4.29
VF52*	14	0.273291925	9	0.4478951	5	0.278812974	4.46

VF53	14	0.268839104	9	0.44059742	7	0.290563476	4.49
VF54	14	0.258148631	9	0.423076923	8	0.318774446	4.75
VF55*	14	0.258654474	9	0.423905944	3	0.317439582	4.64
VF56	14	0.25848564	9	0.423629243	6	0.317885117	4.65
VF57	14	0.265771812	9	0.43557047	1	0.298657718	4.57
VF58	14	0.262599469	9	0.430371353	4	0.307029178	4.6
VF59	14	0.267386901	9	0.438217421	2	0.294395679	4.49
VF60	13	0.295180723	10	0.434404284	5	0.270414993	4.6
VF61	13	0.288990826	10	0.425294889	2	0.285714286	4.9
VF62	16	0.245858184	11	0.431411531	5	0.322730285	4.31
VF63	16	0.255862069	11	0.448965517	7	0.295172414	4.37
VF64*	16	0.24553276	11	0.430840503	8	0.323626737	4.44
VF65	16	0.24602122	11	0.431697613	3	0.322281167	4.34
VF66	16	0.245858184	11	0.431411531	6	0.322730285	4.38
VF67	16	0.252897069	11	0.443762781	1	0.30334015	4.35
VF68	16	0.24983165	11	0.438383838	4	0.311784512	4.31
VF69*	16	0.254458162	11	0.446502058	2	0.299039781	4.25
VF70	15	0.327407407	12	0.373333333	5	0.299259259	4.79
VF71	15	0.321688501	12	0.366812227	7	0.311499272	4.72
VF72*	15	0.308013937	12	0.351219512	8	0.340766551	4.97
VF73	15	0.308659218	12	0.351955307	3	0.339385475	4.89
VF74	15	0.308443824	12	0.3517097	6	0.339846476	4.86
VF75*	15	0.317757009	12	0.36232926	1	0.319913731	4.75
VF76*	15	0.313697658	12	0.357700497	4	0.328601845	4.81

*Test set mixtures of modeling division

Table 3.20. Compositions and toxicity data (pEC50 (Mol/L)) of 67 mixtures against *E. coli*

SL. NO.	Component 1	Fraction 1	Component 2	Fraction 2	Component 3	Fraction 3	pEC50M (Mol/L)
EC1	9	0.562688822	6	0.437311178			6.51
EC2	9	0.578416149	2	0.421583851			6.22
EC3*	10	0.545144804	5	0.454855196			6.21
EC4	10	0.535117057	7	0.464882943			6.44
EC5*	10	0.527617477	8	0.472382523			6.43
EC6	10	0.525020509	6	0.474979491			6.72
EC7	10	0.539174389	4	0.460825611			6.36
EC8	11	0.543062201	8	0.456937799			6.46
EC9	11	0.562345169	3	0.437654831			6.38
EC10	11	0.565614618	1	0.434385382			6.29
EC11*	14	0.577531646	5	0.422468354			5.46
EC12*	14	0.567651633	7	0.432348367			5.73
EC13	14	0.579365079	3	0.420634921			5.45
EC14*	14	0.557677617	6	0.442322383			5.97
EC15	14	0.582601756	1	0.417398244			5.41
EC16	14	0.57165231	4	0.42834769			5.61
EC17	14	0.573448547	2	0.426551453			5.55
EC18	15	0.590804598	5	0.409195402			5.4
EC19	15	0.581009797	7	0.418990203			5.62
EC20	15	0.573660714	8	0.426339286			5.71
EC21	15	0.592621061	3	0.407378939			5.46
EC22	15	0.595826893	1	0.404173107			5.44
EC23	15	0.584977238	4	0.415022762			5.61
EC24*	15	0.586757991	2	0.413242009			5.55

EC25	16	0.573482428	5	0.426517572			5.44
EC26	16	0.563579278	7	0.436420722			5.73
EC27	16	0.556158017	8	0.443841983			5.85
EC28	16	0.575320513	3	0.424679487			5.51
EC29	16	0.553585197	6	0.446414803			5.9
EC30	16	0.567588933	4	0.432411067			5.6
EC31	13	0.594224924	5	0.405775076			5.41
EC32*	13	0.596036585	3	0.403963415			5.42
EC33*	13	0.599233716	1	0.400766284			5.36
EC34	13	0.58841234	4	0.41158766			5.67
EC35*	13	0.590188679	2	0.409811321			5.61
EC36	14	0.494915254	9	0.505084746			7.22
EC37	14	0.532846715	10	0.467153285			6.41
EC38	14	0.517363572	11	0.482636428			6.73
EC39	15	0.546420978	10	0.453579022			6.59
EC40	16	0.490772386	9	0.509227614			7.17
EC41*	16	0.528718704	10	0.471281296			6.5
EC42	16	0.513223731	11	0.486776269			6.76
EC43	13	0.549929677	10	0.450070323			6.53
EC44	14	0.363364858	9	0.370831259	5	0.2658039	6.16
EC45	14	0.359428853	9	0.366814377	7	0.2737568	6.42
EC46*	14	0.356445313	9	0.363769531	8	0.2797852	6.55
EC47	14	0.364089776	9	0.371571072	3	0.2643392	6.16
EC48	14	0.35540409	9	0.362706913	6	0.281889	6.65
EC49	14	0.365365365	9	0.372872873	1	0.2617618	6.21
EC50	14	0.361028684	9	0.368447082	4	0.2705242	6.3
EC51*	14	0.361744301	9	0.369177403	2	0.2690783	6.28
EC52	13	0.399795501	10	0.327198364	5	0.2730061	6.06

EC53*	13	0.395348837	10	0.323559151	7	0.281092	6.31
EC54	13	0.39197995	10	0.320802005	8	0.287218	6.45
EC55	13	0.400614754	10	0.327868852	3	0.2715164	6.06
EC56	13	0.390804598	10	0.31984008	6	0.2893553	6.57
EC57	13	0.402056555	10	0.329048843	1	0.2688946	6.11
EC58	13	0.397155917	10	0.32503809	4	0.277806	6.31
EC59*	13	0.397964377	10	0.325699746	2	0.2763359	6.38
EC60*	16	0.371443352	11	0.352302121	5	0.2762545	6.15
EC61	16	0.367263427	11	0.348337596	7	0.284399	6.51
EC62	16	0.364097363	11	0.345334686	8	0.290568	6.6
EC63	16	0.372213582	11	0.353032659	3	0.2747538	6.24
EC64*	16	0.362992922	11	0.344287159	6	0.2927199	6.63
EC65	16	0.373569199	11	0.354318418	1	0.2721124	6.16
EC66	16	0.368961973	11	0.349948613	4	0.2810894	6.38
EC67	16	0.369721936	11	0.350669413	2	0.2796087	6.28

*Test set mixtures of modeling division

Table 3.21. Compositions and toxicity data (pEC50 (Mol/L)) of 42 mixtures against *B. subtilis*

SL. NO.	Component 1	Fraction 1	Component 2	Fraction 2	Component 3	Fraction 3	pEC50M (Mol/L)
BS1	9	0.512866016	8	0.487133984			6.08
BS2	9	0.548907882	3	0.451092118			5.73
BS3	9	0.544769086	4	0.455230914			5.74
BS4*	10	0.548128342	5	0.451871658			5.87
BS5	10	0.546181172	7	0.453818828			5.8
BS6	11	0.57168784	2	0.42831216			5.63
BS7*	14	0.531095755	3	0.468904245			4.9
BS8*	14	0.533201189	1	0.466798811			4.99
BS9	14	0.526934378	4	0.473065622			5.02
BS10	14	0.532673267	2	0.467326733			4.95
BS11	15	0.569356301	3	0.430643699			5.1
BS12	15	0.571428571	1	0.428571429			5.08
BS13*	15	0.570909091	2	0.429090909			5.01
BS14	16	0.477857878	5	0.522142122			4.85
BS15	16	0.477366255	6	0.522633745			4.77
BS16*	16	0.489968321	4	0.510031679			4.8
BS17	16	0.495726496	2	0.504273504			4.78
BS18*	13	0.556423611	7	0.443576389			5.1
BS19*	13	0.570284698	4	0.429715302			4.97
BS20	14	0.482078853	9	0.517921147			5.28
BS21	14	0.466608846	10	0.533391154			5.28
BS22	14	0.460616438	11	0.539383562			5.42
BS23	15	0.520729685	9	0.479270315			5.7
BS24	15	0.505229284	10	0.494770716			5.98

BS25*	15	0.499205087	11	0.500794913			6.07
BS26	16	0.445297505	9	0.554702495			4.82
BS27*	16	0.430027804	10	0.569972196			4.82
BS28	13	0.525840853	9	0.474159147			5.86
BS29	13	0.510350318	10	0.489649682			6.11
BS30	14	0.331484904	9	0.356130622	5	0.312384473	6.09
BS31*	14	0.330669945	9	0.355255071	7	0.314074985	6.06
BS32	14	0.323123123	9	0.347147147	8	0.32972973	6.37
BS33	14	0.338152106	9	0.363293526	3	0.298554368	5.99
BS34	14	0.331280788	9	0.35591133	6	0.312807882	6.15
BS35	14	0.338790932	9	0.363979849	2	0.297229219	5.93
BS36	13	0.370306181	10	0.355285962	3	0.274407857	6.03
BS37	13	0.371163868	10	0.356108859	1	0.272727273	5.96
BS38	13	0.368602645	10	0.353651524	4	0.277745831	6
BS39	16	0.289818863	11	0.39350406	5	0.316677077	5.97
BS40	16	0.289096573	11	0.392523364	7	0.318380062	5.94
BS41	16	0.295729764	11	0.401529637	3	0.302740599	6.08
BS42	16	0.296296296	11	0.402298851	2	0.301404853	6

*Test set mixtures of modeling division

3.5.2. Mixture descriptor calculation considering three different mixing rules

In this study, 9 classes of 0D-2D descriptors (total count = 2400) such as constitutional indices, ring descriptors, connectivity index, electro-topochemical atom indices, functional group counts, atom-centred fragments, atom-type E-state indices, 2D atom pairs, and molecular properties were used. Molecular descriptors of individual mixture components (presented in **Table 3.18**) were computed first using "AlvaDesc" software (Mauri, 2020) from the properly aromatized and explicit H-containing two-dimensional structure inputs (in the *.sdf format). All the 2D structures were drawn using "Marvin Sketch" software (ChemAxon, 2018) without energy minimization and conformational analysis (energy minimization and conformational analysis are important prerequisites for 3D and higher-order descriptor calculation). These classes of descriptors were previously used for modeling in various instances for both single and mixture toxicity modeling (Chatterjee and Roy, 2021, 2022b; Mukherjee et al., 2022; Paul et al., 2022). In a successive step, a pretreatment was performed to the obtained descriptor pool (2400 descriptors) of 16 mixture components with the in-built tool of "AlvaDesc" software. In this step, the redundant descriptors (descriptors with a constant value, descriptors with high inter-correlation, descriptors with very low diversity, etc) were removed. After the pretreatment, we computed the mixture descriptors from the pretreated pool by employing 3 different mixing rules i.e., (i) linear combination of molecular contributions, (ii) the squared molecular contributions, and (iii) norm of molecular contributions with a freely available Java-based tool namely "Mix_Desc_Calc" version 1.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1). The linear combination of molecular contributions was used to consider the probable additive effects of mixture components and the other two mixing rules i.e., squared molecular contributions and norm of molecular contributions were specified for accounting interactions. The mixing rules have been explained in section "***3.4.2. Mixture descriptors***".

3.5.3. Dataset division and proper mixing rules identification

The dataset division was performed in 3 different techniques i.e., Sorted activity-based division, Euclidean distance-based division, and Kennard-Stone division using the "Dataset Division GUI" version 1.2 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). To avoid the probable selection bias of training and test mixtures, both descriptor-based (Euclidean distance-based division and Kennard-Stone division) and response-based (Sorted activity-based division) techniques were utilized. Nine different training-test division sets were obtained (4 seeds of sorted activity-based divisions, 4 seeds of Euclidean distance-based divisions, and 1 Kennard-Stone division) for each descriptor matrix; therefore a total of 27 division sets have been generated from each mixture dataset (from each dataset, we can get 3 different descriptor matrix using 3 mixing rules). In this study, we divided the mixture data into training and test sets of 75% and 25% respectively of the whole data set size; the training sets were used for model development and internal validation and the test sets were exclusively used for external validation of developed models. Before the initial modeling, all the training and test set pairs were subjected to a data pretreatment using the "DataPreTreatmentTrainTest" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) to further reduce the redundant data generated after division. We selected the variance cutoff of 0.001 and the intercorrelation cutoff of 0.95 for the pretreatment.

The selection of the best division and best mixing rules were performed based on the statistical quality of preliminary Genetic algorithm–based multiple linear regression (GA-MLR) models. The "Double Cross-Validation" Version 2.0 tool (Roy and Ambure, 2016)

(available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for developing the preliminary GA-MLRs and the quality metrics (Q^2_{LOO} and Q^2_{F1}) have been tabulated in **Table 3.22**. In Double cross-validation, the training set is further divided into calibration and validation sets; the important descriptors are extracted by Genetic algorithm from the calibration set and modeled using MLR and then further validated with the validation set. Finally, those extracted descriptors are used for modeling the actual training set and then validated with the test set. For finding the best mixing rule, we calculated the mean of cross-validated correlation coefficients (Q^2_{LOO}) (which are based on solely the training sets) obtained from models of 9 different divisions. It was observed that mixing rule 2 (the squared molecular contributions), mixing rule 1 (linear combination of molecular contributions), and mixing rule 3 (norm of molecular contributions) produce the highest mean Q^2_{LOO} values for MD1 ($\overline{Q^2_{LOO}}$=0.920), MD2 ($\overline{Q^2_{LOO}}$=0.919), and MD3 ($\overline{Q^2_{LOO}}$=0.810), respectively; therefore these were selected as best mixing rules for MD1, MD2, and MD3 for further analysis. Besides the mixing rules identification, we also selected the best division based on the overall performance of preliminary GA-MLR in terms of balanced Q^2_{LOO} and Q^2_{F1} values. Two different seeds of Euclidean distance-based divisions were selected for MD1 (4-2 DBD) and MD3 (4-3 DBD), whereas, a sorted response-based division was selected for MD2 (4-1 ABD), and we proceeded further with the selected divisions.

Table 3.22. Preliminary GA-MLR results for three datasets

Mixture Dataset 1 (*Vibrio fischeri*)			
Mixing rule 1			
Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.896	0.816	6
4-1 DBD	0.864	0.899	6
4-2 ABD	0.880	0.882	6
4-2 DBD	0.888	0.839	6
4-3 ABD	0.904	0.813	6
4-3 DBD	0.891	0.896	6
4-4 ABD	0.881	0.882	6
4-4 DBD	0.902	0.818	6
75-25 KS	0.884	0.843	6
MEAN	0.888	0.854	
Mixing rule 2			
Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.912	0.907	6
4-1 DBD	0.910	0.844	6

4-2 ABD	0.910	0.932	6
4-2 DBD	**0.912**	**0.940**	**6**
4-3 ABD	0.929	0.891	6
4-3 DBD	0.926	0.853	6
4-4 ABD	0.933	0.844	6
4-4 DBD	0.910	0.899	6
75-25 KS	0.934	0.807	6
MEAN	0.920	0.880	

Mixing rule 3

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.900	0.865	6
4-1 DBD	0.907	0.867	6
4-2 ABD	0.892	0.886	6
4-2 DBD	0.903	0.895	6
4-3 ABD	0.907	0.840	6
4-3 DBD	0.915	0.844	6
4-4 ABD	0.905	0.887	6
4-4 DBD	0.895	0.835	6
75-25 KS	0.916	0.827	6
MEAN	0.904	0.861	

Mixture Dataset 2 (*Escherichia coli*)

Mixing rule 1

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	**0.915**	**0.919**	**6**
4-1 DBD	0.924	0.886	6
4-2 ABD	0.913	0.933	6
4-2 DBD	0.917	0.913	6
4-3 ABD	0.919	0.873	6
4-3 DBD	0.916	0.907	6
4-4 ABD	0.911	0.939	6

4-4 DBD	0.916	0.912	6
75-25 KS	0.943	0.873	6
MEAN	0.919	0.906	

Mixing rule 2

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.909	0.939	6
4-1 DBD	0.915	0.887	6
4-2 ABD	0.908	0.938	6
4-2 DBD	0.919	0.893	6
4-3 ABD	0.927	0.871	6
4-3 DBD	0.901	0.918	6
4-4 ABD	0.901	0.931	6
4-4 DBD	0.918	0.888	6
75-25 KS	0.921	0.832	6
MEAN	0.913	0.900	

Mixing rule 3

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.901	0.932	6
4-1 DBD	0.902	0.876	6
4-2 ABD	0.909	0.932	6
4-2 DBD	0.913	0.888	6
4-3 ABD	0.911	0.860	6
4-3 DBD	0.921	0.867	6
4-4 ABD	0.896	0.928	6
4-4 DBD	0.898	0.912	6
75-25 KS	0.933	0.842	6
MEAN	0.909	0.893	

Mixture Dataset 3 (*Bacillus subtilis*)

Mixing rule 1

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**

4-1 ABD	0.439	-0.259	5
4-1 DBD	0.466	-0.252	5
4-2 ABD	0.338	0.350	5
4-2 DBD	0.433	0.084	5
4-3 ABD	0.462	-1.476	5
4-3 DBD	0.417	0.247	5
4-4 ABD	0.468	-0.636	5
4-4 DBD	0.333	0.302	5
75-25 KS	0.490	-0.033	5
MEAN	0.419	-0.205	

Mixing rule 2

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.675	0.106	5
4-1 DBD	0.796	0.167	5
4-2 ABD	0.625	0.451	5
4-2 DBD	0.677	0.647	5
4-3 ABD	0.693	0.380	5
4-3 DBD	0.785	0.709	5
4-4 ABD	0.855	0.320	5
4-4 DBD	0.696	0.592	5
75-25 KS	0.776	0.789	5
MEAN	0.725	0.422	

Mixing rule 3

Division	**Q^2_{LOO}**	**Q^2_{F1}**	**No. of Descriptors**
4-1 ABD	0.794	0.650	5
4-1 DBD	0.775	0.830	5
4-2 ABD	0.782	0.779	5
4-2 DBD	0.762	0.741	5
4-3 ABD	0.784	0.849	5
4-3 DBD	**0.787**	**0.805**	**5**

4-4 ABD	0.870	0.352	5
4-4 DBD	0.872	0.138	5
75-25 KS	0.868	0.609	5
MEAN	0.803	0.643	

3.5.4. Multilayered descriptor selection and QSAR/QSAAR model development

In this work, we performed a multi-layered feature selection technique to extract valuable descriptors for modeling. The initial pool reduction was commenced by stepwise selection (using "MINITAB" version 14 software ("Data Analysis, Statistical & Process Improvement Tools | Minitab," 2004) in multiple iterations) and genetic algorithm (GA) (using the "Genetic algorithm: version 4.1 tool; available from http://teqip.jdvu.ac.in/QSAR_Tools/). Finally, the most significant descriptors were extracted by checking all possible subsets from the reduced pool of descriptors by employing the "BestSubsetSelection" version 2.1 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). From the extracted descriptors of MD1, MD2, and MD3, three partial least squares (PLS) regression models (QSAR) were developed by employing the "PLS_single_Y" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) in a successive step.

To evaluate the impact of environmental degradation if any on the mixture toxicity, we incorporated the dissociation constant (pKa) of antibiotics for modeling. The pKa (strongly acidic and strongly basic) data of 13 antibiotics were collected from Drug Bank Database (available from: https://go.drugbank.com/), and pKa (strongly acidic and strongly basic) data of 3 antibiotics (sulfamonomethoxine, diaveridine, and ormetoprim) were computed from "MarvinSketch" software (ChemAxon, 2018). The pKa data used for modeling has been provided in **Table 3.23**. From the pKa data of individual components, we computed the mixture descriptors using three mixing rules discussed in the "Mixture descriptor calculation considering three different mixing rules" section. Finally, another three QSAR models were developed using PLS regression with the identified descriptors in previous steps and the pKa (strongly acidic and strongly basic) descriptors from three datasets.

Table 3.23. Collected and computed pKa values of individual mixture components

Sl. No.	Name	CAS No.	pKa (Strong acidic)	pKa (Strong basic)
1	Sulfamethazine	57-68-1	6.99	2
2	Sulfameter	651-06-9	7.06	1.98
3	Sulfamethoxypyridazine	80-35-3	6.84	2.02
4	Sulfisoxazole	127-69-5	5.8	2.17
5	Sulfadoxine	2447-57-6	6.12	3.44
6	Sulfamethoxazole	723-46-6	6.16	1.97
7	Sulfadiazine	68-35-9	6.99	2.01
8	Sulfamonomethoxine*	1220-83-3	7.15	1.76

9	Trimethoprim	738-70-5	17.33	7.16
10	Diaveridine*	5355-16-8	17.34	7.16
11	Ormetoprim*	6981-18-6	17.34	7.16
12	Pyrimethamine	58-14-0	17.22	7.77
13	Doxycycline hyclate	24390-14-5	3.27	8.33
14	Tetracycline hydrochloride	64-75-5	3.26	9.25
15	Chlortetracycline hydrochloride	64-72-2	2.99	9.04
16	Oxytetracycline hydrochloride	2058-46-0	2.84	7.41

*pKa value has been computed from MarvinSketch software

For the development of interspecies QSAAR models, common data points for two toxicity endpoints are needed. But in this study, there were no common existing mixture data points which can be used for this purpose. Hence, we employed the predicted toxicity values obtained from the developed QSAR models in this study as predictors along with additional structural descriptors. Two QSAAR models were developed for each dataset; therefore, a total of 6 interspecies QSAARs, i.e., (i) QSAAR of *V. fischeri* with the predicted toxicity of *E. coli* as a predictor, (ii) QSAAR of *V. fischeri* with the predicted toxicity of *B. subtilis* as a predictor, (iii) QSAAR of *E. coli* with the predicted toxicity of *V. fischeri* as a predictor, (iv) QSAAR of *E. coli* with the predicted toxicity of *B. subtilis* as a predictor, (v) QSAAR of *B. subtilis* with the predicted toxicity of *V. fischeri* as a predictor, and (vi) QSAAR of *B. subtilis* with the predicted toxicity of *E. coli* as a predictor were obtained. Feature selection was performed in a manner similar to the previous QSAR modeling. The extracted descriptors along with the predicted toxicity of other species were modeled against the corresponding endpoint using PLS regression by employing the "PLS_single_Y" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) in all 6 cases. The general workflow has been schematically represented in the following **Figure 3.8**.

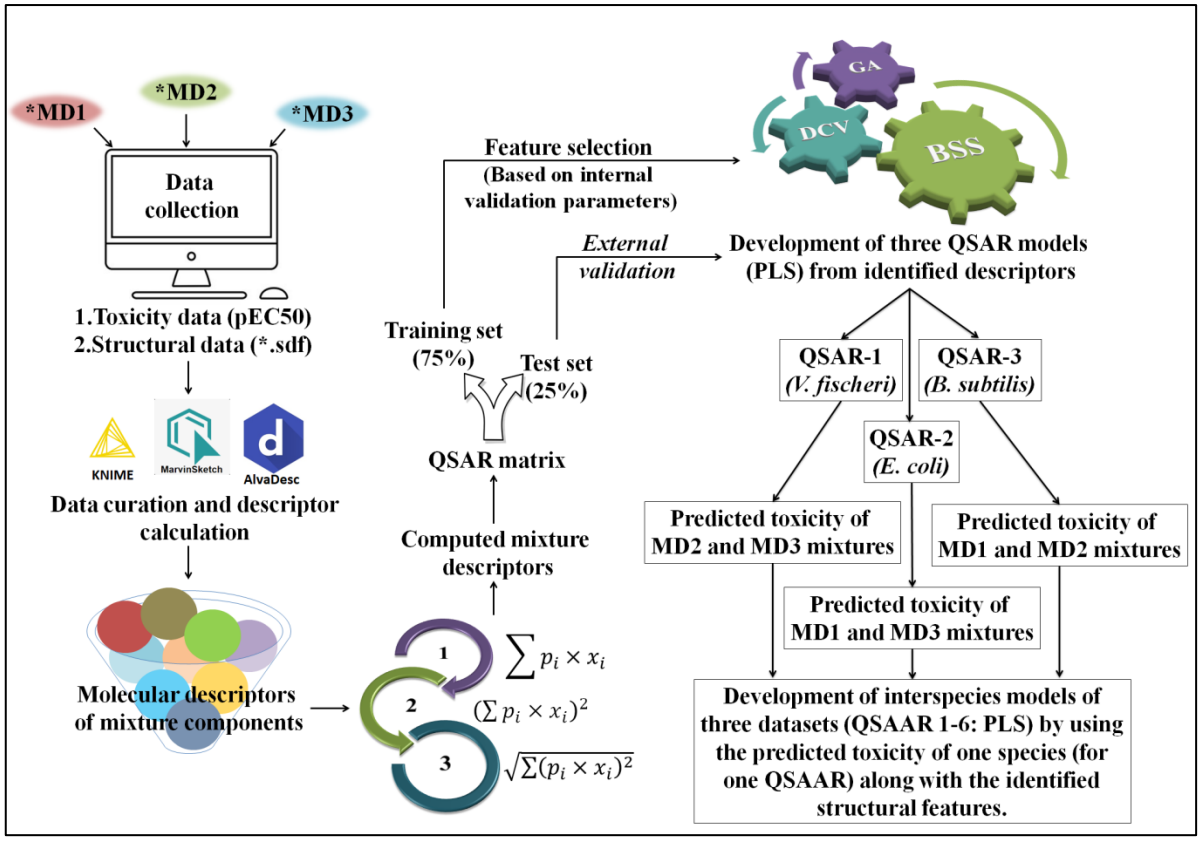


Figure 3.8. General QSAR development procedures used in this study; (*MD1: mixture dataset 1, comprised of toxicity data 76 mixtures against *V. fischeri*; MD2: mixture dataset 2, comprised of toxicity data 67 mixtures against *E. coli*; MD3: mixture dataset 3, comprised of toxicity data 42 mixtures against *B. subtilis*)

3.5.5. Application of q-RASAR in QSAAR development

We applied the concept of q-RASAR in this study to incorporate the biological descriptors in terms of the predicted toxicity of other organisms during QSAAR modeling. Three different RASAR-based QSAAR models were developed in this study from three mixture datasets. The descriptors obtained during preliminary PLS-QSAAR models were clubbed together into 3 sets (unique descriptors of QSAAR 1 and 2 are used for the RASAR descriptor calculation of MD1; unique descriptors of QSAAR 3 and 4 are used for the RASAR descriptor calculation of MD2; unique descriptors of QSAAR 5 and 6 are used for the RASAR descriptor calculation of MD3) for the RASAR descriptor calculation (**Table 3.24**).

Table 3.24. Features used for RASAR descriptor calculation

Dataset	Preliminary PLS-QSAAR models	Number of descriptors in preliminary PLS-QSAARs	Number of clubbed features used for RASAR descriptor calculation	Clubbed features used for RASAR descriptor calculation
MD1 (Organism: *V. fischeri*)	QSAAR of *V. fischeri* with predicted toxicity of *E. coli* as a predictor	6	11	C-032, nCL, nHDon, minssO, F03[C-S], F04[C-Cl], C-029, ALOGP2, SAacc, Predicted toxicity (*E. coli*). Predicted toxicity (*B. subtilis*)
	QSAAR of *V. fischeri* with predicted toxicity of *B. subtilis* as a predictor	6		
MD2 (Organism: *E. coli*)	QSAAR of *E. coli* with predicted toxicity of *V. fischeri* as a predictor	5	8	B06[O-O], F02[N-O], F04[C-N], ALOGP2, B05[N-O], F08[C-N], Predicted toxicity (*V. fischeri*), Predicted toxicity (*B. subtilis*)
	QSAAR of *E. coli* with predicted toxicity of *B. subtilis* as a predictor	5		
MD3 (Organism: *B. subtilis*)	QSAAR of *B. subtilis* with predicted toxicity of *V. fischeri* as a predictor	5	10	F02[C-N], nHM, nAA, F04[N-N], T(N..Cl), ALOGP2, nOHt, minaaCH, Predicted toxicity (V. fischeri), Predicted toxicity (E. coli)
	QSAAR of *B. subtilis* with predicted toxicity of *E. coli* as a predictor	5		

The RASAR descriptors are the read-across derived similarity, error, concordance, and RA-based prediction functions which may be clubbed together with additional structural and physicochemical descriptors to form the RASAR model (Banerjee and Roy, 2022). Read-across predictions and corresponding hyperparameter optimization are the prerequisites of RASAR descriptor calculation. Read-across is a similarity- based prediction technique where the exerted response of any unknown compound (target) can be predicted from the known response values of structural analogues (source). Chatterjee et al. have introduced a novel read-across prediction workflow in a previous study where they used. Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity estimation techniques for identifying the close structural analogues against query compounds (Chatterjee et al., 2022). The hyperparameters sigma (σ) and gamma (γ) are the radial basis functions used in Gaussian and Laplacian kernel similarity estimations respectively. The number of close-training compounds is another hyperparameter which also affects the prediction quality of all three techniques. In this study, we optimized the value of sigma, gamma, and the number of close training compounds using 5-fold cross-validation by keeping 25% of training data out in the validation sets (different from the test sets). Different sigma values (0.25 to 2), different gamma values (0.25 to 2), and 2 to 10 number of close training compounds were used for this optimization (based on the quality of predictions in terms of Q^2_{F1} value for a validation set). This computation was performed using an in-house tool "Auto RA Optimizer" version 1.0 and the optimized settings for three mixture datasets have been reported in **Table 3.25**.

Table 3.25. Optimized setting for RASAR descriptor calculation

Dataset	Optimized RA method	Optimized setting
MD1	Laplacian kernel function-based read-across	Gamma = 1, Number of Close training compounds = 3, Similarity threshold = 0
MD2	Euclidean distance-based read-across	Number of close training compounds = 9, Distance threshold = 1
MD3	Laplacian kernel function-based read-across	Gamma = 1.5, Number of Close training compounds = 5, Similarity threshold = 0

Finally, the RASAR descriptors were calculated using the optimized hyperparameters by employing "RASAR Descriptor Calculator" version 2.0 (available from:http://teqip.jdvu.ac.in/QSAR_Tools/). The training set itself was used as both source and target data for the RASAR descriptor calculation of the training sets, whereas, the training and test sets were used as source and target data respectively for the RASAR descriptor calculation of the test sets. There are chances of overfitting during the RASAR descriptor calculation of the training set because the same training set is used as both source and target data. To avoid this, the "leave-same-out" strategy was used where the same training data point has been removed from the close training compounds' list. After the RASAR descriptor calculation, these were clubbed together with structural and physicochemical descriptors of corresponding datasets (MD1, MD2, and MD3) and then the best descriptor combinations were obtained by using the "BestSubsetSelection" version 2.1

tool. In a successive step, the final RASAR models for MD1 and MD3 were developed using single descriptors by MLR, whereas, the RASAR model of MD2 was developed with 4 descriptors and 3 latent variables using PLS regression. In the current work, the RASAR descriptors were derived by "data fusion" between structural/physicochemical descriptors with toxicity values for other endpoints. Thus, here, the q-RASAR models may more appropriately be termed as q-RASAAR models. We have schematically represented the specific RASAR development procedure (q-RASAAR-2 has been used as an example for the scheme) in **Figure 3.9**.

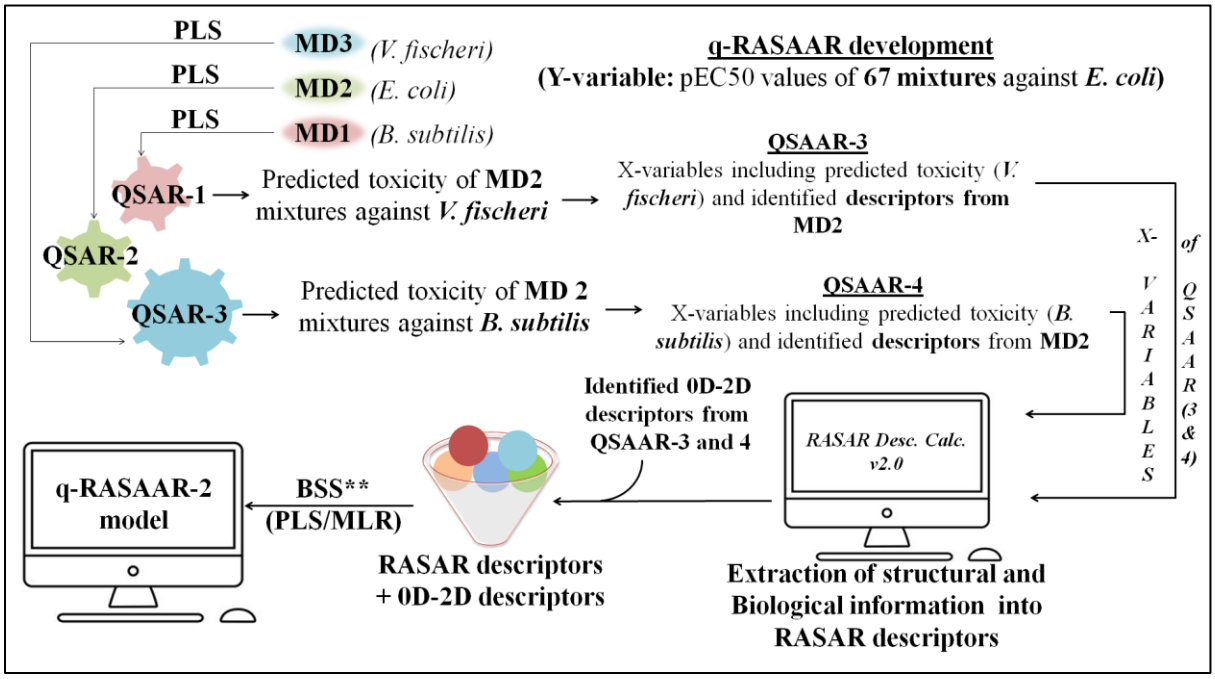


Figure 3.9. Specific q-RASAAR development procedures used in this study (q-RASAAR-2 has been taken as an example for creating the scheme)(**BSS: best subset selection)

3.5.6. Statistical model validation

Statistical validation is one of the essential aspects of any predictive operation because it establishes the reliability and relevance of the predicted response obtained by the developed model (Roy, 2007). The quality of QSAR models should be accurately measured in terms of the goodness of fit, robustness, and predictability, as suggested by OECD's Principle 4 (OECD, 2007, 2004). In the conventional practice of QSAR modeling, rigorous internal and external validations are carried out with training and test datasets respectively, for evaluating the performance of the model. In this study, we computed different internal quality and validation metrics such as determination coefficient (R^2), adjusted determination coefficient (R^2_{adj}), variance ratio (F with degrees of freedom), leave-one-out (LOO) cross-validated correlation coefficient (Q^2_{LOO}), mean absolute error of training set predictions (MAE_{Train}), etc. and external validation metrics such as external correlation coefficients (Q^2_{F1}, Q^2_{F2}), concordance correlation coefficient (CCC), mean absolute error of test set predictions (MAE_{Test}) to adjudge whether the developed models are satisfying the acceptability criteria or not. There are certain thresholds for both the internal and external validation parameters; i.e., R^2 and R^2_{adj} should have computed values more than 0.6, whereas Q^2_{LOO}, Q^2_{F1}, and Q^2_{F2} should have computed values more than 0.5, and CCC should have a computed value more than 0.85 for acceptable models. The valiance ratio should be statistically significant in 95%

confidence level at the corresponding degrees of freedom. The values of Q^2_{LOO} and Q^2_{F1} or Q^2_{F2} close to near 1 may not be the prominent indicators of a good model; therefore, the r^2_m metrics ($\overline{r^2_{m(Train)}}$, $\Delta r^2_{m(Train)}$, $\overline{r^2_{m(Test)}}$, $\Delta r^2_{m(Test)}$) were additionally calculated to better reflect the quality of predictions (Roy et al., 2012). Roy et al. have suggested the thresholds for r^2_m metrics; the $\overline{r^2_{m(Train)}}$ and $\overline{r^2_{m(Test)}}$ should have computed values greater than 0.5, whereas, for $\Delta r^2_{m(Train)}$ and $\Delta r^2_{m(Test)}$ should have computed values less than 0.2 for acceptable models (Roy et al., 2012). Besides these, all the error measures should have minimum computed values. The Y-randomization test (Topliss and Edwards, 1979) was another validation test which confirm whether the model is built by any chance correlation or not. It is carried out by 100 times random shuffling of the response variable (pEC50) of the training data using "SIMCA-P software" (Wu et al., 2010). 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. The Akaike information criteria (AIC) values were also computed for the QSARs and q-RASAARs in this study and compared. This metric estimates the prediction error of the models developed from a single dataset and thereby, judges the relative efficiency of the models. The AIC was named after a Japanese scientist Hirotsugu Akaike who formulated it. For a regression model AIC is computed by employing the following equation 3.12 (Burnham and Anderson, 2002):

$$AIC = n \times log(\sigma^2) + 2 \times K \qquad (3.12)$$

where, n = number of training data, K = number of descriptors, and σ^2 = mean sum squared residual. The least AIC value of a model is the indicator of more relative efficiency among various models developed from a dataset.

3.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach

3.6.1. Toxicity data collection

We have collected a toxicity dataset of 78 binary fluoroquinolone mixtures from the recently published article of Wang et al. and have used those for toxicity modeling (Wang et al., 2020). The Log (1/IC50) values (in molar scale) have been taken as the endpoint of this study where IC50 is defined as the median inhibitory concentration of binary mixtures in *E. coli*. The mixtures were prepared from 13 different fluoroquinolone antibiotics namely ciprofloxacin (CIP), danofloxacin (DAN), enoxacin (ENO), enrofloxacin (ENR), levofloxacin (LEV), marbofloxacin (MAR), moxifloxacin (MOX), norfloxacin (NOR), ofloxacin (OFL), orbifloxacin (ORB), pefloxacin (PEF), sarafloxacin (SAR), and sparfloxacin (SPA). All of these component chemicals have been tabulated along with their 2D structures and experimental toxicity data [Log (1/IC50)] in **Table 3.26**. According to the report of Wang et al., equitoxic amounts of component chemicals were mixed to obtain the desired mixtures and then used for generating the toxicity data; we have tabulated the compositions (same as collected) of those 78 binary mixtures in **Table 3.27**. The IC50s of the studied mixtures were determined by a logistically fitted concentration-inhibition curve. The growth inhibition of *E. coli* was determined by comparing the optical density of inoculated lysogeny broth (LB) before and after the addition of sample mixtures. The bacterial strain of *E. coli* was inoculated in sterile LB medium and cultured at 37ºC for toxicity assay. The optical density of the LB medium was tested using a Multiskan™ FC Microplate Photometer (Thermo-Fisher Scientific, USA).

Table 3.26. Toxicity data [Log (1/IC50) in molar scale] of mixture components (fluoroquinolones) along with their 2D structures

Serial No.	Name (Abbreviation)	CAS No.	Experimental toxicity [Log (1/IC50)]
C1	Ciprofloxacin (CIP)	85721-33-1	7.30
C2	Danofloxacin (DAN)	112398-08-0	6.90
C3	Enoxacin (ENO)	74011-58-8	6.13
C4	Enrofloxacin (ENR)	93106-60-6	6.98
C5	Levofloxacin (LEV)	100986-85-4	6.42
C6	Marbofloxacin (MAR)	115550-35-1	7.27
C7	Moxifloxacin (MOX)	151096-09-2	6.71
C8	Norfloxacin (NOR)	70458-96-7	6.18
C9	Ofloxacin (OFL)	82419-36-1	6.74
C10	Orbifloxacin (ORB)	113617-63-3	6.96
C11	Pefloxacin (PEF)	70458-92-3	6.25
C12	Sarafloxacin (SAR)	98105-99-8	7.10
C13	Sparfloxacin (SPA)	110871-86-8	7.32

Table 3.27. Composition and experimental toxicity data [Log (1/IC50) in molar scale] of studied mixtures

Mixture No.	Comp. A	Mole Fraction	Comp. B	Mole Fraction	Experimental toxicity [Log (1/IC50)]
M1	C1	0.28	C2	0.72	7.05
M2*	C1	0.06	C3	0.94	6.43
M3*	C1	0.32	C4	0.68	7.27
M4	C1	0.12	C5	0.88	6.71
M5	C1	0.48	C6	0.52	7.28
M6	C1	0.2	C7	0.8	6.84

M7	C1	0.07	C8	0.93	6.34
M8*	C1	0.22	C9	0.78	6.99
M9	C1	0.31	C10	0.69	7.27
M10	C1	0.08	C11	0.92	6.56
M11*	C1	0.39	C12	0.61	7.2
M12	C1	0.51	C13	0.49	7.4
M13*	C2	0.15	C3	0.85	6.51
M14	C2	0.55	C4	0.45	7
M15	C2	0.25	C5	0.75	6.64
M16*	C2	0.7	C6	0.3	7.02
M17*	C2	0.39	C7	0.61	6.96
M18	C2	0.16	C8	0.84	6.48
M19*	C2	0.41	C9	0.59	6.78
M20	C2	0.53	C10	0.47	7.18
M21	C2	0.18	C11	0.82	6.52
M22	C2	0.61	C12	0.39	7.27
M23	C2	0.72	C13	0.28	7
M24	C3	0.88	C4	0.12	6.5
M25	C3	0.66	C5	0.34	6.47
M26	C3	0.93	C6	0.07	6.42
M27	C3	0.79	C7	0.21	6.25
M28	C3	0.53	C8	0.47	6.16
M29*	C3	0.8	C9	0.2	6.37
M30	C3	0.87	C10	0.13	6.39
M31	C3	0.57	C11	0.43	6.39
M32	C3	0.9	C12	0.1	6.47
M33	C3	0.94	C13	0.06	6.41

M34*	C4	0.22	C5	0.78	6.82
M35	C4	0.66	C6	0.34	7.05
M36*	C4	0.35	C7	0.65	6.87
M37	C4	0.14	C8	0.86	6.54
M38	C4	0.37	C9	0.63	6.98
M39	C4	0.49	C10	0.51	6.93
M40	C4	0.16	C11	0.84	6.8
M41	C4	0.57	C12	0.43	7.2
M42	C4	0.69	C13	0.31	7.21
M43	C5	0.88	C6	0.12	6.52
M44	C5	0.66	C7	0.34	6.78
M45*	C5	0.37	C8	0.63	6.24
M46*	C5	0.68	C9	0.32	6.46
M47	C5	0.78	C10	0.22	6.77
M48	C5	0.4	C11	0.6	6.39
M49	C5	0.83	C12	0.17	6.49
M50*	C5	0.89	C13	0.11	6.54
M51	C6	0.22	C7	0.78	6.94
M52	C6	0.08	C8	0.92	6.36
M53	C6	0.23	C9	0.77	6.98
M54	C6	0.33	C10	0.67	7.06
M55	C6	0.09	C11	0.91	6.86
M56	C6	0.4	C12	0.6	7.17
M57	C6	0.53	C13	0.47	7.24
M58	C7	0.23	C8	0.77	6.43
M59	C7	0.52	C9	0.48	6.69
M60	C7	0.64	C10	0.36	6.94

M61*	C7	0.26	C11	0.74	6.62
M62	C7	0.71	C12	0.29	6.71
M63*	C7	0.8	C13	0.2	6.7
M64	C8	0.78	C9	0.22	6.43
M65*	C8	0.86	C10	0.14	6.48
M66	C8	0.54	C11	0.46	6.4
M67	C8	0.89	C12	0.11	6.44
M68*	C8	0.93	C13	0.07	6.39
M69	C9	0.62	C10	0.38	6.78
M70	C9	0.24	C11	0.76	6.57
M71	C9	0.7	C12	0.3	6.85
M72	C9	0.79	C13	0.21	6.98
M73	C10	0.16	C11	0.84	6.6
M74	C10	0.58	C12	0.42	7.01
M75	C10	0.7	C13	0.3	7.09
M76	C11	0.88	C12	0.12	6.5
M77	C11	0.92	C13	0.08	6.63
M78*	C12	0.62	C13	0.38	7.07

*Test set mixtures of modeling division

3.6.2. Mixture descriptor calculation

In any mixture-QSAR model, each mixture is considered as an individual data point represented by the unique structural features that are numerically encoded in the mixture descriptors. Generally, molecular descriptors are defined as "*the final result of a logical 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*"; (Consonni and Todeschini, 2010) however, mixture descriptors are different from the individual molecular descriptors. Computation of mixture descriptor is a two-step process where the molecular descriptors of the mixture components are computed first. In the next step, the mixture descriptors are calculated from the structural information of the component chemicals and their corresponding concentration or mass ratio. Different mixing rules are found in the literature for calculating the mixture descriptors among which linear combination of molecular contributions is mostly used. In linear combination, the mixture effect is assumed to be additive; however, the mixture components may also produce supra-

additive/sub-additive responses due to biological interactions namely, synergism and antagonism respectively. Qin and coworkers (Qin et al., 2018) have enlisted numerous mixing rules apart from the linear combination of molecular contributions among which the norm of molecular contributions (Gaudin et al., 2015) and square molecular contributions (Saldana et al., 2013) are most prominent. Therefore, to consider the probable additive, supra-additive, and sub-additive mixture effects, we have employed the linear combination of molar contributions, the square molar contributions, and the norm of molar contributions respectively in the present work. These three mixing rules have been presented in the previous section "***3.4.2. Mixture descriptors***" in detail. We have used only 0D-2D descriptors to make it simple and easily interpretable. Nine classes of 0D-2D descriptors namely constitutional indices, molecular properties, ring descriptors, connectivity index, electro-topochemical atom indices, functional group counts, atom-centred fragments, atom-type E-state indices, and 2D atom pairs (total descriptor count = 2400) have been computed from the AlvaDesc software (Mauri, 2020) and accordingly pretreated to omit the redundant descriptors (descriptors with same values, descriptors with strong inter-correlation). Finally, three classes of pretreated mixture descriptors (363 descriptors in each class) have been obtained by using a Java-based software tool "Mixture-Desc-Calc v1.0" (freely available online: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1).

3.6.3. Dataset division and optimization of mixing rule

Ideally, the QSAR models should be developed with the available experimental data and then validated with an external experimental dataset. However, the availability of such experimental data is very limited, and therefore, the complete dataset gets divided into training and test sets to serve the purpose. The training data points are used for model development, whereas, the test data points are separately employed for external validation. Here in this study, we have developed QSAR models using 75% of the total data, and the rest has been used to test the developed model (as test set). There are many reported research in the literature where well-validated QSAR models were obtained with 75:25 training-test divisions (Chatterjee and Roy, 2023; Ghosh et al., 2023; Paul et al., 2022). Both the response-based (sorted-activity-based division) and descriptor-based (Euclidean distance-based division and Kennard-Stone division) division techniques have been applied to get the optimum division. The division has been performed using the "Dataset Division GUI" version 1.2 tool (freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/) and nine different training-test division sets have been obtained (4 seeds of sorted activity-based divisions, 4 seeds of Euclidean distance-based divisions, and 1 Kennard-Stone division) for each descriptor matrix; therefore a total of 27 division sets were generated.

For optimizing the mixing rule and selecting the best division set, we have employed the "Double Cross-Validation v2.0" (DCV) tool(Roy and Ambure, 2016) (freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/) and have developed preliminary genetic algorithm-based multiple linear regression (GA-MLR) models from each of the 27 training sets for respective mixing rule. Another round of data pretreatment has been carried out (using "dataPreTreatmentTrainTest v1.0" tool; freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/) before starting the preliminary GA-MLR modeling for removing the redundant descriptors generated after division. The preliminary models have been validated in the same tool ("Double Cross-Validation v2.0") and validation metrics such as R^2, Q^2_{LOO}, $MAE_{Train95\%}$, Q^2_{F1}, and $MAE_{Test95\%}$ have been calculated. To optimize the best mixing rule, the averages of the above-mentioned metrics have been calculated and based on the average values, the optimum mixing rule has been selected. After optimization of the

mixing rule, we have compared the balance of internal and external validation metrics of different division sets under the optimized mixing rule and based on that, the best division was selected.

3.6.4. Feature selection and model development

A multi-layered algorithm has been adopted in this study for extracting valuable structural features for modeling. The most suitable descriptor combination has been finalized by comparing the determination coefficient (R^2) and cross-validated correlation coefficient (Q^2_{LOO}) values of all possible subsets of descriptors. Generally in the regression-based model, a 5:1 ratio is maintained between the number of training data points and the number of modeled descriptors; the same rule has been followed here too and the combinations with minimum descriptors have been prioritized. We have generated all possible subsets from the descriptor pool using "BestSubsetSelection v2.1" tool. However, all possible subset selection is computationally very exhaustive and it needs a lot of time to compute the validation metrics of the MLR models developed from the descriptor combinations. Thus, a pool reduction is carried out (from 138 descriptors to 40 descriptors) before all possible subset selections. The initial pool reduction has been commenced by genetic algorithm (GA) (using the "Genetic algorithm v4.1" tool; available from http://teqip.jdvu.ac.in/QSAR_Tools/) in this study. The important descriptors have been taken into the reduced pool based on the R^2 and Q^2_{LOO} values of the incoming descriptor combinations.

After selecting the best descriptor combination, we have performed the final regression analysis by PLS algorithm by employing the "PLS_single_Y v1.0" tool (freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/). The PLS is the generalized form of MLR that can handle strongly correlated and colinear data too (Wold et al., 2001). In the PLS algorithm, latent variables (LVs) are computed from the descriptors and they take part in the regression analysis, unlike MLR. We have optimized the number of latent variables (LVs) by comparing the calculated x-variances after incorporating each incoming LVs. Finally, a PLS model has been obtained with the optimum number of LVs from the selected descriptors.

3.6.5. Machine-learning models

Machine learning (ML) is the Science of programming computers so that they can learn from the existing data and use that knowledge to solve future assignments more efficiently.(Vamathevan et al., 2019) Its applications span across different sectors including regulatory decision-making and *in silico* predictive modeling also. The computers are trained by the significant patterns of large data against a specific learning algorithm and then use that learning experience for predicting the corresponding properties of query compounds. Based on the learning algorithm, ML methods are categorized into three major classes namely supervised learning, unsupervised learning, and reinforcement. Supervised and unsupervised ML are trained by labeled and unlabeled data respectively (Jordan and Mitchell, 2015). Reinforcement is different from these two and it is trained by the feedback (either reward or penalty) of performed action (Jordan and Mitchell, 2015; Sarker, 2021). In the present study, we have applied supervised machine learning by employing algorithms like random forest (RF),(Chase et al., 2023) ada boost,(Geron, 2019) gradient boosting (GB),(Geron, 2019) extreme gradient boosting (XGB),(Chen and Guestrin, 2016) linear support vector machine (LSVM),(Burbidge et al., 2001) support vector machine (SVM),(Chase et al., 2023) and ridge regression (RR),(Geron, 2019; McDonald, 2009) to develop a regression model.

Generally, different models are trained by a specific ML algorithm for a dataset and the best one is selected based on the performance. However, there may be some possibilities for further improvement of the model and to do so, tuning of hyperparameters is carried out. In this study, we have employed our in-house Python-based tool "Optimization and Cross-validation v1.0" (available online: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1) for tuning the hyperparameters for all the ML algorithms. The GridSearhCV and negative mean squared error have been used as the optimizing algorithm and the scoring function respectively in this study. The same training set containing the scaled matrix of previously identified descriptors and endpoint values has been used for hyperparameter optimization. After obtaining the optimized setting, the ML models have been developed using the Python-based "Machine Learning Regression v1.0" tool (available online: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1). For obtaining further details of the ML method used, one may consult the article published by Pore *et al.*(Pore et al., 2023) The same standardized training and test sets have been used for all ML-based models [standardized/scaled value = (original value-mean of the respective column)/standard deviation of the same column; training set means have been used for the scaling of both training and test sets] as used for the PLS model development. The entire modeling process has been schematically represented in the following **Figure 3.10**.

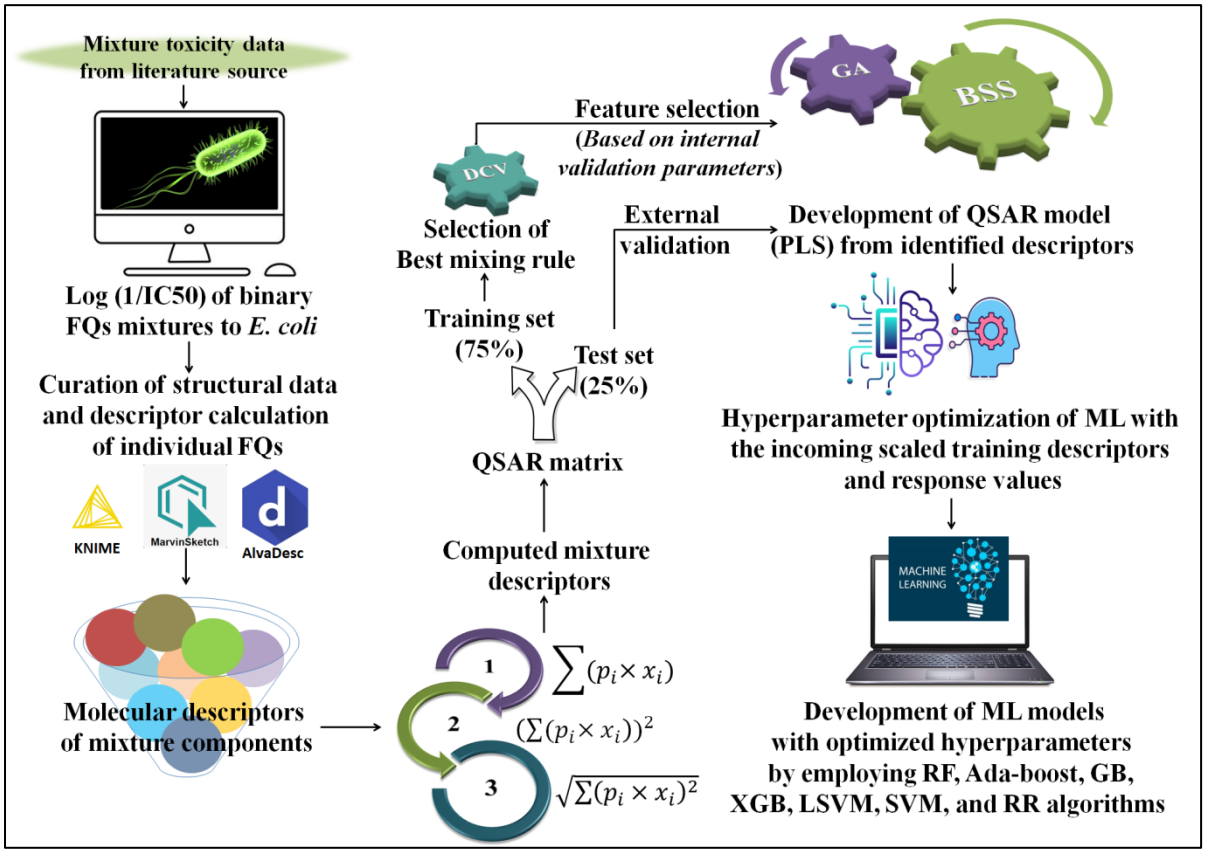


Figure 3.10. Schematic representation of the entire modeling process [*FQs = fluoroquinolones, DCV = double cross-validation, GA = genetic algorithm, BSS = best subset selection, ML = machine learning, PLS = partial least squares, RF = random forest, GB = gradient boost, XGB = extreme gradient boost, LSVM = linear support vector machine, SVM = support vector machine (non-linear), RR = ridge regression]

3.6.6. Statistical validation

Statistical validation is an inevitable part of any predictive modeling. As per the OECD guideline 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 coefficients (R^2 and/or R^2_{Adj}) are the measurement parameters of goodness of fit, 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 cross-validated correlation coefficient (Q^2_{LOO}) is measured for adjudging the robustness of the statistical model. The difference between Q^2_{LOO} and R^2 is a marker of robustness and this difference should ideally be a smaller one for a robust 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).

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. The 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.

3.6.7. Toxicity prediction of binary mixture dataset designed from parent FQs and their degradates: Use of the PRI tool

As found in the literature, the chemical species get degraded under open sunlight and other environmental stresses (Sturini et al., 2015). There are many instances where photodegradation/chemical degradation of FQs has been reported in the literature.(Rusch et al., 2019; Zhang and Huang, 2005) As we are mostly focused on the environmental aspects of FQs, we have concentrated on the photodegradation products only. Sturini et al., Babic et al., Ye et al., and Ou et al. reported different photodegradation pathways of FQs which have been graphically represented in **Figure 3.11** by taking ciprofloxacin as a prototype compound. (Babić et al., 2013; Ou et al., 2016; Sturini et al., 2015; Ye et al., 2016). Reductive halogenation at the 6^{th} position, hydroxy substitution at the 6^{th} position, and hydroxy addition at the 2^{nd} position are the important transformation reactions that may take place under UV light in the quinolone ring whereas, H-elimination, oxidation, and photolysis are the important transformation reaction that may take place at the ring attached at 7^{th} position of quinolone. We have not considered any reaction where the structural integrity of the core quinolone ring of FQ (important for the toxic response of FQs) is compromised. We have designed a hypothetical mixture dataset from nine parent FQs (ciprofloxacin, danofloxacin, enoxacin, enrofloxacin, levofloxacin, marbofloxacin, norfloxacin, pefloxacin, and sarafloxacin) and their corresponding degradation products as the component chemicals of

50:50 binary mixtures. All possible combinations of components (45 mixtures) have been considered for each FQs. The composition of all the hypothetical mixtures has been tabulated in **Table 3.34**.

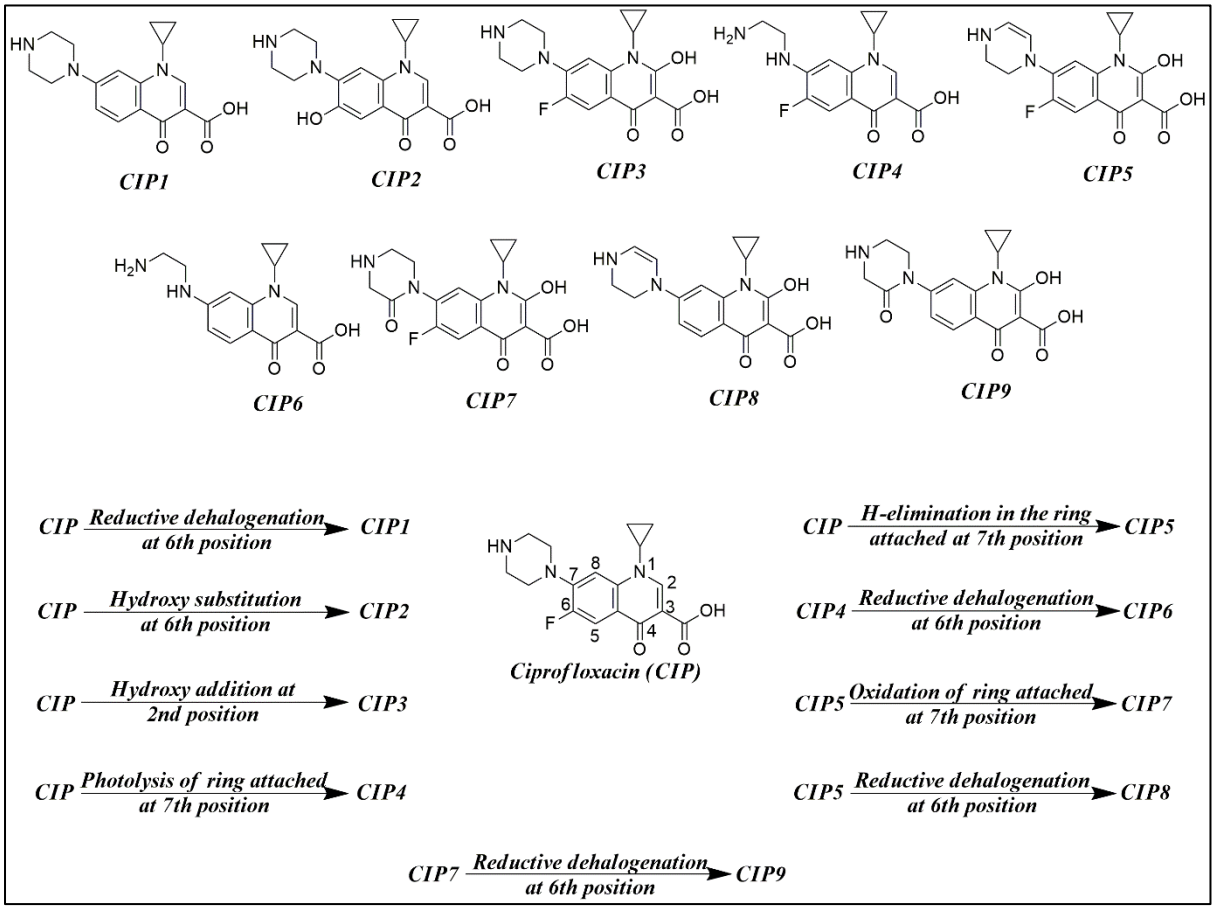


Figure 3.11. Different transformation reactions (UV light-mediated) and the probable photodegradation degradation products

Table 3.34. Composition of the hypothetical mixtures designed from the FQs and their corresponding degradation products

Hypothetical Mixture No.	Component 1	Fraction 1	Component 2	Fraction 2
HM1	CIP	0.5	CIP1	0.5
HM2	CIP	0.5	CIP2	0.5
HM3	CIP	0.5	CIP3	0.5
HM4	CIP	0.5	CIP4	0.5
HM5	CIP	0.5	CIP5	0.5
HM6	CIP	0.5	CIP6	0.5
HM7	CIP	0.5	CIP7	0.5
HM8	CIP	0.5	CIP8	0.5

HM9	CIP	0.5	CIP9	0.5
HM10	CIP1	0.5	CIP2	0.5
HM11	CIP1	0.5	CIP3	0.5
HM12	CIP1	0.5	CIP4	0.5
HM13	CIP1	0.5	CIP5	0.5
HM14	CIP1	0.5	CIP6	0.5
HM15	CIP1	0.5	CIP7	0.5
HM16	CIP1	0.5	CIP8	0.5
HM17	CIP1	0.5	CIP9	0.5
HM18	CIP2	0.5	CIP3	0.5
HM19	CIP2	0.5	CIP4	0.5
HM20	CIP2	0.5	CIP5	0.5
HM21	CIP2	0.5	CIP6	0.5
HM22	CIP2	0.5	CIP7	0.5
HM23	CIP2	0.5	CIP8	0.5
HM24	CIP2	0.5	CIP9	0.5
HM25	CIP3	0.5	CIP4	0.5
HM26	CIP3	0.5	CIP5	0.5
HM27	CIP3	0.5	CIP6	0.5
HM28	CIP3	0.5	CIP7	0.5
HM29	CIP3	0.5	CIP8	0.5
HM30	CIP3	0.5	CIP9	0.5
HM31	CIP4	0.5	CIP5	0.5
HM32	CIP4	0.5	CIP6	0.5
HM33	CIP4	0.5	CIP7	0.5
HM34	CIP4	0.5	CIP8	0.5
HM35	CIP4	0.5	CIP9	0.5
HM36	CIP5	0.5	CIP6	0.5
HM37	CIP5	0.5	CIP7	0.5
HM38	CIP5	0.5	CIP8	0.5

HM39	CIP5	0.5	CIP9	0.5
HM40	CIP6	0.5	CIP7	0.5
HM41	CIP6	0.5	CIP8	0.5
HM42	CIP6	0.5	CIP9	0.5
HM43	CIP7	0.5	CIP8	0.5
HM44	CIP7	0.5	CIP9	0.5
HM45	CIP8	0.5	CIP9	0.5
HM46	DAN	0.5	DAN1	0.5
HM47	DAN	0.5	DAN2	0.5
HM48	DAN	0.5	DAN3	0.5
HM49	DAN	0.5	DAN4	0.5
HM50	DAN	0.5	DAN5	0.5
HM51	DAN	0.5	DAN6	0.5
HM52	DAN	0.5	DAN7	0.5
HM53	DAN	0.5	DAN8	0.5
HM54	DAN	0.5	DAN9	0.5
HM55	DAN1	0.5	DAN2	0.5
HM56	DAN1	0.5	DAN3	0.5
HM57	DAN1	0.5	DAN4	0.5
HM58	DAN1	0.5	DAN5	0.5
HM59	DAN1	0.5	DAN6	0.5
HM60	DAN1	0.5	DAN7	0.5
HM61	DAN1	0.5	DAN8	0.5
HM62	DAN1	0.5	DAN9	0.5
HM63	DAN2	0.5	DAN3	0.5
HM64	DAN2	0.5	DAN4	0.5
HM65	DAN2	0.5	DAN5	0.5
HM66	DAN2	0.5	DAN6	0.5
HM67	DAN2	0.5	DAN7	0.5
HM68	DAN2	0.5	DAN8	0.5

HM69	DAN2	0.5	DAN9	0.5
HM70	DAN3	0.5	DAN4	0.5
HM71	DAN3	0.5	DAN5	0.5
HM72	DAN3	0.5	DAN6	0.5
HM73	DAN3	0.5	DAN7	0.5
HM74	DAN3	0.5	DAN8	0.5
HM75	DAN3	0.5	DAN9	0.5
HM76	DAN4	0.5	DAN5	0.5
HM77	DAN4	0.5	DAN6	0.5
HM78	DAN4	0.5	DAN7	0.5
HM79	DAN4	0.5	DAN8	0.5
HM80	DAN4	0.5	DAN9	0.5
HM81	DAN5	0.5	DAN6	0.5
HM82	DAN5	0.5	DAN7	0.5
HM83	DAN5	0.5	DAN8	0.5
HM84	DAN5	0.5	DAN9	0.5
HM85	DAN6	0.5	DAN7	0.5
HM86	DAN6	0.5	DAN8	0.5
HM87	DAN6	0.5	DAN9	0.5
HM88	DAN7	0.5	DAN8	0.5
HM89	DAN7	0.5	DAN9	0.5
HM90	DAN8	0.5	DAN9	0.5
HM91	ENO	0.5	ENO1	0.5
HM92	ENO	0.5	ENO2	0.5
HM93	ENO	0.5	ENO3	0.5
HM94	ENO	0.5	ENO4	0.5
HM95	ENO	0.5	ENO5	0.5
HM96	ENO	0.5	ENO6	0.5
HM97	ENO	0.5	ENO7	0.5
HM98	ENO	0.5	ENO8	0.5

HM99	ENO	0.5	ENO9	0.5
HM100	ENO1	0.5	ENO2	0.5
HM101	ENO1	0.5	ENO3	0.5
HM102	ENO1	0.5	ENO4	0.5
HM103	ENO1	0.5	ENO5	0.5
HM104	ENO1	0.5	ENO6	0.5
HM105	ENO1	0.5	ENO7	0.5
HM106	ENO1	0.5	ENO8	0.5
HM107	ENO1	0.5	ENO9	0.5
HM108	ENO2	0.5	ENO3	0.5
HM109	ENO2	0.5	ENO4	0.5
HM110	ENO2	0.5	ENO5	0.5
HM111	ENO2	0.5	ENO6	0.5
HM112	ENO2	0.5	ENO7	0.5
HM113	ENO2	0.5	ENO8	0.5
HM114	ENO2	0.5	ENO9	0.5
HM115	ENO3	0.5	ENO4	0.5
HM116	ENO3	0.5	ENO5	0.5
HM117	ENO3	0.5	ENO6	0.5
HM118	ENO3	0.5	ENO7	0.5
HM119	ENO3	0.5	ENO8	0.5
HM120	ENO3	0.5	ENO9	0.5
HM121	ENO4	0.5	ENO5	0.5
HM122	ENO4	0.5	ENO6	0.5
HM123	ENO4	0.5	ENO7	0.5
HM124	ENO4	0.5	ENO8	0.5
HM125	ENO4	0.5	ENO9	0.5
HM126	ENO5	0.5	ENO6	0.5
HM127	ENO5	0.5	ENO7	0.5
HM128	ENO5	0.5	ENO8	0.5

HM129	ENO5	0.5	ENO9	0.5
HM130	ENO6	0.5	ENO7	0.5
HM131	ENO6	0.5	ENO8	0.5
HM132	ENO6	0.5	ENO9	0.5
HM133	ENO7	0.5	ENO8	0.5
HM134	ENO7	0.5	ENO9	0.5
HM135	ENO8	0.5	ENO9	0.5
HM136	ENR	0.5	ENR1	0.5
HM137	ENR	0.5	ENR2	0.5
HM138	ENR	0.5	ENR3	0.5
HM139	ENR	0.5	ENR4	0.5
HM140	ENR	0.5	ENR5	0.5
HM141	ENR	0.5	ENR6	0.5
HM142	ENR	0.5	ENR7	0.5
HM143	ENR	0.5	ENR8	0.5
HM144	ENR	0.5	ENR9	0.5
HM145	ENR1	0.5	ENR2	0.5
HM146	ENR1	0.5	ENR3	0.5
HM147	ENR1	0.5	ENR4	0.5
HM148	ENR1	0.5	ENR5	0.5
HM149	ENR1	0.5	ENR6	0.5
HM150	ENR1	0.5	ENR7	0.5
HM151	ENR1	0.5	ENR8	0.5
HM152	ENR1	0.5	ENR9	0.5
HM153	ENR2	0.5	ENR3	0.5
HM154	ENR2	0.5	ENR4	0.5
HM155	ENR2	0.5	ENR5	0.5
HM156	ENR2	0.5	ENR6	0.5
HM157	ENR2	0.5	ENR7	0.5
HM158	ENR2	0.5	ENR8	0.5

HM159	ENR2	0.5	ENR9	0.5
HM160	ENR3	0.5	ENR4	0.5
HM161	ENR3	0.5	ENR5	0.5
HM162	ENR3	0.5	ENR6	0.5
HM163	ENR3	0.5	ENR7	0.5
HM164	ENR3	0.5	ENR8	0.5
HM165	ENR3	0.5	ENR9	0.5
HM166	ENR4	0.5	ENR5	0.5
HM167	ENR4	0.5	ENR6	0.5
HM168	ENR4	0.5	ENR7	0.5
HM169	ENR4	0.5	ENR8	0.5
HM170	ENR4	0.5	ENR9	0.5
HM171	ENR5	0.5	ENR6	0.5
HM172	ENR5	0.5	ENR7	0.5
HM173	ENR5	0.5	ENR8	0.5
HM174	ENR5	0.5	ENR9	0.5
HM175	ENR6	0.5	ENR7	0.5
HM176	ENR6	0.5	ENR8	0.5
HM177	ENR6	0.5	ENR9	0.5
HM178	ENR7	0.5	ENR8	0.5
HM179	ENR7	0.5	ENR9	0.5
HM180	ENR8	0.5	ENR9	0.5
HM181	LEV	0.5	LEV1	0.5
HM182	LEV	0.5	LEV2	0.5
HM183	LEV	0.5	LEV3	0.5
HM184	LEV	0.5	LEV4	0.5
HM185	LEV	0.5	LEV5	0.5
HM186	LEV	0.5	LEV6	0.5
HM187	LEV	0.5	LEV7	0.5
HM188	LEV	0.5	LEV8	0.5

HM189	LEV	0.5	LEV9	0.5
HM190	LEV1	0.5	LEV2	0.5
HM191	LEV1	0.5	LEV3	0.5
HM192	LEV1	0.5	LEV4	0.5
HM193	LEV1	0.5	LEV5	0.5
HM194	LEV1	0.5	LEV6	0.5
HM195	LEV1	0.5	LEV7	0.5
HM196	LEV1	0.5	LEV8	0.5
HM197	LEV1	0.5	LEV9	0.5
HM198	LEV2	0.5	LEV3	0.5
HM199	LEV2	0.5	LEV4	0.5
HM200	LEV2	0.5	LEV5	0.5
HM201	LEV2	0.5	LEV6	0.5
HM202	LEV2	0.5	LEV7	0.5
HM203	LEV2	0.5	LEV8	0.5
HM204	LEV2	0.5	LEV9	0.5
HM205	LEV3	0.5	LEV4	0.5
HM206	LEV3	0.5	LEV5	0.5
HM207	LEV3	0.5	LEV6	0.5
HM208	LEV3	0.5	LEV7	0.5
HM209	LEV3	0.5	LEV8	0.5
HM210	LEV3	0.5	LEV9	0.5
HM211	LEV4	0.5	LEV5	0.5
HM212	LEV4	0.5	LEV6	0.5
HM213	LEV4	0.5	LEV7	0.5
HM214	LEV4	0.5	LEV8	0.5
HM215	LEV4	0.5	LEV9	0.5
HM216	LEV5	0.5	LEV6	0.5
HM217	LEV5	0.5	LEV7	0.5
HM218	LEV5	0.5	LEV8	0.5

HM219	LEV5	0.5	LEV9	0.5
HM220	LEV6	0.5	LEV7	0.5
HM221	LEV6	0.5	LEV8	0.5
HM222	LEV6	0.5	LEV9	0.5
HM223	LEV7	0.5	LEV8	0.5
HM224	LEV7	0.5	LEV9	0.5
HM225	LEV8	0.5	LEV9	0.5
HM226	MAR	0.5	MAR1	0.5
HM227	MAR	0.5	MAR2	0.5
HM228	MAR	0.5	MAR3	0.5
HM229	MAR	0.5	MAR4	0.5
HM230	MAR	0.5	MAR5	0.5
HM231	MAR	0.5	MAR6	0.5
HM232	MAR	0.5	MAR7	0.5
HM233	MAR	0.5	MAR8	0.5
HM234	MAR	0.5	MAR9	0.5
HM235	MAR1	0.5	MAR2	0.5
HM236	MAR1	0.5	MAR3	0.5
HM237	MAR1	0.5	MAR4	0.5
HM238	MAR1	0.5	MAR5	0.5
HM239	MAR1	0.5	MAR6	0.5
HM240	MAR1	0.5	MAR7	0.5
HM241	MAR1	0.5	MAR8	0.5
HM242	MAR1	0.5	MAR9	0.5
HM243	MAR2	0.5	MAR3	0.5
HM244	MAR2	0.5	MAR4	0.5
HM245	MAR2	0.5	MAR5	0.5
HM246	MAR2	0.5	MAR6	0.5
HM247	MAR2	0.5	MAR7	0.5
HM248	MAR2	0.5	MAR8	0.5

HM249	MAR2	0.5	MAR9	0.5
HM250	MAR3	0.5	MAR4	0.5
HM251	MAR3	0.5	MAR5	0.5
HM252	MAR3	0.5	MAR6	0.5
HM253	MAR3	0.5	MAR7	0.5
HM254	MAR3	0.5	MAR8	0.5
HM255	MAR3	0.5	MAR9	0.5
HM256	MAR4	0.5	MAR5	0.5
HM257	MAR4	0.5	MAR6	0.5
HM258	MAR4	0.5	MAR7	0.5
HM259	MAR4	0.5	MAR8	0.5
HM260	MAR4	0.5	MAR9	0.5
HM261	MAR5	0.5	MAR6	0.5
HM262	MAR5	0.5	MAR7	0.5
HM263	MAR5	0.5	MAR8	0.5
HM264	MAR5	0.5	MAR9	0.5
HM265	MAR6	0.5	MAR7	0.5
HM266	MAR6	0.5	MAR8	0.5
HM267	MAR6	0.5	MAR9	0.5
HM268	MAR7	0.5	MAR8	0.5
HM269	MAR7	0.5	MAR9	0.5
HM270	MAR8	0.5	MAR9	0.5
HM271	NOR	0.5	NOR1	0.5
HM272	NOR	0.5	NOR2	0.5
HM273	NOR	0.5	NOR3	0.5
HM274	NOR	0.5	NOR4	0.5
HM275	NOR	0.5	NOR5	0.5
HM276	NOR	0.5	NOR6	0.5
HM277	NOR	0.5	NOR7	0.5
HM278	NOR	0.5	NOR8	0.5

HM279	NOR	0.5	NOR9	0.5
HM280	NOR1	0.5	NOR2	0.5
HM281	NOR1	0.5	NOR3	0.5
HM282	NOR1	0.5	NOR4	0.5
HM283	NOR1	0.5	NOR5	0.5
HM284	NOR1	0.5	NOR6	0.5
HM285	NOR1	0.5	NOR7	0.5
HM286	NOR1	0.5	NOR8	0.5
HM287	NOR1	0.5	NOR9	0.5
HM288	NOR2	0.5	NOR3	0.5
HM289	NOR2	0.5	NOR4	0.5
HM290	NOR2	0.5	NOR5	0.5
HM291	NOR2	0.5	NOR6	0.5
HM292	NOR2	0.5	NOR7	0.5
HM293	NOR2	0.5	NOR8	0.5
HM294	NOR2	0.5	NOR9	0.5
HM295	NOR3	0.5	NOR4	0.5
HM296	NOR3	0.5	NOR5	0.5
HM297	NOR3	0.5	NOR6	0.5
HM298	NOR3	0.5	NOR7	0.5
HM299	NOR3	0.5	NOR8	0.5
HM300	NOR3	0.5	NOR9	0.5
HM301	NOR4	0.5	NOR5	0.5
HM302	NOR4	0.5	NOR6	0.5
HM303	NOR4	0.5	NOR7	0.5
HM304	NOR4	0.5	NOR8	0.5
HM305	NOR4	0.5	NOR9	0.5
HM306	NOR5	0.5	NOR6	0.5
HM307	NOR5	0.5	NOR7	0.5
HM308	NOR5	0.5	NOR8	0.5

HM309	NOR5	0.5	NOR9	0.5
HM310	NOR6	0.5	NOR7	0.5
HM311	NOR6	0.5	NOR8	0.5
HM312	NOR6	0.5	NOR9	0.5
HM313	NOR7	0.5	NOR8	0.5
HM314	NOR7	0.5	NOR9	0.5
HM315	NOR8	0.5	NOR9	0.5
HM316	PEF	0.5	PEF1	0.5
HM317	PEF	0.5	PEF2	0.5
HM318	PEF	0.5	PEF3	0.5
HM319	PEF	0.5	PEF4	0.5
HM320	PEF	0.5	PEF5	0.5
HM321	PEF	0.5	PEF6	0.5
HM322	PEF	0.5	PEF7	0.5
HM323	PEF	0.5	PEF8	0.5
HM324	PEF	0.5	PEF9	0.5
HM325	PEF1	0.5	PEF2	0.5
HM326	PEF1	0.5	PEF3	0.5
HM327	PEF1	0.5	PEF4	0.5
HM328	PEF1	0.5	PEF5	0.5
HM329	PEF1	0.5	PEF6	0.5
HM330	PEF1	0.5	PEF7	0.5
HM331	PEF1	0.5	PEF8	0.5
HM332	PEF1	0.5	PEF9	0.5
HM333	PEF2	0.5	PEF3	0.5
HM334	PEF2	0.5	PEF4	0.5
HM335	PEF2	0.5	PEF5	0.5
HM336	PEF2	0.5	PEF6	0.5
HM337	PEF2	0.5	PEF7	0.5
HM338	PEF2	0.5	PEF8	0.5

HM339	PEF2	0.5	PEF9	0.5
HM340	PEF3	0.5	PEF4	0.5
HM341	PEF3	0.5	PEF5	0.5
HM342	PEF3	0.5	PEF6	0.5
HM343	PEF3	0.5	PEF7	0.5
HM344	PEF3	0.5	PEF8	0.5
HM345	PEF3	0.5	PEF9	0.5
HM346	PEF4	0.5	PEF5	0.5
HM347	PEF4	0.5	PEF6	0.5
HM348	PEF4	0.5	PEF7	0.5
HM349	PEF4	0.5	PEF8	0.5
HM350	PEF4	0.5	PEF9	0.5
HM351	PEF5	0.5	PEF6	0.5
HM352	PEF5	0.5	PEF7	0.5
HM353	PEF5	0.5	PEF8	0.5
HM354	PEF5	0.5	PEF9	0.5
HM355	PEF6	0.5	PEF7	0.5
HM356	PEF6	0.5	PEF8	0.5
HM357	PEF6	0.5	PEF9	0.5
HM358	PEF7	0.5	PEF8	0.5
HM359	PEF7	0.5	PEF9	0.5
HM360	PEF8	0.5	PEF9	0.5
HM361	SAR	0.5	SAR1	0.5
HM362	SAR	0.5	SAR2	0.5
HM363	SAR	0.5	SAR3	0.5
HM364	SAR	0.5	SAR4	0.5
HM365	SAR	0.5	SAR5	0.5
HM366	SAR	0.5	SAR6	0.5
HM367	SAR	0.5	SAR7	0.5
HM368	SAR	0.5	SAR8	0.5

HM369	SAR	0.5	SAR9	0.5
HM370	SAR1	0.5	SAR2	0.5
HM371	SAR1	0.5	SAR3	0.5
HM372	SAR1	0.5	SAR4	0.5
HM373	SAR1	0.5	SAR5	0.5
HM374	SAR1	0.5	SAR6	0.5
HM375	SAR1	0.5	SAR7	0.5
HM376	SAR1	0.5	SAR8	0.5
HM377	SAR1	0.5	SAR9	0.5
HM378	SAR2	0.5	SAR3	0.5
HM379	SAR2	0.5	SAR4	0.5
HM380	SAR2	0.5	SAR5	0.5
HM381	SAR2	0.5	SAR6	0.5
HM382	SAR2	0.5	SAR7	0.5
HM383	SAR2	0.5	SAR8	0.5
HM384	SAR2	0.5	SAR9	0.5
HM385	SAR3	0.5	SAR4	0.5
HM386	SAR3	0.5	SAR5	0.5
HM387	SAR3	0.5	SAR6	0.5
HM388	SAR3	0.5	SAR7	0.5
HM389	SAR3	0.5	SAR8	0.5
HM390	SAR3	0.5	SAR9	0.5
HM391	SAR4	0.5	SAR5	0.5
HM392	SAR4	0.5	SAR6	0.5
HM393	SAR4	0.5	SAR7	0.5
HM394	SAR4	0.5	SAR8	0.5
HM395	SAR4	0.5	SAR9	0.5
HM396	SAR5	0.5	SAR6	0.5
HM397	SAR5	0.5	SAR7	0.5
HM398	SAR5	0.5	SAR8	0.5

HM399	SAR5	0.5	SAR9	0.5
HM400	SAR6	0.5	SAR7	0.5
HM401	SAR6	0.5	SAR8	0.5
HM402	SAR6	0.5	SAR9	0.5
HM403	SAR7	0.5	SAR8	0.5
HM404	SAR7	0.5	SAR9	0.5
HM405	SAR8	0.5	SAR9	0.5

*N.B. HM = hypothetical mixtures, CIP =ciprofloxacin, DAN = danofloxacin, ENO = enoxacin, ENR = enrofloxacin, LEV = levofloxacin, MAR = marbofloxacin, NOR = norfloxacin, PEF = pefloxacin, and SAR = sarafloxacin

The 2D structures of all degradation products have been made in MarvinSketch software with explicit hydrogen addition and proper aromatization and then subjected to the AlvaDesc software for descriptor calculation. The mixture descriptors have been computed from the computed descriptors (modeled descriptors) of individual components by employing the linear combination of molecular contribution in a similar manner as discussed in the previous "***3.6.2. Mixture descriptor calculation***" section. Finally, the "Prediction reliability indicator" (PRI) tool (available online: http://teqip.jdvu.ac.in/QSAR_Tools/) has been employed for the prediction of mixture toxicity and the reliability of the predictions as well. This tool was developed by Roy et al. (Roy et al., 2018b) that predicts the toxicity of untested mixtures using the developed model and also categorizes the reliability of the predicted responses into 'Good', 'Moderate', and 'Poor', depending upon the composite score. The reliability of predictions obtained from the PRI tools is essentially based on three sets of principles i.e., "*(i) mean absolute error of prediction (leave-one-out) of a query compound from 10 most close training compounds; (ii) applicability domain in terms of similarity, by using standardization approach; (iii) closeness of predicted values to the mean of the observed response of training set compounds*". This tool also provides the applicability domain status of the untested mixtures based on the Euclidean distance analysis of query mixtures with the training set mixture

CHAPTER - 4

RESULTS AND DISCUSSIONS

Chapter 4: Results and Discussions

4.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors

Four PLS models were constructed from the four data sets mentioned in the materials and methods section. The first two models were for the individual chemical datasets (**Model 1** and **Model 2**) whereas the next two were for the mixtures (**Model 3** and **Model 4**). The details of latent variables and descriptors of the models are enlisted in **Table 4.1**. All the models were evaluated rigorously by internationally accepted internal and external validation criteria**.** The models were robust and statistically sound considering the stipulated threshold values of the validation parameters (**Table 4.2**). The scatter plots of observed vs. predicted response (pEC50) of the models are given in **Figure 4.1**.

Table 4.1. Descriptors and LVs of the constructed PLS models

PLS Models	No. of LVs	No. of Descriptors	Descriptors
1	3	4	B04[C-C], X5A, ETA_dBetaP, Mv
2	2	3	Hy, F07[C-C], B04[N-Cl]
3	2	6	B03[N-O], nR=Cp, nROH, N-074, C-006, F07[O-O]
4	2	3	ETA_Alpha, F06[C-S], nCsp3

LVs: Latent variables; B04[C-C]: Presence/absence of C - C at the topological distance 4; X5A: average connectivity index of order 5; ETA_dBetaP: A measure of relative unsaturation content with respect to the molecular size; Mv: mean atomic van der Waals volume (scaled on Carbon atom); Hy: hydrophilic factor; F07[C-C: Frequency of C - C at the topological distance 7; B04[N-Cl]: Presence/absence of N - Cl at the topological distance 4; B03[N-O]: Presence/absence of N - O at the topological distance 3; nR=Cp: number of terminal primary C(sp^2); nROH: number of hydroxyl groups; N-074: atom centered fragment containing -C#N/-CH=N- groups (#: triple bond); C-006: atom centered fragment containing $–CH_2RX$ group (R: any group linked through C; X: heteroatoms like O,N,S,P, Se, Halogen); F07[O-O]: Frequency of O - O at the topological distance 7; ETA_Alpha: Summed α values of all non-hydrogen vertices of a molecule (or fragment/substructure); F06[C-S]: Frequency of C - S at the topological distance 6; $nCsp^3$: number of sp^3 hybridized Carbon atoms.

The datasets of Series 2 compounds and type 2 mixtures (presented in **Table 3.3** and **3.4** respectively in the thesis) contain 20 data points each. In both cases, 15 data points were used for model development (**Model 2** and **Model 4**) and rest 5 data points were used for rigorous external validation. In both these cases, the models were developed with limited data due to the lack of availability of additional experimental data points. The models were properly validated and were acceptable in terms of goodness-of-fit, robustness, and external

predictivity (the metric values were discussed in section "***4.1.2*** and ***4.1.4***"). The acceptable result of validation metrics indicates reliability of predictions. However, if the additional experimental data are available, more reliable models can be developed.

Table 4.2. Statistical metric values of the developed PLS models

Statistical Metrics		Model 1 (Set1 Compounds)	Model 2 (Set2 Compounds)	Model 3 (Type-1 Mixtures)	Model 4 (Type-2 Mixtures)
Internal validation	N_{Train}	27	15	60	15
	R^2	0.84	0.80	0.76	0.78
	R^2_{adj}	0.81	075	0.73	0.72
	Q^2_{LOO}	0.78 (3 LVs)	0.65 (2 LVs)	0.70 (2 LVs)	0.73 (2 LVs)
	$avg\ r_m^{2}{}_{(LOO)}$	0.68	0.56	0.60	0.58
	$\Delta r_m^{2}{}_{(LOO)}$	0.18	0.10	0.13	0.10
External validation	N_{Test}	8	5	19	5
	Q^2_{F1}	0.87	0.68	0.82	0.74
	Q^2_{F2}	0.87	0.61	0.82	0.62
	$avg\ r_m^{2}{}_{(Test)}$	0.79	0.64	0.70	0.79
	$\Delta r_m^{2}{}_{(Test)}$	0.11	0.17	0.13	0.03

Optimum values of all studied statistical metrics are as follows: [R^2, R^2_{adj}, Q^2_{LOO}, Q^2_{F1}, Q^2_{F2}, $avgr_m^{2}{}_{(LOO)}$, $avg\ r_m^{2}{}_{(Test)}$> 0.5] and [$\Delta r_m^{2}{}_{(LOO)}$, $\Delta r_m^{2}{}_{(Test)}$ < 0.2

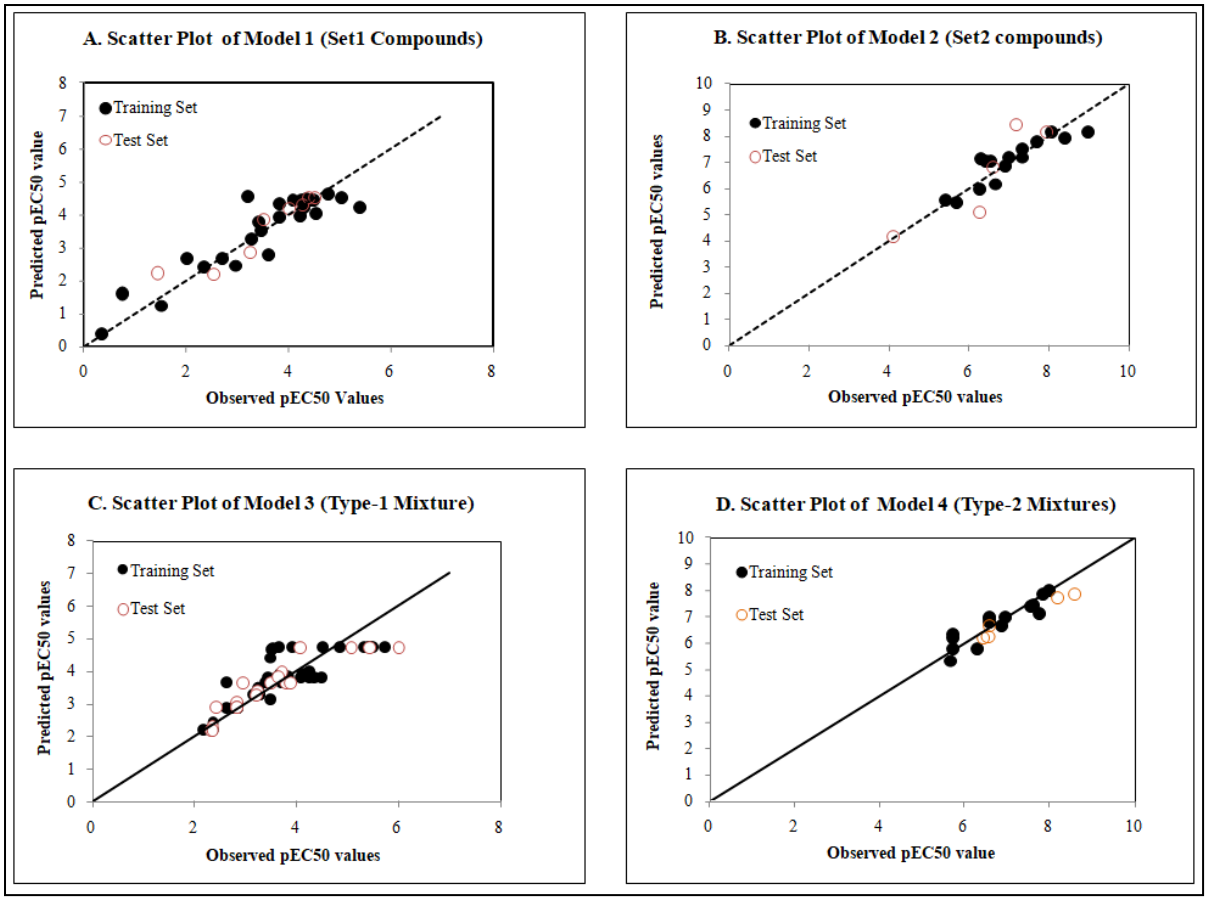


Figure 4.1 Scatter plots of four PLS models; [A: Scatterd plot of Model 1 (Set1 compounds); B: Scatter plot of Model 2 (Set2 compounds); C: Scatter plot of Model 3 (Type-1 mixtures); D: Scatter plot of Model 4 (Type-2 mixtures)]

4.1.1. QSAR study of toxicity of individual chemicals (Series 1) to P. phosphoreum

Model 1 (Series 1) has been developed with four descriptors using three latent variables (LVs) and can be illustrated in the following equation 4.1-

$$pEC50 = -1.00784 + 2.08722 \times B04[C-Cl] - 7.3775 \times X5A - 2.36678 \times ETA_{dBetaP} + 6.70342 \times Mv \qquad (4.1)$$

N_{train}= 27; R^2= 0.84; R^2_{adj}= 0.81; Q^2_{LOO}= 0.78 (3 LV); avg $r_m^2{}_{(LOO)}$=0.68; $\Delta r_m^2{}_{(LOO)}$=0.18; MAE_{Train}= 0.42; N_{test}= 8; Q^2_{F1}= 0.87; Q^2_{F2}=0.8; avg $r_m^2{}_{(test)}$=0.79; $\Delta r_m^2{}_{(test)}$=0.11; MAE_{Test}=0.27

Model 1 showed acceptable values of validation metrics (both internal and external). The R^2, R^2_{adj}, and Q^2_{LOO}values are 0.84, 0.81, and 0.78 (3 LVs), respectively signifying good statistical reliability and robustness of the model. The predictive ability of the model was judged by means of predictive R^2 (R^2_{pred}) or Q^2_{F1}, which shows good predictive ability of the model (Q^2_{F1} = 0.87). A regression coefficient plot (Wold et al., 2001) (**Figure 4.2A**) suggests the positive and negative contribution of descriptors towards the modeled response (pEC50). The descriptors with positive contributions (B04[C-C], Mv) suggest an increase in the response with an increase in the values of descriptors whereas descriptors with negative contributions (X5A, ETA_dBetaP) suggest the inverse relationship. The variable importance plot (VIP) (**Figure 4.2B**) signifies the importance of each descriptor obtained in the final PLS

model to the modeled response (pEC50). The descriptors having VIP score > 1 show higher statistical significance towards toxicity of the compounds (Akarachantachote et al., 2014); and the variables are arranged in a descending order of importance from left to right of the plot. The significance of the descriptors towards modeled response (Model 1) is in following order: B04[C-C] > X5A >ETA_dBetaP>Mv. B04[C-C] is highly significant descriptor of this model. The rest three descriptors are having almost similar level of significance towards the model.

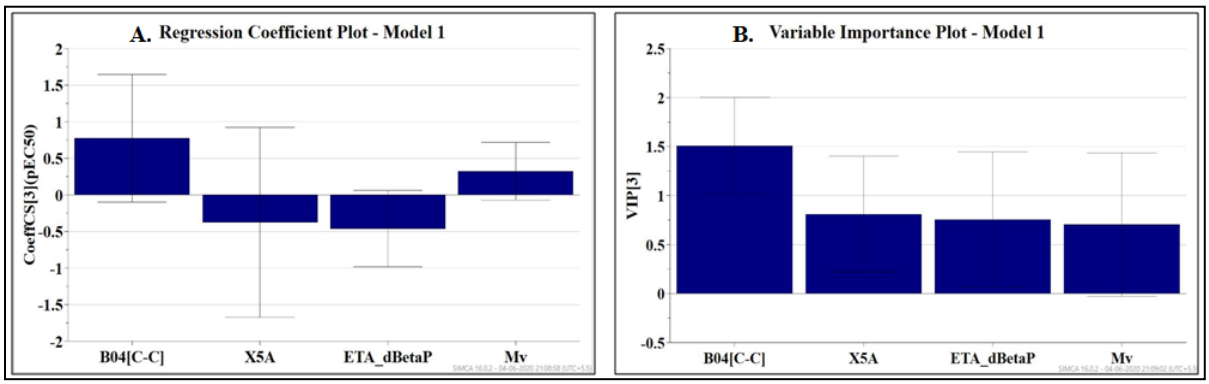


Figure 4.2. A. Regression Coefficient Plot of Model 1; B. Variable Importance Plot of Model 1

The most contributing descriptor towards the toxicity of compounds is B04[C-C] which is basically a 2D atom pair descriptor, which checks whether two carbon atoms present at topological distance 4 or not. It has positive correlation with the response which signifies an increase of the response in presence of C…C at the topological distance 4 and vice-versa. **Figure 4.3** depicts that propionaldehyde (Comp. No. **2**) and succinonitrile (Comp. No. **22**), which are devoid of C…C at the topological distance 4, have lower pEC50 values than valeraldehyde (Comp. No. **4)** and phthalonitrile (Comp. No.**23)** with C…C at the topological distance 4.

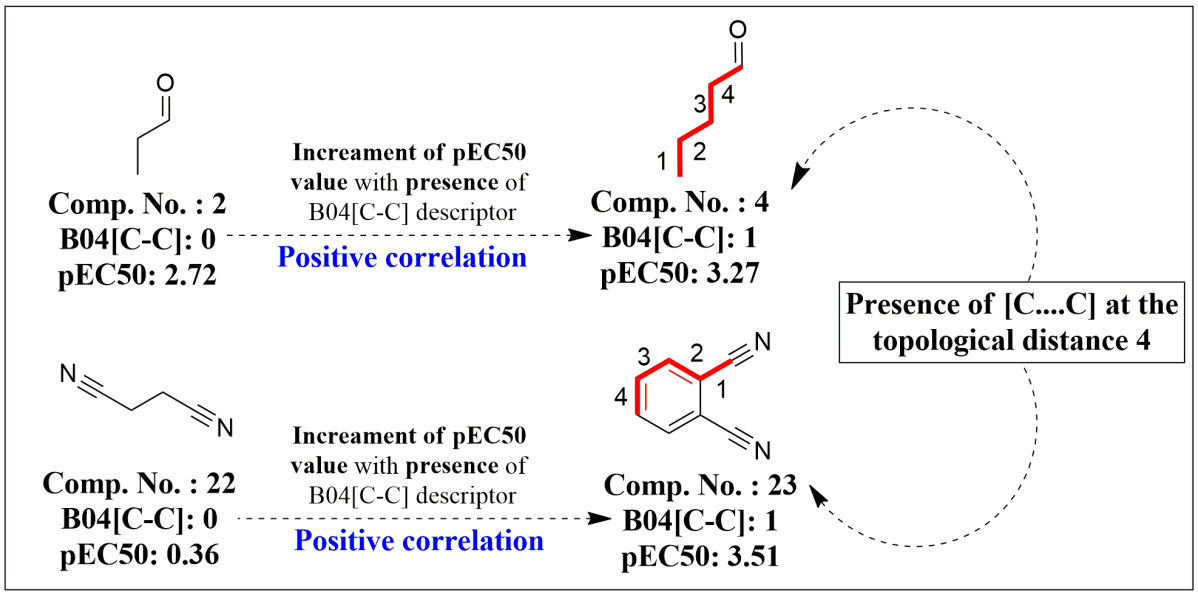


Figure 4.3. Contribution of B04[C-C] descriptor to pEC50

The next important descriptor X5A is the average connectivity index of order five which has a negative contribution towards the modeled response. In **Figure 4.4**, it has been clearly depicted that pEC50 value gets lowered with an increase in the X5A value from α-Hydroxyisobutyronitrile (Comp.No. **16**) to succinonitrile (Comp. No. **22**).

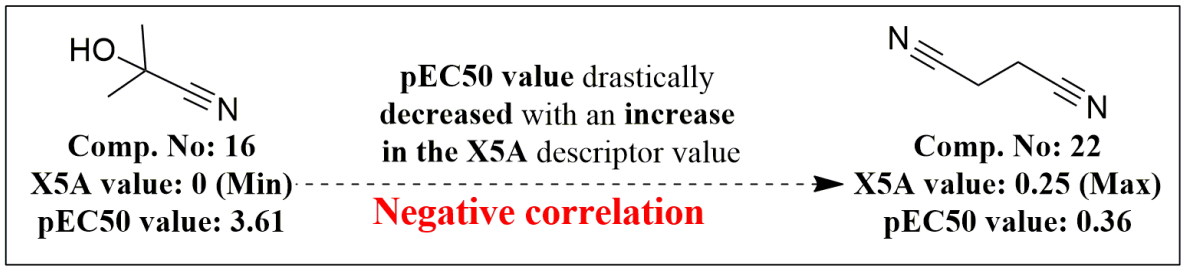


Figure 4.4. Contribution of X5A descriptor to pEC50

ETA_dBetaP (Δβ') is an important feature of Model 1 (Set 1), which is a 2nd generation Extended Topochemical Atom (ETA) descriptor. It is a measure of unsaturation content relative to the molecular size. It has a negative contribution towards the modeled response. The mathematical form of this descriptor is depicted in **Figure 4.5**, which clearly suggests the proportional relationship of ETA_dBetaP with the unsaturation content (Δβ) and inversely proportional relationship of this particular descriptor with molecular size (N*ν*). Among the Set1 compounds, sulfapyridine (Comp. no. **26**) and sulfachloropyridazine (Comp. No. **31)** have similar unsaturation count, but the latter compound has a higher molecular size and correspondingly lower ETA_dBetaP value resulting in a higher pEC50 value. A drastic decrease of pEC50 value has been observed from lactonitrile (Comp.No. **24**) to succinonitrile (Comp.No. **22**) with an increase in the unsaturation content of succinonitrile which is clearly depicted in **Figure 4.5**.

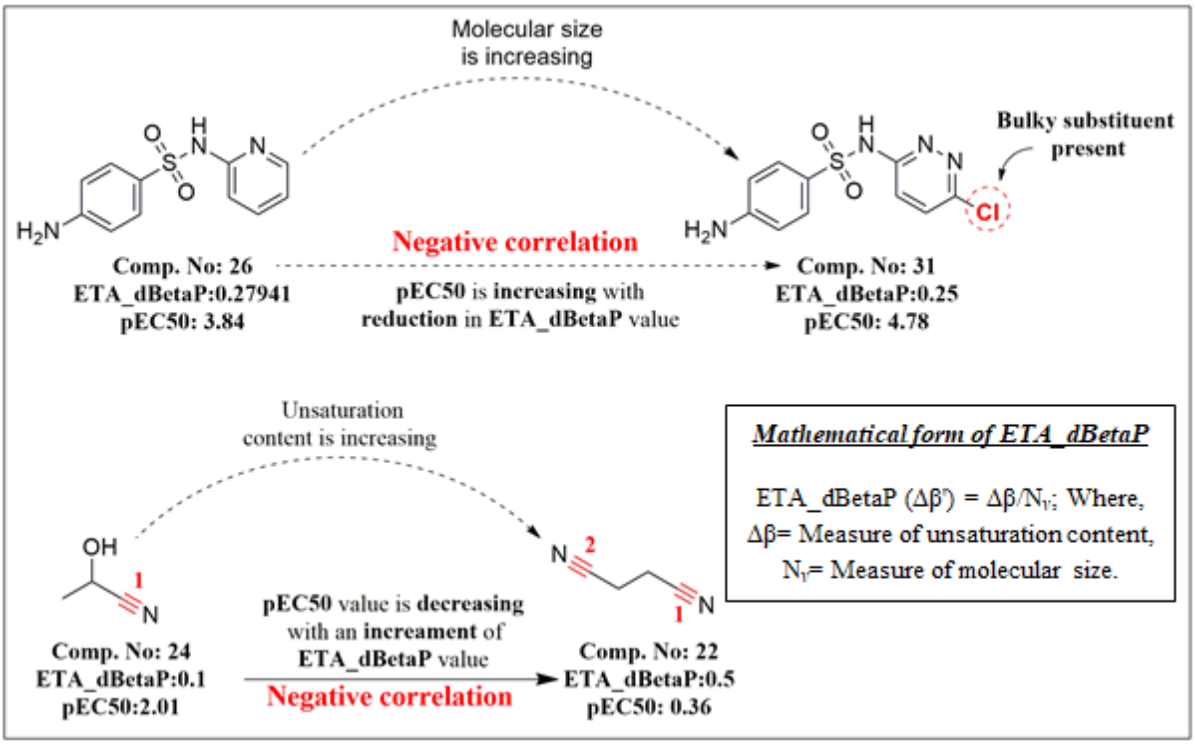


Figure 4.5. Contribution of ETA_dBetaP descriptor to pEC50

Mv (Mean atomic van der Waals volume) is the least important descriptor, which is a measure of molar volume. The positive contribution of this descriptor suggests an importance of molar bulk to the response as seen in more bulky compounds like *p*-bromobenzaldehyde (Comp. No. **9)** or benzyl cyanide (Comp.No. **19)**. In contrast, less bulky compounds like *p*-chlorobenzaldehyde (Comp. No. **8)** and benzonitrile (Comp. No. **18)** have lower Mv values and hence lower pEC50 values also (**Figure 4.6**).

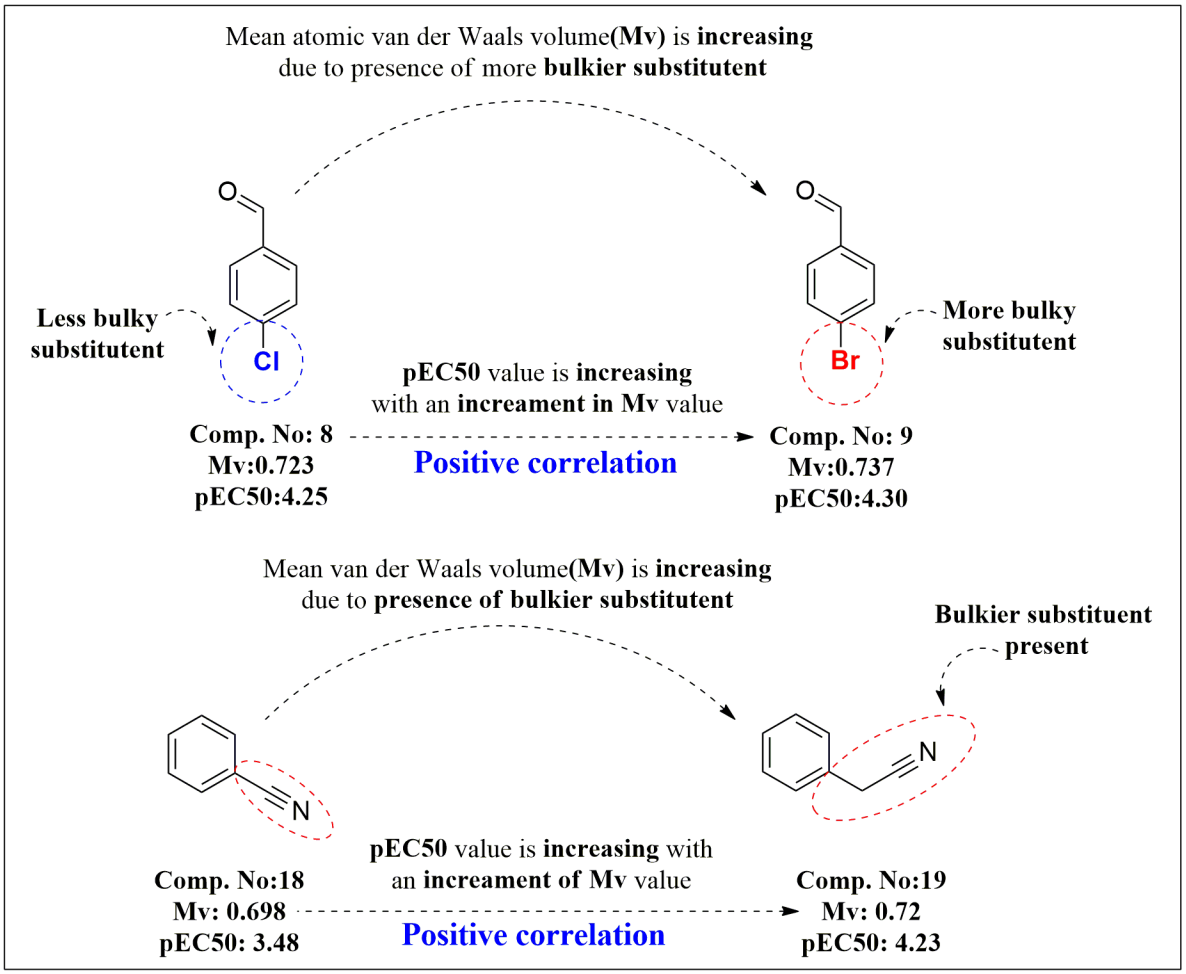


Figure 4.6. Contribution of Mv descriptor to pEC50

4.1.2. QSAR study of toxicity of individual chemicals (Series 2) to S. capricornutum

The developed Model 2 (Series 2 compounds) is composed of three descriptors and 2 latent variables (LV). The PLS model (Model 2) is shown in the equation 4.2-

$$pEC50 = 5.8289 - 2.62934 \times H_y + 0.4797 \times F07[C-C] + 0.73373 \times B04[N-Cl] \quad (4.2)$$

N_{Train}= 15; R^2= 0.80; R^2_{adj}=0.75; Q^2LOO= 0.65 (2 LV); avg $r_m{}^2_{(LOO)}$=0.56; $\Delta r_m{}^2_{(LOO)}$=0.10; MAE_{Train}= 0.46; N_{Test}= 5; Q^2_{F1}= 0.68; Q^2_{F2}= 0. 61; $avgr_m{}^2_{(test)}$=0.64;$\Delta r_m{}^2_{(test)}$=0.17; MAE_{Test}= 0.58

This model showed acceptable values of validation metrics (both internal and external). The coefficient of determination R^2, adjusted coefficient of determination R_a^2, and the cross-validated correlation coefficient LOO-Q^2 values are 0.80, 0.75, and 0.65 (2 LVs), respectively signifying good statistical reliability and robustness of the model. The predictive ability of the model was judged by means of predictive R^2 (R^2_{pred}) or Q^2_{F1}, which shows good predictive ability of the model (Q^2_{F1} = 0.68). The regression coefficient plot of Model 2 (**Figure 4.7A**) suggests the positive contributions of F07[C-C] and B04[N-Cl] descriptors whereas negative contribution of the Hy descriptor to the response. Among the modeled descriptors, F07[C-C] is the most important one towards the response, showing a VIP score more than 1. The next two important descriptors are Hy and B04[N-Cl] respectively (**Figure 4.7B**).

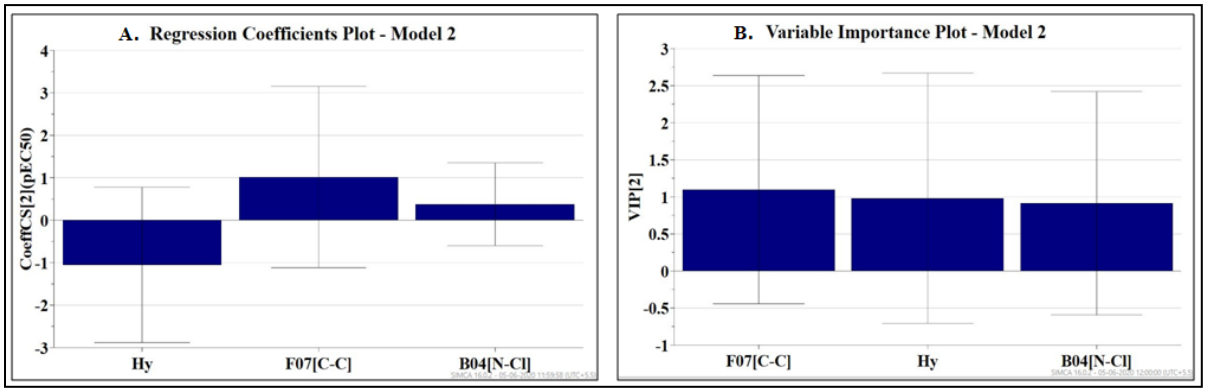


Figure 4.7. A. Regression Coefficient Plot of Model 2; B. Variable Importance Plot of Model 2

F07[C-C] is the frequency of presence of two carbon atoms at the topological distance seven. It is positively correlated with the modeled response which means that an increase inpEC50 value will be observed if the F07[C-C] descriptor value increases and vice-versa. Simazine (Comp. No. **40**) and atrazine (Comp. No. **36**) of Set 2 are similar to each other in structure. However as simazine has a less F07[C-C] value than that of atrazine, the toxicity of the former is also less. The occurrences of two carbons at the topological distance 7 in both the compounds have been depicted in **Figure 4.8**.

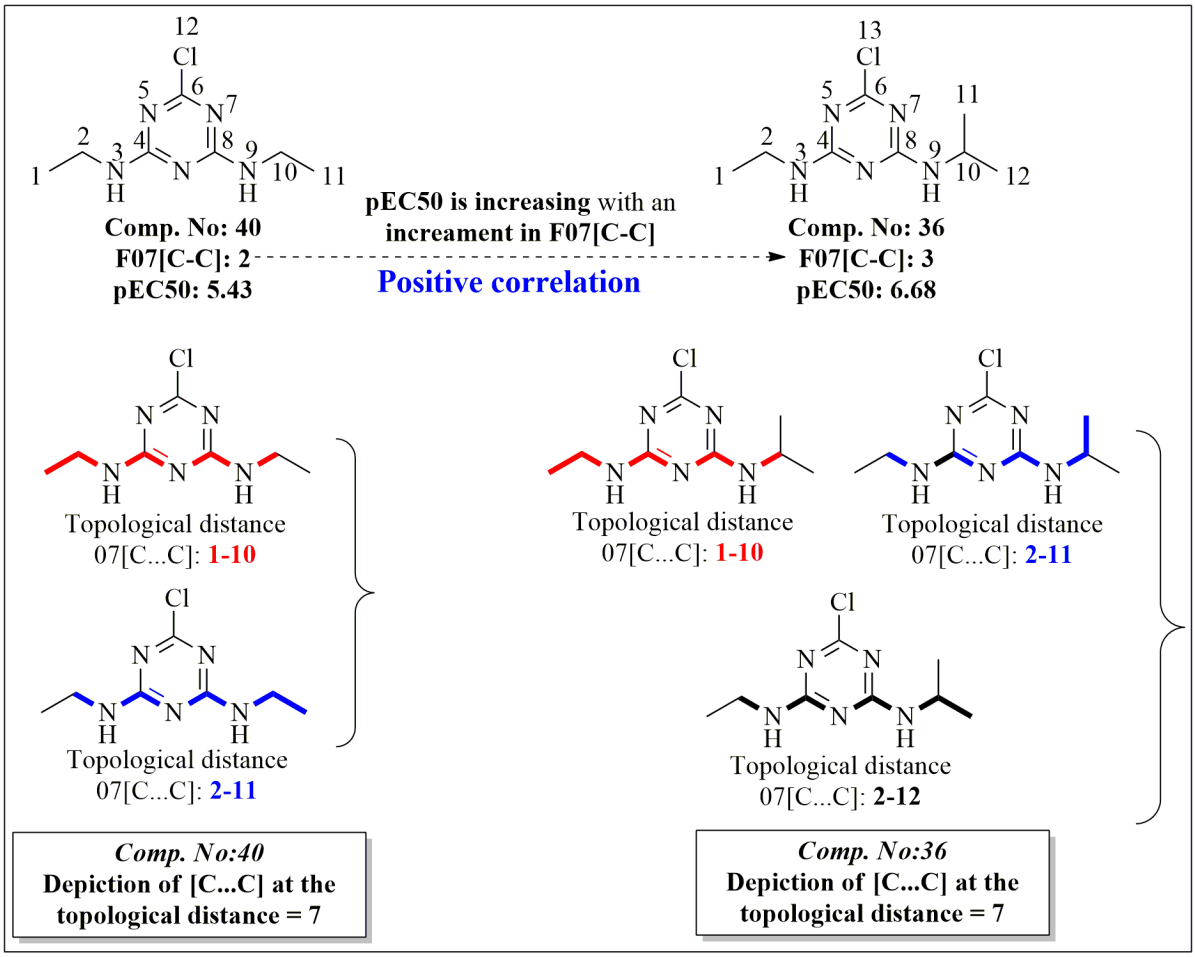


Figure 4.8. Contribution of F07[C-C] to pEC50

Toxicity of chemicals to microorganisms mainly depends on the penetration into the cell and the effective binding to the target protein. Hy is the next significant descriptor of Model 2; this is a measure of hydrophilic factor. More hydrophilic nature is detrimental to the penetration of organic molecules into cell through the hydrophobic cell membrane and hence Hy is negatively correlated with the modeled response. This negative correlation can be justified if we compare the descriptor values of prometon (Comp. No. **39**) and simazine (Comp. No. **40**).We can see the presence of two H-bond donor groups (-NH) in both of the compounds. However, due to the presence of more sterically hindering groups (-isopropyl) in prometon, the chances of H-bond formation are less in prometon and hence the corresponding hydrophilic factor will also be less. In contrast, because of the presence of less sterically hindering substituent (-ethyl), simazine has more chances to form H-bond and hence be more hydrophilic. In another pair of compounds, chlorotoluron (Comp. No. **47**) and chlorosulfuron (Comp. No. **54**), an increase of hydrophilic factor as observed with an increase in the number of H-bond donor groups in chlorosulfuron results in a corresponding decrease in the toxicity (**Figure 4.9**).

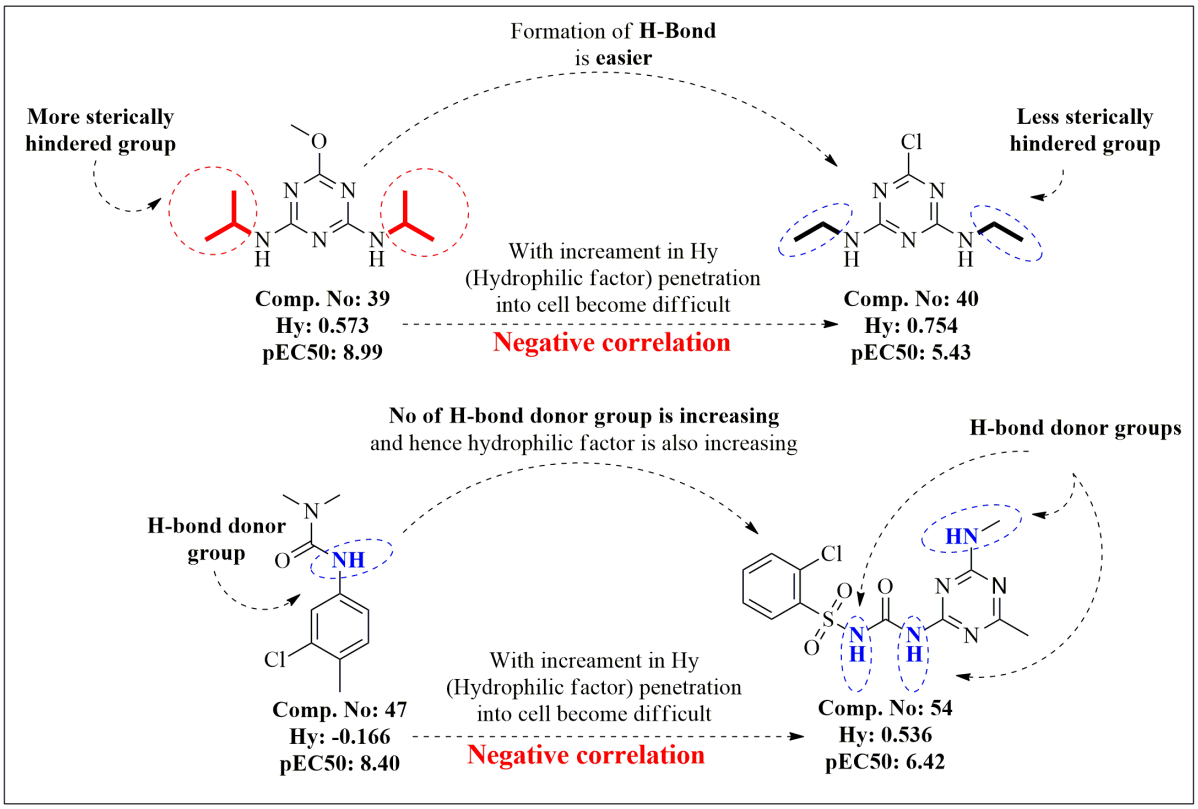


Figure 4.9. Contribution of Hy into pEC50

B04[N-Cl] is the least important descriptor which is negatively correlated with the modeled response. This descriptor identifies the presence of nitrogen and chlorine atoms at the topological distance four. The B04[N-Cl] values, e.g., 1 and 0, suggest the presence and absence respectively of the N and Cl at the topological distance four. In the dataset, compounds like metribuzine (Comp. No. **41**) and monuron (Comp. No. **49**) do not have this feature, and hence are more toxic. On contrary, compounds like terbuthylazine (Comp. No. **44**) and chlorotoluron (Comp. No. **47**) having nonzero descriptor values are less toxic. Therefore, the negative correlation of B04[N-Cl] with pEC50 value is demonstrated (**Figure 4.10**).

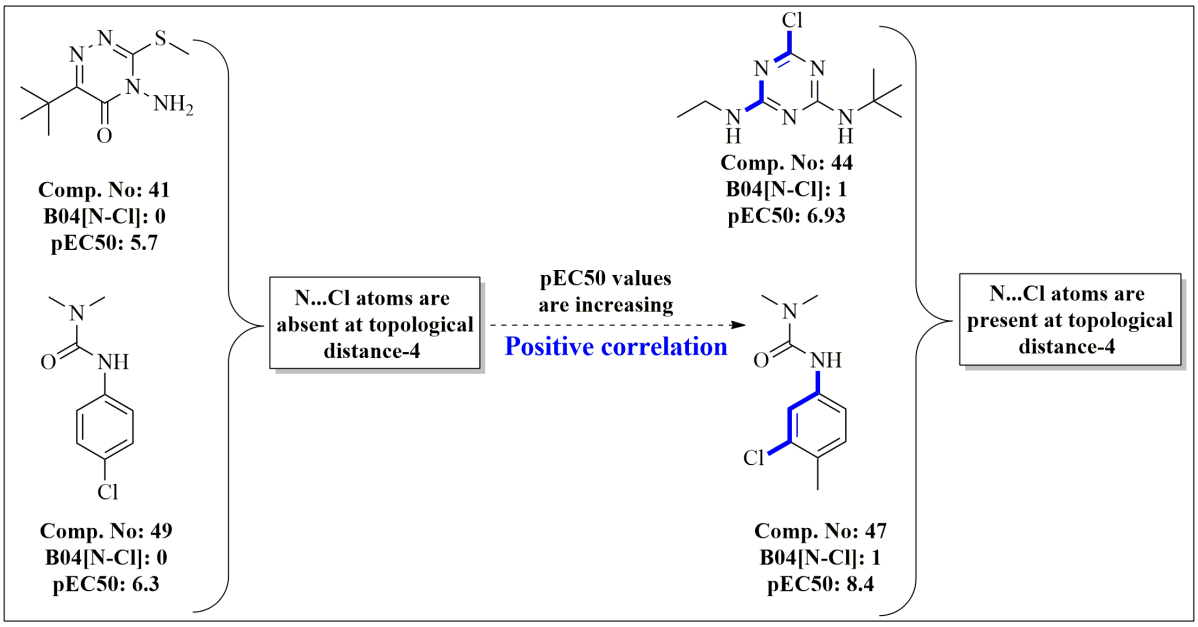


Figure 4.10. Contribution of B04[N-Cl] to pEC50

4.1.3. QSAR study of toxicity of mixtures (Type-1) to P. phosphoreum

For the Type-1 mixture dataset, we have developed a PLS regression model (Model 3) with 2 latent variables. This model is composed of six descriptors and can be illustrated in the following equation 4.3-

$$pEC50 = 4.75139 - 0.00012 \times B03[N-O] - 1.58803 \times nR{=}Cp - 0.16606 \times nROH - 0.94445 \times N-074 - 0.35592 \times C-006 + 0.10023 \times F07[O-O] \quad (4.3)$$

N_{Train}= 60; R^2= 0. 76; R^2_{adj}= 0.73; Q^2LOO= 0.70 (2 LV); $avgr_m{}^2_{(LOO)}$=0.60; $\Delta r_m{}^2_{(LOO)}$=0.13; MAE_{Train}= 0.35; N_{Test}= 19; Q^2_{F1}= 0.82; Q^2_{F2}= 0.82; $avgr_m{}^2_{(test)}$=0.70;$\Delta r_m{}^2_{(test)}$=0.13; MAE_{Test}= 0. 31

The well accepted internal validation parameter values (R^2= 0.76, R^2_{adj}= 0.73, Q^2_{LOO}= 0.70) suggest the statistical reliability and robustness of this model. Model 3 has good predictive ability also as it shows acceptable values of external validation parameters (Q^2_{F1}= 0.82, Q^2_{F2}= 0.82). Among the modeled descriptors, B03[N-O], nR=Cp, nROH, N-074, and C-006 have negative contributions whereas only F07[O-O] has a positive contribution towards the modeled response as suggested by the regression coefficient plot of Model 3 (**Figure 4.11A**). The variable importance plot (**Figure 4.11B**) of this model suggests the order of importance of the modeled descriptors and the order is: N-074> nR=Cp> B03[N-O]> C-006> F07[O-O]> nROH.

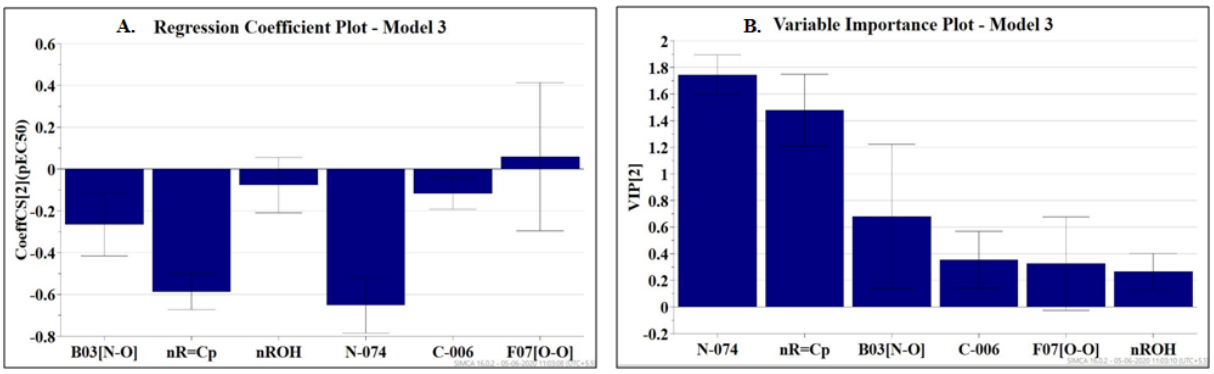


Figure 4.11. A. Regression Coefficient Plot of Model 3; B. Variable Importance Plot of Model 3

N-074 is an atom centered fragment descriptor and the most significant one to demonstrate the role of structural features of mixture components concerning the toxicity end point (pEC50) in this PLS model. This descriptor computes the number of –R#N/-R=N- fragments into the molecule (#: triple bond; =: double bond; R: any group linked through carbon atom). Mixture components like malononitrile (Comp. No. **14**) and glyconitrile (Comp. No. **15**) have this structural feature, whereas propionaldehyde (Comp. No. **2**) and benzaldehyde (Comp. No. **5**) have no such fragment in their structures (**Figure 4.12**). Hence mixtures containing malononotrile or glyconitrile have non-zero descriptor values (e.g., Mix. No. **2**, **14**). It has been observed that a mixture (e.g. Mix. No. **2**) with a higher descriptor value is less toxic than the mixture (e.g.,Mix. No. **14**) with a lower descriptor value which clearly justifies the negative correlation of N-074 with the modeled response (**Table 4.3**).

nR=Cp is another important descriptor of Model 3. It is a functional group count descriptor which computes the number of terminal sp^2-hybridized carbon atom in the molecular structure of corresponding compounds. The negative contribution of this descriptor towards the response suggests that a mixture with lower descriptor value is more toxic than a mixture with higher descriptor value. Type-1 mixtures like **38**, **40**, **41**, and **15** are the equimolar (1:1) mixtures of the Series 1 compounds (**Table 4.3**). Among the components of above mentioned mixtures, allyl cyanide (Comp. No. **17**) contains terminal C (sp^2) while the rest have no such terminal carbon (**Figure 4.12**). It has been noted that only the allyl cyanide containing mixtures (e.g. Mix. No. **38**, **40**, **41**) have non-zero descriptor values and they show similar and less toxicity than the mixtures (e.g. Mix. No. **15**) with a lower (zero) descriptor value.
B03[N-O] is a 2D atom pair descriptor which describes the presence or absence of nitrogen and oxygen atoms at the topological distance 3. This descriptor is negatively correlated with the modeled response which means that the pEC50 value rises with an increase in the descriptor value and vice-verse. Type-1 mixtures such as **67**, **72**, **75**, and **78** are the mixtures of Set 1 compounds sulfamethoxazole (Comp. No. **27**) and trimethoprim (Comp. No. **35**) in different toxic ratios (**Table 4.3**). Sulfamethoxazole contains a pair of nitrogen and oxygen at the topological distance 3 in its molecular structure, whereas trimethoprim is completely devoid of this feature (**Figure 4.12**). It has been observed that the B03[N-O] values get increased with an increase in the percentage of sulfamethoxazole in the mixtures and the corresponding toxicity values get decreased (e.g., Mix. No. **67**, **72**, and **75)**. However, the presence of higher percentage of trimethoprim than sulfamethoxazole in the mixture does not affect the descriptor value as seen in mixture **78**.

C-006 is the next important descriptor of Model 3 which is an atom centered fragments count, and it particularly suggests the number of $–CH_2RX$ fragment (R: any group linked through C; X: heteroatoms such as O, N, P, S, Se, halogens) in the molecule. The Set 1 compounds except glycolonitrile (Comp. No. **15**) are devoid of this feature (**Figure 4.13**). C-006 is negatively correlated with the response which can be clearly demonstrated by the data enlisted in **Table 4.3**. The mixtures with higher descriptor values contain glycolonitrile as a component as seen Mixture no. **15** and **18**. It has been observed that the mixtures with higher descriptor values (e.g.,Mix. No. **15** and **18**) showed less toxicity than the mixtures with lower descriptor values (Mix. No. **27** and **52**).

F07[O-O] is again a 2D atom pair descriptor of the developed Model 3. It is the frequency count of oxygen atom pairs at the topological distance 7. Among the type-1 mixture components, *p*-nitrobenzaldehyde (Comp. No. **6**) and *p*-terephthaldehyde (Comp. No. **7**) have this feature in their structures (**Figure 4.13**) and hence the mixtures containing these two components (e.g., Mix. No. **27** and **54**) have non-zero descriptor values. F07[O-O] is the only descriptor of Model 3 which have positive contribution towards modeled response. It has been observed that the toxicity increases with an increase in the descriptor value (**Table 4.3**) which clearly justifies the positive contribution of this descriptor.

The least important descriptor of Model 3 is nROH which is a functional group count descriptor. It computes the number of hydroxyl groups present in the chemical structures of compounds. It is a negatively contributing descriptor, i.e., the mixtures with lower descriptor values are more toxic than the mixtures with lower descriptor values. This feature can be identified in type-1 mixture component like glycolonitrile (Comp. No. **15**) (**Figure 4.13**). **Table 4.3** demonstrates that the mixtures (e.g., Mix No. **15**, **18**) with higher nROH values have lower toxicity as compared to the mixtures (e.g., Mix. No. **69**) with lower descriptor values and these observations clearly address the negative correlation of nROH with pEC50.

Table 4.3. Demonstration of the role of descriptors to the toxicity of Type-1 mixtures

Descriptor: N-074				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
2	#2:#14	1:1	2	2.63
14	#5:#15	1:1	1	3.15
Descriptor: nR=Cp				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
38	#6:#17	1:1	1	2.36
40	#8:#17	1:1	1	2.36
41	#10:#17	1:1	1	2.36
15	#6:#15	1:1	0	3.26
Descriptor: B03[N-O]				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
67	#27:#35	2747:1	2747	3.49
72	#27:#35	85:1	85	3.67

75	#27:#35	4:1	4	4.52
78	#27:#35	1:37	1	5.45
Descriptor: C-006				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
15	#6:#15	1:1	1	3.26
18	#9:#15	1:1	1	3.26
27	#7:#16	1:1	0	3.78
52	#13:#18	1:1	0	3.78
Descriptor: F07[O-O]				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
22	#1:#16	1:1	0	2.64
27	#7:#16	1:1	1	3.78
54	#6:#19	1:1	2	4.25
Descriptor: nROH				
Mix. No.	**Components in mixture**	**Toxic ratio**	**Hypothetical descriptor**	**pEC50**
15	#6:#15	1:1	1	3.26
18	#9:#15	1:1	1	3.26
69	#27:#35	2747:1	0	3.49

#Individual component

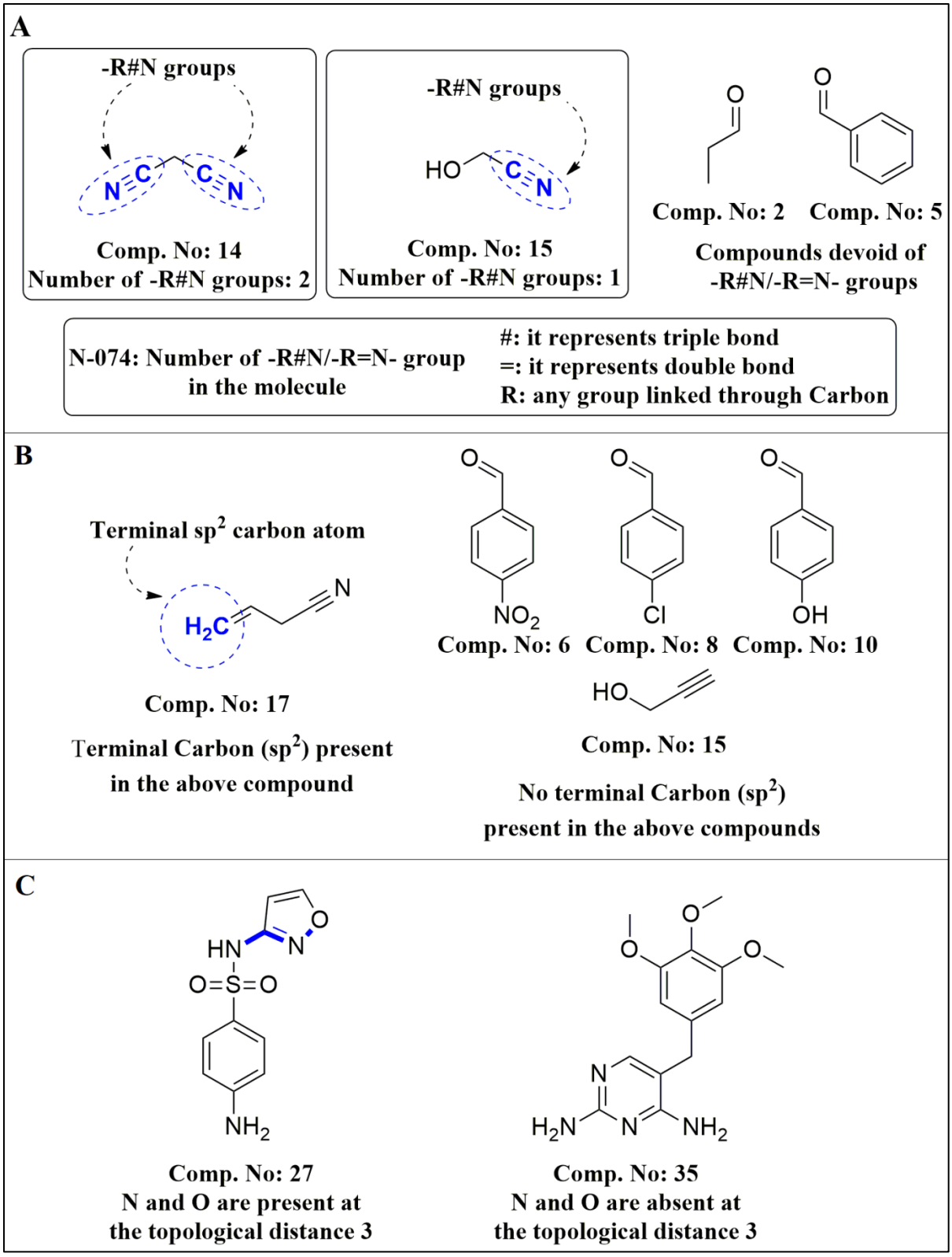


Figure 4.12. Depiction of feature N-074, nR=Cp, and B03[N-O] in the mixture components (A. represents N-074; B. represents nR=Cp; C. represents B03[N-O])

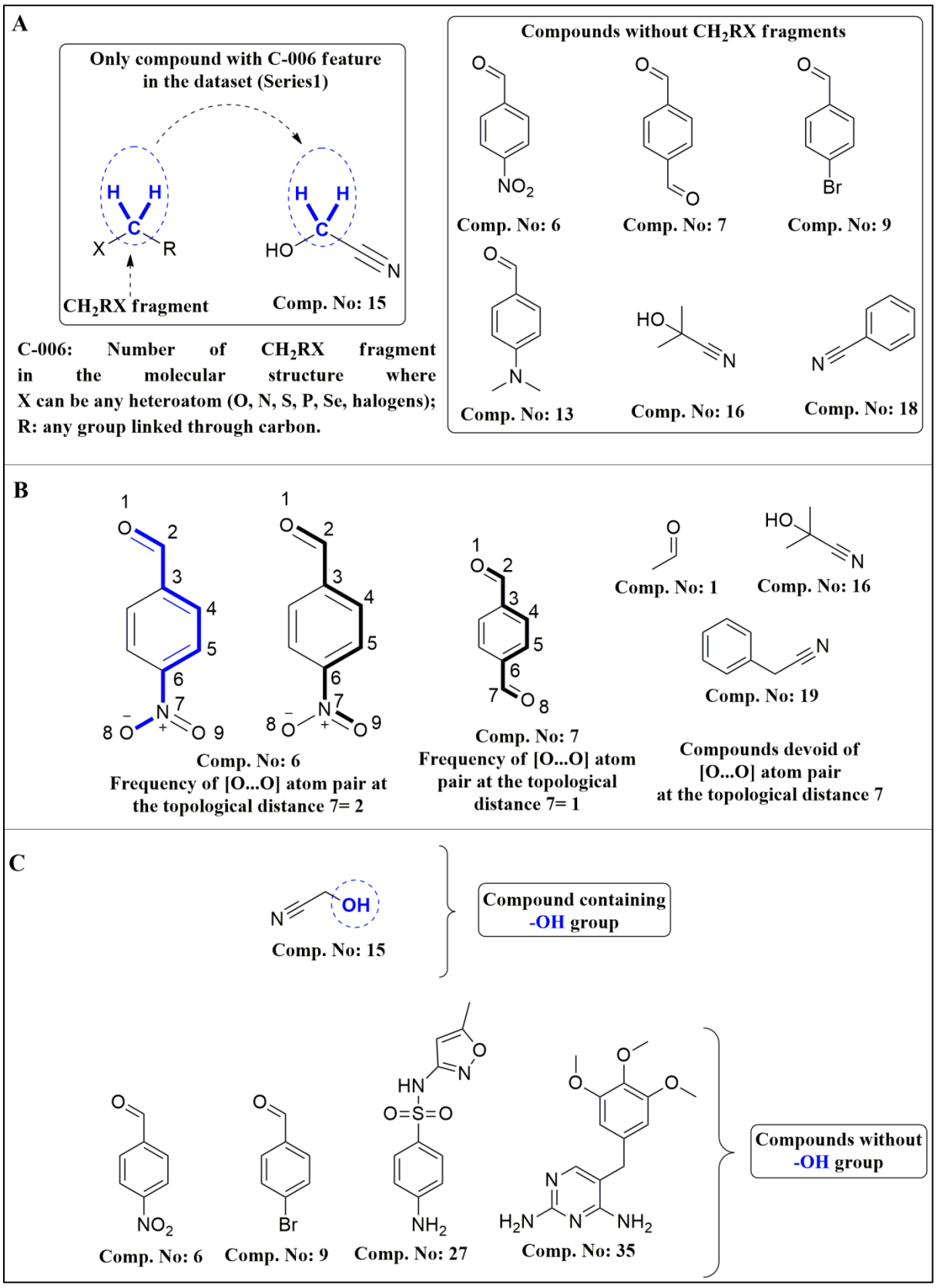


Figure 4.13. Depiction of feature C-006, F07[O-O], and nROH in the mixture components (A. represents C-006; B. represents F07[O-O]; C. represents nROH)

4.1.4. QSAR study of toxicity of mixtutes (Type-2) to S. capricornutum

The 4th and final model has been composed of 3 descriptors with 2 latent variables and is shown in equation 4.4-

$$pEC50 = -12.38041 + 1.34717 \times ETA_{Alpha} - 0.25978 \times F06[C-S] + 0.13686 \times nCsp^3 \quad (4.4)$$

N_{Train}= 15; R^2= 0.82; R^2adj= 0.72; Q^2LOO= 0.73 (2 LV); $avgr_m{}^2{}_{(LOO)}$=0.64; $\Delta r_m{}^2{}_{(LOO)}$=0.09; MAE_{Train}= 0.35; N_{Test}= 5; $Q^2{}_{F1}$= 0.82; $Q^2{}_{F2}$= 0. 77; $avgr_m{}^2{}_{(test)}$=0.79;$\Delta r_m{}^2{}_{(test)}$=0.04; MAE_{Test}= 0.37

This model was constructed for theType-2 mixture dataset. The acceptable values of validation parameters (*R^2= 0.82; R^2adj= 0.72; Q^2LOO= 0.73*) confirm the statistical reliability and robustness of Model 4. Among the three descriptors, ETA_Alpha and $nCsp^3$ have positive contributions whereas F06[C-S] has a negative contribution towards the toxicity of the mixtures (**Figure 4.14A**). As per the variable importance plot of this model (**Figure 4.14B**), the significance level of the model descriptors are found to be in the following order: ETA_Alpha> F06[C-S]> $nCsp^3$.

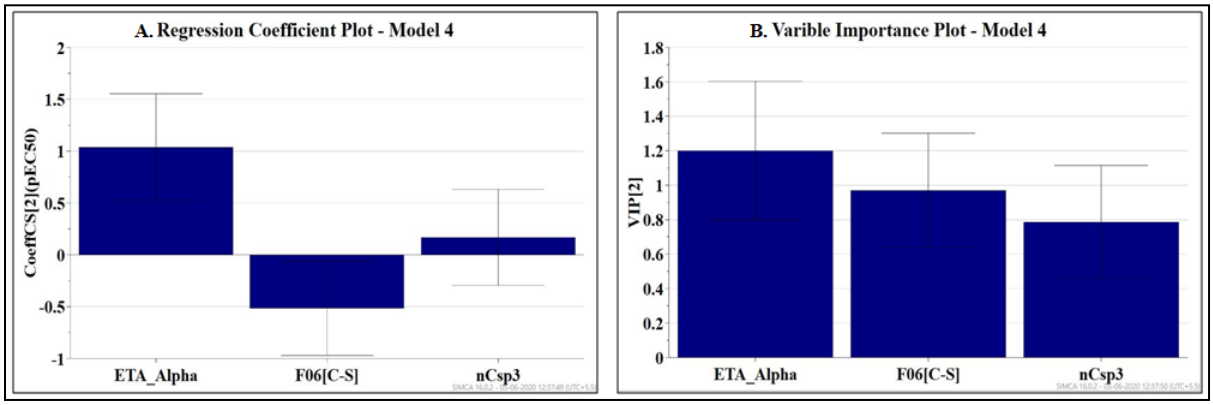


Figure 4.14. A. Regression Coefficient Plot of Model 4; B. Variable Importance Plot of Model 4

ETA_Alpha (∑α), a 1st generation Extended Topochemical Atom (ETA) descriptor, is a measure of molecular bulk of compounds. Molecular bulk is an important feature of chemical compounds for penetrating the cell wall of microorganisms and at the same time for binding to the receptor also. ∑α has a maximum and positive contribution to the response which suggests a proportional relationship between ∑α and pEC50. Mixture No. **82** and **98** are two equimolar Type-2 mixtures containing monolinuron **(**Comp. No. **48)** as a constant component. It has been observed that Mix. No. **82** has a higher ∑α value and correspondingly a higher pEC50 value due to the presence of more bulky 2nd component (Comp. No. **36**) than Mix. No. **98** (**Figure 4.15**). Another pair of mixtures (Mix. No. **84** and **87**) have also shown the proportional relationship of ETA_Alpha and the toxicity.

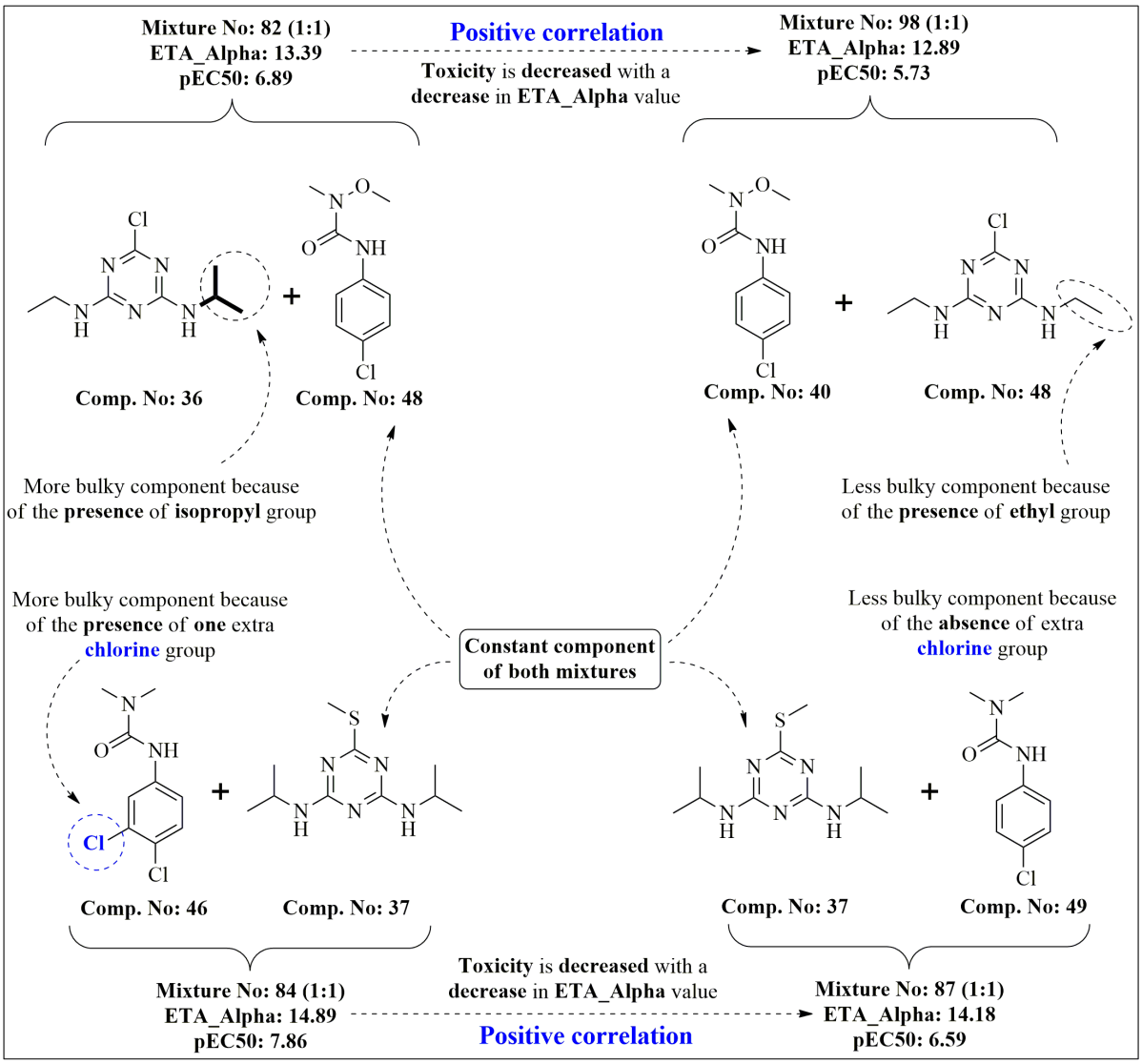


Figure 4.15. Contribution of ETA_Alpha to pEC50 of Type-2 mixtures

The next important descriptor of Model 4 is F06[C-S]. It is a 2D atom pair descriptor which calculates the number of occurrences of carbon-sulphur atom pair at the topological distance six. It is the only negative contributing descriptor of Model 4; so, it has an inversely proportional relationship to the modeled response. The feature has been identified in the mixture component like symetryn (Comp. No. **38**) and is depicted in **Figure 4.16.** Components having [C-S] atom pair at the topological distance 6 have the key effect to raise the descriptor values as seen in **Table 4.4**. We have seen that more toxic mixtures (e.g., Mix. No. **92**) have lower descriptor values than less toxic mixtures (e.g., Mix. No. **88**).

Table 4.4. Effect of F06[C-S] to the pEC50

Mix. No.	Components in mixture	Toxic ratio	Hypothetical descriptor (F06[C-S]) value	pEC50
88	#38:#46	1:1	2	6.58
92	#39:#46	1:1	0	8.00

#: individual compound

The last significant descriptor of Model 4 is $nCsp^3$ which computes the number of sp^3-hybridized carbon atoms. It is a constitutional descriptor which has a positive contribution to the toxicity of type-2 mixtures. Two observations have been depicted in **Figure 4.16** where the sp^3- hybridized carbon atoms are marked in the mixture components. An increase in pEC50 value has been observed with an increase in the descriptor value (e.g., Mix. nos. **99** and **95**).

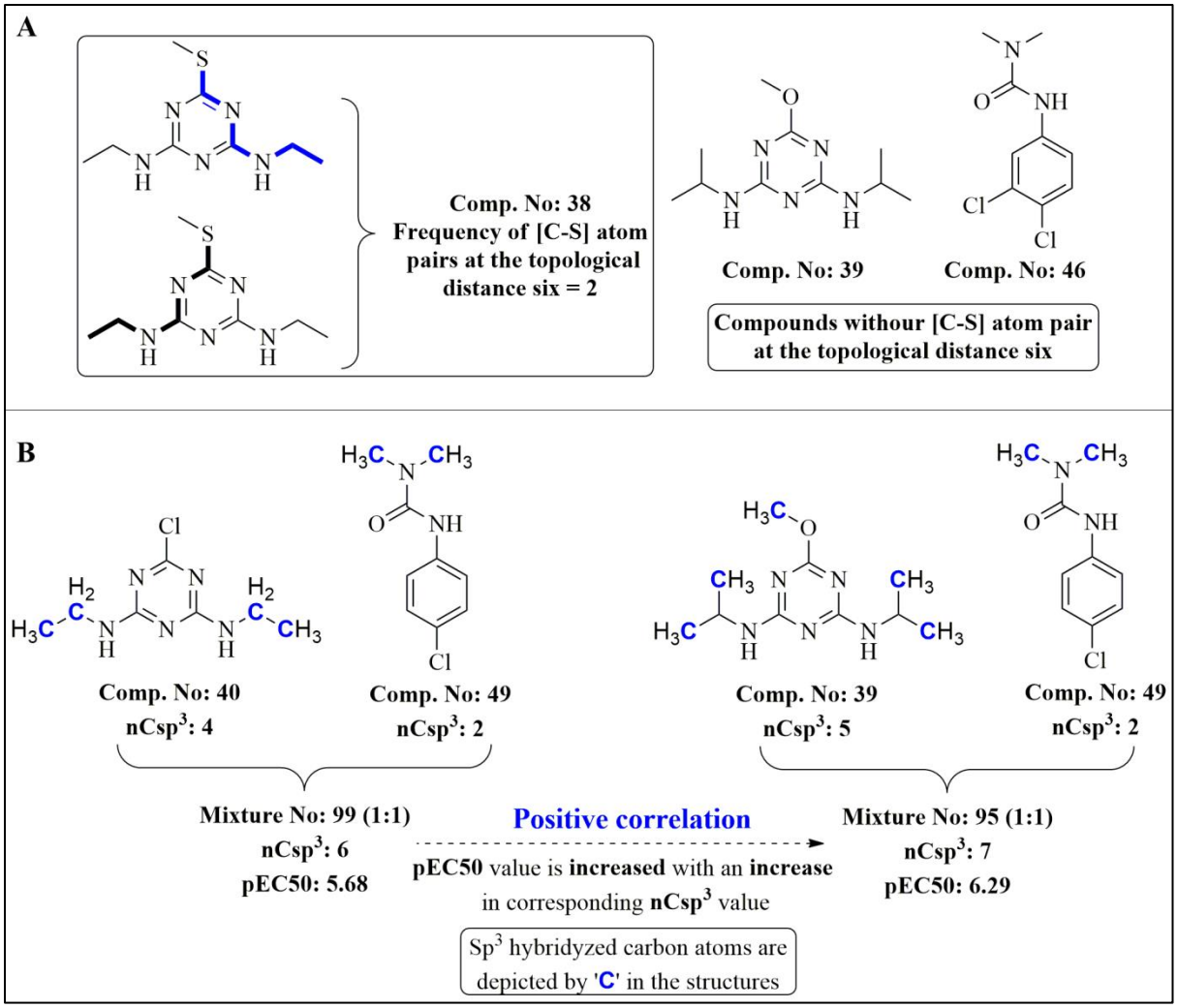


Figure 4.16. Depiction of features F06[C-S] and $nCsp^3$ in type 2 mixture components (A. represents F06[C-S]; B. represents $nCsp^3$)

4.1.5. Applicability domain and Y-randomization of the developed PLS models

From **Figure 4.17**, we can see that the test set compounds of each developed model have fallen into the applicability domain of corresponding training sets at a 99% confidence interval (D-critical = 0.009999).

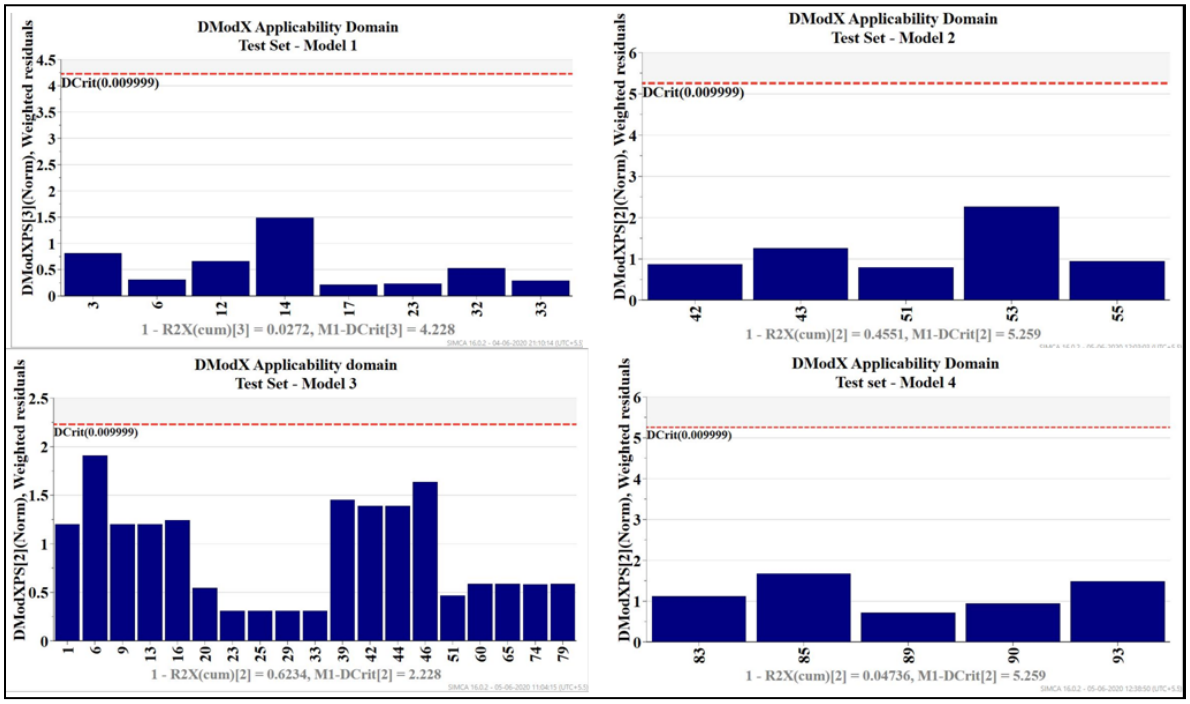


Figure 4.17. Applicability domain analysis of the test set compounds of the developed four models

We have also analyzed the models by randomization plots in order to confirm that the developed models are not the results of any chance correlation. In the randomization approach, a number of models are constructed randomly by permuting different combinations of descriptors (X) or response (Y) variables based on the fit of already recorded models (Topliss and Edwards, 1979). In this work, for the training set, the X variable data remained unaltered and the Y data were shuffled randomly (Y-randomization). The recorded model was fitted to the permuted data and compared with the best fit. The R^2 and Q^2 values of the random models (Y-axis) are plotted against the correlation coefficient between the original Y values and the permuted Y values (X axis). The statistical metrics (R^2 and Q^2) of the randomization model should be poor (R_Y^2 intercept <0.3; Q_Y^2 intercept <0.05) and not within the acceptable range of those for regression models. Otherwise the constructed models may be considered as the results of chance correlation (Rücker et al., 2007). Here we have used 100 permutations for the analysis and the obtained plots of four models have been shown in **Figure 4.18**. The R_Y^2 intercept and Q_Y^2 intercept values have been mentioned in **Table 4.5**, which suggest that all the models are statistically significant and not the results of any chance correlation.

Table 4.5. Statistical parameters of Y-randomization models

Model	Model 1		Model 2		Model 3		Model 4	
Statistical parameters after Y-randomization	R_Y^2	Q_Y^2	R_Y^2	Q_Y^2	R_Y^2	Q_Y^2	R_Y^2	Q_Y^2
	0.03	-0.35	0.05	-0.26	0.04	-0.26	0.05	-0.30

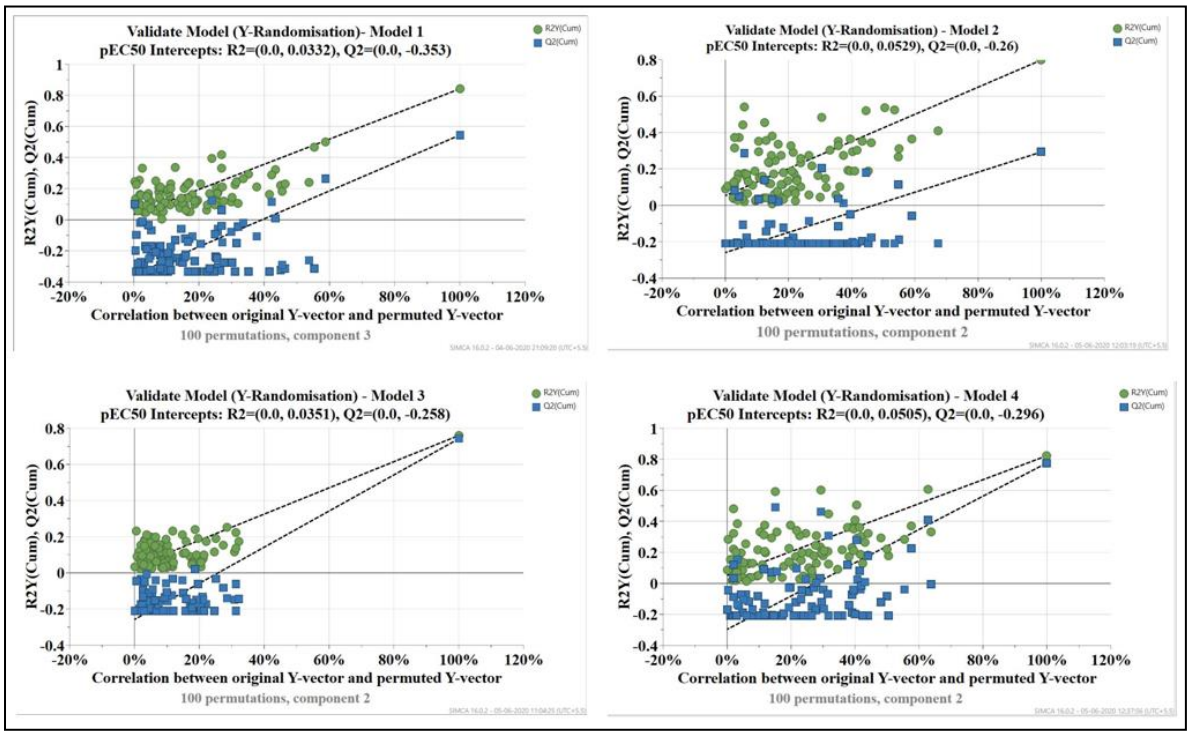


Figure 4.18. Y-randomization plots of Model 1, Model 2, Model 3, and Model 4

4.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

4.2.1. The developed model and its validation

A PLS model (with 3 latent variables) was developed from 63 data points (training set), reported in the materials and methods section. The model essentially reports 4 important structural features, responsible for the mixture toxicity. The model has been represented in the following equation 4.5-

$$\log 1/EC50 = 2.34406 - 0.16478 \times N\% + 0.22706 \times T(N..Cl) + 0.98204 \times B03[O-Cl] + 0.1175 \times MLOGP2 \qquad (4.5)$$

The reported model was validated rigorously following the OECD Principle 4. We have computed both the statistical validation metrics and reported them in the following **Table 4.6**. All the computed metric values are in the acceptable range (as per the OECD protocol), and it suggests the statistical credibility of the model. Statistical quality metrics like the R^2 (0.82) and R^2_{adj} (0.81) values indicate the goodness-of-fit whereas the Q^2_{LOO} (0.78) value indicates

the robustness of the developed model. The good predictive ability of the model can be justified by the computed external validation metric values such as Q^2_{F1} (0.87). The model shows low prediction errors which can be justified by the small values of $MAE_{Train;\ LOO}$ (0.24) and $RMSE_P$ (0.26).

Table 4.6. Computed values of statistical validation metrics

Internal Validation Metrics	**Computed Value**
Data points in the training set	63
R^2_{Train}	0.82
R^2_{adj}	0.81
Q^2_{LOO}	0.78
Scaled Average $r^2_{m\ (Train;\ LOO)}$	0.70
Scaled $\Delta r^2_{m\ (Train;\ LOO)}$	0.14
MAE (Train; LOO)	0.24
External validation metrics	**Computed values**
Data points in the test set	21
Q^2_{F1}	0.87
Q^2_{F2}	0.87
Scaled average $r^2_{m\ (Test)}$	0.83
Scaled $\Delta r^2_{m\ (Test)}$	0.005
MAE (Test)	0.22
$RMSE_P$	0.26

LOO: Leave-one-out; MAE: Mean absolute error; $RMSE_P$: Root mean square error of prediction

Various plots have been incorporated in this study to graphically represent various aspects of the developed model. In **Figure 4.19**, the scatter plot (A), the Y-randomization plot (B), the applicability domain column plot (C), and the score plot (D) have been reported. An observed vs. predicted response scatter plot indicates the overall correlation between the observed and predicted response values. A good correlation can be identified by the scattering of plotted observed vs. predicted data at a particular region of the chemical space. It has been observed in **Figure 4.19A** that the plotted observed vs. predicted response data are scattered (for both the training and test sets) within a limited region along the axis passing through the origin. It justifies a good correlation between the observed and predicted response and hence the goodness-of-fit of the model. We have performed the Y-randomization test, computed the resulting R^2 and Q^2, and shown the corresponding plot in **Figure 4.19B**. The poor R^2 (- 0.0195) and Q^2 (- 0.325) intercepts suggest that the developed model is not the result of any chance correlation. Finally, we have also identified the applicability domain of the developed model and represented in **Figure 4.19C**, to comply with the OECD principle 3. It has been observed from the applicability domain column plot that all the test mixtures are inside the

applicability domain of the training mixtures of the developed model. Therefore, the predicted responses of the test set compounds are reliable enough to decide the predictive ability of the developed model. We have also reported a score plot (Wold et al., 2001) of the PLS model (**Figure 4.19D**) where the training data points are projected in 1st 2LV space of the developed model. It has been observed that the data points are scattered throughout the LV space and they are not clustered as such. This observation explains the structural diversity of the data points.

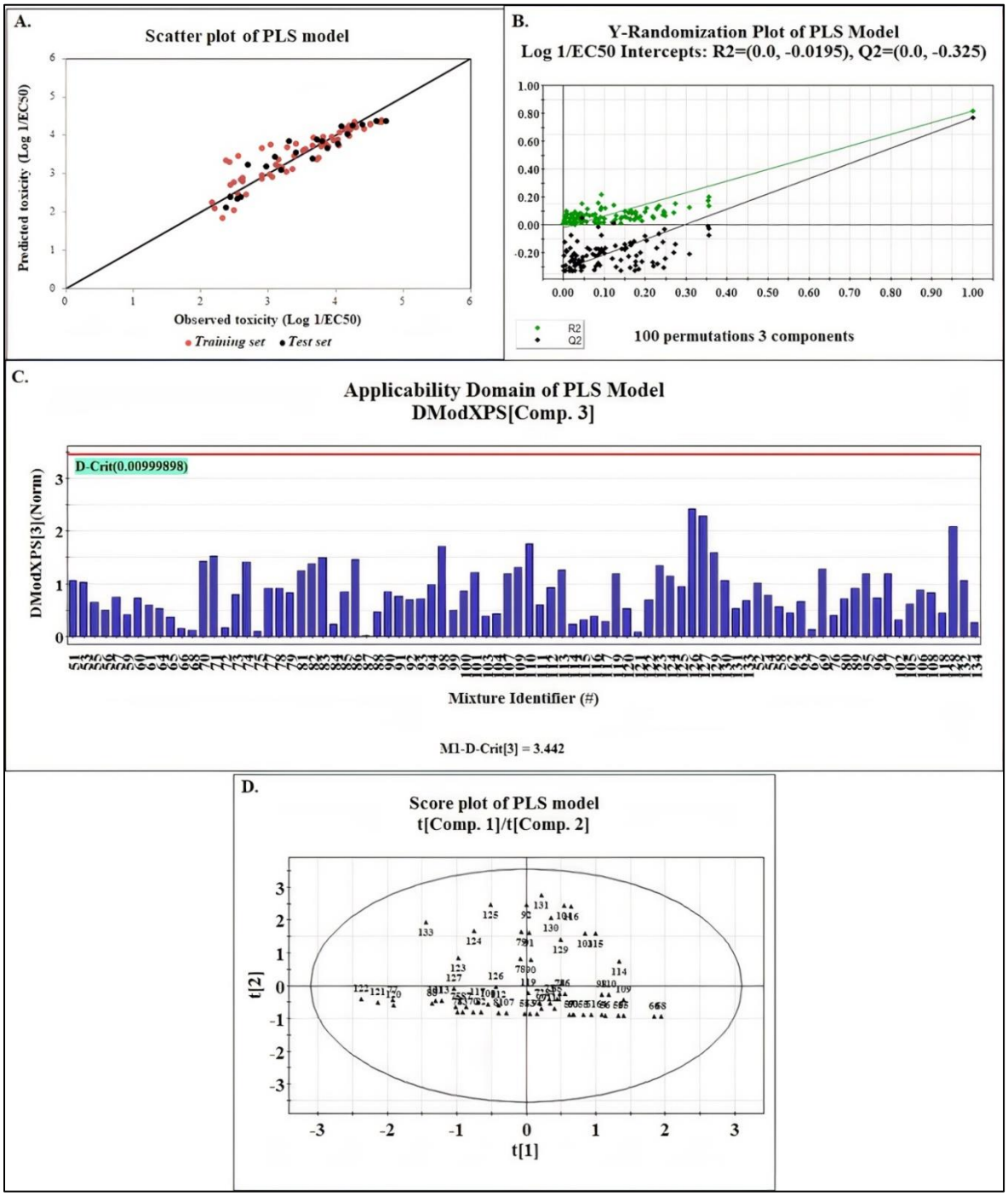


Figure 4.19. Various plots of developed PLS model; (A) Scatter plot (observed Vs. predicted toxicity); (B) Y-randomization plot; (C) Applicability domain column plot (D) Score plot of PLS model

4.2.2. Result of cross-validation and prediction reliability indicator

We have performed two rigorous cross-validations in this study as discussed in the section '***3.2.6. Cross-validation***'. To obtain the cross-validated metrics, the new divisions (available at online repository: https://github.com/mchatt1012/PhD-Thesis_QSAR-Modeling-of-Toxicity-of-Chemical-Mixtures.git) have been used. We have computed the R^2, R^2_{adj}, and Q^2_{LOO}, and Q^2_{F1} as the quality measures. All cross-validated metric values have been tabulated in **Table 4.7** along with the validation parameters of the original PLS model (Equation 4.5). It has been found that the values of cross-validated metrics are very similar to the original PLS model. Although the Q^2_{F1} of compounds out cross-validation is inferior (0.73) as compared to the original PLS model, but still it is in the acceptable range as per the standard protocol. Therefore, the statistical credibility, robustness, and predictive ability of the developed model have been confirmed again by these results of cross-validations.

Table 4.7. Cross-validated metric values

Cross-validated metrics	Mixtures out Cross-validation	Compounds out Cross-validation	Developed PLS model
N_{Train}	60	63	63
R^2	0.82	0.85	0.82
R^2_{adj}	0.81	0.84	0.81
Q^2_{LOO}	0.78	0.83	0.78
N_{Test}	24	21	21
Q^2_{F1}	0.87	**0.73**	0.87

LOO: Leave-one-out; adj: adjusted

We did not get enough mixture toxicity data for this particular endpoint for true external validation. Therefore, we have used the PRI tool to check the reliability of predictions of the developed PLS model with the newly designed mixture dataset (Section "***3.2.8. Prediction reliability indicator and designing of empirical dataset***"). It has been found that all the designed mixtures can be predicted well with the developed model. As far as the composite score is concerned, all the predictions have shown "Good" reliability. Apart from that, the new mixtures were inside the applicability domain of the developed model; therefore, the predicted results obtained by the PRI tool are reliable.

4.2.3. Mechanistic interpretation of modeled descriptors

OECD suggested five main guidelines to the QSAR modelers for developing unbiased, properly validated and mechanistically interpretable models. OECD Principle 5 suggests interpreting the modeled descriptors as far as possible and identifying the effects of structural features towards the response being modeled (molecular properties in general). Therefore, to comply with the 5th guideline, we have identified the modeled descriptors in the dataset and mechanistically shown how the structural features affect the toxicity. Four descriptors were identified during modeling, and they are MLOGP2, T(N..Cl), N%, and B03[O-Cl]. It has been seen from the coefficient plot (**Figure 4.20A**) that among the modeled descriptors-MLOGP2, T(N..Cl), and B03[O-Cl] are positively contributing whereas only N% is negatively contributing to the toxicity. We have also got the knowledge about the relative

importance of the descriptors from the variable importance plot (**Figure 4.20B**) and the relative order of importance is MLOGP2 > T(N..Cl) > N% > B03[O-Cl].

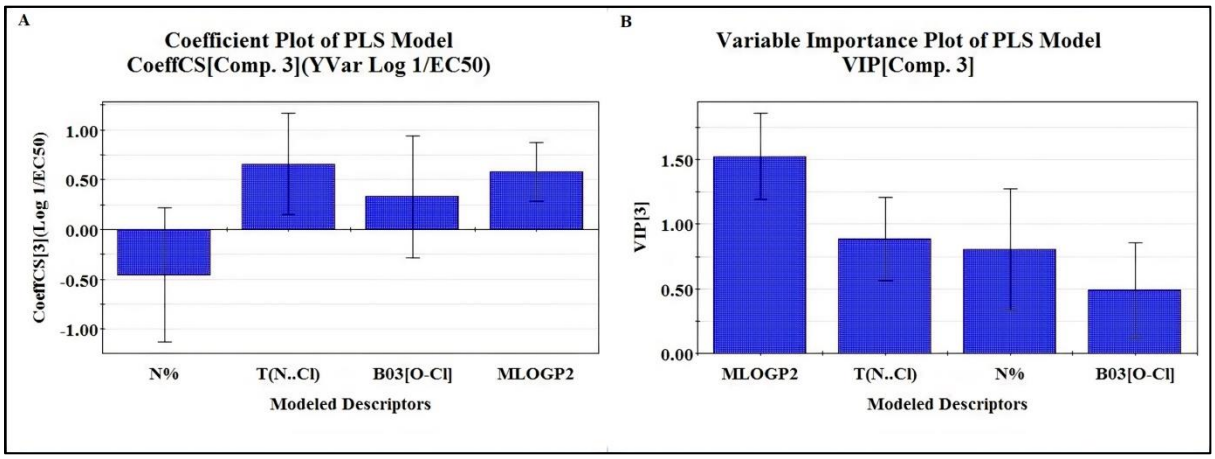


Figure 4.20. A. Coefficient plot and B. Variable importance plot of the developed PLS model

MLOGP2 is a molecular property descriptor that is defined as the value of squared Moriguchi octanol-water partition coefficient of the chemical compounds (chemical mixture in this study), and it essentially suggests the lipophilicity of the mixtures. The value of this descriptor increases with the increment of the lipophilic property of test substances. It has been identified as a positively correlated and the most important descriptor of the developed model (**Figure 4.21**) and hence affects the toxicity to a large extent. In **Figure 4.21**, it has been observed that the *MLOGP2* and *Log 1/EC50* (toxicity) values decrease simultaneously from mixture number 96 to 98 which justifies the positive correlation of the descriptor with the toxicity. Mixture 96 contains a higher percentage of more lipophilic component 1,2,3-trichlorobenzene (#2) and on the contrary, mixture 98 contains a higher percentage of more hydrophilic component 2,4-dichlorophenol (#6). Therefore, it can be concluded from these observations that the lipophilicity of the mixture is an important feature that has a proportional relationship with the toxicity and the presence of more lipophilic components can enhance the aquatic toxicity of mixtures.

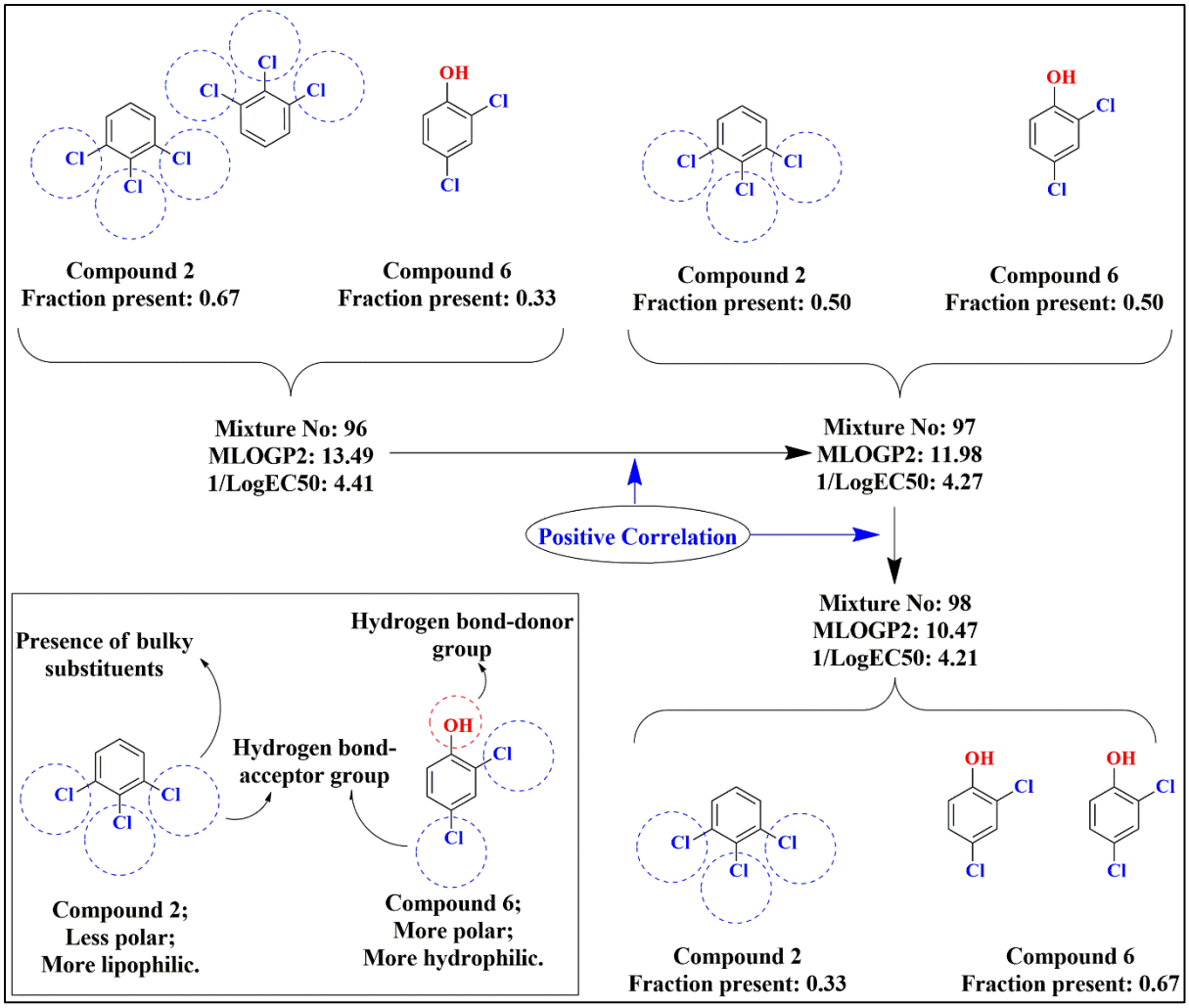


Figure 4.21. Demonstration of the effect of descriptor MLOGP2 towards aquatic toxicity of the mixtures

The next important descriptor in this model is T(N..Cl) (2D atom pairs descriptor), which signifies the collective topological distance between nitrogen and chlorine. Among various mixture components of this study, only 3,4-dichloroaniline (#8) contains this feature and has been depicted in **Figure 4.22**. Nitrogen and chlorine are H-bond donor and H-bond acceptor groups respectively; hence T(N..Cl) represents the hydrogen bonding propensity of molecules or mixture components. The value of this descriptor increases due to the presence of more than one nitrogen or chlorine atom in the structure (as seen in 3,4-dichloroaniline) and this may increase the hydrogen bonding propensity and hence the solubility in the aquatic medium. It has been observed that the mixtures containing the component 3,4-dichloroaniline (#8) only have non-zero values of T(N..Cl) (e.g., mixture **78**, **79**, **80**). The simultaneous increase of the observed toxicity of mixtures with the increment of the descriptor value justifies the positive contributing effect of the descriptor. Therefore, it can be said that the presence of components devoid of the feature T(N..Cl) or with its small value can help to reduce the aquatic toxicity of mixtures by reducing the solubility in the aquatic environment.

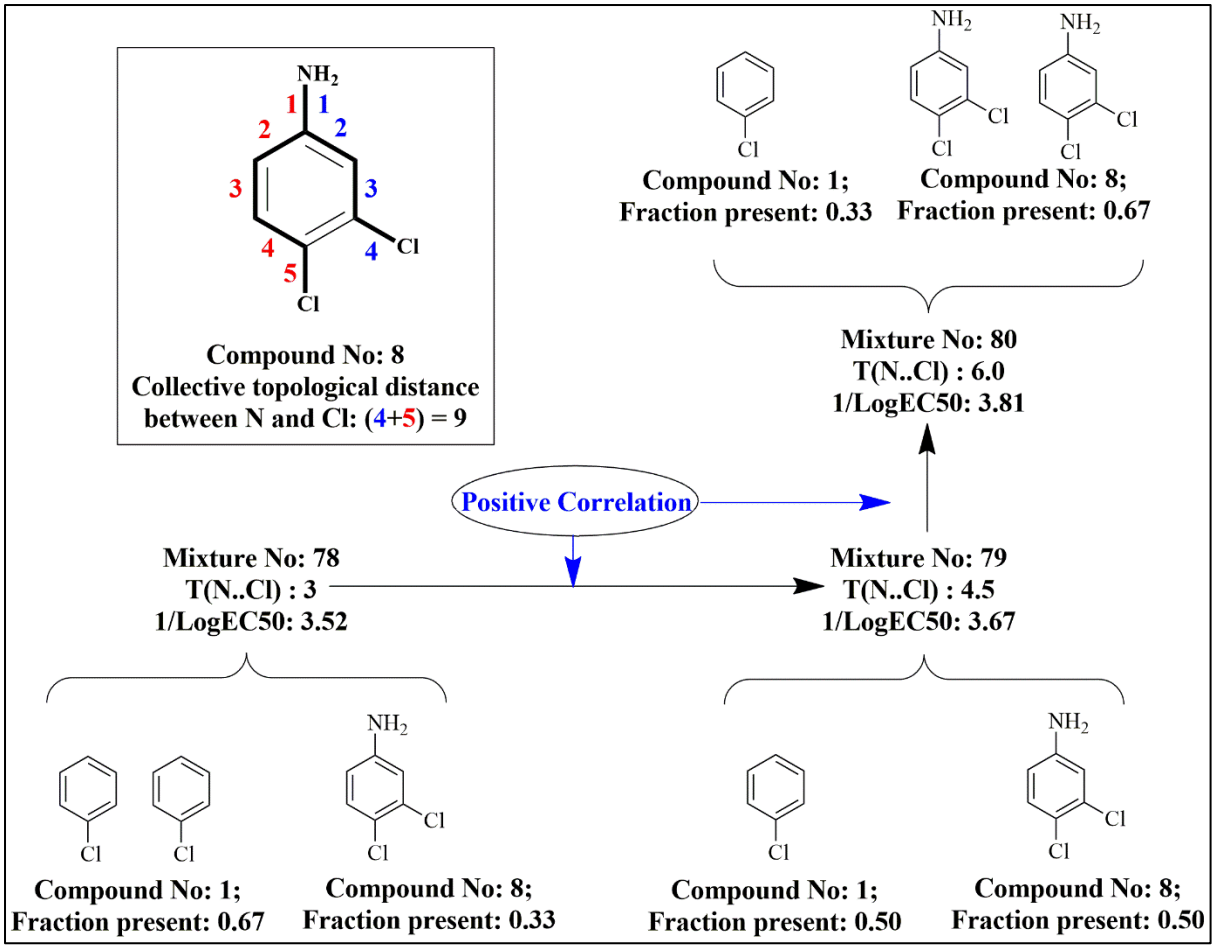


Figure 4.22. Demonstration of the effect of descriptor T(N..Cl) towards aquatic toxicity of the mixtures

N% is a constitutional descriptor, and it contributes negatively to the aquatic toxicity of mixtures. The descriptor is defined as the percentage of nitrogen atoms present in the molecules or mixtures components. Nitrogen is a hydrogen bond donor group, and it can enhance the hydrophilic nature of a molecule by forming H-bonds with water. Hydrophilic compounds are dissolved in aquatic medium but cannot cross the hydrophobic plasma membrane of the organisms easily and hence show less toxicity. In **Figure 4.23**, we have depicted some mixture components like chlorobenzene (#1), 2,4-dichlorophenol (#6), etc. which are devoid of any nitrogen atom, whereas a component like aniline (#7) contains nitrogen. Mixtures containing nitrogen-enriched components (e.g., mixture **76**, **77**) have shown less toxicity than the mixture without nitrogen-enriched components (e.g., mixture **73**). This observation clearly shows the inverse relationship of the mentioned descriptor with toxicity.

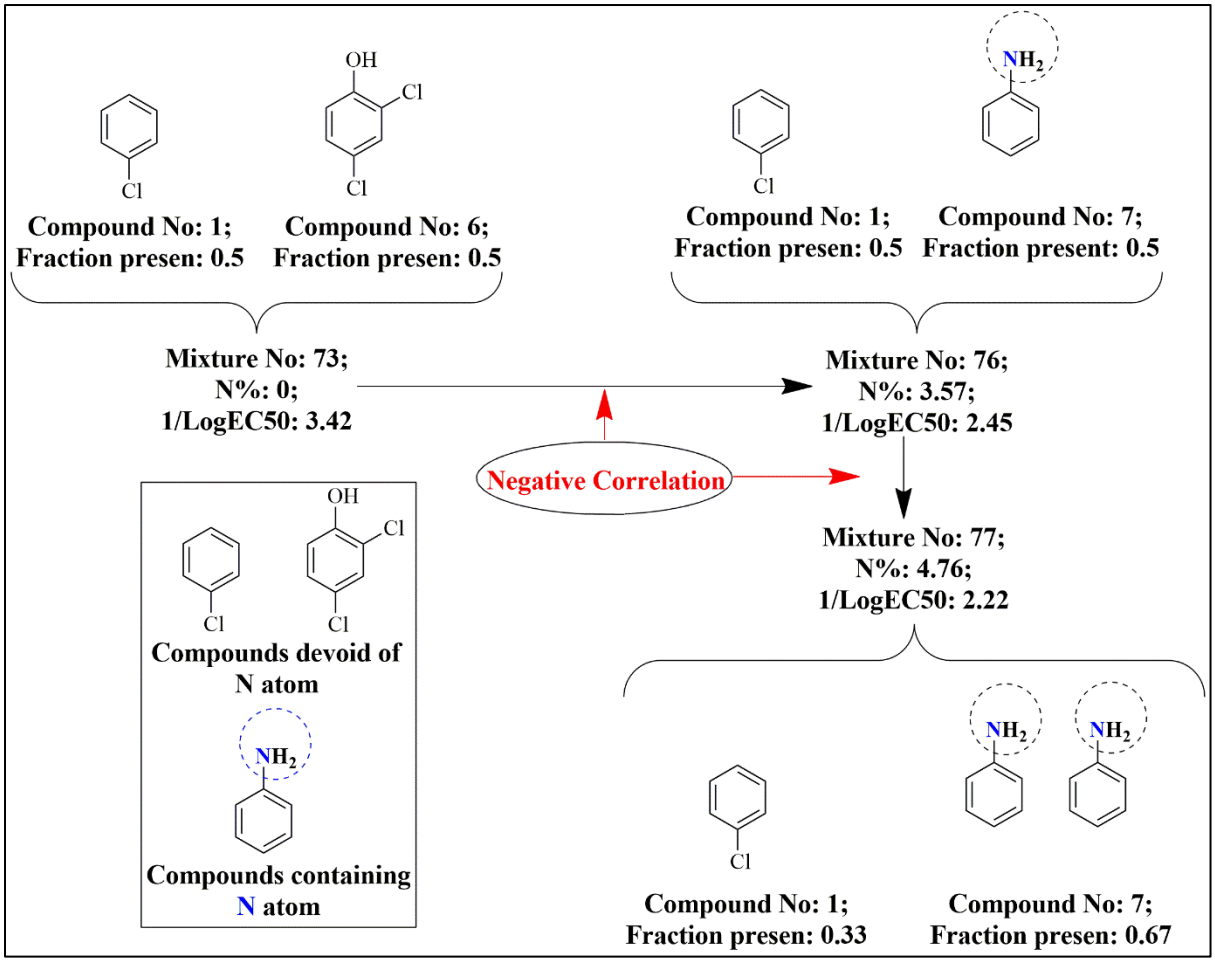


Figure 4.23. Demonstration of the effect of descriptor N% towards aquatic toxicity of the mixtures

The least important descriptor of the developed model is B03[O-Cl] which is a 2D atom pairs descriptor. It is a binary descriptor that indicates the presence (descriptor value 1) or absence (descriptor value 0) of O-Cl atom pairs at topological distance 3. As per the coefficient plot of the PLS model (**Figure 4.20A**), B03[O-Cl] is a positively contributing descriptor. This feature can be seen in the mixture component 2,4-Dichlorophenol (#6) among the studied dataset (**Figure 4.24**). We have depicted the individual components of mixture **106** and **109** in **Figure 4.24**. Oxygen and chlorine can form hydrogen bonds and can be dissolved easily in the aquatic medium. It has been observed that this feature is absent in the components of mixture **106** and that is why it shows mixture descriptor value 0; on the contrary, mixture **109** contains the feature and shows higher toxicity than mixture **106**. This observation indicates the positive correlation of the descriptor with the studied response and shows the importance of solubility to show toxicity against aquatic organisms.

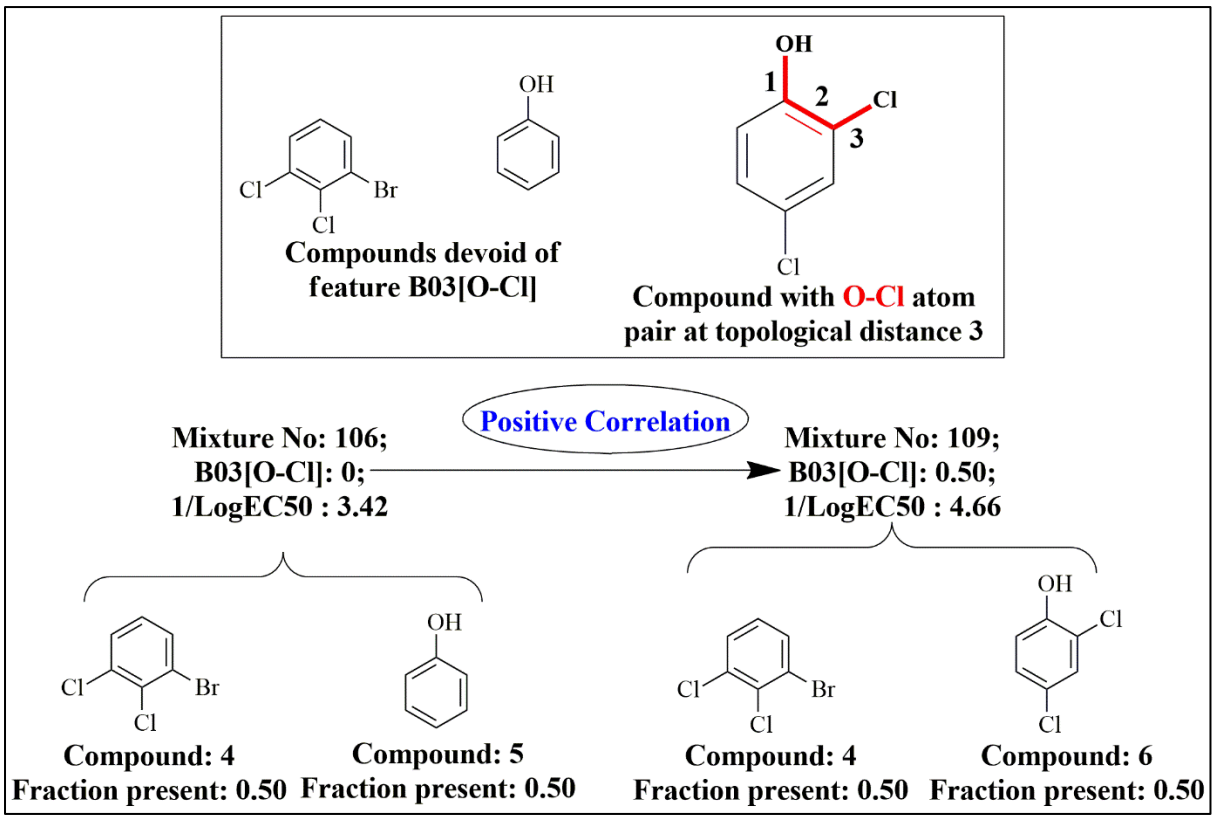


Figure 4.24. Demonstration of the effect of descriptor B03[O-Cl] towards aquatic toxicity of the mixtures

4.2.4. Result of chemical read-across

The result of the chemical read-across has been presented in two steps in this study. The optimization of hyperparameters σ and γ for Gaussian kernel and Laplacian kernel similarity estimations respectively has been performed as the 1st step of machine learning as discussed in the section "***3.2.7. Chemical read-across***". The result of the optimization has been presented in **Table 4.8**. From that table, the optimized values of σ and γ have been obtained and those are 0.5 and 2 respectively.

Table 4.8. Optimization of sigma and gamma values for chemical read-across

σ value	γ value	Q^2_{F1} (GK)	Q^2_{F1} (LK)
0.25	0.25	0.777	0.578
0.5	0.5	**0.789**	0.593
0.75	0.75	0.706	0.607
1.0	1.0	0.626	0.622
1.25	1.25	0.584	0.635
1.5	1.5	0.561	0.647
1.75	1.75	0.548	0.658
2.0	**2.0**	0.539	**0.668**

The optimized setting has been used in the next step to get the predictions of the actual test set. We have applied the number of close training compounds as a determinant of prediction; thus nine predictions have been obtained with 10 to 2 close training compounds. The best prediction has been obtained with 3 close training compounds by Gaussian kernel function similarity and is given in **Table 4.9**. The calculated values of Q^2_{F1} and Q^2_{F2} were 0.94 and 0.94 respectively which suggest good predictivity. The error measures such as RMSEp and MAE have also been computed and the values were 0.18 and 0.15 respectively which are on the lower side. We have given the predicted values of each test mixture along with the standard error of predictions in **Table 4.9**. From the results of chemical read-across predictions, an improvement in the prediction quality of binary mixture has been clearly observed in this study.

Table 4.9. Observed and predicted toxicity (pEC50) values of test set by both QSAR model and read-across prediction (GK)

Mixture No.	**$Y_{Observed}$ (pEC50) (mol/l)**	**$Y_{Predicted}$ (QSAR) (log 1/EC50) (mol/l)**	**$Y_{Predicted}$_Read-across_GK (Standard error of prediction)**
52	3.67	3.37	3.63 (0.12)
54	3.89	3.64	3.68 (0.31)
58	3.74	3.87	3.93 (0.04)
62	4.19	4.00	4.08 (0.08)
63	4.04	3.77	3.65 (0.31)
67	4.76	4.36	4.60 (0.15)
69	3.2	3.08	3.16 (0.07)
76	2.45	2.37	2.52 (0.15)
80	3.81	3.83	3.91 (0.20)
89	2.38	2.10	2.30 (0.08)
95	2.98	3.17	3.18 (0.05)
96	4.41	4.26	4.56 (0.49)
97	4.27	4.24	4.55 (0.15)
102	4.1	4.22	4.32 (0.10)
105	3.32	3.83	3.65 (0.29)
106	3.42	3.52	3.61 (0.12)
108	4.61	4.36	4.58 (0.31)
118	3.11	3.41	3.30 (0.06)
128	2.61	2.36	2.54 (0.20)
132	2.55	2.32	2.50 (0.17)
134	2.71	3.20	2.74 (0.68)

GK: Gaussian kernel similarity estimation; SE: standard error of prediction

4.2.5. Comparison with a previously published model

Lin et al. reported an MLR model in 2003 on the mentioned dataset (Lin et al., 2003). The model was made from the experimentally derived octanol-water partition coefficient based hydrophobicity descriptors. Experimental estimation of octanol-water partition coefficient needs more resources, time, and manpower. To predict the toxicity of untested mixtures, a laboratory setup is also needed. Though the model quality was good in terms of the R^2 (0.948) and standard error (0.166), the model was not externally validated. Therefore, the predictivity of the mentioned model for new untested mixtures has not been sufficiently established. In the present work, our objective was to use solely computational descriptors which can be derived in no time given the structure of query compounds. The new PLS model has been developed and validated following the OECD protocol. Various internal and external validation metrics have been computed and reported in this communication (Section "**4.2.1. *The developed model and its validation***"). Acceptable R^2 (0.82) and Q^2_{LOO} (0.8) values prove the robustness of the model, whereas the acceptable Q^2_{F1} (0.87) value suggests its predictive ability. The predictive ability of the binary mixtures has been further improved (Q^2_{F1} = 0.94, MAE = 0.15) by the use of chemical read-across with the modeled descriptors. To avoid the complexity associated with the experimentally derived descriptors, only 2D structural descriptors have been utilized in this study. So to check the toxicity of untested mixtures, the prior formulation is not needed.

4.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against *Aliivibrio fischeri*

4.3.1. The developed model and its validation

In the present communication, we have developed a PLS regression model with the 2D structural features selected by the BSS algorithm from the reduced pool of descriptors after using GA and stepwise selection, as discussed in section "***3.3.4. Feature selection and model development***". The statistical soundness of the model is proven by the computed validation metric values (OECD, 2004). The final PLS model was developed with seven selected descriptors using five LVs, as depicted in the following equation 4.6 along with the computed internal and external validation metrics:

$$pEC50 = -1.06168 + 3.21035 \times B03[N-0] + 0.42675 \times LOGP99 - 0.6651 \times B04[N-N] - 8.67482 \times Eta_D_epsiC + 0.49428 \times B07[N-Cl] - 1.44239 \times B08[C-Cl] - 1.65409 \times H-054 \quad (4.6)$$

$n_{Train} = 149$, $LV_s = 5$, $R^2 = 0.942$, $R^2_{adj} = 0.939$, $Q^2_{LOO} = 0.932$, $\overline{r^2_{m(LOO)}} = 0.904$, $\Delta r^2_{m(Train)} = 0.042$, $S_{Train} = 0.254$, MAE_{Train} = 0.172, $RMSE_C = 0.247$, Training set prediction quality (based on *MAE-based criteria)* = Good; $n_{Test} = 49, Q^2_{F1} = 0.95, Q^2_{F2} = 0.95, \overline{r^2_{m(Test)}} = 0.932, \Delta r^2_{m(Test)} = 0.014, CCC = 0.975$, $S_{Test} = 0.269$, $MAE_{Test} = 0.183, RMSE_P = 0.246$,Test set prediction quality *(based on MAE-based criteria)* = Good

The model showed a high coefficient of determination (R^2 = 0.942) with statistically significant R^2_{adj} (0.939) values which suggest the goodness-of-fit of the model. The high value of the cross-validated correlation coefficient (Q^2_{LOO} = 0.932), and a low standard error of estimate (S_{Train} = 0.254) signifies the robustness of the developed model. The reliability of predictions of the model was adjudged by the computed external correlation coefficient

values (Q^2_{F1} = 0.95 and Q^2_{F2} = 0.95), which show good predictivity. The training and test set prediction qualities of the developed model based on the MAE-based criteria were proved to be "Good". Acceptable values of other validation parameters such as r_m^2 metrics (Ojha et al., 2011) ($\overline{r^2_{m(LOO)}} = 0.904, \Delta r^2_{m(Train)} = 0.042, \overline{r^2_{m(Test)}} = 0.932, \Delta r^2_{m(Test)} = 0.014$), error measures ($RMSE_C$ = 0.247, $RMSE_P$ = 0.247, MAE_{Train} = 0.172, MAE_{Test} = 0.183), and concordance correlation coefficient (CCC = 0.975) etc. suggests the statistical quality and reliability of prediction of the studied model.
B03[N-O], LOGP99, B04[N-N], Eta_D_epsiC, B07[N-Cl], B08[C-Cl], and H-054 are the descriptors of interest in this research. The definition, class, and the type of contribution of these modeled descriptors (See the regression coefficient plot in **Figure 4.25A**) have been tabulated in **Table 4.10**.

Table 4.10 Modeled descriptors: definition, class, and type of contribution to toxicity

Descriptor	Definition	Class	Contribution to toxicity
B03[N-0]	Presence/absence of N-O atom pair at topological distance 3	2D-atom pairs	Positive
LOGP99	Wildmann-Crippen octanol-water partition coefficient (logP)	Molecular properties	Positive
B04[N-N]	Presence/absence of N-N atom pair at topological distance 4	2D-atom pairs	Negative
Eta-D-epsiC	A measure of the contribution of electronegativity	ETA indices	Negative
B07[N-Cl]	Presence/absence of N-Cl atom pair at topological distance 4	2D-atom pairs	Positive
B08[C-Cl]	Presence/absence of C-Cl atom pair at topological distance 8	2D-atom pairs	Negative
H-054	H-attached sp^3 hybridised carbon with 3 electronegative atoms at the next carbon/carbons	Atom-centred fragments	Negative

The additional loading and score plots (Ergon, 2004) of the PLS model were drawn using SIMCA-P software (Wu et al., 2010) and depicted in **Figure 4.25B** and **4.25C** respectively. The score vectors of 1st two components t1 and t2 are plotted in the Score plot. The ellipse signifies the applicability domain of the model with 1st two latent variables; thus four mixtures (**161**, **166**, **167**, and **173**) are found to be the outliers with respect to LV1 and LV2. The loading plot provides a relationship between descriptors (X-variables) and toxicity (Y-variable) in the 1st two latent variable spaces (w1 and w2). Here in the loading plot, pEC50 along with seven descriptors have been drawn. If we look closely at the loading plot, then it is

observed that the Eta_D_epsiC and B07[N-Cl] are placed at the farthest and nearest points from the origin, hence they are signified as the most important and least important descriptors respectively. We performed the Y-randomization test(Rucker et al., 2007) with the developed model using SIMCA-P software (Wu et al., 2010) to confirm whether the model is the result of any chance correlation or not. Y-randomization is performed by generating n number of models with randomly permuted response (Y) variables. The descriptors and the modelling procedure are kept constant during the random model generation and then fitted with the original one. The number of permutations can vary, and we have used 100 such permutations in this study. The coefficient of determination (R^2) and the correlation coefficient (Q^2) values of the random models are plotted (Y-axis) against the 100 resultant correlation coefficients between the original Y values and the permuted Y values (X-axis). The best-fitting lines have been drawn with both R^2 and Q^2 and their respective intercepts (R_Y^2 and Q_Y^2) have been noted. In the present study, the Y-randomization plot has been obtained by using SIMCA-P software and showcased in **Figure 4.25D**. The small values of R^2 and Q^2 intercepts ((R_Y^2 = -0.01 and Q_Y^2 = -0.33)) justify the robustness of the reported model, and also cancel out the possibilities of chance correlation.

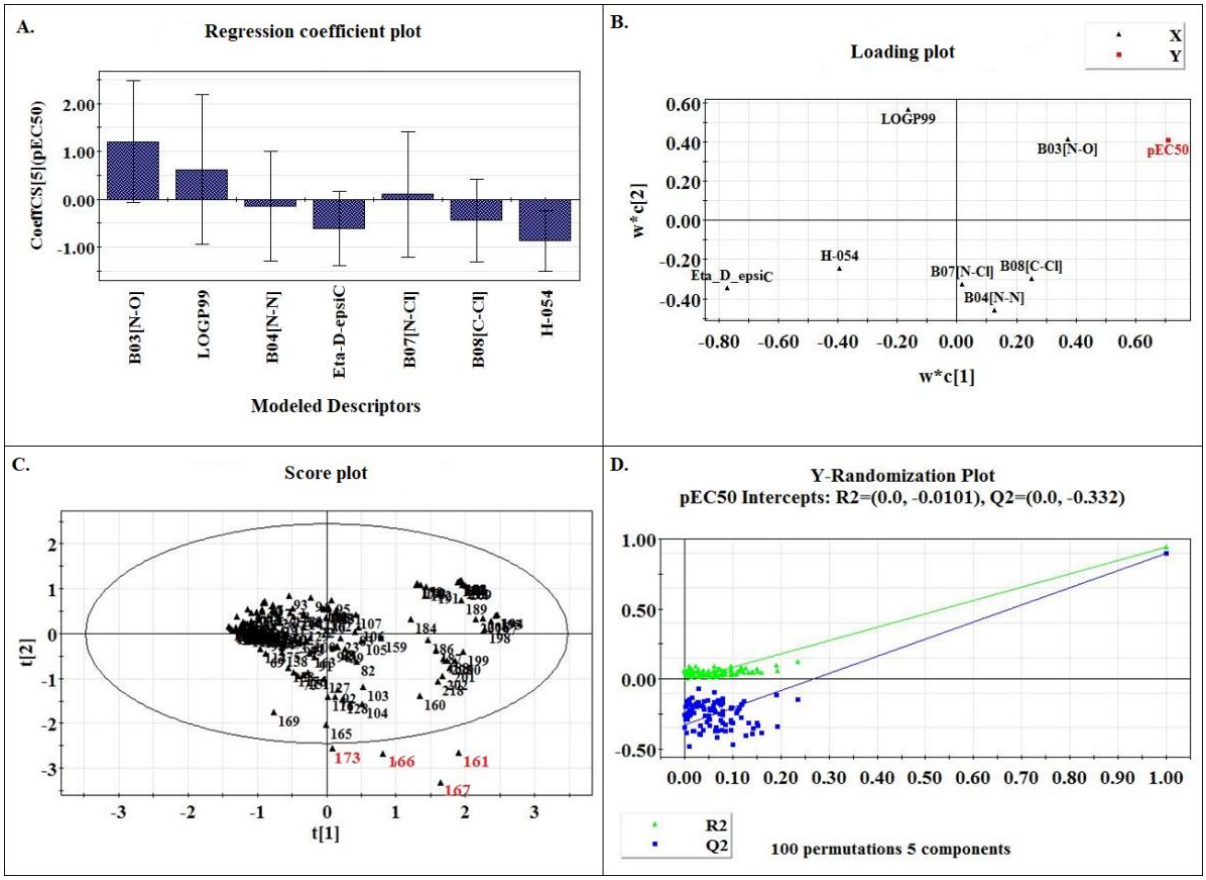


Figure 4.25. Various PLS plots drawn using SIMCA-P software (A. Coefficient plot; B. Loading plot; C. Score plot; D. Y-randomization plot)

We have also analysed the applicability domain (Gadaleta et al., 2016) of the developed PLS model by DModX (distance to model X) approach (De et al., 2018) using SIMCA-P software (Wu et al., 2010) with a 99% confidence interval. Three training mixtures (a 0.25:0.75 mixture of omeprazole and imatinib, a 0.75:0.25 mixture of imatinib and tobramycin, and a 0.149:0.851 mixture of tetracycline hydrochloride and linuron) were found outside the

standard 99% D-critical line (D-Crit = 0.00999898) of **Figure 4.26**, whereas all the test mixtures were inside the threshold.

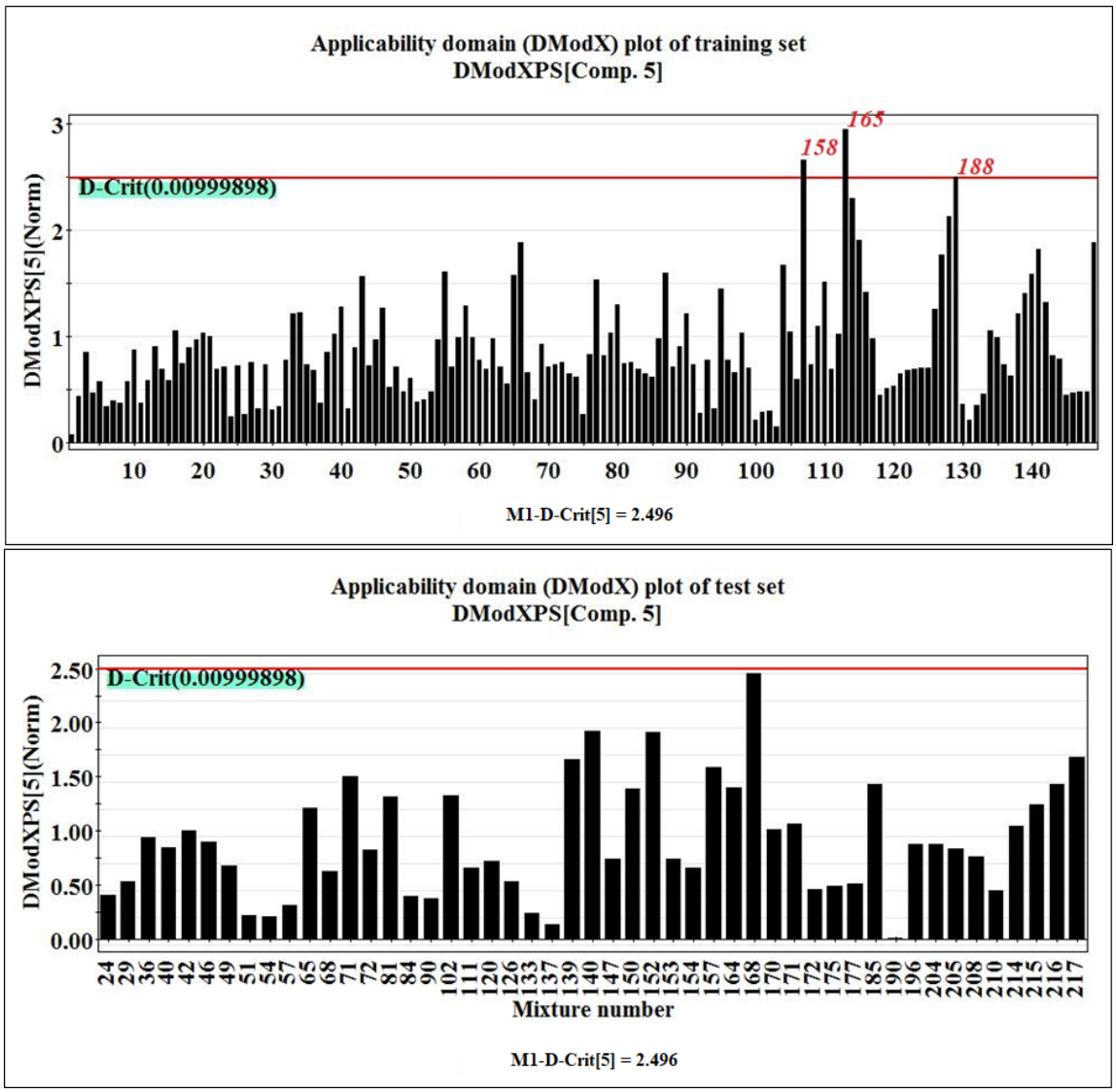


Figure 4.26. Applicability domain plot of training and test sets

4.3.2. Results of cross-validation

The developed model is proved to be robust and extremely predictive as per the OECD guidelines for QSAR modeling of individual data points. However, we have also performed the mixtures out, compounds out, and everything out cross-validations to penalize the overestimation of validation metrics if any. To perform the cross-validation, we shuffled the data points of the original training-test pairs (4-4 Euclidean distance-based division) as discussed in section "***3.3.4. Feature selection and Model development***". The internal and external validation parameters of the cross-validated models were computed with the shuffled training-test pairs and given in **Table 4.11**.

Table 4.11. Computed metrics for mixtures out, compounds out, and everything out cross-validations

Internal validation metrics	**Original PLS model**	**Mixture out cross-validation**	**Compounds out cross-validation**	**Everything out cross-validation**
R^2	0.942	0.958	0.912	0.807
Q^2_{LOO}	0.932	0.950	0.883	0.777
MAE-LOO (Train)	0.205	0.188	0.185	0.172
External validation metrics				
Q^2_{F1}	0.950	0.913	0.968	0.859
MAE_{Test}	0.183	0.208	0.234	**0.653**

All three cross-validated models have shown acceptable validation metric values (except MAE_{Test} in everything out CV) which signify the statistical reliability of the developed model and thus counter the chances of misjudgment of model quality. An expected deterioration of the model quality has been observed in everything out cross-validation (more profound in the case of the MAE_{Test} value (0.653)) which justifies the complete structural exclusiveness among training and test set mixtures. The observed vs. predicted scatter plots of cross-validated models along with the original PLS model have been given in **Figure 4.27**. A uniform scattering has been observed around the lines passing through the origins in all the cases which suggest the goodness-of-fit and robustness of the developed model.

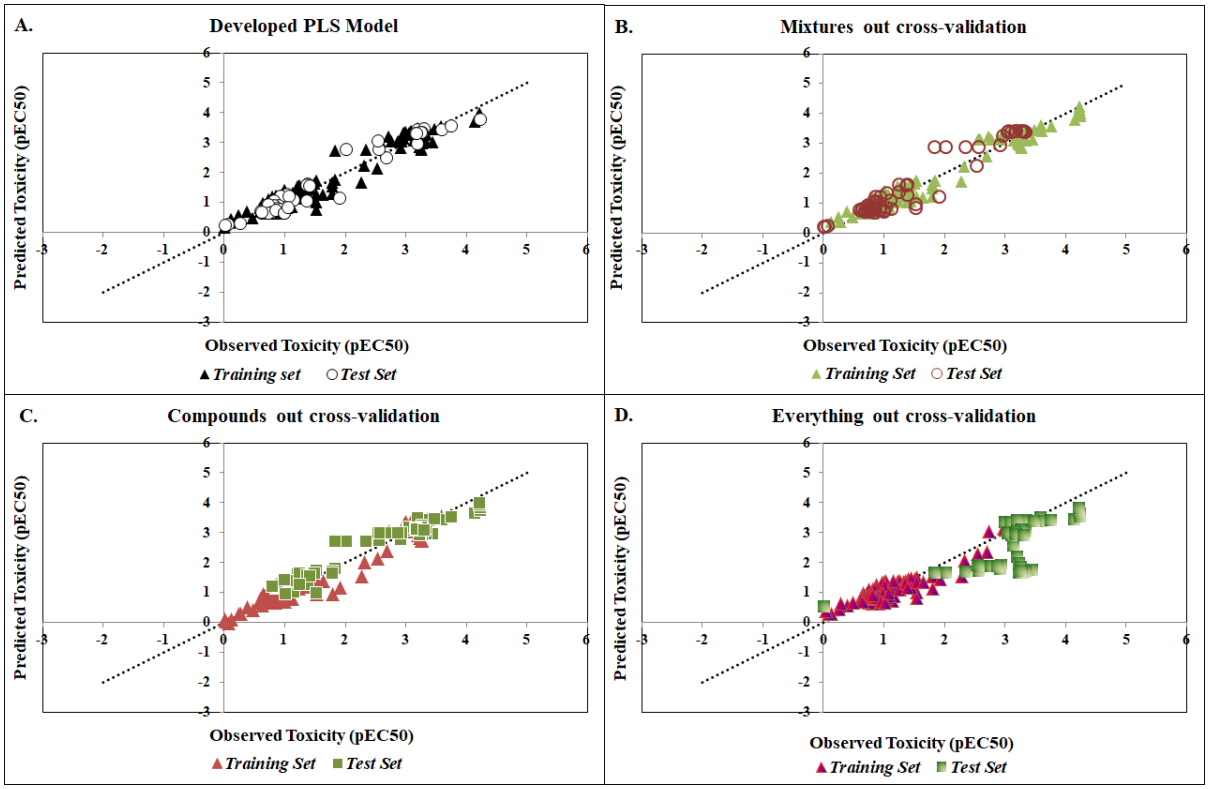


Figure 4.27. Observed vs. predicted scatter plots (A. Original PLS model, B. Mixtures out cross-validation, C. Compounds out cross-validation, D. Everything out cross-validation)

4.3.3. Study of modeled descriptors

In this present research, the toxicity of binary and multi-component mixtures has been modeled with seven 2D descriptors as seen in section "***4.3.1. The developed model and its validation***". These modeled descriptors essentially represent seven unique structural features which are responsible for the toxic action. These descriptors are depicted in **Figure 4.28**. as the variable importance plot (Wold et al., 2001) with their relative importance towards the toxic endpoint (pEC50). A descriptor with VIP score >1 suggests a significant impact on toxicity. In this study, Eta_D_epsiC and B03[N-O] have VIP scores of more than 1, among which *Eta_D_epsiC* is the most important. The importance of the rest descriptors such as H-054, B08[C-Cl], LOGP9, B04[N-N], and B07[N-Cl] is gradually declining. The relationship of the descriptors with toxicity is also depicted in **Figure 4.28**. B03[N-O], LOGP99, B07[N-Cl] show a positive contribution to the modeled response whereas, Eta_D_epsiC, H-054, and B08[C-Cl], and B04[N-N] have negative contributions towards the toxicity. We have depicted all the defined structural fragments as 2D diagrams in the different studied mixture components in the following **Figure 4.28**.

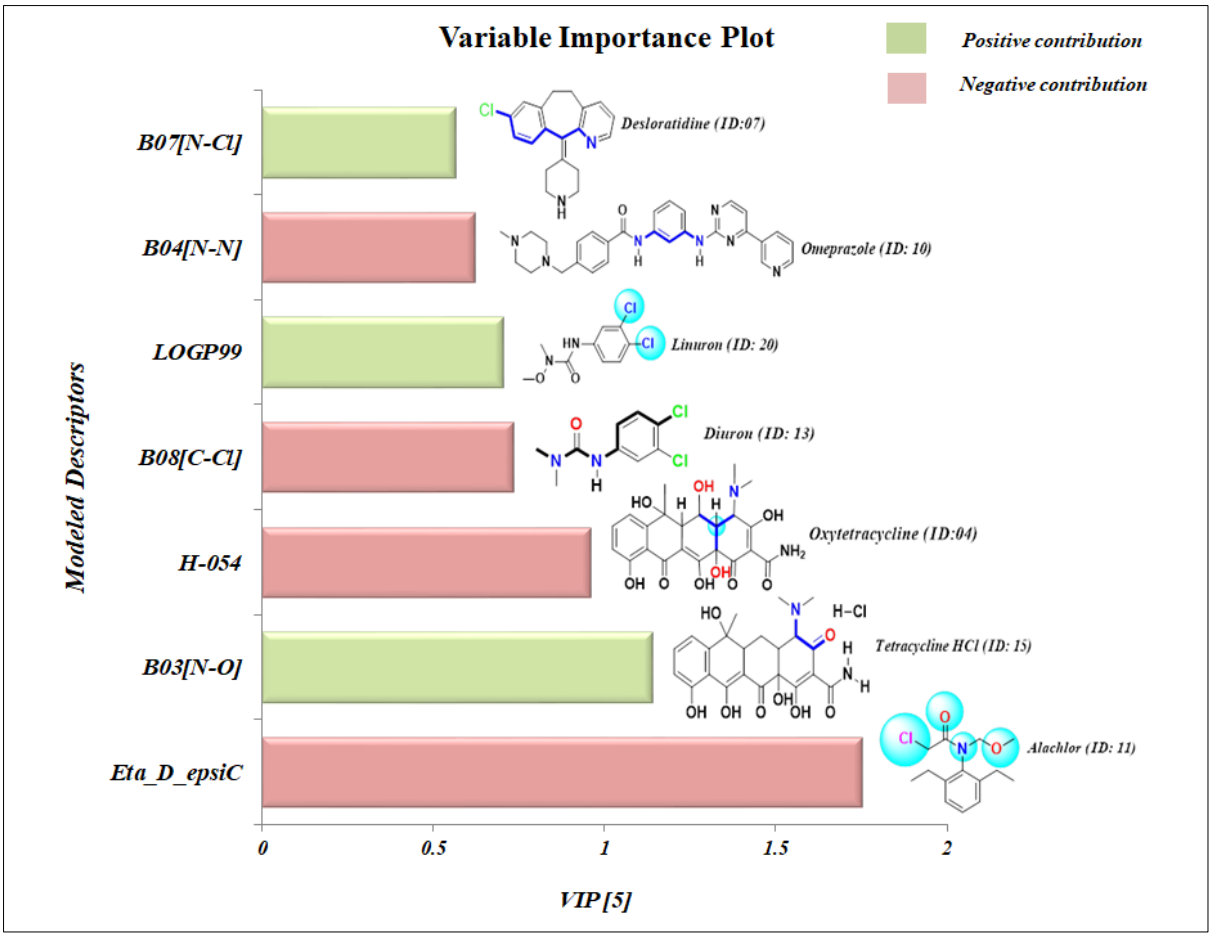


Figure 4.28. Variable importance plot with the 2D representation of the modeled descriptors

Eta_D_epsiC ($\Delta\varepsilon_C$) is an extended topochemical atom indices descriptor which measures the effect of electronegativity (Roy and Das, 2015). The descriptor can be mathematically expressed as $\boldsymbol{\Delta\varepsilon_C = \varepsilon_3 - \varepsilon_4}$, where $\boldsymbol{\varepsilon_3}$ and $\boldsymbol{\varepsilon_4}$ are defind in the equation 4.7 and 4.8 respectively.

$$\varepsilon_3 = \frac{[\sum \varepsilon]_R}{N_R} \qquad (4.7)$$

$$\varepsilon_4 = \frac{[\sum \varepsilon]_{SS}}{N_{SS}} \qquad (4.8)$$

According to Roy et al., ε_3 is "*the sum of epsilon (ε) relative to the total number of atoms present in the molecule(N_R) including hydrogen in the connected molecular graph of reference alkane*", whereas, ε_4 is "*the sum of epsilon (ε) relative to the total number of atoms present in the molecule (N_{SS}) including hydrogen for a saturated carbon skeleton of the molecule*" (Roy and Ghosh, 2004). The reference alkane can be defined as the molecular graph where all the heteroatoms and covalent multiple bonds are replaced with carbons and single bonds respectively; on the contrary, the saturated carbon skeleton is the saturated molecular graph of the corresponding molecule without changing the heteroatoms with carbon (De et al., 2018). From this definition, it is understood that the ε_3 reduces with an increasing number of electronegative atoms in the molecule (due to the replacement of heteroatoms (electronegative) with carbon), however ε_4 remains unaltered. Thus, the resultant $\Delta\varepsilon_C$ becomes more negative with an increasing number of electronegative atoms. It

has been observed in the studied mixtures also. We have compared the change of toxicity from mixture **67** to mixture **80** in **Figure 4.29**. Mixture 67 (Eta_D_epsiC = -0.075) is an equimolar combination of isoproturon (#**12**) and alachlor (#**11**) whereas mixture 80 (Eta_D_epsiC = -0.153) is a 0.25:0.75 combination of alachlor (#**11**) and chlorfenvinphos (#**14**). Chlorfenvinphos has more electronegative atoms in its structure than isoproturon and this structural difference makes the value of Eta_D_epsiC more negative in mixture **80** as compared to mixture **67** (as alachlor is a constant component between the two mixtures). An increament of the toxicity has also been observed from mixture **67** (pEC50 = 0.88) to mixture **80** (pEC50 = 1.03), which justifies the negative contribution of the descriptor Eta_D_epsiC to the toxicity.

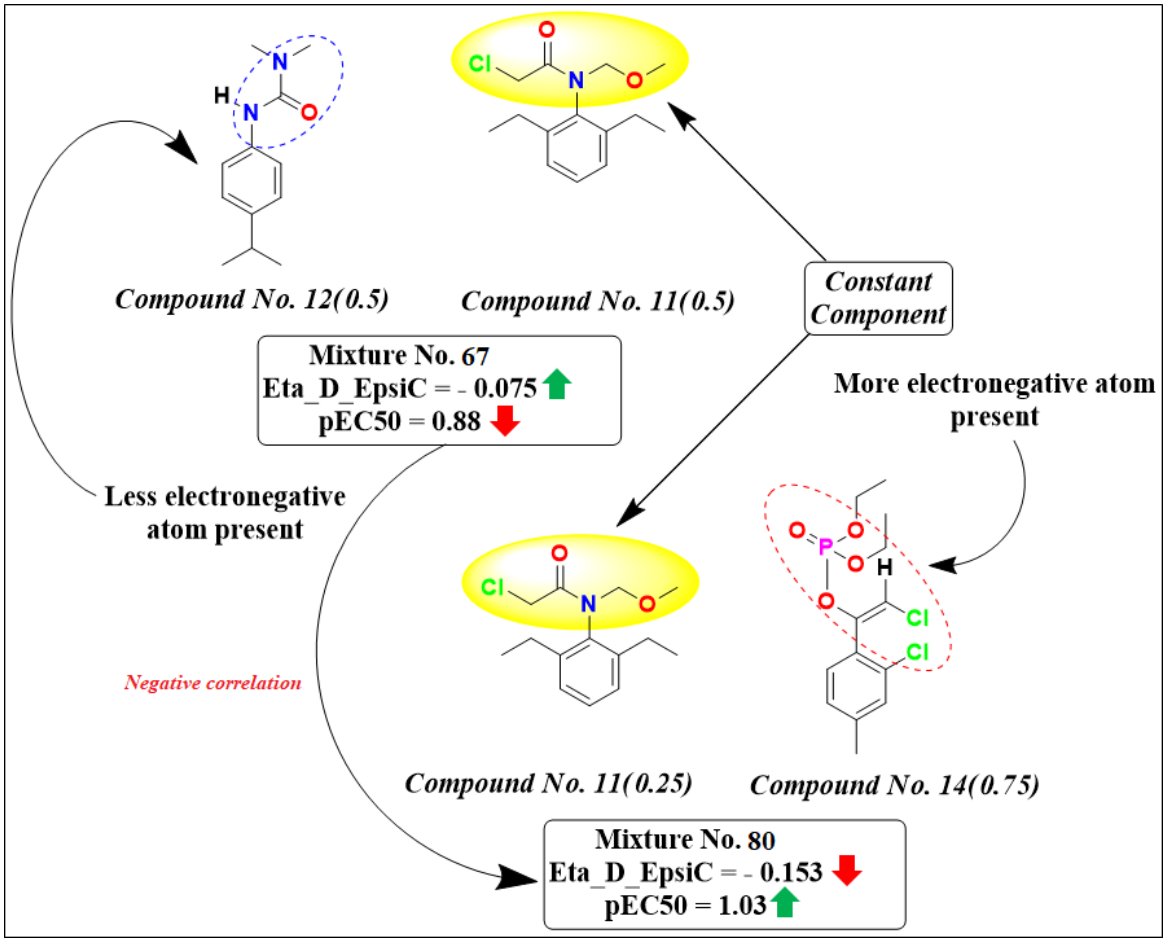


Figure 4.29. Effect of Eta_D_epsiC on the toxicity of pharmaceutical/pesticide mixtures

The second most important descriptor is the B03[N-O] which belongs to the 2D atom pairs class of descriptors. It can have binary values 0 and 1 which essentially signifies the absence and presence of the [nitrogen-oxygen] atom pair at topological distance 3 respectively. The presence of this feature in a mixture increases the toxicity as shown in **Figure 4.30**. This observation justifies the positive correlation of B03[N-O] with the toxic endpoint. A mixture (#**169**) of imatinib (#**8**) and desloratadine (#**7**) does not have this feature and hence shows less toxicity (pEC50 = 0.49) than another mixture **163** (pEC50 = 2.55) with the B03[N-O] feature in the component's structures desloratadine (#**7**) and omeprazole (#**10**). Omeprazole contains this particular feature and hence makes the key difference here (desloratadine is the constant component between two mixtures).

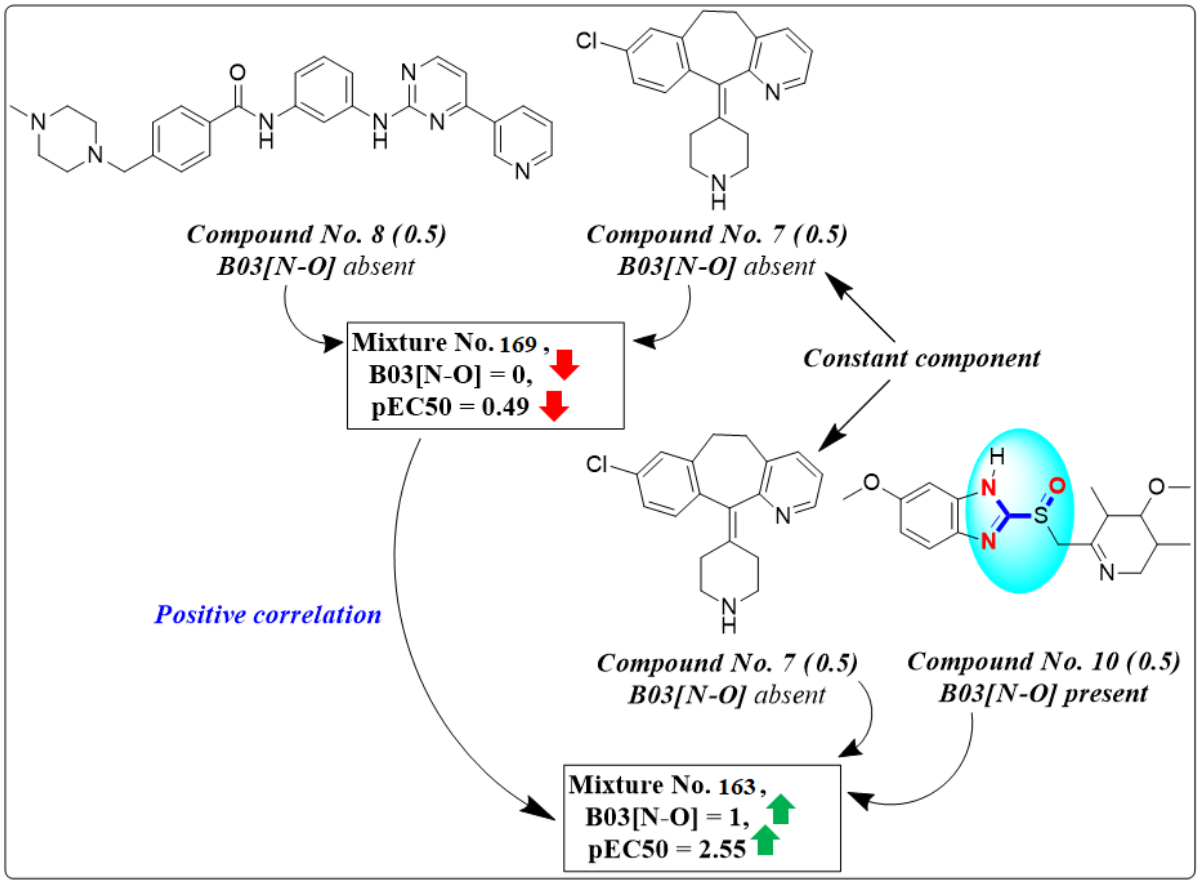


Figure 4.30. Effect of B03[N-O] on the toxicity of pharmaceutical/pesticide mixtures

H-054 has grabbed the third place in the variable importance plot (**Figure 4.28**) of the current model. It is an atom-centred fragment descriptor which is defined by a hydrogen-attached sp^3 hybridized carbon-containing three electronegative atoms in the adjacent carbon/carbons. This descriptor has shown a negative contribution to the toxicity of binary and multi-component mixtures. Oxytetracycline (#**4**) is a mixture component containing the specified fragment that is depicted in **Figure 4.31**. It has been observed that the mixture (#**105**) containing oxytetracycline has greater H-054 value and smaller toxicity value (pEC50 = 1.65) in comparison with the mixture (#**186:** H-054 = 0, pEC50 = 4.25) devoid of oxytetracycline. Therefore, the inverse relationship between H-054 and the mixtures' toxicity can be explained by these observations.

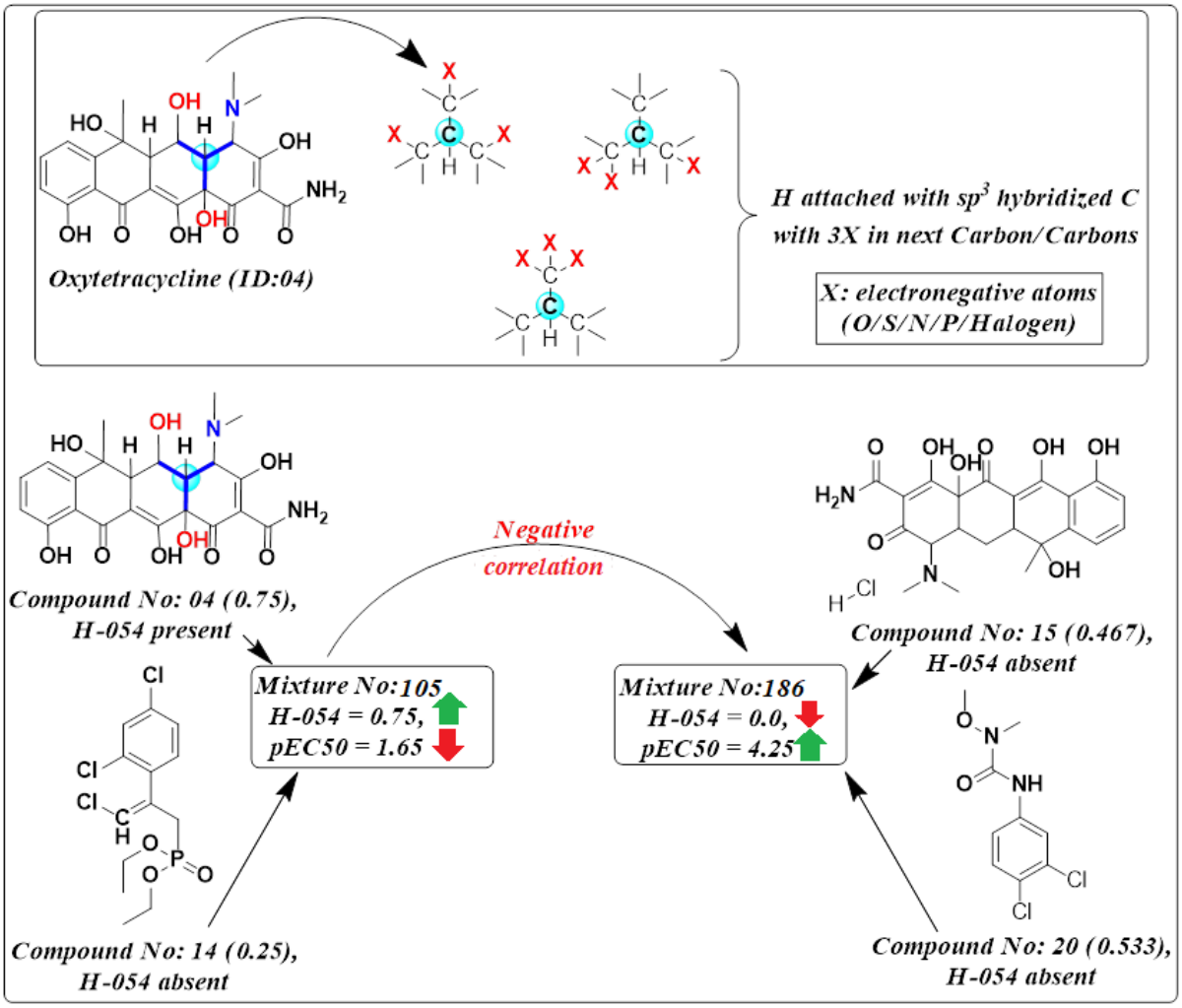


Figure 4.31. Effect of H-054 on the toxicity of pharmaceutical/pesticide mixtures

B08[C-Cl] is another descriptor of the model, which signifies the presence or absence of carbon-chlorine atom pair at topological distance 8. It belongs to the 2D atom pairs class of descriptors. In this study, B08[C-Cl] was negatively correlated to the modeled response. It has been observed that a less toxic mixture **173** (pEC50 = 0.02) is having higher B08[C-Cl] value (0.75). The opposite situation has been found in a more toxic mixture **180** (pEC50 = 3.35) where B08[C-Cl] is zero (**Fig4.32**). Thus, the negative contribution of this descriptor is well explained by these observations. From a close look at the components of mixtures **173** and **180**, it has been found that desloratadine (#**7**) is responsible for the high B08[C-Cl] value of mixture **173** (**Figure 4.32**). It may be concluded from these observations that the presence of chlorine atom with a large and fused ring system is beneficial from the regulatory point of view because this feature will help to reduce toxicity.

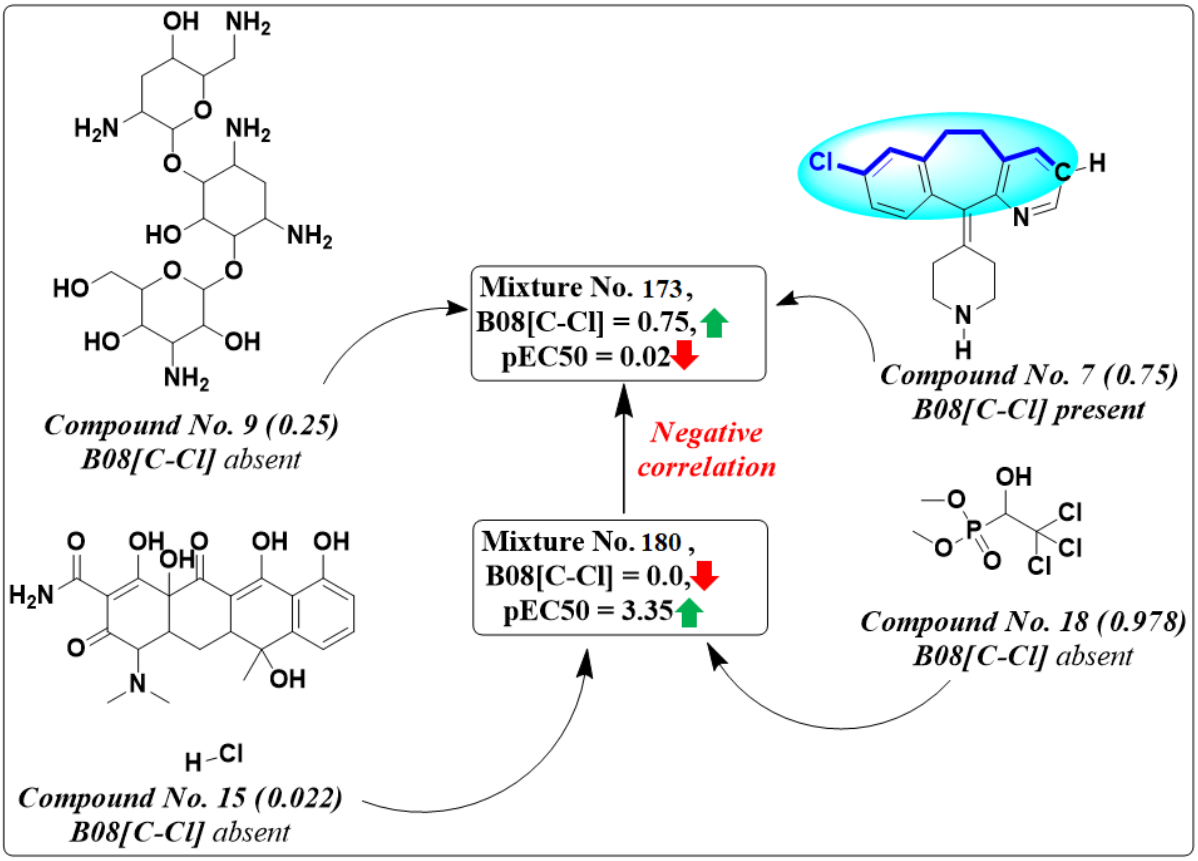


Figure 4.32. Effect of B08[C-Cl] on the toxicity of pharmaceutical/pesticide mixtures

LOGP99 is a molecular property descriptor which represents the Wildmann-Crippen octanol-water partition coefficient (logP). It is a measure of lipophilicity and contributes positively to the mixtures' toxicity in this research. It has been observed from the studied mixtures that the chlorine or other halogen-containing mixtures are more lipophilic. The halogen atoms increase the molecular size and London dispersion forces to some extent which indirectly increases the lipophilicity(Pliska et al., 2008). More lipophilic compounds can cross the lipopolysaccharide-containing cell wall of *Aliivibrio fischeri*, and hence increase the toxicity. Mixture **186** contains tetracycline hydrochloride (#15) and a chlorine-containing component linuron (#20), which have shown more LOGP99 (1.63) and pEC50 (4.25) values (**Figure 4.33**). On the other hand, a combination (mixture **167**) of imatinib (#**8**) and tobramycin (#**9**) has shown less LOGP99 (-3.57) and pEC50 (0.13) values, which justify the proportional relationship between them.

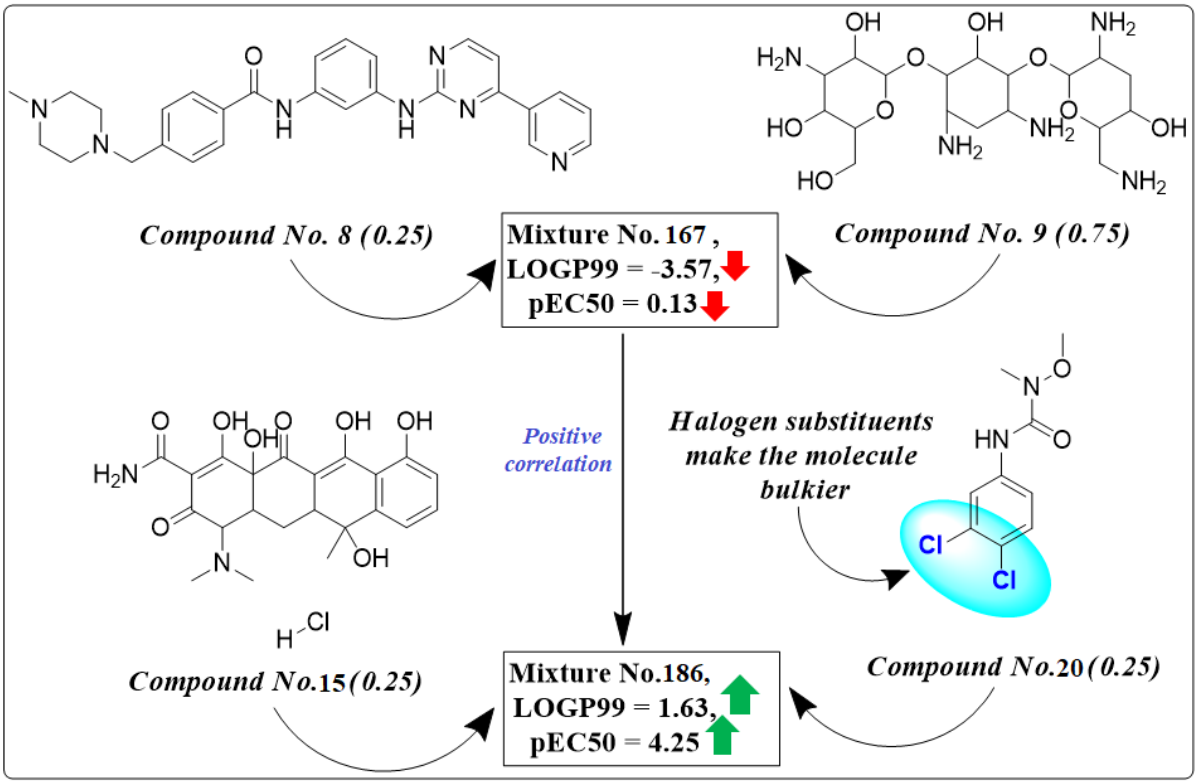


Figure 4.33. Effect of LOGP99 on the toxicity of pharmaceutical/pesticide mixtures

B04[N-N] is a 2D atom pair descriptor which is found to be inversely correlated with the toxic endpoint of this study. It is defined by the presence of nitrogen atom pairs in topological distance 4. This feature has been observed in mixture components such as imatinib (#**8**) and tobramycin (#**9**) as seen in **Figure 4.34**. In both cases, two amino (-NH2) or substituted amino (-NH-) groups are found at alternative carbons concerning a benzene or cyclohexane ring. In the above-mentioned figure, it has been observed that the presence of this feature reduces toxicity to a large extent. This negative relationship is also evident in a less toxic mixture **166** with nitrogen atom pair at topological distance 4 (B04[N-N] = 1, pEC50 = 0.26) and a more toxic mixture **186** without this feature (B04[N-N] = 0, pEC50 = 4.25).

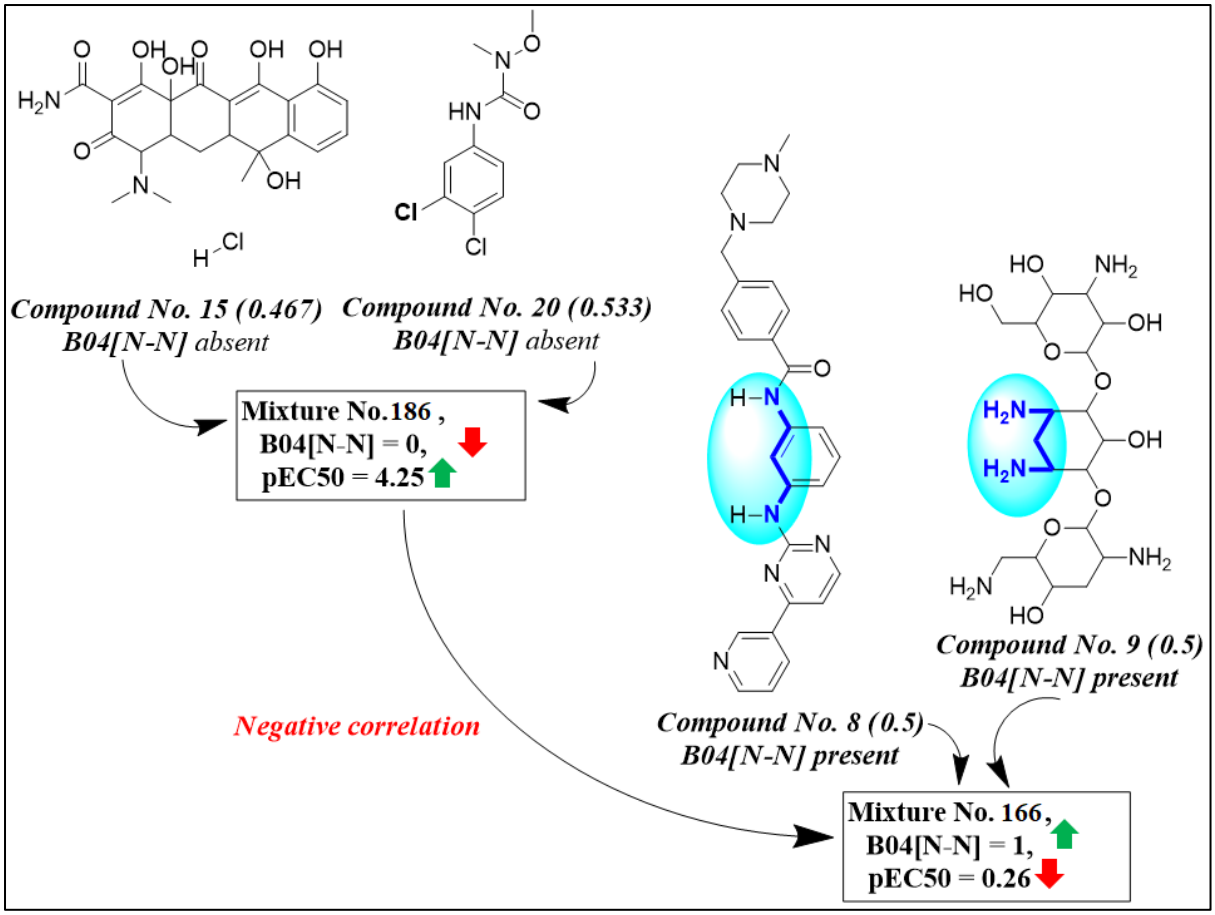


Figure 4.34. Effect of B04[N-N] on the toxicity of pharmaceutical/pesticide mixtures

The least important descriptor in this research to explain the toxicity is B07[N-C]. It is again a 2D atom pair descriptor which signifies the nitrogen and chlorine atom pair at the topological distance 7. This feature has been shown in the structure of desloratadine (#**7**) in **Figure 4.28** and **Figure 4.35**. The coefficient plot (**Figure 4.25A**) of this model suggests a positive correlation of B07[N-N] with the toxicity. To explain this, we have taken up two mixtures **122** and **162** (**Figure 4.35**). Mixture **162** is a 0.75: 0.25 mass combination of omeprazole (#**10**) and desloratadine (#**7**) whereas, mixture **122** is a 0.25: 0.75 mass combination of dexamethasone (#**5**) and isoproturon (#**12**). Among these four components, only desloratadine possesses N-Cl atom pair at a topological distance of 7 and hence, mixture **162** shows a positive descriptor value (B07[N-N] = 0.25). As per the expectation, mixture **162** has imparted a greater toxic response (pEC50 = 2.98) than mixture **122** (pEC50 = 0.63) which does not have this feature. Therefore, the proportional relationship between pEC50 and B07[N-N] is justified.

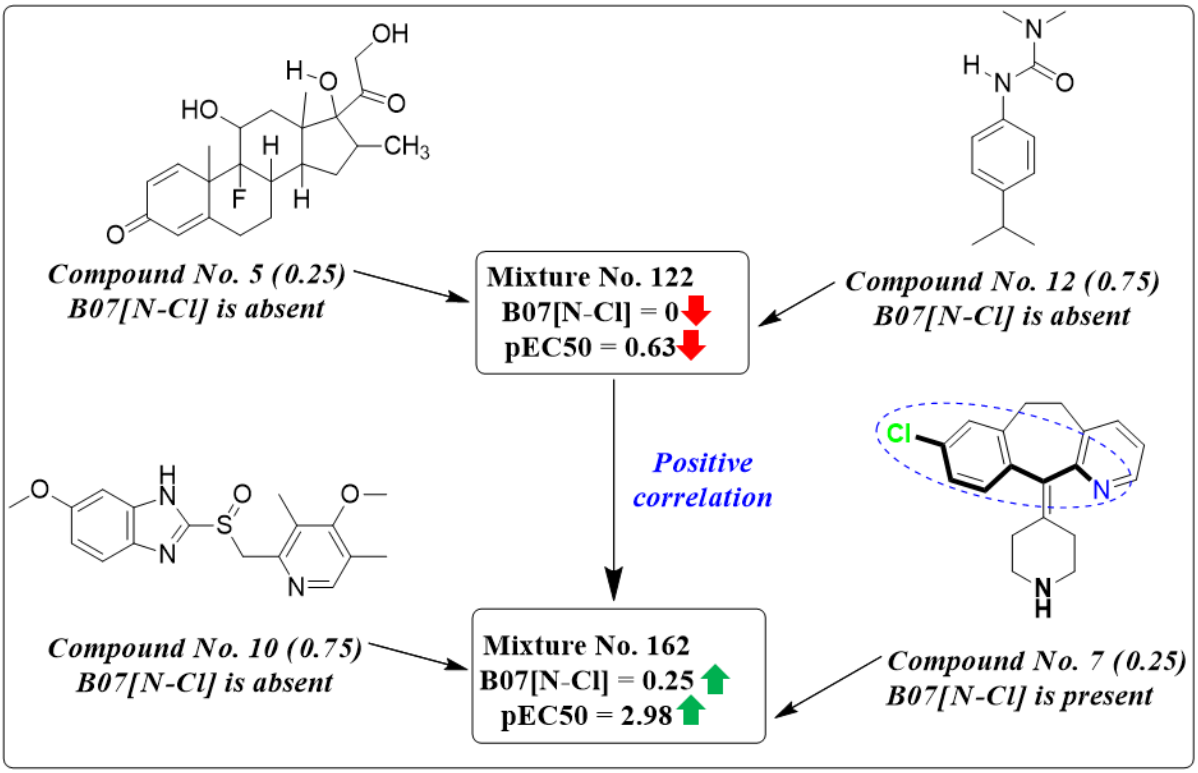


Figure 4.35. Effect of *B07[N-Cl]* on the toxicity of pharmaceutical/pesticide mixtures

4.3.4. Result of chemical read-across

The machine learning-based chemical read-across has been performed in this study as an additional tool to cross-check the relevance of identified descriptors. To do so, the hyperparameters σ and γ were optimized first as discussed in the section "***3.3.5. Chemical read-across***". The optimization of the parameters is given in **Table 4.12**. From the previously mentioned table, σ = 2 and γ = 0.5 have been identified as the optimized values as in these two conditions, the best Q^2_{F1} value (0.644 for both GK and LK) is obtained without changing the other determinants. In the next step, these optimized settings are used to get the best prediction of the test set from the training set. The best prediction has been obtained in the setting σ = 2, γ = 0.5, number of close training compounds = 2, distance threshold = 1, and similarity threshold = 0. The computed external validation metrics are given in the following **Table 4.13** as the quality measure of prediction. It has been observed that the external predictive variance values obtained by the three methods are high enough to justify the relevance of the selected descriptors. The error measures RMSEp and MAE are also on the lower side which shows the reliability of this justification. Among the three similarity estimation techniques of read-across, the best prediction has been obtained by the Laplacian kernel function similarity (**Q^2_{F1} = 0.919**).

Table 4.12. Optimization of hyperparameters (sigma and gamma) to obtain the best prediction

Sigma value (σ)	Gamma value (γ)	Number of C.T.C	Distance threshold	Similarity threshold	Q^2_{F1} (GK)	Q^2_{F1} (LK)
0.25	0.25	10	1	0	0.495	0.632
0.5	**0.5***	10	1	0	0.557	**0.644**
0.75	0.75	10	1	0	0.578	0.628
1	1	10	1	0	0.591	0.608
1.25	1.25	10	1	0	0.611	0.589
1.5	1.5	10	1	0	0.630	0.572
1.75	1.75	10	1	0	0.641	0.559
2*	2	10	1	0	**0.644**	0.548

*Optimized setting; C.T.C: close training compounds; GK: Gaussian kernel function; LK: Laplacian kernel function.

Table 4.13. Computed external validation metrics of read-across prediction

External validation metrics	Euclidean distance-based similarity	Gaussian kernel function similarity	Laplacian kernel function similarity
Q^2_{F1}	0.914	0.913	**0.919**
Q^2_{F2}	0.914	0.913	**0.919**
$RMSE_P$	0.324	0.326	**0.315**
MAE	0.194	0.194	**0.189**

4.3.5. A comparative assessment of the model with previous reports

To justify the merit of this study, we have compared the proposed QSAR model with two previously published literature reports (Bures et al., 2021; Qin et al., 2018) for the aquatic toxicity of binary and multi-component mixtures of pharmaceuticals and pesticides against *Aliivibrio fischeri* in terms of the quality measures. The validation metrics (R^2, Q^2_{LOO}, and Q^2_{F1}) of these three models are summarized in the following **Table 4.14**. The table clearly depicts the better quality of the proposed QSAR model than the previous reports in terms of goodness-of-fit, robustness, and external predictivity. The chemical read-across also provides better prediction than the previous results but is inferior to the QSAR model. The current model covers a comparatively large spectrum of molecular diversity than the previous

models which is evident from the number of training data (n_{Train} = 153, 45, 123 in the proposed model, model of Qin et al., and model of Bures et al. respectively). The previous models were essentially composed of 3D features along with 2D descriptors; hence calculation and interpretation were not so easy. On the contrary, the newly developed model has been constructed with simple 2D structural descriptors, which makes the model easily interpretable and usable for quick predictions. Therefore, this comparative assessment strongly suggests better quality, interpretability, and transferability of the proposed model with fewer and easier necessary prerequisites for prediction.

Table 4.14. Comparative assessment of the proposed model with previous reports of toxicity (pEC50) of binary and multi-component mixtures against *Aliivibrio fischeri*

Variables	**Model reported by Qin et al. in 2018**	**Model reported by Bures et al. in 2021**	**The proposed QSAR model in this study**	**Similarity-based read-across in this study**
Data points	45	153	198	198
Type of model	GA-MLR	GA-MLR	PLS	WAP (Laplacian kernel)
No. of descriptors	3	4	**7 (LVs = 5)**	**7**
Internal validation				
n_{Train}	31	123	149	149
R^2	0.936	0.901	**0.941**	**-**
Q^2_{LOO}	0.908	0.863	**0.931**	**-**
External Validation				
n_{Test}	14	30	49	49
Q^2_{F1}	0.748	0.863	**0.950**	**0.919**

GA-MLR: genetic algorithm-multiple linear regression; PLS: partial least squares; LVs: latent variables; WAP = weighted average prediction

4.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

4.4.1. Developed q-RASAR models and best mixing rule identification

Three distinct q-RASAR models have been developed by partial least squares approach with the optimum number of latent variables (LVs) in this study from the mixture descriptors obtained by three different mixing rules. The optimum number of latent variables has been selected based on the X-variance during PLS regression and those variance values have been showcased in **Table 4.15**. The developed models have also been represented in **Table 4.15**. Each model has been rigorously validated following the OECD protocols. The computed internal and external validation metrics along with the identified descriptors and the optimum number of latent variables have been shown in the following **Table 4.16**.

The computed validation metrics (as shown in **Table 4.16**) confirm the fitness, robustness, and predictive performance of the developed models. Q^2_{F1} value is an exclusive indicator of predictability as discussed in the section "2.6.Validation protocols". All three models have good predictability, but the model obtained from "Mixture descriptor 3" is superior in terms of Q^2_{F1} value (0.91). Besides this, least AIC value has been observed in the model obtained from "Mixture descriptor 3" ($AIC_{Mix.\ Desc.\ 3}$ = -31.86) which also supports the supremacy of the model. This superiority of predictability indicates the efficiency of mixing rule 3 (norm of molecular contributions) for calculating the mixture descriptors. Therefore, the 3rd model has been identified as the most predictive one for mixture toxicity and has been reported as the ultimate model of this study.

Table S4.15. PLS-q-RASAR models obtained from MD1, MD2, and MD3 respectively and the optimum number of latent variables used based on the X-variance

Mixture Descriptor	**1**	**2**	**3**
PLS-q-RASAR model	$pLD_{50} = -1.93321 + 1.33044 \times B05[C-O] + 0.05463 \times D/Dtr03 - 0.10333 \times NaaCH + 2.38($ $+ 0.01771 \times D/Dtr05 + 0.40938 \times F05[C-F] + 0.354 \times Abs.MaxPos - MaxNeg$	$pLD_{50} = -6.5130 + 1.3455 \times B06[O-Cl] + 0.0019 \times D/Dtr03 + 12.7745 \times Eta_{betaS_A} + 0.0913 \times mindsN - 0.0053 \times nCsp2 + 0.1166 \times RDCHI + 1.4520 \times SD\ Similarity\ (GK)$	$pLD_{50} = -16.9846 - 1.252 \times B06[C-Cl] + 2.6563 \times B06[C-F] + 36.9044 \times Eta_betaS_A + 0.1754 \times H - 046 + 0.9453 \times NssssC + 0.3549 \times gm * Avg.Sim - 3.2626 \times gm * SDSimilarity$ (presented in equation 4.9)
X-variance calculated after adding each latent variable	LV 1 = 0.32	LV 1 = 0.23	LV 1 = 0.42
	LV 2 = 0.47	LV 2 = 0.45	LV 2 = 0.58
	LV 3 = 0.66	**LV 3 = 0.61 (optimum)**	LV 3 = 0.73
	LV 4 = 0.75	-	**LV 4 = 0.79 (optimum)**
	LV 5 = 0.84	-	-
	LV 6 = 0.93 (optimum)	-	-
	LV 7 = 1.00 (Cannot be taken because no. of LVs = no. of descriptors)	-	-

Table 4.16. Modeled descriptors, and computed validation metrics of three PLS q-RASAR models obtained from three different classes of mixture descriptors

Validation metrics	Mixture descriptor 1	Mixture descriptor 2	Mixture descriptor 3
Identified descriptors	**Seven descriptors:** B05[C-O], D/Dtr03, NaaCH, nRNR2, D/Dtr05, F05[C-F], Abs MaxPos-MaxNeg	**Seven descriptors:** B06[O-Cl], D/Dtr03, Eta_betaS_A, mindsN, nCsp2, RDCHI, SD similarity(GK)	**Seven descriptors:** B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC, gm*Avg.Sim, gm*SD Similarity
No. LVs	6	3	4
R^2	0.90	0.88	0.89
Q^2_{LOO}	0.85	0.84	0.84
MAE_{Train}	0.31	0.33	0.31
$\overline{r_m^2(Train)}$	0.80	0.78	0.78
$\Delta r_m^2(Train)$	0.07	0.09	0.08
AIC	-29.72	-31.72	**-31.86**
Q^2_{F1}	0.89	0.88	**0.91**
Q^2_{F2}	0.88	0.88	**0.91**
Q^2_{F3}	0.87	0.87	**0.90**
MAE_{Test}	0.32	0.34	0.32
CCC	0.93	0.93	**0.95**
$\overline{r_m^2(Test)}$	0.78	0.85	0.87
$\Delta r_m^2(Test)$	0.08	0.06	0.06
MAE-based prediction quality	Moderate	Good	**Good**

LVs: Latent variables; MAE: mean absolute error; AIC: Akaike information criteria; CCC: concordance correlation coefficient.

4.4.2. Best model and result of the validation

The best model is developed with the mixture descriptors obtained by mixing rule 3, as discussed in the previous section "***4.4.1.Developed q-RASAR models and best mixing rule identification***". This PLS q-RASAR model is composed of seven descriptors (two RASAR descriptors and five 2D structural descriptors) in the 4 latent variable space. The mathematical form of the final model has been shown in equation 4.9-

$$pLD_{50} = -16.9846 - 1.252 \times B06[C - Cl] + 2.6563 \times B06[C - F] + 36.9044 \times Eta_betaS_A + 0.1754 \times H - 046 + 0.9453 \times NssssC + 0.3549 \times gm * Avg.Sim - 3.2626 \times gm * SDSimilarity \quad (4.9)$$

Rigorous internal and external validation tests have been performed, and the computed validation metric values have been showcased in **Table 4.16**. The goodness-of-fit, robustness, and predictability of the model have been confirmed by the determination coefficient (R^2 = 0.89), cross-validated correlation coefficient (Q^2_{LOO} = 0.84), and external correlation coefficient (Q^2_{F1} = 0.91) values respectively. The mean absolute errors of both training and test sets are also on the lower side which signifies the correctness of predictions obtained by the model (MAE_{Train} = 0.31, MAE_{Test} = 0.32). Besides these, acceptable values of additional r^2_m metrics ($\overline{r^2_m(Train)} = 0.78, \Delta r^2_m(Train) = 0.08, \overline{r^2_m(Test)} = 0.87, \Delta r^2_m(Test) = 0.06$) and CCC (0.95) indicate the statistical soundness of the model. We have also performed the Y-randomization test with 100 times randomization of response values by employing "SIMCA-P" software; the resultant plot has been given in the following **Figure 4.36D**. The least R^2 and Q^2 intercept values (R^2 intercept = 0.0596; Q^2 intercept = -0.524) obtained by resultant plot discard the possibility of chance correlation. The observed vs. predicted toxicity scatter plot, variable importance plot, and score plot with 2 latent variable spaces have been given in the following **Figure 4.36A**, **4.36B**, and **4.36C** respectively. A uniform scattering has been observed around the line passing through the origin of Cartesian coordinates (**Figure 4.36A**) by plotting the observed and predicted toxicity values in the X and Y axis respectively, which suggests the good fit and predictability of the model.

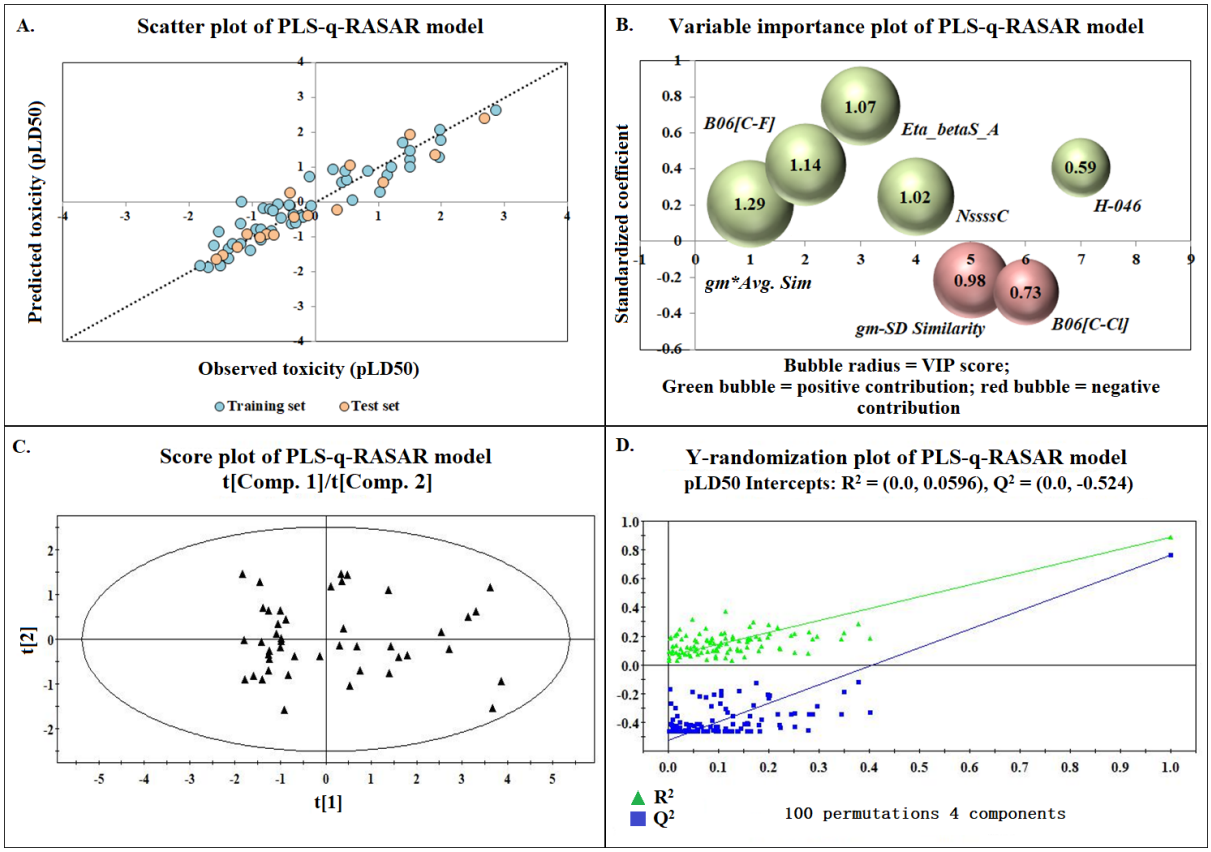


Figure 4.36. A. Scatter plot, B. variable importance plot, C. score plot, and D. Y-randomization plot of the developed PLS q-RASAR model

A true external validation has also been performed as discussed in the "***3.4.6. Validation protocols***" with true external set mixtures. Several classification-based metrics have been computed from the predicted responses obtained by the PLS q-RASAR model and reported in the following **Table 4.17**. Good sensitivity, specificity, and accuracy values indicate good classifiability of the model. The computed G-means value (0.89) indicates the balanced performance of the model between toxic and non-toxic classes. The computed values of Matthews correlation coefficient and Cohen's κ are 0.69 and 0.65 respectively which suggest an acceptable prediction and an agreement between observed and predicted classification.

Table 4.17. Classification-based metrics* obtained from true external validation

Classification-based metrics	Computed values
Sensitivity	1.00
Specificity	0.80
Accuracy	0.85
Precision	0.60
F-measure	0.75
G-means	0.89

MCC	0.69
Cohen's κ	0.65

MCC: Matthews correlation coefficient; *The threshold for the classification is the response mean of the training set (-0.127 μg active substance/bee)

To further enhance the predictability of the developed q-RASAR model, ML algorithms have been used in this study. Linear ML approaches such as ridge regression and linear SVM and non-linear ML approaches like non-linear SVM were employed to develop models using the identified descriptors by the previous best subset selection algorithm. After obtaining the predictions, we computed the Q^2_{F1} value as an identifier of external predictability. The computed Q^2_{F1} values of linear SVM, ridge regression, and non-linear SVM were 0.913, 0.905, and 0.935 respectively. The predictability of non-linear SVM was slightly higher than the other ML approaches and the previous PLS q-RASAR model (PLS q-RASAR: Q^2_{F1} = 0.91) as well. The mean absolute error of prediction of test set (MAE_{Test}) has also been computed for three ML methods. Least error has been observed in non-linear SVM (MAE_{Test}: non-linear SVM = 0.195, linear SVM = 0.256, ridge regression = 0.265). The computed validation metrics of these three ML methods were computed from the predicted responses and reported in the following **Table 4.18**.

Table 4.18. Computed validation metrics from the predictions obtained by the ML-based q-RASAR models

Internal validation metrics	Linear SVM	Ridge rigression	Non-linear SVM
R^2_{Train}	0.874	0.888	**0.959**
Q^2_{LOO}	0.809	0.831	0.735
MAE_{Train}	0.251	0.260	**0.147**
External validation metrics			
R^2_{Test}	0.919	0.914	**0.947**
Q^2_{F1}	0.913	0.905	**0.935**
Q^2_{F2}	0.912	0.904	**0.934**
Q^2_{F3}	0.902	0.894	**0.927**
CCC	0.952	0.950	**0.962**
$\overline{r_m^2(Test)}$	0.884	0.883	0.836
$\Delta r_m^2(Test)$	0.061	0.055	0.051
MAE_{Test}	0.256	0.265	**0.195**
$RMSE_P$	0.310	0.323	**0.268**

$RMSE_P$: root mean squared error of predictions

4.4.3. Applicability domain analysis of the PLS q-RASAR model

The applicability domain of a statistical model is the theoretical region in the chemical space (defined by the modeled descriptors and response values) in which the predictions obtained by the model are reliable (Gadaleta et al., 2016). The prediction of toxicity/activity/property of all possible molecules in the universe with a single statistical model is practically impossible; the query compounds must have optimum structural similarity with the training compounds for obtaining a reliable prediction. Therefore, checking the applicability domain is a mandatory requirement of any statistical model, and it is also suggested by OECD principle 3 (OECD, 2004). To comply with the OECD protocol, applicability domain analysis of the newly developed PLS q-RASAR model has been performed with the SIMCA-P software by employing the DModX approach with 99% confidence level. The DModX (distance to model in descriptor space) plots for both the training and test set mixtures have been showcased in **Figure 4.37**. A 99% threshold line (as standard; D-Crit = 0.00999898) with respect to 4 latent variable space of the training data is considered as the D-critical line, and it indicates the applicability domain of the corresponding model beyond which the predictions are unreliable. In this study, all the training set mixtures (**Figure 4.37A**) and test set mixtures (**Figure 4.37B**) are inside the D-Critical line which indicates the reliability of predictions by the model. A score plot of the training set has also been constructed (**Figure 4.36C**) in this study by using the SIMCA-P software which tells us about the outlier mixture in the 2 latent variable space. However, there is no mixture beyond the threshold line defined by the ellipse.

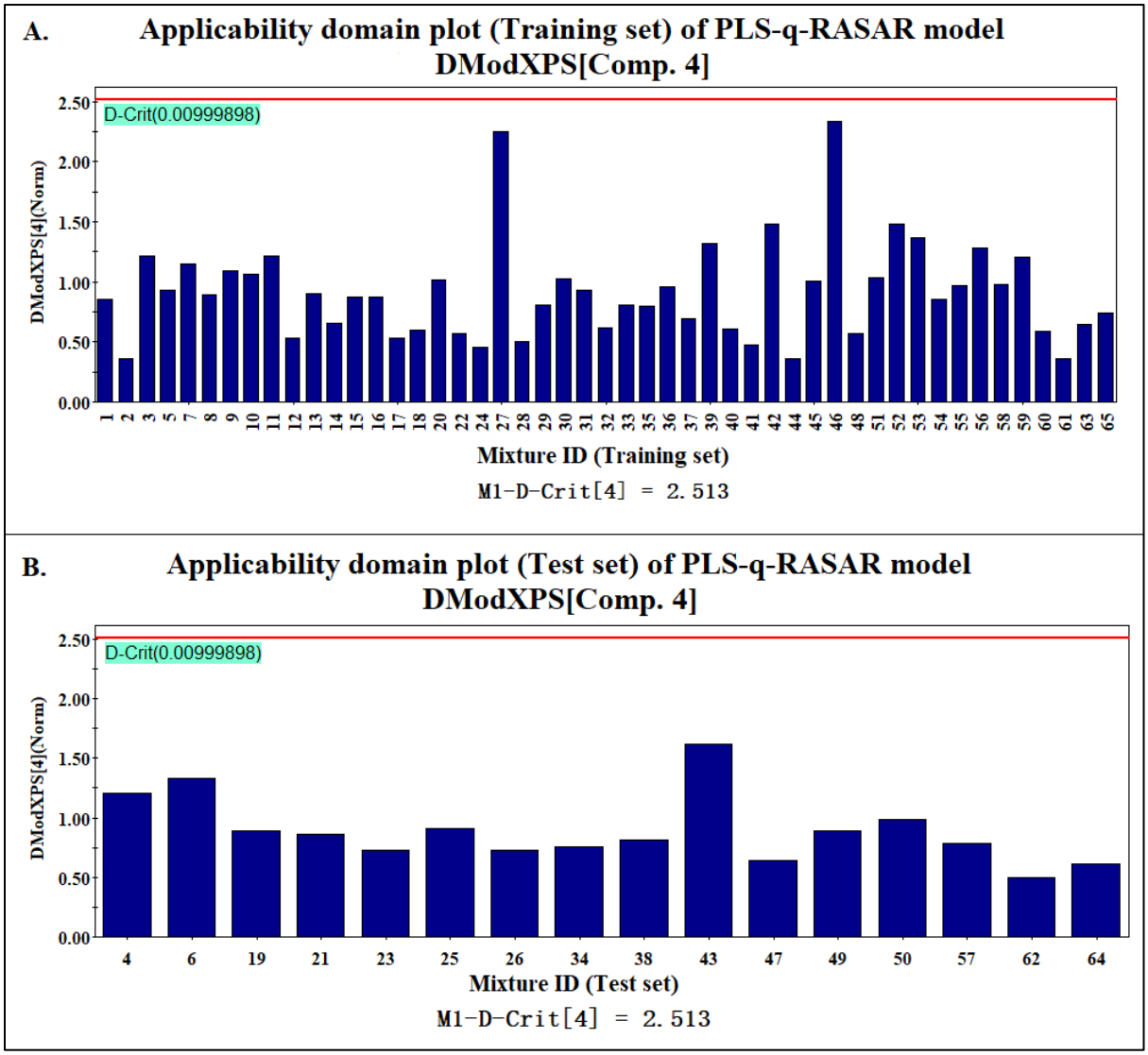


Figure 4.37. Applicability domain (DModX) plot of the PLS q-RASAR model obtained by "SIMCA-P" software (A = Training set; B = Test set)

4.4.4. Insights on modeled descriptors

In the section "***4.4.1.Developed q-RASAR models and best mixing rule identification***", seven descriptors of class "Mixture Descriptors 3" have been identified as the important features which affect the toxicity to a great extent; thus, B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC, g_m*Avg.Sim*, g_m*SD Similarity* are the descriptors of interest in this study. Among these seven descriptors, B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC are the 2D-structural descriptors computed by "AlvaDesc" software whereas, g_m*Avg.Sim* and g_m*SD Similarity* are the RASAR descriptors computed by the "RASAR Descriptor Calculator v2.0" tool. The algebraic sign (positive/negative) of a regression coefficient is a major diagnostic feature of the type of descriptor contribution. A positive regression coefficient generally suggests a proportional relationship between the descriptor and the endpoint (provided other contributing descriptor values remain the same), whereas a negative regression coefficient suggests an inversely proportional relationship (given that other contributing descriptor values remain the same). This gives a general idea about the contribution of the modeled descriptors toward the endpoint. However, with changes in

structural features, values of multiple descriptors (with positive and negative contributions) change simultaneously, and the observed endpoint value is a function of all (positive and negative) contributing features. Thus, to demonstrate the effect of a particular descriptor toward the endpoint, only representative and most suitable examples might be given which may not be universally applicable for all possible examples. The definitions, type, and the type of contribution towards toxicity (positive or negative) of modeled descriptors have been given in the following **Table 4.19**.

Table 4.19. Definitions, type, and direction of the coefficient (positive or negative) of modeled descriptors

Descriptors	Definition	Type	Type of contribution toward toxicity
B06[C-Cl]	Presence or absence of C-Cl atom pair at topological distance 6	2D-atom pairs descriptors	Negative (-)
B06[C-F]	Presence or absence of C-F atom pair at topological distance 6	2D-atom pairs descriptors	Positive (+)
Eta_betaS_A	Eta Sigma average VEM count	Electro topochemical atom index descriptors	Positive (+)
H-046	H attached to CO (sp^3) no X attached to next C	Atom-centred fragments	Positive (+)
NssssC	Number of atom/s of type ssssC	Atom-type E-state indices	Positive (+)
g_m*Avg.Sim	g_m × average similarity of closed source compounds where gm is a concordance measure	Concordance measure: RASAR descriptors	Positive (+)
g_m*SD Similarity	g_m × standard deviation of closed source compounds where g_m is a concordance measure	Concordance measure: RASAR descriptors	Negative (-)

The relative importance of modeled descriptors towards toxicity has been checked with the SIMCA-P software and represented as a bubble plot in **Figure 4.36B**. Each descriptor has been depicted as the bubbles with variable diameters where the diameter of the bubbles corresponds to the VIP score of individual descriptors. According to Wold et al., descriptors

with VIP scores greater than 1 represent higher importance towards toxicity prediction by the developed model (Wold et al., 2001). Therefore, g_m**Avg.Sim*, B06[C-F], Eta_betaS_A, and NssssC affect the toxicity prediction of organic mixtures against honey bees to a greater extent in this newly developed q-RASAR model. In **Figure 4.36B**, the relative importance of modeled descriptors has been plotted along X axis sequentially in descending order from left to right; therefore, the order of importance towards toxicity is g_m**Avg.Sim* > B06[C-F] > Eta_betaS_A> NssssC > g_m**SD Similarity* > B06[C-Cl] > H-046. Among these seven descriptors, B06[C-Cl] and g_m**SD Similarity* have negative contributions towards toxicity prediction; the rest are positively contributed descriptors. One can analyze the dataset mixtures for understanding the type of contribution which further justifies the appropriateness of the model. Ideally, to know a more comprehensive contributing trend of a specific descriptor, a comparative analysis should be performed among dataset mixtures by keeping the values of other descriptors constant. But the prediction of the endpoint (here pLD50) is derived from all the modeled descriptors simultaneously and with changes in structural features, different descriptor values change simultaneously; therefore, the other descriptor values cannot be kept constant to study the effect of a particular descriptor. Thus, the general relationship between the respective descriptor and endpoint identified by the model may not be followed simultaneously by all the mixtures in a dataset (although the general trend can be identified by a comparative analysis of dataset mixtures which has been done in **Table 4.20**).

Table 4.20. Comparison of descriptor values and toxicity of different mixtures

g_m * *Avg. Sim*. (positive contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**41**	0.687	2.886
	60	0.496	1.523
2	**41**	0.687	2.886
	20	0.0	0.444
3	**41**	0.687	2.886
	44	-0.967	-1.806
B06[C-F] (positive contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**41**	0.5	2.886
	60	0	1.523
2	**56**	0.5	2.886
	01	0	-0.068

3	**41**	0.5	2.886
	44	0	-1.806

Eta_betaS_A (positive contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**56**	0.479	2.013
	60	0.475	1.523
2	**10**	0.467	1.046
	48	0.458	0.602
3	**13**	0.446	-0.215
	44	0.391	-1.806

NssssC (positive contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**41**	1	2.886
	56	0.5	2.013
2	**41**	1	2.886
	35	0.33	0.854
3	**41**	1	2.886
	44	0	-1.806

g_m*SD Similarity **(negative contribution)**

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**41**	0.094	2.886
	40	0.144	1.523
2	**56**	0.196	2.013
	07	0.203	0.301

3	**59**	0.007	2
	05	0.032	1.155

B06[C-Cl] (negative contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**41**	0	2.886
	56	0.5	2.013
2	**41**	0	2.886
	35	0.67	0.854
3	**41**	0	2.886
	22	0.707	-1.354

H-046 (positive contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**41**	6.946	2.886
	39	3.5	1.398
2	**41**	6.946	2.886
	05	2.5	1.155
3	**41**	6.946	2.886
	03	0	-0.841

The g_m*Avg.Sim* is a RASAR descriptor represented by the product of concordance measure g_m with the average of similarity values of closed source mixtures against each query mixture (Banerjee and Roy, 2023). A sharp decrease in descriptor values has been observed (Mixture **41**: g_m*Avg.Sim* = 0.687; Mixture **44**: g_m*Avg.Sim* = -0.967) from the most toxic mixture to the least toxic mixture in the dataset (Mixture **41**: pLD_{50}= 2.886; Mixture **44**: pLD_{50} = -1.806). There are several other examples in the dataset (See **Table 4.20**) which justify the proportional relationship (positive contribution) of the most important descriptor of this study.

B06[C-F] is another descriptor of greater relative importance. It is a 2D atom pair descriptor which indicates the presence or absence of C-F atom pairs at the topological distance 6. The presence of the specified fragment has been depicted in the 2D structure of the mixture component cyhalothrin (**C7**) in **Figure 4.38**. A proportional relationship between B06[C-F]

descriptor and toxicity has been observed in this study. There are several examples represented in **Table 4.20** which agree with this proportional relationship. It has also been observed that the least toxic mixture **44** does not have a non-zero descriptor value (mixture **44**: composed of thymol and diethyl but-2-enedioate; pLD_{50}= -1.806; B06[C-F] = 0) whereas, the most toxic mixture **41** (mixture of cyhalothrin and piperonyl butoxide; pLD_{50} = 2.886; B06[C-F] = 0.5) is having a positive descriptor value. This descriptor indicates a positive contribution as well as the presence of lipophilic F as a structural componenrt. Therefore, lipophilicity may enhance the mixture toxicity to honeybee.

Eta_betaS_A ($\sum \beta_S'$) is an electro-topochemical atom descriptor which is the next important identified structural feature of this study. Roy and Das defined this descriptor as "*summed β values for all the sigma bonds (VEM sigma contribution) relative to the number of vertices*" (Roy and Das, 2015). It represents the electron richness of any chemical relative to the molecular size. We have depicted the presence of electronegative atoms with coloured symbols in the structure of a mixture component imidacloprid (**C13**) in **Figure 4.38**. It has been observed that a mixture of chemicals with less electron richness (e.g., mixture **44**: composed with thymol and diethyl but-2-enedioate; pLD_{50} = -1.806; Eta_btaS_A = 0.392) has less toxicity than a mixture of chemicals with greater electron richness (e.g., mixture **56**: composed of imidacloprid and triflumizole; pLD_{50} =2.03; Eta_betaS_A = 0.479). We have presented several other pairs of mixtures (see **Table 4.20**) which signify this proportional relationship. Therefore the presence of electronegative atoms in mixture components as identified by Eta_betaS_A descriptor positively contributes towards mixture toxicity against honey bees.

The next important descriptor is NssssC (an atom type E-state indices) which represents the number of atoms of type ssssC; where "s" indicates the attachment of single-bonded atoms apart from H. In mixture component triflumizole (**C26**), the ssssC type fragment (C-CF_3) has been pointed and has been depicted in **Figure 4.38**. It also has a positive contribution towards toxicity as identified by the standardized coefficient plotted in **Figure 4.36B**, and is evident from the comparative analysis between 3 pairs of studied mixtures represented in **Table 4.20**. The most toxic mixture has shown a greater NssssC value (e.g., mixture **41**, composed of cyhalothrin and piperonyl butoxide; pLD_{50} = 2.886; NssssC = 1) than the least toxic mixture (e.g., mixture **44**, composed of thymol and diethyl but-2-enedioate; pLD_{50} = -1.806; NssssC = 0).

The g_m*SD Similarity* is another RASAR descriptor with comparatively less importance (than the other RASAR descriptor). It is the product of concordance measure g_m with the standard deviation of similarity values of closed source mixtures against each query mixture of read-across prediction (Banerjee and Roy, 2023). Unlike the other RASAR descriptor of this study (g_m*Avg.Sim*), g_m*SD Similarity* has a negative contribution towards toxicity prediction. This inverse relationship can be seen in several mixtures; we have represented 3 such mixture pairs in **Table 4.20**. It has also been observed that the most toxic mixture of the dataset has a lower *gm* SD Similarity* value (mixture **41**, composed of cyhalothrin and piperonyl butoxide; pLD_{50} = 2.886; *gm* SD Similarity* = 0.094) than a comparatively less toxic mixture (mixture **40**, composed of cyhalothrin and tribufos; pLD_{50} = 1.523; *gm* SD Similarity* = 0.144).
B06[C-Cl] is a less important structural descriptor in this study which indicates the presence or absence of C-Cl atom pair at topological distance 6. In **Figure 4.38**, the specified fragment has been depicted in the structure of the mixture component boscalid (**C3**). Mixtures containing this fragment have higher mixture descriptor values and thus less toxicity as observed in mixtures represented in **Table 4.20**. There is also no C-Cl atom pair at the

topological distance 6 in the structural components of the most toxic mixture in the dataset (mixture **41** which is composed of cyhalothrin and piperonyl butoxide), which again justifies the negative correlation between B06[C-Cl] and exerted toxicity.

Among the modeled descriptors, H-046 has the least importance. It is an atom-centred fragment descriptor which indicates the number of H attached to sp^3 hybridized carbon atoms with no attachment of heteroatom-connected adjacent carbon. Such 'H's have been marked in the structure of a mixture component thymol (**C23**) in **Figure 4.38**. It possesses a positive coefficient, and thus has a positive contribution towards toxicity. A marked decrement in both H-046 and toxicity values has been observed in several mixtures noted in **Table 4.20** (e.g., mixture **41**, composed of cyhalothrin and piperonyl butoxide; H-046 =6.946; pLD_{50} = 2.886; and mixture **3**, composed of acetamiprid and diethyl but-2-enedioate; H-046 = 0, pLD_{50} = -0.841) which justify the proportional relationship between H-046 and toxicity (pLD_{50}).

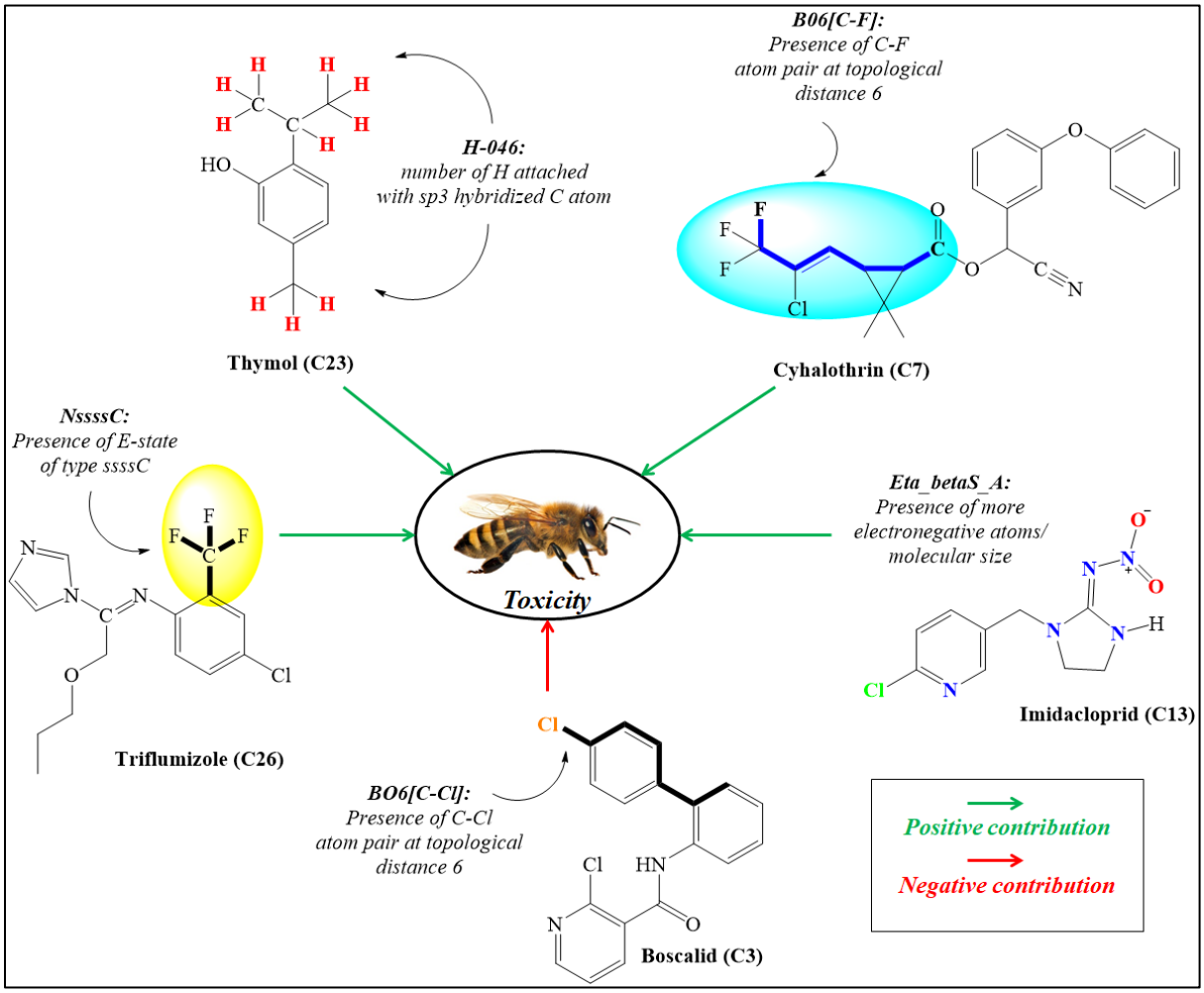


Figure 4.38. Depiction of the presence of 2D-structural descriptor fragments in 2D structure of mixture components

4.4.5. Prediction of toxicity of unknown mixtures and reliability analysis

In this communication, the toxicity of 218 untested binary mixtures against honey bees was predicted by the developed PLS q-RASAR model and also with the ML-based q-RASAR. We got four sets of predictions from PLS q-RASAR, linear SVM q-RASAR, ridge regression, and non-linear SVM q-RASAR respectively. The averages of these four sets of predictions have also been computed. All these predictions along with their averages have been reported in **Table 4.21**.

Table 4.21. Prediction of untested mixtures obtained by PLS-q-RASAR model and three machine learning-based q-RASAR models; reliability of predictions and applicability domain status with respect to PLS-q-RASAR model

Sl. No.	Prediction by PLS-q-RASAR	Prediction by Linear SVM-q-RASAR	Prediction by Ridge-q-RASAR	Prediction by non-linear SVM-q-RASAR	Consensus (average) of all predictions	AD od PLS-q-RASAR model	Reliability of prediction
UM 1	-0.539	-0.525	-0.291	-0.418	-0.443	In	Good
UM 2	0.718	0.639	0.768	0.631	0.689	In	Good
UM 3	-0.053	-0.087	0.133	-0.373	-0.095	In	Good
UM 4	0.623	0.580	0.614	1.294	0.778	In	Good
UM 5	0.676	0.585	0.679	0.851	0.698	In	Good
UM 6	-0.816	-0.499	-0.543	-1.004	-0.716	In	Good
UM 7	1.325	1.032	1.147	0.878	1.096	In	Good
UM 8	1.539	1.220	1.342	1.059	1.290	In	Good
UM 9	1.457	1.186	1.243	0.926	1.203	In	Good
UM 10	-1.315	-0.869	-0.965	-1.042	-1.048	In	Good
UM 11	-1.085	-0.636	-0.759	-0.991	-0.868	In	Good
UM 12	-0.854	-0.514	-0.630	-0.927	-0.731	In	Good

UM 13	-1.617	-1.042	-1.181	-1.166	-1.251	In	Good
UM 14	-1.959	-1.304	-1.468	-1.397	-1.532	In	Good
UM 15	-1.453	-0.916	-1.043	-1.027	-1.110	In	Good
UM 16	-0.865	-0.484	-0.553	-0.527	-0.607	In	Good
UM 17	-0.261	-0.127	-0.071	-0.098	-0.139	In	Good
UM 18	-1.618	-1.044	-1.201	-1.406	-1.317	In	Good
UM 19	0.386	0.426	0.403	0.385	0.400	In	Good
UM 20	0.790	0.686	0.722	0.369	0.642	In	Good
UM 21	0.877	0.803	0.771	0.616	0.767	In	Good
UM 22	0.523	0.852	0.638	0.595	0.652	In	Moderate
UM 23	1.728	1.624	1.610	1.698	1.665	In	Good
UM 24	1.718	1.568	1.524	1.152	1.491	In	Good
UM 25	2.158	1.903	1.959	1.820	1.960	In	Good
UM 26	3.297	2.669	2.891	1.951	2.702	In	Good
UM 27	2.030	1.789	1.761	1.376	1.739	In	Good
UM 28	2.437	2.093	2.126	1.607	2.066	In	Good
UM 29	2.594	2.210	2.250	1.544	2.150	In	Good
UM 30	2.292	2.014	1.987	1.626	1.980	In	Good
UM 31	2.573	2.172	2.297	2.021	2.265	In	Good
UM 32	3.336	2.774	2.911	1.433	2.614	Outside AD	Bad/Unreliable
UM 33	3.353	2.718	2.940	1.995	2.751	In	Good
UM 34	1.235	1.313	1.213	1.278	1.260	In	Good

UM 35	3.065	2.599	2.728	1.632	2.506	Outside AD	Bad/Unreliable
UM 36	1.951	1.734	1.725	1.311	1.680	In	Good
UM 37	2.018	1.758	1.756	1.320	1.713	In	Good
UM 38	2.670	2.216	2.360	1.945	2.298	In	Good
UM 39	2.724	2.278	2.421	2.043	2.367	In	Good
UM 40	2.292	2.014	1.987	1.626	1.980	In	Good
UM 41	-0.281	-0.080	-0.135	-0.250	-0.187	In	Good
UM 42	-0.181	-0.007	-0.107	-0.586	-0.220	In	Good
UM 43	-0.318	-0.108	-0.159	-0.175	-0.190	In	Good
UM 44	-0.741	-0.408	-0.484	-0.345	-0.494	In	Good
UM 45	0.185	0.214	0.332	0.169	0.225	In	Good
UM 46	-0.978	-0.681	-0.621	-0.548	-0.707	In	Good
UM 47	0.385	0.404	0.492	0.569	0.463	In	Good
UM 48	0.298	0.325	0.454	0.433	0.377	In	Good
UM 49	0.764	0.618	0.693	0.336	0.602	In	Good
UM 50	1.520	1.228	1.322	1.202	1.318	In	Good
UM 51	-0.337	-0.097	-0.217	-0.710	-0.340	In	Good
UM 52	-0.654	-0.312	-0.417	-0.235	-0.404	In	Good
UM 53	-1.336	-0.833	-0.988	-0.984	-1.035	In	Good
UM 54	0.178	0.296	0.218	0.137	0.207	In	Good
UM 55	0.212	0.290	0.244	0.193	0.235	In	Good
UM 56	-0.946	-0.534	-0.660	-0.716	-0.714	In	Good

UM 57	0.394	0.431	0.413	0.404	0.410	In	Good
UM 58	0.010	0.139	0.083	-0.164	0.017	In	Good
UM 59	-0.292	-0.089	-0.157	-0.513	-0.263	In	Good
UM 60	-1.441	-0.914	-1.077	-1.036	-1.117	In	Good
UM 61	0.013	0.144	0.099	0.029	0.071	In	Good
UM 62	0.718	0.679	0.630	0.391	0.605	In	Good
UM 63	0.186	0.277	0.245	0.355	0.266	In	Good
UM 64	0.100	0.044	0.266	-0.256	0.038	In	Good
UM 65	-0.096	0.077	0.067	-0.066	-0.005	In	Good
UM 66	3.297	2.669	2.891	1.951	2.702	In	Good
UM 67	0.979	0.851	0.851	0.762	0.861	In	Good
UM 68	-0.567	-0.013	-0.250	0.438	-0.098	Outside AD	Bad/Unreliable
UM 69	0.523	0.476	0.615	1.230	0.711	In	Good
UM 70	3.112	2.607	2.736	2.607	2.766	In	Good
UM 71	-0.096	0.077	0.067	-0.066	-0.005	In	Good
UM 72	2.028	1.625	1.760	1.435	1.712	In	Good
UM 73	0.284	0.197	0.423	-0.091	0.203	In	Good
UM 74	-0.412	-0.158	-0.177	-0.153	-0.225	In	Good
UM 75	-0.281	0.184	-0.015	0.516	0.101	Outside AD	Bad/Unreliable
UM 76	1.280	1.006	1.129	0.988	1.101	In	Good
UM 77	1.372	1.492	1.395	0.480	1.184	In	Moderate
UM 78	2.191	1.901	1.972	1.611	1.919	In	Good

UM 79	0.005	0.375	0.218	0.623	0.305	Outside AD	Bad/Unreliable
UM 80	0.156	0.250	0.171	-0.222	0.089	In	Good
UM 81	0.114	0.267	0.191	-0.031	0.135	In	Good
UM 82	1.009	0.822	0.891	0.628	0.837	In	Good
UM 83	-0.406	-0.149	-0.248	-0.535	-0.334	In	Good
UM 84	-1.116	-0.660	-0.759	-0.710	-0.811	In	Good
UM 85	-0.695	-0.075	-0.437	0.378	-0.207	In	Moderate
UM 86	0.427	0.460	0.428	0.426	0.435	In	Good
UM 87	1.000	0.836	0.937	1.262	1.009	In	Good
UM 88	-0.518	0.027	-0.208	0.448	-0.063	Outside AD	Bad/Unreliable
UM 89	-0.005	0.347	0.180	0.670	0.298	In	Moderate
UM 90	1.008	0.901	0.898	0.739	0.887	In	Good
UM 91	-1.494	-1.023	-1.162	-1.375	-1.263	In	Good
UM 92	-2.010	-1.413	-1.534	-1.650	-1.652	In	Good
UM 93	-2.010	-1.413	-1.534	-1.650	-1.652	In	Good
UM 94	0.109	0.126	0.143	0.419	0.199	In	Good
UM 95	0.444	0.365	0.431	0.399	0.410	In	Good
UM 96	0.388	0.363	0.351	0.285	0.347	In	Good
UM 97	2.124	1.869	1.929	1.905	1.957	In	Good
UM 98	-0.451	-0.241	-0.288	-0.045	-0.256	In	Good
UM 99	-0.442	-0.237	-0.281	-0.048	-0.252	In	Good
UM 100	-0.675	-0.442	-0.393	-0.340	-0.463	In	Good

UM 101	-0.983	-0.415	-0.620	0.282	-0.434	In	Moderate
UM 102	0.479	0.893	0.606	-0.148	0.457	In	Moderate
UM 103	1.751	1.412	1.500	0.534	1.299	Outside AD	Bad/Unreliable
UM 104	-0.403	-0.186	-0.169	-0.072	-0.208	In	Good
UM 105	0.267	0.338	0.302	0.300	0.302	In	Good
UM 106	0.502	0.516	0.491	0.489	0.500	In	Good
UM 107	1.150	0.950	1.091	1.666	1.214	In	Good
UM 108	0.433	0.458	0.400	0.328	0.404	In	Good
UM 109	0.256	0.286	0.303	0.857	0.425	In	Good
UM 110	-1.396	-0.872	-0.995	-0.975	-1.060	In	Good
UM 111	-0.335	-0.336	-0.111	-0.482	-0.316	In	Good
UM 112	0.593	0.530	0.674	1.304	0.775	In	Good
UM 113	0.390	0.369	0.424	0.210	0.348	In	Good
UM 114	2.037	1.759	1.827	1.955	1.895	In	Good
UM 115	1.869	1.656	1.622	1.251	1.600	In	Good
UM 116	2.677	2.226	2.358	2.231	2.373	In	Good
UM 117	-0.906	-0.503	-0.627	-0.683	-0.680	In	Good
UM 118	-0.728	-0.386	-0.439	-0.417	-0.493	In	Good
UM 119	0.368	0.443	0.381	0.539	0.433	In	Good
UM 120	-0.245	-0.021	-0.111	0.090	-0.072	In	Good
UM 121	-1.176	-0.768	-0.849	-0.942	-0.933	In	Good
UM 122	-1.327	-0.820	-0.937	-0.910	-0.999	In	Good

UM 123	-0.498	-0.270	-0.304	-0.676	-0.437	In	Good
UM 124	2.322	1.992	2.030	1.438	1.945	In	Good
UM 125	-0.474	-0.214	-0.285	-0.532	-0.376	In	Good
UM 126	-0.080	0.057	0.023	0.008	0.002	In	Good
UM 127	-0.156	-0.024	-0.043	-0.544	-0.192	In	Good
UM 128	0.797	0.732	0.705	0.833	0.767	In	Good
UM 129	-0.673	-0.389	-0.428	-0.542	-0.508	In	Good
UM 130	-1.374	-0.890	-1.049	-1.349	-1.166	In	Good
UM 131	-0.081	0.106	0.048	0.576	0.162	In	Good
UM 132	-0.709	-0.373	-0.423	-0.401	-0.477	In	Good
UM 133	2.282	1.988	1.984	1.643	1.975	In	Good
UM 134	0.179	0.278	0.283	1.026	0.441	In	Good
UM 135	-0.393	-0.133	-0.184	-0.368	-0.270	In	Good
UM 136	1.962	1.577	1.721	0.684	1.486	Outside AD	Bad/Unreliable
UM 137	-0.003	0.130	0.066	-0.300	-0.027	In	Good
UM 138	-0.356	-0.084	-0.200	-0.418	-0.264	In	Good
UM 139	0.535	0.539	0.460	0.187	0.430	In	Good
UM 140	-0.963	-0.553	-0.633	-0.602	-0.688	In	Good
UM 141	0.678	0.675	0.603	0.258	0.553	In	Good
UM 142	-1.077	-0.657	-0.790	-0.656	-0.795	In	Good
UM 143	-0.144	0.051	-0.050	-0.327	-0.118	In	Good
UM 144	0.097	0.234	0.133	-0.275	0.047	In	Good

UM 145	-0.172	0.335	-0.026	0.713	0.213	Outside AD	Bad/Unreliable
UM 146	-0.581	-0.212	-0.390	0.339	-0.211	In	Good
UM 147	-0.084	0.175	0.030	0.865	0.246	In	Good
UM 148	-0.153	-0.024	-0.040	0.329	0.028	In	Good
UM 149	0.056	0.130	0.134	0.589	0.227	In	Good
UM 150	-1.291	-0.859	-0.987	-0.991	-1.032	In	Good
UM 151	0.678	0.675	0.603	0.258	0.553	In	Good
UM 152	0.283	0.449	0.336	1.121	0.548	In	Good
UM 153	-0.175	0.000	-0.057	-0.273	-0.126	In	Good
UM 154	-0.540	-0.237	-0.306	-0.491	-0.393	In	Good
UM 155	0.281	0.355	0.343	0.903	0.470	In	Good
UM 156	0.290	0.296	0.392	0.810	0.447	In	Good
UM 157	0.407	0.382	0.434	0.164	0.347	In	Good
UM 158	0.073	0.130	0.210	0.451	0.216	In	Good
UM 159	1.568	1.524	1.505	1.208	1.451	In	Good
UM 160	0.048	0.171	0.118	-0.093	0.061	In	Good
UM 161	0.218	0.299	0.263	0.209	0.247	In	Good
UM 162	0.757	0.597	0.715	0.333	0.601	In	Good
UM 163	2.240	1.794	1.920	0.661	1.654	Outside AD	Bad/Unreliable
UM 164	-0.352	0.230	-0.145	0.554	0.072	Outside AD	Bad/Unreliable
UM 165	0.256	0.286	0.303	0.857	0.425	In	Good
UM 166	-1.477	-0.938	-1.107	-1.415	-1.234	In	Good

UM 167	-0.731	-0.463	-0.444	-0.415	-0.513	In	Good
UM 168	-1.309	-0.814	-0.982	-0.965	-1.017	In	Good
UM 169	2.024	1.615	1.861	0.701	1.550	Outside AD	Bad/Unreliable
UM 170	4.082	3.300	3.532	1.004	2.980	Outside AD	Bad/Unreliable
UM 171	2.059	1.642	1.891	0.704	1.574	Outside AD	Bad/Unreliable
UM 172	4.115	3.327	3.560	1.005	3.002	Outside AD	Bad/Unreliable
UM 173	1.998	1.605	1.751	0.690	1.511	Outside AD	Bad/Unreliable
UM 174	3.511	2.728	3.067	0.442	2.437	Outside AD	Bad/Unreliable
UM 175	-1.147	-0.706	-0.880	-0.999	-0.933	In	Good
UM 176	1.244	1.411	1.205	0.504	1.091	In	Good
UM 177	0.609	1.062	0.688	0.521	0.720	Outside AD	Bad/Unreliable
UM 178	2.223	2.008	1.987	2.264	2.121	In	Good
UM 179	3.072	2.375	2.553	0.867	2.216	Outside AD	Bad/Unreliable
UM 180	2.839	2.222	2.372	1.216	2.162	In	Good
UM 181	2.776	2.152	2.301	1.023	2.063	In	Good
UM 182	0.144	0.129	0.187	-0.277	0.046	In	Good
UM 183	2.350	1.944	2.066	1.789	2.037	In	Good
UM 184	-0.798	-0.425	-0.564	-0.379	-0.541	In	Good
UM 185	2.162	1.845	1.906	1.937	1.963	In	Good
UM 186	0.697	0.575	0.652	0.319	0.561	In	Good
UM 187	0.519	0.434	0.468	0.218	0.410	In	Good
UM 188	-1.447	-0.911	-1.037	-1.021	-1.104	In	Good

UM 189	-1.870	-1.236	-1.393	-1.345	-1.461	In	Good
UM 190	-1.112	-0.726	-0.792	-1.306	-0.984	In	Good
UM 191	-0.418	-0.217	-0.260	-0.003	-0.225	In	Good
UM 192	-0.026	0.115	0.067	-0.042	0.028	In	Good
UM 193	0.083	0.244	0.168	-0.054	0.110	In	Good
UM 194	-1.077	-0.660	-0.827	-0.931	-0.874	In	Good
UM 195	0.040	0.162	0.083	-0.354	-0.017	In	Good
UM 196	-0.417	-0.198	-0.181	-0.091	-0.222	In	Good
UM 197	0.802	0.713	0.715	0.657	0.722	In	Good
UM 198	-0.084	-0.114	0.107	-0.392	-0.121	In	Good
UM 199	0.165	0.197	0.235	-0.172	0.106	In	Good
UM 200	1.218	1.193	1.222	0.566	1.050	In	Good
UM 201	-1.727	-1.126	-1.273	-1.249	-1.344	In	Good
UM 202	-0.034	0.047	0.120	0.273	0.102	In	Good
UM 203	-0.123	-0.031	0.024	-0.311	-0.110	In	Good
UM 204	-0.330	-0.331	-0.107	-0.481	-0.313	In	Good
UM 205	1.576	1.324	1.340	0.950	1.298	In	Good
UM 206	-0.310	0.000	-0.161	0.648	0.044	In	Good
UM 207	2.381	2.290	2.141	1.826	2.160	In	Good
UM 208	0.312	0.666	0.345	0.938	0.565	In	Moderate
UM 209	0.682	0.948	0.682	1.184	0.874	In	Moderate
UM 210	0.641	0.730	0.639	1.268	0.819	In	Good

UM 211	0.623	0.580	0.614	1.294	0.778	In	Good
UM 212	0.261	0.440	0.319	1.115	0.534	In	Good
UM 213	0.745	0.677	0.718	1.401	0.885	In	Good
UM 214	0.064	0.214	0.117	0.011	0.101	In	Good
UM 215	1.450	1.266	1.252	0.840	1.202	In	Good
UM 216	0.959	0.845	0.899	1.532	1.059	In	Good
UM 217	-0.450	-0.269	-0.203	-0.047	-0.242	In	Good
UM 218	1.699	1.605	1.559	1.869	1.683	In	Good

*UM: Untested mixtures; C: components

The reliability of PLS q-RASAR predictions of these untested mixtures has been checked with the prediction reliability indicator (PRI) tool as discussed in the section "2.7.Prediction of unknown mixtures with PLS and ML-based q-RASAR models and reliability analysis with PRI tool". The PRI tool also provides information about the applicability domain of the untested mixtures/compounds along with the reliability of predictions. Both the reliability of predictions and applicability domain status of untested mixtures have been given in **Table 4.21**. It has been observed that 91.28% of untested mixtures were inside the domain of applicability of the developed PLS q-RASAR model; therefore, the prediction of those 91.28% mixtures is reliable. The rest 8.72% of the mixtures were outside the AD of the new model; thus unreliable predictions were obtained (**Figure 4.39**). The reliability data obtained from the PRI tool follows the same pattern; "unreliable" predictions have been obtained from 8.72% of query mixtures. Predictions of 87.61% of total mixtures have been obtained with "Good" reliability score, whereas, predictions of 3.67% untested mixtures among total have shown "Moderate" reliability score (**Figure 4.39**).

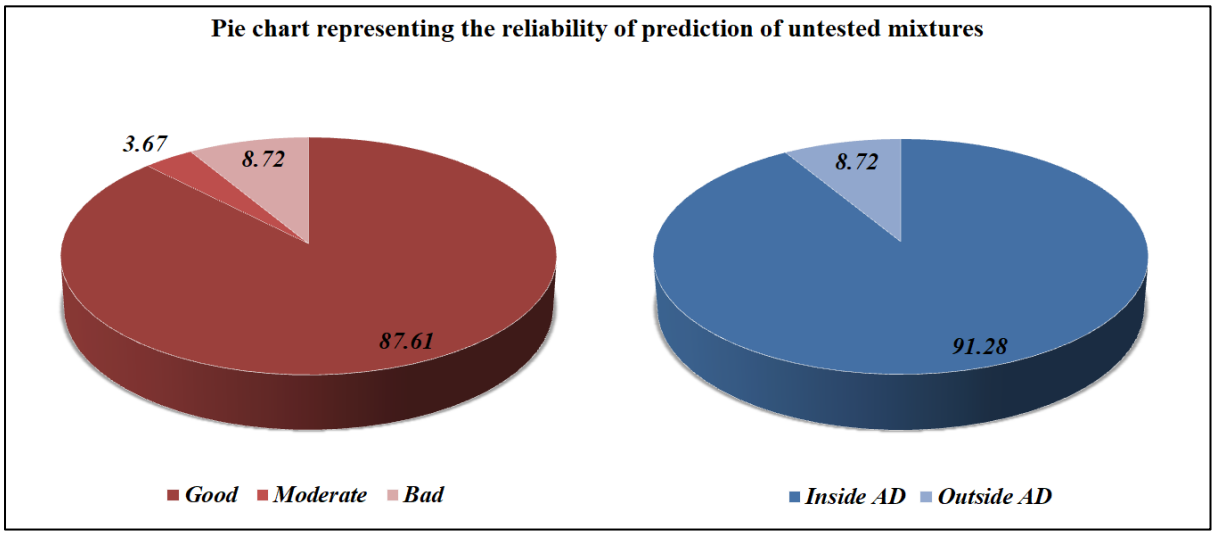


Figure 4.39 Reliability analysis of the predictions of untested mixtures obtained by the PLS q-RASAR model (*All numbers represent the percentage (%) of prediction)

4.4.6. Comparative assessment of the quality of q-RASAR model with previous work

To justify the efficiency of q-RASAR modeling in mixture toxicity predictions, we have compared the quality of the developed models in this study with a previously reported quasi-SMILES-based innovative QSAR model. Carnesecchi et al. reported two different approaches of toxicity predictions of binary mixtures using innovative QSAR modeling by considering the toxic unit data into account (Approach A: without considering the toxic unit; Approach B: considering the toxic unit) (Carnesecchi et al., 2020). Both of these studies (current and previous) used the same mixture dataset against honey bees and thus, they are comparable in terms of goodness-of-fit, robustness, and predictability. In the following **Table 4.22**, the quality metrics (such as R^2, Q^2_{LOO}, Q^2_{F1}, and CCC) of RASAR and the previously reported models have been shown. Both the RASAR and previous QSAR models have comparable goodness-of-fit in terms of R^2 value (R^2 of SVM q-RASAR = 0.96; R^2 of the previous QSAR model-Approach B = 0.97). But, with respect to the robustness and external predictability, both PLS q-RASAR and non-linear SVM q-RASAR outperform the previous QSAR model. The computed Q^2_{LOO} value of PLS q-RASAR model is 0.84 whereas, the Q^2_{CLB} value of the previous QSAR model (Approach B) is 0.81. The SVM q-RASAR model

shows a greater Q^2_{F1} value (0.935) than the previously reported QSAR model (Approach B: R^2_{VLD}=0.89), which indicates better predictability. The PLS q-RASAR model also has better predictability (Q^2_{F1} = 0.91) than the previous model (both Approach A and B). The RASAR models are more precise as indicated by the MAE_{Test} values in comparison to the previous QSAR model (MAE_{Test} of PLS q-RASAR and SVM q-RASAR are 0.32 and 0.23 respectively; MAE_{VLD} of approach A and approach B of previous QSAR models are 0.34 and 0.53 respectively). Besides the quality measures, the q-RASAR models have been developed following the OECD protocols and have also shown reliability during predictions of untested mixtures. The current study has showcased the good classifiability of PLS q-RASAR model which makes it a more appropriate approach for the risk assessment of binary and multicomponent mixtures against honey bees. We may add here that mixture toxicity can also be determined by experiment dependent models like Concentration Addition (CA). In such cases, experimental toxicity data of mixture components is one of the necessary prerequisites. But QSARs or q-RASARs are the mathematical models developed from the existing experimental data and the structural properties of the mixture components. Therefore, one can predict the toxicity of the new/untested mixtures using the q-RASAR model without any additional experimentation. Besides this, the CA model always considers the theory of additivity to obtain the mixture effect; however, there are chances of synergism and antagonism due to interactive nature of chemicals. In this study, we have considered two other mixing rules apart from the simple addition of toxicity. Therefore, more precise predictions are expected from the newly developed q-RASAR model.

Table 4.22. Comparative assessments of quality metrics between q-RASAR and previously reported quasi-SMILES-based innovative QSAR models

Model	*Mix. Desc, Formula*	R^2	Q^2_{LOO}	Q^2_{F1}	*CCC*	MAE_{Test}	*No. of mixtures*
PLS q-RASAR-	$\boldsymbol{Mix.Desc.3 = \sqrt{\sum (p_i \times x_i)^2}}$	0.89	0.84	**0.91**	**0.95**	0.32	65 (Training = 49; Test = 16) Excluding duplicates.
Non-linear SVM q-RASAR	$\boldsymbol{Mix.Desc.3 = \sqrt{\sum (p_i \times x_i)^2}}$	0.96	0.74	**0.935**	**0.96**	0.20	65 (Training = 49; Test = 16) Excluding duplicates
Quasi-SMILES-based QSAR model (Carnesecchi et al., 2020) **Approach A***	Quasi-SMILES fragments	TRN=0.62 iTRN=0.62 **CLB=0.79**	TRN=0.57 iTRN=0.58 **CLB=0.75**	**VLD=0.87**	TRN=0.76 iTRN=0.75 **CLB=0.89**	MAE_{VLD}= 0.34	123 (including duplicate entries)
Quasi-SMILES-based QSAR model (Carnesecchi et al., 2020) **Approach B****	Quasi-SMILES fragments	**TRN= 0.97**(Split 1)**; 0.94** (Split 4)	**CLB=0.81** (Split 1)**; 0.84** (Split 4)	**VLD=0.89** (Split 1); **0.78** (Split 4)	**CLB=0.92** (both in Split 1 and Split 4)	MAE_{VLD}= 0.53 (Split 1); 0.51 (Split 4)	97 (Split 1: TRN=25 iTRN=24 CLB=24 VLD=24 Split 4: TRN=24 iTRN=24 CLB=24 VLD=25)

*Approach A: without considering the toxic unit data; **Approach B: considering the toxic unit data; p_i = descriptor value of i^{th} component; x_i = mass ratio of i_{th} component; *LV*= latent variables; Mix. Desc. = mixture descriptors; TRN: training set; iTRN: invisible training set; CLB: calibration set; VLD: validation set.

4.4.7. Limitations of the developed q-RASAR models

We have computed the mixture descriptors using the assumed mass ratio (50:50) and based on that, we have developed the new PLS/LSVM-based q-RASAR models. A very good correlation between experimental and predicted responses (along with a very low level of model error) has been observed which has been confirmed by the enhanced determination coefficient R^2 (0.89) value and low MAE_{Train} (0.31) value. This justifies the validity of our assumption. Different other validation metrics also suggest the goodness-of-fit, robustness, and good predictability of the developed PLS model. Therefore, we have used this model as the optimum one to predict the toxicity of untested pesticide mixtures to honey bees. The developed model should also be applicable for mixtures with different mass ratios as the descriptors will be computed accordingly reflecting the actual mass ratios. In our understanding, it is better to fill data gaps of mixture toxicity using a model with some approximations and assumptions than having no precise quantitative toxicity data for different untested mixtures.

4.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species

4.5.1. Developed QSAR models and their validations

In the present work, we have developed three QSAR models from MD1, MD2, and MD3 datasets against *V. fischeri*, *E. coli*, and *B. subtilis* respectively, by employing PLS regression with the optimum number of latent variables. As discussed earlier in the "Dataset division and best mixing rule identification" section, mixing rule 2 (the squared molecular contributions), mixing rule 1 (linear combination of molecular contributions), and mixing rule 3 (norm of molecular contributions) were selected as the best mixing rules for QSAR-1, QSAR-2, and QSAR-3 respectively. The developed QSAR models have been represented in the following equation 4.10, 4.11, and 4.12.

QSAR-1

$$pEC50\ (Mol/L) = 6.49686 - 0.01966 \times ALOGP2 - 0.00004 \times SAacc + 0.53033 \times nCL - 0.31642 \times C - 002 - 0.55166 \times C - 032 - 1.09605 \times nR05 \quad (4.10)$$

QSAR-2

$$pEC50\ (Mol/L) = 8.43353 - 0.75722 \times F02[N - O] - 0.28182 \times F04[C - N] + 0.04325 \times MLOGP2 + 0.17719 \times C - 028 + 0.1223 \times C - 029 - 0.22751 \times B03[N - N] \quad (4.11)$$

QSAR-3

$$pEC50\ (Mol/L) = 9.36117 - 2.93858 \times Eta_{betaP_A} - 0.99832 \times B05[C - S] - 0.46666 \times NsCH3 - 8.3358 \times minaasC + 0.06123 \times T(N..Cl) \quad (4.12)$$

All these models were thoroughly validated and proved to be acceptable in terms of robustness and predictability. The computed validation metrics of these three models have been reported in **Table 4.23**. We also performed the Y-randomization test for the developed QSAR models; the Y-randomization plots have been given in **Figure 4.40**. The intercept values of QSAR-1, QSAR-2, and QSAR-3 respectively are R^2: 0.0192, 0.0315, 0.0307, and

Q^2: -0.628, -0.651, -0.589 which indicate that the models are not developed by any chance correlation. We have checked the applicability domain of these developed QSARs (PLS) by DModX approach in 99% confidence level (Gadaleta et al., 2016) using SIMCA-P software (Wu et al., 2010) and represented the resultant plots in **Figure 4.41**. All the mixtures (both training and test set) of these three datasets are falling within the applicability domain (represented by the D-critical line) of corresponding training sets.

Table 4.23. Computed validation metrics of the developed PLS-QSAR models

Validation metrics	**QSAR-1 (Organism: *V. fischeri*)**	**QSAR-2 (Organism: *E. coli*)**	**QSAR-3 (Organism: *B. subtilis*)**
n_{Train}	57	51	32
n_{Test}	19	16	10
No. of descriptors (No. of LVs)	6 (5 LVs)	6 (5 LVs)	5 (4 LVs)
	Internal validation metrics		
R^2	0.942	0.937	0.894
R^2_{adj}	0.935	0.928	0.873
Q^2_{LOO}	0.925	0.922	0.851
MAE_{Train}	0.097	0.089	0.123
F value(*df1,df2)	167.856 (5, 51)	133.590 (5, 45)	56.920 (4, 27)
$\overline{r^2_{m(Train)}}$	0.895	0.891	0.791
$\Delta r^2_{m(Train)}$	0.027	0.050	0.089
AIC	-71.238	-82.315	-41.726
	External validation metrics		
Q^2_{F1}	0.944	0.919	0.841
Q^2_{F2}	0.944	0.915	0.745
CCC	0.969	0.959	0.845
$\overline{MAE_{Test}}$	0.112	0.098	0.196
$\overline{r^2_{m(Test)}}$	0.899	0.868	0.569
$\Delta r^2_{m(Test)}$	0.037	0.062	0.180

*df = degree of freedom

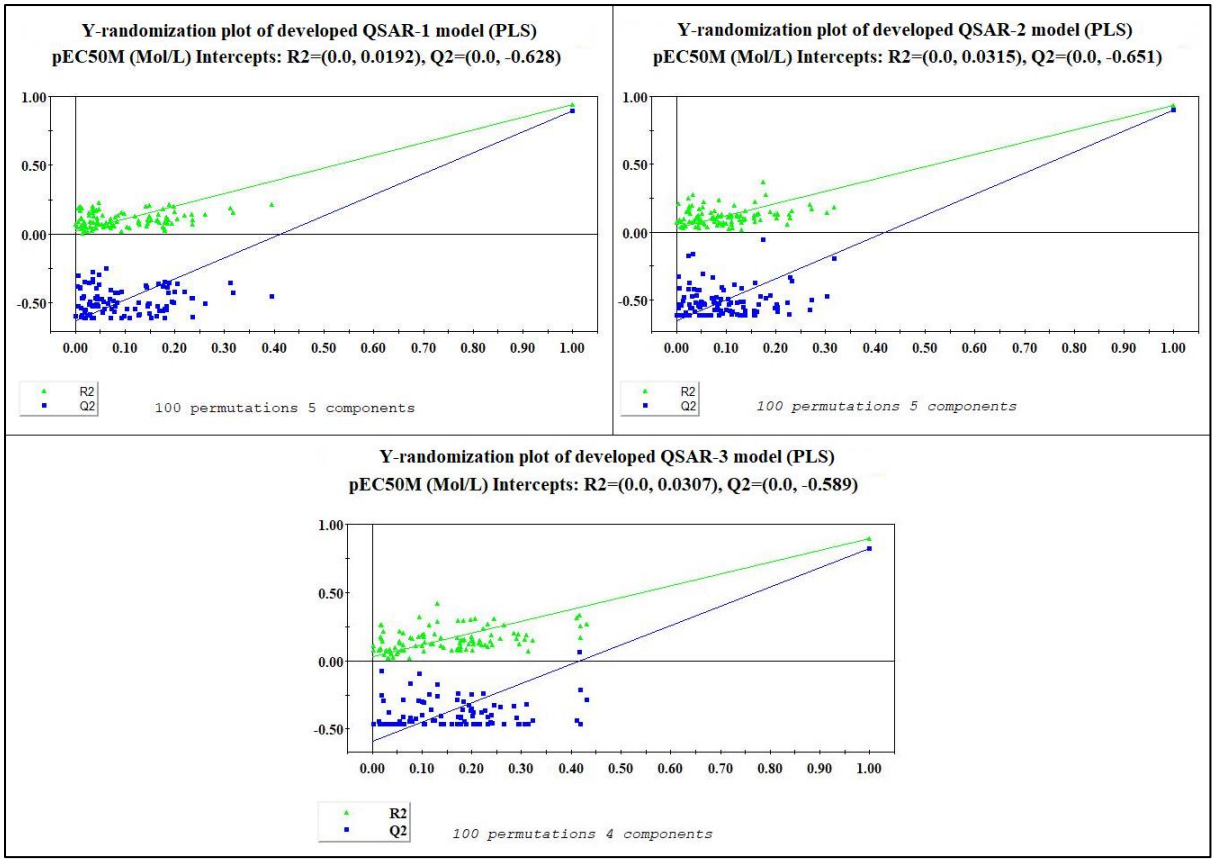


Figure 4.40. Y-randomization plots of developed QSAR-1, QSAR-2, and QSAR-3 models

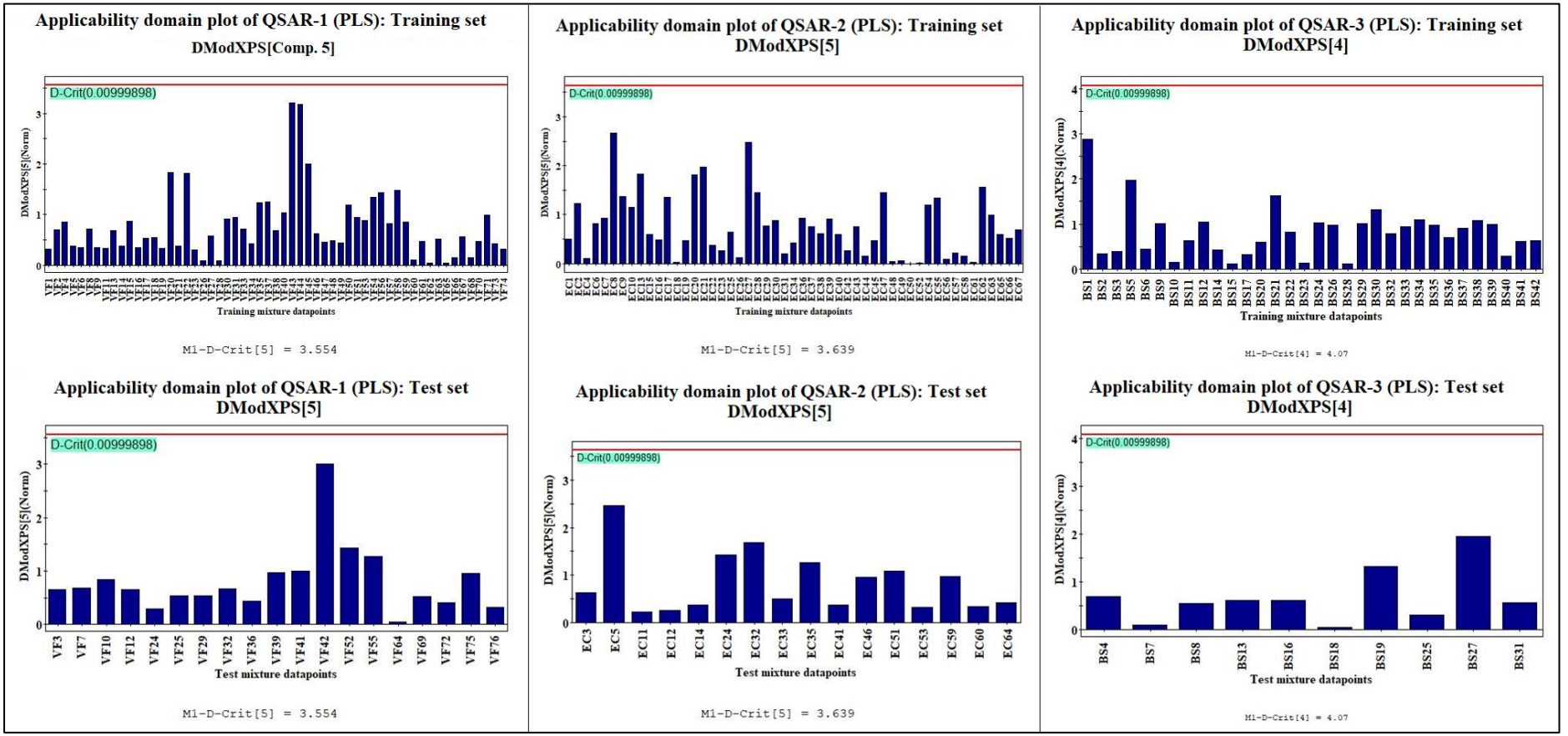


Figure 4.41. Applicability domain plots (DModX: for training and test set) of developed QSAR-1, QSAR-2, and QSAR-3

To check the impact of environmental degradation of antibiotics on the mixture toxicity predictions, three different QSAR models were developed incorporating the pKa (strong acidic and strong basic) values with the previously identified features in QSAR-1, QSAR-2, and QSAR-3. The new models along with the computed validation metrics have been tabulated in **Table 4.24** (Equations 4.13, 4.14, and 4.15). The computed validation metrics of all these three QSAR models are in acceptable range as per OECD criteria. To check the models' performance, we computed Akaike information criteria (AIC) for both QSARs and QSARs with pKa descriptors. It has been observed in each of the three

datasets that AICs of QSARs (AIC: QSAR-1 = -71.238, QSAR-2 = -82.315, QSAR-3 -41.726) are less than the AICs of QSARs with pKa (AIC: MD1 = -65.995, MD2 = -76.827, MD3 = -38.388) which indicates the better relative efficiency of QSARs in terms of performance and descriptor combination. Therefore, it infers the less importance of acidic dissociation constant values in the predictive modeling of mixture toxicity of antibiotics.

Table 4.24. Developed QSAR (PLS) models incorporating pKa (strong acidic and strong basic) as descriptors, and their computed validation metrics

Dataset	Model and computed validation metrics
MD1 *(Organism: V. fischeri)*	$pEC50 = 5.98651 - 0.02408 \times ALOGP2 - 0.00003 \times SAacc + 0.78277 \times nCL - 0.4019 \times C - 002 - 0.65074 \times C - 032 - 0.86442 \times nR05 + 0.00346 \times pKa\ (Strong\ acidic) - 0.00342 \times pKa(Strong\ basic)$ (4.13) *nTrain = 57, LVs = 7,* R^2 *= 0.944,* Q^2_{LOO} *= 0.920,* MAE_{Train} *= 0.010,* $\overline{r^2_{m(Train)}} = 0.888$, $\Delta r^2_{m(Train)}$ *=0.030,* Q^2_{F1} *= 0.959,* Q^2_{F2} *= 0.959 , CCC = 0.978,* MAE_{Test} *= 0.100,* $\overline{r^2_{m(Test)}} = 0.921$, $\Delta r^2_{m(Test)}$*=0.029, AIC = -65.995*
MD2 *(Organism: E. coli)*	$pEC50 = 7.73854 - 0.53759 \times F02[N-0] - 0.28465 \times F04[C-N] + 0.05605 \times MLOGP2 + 0.16852 \times C - 028 + 0.04731 \times C - 029 - 0.37224 \times B03[N-N] + 0.02558 \times pKa\ (Strong\ acidic) + 0.04344 \times pKa\ (Strong\ basic)$ (4.14) *nTrain = 51, LVs = 6,* R^2 *= 0.932,* Q^2_{LOO} *= 0.909,* MAE_{Train} *= 0.085,* $\overline{r^2_{m(Train)}} = 0.873$, $\Delta r^2_{m(Train)}$ *=0.050,* Q^2_{F1} *= 0.917,* Q^2_{F2} *= 0.913 , CCC = 0.959,* MAE_{Test} *= 0.096,* $\overline{r^2_{m(Test)}} = 0.862$, $\Delta r^2_{m(Test)}$*=0.064, AIC = -76.827*
MD3 *(Organism: B. subtilis)*	$pEC50 = 9.52681 - 3.44252 \times Eta_{betaP_A} - 0.8787 \times B05[C-S] - 0.39533 \times NsCH3 - 8.01893 \times minaasC + 0.06365 \times T(N..Cl) + 0.01021 \times pKa(strong\ acidic) - 0.05056 \times pKa(strong\ basic)$ (4.15) *nTrain = 32, LVs = 6,* R^2 *= 0.899,* Q^2_{LOO} *= 0.8220,* MAE_{Train} *= 0.113,* $\overline{r^2_{m(Train)}} = 0.755$, $\Delta r^2_{m(Train)}$ *=0.087,* Q^2_{F1} *= 0.858,* Q^2_{F2} *= 0.774 , CCC = 0.861,* MAE_{Test} *= 0.181,* $\overline{r^2_{m(Test)}} = 0.601$, $\Delta r^2_{m(Test)}$*=0.164, AIC = -38.388*

As we discussed in the "***3.5.4. Multilayered descriptor selection and QSAR/QSAAR model development***" section, QSAAR models were also developed in this study to find whether there is any intercorrelation among interspecies toxicity or not. First, 6 QSAAR models were developed using PLS regression and then, the RASAR algorithm was used to develop three final RASAR-based QSAAR (q-RASAAR) models from 3 datasets. The preliminary PLS-QSAAR models have been tabulated along with their validation metrics in **Table 4.25**. All these models are acceptable in terms of goodness of fit, robustness, and predictability. The final RASAR-based QSAARs (i.e., q-RASAARs) were developed using both MLR (q-RASAAR 1: equation 4.22 and q-RASAAR-3: equation 4.24) and PLS regression (q-RASAAR-2: equation 4.23) with an optimum number of LVs and were validated thoroughly. The three models along with computed validation metrics have been reported in the following **Table 4.26**. The acceptable values of computed validation metrics indicate the statistical soundness and predictability of these three q-RASAAR models. One point should be noted here that both q-RASAAR-1 and 3 are univariate MLR models with excellent statistical quality especially in terms of external predictability. We have plotted the observed vs. predicted toxicity scatter plots of both QSAR and q-RASAAR models and showcased those in the following **Figure 4.42**. A uniform scattering surrounding the trend line has been observed in all the plots.

Table 4.25. Developed PLS-QSAAR models and their computed validation metrics

Dataset	Model and computed validation metrics
MD1 *(Organism: V. fischeri)*	**QSAAR-1** (Predictor sp.: ***E. coli***) $pEC50 = 5.8887 - 0.55941 \times C - 032 + 0.59119 \times nCL - 0.09764 \times nHDon + 0.00482 \times minssO - 0.14749 \times F03[C - S] + 0.1505 \times Predicted\ toxicity\ (E.coli)$ (4.16) *nTrain = 57, LVs = 5, R^2 = 0.929, Q^2_{LOO} = 0.910, MAE_{Train} = 0.111,* $\overline{r^2_{m(Train)}} = 0.874$, $\Delta r^2_{m(Train)}$ *=0.057, Q^2_{F1} = 0.938, Q^2_{F2} = 0.938 , CCC = 0.965, MAE_{Test} = 0.121,* $\overline{r^2_{m(Test)}} = 0.886$, $\Delta r^2_{m(Test)}$*=0.040*
	QSAAR-2 (Predictor sp.: ***B. subtilis***) $pEC50 = 5.64411 - 0.23374 \times C - 032 + 0.05687 \times F04[C - Cl] + 0.12995 \times C - 029 - 0.01621 \times ALOGP2 - 0.00004 \times SAacc + 0.04288 \times Predicted\ toxicity\ (B.subtilis)$ (4.17) *nTrain = 57, LVs = 5, R^2 = 0.928, Q^2_{LOO} = 0.904, MAE_{Train} = 0.115,* $\overline{r^2_{m(Train)}} = 0.873$, $\Delta r^2_{m(Train)}$ *=0.041, Q^2_{F1} = 0.942, Q^2_{F2} = 0.942 , CCC = 0.967, MAE_{Test} = 0.124,* $\overline{r^2_{m(Test)}} =$

	0.873, $\Delta r^2_{m(Test)}$=*0.041*
MD2 *(Organism: E. coli)*	**QSAAR-3** (Predictor sp.: ***V. fischeri***) $pEC50 = 7.73918 + 0.22477 \times B06[O-O] - 0.80761 \times F02[N-O] - 0.25623 \times F04[C-N] + 0.07095 \times ALOGP2 + 0.11987 \times Predicted\ toxicity\ (V.fischeri)$ (4.18) *nTrain = 51, LVs = 4, R^2 = 0.936, Q^2_{LOO} = 0.916, MAE_{Train} = 0.084,* $\overline{r^2_{m(Train)}} = 0.883$, $\Delta r^2_{m(Train)}$ *=0.046, Q^2_{F1} = 0.929, Q^2_{F2} = 0.925 , CCC = 0.964, MAE_{Test} = 0.088,* $\overline{r^2_{m(Test)}} = 0.877$, $\Delta r^2_{m(Test)}$*=0.056*
	QSAAR-4 (Predictor sp.: ***B. subtilis***) $pEC50 = 8.2011 - 0.66979 \times F02[N-O] - 0.2391 \times F04[C-N] - 0.22401 \times B05[N-O] + 0.06773 \times F06[C-N] - 0.01508 \times Predicted\ toxicity\ (B.subtilis)$ (4.19) *nTrain = 51, LVs = 4, R^2 = 0.932, Q^2_{LOO} = 0.909, MAE_{Train} = 0.095,* $\overline{r^2_{m(Train)}} = 0.873$, $\Delta r^2_{m(Train)}$ *=0.055, Q^2_{F1} = 0.936, Q^2_{F2} = 0.932 , CCC = 0.966, MAE_{Test} = 0.081,* $\overline{r^2_{m(Test)}} = 0.905$, $\Delta r^2_{m(Test)}$*=0.052*
MD3 *(Organism: B. subtilis)*	**QSAAR-5** (Predictor sp.: ***V. fischeri***) $pEC50 = 4.5697 - 0.27102 \times F02[C-N] + 0.24567 \times nHM - 0.3221 \times nAA + 0.40031 \times F04[N-N] + 0.81601 \times Predicted\ toxicity\ (V.fischeri)$ (4.20) *nTrain = 32, LVs = 3, R^2 = 0.886, Q^2_{LOO} = 0.844, MAE_{Train} = 0.120,* $\overline{r^2_{m(Train)}} = 0.783$, $\Delta r^2_{m(Train)}$ *=0.105, Q^2_{F1} = 0.853, Q^2_{F2} = 0.765 , CCC = 0.877, MAE_{Test} = 0.198,* $\overline{r^2_{m(Test)}} = 0.699$, $\Delta r^2_{m(Test)}$*=0.135*
	QSAAR-6 (Predictor sp.: ***E. coli***) $pEC50 = 5.51681 + 0.02686 \times T(N..Cl) - 0.1787 \times ALOGP2 - 0.87472 \times nOHt - 3.24308 \times minaaCH + 0.63988 \times Predicted\ toxicity\ (E.coli)$ (4.21)

	nTrain = 32, LVs = 4, R^2 = 0.882, Q^2_{LOO} = 0.820, MAE_{Train} = 0.127, $\overline{r^2_{m(Train)}} = 0.750$, $\Delta r^2_{m(Train)}$ *=0.109, Q^2_{F1} = 0.775, Q^2_{F2} = 0.641 , CCC = 0.804, MAE_{Test} = 0.253,* $\overline{r^2_{m(Test)}} = 0.565$, $\Delta r^2_{m(Test)}$*=0.164*

Table 4.26. The developed q-RASAAR models and the computed validation metrics

Models	Equation	Computed validation metrics
q-RASAAR-1 (Univariate MLR)	$pEC50 = -0.02009 + 1.00347 \times RAfunction\ (LK)$ (4.22)	*nTrain = 57, R^2 = 0.929, Q^2_{LOO} = 0.925, MAE_{Train} = 0.110, F value = 722.68 (df: 1, 55),* $\overline{r^2_{m(Train)}} = 0.894$, $\Delta r^2_{m(Train)}$ *=0.058, Q^2_{F1} = 0.975, Q^2_{F2} = 0.975 , CCC = 0.988, MAE_{Test} = 0.080,* $\overline{r^2_{m(Test)}} = 0.963$, $\Delta r^2_{m(Test)}$*=0.019, AIC =-94.858*
q-RASAAR-2 (PLS with 3 LVs)	$pEC50 = 8.62151 - 0.77254 \times FO2[N-O] - 0.26976 \times F04[C-N] - 0.1759 \times B05[N-O] + 0.04919 \times Neg.Avg.Sim$ (4.23)	*nTrain = 51, R^2 = 0.925, Q^2_{LOO} = 0.906, MAE_{Train} = 0.091, F value = 192.55 (df: 3, 47f),* $\overline{r^2_{m(Train)}} = 0.869$, $\Delta r^2_{m(Train)}$ *=0.056, Q^2_{F1} = 0.944, Q^2_{F2} = 0.944, CCC = 0.970, MAE_{Test} = 0.080,* $\overline{r^2_{m(Test)}} = 0.918$, $\Delta r^2_{m(Test)}$*=0.010, AIC = -82.424*
q-RASAAR-3 (Univariate MLR)	$pEC50 = 0.19879 + 0.96379 \times RAfunction(LK)$ (4.24)	*nTrain = 32, R^2 = 0.847, Q^2_{LOO} = 0.830, MAE_{Train} = 0.129, F value = 165.49 (df: 1, 30),* $\overline{r^2_{m(Train)}} = 0.763$, $\Delta r^2_{m(Train)}$ *=0.118, Q^2_{F1} = 0.965, Q^2_{F2} = 0.944, CCC = 0.969, MAE_{Test} = 0.093,* $\overline{r^2_{m(Test)}} = 0.836$, $\Delta r^2_{m(Test)}$*=0.042, AIC = -44.586*

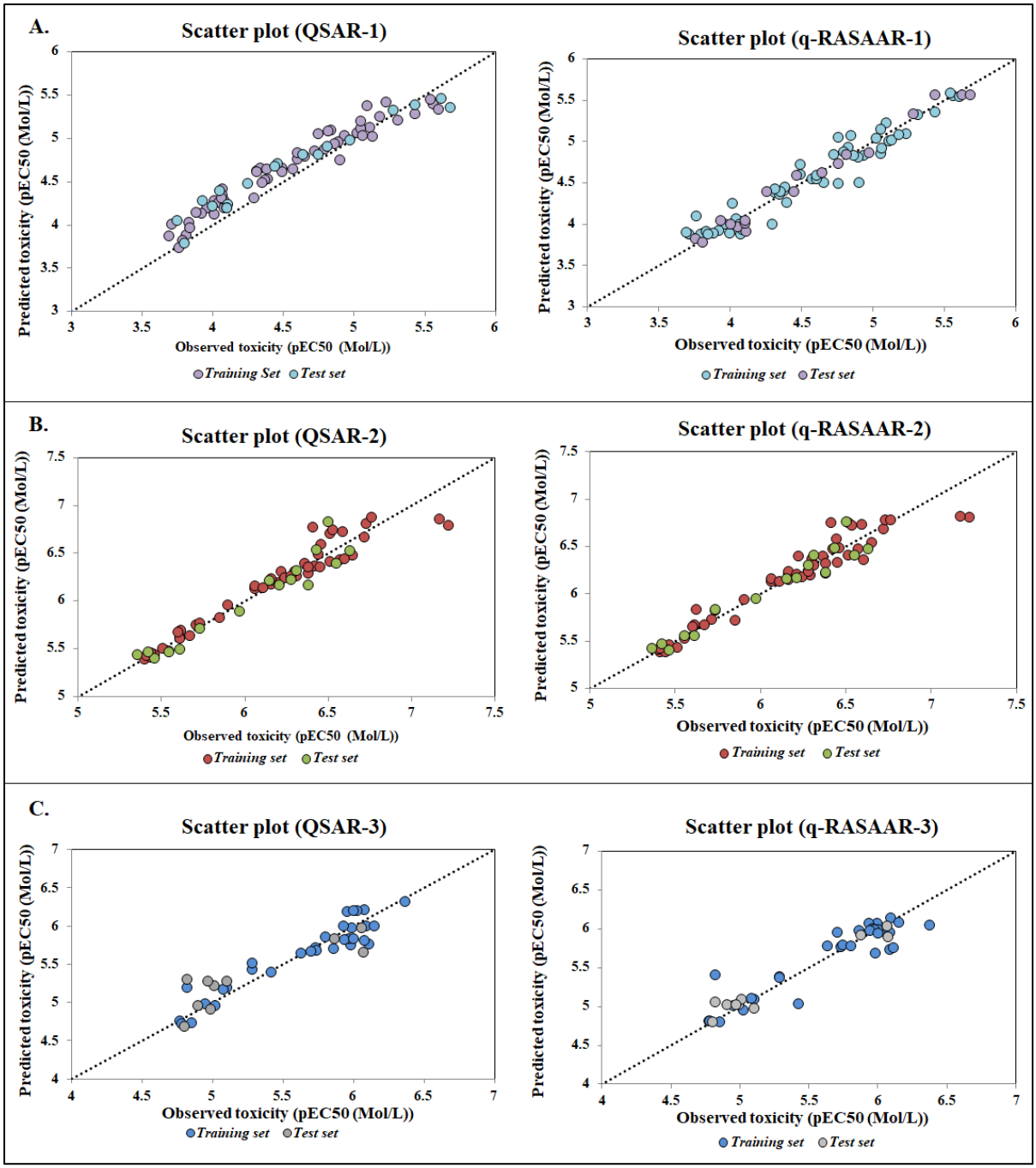


Figure 4.42. Observed vs. predicted toxicity scatter plots; A. Scatter plots of QSAR-1 (left) and q-RASAAR-1 (right) models developed from MD1 dataset against toxicity of *V. fischeri*, B. Scatter plots of QSAR-2 (left) and q-RASAAR-2 (right) models developed from MD2 dataset against toxicity of *E. coli*, C. Scatter plots of QSAR-3 (left) and q-RASAAR-3 (right) models developed from MD3 dataset against toxicity of *B. subtilis*

To comply with the OECD principle 3 (a defined domain of applicability), we have analyzed the applicability domain of the final q-RASAAR models. The applicability domain is the theoretical region in descriptor space defined by the structural/biological/physicochemical

features of training set compounds in which the predictions obtained by the developed model is reliable (Gadaleta et al., 2016). In this study, we have used the William's plot to define the domain of applicability of q-RASAAR models. The William's plots are obtained by plotting the cross-validated standardized residual in Y-axis and computed leverages (hat value) in X-axis of all training and test set compounds (**Figure 4.43**). The leverage values have been computed using "Hi-Calc" version 2.0 (available from http://teqip.jdvu.ac.in/QSAR_Tools/). The applicability domain is represented by the leverage threshold h* line (h* = 3(p'+1)/n: p'=number of predictor variable, n = number of training compounds) drawn in the William's plot (Gramatica et al., 2013). There is no outlier noted in q-RASAAR-1 and q-RASAAR-3 models. In q-RASAAR-2, test mixture **EC 5** (binary mixture of diaveridine and sulfamonomethoxine) has been noted outside the AD (**Figure 4.43**).

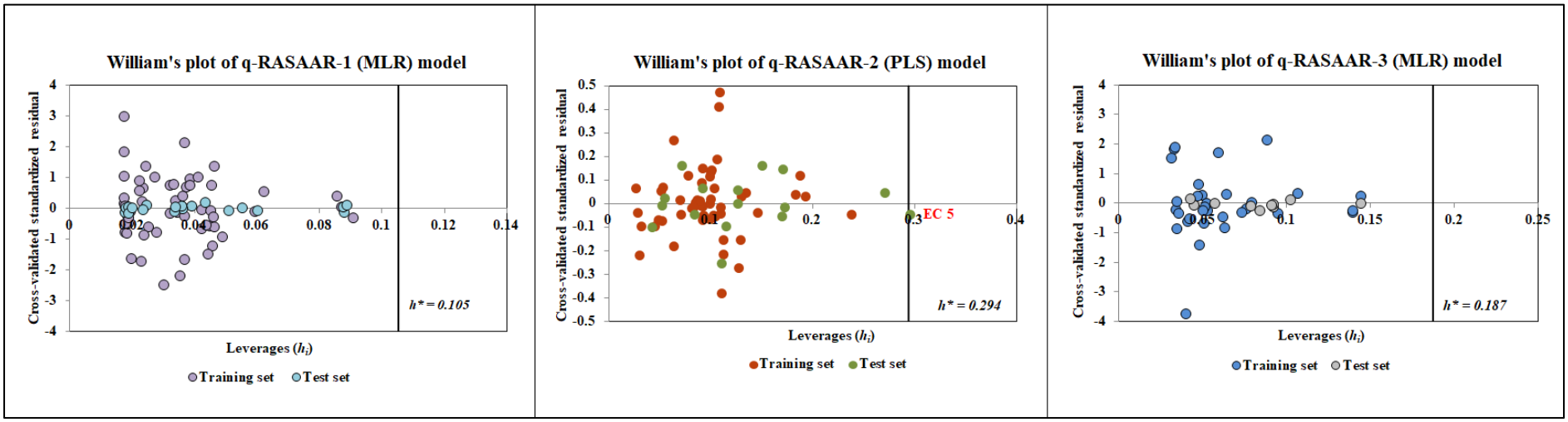


Figure 4.43. William's plots of q-RASAAR-1, q-RASAAR-2 (Test mixture EC 5: outside AD), and q-RASAAR-3 models

4.5.2. Insights of modeled descriptors identified in the QSAR analysis

The important descriptors for three QSAR models were identified by multilayered feature selection strategy as discussed in the "Multilayered descriptor selection and QSAR/QSAAR model development" section. The identified descriptors have been tabulated along with their class and definition in **Table 4.27**. For further analysis of modeled descriptors, variable importance plots (VIP) and coefficient plots of the developed PLS-QSAR models were obtained using "SIMCA-P" software (Wu et al., 2010) and represented as bubble plots in **Figure 4.44.** The VIP scores of modeled descriptors have been represented as bubble radius whereas, the type of contributions (positive/negative) towards toxicity have been indicated by the colour of bubbles (green bubble = positive contribution; red bubble = negative contribution). The VIP score is the indicator of the relative importance of descriptors; in standard practice, descriptors with VIP scores > 1 show significant importance in toxicity prediction.

Table 4.27. Definition and class of identified descriptors in QSAR-1, QSAR-2, and QSAR-3

QSAR-1		
Descriptor	**Class**	**Definition**
SAacc	Molecular properties	Surface area of acceptor atoms from P_VSA-like descriptors
C-032	Atom centred fragment	Presence of X–CX–X
nR05	Ring descriptors	Number of 5-membered rings
C-002	Atom centred fragment	Presence of CH2R2
ALOGP2	Molecular properties	Squared Ghose-Crippen octanol-water partition coeff. (logP^2)
nCL	Constitutional indices	Number of Chlorine atoms
QSAR-2		
Descriptor	**Class**	**Definition**
F02[N-O]	2D atom pairs	Number of occurrences of N-O atom pair at topological distance 2
F04[C-N]	2D atom pairs	Number of occurrences of C-N atom pair at topological distance 4
C-028	Atom centred fragment	Presence of R–CR–X
MLOGP2	Molecular properties	Squared Moriguchi octanol-water partition coeff. (logP^2)
B03[N-N]	2D atom pairs	Presence/absence of N-N atom pair at topological distance 3
C-029	Atom centred fragment	Presence of R–CX–X

QSAR-3		
Descriptor	**Class**	**Definition**
MinaasC	Atom-type E-state indices	Mimimum aasC
NsCH3	Atom-type E-state indices	Number of atoms of type sCH3
T(N..Cl)	2D atom pairs	Sum of topological distances between N..Cl
B05[C-S]	2D atom pairs	Presence/absence of C_S atom pair at topological distance 5
Eta_betaP_A	Electro-topochemical atom (ETA) indices	Eta pi and lone pair average VEM count

4.5.2.1. Features responsible for toxicity against V. fischeri identified by the QSAR-1 model

The descriptors of interest in QSAR-1 are SAacc, C-032, nR05, C-002, ALOGP2, and nCl. Among these 6 descriptors, all except nCl (positive contribution) have negative contributions towards cellular toxicity against *V. fischeri*. According to the VIP represented in the "Bubble plot" of **Figure 4.44A**, the order of relative importance of modeled descriptors is SAacc > C-032 > nR05 > C-002 > ALOGP2 > nCl. We have also depicted the fragment/s responsible for the non-zero value of descriptors (QSAR-1) in **Figure 4.44A**. To understand the relationship between individual descriptors and toxicity, we have systematically compared 3 pairs of mixtures in **Table 4.28**. The SAacc is the most important descriptor of QSAR-1 and is categorised as a molecular property descriptor which represents the surface area of H-bond acceptor atoms. It has shown a negative contribution towards toxicity; thus, higher availability of electron acceptor atoms decreases the mixture toxicity against *V. fischeri*. If we compare the extreme points of the datasets, then it can be observed that the most toxic mixture has a lower SAacc value than the least toxic mixtures due to the presence of components with more H-bond acceptor atoms (oxytetracycline present in **VF 34** as shown in **Figure 4.44A**). C-032 is another important descriptor of QSAR-1, which reduces the toxic response. The presence of X-CX-X fragments (X= electronegative atoms; i.e., O, N, S, Halogens, etc.) in the 2D structure of mixture components are identified by the C-032 descriptor. The presence of X-CX-X fragment has been marked in a component (ormetoprim) of a comparatively less toxic mixture **VF 67** in **Figure 4.44A**. nR05 is the next important descriptor of QSAR-1, which gives the number of 5-membered rings in a molecular structure. The negative coefficient of nR05 signifies an inversely proportional relationship with toxicity against *V. fischeri*. The presence of 5-membered ring has been observed in the mixture component of the least toxic mixture **VF 34** (sulfamethoxazole is a 5-membered ring-containing component; see **Figure 4.44A**). The C-002 is the 4th important descriptor of QSAR-1, which signifies the presence/absence of CH2R2 fragment (R is any group connected with carbon) in a molecular structure (this feature has been marked on the structural representation of pyrimethamine: component of a comparatively less toxic mixture **VF 45**; see **Figure 4.44A**). It has a negative contribution towards the toxicity endpoint which can be justified by comparing the mixtures presented in **Table 4.28**. The next important structural feature of QSAR-1 is the molecular property descriptor ALOGP2 which computes the value of the squared Ghose-Crippen octanol-water partition coefficient (log P^2) of chemical substances. It also has a negative contribution towards toxicity against *V. fischeri*. A

systematic decrease in toxicity has been observed with an increase in the ALOGP2 values (see **Table 4.28**), which supports the negative contribution of this descriptor. The only positive contributing descriptor of QSAR-1 is nCl which denotes the number of chlorine atoms present in a molecule. Among the mixture components in this study, pyrimethamine and chlortetracycline contain Cl in their structures. Therefore, Cl-containing mixtures like **VF 70**, **VF 51**, **VF 31** etc. have comparatively higher nCl values and toxicity than the mixtures without Cl atoms (e.g., **VF 34**: least toxic mixture; see **Table 4.28**).

Table 4.28. Relationship of descriptors with toxicity against *V. fischeri* (represented by comparative assessment between different mixtures of MD1)

SAacc (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	16499.50	5.60
	VF 5 (Trimethoprim + sulfamethazine)	16967.28	5.02
2	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	16499.50	5.60
	VF 44 (Tetracycline + ormetoprim)	39337.42	4.08
3	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	16499.50	5.60
	VF 34 (Oxytetracycline + sulfamethoxazole)	52310.85	3.69
C-032 (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 17** (Pyrimethamine + sulfamonomethoxine)	0.26	5.23
	VF 21 (Pyrimethamine + sulfisoxazole)	0.27	5.05
2	**VF 17** (Pyrimethamine + sulfamonomethoxine)	0.26	5.23
	VF 61 (Doxycycline + diaveridine + sulfameter)	0.51	4.90
3	**VF 17** (Pyrimethamine + sulfamonomethoxine)	0.26	5.23
	VF 67 (Oxytetracycline + ormetoprim + sulfamethazine)	0.56	4.35
nR05 (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.00	5.60
	VF 15 (Ormetoprim + sulfisoxazole)	0.17	5.31
2	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.00	5.60

	VF 27 (Tetracycline + sulfisoxazole)	0.29	3.99
3	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.00	5.60
	VF 34 (Oxytetracycline + sulfamethoxazole)	0.32	3.69

C-002 (Negative contribution)

Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.33	5.60
	VF 73 (Chlortetracycline + pyrimethamine + sulfamethoxypyridazine)	0.44	4.89
2	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.33	5.60
	VF 59 (Tetracycline + trimethoprim + sulfameter)	0.50	4.49
3	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	0.33	5.60
	VF 45 (Tetracycline + pyrimethamine)	1.00	4.01

ALOGP2 (Negative contribution)

Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	11.17	5.60
	VF 17 (Pyrimethamine + sulfamonomethoxine)	22.15	5.23
2	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	11.17	5.60
	VF 45 (Tetracycline + pyrimethamine)	24.36	4.01
3	**VF 13** (Ormetoprim + Sulfamethoxypyridazine)	11.17	5.60
	VF 50 (Oxytetracycline + pyrimethamine)	42.55	3.76

nCl (Positive contribution)

Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**VF 70** (Chlortetracycline + pyrimethamine + sulfadoxine)	0.49	4.79
	VF 34 (Oxytetracycline + sulfamethoxazole)	0.00	3.69
2	**VF 51** (Doxycycline + pyrimethamine)	0.28	4.29
	VF 34 (Oxytetracycline + sulfamethoxazole)	0.00	3.69
3	**VF 31** (Chlortetracycline + sulfameter)	0.25	4.04

VF 34 (Oxytetracycline + sulfamethoxazole)	0.00	3.69

4.5.2.2. Features responsible for toxicity against E. coli identified by the QSAR-2 model

Just like the previous QSAR-1, the important structural features of MD2 were identified by multilayered descriptor selection and were modeled by PLS regression. Thus, the descriptors of interest in QSAR-2 are F02[N-O], F02[C-N], MLOGP2, C-028, C-029, B03[N-N]. An equal distribution between positive (C-028, MLOGP2, and C-029) and negative (F02[N-O], F04[C-N], and B03[N-N]) contributing descriptors has been observed in QSAR-2. The order of relative importance of the modeled descriptors is F02[N-O] > F02[C-N] > C-028 > MLOGP2 > B03[N-N] > C-029, according to the VIP represented in the "Bubble plot" of **Figure 4.44B**. Among these, F02[N-O] and F04[C-N] have VIP scores greater than 1; therefore, their relative importance is more significant. Both of these negatively contributing descriptors are from the class of 2D atom pairs, and they compute the frequency of atom pairs N-O (former) and C-N (latter) at topological distances 2 and 4 respectively. We have systematically compared 3 pairs of mixtures for each descriptor (in **Table 4.29**) to understand the relationship between structural features and toxicity. F02[N-0] is the most important descriptor of QSAR-2 which inversely impacts the toxicity against *E. coli*. We have depicted the N-O atom pair at topological distance 2 within the structure of the mixture component chlortetracycline in **Figure 4.44B**. It has been observed that the mixtures with greater F02[N-O] values are less toxic (e.g., **EC 18**, **EC 48**, **EC 29**) than the mixture with less value (**EC 36**: most toxic mixture of MD2) which justifies the inversely proportional relationship between F02[N-O] and toxicity. A similar observation has been found in the case of descriptor F04[C-N] also. The C-N atom pair at topological distance 4 has been depicted in the structural representation of sulfadoxine (see **Figure 4.44B**) which is a component of the least toxic mixture of MD2 (**EC 18**: chlortetracycline + sulfadoxine). In **Table 4.29**, we have seen a systematic decrease in mixture toxicity with an increase in F04[C-N] values which agrees with the negative contribution of this descriptor. The next important descriptor is C-028 which belongs to the class 2D atom pair and indicates the presence of R-CR-X fragment in the molecular structure (R = any group attached via C atom; X = electronegative atoms i.e., O, N, S, Halogens, etc.). It has a positive contribution towards the toxic endpoint; therefore, it increases the mixture toxicity against *E. coli*. We have shown the presence of R-CR-X fragment in the structural representation of sulfamethazine (see **Figure 4.44B**) which is present in different toxic mixtures as a component (e.g., **EC 10**, **EC 57**). Three pairs of mixtures have been tabulated along with their toxicity and descriptor values in **Table 4.29** for a systematic comparison among mixtures. Increased toxicity has been observed with an increase in the C-028 value in mixtures which signifies a positive contribution. MLOGP2 is a molecular property and the next positive contributing descriptor of QSAR-2. It computes the squared Moriguchi octanol-water partition coeff. (log P^2) which is a measure of the lipophilicity of the mixture. Its positive contribution shows the impact of lipophilicity towards mixture toxicity against *E. coli*. A distinct increase of both MLOGP2 and toxicity has been observed from the least toxic mixture (**EC 18**) to the most toxic mixture (**EC 36**) which aligns with their proportional relationship. B03[N-N] is again a negative contributing descriptor which indicates the presence or absence of N-N atom pair at topological distance 3. The most toxic mixture (**EC 36**: pEC50 = 7.22) does not contain components with N-N atom pair at topological distance 3; however, we have seen the presence of N-N atom (3 topological distance apart) pairs in various less toxic mixtures (**EC 66**, **EC 34**, **EC 13**; see **Table 4.29**). We have depicted the B03[N-N] fragment in the mixture component (sulfamethoxypyridazine) of one of the least toxic mixtures (**EC 13**) in **Figure 4.44B**. The

final descriptor of QSAR-2 is the C-029 which positively contributes toward the toxicity against E. coli. It indicates the presence/absence of the R-CX-X fragment (R = any group attached by a C atom; X = electronegative atoms e.g., O, N, S, Halogens, etc.) in the molecule. The presence of the R-CX-X fragment has been observed in more toxic mixtures like **EC 6** (pEC50 = 6.72) and has been depicted in the molecular structure of sulfamethoxazole (a component of **EC 6**) in **Figure 4.44B**. A marked decrease in toxicity and C-029 values has been observed in the mixtures tabulated in **Table 4.29** which justifies the positive contribution of C-029 towards toxicity against *E. coli*.

Table 4.29. Relationship of descriptors with toxicity against *E. coli* (represented by comparative assessment between different mixtures of MD2)

F02 [N-O] (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 36** (Tetracycline + trimethoprim)	0.49	7.22
	EC 48 (Tetracycline + trimethoprim + Sulfamethoxazole)	0.92	6.65
2	**EC 36** (Tetracycline + trimethoprim)	0.49	7.22
	EC 29 (Oxytetracycline + sulfamethoxazole)	1.45	5.90
3	**EC 36** (Tetracycline + trimethoprim)	0.49	7.22
	EC 18 (Chlortetracycline + sulfadoxine)	1.82	5.40
FO4 [C-N] (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 36** (Tetracycline + trimethoprim)	4.99	7.22
	EC 39 (Chlortetracycline + diaveridine)	5.09	6.59
2	**EC 36** (Tetracycline + trimethoprim)	4.99	7.22
	EC 44 (Tetracycline + trimethoprim + sulfadoxine)	5.52	6.16
3	**EC 36** (Tetracycline + trimethoprim)	4.99	7.22
	EC 18 (Chlortetracycline + sulfadoxine)	6.41	5.40
C-028 (Positive contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 10** (Ormetoprim + sulfamethazine)	0.87	6.29
	EC 18 (Chlortetracycline + sulfadoxine)	0.00	5.40
2	**EC 57** (Doxycycline + diaveridine + sulfamethazine)	0.54	6.11
	EC 18 (Chlortetracycline + sulfadoxine)	0.00	5.40
3	**EC 30** (Oxytetracycline + sulfisoxazole)	0.43	5.60

	EC 18 (Chlortetracycline + sulfadoxine)	0.00	5.40
MLOGP2 (Positive contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 36** (Tetracycline + trimethoprim)	1.75	7.22
	EC 48 (Tetracycline + trimethoprim + Sulfamethoxazole)	1.53	6.65
2	**EC 36** (Tetracycline + trimethoprim)	1.75	7.22
	EC 20 (Chlortetracycline + sulfamonomethoxine)	1.43	5.71
3	**EC 36** (Tetracycline + trimethoprim)	1.75	7.22
	EC 18 (Chlortetracycline + sulfadoxine)	0.92	5.40
B03[N-N] (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 36** (Tetracycline + trimethoprim)	0.00	7.22
	EC 66 (Oxytetracycline + ormetoprim + sulfisoxazole)	0.28	6.38
2	**EC 36** (Tetracycline + trimethoprim)	0.00	7.22
	EC 34 (Doxycycline + sulfisoxazole)	0.41	5.67
3	**EC 36** (Tetracycline + trimethoprim)	0.00	7.22
	EC 13 (Tetracycline + sulfamethoxypyridazine)	0.42	5.45
C-029 (Positive contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**EC 6** (Diaveridine + sulfamethoxazole)	1.00	6.72
	EC **54** (Doxyxline + diaveridine + sulfamonomethoxine)	0.90	6.45
2	**EC 6** (Diaveridine + sulfamethoxazole)	1.00	6.72
	EC 65 (Oxytetracycline + ormetoprim + sulfamethazine)	0.35	6.16
3	**EC 6** (Diaveridine + sulfamethoxazole)	1.00	6.72
	EC 15 (Tetracycline + sulfamethazine)	0.00	5.41

4.5.2.3. Features responsible for toxicity against B. subtilis identified by the QSAR-3 model

Five distinct descriptors namely Eta_betaP_A, B05(C-S), NsCH3, minaasC, T(N..Cl) were identified in QSAR-3. Among these descriptors, all except T(N..Cl) are negatively contributing towards modeled endpoint (pEC50). According to the "Bubble plot" presented in **Figure 4.44C**, the order of relative importance of modeled descriptors is minaasC > NsCH3 > T(N..Cl) > B05(C-S) > Eta_betaP_A; therefore, minaasC (VIP score = 1.44) and

Eta_betaP_A (VIP score = 0.74) are identified as the most important and least important descriptors respectively. The minaasC is a negative contributing descriptor of QSAR-3 which indicates the electronic and topological state of fragments containing 2 aromatic and 1 single bond (Kier and Hall, 1990). We have depicted 4 such fragments inside a mixture component sulfamethoxazole (a component of the least toxic mixture **BS 15** of MD3) in **Figure 4.44C**. It has been observed in the studied mixtures that the toxicity decreases with an increase in minaasC value which indicates the inverse relationship. We have systematically compared 3 pairs of mixtures in terms of their pEC50 and minaasC values in **Table 4.30;** the result agrees with the negative contribution of minaasC towards the toxicity endpoint. NsCH3 is the next important descriptor with a VIP score of 0.9, which computes the number of single-bonded CH_3 groups in a molecular structure (see **Figure 4.44C**). In the comparative analysis represented in **Table 4.30**, various mixtures such as **BS 35**, **BS 21**, and **BS 26** have shown more NsCH3 values and thus, less toxicity than comparatively more toxic mixtures like **BS 32**. These observations justify the negative contribution of the mentioned descriptor. T(N..Cl) is the only positive contributing descriptor of QSAR-3 which indicates the total topological distance between the N-Cl atom pair in a molecule. In **Figure 4.44C**, we have depicted the presence of N-Cl atom pair in the structure of chlortetracycline which is a component of a comparatively more toxic mixture (**BS 24**; pEC50 = 5.98, T(N..Cl) = 9.60) of MD3. The mixtures without the N-Cl atom pair (e.g., **BS 2**, **BS 21**, **BS 15**) have shown less toxicity (see **Table 4.30**) which signifies the positive contribution of T(N..Cl). The B05[C-S] and Eta_betaP_A are the final two descriptors of QSAR-3 which negatively affect the toxic endpoint. B05[C-S] is a 2D atom pair descriptor which indicates the presence/absence of C-S atom pair at topological distance 5, whereas the Eta_betaP_S is an electrotopochemical atom index descriptor which is a measure of the multiple bonded electronegative atoms in the molecule. In both cases, a marked increase in the corresponding descriptor values has been observed from the most toxic mixture (**BS 32**: pEC50 = 6.37, B05[C-S] = 0, Eta_betaP_S = 0.40) to the least toxic mixture (**BS 15**: pEC50 = 4.77, B05[C-S] = 0.52, Eta_betaP_S = 0.48) of MD3 dataset which indicates the inverse relationship with toxicity against *B. subtilis*. The comparative analysis among 3 pairs of mixtures represented in **Table 4.30** also agrees with the negative contribution of these two descriptors.

Table 4.30. Relationship of descriptors with toxicity against *B. subtilis* (represented by comparative assessment between different mixtures of MD3)

minaasC (Negative contribution)			
Mixture pair	**Mix. No.**	**Descriptor Value**	**Toxicity (pLD50)**
1	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.14	6.37
	BS 42 (Oxytetracycline + ormetoprim + sulfameter)	0.16	6.00
2	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.14	6.37
	BS 21 (Tetracycline + diaveridine)	0.21	5.28
3	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.14	6.37
	BS 15 (Oxytetracycline + sulfamethoxazole)	0.24	4.77
NsCH3 (Negative contribution)			

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	1.46	6.37
	BS 35 (Tetracycline + trimethoprim + sulfameter)	1.52	5.93
2	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	1.46	6.37
	BS 21 (Tetracycline + diaveridine)	1.76	5.28
3	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	1.46	6.37
	BS 26 (Oxytetracycline + trimethoprim)	2.13	4.82

T(N..Cl) (Positive contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**BS 24** (Chlortetracycline + diaveridine)	9.60	5.98
	BS2 (Trimethoprim + sulfamethoxypyridazine)	0.00	5.73
2	**BS 24** (Chlortetracycline + diaveridine)	9.60	5.98
	BS 21 (Tetracycline + diaveridine)	0.00	5.28
3	**BS 24** (Chlortetracycline + diaveridine)	9.60	5.98
	BS 15 (Oxytetracycline + sulfamethoxazole)	0.00	4.77

B05[C-S] (Negative contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.00	6.37
	BS 36 (Sulfadoxine + diaveridine + sulfamethoxypyridazine)	0.27	6.03
2	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.00	6.37
	BS 2 (Trimethoprim + sulfamethoxypyridazine)	0.45	5.73
3	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.00	6.37
	BS 15 (Oxytetracycline + sulfamethoxazole)	0.52	4.77

Eta_betaP_A(Negative contribution)

Mixture pair	Mix. No.	Descriptor Value	Toxicity (pLD50)
1	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.40	6.37
	BS 24 (Chlortetracycline + diaveridine)	0.42	5.98

2	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.40	6.37
	BS 11 (Chlortetracycline + sulfamethoxypyridazine)	0.45	5.10
3	**BS 32** (Tetracycline + trimethoprim + sulfamonomethoxine)	0.40	6.37
	BS 15 (Oxytetracycline + sulfamethoxazole)	0.48	4.77

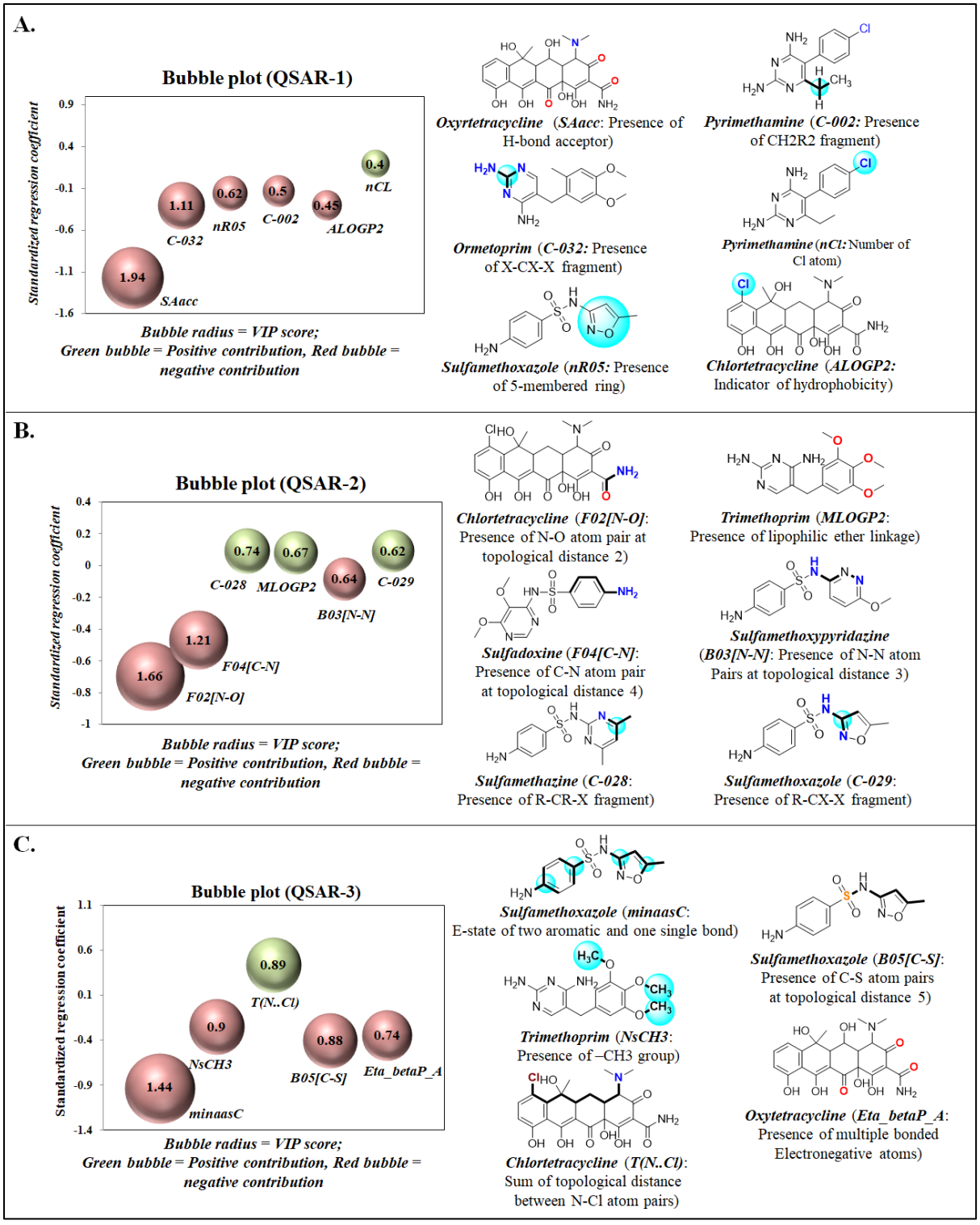


Figure 4.44. Variable importance plots of QSAR models along with the structural representation of identified descriptors; A. VIP of QSAR-1, B. VIP of QSAR-2, C. VIP of QSAR-3 (Bubble radius = VIP Score; Colour coding: Green = positive contribution, Red = negative contribution)

4.5.3. Insights of modeled descriptors identified in the QSAAR analysis

To understand the correlation of toxicity among different bacterial species, six preliminary QSAAR models (PLS) have been developed here. The impact of predicted toxicity of a species for toxicity prediction of another has been analysed by the VIPs and coefficient plots of QSAAR models (developed using SIMCA-P software). The VIP and coefficient plot of QSAAR-1 or -2, QSAAR-3 or -4, and QSAAR-5 or -6 have been showcased in **Figure 4.45**, **4.46**, and **4.47** respectively. From the VIPs of QSAAR-1 and 2 (**Figure 4.45**), it has been found that the predicted toxicity (*B. subtilis*) possesses positive contribution and significant relative importance towards the toxicity prediction against *V. fischeri*. In the VIP of QSAAR-5 (**Figure 4.47**), the predicted toxicity (*V. fischeri*) has shown maximum relative importance among other predictors for the toxicity prediction against *B. subtilis*. It is also a positively contributing descriptor which suggests that a toxic mixture against *V. fischeri* may also be toxic against *B. subtilis*. Therefore, it may be inferred from these observations that *V. fischeri* and *B. subtilis* have similar toxicity mechanisms. The predicted toxicity (*B. subtilis*) does not possess significant importance towards the toxicity prediction against *E. coli*. Among the predicted toxicities, *B. subtilis* have shown negative contribution to the toxicity prediction against *E. coli* in QSAAR-4. Therefore, a more toxic mixture against *B. subtilis* may be less toxic towards *E. coli*. The predicted toxicity (*E. coli*) has not shown much importance towards the toxicity prediction against *V. fischeri* also, and it has been confirmed by the VIP of QSAAR-3 (**Figure 4.46**). Therefore, it may be inferred that the *E. coli* does not have a similar mechanism of toxicity with the other two species (*V. fischeri* and *B. subtilis*).
By analyzing the descriptors of QSAAR-1 and 2, it has been observed that QSAAR-1 and 2 have two (C-032 and nCL) and three (SAacc, C-032, and ALOGP2) common descriptors respectively with QSAR-1. These common descriptors have similar impact towards toxicity as seen in QSAR-1. The QSAAR-3 and 4 share two common descriptors (F02[N-O] and F04[C-N]) with QSAR-2. Both of these descriptors are negatively contributing in QSAR-2, QSAAR-3, and QSAAR-4. Among the descriptors of QSAAR-5 and 6, only T(N..Cl) is seen as the common predictor with a positive coefficient in QSAR-3 and QSAAR-6.

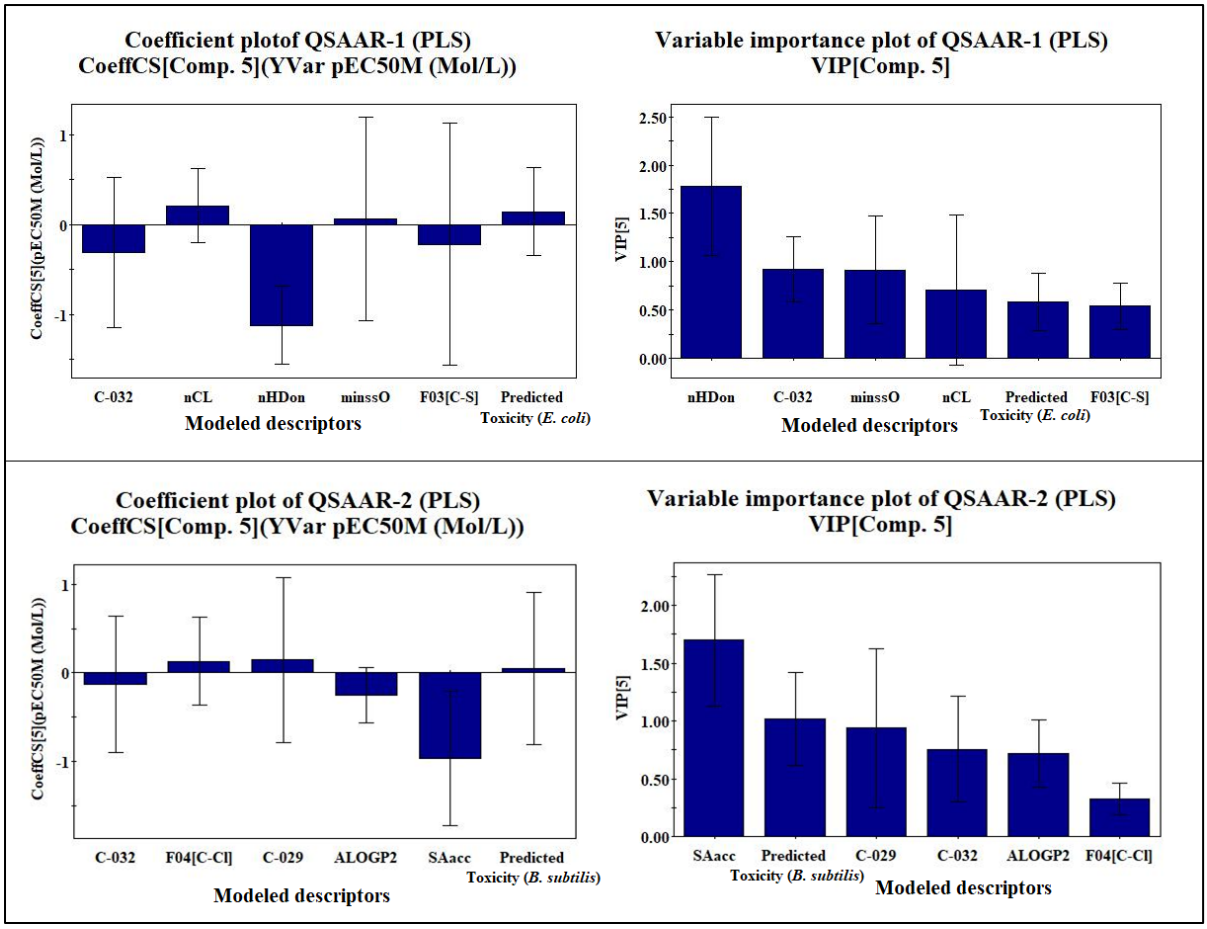


Figure 4.45. Coefficient plot and variable importance plots of QSAAR-1 and QSAAR-2

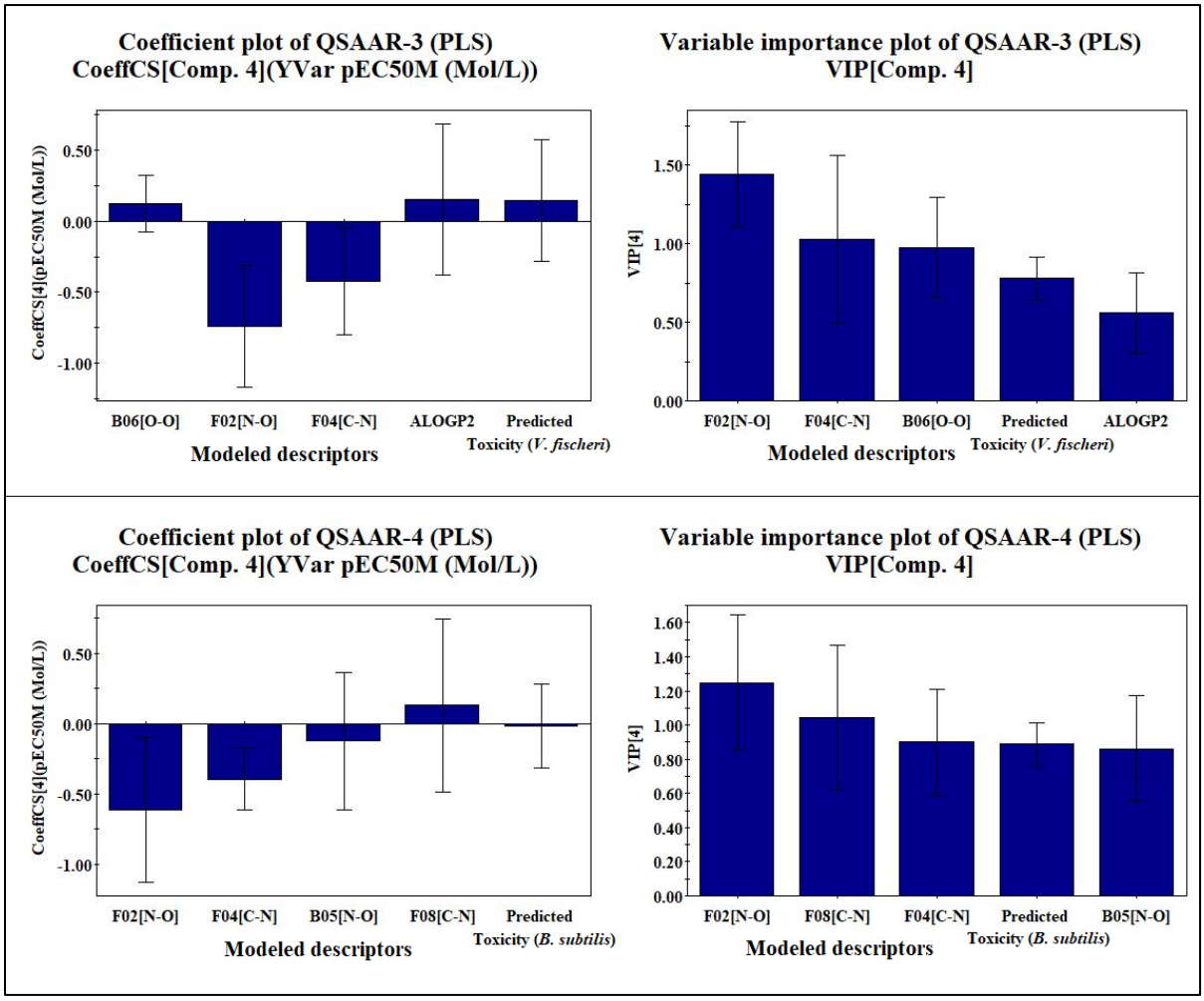


Figure 4.46 Coefficient plot and variable importance plots of QSAAR-3 and QSAAR-4

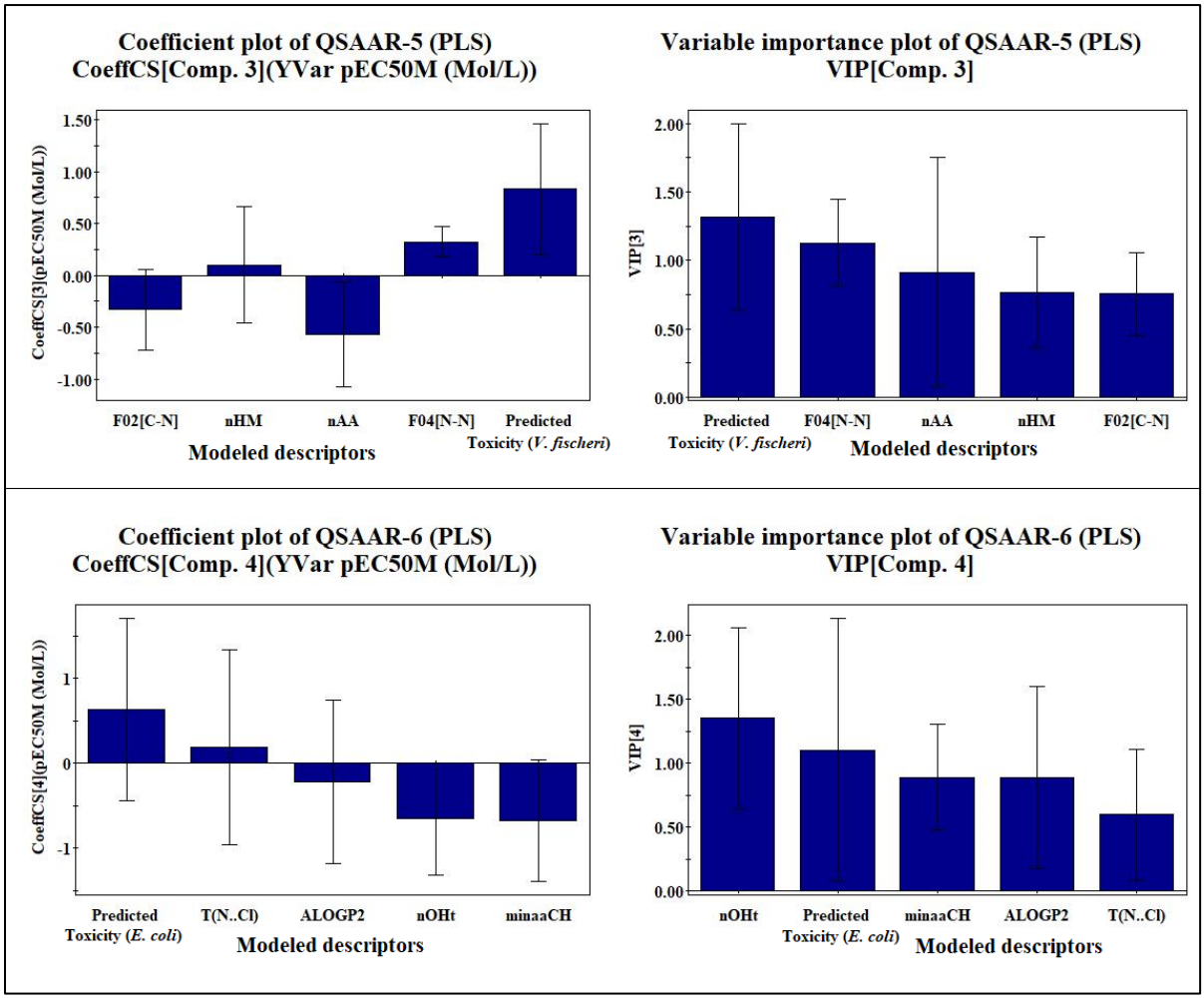


Figure 4.47. Coefficient plot and variable importance plots of QSAAR-5 and QSAAR-6

4.5.4. Importance of q-RASAR algorithm in interspecies mixture toxicity modeling

The q-RASAR algorithm has played a crucial role in developing the QSAAR models for the toxicity prediction against three bacterial endpoints in this study. As discussed earlier that the q-RASAR is a mixed approach of non-linear read-across predictions and a linear QSAR framework. Therefore, the non-linearity obtained during the similarity estimation by Gaussian kernel function/Laplacian kernel function are applied to form a linear QSAR equation. In this study, we have used the optimized Laplacian kernel function (Chatterjee et al., 2022) for the RASAR descriptor calculation of MD1 and MD3, and the optimized Euclidean distance-based similarity (Chatterjee et al., 2022) for the MD2 (mathematical representations of Laplacian kernel function similarity and Euclidean distance-based similarity have been discussed in "***1.6.2.1. Similarity estimation techniques used in read-across predictions***" section). The RASAR descriptors are derived by extracting structural information obtained from the dataset; here both structural and biological information (in form of the predicted toxicity of other bacterial species) are used to develop the final RASAR-based QSAARs or q-RASAARs (see **Figure 4.48**). We got two univariate MLR models (q-RASAAR-1 and -3) with *RA function (LK)* as the predictor, and 1 PLS model (q-RASAAR-2) with 4 descriptors (including RASAR descriptor *Neg. Avg. Sim*) and 3 LVs. The RASAR descriptor *RA function (LK)* is the read-across derived prediction function obtained by weighted averaging of response values of identified similar structural analogues (source

compounds) during non-linear similarity estimation (Laplacian kernel function), whereas, the *Neg. Avg. Sim.* is the the mean of similarity values of non-toxic (with respect to the training set response mean) structural analogues identified by Euclidean distance-based similarity estimation (Banerjee et al., 2023). Both of these descriptors have proportional relationship with toxic endpoints as evidenced from the positive regression coefficients (see **Table 4.26**). The statistical quality we got is comparable with QSAR models in terms of internal validation metrics and superior in terms of external validation metrics (**Figure 4.49**). The variance ratio (F value) has been improved from QSAR (F value: QSAR-1 = 167.856, dfs: 5, 51; QSAR-2 = 133.59, dfs: 5, 45; QSAR-3 = 56.92, dfs: 4, 27) to RASAR (F value: q-RASAAR-1 = 722.68, dfs: 1, 55; q-RASAAR-2 = 192.55, dfs: 3, 47; q-RASAAR-3 = 165.49, dfs: 1, 30) which can be a marker of better statistical significance of RASAR models. For better understanding, we have plotted the F values of QSARs and q-RASAARs in **Figure 4.49**. The MAE_{Test} values of QSAR, preliminary QSAARs, and q-RASAAR models have also been plotted in Y-axis of column plots (3 column plots for three datasets) and have given in **Figure 4.49**. An increase in prediction error in terms of MAE_{Test} has been observed from QSAR-1 to QSAAR-1/-2 and QASR-3 to QSAAR-5/-6 which indicates inferior prediction by PLS-based interspecies modeling. But during the q-RASAAR model development, structural (descriptors) and biological features (predicted endpoint values against other two bacterial species) of two QSAARs (QSAAR-1/-2 to q-RASAAR-1; QSAAR-3/-4 to q-RASAAR-2; QSAAR-5/-6 to q-RASAAR-3) are extracted into the corresponding RASAR descriptors which significantly reduce the prediction errors. It is clearly seen that the MAE_{Test} is minimum in q-RASAAR-1 (MAE_{Test} = 0.080), q-RASAAR-2 (MAE_{Test} = 0.080), and q-RASAAR-3 (MAE_{Test} = 0.093) compared to the other QSARs (MAE_{Test}: QSAR-1 = 0.112; QSAR-2 = 0.098; QSAR-3 = 0.198) and preliminary QSAARs (MAE_{Test}: QSAAR-1 = 0.121; QSAAR-2 = 0.124; QSAAR-3 = 0.088; QSAAR-4 = 0.081; QSAAR-5 = 0.198; QSAAR-6 = 0.253). We have also noticed a decrease in CCC values from QSAR-3 (CCC = 0.845: marginal to the threshold) to QSAAR-6 (CCC = 0.807: inferior) and then a marked increase in q-RASAAR-3 (CCC = 0.969: superior) which indicates the efficiency of RASAR algorithm in interspecies modeling. In each of the three datasets, AICs of q-RASAAR (AIC: q-RASAAR-1 = -94.858, q-RASAAR-2 = -82.424, q-RASAAR-3 = -44.586) were less than the AICs of QSARs (AIC: QSAR-1 = -71.238, QSAR-2 = -82.315, QSAR-3 = -41.726) which indicates the better relative efficiency of q-RASAARs in terms of performance and descriptor combination.

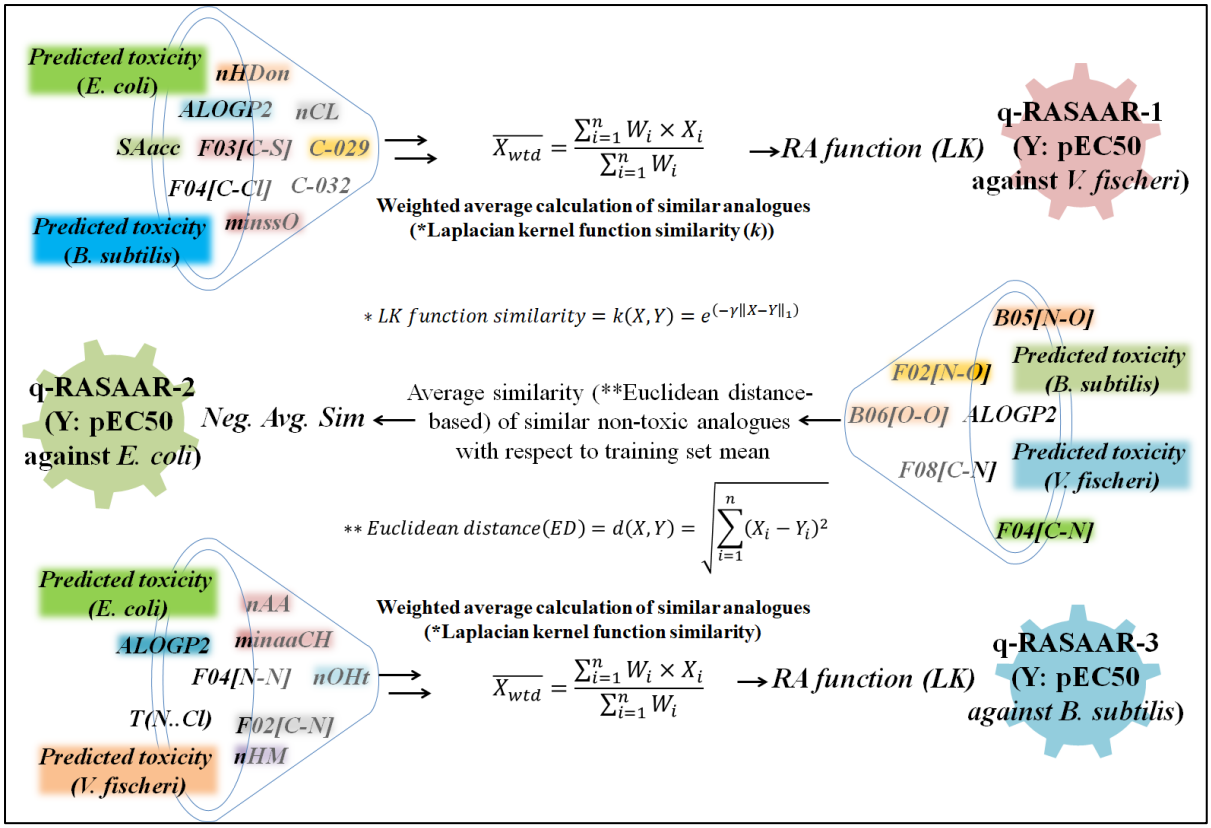


Figure 4.48. Representation of the features involved to obtain relevant RASAR descriptors (for q-RASAAR modeling)

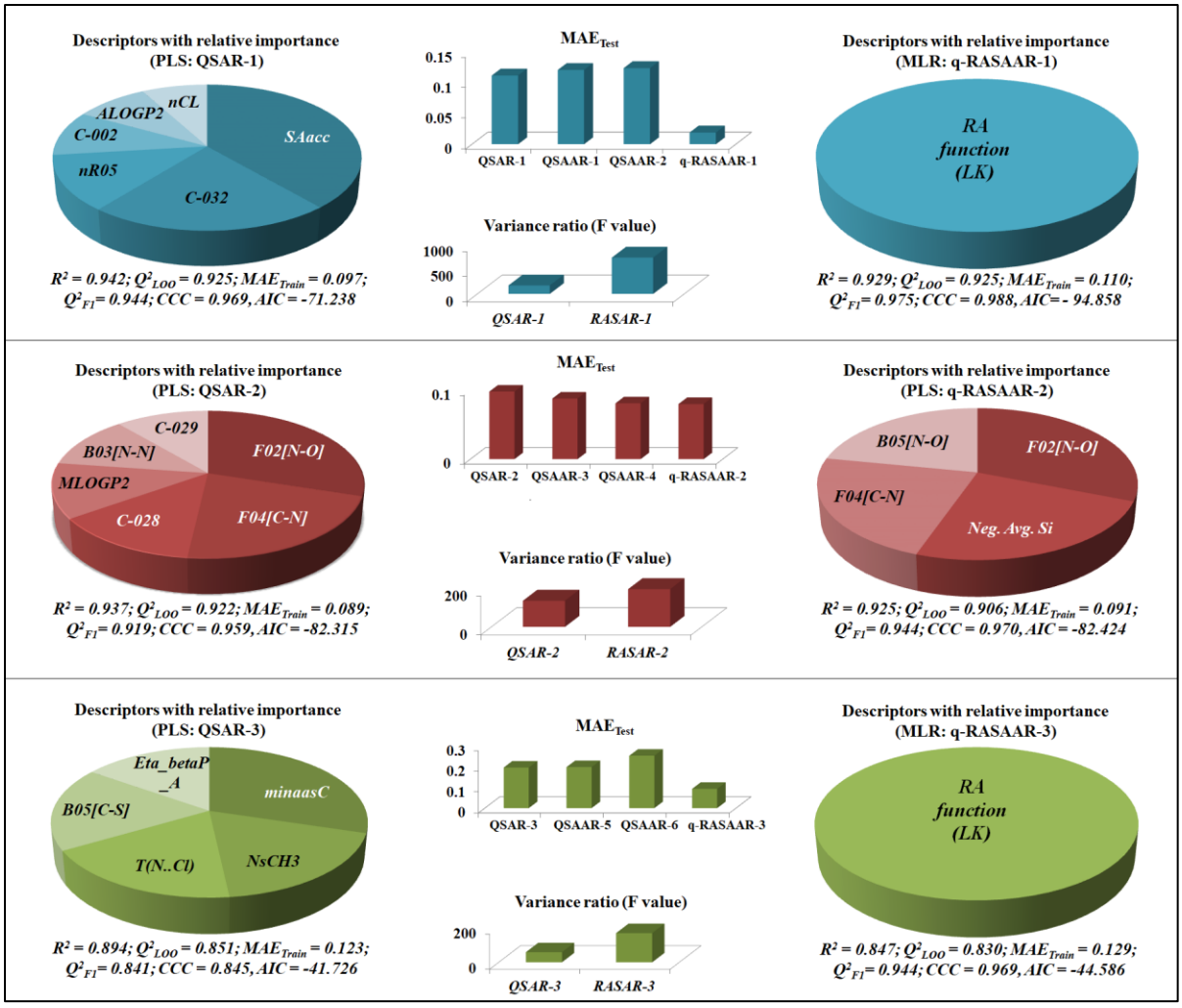


Figure 4.49. Comparison of (i) importance of descriptors used in QSAR and q-RASAAR models, (ii) MAETest among QSAR, QSAAR, and q-RASAAR models, (iii) Variance ratio (F value) between QSAR and q-RASAAR models

4.5.5. Comparative analysis with the previous report

To judge the merit of any predictive model, comparative analysis with previous reports can be a useful tool. The use of the interspecies model in mixture toxicity prediction is a new approach and hence, we did not get any previous QSAAR models of mixtures for the studied endpoints. However, we have compared the quality of newly developed QSAR and q-RASAAR models with previously published QSAR models by Wang *et. al*. (D. Wang et al., 2018). A total of 12 MLR models (4 models against each dataset) were reported by Wang et. al. where they have used molecular docking-derived E_{bind} descriptors as the predictor variables. They separately developed more specific toxicity models for binary (SA + SAP, SA + TC, SAP + TC) and ternary mixtures (SA + SAP + TC) against each bacterial species with comparatively smaller data points (see **Table 4.31**). However, a more general strategy has been taken for the development of QSAR and QSAAR in the present study, and therefore, we have used both binary and ternary mixtures together for the model development. Comparatively greater number of data points have been used for model development in this study (QSAR-1/q-RASAAR-1: n_{Train} = 57, QSAR-2/q-RASAAR-2: n_{Train} = 51, QSAR-3/q-RASAAR-3: n_{Train} = 32) which is desirable for any statistical modeling. We have also compared the robustness and predictability in terms of computed R^2, Q^2_{LOO}, Q^2_{F1}

values and tabulated them into the following **Table 4.31**. The QSAR-1 and q-RASAAR-1 clearly exceed (QSAR-1: R^2 = 0.942, Q^2_{LOO} = 0.925, Q^2_{F1}=0.944; q-RASAAR-1: R^2 = 0.929, Q^2_{LOO} = 0.924, Q^2_{F1}=0.975) the validation results of all 4 previous QSARs for predicting toxicity to *V. fischeri*. Similar observations have been obtained for the rest two dataset also. However, we have also found two previous models with greater R^2 and Q^2_{LOO} from MD2 and MD3 than the new models of corresponding datasets. The previous QSAR model developed for the prediction of toxicity of mixtures (comprised of SAP and TC) against *E. coli* showed a R^2 value (0.951) greater than the corresponding QSAR-2 (R^2 = 0.936) and q-RASAAR-2 (R^2 = 0.924). However, the previous model was developed with 6 data points and thus, it is statistically less significant than the QSAR-2 and q-RASAAR-2 (n_{Train} = 51). Another previous model with greater R^2 (0.999) and Q^2_{LOO} (0.945) values was developed for the prediction of binary mixtures (SA-SAP) against *B. subtilis* from only 5 datapoints; thus statistically less significant. On the contrary, the newly developed QSAR-3 (R^2 = 0.893, Q^2_{LOO} = 0.850) and q-RASAAR-3 (R^2 = 0.846, Q^2_{LOO} = 0.830) models have been developed with a higher number of training data points (n_{Train} = 32), thus having higher statistical significance. This study aimed to develop simple and transferable prediction models for the efficient toxicity prediction of antibiotic mixtures. In this direction, 0D-2D descriptors have been used due to their simplicity, reproducibility, and interpretability; the developed models have been proved to be good-fit, robust, and highly predictive also in comparison with the previous report (Table 3). Although the use of 0D-2D descriptors limit the ability of model to explain the toxicity mechanism to some extent, but it enhances the reproducibility and transferability of the models. Molecular docking-based descriptors have been used for modeling the mixture toxicity in some of the previous studies; however, those molecular docking-based descriptors do not appear appropriate for modeling nonspecific interactions as no particular target is being considered here for the toxicity. Apart from that, the exact toxicity mechanism is also uncertain, because the toxicity data was obtained by simple assay where the concentration of sample broth is the determinant of lethality. The targets for the toxicity may vary depending on the chemical classes and the organism being considered; therefore, the use of a fixed target receptor or protein for docking-based descriptor calculation potentially limits the efficiency of predictions. The predicted toxicity values of other bacteria have also been incorporated in QSAAR/q-RASAAR models for improving the predictive performance further. We have used the available toxicity data of one bacterial species with other structural descriptors for more efficient prediction of other bacterial endpoints. Therefore, from the perspective of statistical significance, robustness, external predictability, and better usability, the new models (both QSARs and RASARs) perform well and are also able to produce more realistic predictions than the previous models.

Table 4.31. Comparison of results obtained by this study with previously published report in terms of number of descriptors, robustness, and predictability

Work done by	Model	Type of mixtures	Bacterial species	No. of training data	R^2	Q^2_{LOO}	Q^2_{F1}	Descriptors (No. of LVs)	Type of descriptors
Wang, *et al*. (2018)	MLR	SA-SAP	*V. fischeri*	18	0.699	0.6	0.77	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		15	0.813	0.714	0.601	2	
		SAP-TC		8	0.829	0.603	0.760	2	
		SA-SAP-TC		21	0.729	0.601	0.825	3	
Current study	**PLS: QSAR-1**	**Both binary and ternary**	***V. fischeri***	**57**	**0.942**	**0.925**	**0.944**	**6 (5 LVs)**	**0D to 2D**
Current study	**q-RASAAR-1 (MLR)**	**Both binary and ternary**	***V. fischeri***	**57**	**0.929**	**0.924**	**0.975**	**1**	***RASAR* descriptor**
Wang, *et al*. (2018)	MLR	SA-SAP	*E. coli*	9	0.788	0.505	0.500	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		21	0.740	0.649	0.629	2	
		SAP-TC		6	0.951	0.823	0.702	2	
		SA-SAP-TC		20	0.805	0.627	0.640	3	
Current study	**PLS: QSAR-2**	**Both binary and ternary**	***E. coli***	**51**	**0.936**	**0.921**	**0.919**	**6 (5 LVs)**	**0D to 2D**
Current study	**q-RASAAR-2 (PLS)**	**Both binary and ternary**	***E. coli***	**51**	**0.924**	**0.905**	**0.944**	**4 (3 LVs)**	**0D to 2D *and RASAR* descriptor**
Wang, *et al*. (2018)	MLR	SA-SAP	*B. subtilis*	5	0.999	0.945	0.702	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		9	0.893	0.812	0.500	2	
		SAP-TC		8	0.778	0.526	0.906	2	
		SA-SAP-TC		11	0.828	0.508	0.754	3	
Current study	**PLS: QSAR-3**	**Both binary and ternary**	***B. subtilis***	**32**	**0.893**	**0.850**	**0.841**	**5 (4 LVs)**	**0D to 2D**
Current study	**q-RASAAR-3 (MLR)**	**Both binary and ternary**	***B. subtilis***	**32**	**0.846**	**0.830**	**0.964**	**1**	***RASAR* descriptor**

4.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach

4.6.1. Optimized mixing rule and best division

We have optimized the mixing rule based on the statistical quality of the preliminary GA-MLR models developed with 9 sets of division obtained from the descriptor matrix of each mixing rule. The computed validation metrics (R^2, Q^2_{LOO}, $MAE_{Train95\%}$, Q^2_{F1}, and $MAE_{Test95\%}$) of each model have been tabulated in **Table 4.32**. We have taken the average of validation metrics from all the divisions. Among the three mixing rules, the linear combination of molecular contributions has shown the best average metric values (R^2 = 0.883, Q^2_{LOO} = 0.853, $MAE_{Train95\%}$ = 0.085, Q^2_{F1} = 0.827, $MAE_{Test95\%}$ = 0.091) and therefore, it has been selected as the optimum mixing rule for mixture descriptor calculation in this study. To select the best division, we have compared the performance metrics of GA-MLR models obtained from mixing rule 1. It has been observed that the 2nd seed of 75:25 activity-based division (4-2 ABD) yielded a model with the most balanced performance (R^2 = 0.874, Q^2_{LOO} = 0.842, $MAE_{Train95\%}$ = 0.087, Q^2_{F1} = 0.837, $MAE_{Test95\%}$ = 0.088). Thus, we have proceeded to the feature selection step with 4-2 ABD.

Table 4.32. Computed validation metrics of preliminary GA-MLR models (Number of descriptors in each model = 6)

Mixing rule 1 (Linear combination of molecular contributions)

Division	R^2	Q^2_{LOO}	$ME_{Train95\%}$	Q^2_{F1}	$ME_{Test95\%}$
4-1ABD	0.897	0.870	0.081	0.809	0.097
4-1DBD	0.886	0.856	0.082	0.817	0.106
4-2ABD	**0.874**	**0.842**	**0.087**	**0.837**	**0.088**
4-2DBD	0.863	0.828	0.089	0.889	0.084
4-3ABD	0.880	0.850	0.088	0.833	0.092
4-3DBD	0.874	0.844	0.092	0.863	0.072
4-4ABD	0.884	0.857	0.085	0.833	0.086
4-4DBD	0.887	0.858	0.084	0.827	0.092
75-25KS	0.901	0.875	0.082	0.733	0.102
Average	**0.883**	**0.853**	**0.085**	**0.827**	**0.091**

Mixing rule 2 (Square molecular contributions)

Division	R^2	Q^2_{LOO}	$ME_{Train95\%}$	Q^2_{F1}	$ME_{Test95\%}$
4-1ABD	0.890	0.861	0.079	0.774	0.103
4-1DBD	0.858	0.822	0.092	0.840	0.085
4-2ABD	0.863	0.827	0.088	0.825	0.090
4-2DBD	0.872	0.838	0.090	0.751	0.092

4-3ABD	0.877	0.846	0.084	0.810	0.091
4-3DBD	0.878	0.847	0.081	0.734	0.117
4-4ABD	0.865	0.829	0.088	0.815	0.091
4-4DBD	0.869	0.835	0.083	0.801	0.107
75-25KS	0.882	0.852	0.087	0.762	0.091
Average	0.873	0.840	0.086	0.790	0.096

Mixing rule 3 (Norm of molecular contributions)

Division	R^2	Q^2_{LOO}	$ME_{Train95\%}$	Q^2_{F1}	$ME_{Test95\%}$
4-1ABD	0.865	0.829	0.088	0.767	0.099
4-1DBD	0.829	0.778	0.099	0.773	0.098
4-2ABD	0.829	0.783	0.100	0.765	0.102
4-2DBD	0.862	0.823	0.083	0.737	0.141
4-3ABD	0.835	0.782	0.097	0.817	0.095
4-3DBD	0.888	0.858	0.083	0.644	0.121
4-4ABD	0.887	0.857	0.076	0.576	0.130
4-4DBD	0.849	0.806	0.096	0.764	0.089
75-25KS	0.864	0.820	0.089	0.744	0.100
Average	0.856	0.815	0.090	0.732	0.108

4.6.2. Partial least squares (PLS) regression model and validation

As discussed in the section "***3.6.4. Feature selection and model development***", all possible subset selection has been employed to identify the best descriptor combination from the reduced pool of 40 descriptors. Six descriptors namely H-047, Eta_F_A, nR03, F05[C-O], nBnz, and MLOGP2 have been identified as the most prominent descriptors for the prediction of toxicity of fluoroquinolone mixtures. We have moved forward with these six descriptors and developed the QSAR model using the PLS algorithm. To optimise the number of LVs, a PLS regression has been run with the MINITAB software("Data Analysis, Statistical & Process Improvement Tools | Minitab," 2004) and the X-variances of each incoming LV have been noted (see **Table 4.33**).

Table 4.33. Computed X-variances, R^2, and Q^2_{LOO} values after incorporating each LVs

Number of LVs	**X-Variance**	$\mathbf{R^2}$	$\mathbf{Q^2_{LOO}}$	**Inference**
1	0.316	0.599	0.544	-
2	0.623	0.671	0.612	-
3	0.732	0.766	0.657	-

4	0.833	0.864	0.801	-
5	**0.948**	**0.882**	**0.849**	Mximum possible X-variance accounted. Thus, the optimum number of **LVs = 5**
6	1.000	0.885	0.858	In PLS, the number of LVs < number of descriptors

With 5 LVs, maximum X-variance has been extracted (X-variance = 0.948) from the descriptor matrix and therefore, final PLS regression has been performed with 5 LVs. It should be noted that the number of LVs in PLS must be lower than the number of descriptors; otherwise, the PLS model will be converted to an MLR equation. The developed PLS model has been shown in the following equation 4.25 along with the computed validation metrics-

$$Log\ (1/IC50) = -1.09774 + 0.1092 \times H - 047 + 8.10906 \times Eta_F_A + 1.50499 \times nR03 - 0.11551 \times F05[C-O] - 0.33295 \times nBnz - 0.144 \times MLOGP2 \qquad (4.25)$$

n_{Train} = 59, R^2 = 0.882, R^2_{adj} = 0.868, Q^2_{LOO} = 0.849, MAE_{Train} = 0.085, $\overline{r_m^2(Train, LOO)} = 0.788$, $\Delta r_m^2(Train, LOO) = 0.123$, $RMSE_C$ = 0.108; n_{Test} = 19, Q^2_{F1} = 0.859, Q^2_{F2} = 0.858, Q^2_{F3} = 0.873, MAE_{Test} = 0.090, $\overline{r_m^2(Test)} = 0.811$, $\Delta r_m^2(Test) = 0.089$; CCC = 0.927, $RMSE_P$ = 0.112

The developed model is statistically sound and externally predictive as evidenced by the high computed validation metric values (above 0.8) such as R^2, R^2_{adj},Q^2_{LOO}, Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}. The error metrics of both training and test sets (such as MAE_{Train} and $RMSE_C$, MAE_{Test} and $RMSE_P$) are on the lower side (below 0.15) which signifies the reliability of predictions. There is a very minute difference (0.033) between the determination coefficient (R^2) and the cross-validated correlation coefficient (Q^2_{LOO}), which is an indicator of the robustness of the model. The computed values of r^2_m metrics are also in the acceptable range. We have computed the CCC of the developed model and its higher value (CCC = 0.927) represents the proximity of the regression line and the concordance line (line passing through the origin of observed vs. predicted scatter plot) of the developed model; therefore, the accuracy and precision of the predicted responses have been acknowledged. We have shown the observed vs. predicted scatter plot in the following **Figure 4.50A** where the least scattering and uniform distribution of the training and test data points have been observed surrounding the line passed through the origin. The score plot obtained using the SIMCA-P software indicates the uniform representations of the training data points from all the quadrants of the ellipse (**Figure 4.50B**). We have run the Y-randomization test using SIMCA-P software to cut down the possibility of chance correlation. The R^2 and Q^2 intercepts of the Y-randomization plot have been computed and the resultant Y-randomization plot has been given in the following **Figure 4.50C**. These R^2 and Q^2 intercept values (R^2 = 0.0069, Q^2 = -0.577) have not crossed the threshold limit which indicates the statistical soundness of the model and affirms that it is not the result of any chance correlation.

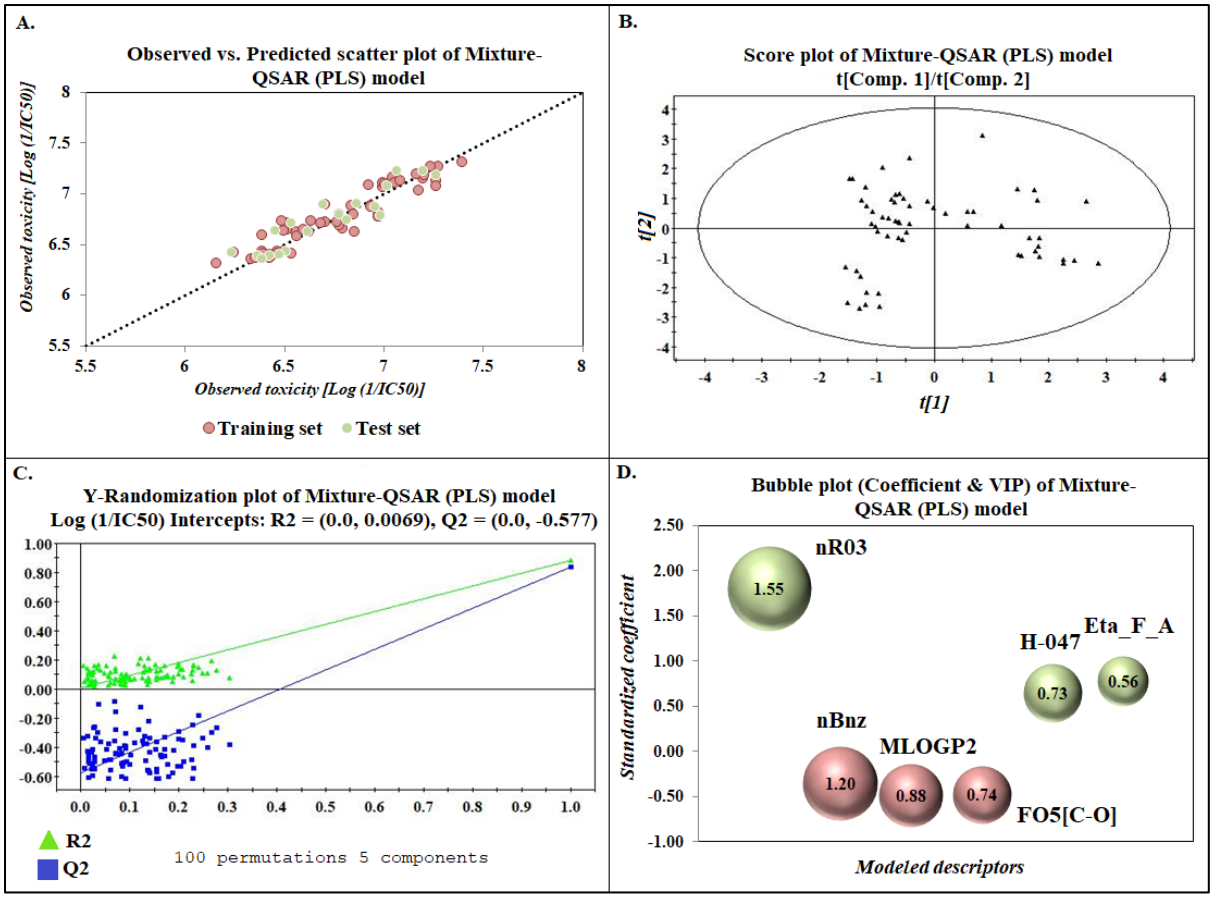


Figure 4.50. Graphical representations [A. Observed vs. predicted toxicity scatter plot, B. Score plot, C. Y-Randomization plot, D. Bubble plot (coefficient & VIP)].

To know the contributory effect and order of importance of the modeled descriptors, we have derived the standardized regression coefficients and the VIP scores using SIMCA-P software. After arranging the descriptors in descending order of the VIP scores and giving them the number of importance (1 to 6), we plotted them against the respective standardized coefficient in a bubble plot (**Figure 4.50D**). The number of importance and the standardized coefficient have been plotted in the bubble plot along the X-axis and Y-axis respectively where the the bubbles' radiuses represent the VIP scores of respective descriptors. As per the bubble plot, nR03 and nBnz are the most important descriptors with VIP scores > 1. Among the 6 descriptors, nR03, H-047, and Eta_F_A are the positively contributory descriptors (represented by green colour in **Figure 4.50D**) whereas, nBnz, MLOGP2, and F05[C-O] are the negatively contributory descriptors (represented by red colour in **Figure 4.50D**).

4.6.3. Machine learning models and the result of validation

To get the best possible predictions from ML, we first optimized the hyperparameter settings of different ML algorithms using the scaled training set. The optimized settings for different ML approaches have been given in **Table 4.34**. Finally, we got the ML-derived predictions using RF, AdaBoost, GB, XGB, SVM, LSVM, and RR algorithms. The validation metrics have been computed and tabulated in the following **Table 4.35**. Among these seven different ML algorithms, the model derived by extreme gradient boost (XGB) has the most balanced performance. The high R^2 (0.885) and Q^2_{LOO} (0.858) values, and the minute difference between them (0.027) confirm the goodness of fit and robustness respectively. On the

contrary, the high value of external correlation coefficients (Q^2_{F1} = 0.858, Q^2_{F2} = 0.857, Q^2_{F3} = 0.872) indicates the good external predictability of the model. The error metrics such as MAE_{Train} (0.267) and MAE_{Test} (0.279) are also on the lower side.

Table 4.34. Optimized hyperparameters of the ML algorithm used in this study

ML algorithm	Optimized setting
RF	*Max-depth = 2, max_leaf_nodes = none, min_impurity_decrease = 0.0, min_sample_split = 2, min_sample_leaf = 3, n_estimator = 200, min_weight_fraction_leaf =0.0*
Ada-Boost	*n-Estimator =70 , loss =square*
GB	*Max-depth = 4, min_sample_split = 2, min_sample_leaf = 1, n_estimator = 50, criterion = friedman_msc, min_weight_fraction_leaf = 0.0, min_impurity_decrease = 0.0, subsample = 1.0, learning_rate = 0.1, loss = squared_error*
XGB	*n_estimator = 200, max_depth = none, booster = gblinear, learning_rate =1.0*
NL-SVM	*C = 5, degree = 2, gamma =auto*
L-SVM	*C =1*
RR	*Alpha (α) =0.5*

Table 4.35. Computed validation metrics obtained from the predictions of ML models

Metrics	RF	AB	GB	XGB	NL-SVM	LSVM	RR
R^2	0.786	0.907	0.996	**0.885**	0.917	0.877	0.881
Q^2_{LOO}	0.657	0.648	0.619	**0.858**	0.713	0.844	0.852
MAE_{Train}	0.349	0.240	0.053	**0.267**	0.198	0.263	0.271
$RMSE_C$	0.458	0.302	0.064	**0.336**	0.286	0.348	0.343
Q^2_{F1}	0.772	0.769	0.701	**0.858**	0.838	0.838	0.847
Q^2_{F2}	0.770	0.767	0.698	**0.857**	0.837	0.837	0.845
Q^2_{F3}	0.794	0.791	0.730	**0.872**	0.854	0.854	0.862

MAE_{Test}	0.345	0.355	0.391	**0.279**	0.284	0.287	0.284
CCC	0.859	0.867	0.839	**0.925**	0.906	0.910	0.918
$RMSE_P$	0.450	0.453	0.515	**0.355**	0.379	0.379	0.369

We have compared the XGB-derived ML model and the PLS model based on their overall performance. Both of these models have almost similar goodness of fit, robustness, and external predictability. However, the PLS model has shown the least errors of predictions (MAE_{Train} = 0.085 and MAE_{Test} = 0.090) as compared to the XGB-derived ML model (MAE_{Train} = 0.2668 and MAE_{Test} = 0.2787). Therefore, in terms of the overall performance, the PLS model has been chosen as the best one.

4.6.4. Insights of modeled descriptors

As we have stated earlier, six 2D descriptors namely, H-047, Eta_F_A, nR03, F05[C-O], nBnz, and MLOGP2 have been identified by best subset selection algorithm and then modeled using PLS regression in this study. We have discussed the contributions of each descriptor in the coming section following their order of importance towards the studied endpoint. The structural features defined by the identified descriptors have been showcased in the following **Figure 4.51**.

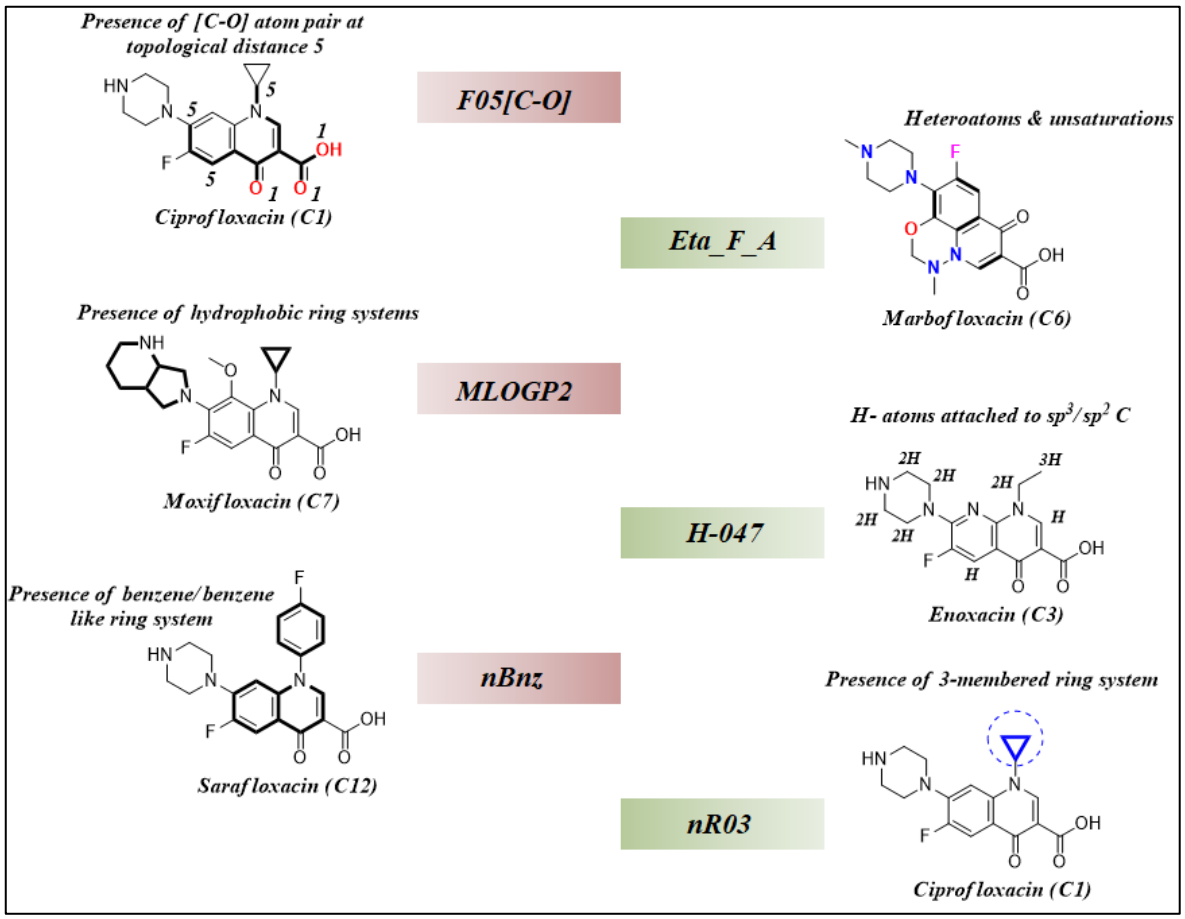


Figure 4.51. Structural representation of modeled descriptors (green blocks: positive contributory descriptors, red blocks = negative contributory descriptors)

4.6.4.1. nR03 descriptor

nR03 is the most important and positive contributing descriptor of this study. It belongs to the class of 'Ring descriptor' and represents the number of three-membered ring systems in the molecule. It has been observed that the presence of the 3-membered ring in the mixture component potentiates mixture toxicity. For better understanding, we have compared a set of mixtures with structural representation in **Figure 4.52**. The most toxic mixture [**Mixture M12:** Log (1/IC50) = 7.4] of this data set consists of the cyclopropyl ring containing components ciprofloxacin and sparfloxacin which enhance the mixture toxicity by a considerable amount. On the contrary, the mixture of enoxacin and norfloxacin (**Mixture M28**) is devoid of the 3-membered ring system, and thus, it is found to be the least toxic (**Mixture M28:** Log (1/IC50) = 6.16) among all the training data points.

4.6.4.2. nBnz descriptor

nBnz is the next important descriptor, which indicates the number of benzene or benzene-like ring systems in a molecule. It is a negative contributing descriptor that seems to decrease the mixture toxicity. We have comparatively analysed two mixtures (**M56** and **M22**) with different toxicities [Log (1/IC50): **M56** = 7.17; **M22** = 7.27] in **Figure 4.53**. It has been observed that the comparatively less toxic mixture **M56** contains marbofloxacin (mole fraction = 0.4) and sarafloxacin (mole fraction = 0.6) as the mixture components, and the more toxic mixture **M22** contains danofloxacin (mole fraction = 0.61) and sarafloxacin (mole fraction = 0.39) as the mixture components. Sarafloxacin has a 4-fluorobenzyl substitution on the nitrogen of the central quinolone ring and it is the constant component of both the mixtures. However, it constitutes 60% of **Mixture M56** and 39% of **Mixture M22**. Among the variable components, marbofloxacin constitutes 40% of **Mixture M56**, and it has a 6-membered fused ring attached to the central quinolone. On the contrary, Danofloxacin of **Mixture M22** constitutes 61% of the mixture and contains a cyclopropyl ring in its structure (cyclopropyl ring helps to increase toxicity). Due to the presence of more benzene-like ring systems attached to the central quinolone ring, less toxic **Mixture M56** has a greater nBnz value (1.6) than the more toxic counterpart Mixture M22 (nBnz = 1.39). This comparative analysis clearly justifies the negative contributory effect of descriptor nBnz and also establishes the positive contributory effect of descriptor nR03.

4.6.4.3. MLOGP2 descriptor

The next important descriptor is the molecular property MLOGP2, which negatively affects the endpoint predictition in this study. MLOGP2 is the squared Moriguchi octanol-water partition coefficient which indicates the lipophilicity of chemicals. Generally, the lipophilic chemicals become more toxic due to the easy transportation in the cell through the lipid bi-layers; however, the increase has a limit. Roy *et al*. reported the parabolic relationship between lipophilicity and cellular toxicity where they showed a continuous decrease of toxicity after reaching the peak level with a continuous increase in lipophilicity.(Roy et al., 2020) Besides this, Wang *et al.* identified the π-π stacking interaction and hydrogen bonding interaction as the most important mechanism for toxic action of fluoroquinolones in a docking study.(Wang et al., 2020) Therefore, it has been established that the lipophilicity is not increasing the toxicity here. We have comparatively analysed the studied mixtures also in the **Figure 4.53**. A more lipophilic **Mixture M27** (MLOGP2 = 4.014) is found to be less toxic [Log (1/IC50) = 6.25] than a comparatively less lipophilic **Mixture M05** [MLOGP2 = 2.331; Log (1/IC50) = 7.28] which justifies the negative contributory effect of MLOGP2 towards the toxic endpoint of this study.

4.6.4.4. F05[C-O] descriptor

F05[C-O] is a 2D-atom pair descriptor which represents the frequency of [C-O] atom pairs present at topological distance 5 in a molecule. Its negative regression coefficient indicates an inverse correlation between the toxic endpoint and the descriptor value. We have carefully analysed the components of two different mixtures (**M59** and **M12**) presented in **Figure 4.53**. It has been observed that the atom pair [C-O] at topological distance 5 mostly represents the central quinolone ring. Therefore, a greater value of this descriptor corresponds to the attachment of extra ring systems to the central quinolone as seen in **Figure 4.53**, and this structural deviation of the main quinolone ring is detrimental to the toxicity. **Mixture M59** has a greater descriptor value (F05[C-O] = 12.04) and comparatively less toxicity [Log (1/IC50) = 6.69] than **Mixture M12** with less descriptor value (F05[C-O] = 8.0) and comparatively more toxicity [Log (1/IC50) = 7.4]. These observations justify the negative contributory effect of F05[C-O] towards the studied endpoint in *E. coli*.

4.6.4.5. H-047 descriptor

H-047 is the penultimate descriptor as per the descending order of importance (see **Figure 2D**). This atom-centered fragment descriptor is a measure of H-atom attached to the sp^3/sp^2 carbons in a molecule. Its positive regression coefficient indicates the linear correlation between the descriptor value and the toxicity. Mixtures from the studied dataset also align with this trend. For example, mixture of enoxacin and norfloxacin (**Mixture M28**) has less H-047 value (11.47) as well as toxicity [Log (1/IC50) = 6.16] as compared to the mixture of ciprofloxacin and marbofloxacin (**Mixture M05**) which has greater H-047 value (13.08) and toxicity [Log (1/IC50) = 7.28]. In **Figure 4.52**, we have analysed the structural components of both Mixturs **M28** and **M05** for a better understanding of the descriptor.

4.6.4.6. Eta_F_A descriptor

Eta_F_A is a 1st generation electro-topochemical atom (ETA) index descriptor which positively contributes towards the prediction of the studied endpoint [Log (1/IC50)] to *E. coli*. It is a measure of heteroatoms and unsaturation present relative to the overall size of a molecule. According to the published report of Roy *et al*., Eta_F_A (η'_F) is calculated by employing the following equation 4.26 (Roy and Ghosh, 2003)-

$$\eta'_{F=}{}^{\eta_F}/_{N_v} \qquad (4.26)$$

where η_F is the functionality index which estimates the unsaturation and heteroatoms of a molecule whereas N_v is the total number of non-hydrogen atoms present in a molecule. For a better understanding of the descriptor, we have compared two studied mixtures (**M7** and **M57**) in **Figure 4.52**. **Mixture M7** is composed of ciprofloxacin and norfloxacin as the component chemicals whereas, **Mixture M57** contains marbofloxacin and sparfloxacin. The unsaturation count of both these mixtures are same; however, **Mixture M57** bears more heteroatoms (8 N, 3 F, 1 O) than **Mixture M7** (6 N, 2 F). Generally, the presence of electronegative atoms influences the formation of intermolecular H-bonds with the amino acid residues of cellular receptor/protein and therefore increases the toxic potential (Wang et al., 2020). A similar observation has been proven in this study also. The mixture toxicity is increased from lower heteroatoms containing **Mixture M7** [Eta_F_A = 0.938; Log (1/IC50) = 6.34] to greater heteroatoms containing **Mixture M57** [Eta_F_A = 0.999; Log (1/IC50) = 7.24].

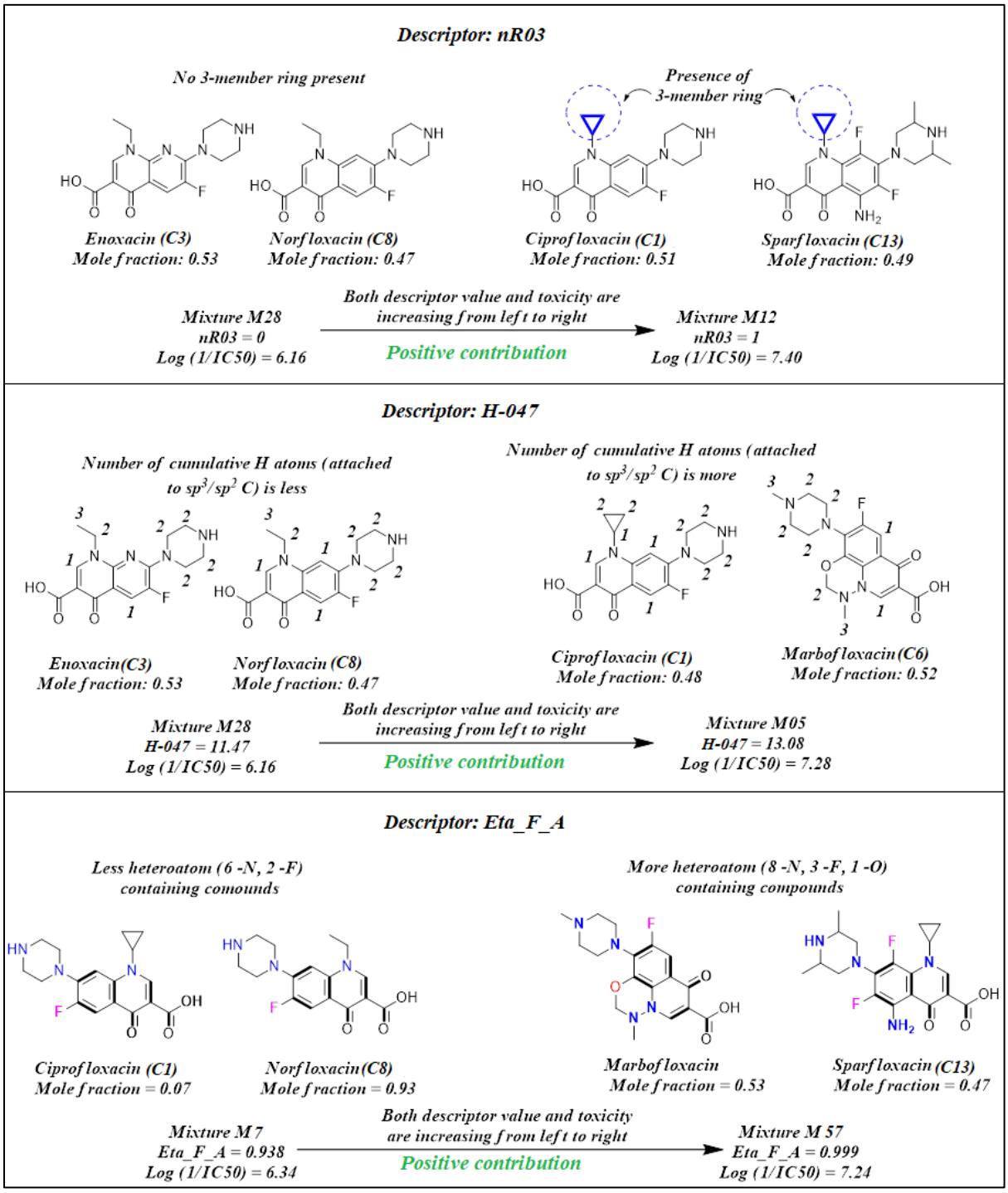


Figure 4.52. Significance of positive contributory descriptors towards toxicity

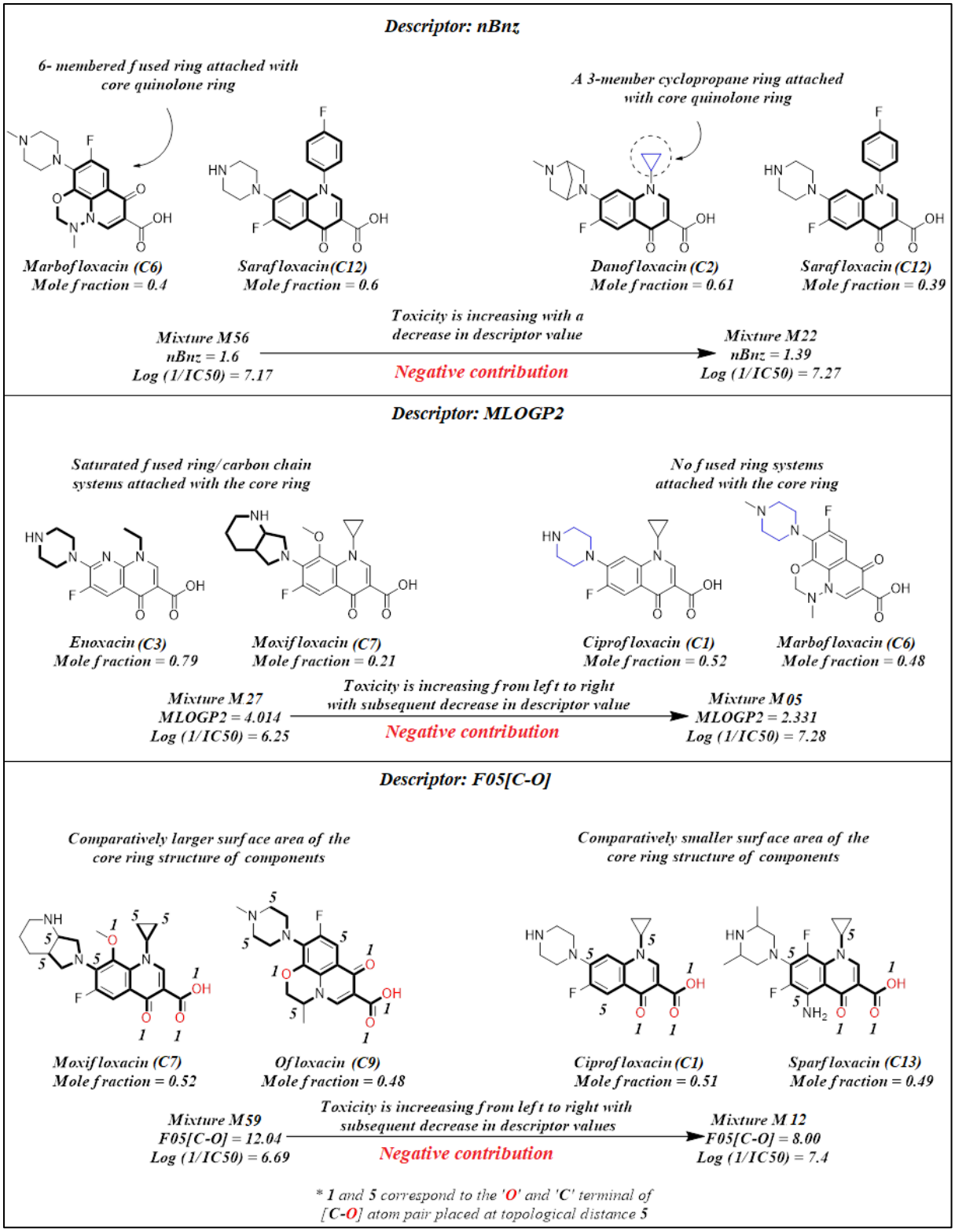


Figure 4.53. Significance of negative contributory descriptors towards toxicity

4.6.5. Applicability domain analysis

The applicability domain (AD) is the theoretical region in descriptor space where the model-derived predictions become reliable (Gadaleta et al., 2016). The reliable predictions of all chemicals can not be derived from a single model and that is why the knowledge of AD

should be provided with a newly developed model. OECD also suggests the same as a form of OECD principle 3 (a defined domain of applicability) (OECD, 2004). To comply with the norm, we have also analyzed the domain of applicability of the developed PLS-based mixture QSAR model in this study. The DModX (distance to model in X-space) approach has been employed in SIMCA-P software to derive the AD of the PLS model. We have given the AD plots in **Figure 4.54**. No outliers have been found (in both training and test set) in 99% confidence interval (D-Crit value = 0.00999898) which again strengthen the performance claim of the new model.

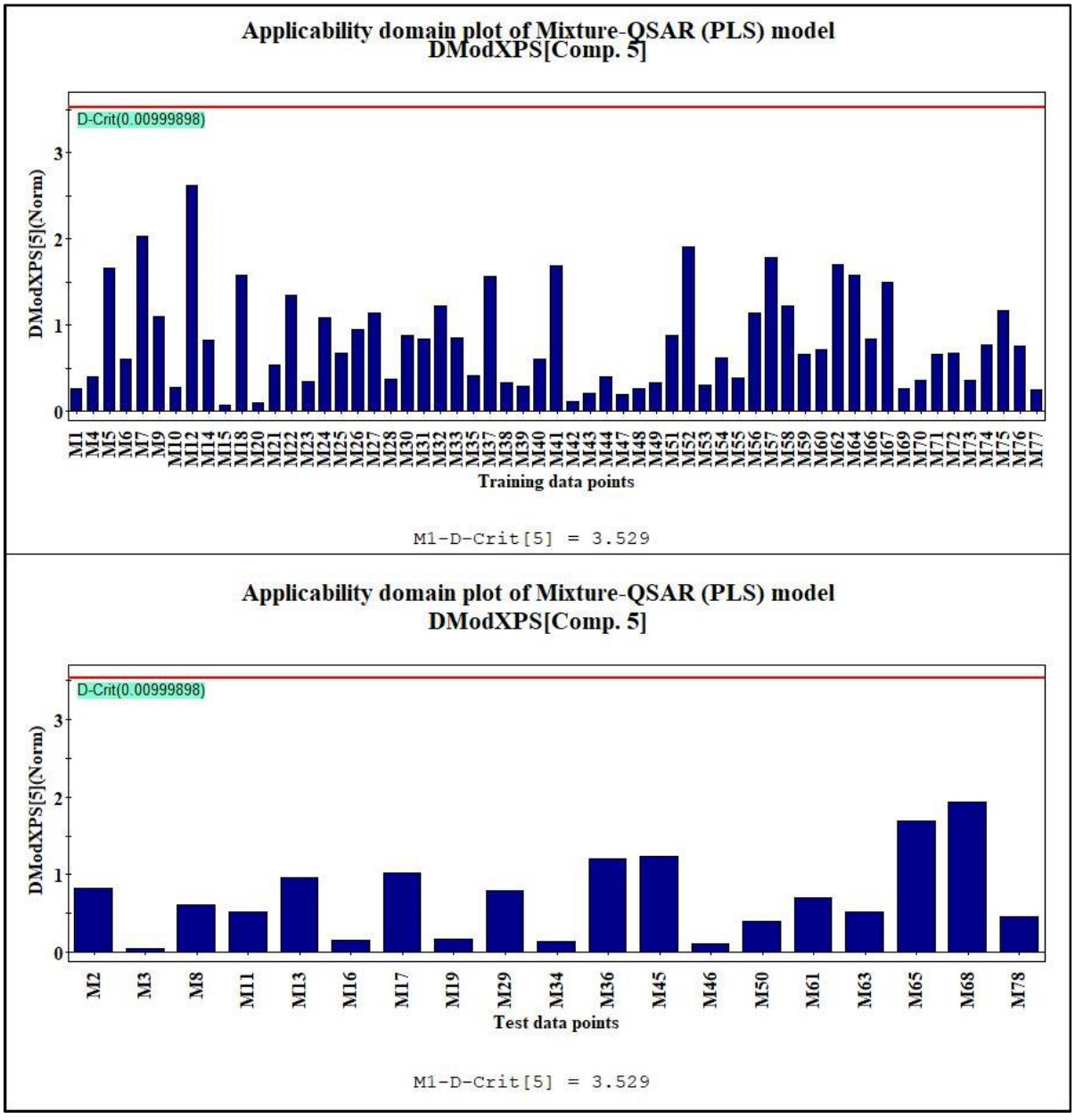


Figure 4.54. Applicability domain plot of developed Mixture-QSAR (PLS) model

4.6.6. Predicted responses of untested mixtures and the result of reliability analysis

We have predicted the toxicity of 405 hypothetical mixtures (designed from the parent FQs and their probable photodegradation products) with the developed PLS model employing the PRI tool. The predicted responses have been tabulated in **Table 4.36** along with their

respective reliability of prediction and AD status. The toxicity of 354 mixtures (87.41%), 42 mixtures (10.37%), and 9 mixtures (2.22%) has been predicted respectively with "Good", "Moderate", and "Poor" reliability. Among these mixtures, 86.67% fall under the domain of applicability of the developed PLS model whereas, 13.33% of mixtures have been identified as outliers. This analysis once again proved the efficiency of the developed model for predicting the toxicity of untested mixtures. We have also graded the response as highly toxic and less toxic after keeping the observed response of prototype FQ ciprofloxacin [Log (1/IC50) = 7.30 Mol/L] as the threshold. According to this grading procedure, 19.75% of mixtures have been classified as highly toxic whereas, the rest have come under the less toxic category.

Table 4. Toxicity prediction of hypothetical binary mixtures along with their reliability of prediction and AD status

Hypothetical Mixtures	Predicted response	Prediction Quality	AD status	Toxic class
HM1	7.277	Good	In	Less toxic
HM2	7.342	Good	In	Highly toxic
HM3	7.313	Good	In	Highly toxic
HM4	7.280	Good	In	Less toxic
HM5	7.288	Good	In	Less toxic
HM6	7.219	Good	In	Less toxic
HM7	7.500	Moderate	In	Highly toxic
HM8	7.250	Good	In	Less toxic
HM9	7.443	Moderate	In	Highly toxic
HM10	7.308	Good	In	Highly toxic
HM11	7.278	Good	In	Less toxic
HM12	7.246	Good	In	Less toxic
HM13	7.254	Good	In	Less toxic
HM14	7.185	Good	In	Less toxic
HM15	7.466	Moderate	In	Highly toxic
HM16	7.216	Good	In	Less toxic
HM17	7.409	Moderate	In	Highly toxic
HM18	7.343	Good	In	Highly toxic
HM19	7.310	Good	In	Highly toxic
HM20	7.319	Good	In	Highly toxic
HM21	7.249	Good	In	Less toxic
HM22	7.530	Moderate	In	Highly toxic
HM23	7.280	Good	In	Less toxic
HM24	7.474	Moderate	In	Highly toxic

HM25	7.281	Good	In	Less toxic
HM26	7.289	Good	In	Less toxic
HM27	7.220	Good	In	Less toxic
HM28	7.501	Moderate	In	Highly toxic
HM29	7.251	Good	In	Less toxic
HM30	7.444	Moderate	In	Highly toxic
HM31	7.257	Good	In	Less toxic
HM32	7.188	Good	In	Less toxic
HM33	7.469	Moderate	In	Highly toxic
HM34	7.219	Good	In	Less toxic
HM35	7.412	Moderate	In	Highly toxic
HM36	7.196	Good	In	Less toxic
HM37	7.477	Moderate	In	Highly toxic
HM38	7.227	Good	In	Less toxic
HM39	7.420	Moderate	In	Highly toxic
HM40	7.408	Moderate	In	Highly toxic
HM41	7.158	Good	In	Less toxic
HM42	7.351	Good	In	Highly toxic
HM43	7.439	Moderate	In	Highly toxic
HM44	7.632	Moderate	In	Highly toxic
HM45	7.382	Good	In	Highly toxic
HM46	7.012	Good	In	Less toxic
HM47	7.051	Good	In	Less toxic
HM48	7.008	Good	In	Less toxic
HM49	7.302	Good	In	Highly toxic
HM50	7.289	Good	In	Less toxic
HM51	7.281	Good	In	Less toxic
HM52	7.525	Moderate	In	Highly toxic
HM53	7.264	Good	In	Less toxic
HM54	7.482	Moderate	In	Highly toxic
HM55	7.042	Good	In	Less toxic
HM56	7.000	Good	In	Less toxic
HM57	7.293	Good	In	Less toxic
HM58	7.280	Good	In	Less toxic

HM59	7.273	Good	In	Less toxic
HM60	7.516	Moderate	In	Highly toxic
HM61	7.255	Good	In	Less toxic
HM62	7.474	Moderate	In	Highly toxic
HM63	7.038	Good	In	Less toxic
HM64	7.332	Good	In	Highly toxic
HM65	7.318	Good	In	Highly toxic
HM66	7.311	Good	In	Highly toxic
HM67	7.555	Moderate	In	Highly toxic
HM68	7.294	Good	In	Less toxic
HM69	7.512	Moderate	In	Highly toxic
HM70	7.289	Good	In	Less toxic
HM71	7.276	Good	In	Less toxic
HM72	7.268	Good	In	Less toxic
HM73	7.512	Moderate	In	Highly toxic
HM74	7.251	Good	In	Less toxic
HM75	7.469	Moderate	In	Highly toxic
HM76	7.570	Moderate	In	Highly toxic
HM77	7.562	Moderate	In	Highly toxic
HM78	7.806	Bad/Unreliable	In	Highly toxic
HM79	7.545	Moderate	In	Highly toxic
HM80	7.763	Bad/Unreliable	In	Highly toxic
HM81	7.549	Moderate	In	Highly toxic
HM82	7.792	Bad/Unreliable	In	Highly toxic
HM83	7.531	Moderate	In	Highly toxic
HM84	7.750	Bad/Unreliable	In	Highly toxic
HM85	7.785	Bad/Unreliable	In	Highly toxic
HM86	7.524	Moderate	In	Highly toxic
HM87	7.742	Bad/Unreliable	In	Highly toxic
HM88	7.768	Bad/Unreliable	In	Highly toxic
HM89	7.986	Bad/Unreliable	In	Highly toxic
HM90	7.725	Bad/Unreliable	In	Highly toxic
HM91	6.319	Good	In	Less toxic
HM92	6.409	Good	In	Less toxic

HM93	6.319	Good	In	Less toxic
HM94	6.348	Good	In	Less toxic
HM95	6.316	Good	In	Less toxic
HM96	6.309	Good	In	Less toxic
HM97	6.600	Good	In	Less toxic
HM98	6.300	Good	In	Less toxic
HM99	6.754	Good	In	Less toxic
HM100	6.397	Good	In	Less toxic
HM101	6.307	Good	In	Less toxic
HM102	6.336	Good	In	Less toxic
HM103	6.304	Good	In	Less toxic
HM104	6.296	Good	In	Less toxic
HM105	6.588	Good	In	Less toxic
HM106	6.287	Good	In	Less toxic
HM107	6.742	Good	In	Less toxic
HM108	6.397	Good	In	Less toxic
HM109	6.426	Good	In	Less toxic
HM110	6.394	Good	In	Less toxic
HM111	6.387	Good	In	Less toxic
HM112	6.678	Good	In	Less toxic
HM113	6.378	Good	In	Less toxic
HM114	6.832	Good	In	Less toxic
HM115	6.337	Good	In	Less toxic
HM116	6.304	Good	In	Less toxic
HM117	6.297	Good	In	Less toxic
HM118	6.588	Good	In	Less toxic
HM119	6.288	Good	In	Less toxic
HM120	6.742	Good	In	Less toxic
HM121	6.333	Good	In	Less toxic
HM122	6.326	Good	In	Less toxic
HM123	6.617	Good	In	Less toxic
HM124	6.317	Good	In	Less toxic
HM125	6.771	Good	In	Less toxic
HM126	6.294	Good	In	Less toxic

HM127	6.585	Good	In	Less toxic
HM128	6.285	Good	In	Less toxic
HM129	6.739	Good	In	Less toxic
HM130	6.578	Good	In	Less toxic
HM131	6.277	Good	In	Less toxic
HM132	6.732	Good	In	Less toxic
HM133	6.569	Good	In	Less toxic
HM134	7.023	Good	In	Less toxic
HM135	6.723	Good	In	Less toxic
HM136	7.115	Good	In	Less toxic
HM137	7.211	Good	In	Less toxic
HM138	7.111	Good	In	Less toxic
HM139	7.066	Good	In	Less toxic
HM140	7.117	Good	In	Less toxic
HM141	7.032	Good	In	Less toxic
HM142	7.370	Good	In	Highly toxic
HM143	7.104	Good	In	Less toxic
HM144	7.340	Good	In	Highly toxic
HM145	7.203	Good	In	Less toxic
HM146	7.102	Good	In	Less toxic
HM147	7.057	Good	In	Less toxic
HM148	7.108	Good	In	Less toxic
HM149	7.023	Good	In	Less toxic
HM150	7.361	Good	In	Highly toxic
HM151	7.096	Good	In	Less toxic
HM152	7.331	Good	In	Highly toxic
HM153	7.199	Good	In	Less toxic
HM154	7.153	Good	In	Less toxic
HM155	7.204	Good	In	Less toxic
HM156	7.119	Good	In	Less toxic
HM157	7.457	Moderate	In	Highly toxic
HM158	7.192	Good	In	Less toxic
HM159	7.427	Moderate	In	Highly toxic
HM160	7.053	Good	In	Less toxic

HM161	7.104	Good	In	Less toxic
HM162	7.019	Good	In	Less toxic
HM163	7.357	Good	In	Highly toxic
HM164	7.092	Good	In	Less toxic
HM165	7.327	Good	In	Highly toxic
HM166	7.059	Good	In	Less toxic
HM167	6.974	Good	In	Less toxic
HM168	7.312	Good	In	Highly toxic
HM169	7.046	Good	In	Less toxic
HM170	7.282	Good	In	Less toxic
HM171	7.025	Good	In	Less toxic
HM172	7.363	Good	In	Highly toxic
HM173	7.097	Good	In	Less toxic
HM174	7.333	Good	In	Highly toxic
HM175	7.278	Good	In	Less toxic
HM176	7.012	Good	In	Less toxic
HM177	7.248	Good	In	Less toxic
HM178	7.350	Good	In	Highly toxic
HM179	7.586	Moderate	In	Highly toxic
HM180	7.321	Good	In	Highly toxic
HM181	6.539	Good	In	Less toxic
HM182	6.541	Good	In	Less toxic
HM183	6.641	Good	In	Less toxic
HM184	6.581	Good	In	Less toxic
HM185	6.613	Good	In	Less toxic
HM186	6.477	Good	In	Less toxic
HM187	6.806	Good	In	Less toxic
HM188	6.516	Good	In	Less toxic
HM189	6.705	Good	In	Less toxic
HM190	6.445	Good	Outside AD	Less toxic
HM191	6.544	Good	In	Less toxic
HM192	6.484	Good	In	Less toxic
HM193	6.517	Good	In	Less toxic
HM194	6.380	Good	In	Less toxic

HM195	6.710	Good	In	Less toxic
HM196	6.420	Good	In	Less toxic
HM197	6.608	Good	In	Less toxic
HM198	6.547	Good	Outside AD	Less toxic
HM199	6.487	Good	In	Less toxic
HM200	6.519	Good	In	Less toxic
HM201	6.383	Good	In	Less toxic
HM202	6.712	Good	Outside AD	Less toxic
HM203	6.422	Good	In	Less toxic
HM204	6.610	Good	Outside AD	Less toxic
HM205	6.586	Good	In	Less toxic
HM206	6.619	Good	In	Less toxic
HM207	6.482	Good	In	Less toxic
HM208	6.811	Good	Outside AD	Less toxic
HM209	6.521	Good	In	Less toxic
HM210	6.710	Good	In	Less toxic
HM211	6.559	Good	In	Less toxic
HM212	6.422	Good	In	Less toxic
HM213	6.751	Good	In	Less toxic
HM214	6.461	Good	In	Less toxic
HM215	6.650	Good	In	Less toxic
HM216	6.455	Good	In	Less toxic
HM217	6.784	Good	In	Less toxic
HM218	6.494	Good	In	Less toxic
HM219	6.682	Good	In	Less toxic
HM220	6.647	Good	In	Less toxic
HM221	6.357	Good	In	Less toxic
HM222	6.546	Good	In	Less toxic
HM223	6.687	Good	In	Less toxic
HM224	6.875	Good	Outside AD	Less toxic
HM225	6.585	Good	In	Less toxic
HM226	7.129	Good	In	Less toxic
HM227	7.153	Good	In	Less toxic
HM228	7.218	Good	In	Less toxic

HM229	7.187	Good	In	Less toxic
HM230	7.197	Good	In	Less toxic
HM231	7.089	Good	In	Less toxic
HM232	7.397	Moderate	In	Highly toxic
HM233	7.106	Good	In	Less toxic
HM234	7.302	Good	In	Highly toxic
HM235	7.063	Good	In	Less toxic
HM236	7.128	Good	In	Less toxic
HM237	7.097	Good	In	Less toxic
HM238	7.107	Good	In	Less toxic
HM239	6.999	Good	In	Less toxic
HM240	7.307	Good	In	Highly toxic
HM241	7.016	Good	In	Less toxic
HM242	7.211	Good	In	Less toxic
HM243	7.152	Good	In	Less toxic
HM244	7.121	Good	In	Less toxic
HM245	7.131	Good	In	Less toxic
HM246	7.023	Good	In	Less toxic
HM247	7.331	Good	Outside AD	Highly toxic
HM248	7.040	Good	In	Less toxic
HM249	7.235	Good	In	Less toxic
HM250	7.186	Good	In	Less toxic
HM251	7.196	Good	In	Less toxic
HM252	7.088	Good	In	Less toxic
HM253	7.396	Moderate	In	Highly toxic
HM254	7.105	Good	In	Less toxic
HM255	7.300	Good	In	Highly toxic
HM256	7.165	Good	In	Less toxic
HM257	7.057	Good	In	Less toxic
HM258	7.365	Good	In	Highly toxic
HM259	7.074	Good	In	Less toxic
HM260	7.269	Good	In	Less toxic
HM261	7.068	Good	In	Less toxic
HM262	7.375	Good	Outside AD	Highly toxic

HM263	7.084	Good	In	Less toxic
HM264	7.280	Good	In	Less toxic
HM265	7.267	Good	In	Less toxic
HM266	6.977	Good	In	Less toxic
HM267	7.172	Good	In	Less toxic
HM268	7.284	Good	In	Less toxic
HM269	7.480	Moderate	Outside AD	Highly toxic
HM270	7.189	Good	In	Less toxic
HM271	6.240	Good	In	Less toxic
HM272	6.289	Good	In	Less toxic
HM273	6.294	Good	In	Less toxic
HM274	6.194	Good	In	Less toxic
HM275	6.267	Good	In	Less toxic
HM276	6.141	Good	In	Less toxic
HM277	6.466	Good	In	Less toxic
HM278	6.216	Good	In	Less toxic
HM279	6.396	Good	In	Less toxic
HM280	6.242	Good	In	Less toxic
HM281	6.246	Good	In	Less toxic
HM282	6.146	Good	In	Less toxic
HM283	6.220	Good	In	Less toxic
HM284	6.094	Moderate	In	Less toxic
HM285	6.418	Good	In	Less toxic
HM286	6.169	Good	In	Less toxic
HM287	6.348	Good	In	Less toxic
HM288	6.295	Good	In	Less toxic
HM289	6.195	Good	In	Less toxic
HM290	6.269	Good	In	Less toxic
HM291	6.143	Good	In	Less toxic
HM292	6.467	Good	In	Less toxic
HM293	6.217	Good	In	Less toxic
HM294	6.397	Good	In	Less toxic
HM295	6.200	Good	In	Less toxic
HM296	6.273	Good	In	Less toxic

HM297	6.147	Good	In	Less toxic
HM298	6.472	Good	In	Less toxic
HM299	6.222	Good	In	Less toxic
HM300	6.402	Good	In	Less toxic
HM301	6.173	Good	In	Less toxic
HM302	6.047	Moderate	In	Less toxic
HM303	6.371	Good	In	Less toxic
HM304	6.122	Good	In	Less toxic
HM305	6.301	Good	In	Less toxic
HM306	6.121	Good	In	Less toxic
HM307	6.445	Good	In	Less toxic
HM308	6.196	Good	In	Less toxic
HM309	6.375	Good	In	Less toxic
HM310	6.319	Good	In	Less toxic
HM311	6.069	Moderate	In	Less toxic
HM312	6.249	Good	In	Less toxic
HM313	6.394	Good	In	Less toxic
HM314	6.573	Good	In	Less toxic
HM315	6.324	Good	In	Less toxic
HM316	6.530	Good	In	Less toxic
HM317	6.595	Good	In	Less toxic
HM318	6.562	Good	In	Less toxic
HM319	6.418	Good	In	Less toxic
HM320	6.554	Good	In	Less toxic
HM321	6.380	Good	In	Less toxic
HM322	6.777	Good	In	Less toxic
HM323	6.516	Good	In	Less toxic
HM324	6.722	Good	In	Less toxic
HM325	6.561	Good	In	Less toxic
HM326	6.528	Good	In	Less toxic
HM327	6.384	Good	In	Less toxic
HM328	6.520	Good	In	Less toxic
HM329	6.346	Good	In	Less toxic
HM330	6.743	Good	In	Less toxic

HM331	6.482	Good	In	Less toxic
HM332	6.688	Good	In	Less toxic
HM333	6.593	Good	In	Less toxic
HM334	6.449	Good	In	Less toxic
HM335	6.585	Good	In	Less toxic
HM336	6.411	Good	In	Less toxic
HM337	6.809	Good	In	Less toxic
HM338	6.547	Good	In	Less toxic
HM339	6.753	Good	In	Less toxic
HM340	6.416	Good	In	Less toxic
HM341	6.552	Good	In	Less toxic
HM342	6.378	Good	In	Less toxic
HM343	6.775	Good	In	Less toxic
HM344	6.514	Good	In	Less toxic
HM345	6.720	Good	In	Less toxic
HM346	6.407	Good	In	Less toxic
HM347	6.234	Good	In	Less toxic
HM348	6.631	Good	In	Less toxic
HM349	6.370	Good	In	Less toxic
HM350	6.576	Good	In	Less toxic
HM351	6.369	Good	In	Less toxic
HM352	6.767	Good	In	Less toxic
HM353	6.505	Good	In	Less toxic
HM354	6.711	Good	In	Less toxic
HM355	6.593	Good	In	Less toxic
HM356	6.332	Good	In	Less toxic
HM357	6.538	Good	In	Less toxic
HM358	6.729	Good	In	Less toxic
HM359	6.935	Good	In	Less toxic
HM360	6.674	Good	In	Less toxic
HM361	7.113	Good	Outside AD	Less toxic
HM362	7.268	Good	Outside AD	Less toxic
HM363	7.048	Good	Outside AD	Less toxic
HM364	7.075	Good	Outside AD	Less toxic

HM365	7.134	Good	Outside AD	Less toxic
HM366	7.017	Good	Outside AD	Less toxic
HM367	7.390	Moderate	Outside AD	Highly toxic
HM368	7.071	Good	Outside AD	Less toxic
HM369	7.320	Good	Outside AD	Highly toxic
HM370	7.208	Good	Outside AD	Less toxic
HM371	6.987	Good	Outside AD	Less toxic
HM372	7.014	Good	Outside AD	Less toxic
HM373	7.073	Good	Outside AD	Less toxic
HM374	6.956	Good	Outside AD	Less toxic
HM375	7.329	Good	Outside AD	Highly toxic
HM376	7.011	Good	Outside AD	Less toxic
HM377	7.259	Good	Outside AD	Less toxic
HM378	7.142	Good	Outside AD	Less toxic
HM379	7.170	Good	Outside AD	Less toxic
HM380	7.229	Good	Outside AD	Less toxic
HM381	7.112	Good	Outside AD	Less toxic
HM382	7.485	Moderate	Outside AD	Highly toxic
HM383	7.166	Good	Outside AD	Less toxic
HM384	7.414	Moderate	Outside AD	Highly toxic
HM385	6.949	Good	Outside AD	Less toxic
HM386	7.008	Good	Outside AD	Less toxic
HM387	6.891	Good	Outside AD	Less toxic
HM388	7.264	Good	Outside AD	Less toxic
HM389	6.945	Good	Outside AD	Less toxic
HM390	7.193	Good	Outside AD	Less toxic
HM391	7.036	Good	Outside AD	Less toxic
HM392	6.918	Good	Outside AD	Less toxic
HM393	7.292	Good	Outside AD	Less toxic
HM394	6.973	Good	Outside AD	Less toxic
HM395	7.221	Good	Outside AD	Less toxic
HM396	6.977	Good	Outside AD	Less toxic
HM397	7.350	Good	Outside AD	Highly toxic
HM398	7.032	Good	Outside AD	Less toxic

HM399	7.280	Good	Outside AD	Less toxic
HM400	7.233	Good	Outside AD	Less toxic
HM401	6.915	Good	Outside AD	Less toxic
HM402	7.163	Good	Outside AD	Less toxic
HM403	7.288	Good	Outside AD	Less toxic
HM404	7.536	Moderate	Outside AD	Highly toxic
HM405	7.217	Good	Outside AD	Less toxic

4.6.7. Comparison of models

We have compared the modeling approaches as well as the performance metrics of this study with a previously published work by Wang and coworkers (Wang et al., 2020). The same toxicity dataset was used in both of these works; however, we have changed the training-test division in the current work. All the models have been developed with 59 training compounds in this study, unlike the 63 training compounds used in the previous work. Besides this, more efficient PLS regression and ML-based modeling approaches have been used in this study instead of the simple linear regression of the previous model. As per the reported validation metrics by Wang et al., the previous linear regression was a robust model with sufficient goodness of fit and moderate external predictability (**Table 4.37**). The present study has achieved better performance in all aspects of the previous model. Both the PLS and XGB models have shown better R^2 (PLS = 0.882, XGB = 0.885), Q^2_{LOO} (PLS = 0.849, XGB = 0.858), and Q^2_{F1} (PLS = 0.859, XGB = 0.858) values than the previous model (R^2 = 0.824, Q^2_{LOO} = 0.778, Q^2_{F1} = 697). The difference between R^2 and Q^2_{LOO} (an indicator of robustness) is less in the PLS (0.882-0.849 = 0.033) and XGB (0885-0.858 = 0.027) models of this study as compared to the linear regression model (0.824-0.778 = 0.046) of the previous work. However, the XGB model has shown greater error of predictions as evidenced by the $RMSE_C$ (0.336) and $RMSE_P$ (0.355) values, but the PLS model clearly outperforms the linear regression of the previous work in this context (PLS: $RMSE_C$ = 0.108, $RMSE_P$ = 0.112; Linear regression: $RMSE_C$ = 0.135, $RMSE_P$ = 0.156). Above all, the present models (PLS and ML) have been developed with simple and easily reproducible 0D-2D descriptors which make the model easily transferable as compared to the linear regression model obtained from the molecular docking-based energy descriptors (docking energy descriptor calculation involves non-reproducible and less transferable steps i.e., energy minimization and conformational analysis of the 3D structures of molecules). Finally, as a part of this work, we have also analyzed the effect of the structural features as well as the mole fraction of each component towards the toxicity of *E. coli* which helps the reader to graphically understand the additive nature of the studied mixtures.

Table 4.37. Comparative assessment of the performance of present models (PLS and XGB) with the previous work

Criteria and metrics of comparison	**Regression model of previous work by Wang *et al.*** (Wang et al., 2020)	**PLS regression model of the present work**	**ML model of the present work**
Dataset size	n_{Train} = 63, n_{Test} = 15	n_{Train} = 59, n_{Test} = 19	n_{Train} = 59, n_{Test} = 19
Modeled descriptors used	Molecular docking-based energy descriptor	2D descriptors	2D descriptors
Type of mixtures	Additive	Additive	Additive
Modeling algorithm	Linear regression	PLS	XGB
R^2	0.824	0.882	0.885
Q^2_{LOO}	0.778	0.849	0.858
$RMSE_C$	0.135	0.108	0.336
Q^2_{F1}	0.697	0.859	0.858
$RMSE_P$	0.156	0.112	0.355

CHAPTER - 5

CONCLUSIONS

Chapter 5: Conclusions

In any scientific exploration, success becomes embedded in the explanation of results obtained and the conclusion drawn from it. The results explain the hidden or unexplored scientific facts and the conclusion briefly discusses the new/unexplored information. Access to this new knowledge enables us to gain a deeper understanding of a specific domain of research. On top of that, results and descriptions are always helpful in clarifying concepts that have already been established and also help to pave the future avenues of research. In the present thesis, the development of in silico studies has been discussed giving special emphasis on the mixtures' toxicity prediction against several ecologically relevant endpoints. Toxicity data of various classes of chemical mixtures i.e., organic chemicals, pharmaceuticals, plant protection products, and different industrial chemicals have been employed. The structural and toxicity data have been curated and then used for the modeling. QSAR, q-RASAR, and ML models have been developed for better and eco-friendly environmental risk assessment of chemical mixtures. Read-across prediction methods have also been employed for the same purpose. The OECD principles of QSAR model development have been followed during the study to avoid any ambiguity in the results obtained. The important structural features (specific for each dataset) have been identified that are responsible for mixture toxicity. All the models have been strictly validated using conventional internal and external validations. Additionally, mixtures-out, compounds-out, and everything-out cross-validations have been employed to avoid overestimation of validation metrics (associated with mixture toxicity modeling). To check the reliability of external validation, the domains of applicability of the developed models have been carried out which also provides the knowledge about the chemical space of respective models. An effort of data gap filling has been made in this thesis where the toxicity of different hypothetical and/or real untested mixtures have been predicted against corresponding ecological species.

The objectives of the dissertation were the collection of experimental toxicity data of various chemical mixtures against different environmental endpoints, the development of QSAR models for predicting toxicity of different classes of chemical mixtures (pharmaceuticals, industrial chemicals, plant protection products, etc.), rigorous validation of the developed models in accordance with the OECD guidelines, identification of various structural features, responsible for the toxicity of chemical mixtures, make use of different other machine learning-based new approach methodologies such as read-across and q-RASAR for enhancing the performance of predictive model further. By keeping in mind the objectives, we have performed all the studies and compiled the results in this thesis. In a nutshell, an effort has been made in this thesis to provide a thorough understanding of (i) the conventional risk assessment of chemical mixtures, (ii) different computational modeling approaches, (iii) the development of the mixtures' QSAR/q-RASAR/ML models for toxicity prediction and risk assessment, (iv) identification of probable interactions among mixture components, (v) identification of structural features responsible for mixture toxicity, and (vi) on prediction of mixture toxicity of new/untested mixtures and subsequent data gap filling. The specific knowledge obtained from each study has been discussed separately in the following section.

5.1. Study 1: Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors

The prediction of toxicity of various individual chemical data sets and corresponding mixture data sets was accomplished by the PLS (partial least squares) regression method. The 2D

descriptors were successfully used to measure the contribution of structural features of individual chemicals as well as the components of mixtures to overall toxicity. In addition, the weighted descriptor generation strategy was successfully employed to calculate the mixture descriptors in this work. It has been found that the structural features associated with the toxicity of individual chemicals have no significant contribution to the toxicity of mixtures. The presence of multiple bonded Nitrogen and terminal sp^2 hybridized carbon atoms in mixture components can help to reduce the toxicity of binary mixtures against the marine bacterium. In the case of freshwater organisms, the molecular bulk of mixture components can enhance the toxicity. The statistical validation results show that the developed models are robust and highly predictive and not the results of any chance correlation as suggested by the Y-randomization test. We should mention that we could not include the data from all types of functional groups in this study due to the scarcity of toxicity data; however, these models can be used efficiently for the prediction of aquatic toxicity and risk assessment of the chemical mixtures under their domain of applicability.

5.2. Study 2: Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

The prediction of aquatic toxicity of 84 binary mixtures has been accomplished by the PLS regression method. Various 2D descriptors have been successfully used to identify the important structural features of mixture components and model the overall mixture toxicity. The weighted descriptor generation approach has also been utilized for the calculation of mixture descriptors. The Y-randomization test has also been successfully executed in this study which confirms the model as not a result of chance co-relation. It has been found that the lipophilicity, hydrogen bonding propensity, and solubility of the mixture components are the important features that affect the aquatic toxicity of mixtures of polar and non-polar narcotic toxicants. Overall, lipophilicity of mixture components enhances the toxicity of the mixtures; at the same time, the solubility of mixture components is important for the dissolution into the aquatic medium which indirectly promotes the toxicity. The presence of nitrogen atoms in the mixture component can reduce the toxicity, whereas the simultaneous occurrence of nitrogen and chlorine can enhance the toxicity. Various newly devised cross-validation strategies have also been employed to avoid the developmental bias and overestimation of the predictive ability of the model in this study. We should mention that the true external validation and everything-out cross-validation could not be performed in this study due to the non-availability of toxicity data against the particular endpoint; instead, the reliability of prediction of the newly designed empirical mixtures has been adjudged by the PRI tool and proved to be reliable. Finally, the prediction quality of the binary mixtures has been successfully improved by using a machine learning-based chemical read-across algorithm with the identified structural features. In a nutshell, the use of similarity-based QSAR with different cross-validation strategies and the machine learning-based chemical read-across algorithm as alternative approaches for mixture risk assessment has been successfully established in this study.

5.3. Study 3: Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against *Aliivibrio fischeri*

This research mainly focuses on the use of alternative in silico approaches (especially QSAR and chemical read-across) to the toxicity prediction of binary and multi-component mixtures. Chemical mixtures are no longer an extended reality in the perspective of environmental

components in this era of chemical advancement. After confirming the presence of pharmaceuticals and pesticides in wastewater streams, and by keeping in mind their probable effect on the environmental species, the study was designed with pharmaceutical and pesticide mixtures against the aquatic bacterium *Aliivibrio fischeri*. A PLS regression model was developed in this study with the 2D structural descriptors, which is proved as an efficient and reliable model. The use of 2D descriptors has sufficiently reduced the computation time and made the model easily transferable and interpretable. The weighted descriptor generation approach was used to impart the mixture effect in this study, and three cross-validation strategies were applied to counter the chances of overestimating validation metrics. The results of these cross-validations clearly suggested the efficiency of the developed model. The external validation has been performed with multicomponent mixtures containing test set and good model quality has been obtained which suggests the efficiency of the model in predicting multicomponent mixtures. This model successfully identified seven structural features that are responsible for the mixture's toxicity and also demonstrated their roles in toxicity enhancement. We compared the quality of the developed QSAR model with previous reports and found the better quality, fast applicability, and easy interpretability of our model. We applied a newly designed tool based on chemical read-across prediction in this study to check the relevance of modelled descriptors. The read-across predictions were found to be good with respect to quality measures and the quality of predictions was also reliable. In a nutshell, a mixture QSAR has been reported in this study with better quality of prediction, machine learning-based quantitative read-across predictions have been successfully used to mixture toxicity assessment with the descriptors obtained in the QSAR, and the relevance of the selected descriptors has been successfully established by the result of machine learning-based chemical read-across algorithm.

5.4. Study 4: Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

This research has explored the application of the q-RASAR modeling approach for the prediction of acute contact toxicity of binary mixtures in honey bees for the first time. 2D-structural descriptors of the mixture components have been used as the preliminary input. The mixture descriptors have been calculated by employing three different mixing rules from the descriptors of individual mixture components and their corresponding mass ratio. Based on the computed quality metrics of three PLS q-RASAR models (obtained with three classes of mixture descriptors), the 3rd model has been identified as the most predictive one. Therefore, mixing rule 3, i.e., the norm of the molecular contributions of mixture components has been selected as the best rule for calculating the mixture descriptors. The PLS q-RASAR model obtained with mixture descriptor 3 is a good fit with the training data, robust, and an extremely predictive one. Good classifiability of the model's predictions in toxic and non-toxic classes has also been established by performing a classification-based external validation with a true-external dataset in this study. Besides these, linear and non-linear ML approaches have been used to enhance the predictability of the developed model. A non-linear SVM-based q-RASAR was the most predictive model of this study in terms of Q^2_{F1} value (Q^2_{F1} = 0.935). We have compared the performance of the q-RASAR model with previously reported innovative QSAR, and the q-RASAR has been proven as a more robust and predictive approach for mixture toxicity assessment of organic mixtures in honey bees. Among the modeled descriptors, RASAR descriptors have a major contribution to enhance the predictability of the model. On the other hand, the structural descriptors have effectively pointed out a specific feature i.e., the presence of electronegative atoms which increases the toxicity of organic mixtures. The descriptors namely, B06[C-F], Eta_betaS_A, and NssssC

mostly indicate the presence of electronegative atoms and all of them contribute positively. Furthermore, the PLS and ML-based q-RASAR models have been employed to predict the acute contact toxicity of 218 untested binary mixtures in honey bees and the predicted responses have been reported along with their consensus (average) prediction. Here we thus attempted to fill the major data gap on binary mixture toxicity. In a nutshell, we proposed this approach as a new, efficient, reproducible, and easily transferable technique that can predict the acute contact toxicity of untested binary mixtures within the applicability domain with sufficient reliability. It can thus be used as a reliable integrative approach to the standard empirical approaches on pesticide mixture toxicity risk assessment.

5.5. Study 5: "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary mixtures towards three bacterial species

The present study involves toxicity modeling of chemical mixtures (both binary and ternary) for three bacterial endpoints (*V. fischeri, E. coli, and B. subtilis*) with QSAR and q-RASAR-based QSAAR (q-RASAAR) approaches. This study was designed to develop more efficient toxicity models for mixtures' risk assessment, to identify the best mixing rule for mixture descriptor calculation, to study the use of toxicity of separate species (as predictor) in mixture toxicity modeling, and to apply RASAR algorithm for refining the QSAAR models further. Three QSAR models have been developed by PLS regression and the important structural features have been identified which affect the mixtures' toxicity. Simple 0D-2D descriptors have been used for encoding the structural features in this study which makes it a cost-effective, less complex, reproducible, interpretable, and transferable approach. Three different mixing rules have been used in this study and the best has been identified (MD1: the squared molecular contributions, MD2: linear combination of molecular contributions, MD3: norm of molecular contributions). Six other interspecies QSAAR models (two QSAARs for each dataset) have also been developed using PLS regression with the predicted toxicity of other bacterial species as descriptors. Finally, three q-RASAAR models have been developed for predicting toxicity against three bacterial species. All these models have been strictly validated and the computed validation metrics have shown the acceptance of the models as per the OECD guidelines. The interspecies toxicity has not shown much impact on the models' quality in preliminary QSAARs (PLS). However, similar toxicity patterns have been observed in *V. fischeri* and *B. subtilis*. The RASAR algorithm has incorporated the structural (identified descriptors) and biological information (in the form of predicted toxicity of two other species) during the RASAR descriptor calculation and it significantly improves the performance of QSAARs. Among the three models, q-RASAAR-1 and q-RASAAR-3 have been developed with a single predictor variable *RA function (LK)* which is a prediction function obtained by weighted averaging of the response variables of similar structural analogues. Therefore, the RASAR algorithm enables the most effective incorporation of the underlined features (both biological and structural) responsible for mixtures' toxicity and thus produces a better quality prediction. In a nutshell, this is the first instance where the interspecies mixture toxicity models coupled with the q-RASAR algorithm have been developed which can predict the cellular toxicity of binary and ternary mixtures against any of the three above-mentioned organisms with sufficient efficiency.

5.6. Study 6: Predictive binary mixture toxicity modeling of fluoroquinolone (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: A machine learning approach

The present work represents the development of QSAR models for toxicity prediction of FQ mixtures to *E. coli* and the use of different linear and non-linear machine learning algorithms for model development. The models were made only with the binary mixtures' toxicity data. The conventional PLS algorithm and the XGB-based ML approach have shown promising efficiency in mixture toxicity prediction. We have successfully applied three different mixing rules to compute the mixture descriptors in this study and the linear combination of molecular contributions (additive effect) has been identified as the best one. Our model outperforms the previous one in terms of goodness of fit, robustness, and external predictivity. The use of 2D-structural descriptors made the model quite simple, reproducible, and easily transferable as compared to the more complicated counterparts. Six different 2D-structural features have been identified that successfully demonstrate the effect of structural features on the mixture toxicity of FQs. The presence of a tricyclic ring on 1st N of quinolone and a greater number of heteroatoms and unsaturation in the mixture component can be detrimental to environmental safety. On the contrary, the lipophilicity and the presence of more benzene/benzene-like ring structure in the mixture component can reduce the greater toxic effect. The toxic effect of different photodegradation products on mixture toxicity has also been assessed in this study. We have predicted the toxicity of 405 hypothetical mixtures (binary) which are empirically designed from 9 FQs and their corresponding photodegradation products. Among these 405 mixtures, 19.75% have shown a higher risk of being toxic towards *E. coli*. However, this projection cannot cover the entire chemical domain of the FQs and their different degradation products which makes us bound to perform more exhaustive work towards diverse ecological species for better risk assessment.

REFERENCES

References

Abbas, M., Adil, M., Ehtisham-ul-Haque, S., Munir, B., Yameen, M., Ghaffar, A., Shar, G.A., Asif Tahir, M., Iqbal, M., 2018. Vibrio fischeri bioluminescence inhibition assay for ecotoxicity assessment: A review. Sci. Total Environ. 626, 1295–1309. https://doi.org/10.1016/j.scitotenv.2018.01.066

Adhikari, C., Mishra, B. kumar, 2018. Quantitative Structure-Activity Relationships of Aquatic Narcosis: A Review. Curr. Comput. Aided. Drug Des. 14, 7–28. https://doi.org/10.2174/1573409913666170711130304

Agresti, A., 1996. An introduction to categorical data analysis. John Wiley & Sons, New Jersey.

Ajay, Walters, W.P., Murcko, M.A., 1998. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J. Med. Chem. 41, 3314–3324. https://doi.org/10.1021/jm970666c

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

Almaliki, Z.A., 2018. Bias-variance dilemma? [WWW Document]. Towar. Data Sci. URL https://towardsdatascience.com/bias-variance-dilemma-74e5f1f52b12 (accessed 10.19.22).

Altenburger, R., Ait-Aissa, S., Antczak, P., Backhaus, T., Barceló, D., Seiler, T.B., Brion, F., Busch, W., Chipman, K., de Alda, M.L., de Aragão Umbuzeiro, G., Escher, B.I., Falciani, F., Faust, M., Focks, A., Hilscherova, K., Hollender, J., Hollert, H., Jäger, F., Jahnke, A., Kortenkamp, A., Krauss, M., Lemkine, G.F., Munthe, J., Neumann, S., Schymanski, E.L., Scrimshaw, M., Segner, H., Slobodnik, J., Smedes, F., Kughathas, S., Teodorovic, I., Tindall, A.J., Tollefsen, K.E., Walz, K.H., Williams, T.D., Van den Brink, P.J., van Gils, J., Vrana, B., Zhang, X., Brack, W., 2015. Future water quality monitoring - Adapting tools to deal with mixtures of pollutants in water resource management. Sci. Total Environ. 512–513, 540–551. https://doi.org/10.1016/j.scitotenv.2014.12.057

Altenburger, R., Backhaus, T., Boedeker, W., Faust, M., Scholze, M., Grimme, L.H., 2000. Predictability of the toxicity of multiple chemical mixtures to Vibrio fischeri: Mixtures composed of similarly acting chemicals. Environ. Toxicol. Chem. 19, 2341–2347. https://doi.org/10.1002/etc.5620190926

Altenburger, R., Greco, W.R., 2009. Extrapolation Concepts for dealing with multiple contamination in environmental risk assessment. Integr. Environ. Assess. Manag. 5, 62–68. https://doi.org/10.1897/IEAM_2008-038.1

Altenburger, R., Nendza, M., Schüürmann, G., 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ. Toxicol. Chem. 22, 1900–1915. https://doi.org/10.1897/01-386

Altenburger, R., Schmitt, H., Schüürmann, G., 2005. Algal toxicity of nitrobenzenes:

combined effect analysis as a pharmacological probe for similar modes of interaction. Environ. Toxicol. Chem. 24, 324.

Altenburger, R., Walter, H., Grote, M., 2004. What contributes to the combined effect of a complex mixture? Environ. Sci. Technol. 38, 6353–6362. https://doi.org/10.1021/es049528k

Álvarez-Muñoz, D., Rodríguez-Mozaz, S., Maulvault, A.L., Tediosi, A., Fernández-Tejedor, M., Van den Heuvel, F., Kotterman, M., Marques, A., Barceló, D., 2015. Occurrence of pharmaceuticals and endocrine disrupting compounds in macroalgaes, bivalves, and fish from coastal areas in Europe. Environ. Res. 143, 56–64. https://doi.org/10.1016/j.envres.2015.09.018

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Comput. 9, 1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545

Andrea, T.A., Kalayeh, H., 1991. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. J. Med. Chem. 34, 2824–2836. https://doi.org/10.1021/jm00113a022

Andres, C., Hutter, M.C., 2006. CNS permeability of drugs predicted by a decision tree. QSAR Comb. Sci. 25, 305–309. https://doi.org/10.1002/QSAR.200510200

Aptula, A.O., Jeliazkova, N.G., Schultz, T.W., Cronin, M.T.D., 2005. The better predictive model: High q2 for the training set or low root mean square error of prediction for the test set? QSAR Comb. Sci. 24, 385–396. https://doi.org/10.1002/QSAR.200430909

Babić, S., Periša, M., Škorić, I., 2013. Photolytic degradation of norfloxacin, enrofloxacin and ciprofloxacin in various aqueous media. Chemosphere 91, 1635–1642. https://doi.org/10.1016/j.chemosphere.2012.12.072

Baek, I.H., Kim, Y., Baik, S., Kim, J., 2019. Investigation of the synergistic toxicity of binary mixtures of pesticides and pharmaceuticals on AliiVibrio fischeri in major river basins in South Korea. Int. J. Environ. Res. Public Health 16, 208. https://doi.org/10.3390/ijerph16020208

Balls, M., Blaauboer, B.J., Fentem, J.H., Bruner, L., Combes, R.D., Ekwall, B., Fielder, R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski, D., Spielmann, H., Zucco, F., 1995. Practial aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM workshop 5. Altern. to Lab. Anim. 23, 129–147.

Banerjee, A., Chatterjee, M., De, P., Roy, K., 2022. Quantitative predictions from chemical read-across and their confidence measures. Chemom. Intell. Lab. Syst. 227, 104613. https://doi.org/10.1016/J.CHEMOLAB.2022.104613

Banerjee, A., Gajewicz-Skretna, A., Roy, K., 2023. A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites. Mol. Inform. 42, 2200261. https://doi.org/10.1002/minf.202200261

Banerjee, A., Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. Chem. Res. Toxicol. 36, 446–464. https://doi.org/10.1021/acs.chemrestox.2c00374

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

Bilal, M., Iqbal, H.M.N., Barceló, D., 2019a. Persistence of pesticides-based contaminants in the environment and their effective degradation using laccase-assisted biocatalytic systems. Sci. Total Environ. 695, 133896. https://doi.org/10.1016/J.SCITOTENV.2019.133896

Bilal, M., Rasheed, T., Nabeel, F., Iqbal, H.M.N., Zhao, Y., 2019b. Hazardous contaminants in the environment and their laccase-assisted degradation – A review. J. Environ. Manage. 234, 253–264. https://doi.org/10.1016/J.JENVMAN.2019.01.001

Bliss, C.I., 1939. The toxicity of poisons applied jointly. Ann. Appl. Biol. 26, 585–615. https://doi.org/10.1111/j.1744-7348.1939.tb06990.x

Borman, S., 1990. New QSAR techniques eyed for environmental assessments. Chem. Eng. News 19, 20–23.

Brack, W., Altenburger, R., Ensenbach, U., Möder, M., Segner, H., Schüürmann, G., 1999. Bioassay-directed identification of organic toxicants in river sediment in the industrial region of Bitterfeld (Germany) - A contribution to hazard assessment. Arch. Environ. Contam. Toxicol. 37, 164–174. https://doi.org/10.1007/s002449900502

Brack, W., Altenburger, R., Schüürmann, G., Krauss, M., López Herráez, D., van Gils, J., Slobodnik, J., Munthe, J., Gawlik, B.M., van Wezel, A., Schriks, M., Hollender, J., Tollefsen, K.E., Mekenyan, O., Dimitrov, S., Bunke, D., Cousins, I., Posthuma, L., van den Brink, P.J., López de Alda, M., Barceló, D., Faust, M., Kortenkamp, A., Scrimshaw, M., Ignatova, S., Engelen, G., Massmann, G., Lemkine, G., Teodorovic, I., Walz, K.H., Dulio, V., Jonker, M.T.O., Jäger, F., Chipman, K., Falciani, F., Liska, I., Rooke, D., Zhang, X., Hollert, H., Vrana, B., Hilscherova, K., Kramer, K., Neumann, S., Hammerbacher, R., Backhaus, T., Mack, J., Segner, H., Escher, B., de Aragão Umbuzeiro, G., 2015. The SOLUTIONS project: Challenges and responses for present and future emerging pollutants in land and water resources management. Sci. Total Environ. 503–504, 22–31. https://doi.org/10.1016/j.scitotenv.2014.05.143

Brown, A.C., Fraser, T.R., 1868. On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. J. Anat. Physiol. 2, 224. https://doi.org/10.1017/S0080456800028155

Bruice, T.C., Kharasch, N., Winzler, R.J., 1956. A correlation of thyroxine-like activity and chemical structure. Arch. Biochem. Biophys. 62, 305–317. https://doi.org/10.1016/0003-9861(56)90129-1

Burbidge, R., Trotter, M., Buxton, B., Holden, S., 2001. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. Comput. Chem. 26, 5–14. https://doi.org/10.1016/S0097-8485(01)00094-8

Bures, M.S., Ukic, S., Cvetnic, M., Prevaric, V., Markic, M., Rogosic, M., Kusic, H., Bolanca, T., 2021. Toxicity of binary mixtures of pesticides and pharmaceuticals toward Vibrio fischeri: Assessment by quantitative structure-activity relationships. Environ.

Pollut. 275, 115885. https://doi.org/10.1016/j.envpol.2020.115885

Burnham, K.P., Anderson, D.R., 2002. Information and Likelihood Theory: A Basis for Model Selection and Inference, in: Burnham, K.P., Anderson, D.R. (Eds.), Model Selection and Multimodel Inference. Springer, New York, NY, pp. 49–97. https://doi.org/10.1007/978-0-387-22456-5_2

Byvatov, E., Fechner, U., Sadowski, J., Schneider, G., 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J. Chem. Inf. Comput. Sci. 43, 1882–1889. https://doi.org/10.1021/ci0341161

Carley, M., Christie, I., 2019. The Ecology of an Industrial Planet, in: Managing Sustainable Development. Routledge, London, pp. 3–24.

Carlson, R., Carlson, J.E., 1992. Data handling in science and technology, in: Design and Optimization in Organic Synthesis. Elsevier, Amsterdam.

Carnesecchi, E., Svendsen, C., Lasagni, S., Grech, A., Quignot, N., Amzal, B., Toma, C., Tosi, S., Rortais, A., Cortinas-Abrahantes, J., Capri, E., Kramer, N., Benfenati, E., Spurgeon, D., Guillot, G., Dorne, J.L.C.M., 2019. Investigating combined toxicity of binary mixtures in bees: Meta-analysis of laboratory tests, modelling, mechanistic basis and implications for risk assessment. Environ. Int. 133, 105256. https://doi.org/10.1016/j.envint.2019.105256

Carnesecchi, E., Toropov, A.A., Toropova, A.P., Kramer, N., Svendsen, C., Dorne, J. Lou, Benfenati, E., 2020. Predicting acute contact toxicity of organic binary mixtures in honey bees (A. mellifera) through innovative QSAR models. Sci. Total Environ. 704, 135302. https://doi.org/10.1016/j.scitotenv.2019.135302

Cassee, F.R., Groten, J.P., Van Bladeren, P.J., Feron, V.J., 1998. Toxicological evaluation and risk assessment of chemical mixtures. Crit. Rev. Toxicol. 28, 73–101.

Cedergreen, N., Sørensen, H., Svendsen, C., 2012. Can the joint effect of ternary mixtures be predicted from binary mixture toxicity results? Sci. Total Environ. 427–428, 229–237. https://doi.org/10.1016/j.scitotenv.2012.03.086

Charton, M., 1969. Nature of the ortho effect. II. Composition of the Taft steric parameters. J. Am. Chem. Soc. 91, 615–618.

Chase, R.J., Harrison, D.R., Lackmann, G.M., McGovern, A., 2023. A machine learning tutorial for operational meteorology, part II: Neural networks and deep learning. Weather Forecast. 1271–1293. https://doi.org/10.1175/waf-d-22-0187.1

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., Roy, K., 2023. "Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary antibiotic mixtures toward three bacterial species. J. Hazard. Mater. 459, 132129. https://doi.org/10.1016/j.jhazmat.2023.132129

Chatterjee, M., Roy, K., 2022a. Computational Modeling of Mixture Toxicity, in: Emilio

Benfenati (Ed.), In Silico Methods for Predicting Drug Toxicity. Humana Press Inc., New York, pp. 561–587. https://doi.org/10.1007/978-1-0716-1960-5_22

Chatterjee, M., Roy, K., 2022b. Recent advances on modelling the toxicity of environmental pollutants for risk assessment: from single pollutants to mixtures. Curr. Pollut. Reports 8, 81–97. https://doi.org/10.1007/s40726-022-00219-6

Chatterjee, M., Roy, K., 2022c. 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

Chatterjee, M., Roy, K., 2022d. 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

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

ChemAxon, 2018. Marvin [WWW Document]. URL https://chemaxon.com/products/marvin (accessed 10.10.22).

ChemAxon, 2013. Marvin Sketch.

Chen, J., Wang, C., Sun, Y., Shen, X., 2011. Semi-supervised Laplacian regularized least squares algorithm for localization in wireless sensor networks. Comput. Networks 55, 2481–2491. https://doi.org/10.1016/J.COMNET.2011.04.010

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chirico, N., 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. J. Chem. Inf. Model. 52, 2044–2058. https://doi.org/10.1021/ci300084j

Chung-Yuan, C., Ch'i-Feng, H., 1996. Toxicity of organic mixtures containing cyanogenic toxicants. Environ. Toxicol. Chem. 15, 1464–1469. https://doi.org/10.1002/etc.5620150906

Cipullo, S., Snapir, B., Prpich, G., Campo, P., Coulon, F., 2019. Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. Chemosphere 215, 388–395.

Cizmas, L., Sharma, V.K., Gray, C.M., McDonald, T.J., 2015. Pharmaceuticals and personal care products in waters: occurrence, toxicity, and risk. Environ. Chem. Lett. 13, 381–394. https://doi.org/10.1007/s10311-015-0524-4

Consonni, V., Todeschini, R., 2010. Molecular Descriptors, in: Puzin, T., Leszczynski, J., Cronin, M. (Eds.), Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics. https://doi.org/10.1007/978-1-4020-9783-6_3

Cramer, R.D., 1993. Partial least squares (PLS): Its strength and limitations. Perspect. Drug Discov. Des. 1, 169–178. https://doi.org/10.1007/bf02174528

Cramer, R.D., Patterson, D.E., Bunce, J.D., 1988. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. 110, 5959–5967. https://doi.org/https://doi.org/10.1021/ja00226a005

Darlington, R.B., 1990. Regression and linear models. McGraw-Hill, NewYork.

Data Analysis, Statistical & Process Improvement Tools | Minitab, 2004.

Daughton, C.G., 2003. Cradle-to-cradle stewardship of drugs for minimizing their environmental disposition while promoting human health. I. Rational for and avenues toward a green pharmacy. Environ. Health Perspect. 111, 757–774. https://doi.org/10.1289/ehp.5947

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 Adv. 8, 4662–4670. https://doi.org/10.1039/c7ra13159c

De Zwart, D., Slooff, W., 1983. The Microtox as an alternative assay in the acute toxicity assessment of water pollutants. Aquat. Toxicol. 4, 129–138. https://doi.org/10.1016/0166-445X(83)90050-4

Dimitrov, S., Mekenyan, O., 2010. An introduction to read-across for the prediction of the effects of chemicals, in: In Silico Toxicology: Principles and Applications. The Royal Society of Chemistry, UK, pp. 372–384.

Dorigo, M., Stutzle, T., 2004. Ant colony optimization. The MIT Press, Cambridge.

Dsikowitzky, L., Schwarzbauer, J., 2014. Industrial organic contaminants: Identification, toxicity and fate in the environment. Environ. Chem. Lett. 12, 371–386. https://doi.org/10.1007/s10311-014-0467-1

E. Ite, A., A. Harry, T., O. Obadimu, C., R. Asuaiko, E., J. Inim, I., 2018. Petroleum hydrocarbons contamination of surface water and groundwater in the Niger delta region of Nigeria. J. Environ. Pollut. Hum. Heal. 6, 51–61. https://doi.org/10.12691/jephh-6-2-2

Elisabeth, H.M., Aude, K., Maurice, W., Andrew, W., 2014. EURL ECVAM strategy to replace, reduce and refine the use of fish in aquatic toxicity and bioaccumulation testing. Publications Office of the European Union. https://doi.org/10.2788/084219

Elliott, S.M., Brigham, M.E., Kiesling, R.L., Schoenfuss, H.L., Jorgenson, Z.G., 2018. Environmentally relevant chemical mixtures of concern in waters of United States tributaries to the Great Lakes. Integr. Environ. Assess. Manag. 14, 509–518. https://doi.org/10.1002/ieam.4041

Ergon, R., 2004. Informative PLS score-loading plots for process understanding and monitoring. J. Process Control 14, 889–897. https://doi.org/10.1016/j.jprocont.2004.02.004

European Commision, 2003. Technical guidance document on risk assessment.

Fairchild, J.F., Ruessler, D.S., Haverland, P.S., Carlson, A.R., 1997. Comparative sensitivity of Selenastrum capricornutum and Lemna minor to sixteen herbicides. Arch. Environ. Contam. Toxicol. 32, 353–357. https://doi.org/10.1007/s002449900196

Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y., Weinstein, J.N., 2001. Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. J. Med. Chem. 44, 3254–3263. https://doi.org/10.1021/jm0005151

Finney, D., 1971. Probit Analysis. Cambridge University Press, London, UK.

Forouzanfar, M.H., Afshin, A., Alexander, L.T., Biryukov, S., Brauer, M., Cercy, K., Charlson, F.J., Cohen, A.J., Dandona, L., Estep, K., Ferrari, A.J., Frostad, J.J., Fullman, N., Godwin, W.W., Griswold, M., Hay, S.I., Kyu, H.H., Larson, H.J., Lim, S.S., Liu, P.Y., Lopez, A.D., Lozano, R., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., Naghavi, M., Reitsma, M.B., Roth, G.A., Sur, P.J., Vos, T., Wagner, J.A., Wang, H., Zhao, Y., Zhou, M., Barber, R.M., Bell, B., Blore, J.D., Casey, D.C., Coates, M.M., Cooperrider, K., Cornaby, L., Dicker, D., Erskine, H.E., Fleming, T., Foreman, K., Gakidou, E., Haagsma, J.A., Johnson, C.O., Kemmer, L., Ku, T., Leung, J., Masiye, F., Millear, A., Mirarefin, M., Misganaw, A., Mullany, E., Mumford, J.E., Ng, M., Olsen, H., Rao, P., Reinig, N., Roman, Y., Sandar, L., Santomauro, D.F., Slepak, E.L., Sorensen, R.J.D., Thomas, B.A., Vollset, S.E., Whiteford, H.A., Zipkin, B., Murray, C.J.L., Mock, C.N., Anderson, B.O., Futran, N.D., Anderson, H.R., Bhutta, Z.A., Nisar, M.I., Akseer, N., Krueger, H., Gotay, C.C., Kissoon, N., Kopec, J.A., Pourmalek, F., Burnett, R., Abajobir, A.A., Knibbs, L.D., Veerman, J.L., Lalloo, R., Scott, J.G., Alam, N.K.M., Gouda, H.N., Guo, Y., McGrath, J.J., Jeemon, P., Dandona, R., Goenka, S., Kumar, G.A., Gething, P.W., Bisanzio, D., Deribew, A., Darby, S.C., Ali, R., Bennett, D.A., Jha, V., Kinfu, Y., McKee, M., Murthy, G.V.S., Pearce, N., Stöckl, H., Duan, L., Jin, Y., Li, Y., Liu, S., Wang, L., Ye, P., Liang, X., Azzopardi, P., Patton, G.C., Meretoja, A., Alam, K., Borschmann, R., Colquhoun, S.M., Weintraub, R.G., Szoeke, C.E.I., Ademi, Z., Taylor, H.R., Wijeratne, T., Batis, C., Barquera, S., Campos-Nonato, I.R., Contreras, A.G., Cuevas-Nasu, L., De, V., Gomez-Dantes, H., Heredia-Pi, I.B., Medina, C., Mejia-Rodriguez, F., Montañez Hernandez, J.C., Razo-García, C.A., Rivera, J.A., Rodríguez-Ramírez, S., Sánchez-Pimienta, T.G., Servan-Mori, E.E., Shamah, T., Mensah, G.A., Hoff, H.J., Neal, B., Driscoll, T.R., Kemp, A.H., Leigh, J., Mekonnen, A.B., Bhatt, S., Fürst, T., Piel, F.B., Rodriguez, A., Hutchings, S.J., Majeed, A., Soljak, M., Salomon, J.A., Thorne-Lyman, A.L., Ajala, O.N., Bärnighausen, T., Cahill, L.E., Ding, E.L., Farvid, M.S., Khatibzadeh, S., Wagner, G.R., Shrime, M.G., Fitchett, J.R.A., Aasvang, G.M., Savic, M., Abate, K.H., Gebrehiwot, T.T., Gebremedhin, A.T., Abbafati, C., Abbas, K.M., Abd-Allah, F., Abdulle, A.M., Abera, S.F., Melaku, Y.A., Abyu, G.Y., Betsu, B.D., Hailu, G.B., Tekle, D.Y., Yalew, A.Z., Abraham, B., Abu-Raddad, L.J., Adebiyi, A.O., Adedeji, I.A., Adou, A.K., Adsuar, J.C., Agardh, E.E., Rehm, J., Badawi, A., Popova, S., Agarwal, A., Ahmad, A., Akinyemiju, T.F., Schwebel, D.C., Singh, J.A., Al-Aly, Z., Aldhahri, S.F., Altirkawi, K.A., Terkawi, A.S., Aldridge, R.W., Tillmann, T., Alemu, Z.A., Tegegne, T.K., Alkerwi, A., Alla, F., Guillemin, F., Allebeck, P., Rabiee, R.H.S., Fereshtehnejad, S.M., Kivipelto, M., Carrero, J.J., Weiderpass, E., Havmoeller, R., Sindi, S., Alsharif, U., Alvarez, E., Alvis-Guzman, N., Amare, A.T., Ciobanu, L.G., Taye, B.W., Amberbir, A., Amegah, A.K., Amini, H., Karema, C.K., Ammar, W., Harb, H.L., Amrock, S.M., Andersen, H.H., Antonio, C.A.T., Faraon, E.J.A., Anwari, P., Ärnlöv, J., Larsson, A., Artaman, A., Asayesh, H., Asghar, R.J., Assadi, R., Atique, S., Avokpaho, E.F.G.A., Awasthi, A.,

Ayala, B.P., Bacha, U., Bahit, M.C., Balakrishnan, K., Barac, A., Barker-Collo, S.L., del Pozo-Cruz, B., Mohammed, S., Barregard, L., Petzold, M., Barrero, L.H., Basu, S., Del, L.C., Bazargan-Hejazi, S., Beardsley, J., Bedi, N., Beghi, E., Sheth, K.N., Bell, M.L., Huang, J.J., Bello, A.K., Santos, I.S., Bensenor, I.M., Lotufo, P.A., Berhane, A., Wolfe, C.D., Bernabé, E., Roba, H.S., Beyene, A.S., Hassen, T.A., Mesfin, Y.M., Bhala, N., Bhansali, A., Biadgilign, S., Bikbov, B., Bjertness, E., Htet, A.S., Boufous, S., Degenhardt, L., Resnikoff, S., Calabria, B., Bourne, R.R.A., Brainin, M., Brazinova, A., Majdan, M., Shen, J., Breitborde, N.J.K., Brenner, H., Schöttker, B., Broday, D.M., Brugha, T.S., Brunekreef, B., Kromhout, H., Butt, Z.A., van Donkelaar, A., Martin, R. V., Cárdenas, R., Carpenter, D.O., Castañeda-Orjuela, C.A., Castillo, J., Castro, R.E., Catalá-López, F., Chang, J., Chiang, P.P., Chibalabala, M., Chimed-Ochir, O., Jiang, Y., Takahashi, K., Chisumpa, V.H., Mapoma, C.C., Chitheer, A.A., Choi, J.J., Christensen, H., Christopher, D.J., Cooper, L.T., Crump, J.A., Poulton, R.G., Damasceno, A., Dargan, P.I., das Neves, J., Davis, A.C., Newton, J.N., Steel, N., Davletov, K., de Castro, E.F., De, D., Dellavalle, R.P., Des, D.C., Dharmaratne, S.D., Dhillon, P.K., Lal, D.K., Zodpey, S., Diaz-Torné, C., Dorsey, E.R., Doyle, K.E., Dubey, M., Rahman, M.H.U., Ram, U., Singh, A., Yadav, A.K., Duncan, B.B., Kieling, C., Schmidt, M.I., Elyazar, I., Endries, A.Y., Ermakov, S.P., Eshrati, B., Farzadfar, F., Kasaeian, A., Parsaeian, M., Esteghamati, A., Hafezi-Nejad, N., Sheikhbahaei, S., Fahimi, S., Malekzadeh, R., Roshandel, G., Sepanlou, S.G., Hassanvand, M.S., Heydarpour, P., Rahimi-Movaghar, V., Yaseri, M., Farid, T.A., Khan, A.R., Farinha, C.S.E.S., Faro, A., Feigin, V.L., Fernandes, J.G., Fischer, F., Foigt, N., Shiue, I., Fowkes, F.G.R., Franklin, R.C., Garcia-Basteiro, A.L., Geleijnse, J.M., Jibat, T., Gessner, B.D., Tefera, W., Giref, A.Z., Haile, D., Manamo, W.A.A., Giroud, M., Gishu, M.D., Martinez-Raga, J., Gomez-Cabrera, M.C., Gona, P., Goodridge, A., Gopalani, S. V., Goto, A., Inoue, M., Gugnani, H.C., Gupta, R., Gutiérrez, R.A., Orozco, R., Halasa, Y.A., Undurraga, E.A., Hamadeh, R.R., Hamidi, S., Handal, A.J., Hankey, G.J., Hao, Y., Harikrishnan, S., Haro, J.M., Hernández-Llanes, N.F., Hoek, H.W., Tura, A.K., Horino, M., Horita, N., Hosgood, H.D., Hoy, D.G., Hsairi, M., Hu, G., Husseini, A., Huybrechts, I., Iburg, K.M., Idrisov, B.T., Kwan, G.F., Ileanu, B. V., Pana, A., Kawakami, N., Shibuya, K., Jacobs, T.A., Jacobsen, K.H., Jahanmehr, N., Jakovljevic, M.B., Jansen, H.A.F., Jassal, S.K., Stein, M.B., Javanbakht, M., Jayaraman, S.P., Jayatilleke, A.U., Jee, S.H., Jonas, J.B., Kabir, Z., Kalkonde, Y., Kamal, R., She, J., Kan, H., Karch, A., Karimkhani, C., Kaul, A., Kazi, D.S., Keiyoro, P.N., Parry, C.D., Kengne, A.P., Matzopoulos, R., Wiysonge, C.S., Stein, D.J., Mayosi, B.M., Keren, A., Khader, Y.S., Khan, E.A., Khan, G., Khang, Y.H., Won, S., Khera, S., Tavakkoli, M., Khoja, T.A.M., Khubchandani, J., Kim, C., Kim, D., Kimokoti, R.W., Kokubo, Y., Koul, P.A., Koyanagi, A., Kravchenko, M., Varakin, Y.Y., Kuate, B., Kuchenbecker, R.S., Kucuk, B., Kuipers, E.J., Lallukka, T., Shiri, R., Meretoja, T.J., Lan, Q., Latif, A.A., Lawrynowicz, A.E.B., Leasher, J.L., Levi, M., Li, X., Liang, J., Lloyd, B.K., Logroscino, G., Lunevicius, R., Pope, D., Mahdavi, M., Malta, D.C., Marcenes, W., Matsushita, K., Nachega, J.B., Tran, B.X., Meaney, P.A., Mehari, A., Tedla, B.A., Memish, Z.A., Mendoza, W., Mensink, G.B.M., Mhimbira, F.A., Miller, T.R., Mills, E.J., Mohammadi, A., Mola, G.L.D., Monasta, L., Morawska, L., Norman, R.E., Mori, R., Mozaff, D., Shi, P., Werdecker, A., Mueller, U.O., Paternina, A.J., Westerman, R., Seedat, S., Naheed, A., Nangia, V., Nassiri, N., Nguyen, Q.L., Nkamedjie, P.M., Norheim, O.F., Norrving, B., Nyakarahuka, L., Obermeyer, C.M., Ogbo, F.A., Oh, I., Oladimeji, O., Sartorius, B., Olusanya, B.O., Olivares, P.R., Olusanya, J.O., Opio, J.N., Oren, E., Ortiz, A., Ota, E., Mahesh, P.A., Park, E., Patel, T., Patil, S.T., Patten, S.B., Wang, J., Pereira, D.M., Cortinovis, M., Giussani, G., Perico, N., Remuzzi, G., Pesudovs, K., Phillips, M.R., Pillay, J.D., Plass, D., Tobollik, M.,

Polinder, S., Pond, C.D., Pope, C.A., Prasad, N.M., Qorbani, M., Radfar, A., Rafay, A., Rana, S.M., Rahman, M., Rahman, S.U., Rajsic, S., Rai, R.K., Raju, M., Ranganathan, K., Refaat, A.H., Rehm, C.D., Ribeiro, A.L., Rojas-Rueda, D., Roy, A., Satpathy, M., Tandon, N., Rothenbacher, D., Saleh, M.M., Sanabria, J.R., Sanchez-Riera, L., Sanchez-Niño, M.D., Sarmiento-Suarez, R., Sawhney, M., Schmidhuber, J., Schneider, I.J.C., Schutte, A.E., Silva, D.A.S., Shahraz, S., Shin, M., Shaheen, A., Shaikh, M.A., Sharma, R., Shigematsu, M., Yoon, S., Shishani, K., Sigfusdottir, I.D., Singh, P.K., Silveira, D.G.A., Silverberg, J.I., Yano, Y., Soneji, S., Stranges, S., Steckling, N., Sreeramareddy, C.T., Stathopoulou, V., Stroumpoulis, K., Sunguya, B.F., Swaminathan, S., Sykes, B.L., Tabarés-Seisdedos, R., Talongwa, R.T., Tanne, D., Tuzcu, E.M., Thakur, J., Shaddick, G., Thomas, M.L., Thrift, A.G., Thurston, G.D., Thomson, A.J., Topor-Madry, R., Topouzis, F., Towbin, J.A., Uthman, O.A., Tobe-Gai, R., Tsilimparis, N., Tsala, Z., Tyrovolas, S., Ukwaja, K.N., van Os, J., Vasankari, T., Venketasubramanian, N., Violante, F.S., Waller, S.G., Uneke, C.J., Wang, Y., Weichenthal, S., Woolf, A.D., Xavier, D., Xu, G., Yakob, B., Yip, P., Kesavachandran, C.N., Montico, M., Ronfani, L., Yu, C., Zaidi, Z., Yonemoto, N., Younis, M.Z., Wubshet, M., Zuhlke, L.J., Zaki, M.E., Zhu, J., 2016. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388, 1659–1724. https://doi.org/10.1016/S0140-6736(16)31679-8

Fourches, D., Muratov, E., Tropsha, A., 2010. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model. 50, 1189. https://doi.org/10.1021/CI100176X

Frank, L.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109–135. https://doi.org/10.1080/00401706.1993.10485033

Franke, R., 1984. Theoritical Drug Design Methods. Elsevier, Amsterdam.

Free, S.M., Wilson, J.W., 1964. A mathematical contribution to structure-activity studies. J. Med. Chem. 7, 395–399. https://doi.org/https://doi.org/10.1021/jm00334a001

Fried, J., Borman, A., 1958. Synthetic derivatives of cortical hormones. Vitam. Horm. 16, 303–374. https://doi.org/10.1016/S0083-6729(08)60320-9

Fujita, T., Ban, T., 1971. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. J. Med. Chem. 14, 148–152. https://doi.org/https://doi.org/10.1021/jm00284a016

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

Gaggi, C., Duccini, M., Bacci, E., Sbrilli, G., Bucci, M., Naby, A.M.H. El, 1995. Toxicity and hazard ranking of *s* -triazine herbicides using microtox® two green algal species and a marine crustacean. Environ. Toxicol. Chem. 14, 1065–1069. https://doi.org/10.1002/etc.5620140618

Gajewicz, A., 2017. Development of valuable predictive read-across models based on "real-life" (sparse) nanotoxicity data. Environ. Sci. Nano 4, 1389–1403.

Gajewicz, A., Jagiello, K., Cronin, M.T.D., Leszczynski, J., Puzyn, T., 2017. Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. Environ. Sci. Nano 4, 346–358.

Gallegos Saliner A., B.S.P., 2006. Mini-Review on Chemical Similarity and Prediction of Toxicity. Curr. Comput. Aided-Drug Des. 2, 105–122. https://doi.org/10.2174/157340906777441681

Gao, L., Shi, Y., Li, W., Liu, J., Cai, Y., 2012. Occurrence, distribution and bioaccumulation of antibiotics in the Haihe River in China. J. Environ. Monit. 14, 1248–1255. https://doi.org/10.1039/c2em10916f

Gasteiger, J., Zupan, J., 1993. Neural networks in chemistry. Angew. Chemie Int. Ed. English 32, 503–527. https://doi.org/10.1002/ANIE.199305031

Gaudin, T., Rotureau, P., Fayet, G., 2015. Mixture descriptors toward the development of Quantitative structure-property relationship models for the flash points of organic mixtures. Ind. Eng. Chem. Res. 54, 6596–6604. https://doi.org/10.1021/acs.jecr.5b01457

Geron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc., USA.

Ghanem, O. Ben, Mutalib, M.I.A., Lévêque, J.M., El-Harbawi, M., 2017. Development of QSAR model to predict the ecotoxicity of Vibrio fischeri using COSMO-RS descriptors. Chemosphere 170, 242–250. https://doi.org/10.1016/j.chemosphere.2016.12.003

Ghosh, S., Chatterjee, M., Roy, K., 2023. Predictive quantitative read-across structure–property relationship modeling of the retention time (Log t R) of pesticide residues present in foods and vegetables. J. Agric. Food Chem. 71, 9538–9548. https://doi.org/10.1021/acs.jafc.3c01438

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

Golbraikh, A., Tropsha, A., 2000. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol. Divers. 2000 54 5, 231–243. https://doi.org/10.1023/A:1021372108686

Gramatica, P., 2020. Principles of QSAR Modeling. Int. J. Quant. Struct. Relationships 5, 61–97. https://doi.org/10.4018/ijqspr.20200701.oa1

Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem. 34, 2121–2132. https://doi.org/10.1002/jcc.23361

Grube, A., Donaldson, D., Kiely, T., Wu, L., 2011. Pesticides industry sales and usage: 2006 and 2007 Market estimates, U.S. Environmental Protection Agency. Washington DC.

Guha, R., Jurs, P.C., 2005. Determining the validity of a QSAR model - A classification approach. J. Chem. Inf. Model. 45, 65–73. https://doi.org/10.1021/ci0497511

Guner, O., 2005. History and evolution of the pharmacophore concept in computer-aided drug design. Curr. Top. Med. Chem. 2, 1321–1332.

https://doi.org/10.2174/1568026023392940

Halder, A.K., Cordeiro, M.N.D.S., 2019. Development of Predictive Linear and Non-linear QSTR Models for Aliivibrio Fischeri Toxicity of Deep Eutectic Solvents. Int. J. Quant. Struct. Relationships 4, 50–69. https://doi.org/10.4018/ijqspr.2019100104

Hammett, L.P., 1970. Physical Organic Chemistry: Reaction rate, equilibria and mechanism, 2nd ed. McGraw-Hill, New York.

Hansch, C., 1969. A quantitative approach of biochemical structure-activity relationship. Acc. Chem. Res. 2, 232–239.

Hansch, C., Fujita, T., 1964. ρ-σ-π analysis. A Method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 86, 1616–1626. https://doi.org/https://doi.org/10.1021/ja01062a035

Hansch, C., Maloney, P.P., Fujita, T., Muir, R.M., 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194, 178–180. https://doi.org/10.1038/194178b0

Hansen, O.R., 1962. Hammett series with biological activity. Acta Chem. Scand. 16, 1593–1600. https://doi.org/10.3891/acta.chem.scand.16-1593

Hao, Y., Sun, G., Fan, T., Tang, X., Zhang, J., Liu, Y., Zhang, N., Zhao, L., Zhong, R., Peng, Y., 2020. In vivo toxicity of nitroaromatic compounds to rats: QSTR modelling and interspecies toxicity relationship with mouse. J. Hazard. Mater. 399, 122981. https://doi.org/10.1016/j.jhazmat.2020.122981

Harell, F.E., 2001. Regression modeling strategies. Springer-Verlag.

Hasan, M.K., Shahriar, A., Jim, K.U., 2019. Water pollution in Bangladesh and its impact on public health. Heliyon 5, e02145. https://doi.org/10.1016/j.heliyon.2019.e02145

Hawkins, D.M., Young, S.S., Rusinko, A., 1997. Analysis of a large structure-activity data set using recursive partitioning. Quant. Struct. Relationships 16, 296–302. https://doi.org/10.1002/QSAR.19970160404

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Their Appl. 13, 18–21. https://doi.org/10.1109/5254.708428

Heys, K.A., Shore, R.F., Pereira, M.G., Jones, K.C., Martin, F.L., 2016. Risk assessment of environmental mixture effects. RSC Adv. 6, 47844–47857. https://doi.org/10.1039/c6ra05406d

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoover, G., Kar, S., Guffey, S., Leszczynski, J., Sepúlveda, M., 2019. In vitro and in silico modeling of per fluoroalkyl substances mixture toxicity in an amphibian fi broblast cell line. Chemosphere 233, 25–33. https://doi.org/10.1016/j.chemosphere.2019.05.065

Hopfinger, A.J., Wang, S., Tokarski, J.S., Jin, B., Albuquerque, M., Madhav, P.J., Duraiswami, C., 1997. Construction of 3D-QSAR models using the 4D-QSAR analysis

formalism. J. Am. Chem. Soc. 119, 10509–10524. https://doi.org/10.1021/JA9718937

Horel, A.E., 1962. Application of ridge analysis to regression problems. Chem. Eng. Prog. 58, 54–59.

Horton, A.A., Barnes, D.K.A., 2020. Microplastic pollution in a rapidly changing world: Implications for remote and vulnerable marine ecosystems. Sci. Total Environ. 738, 140349. https://doi.org/10.1016/j.scitotenv.2020.140349

Howard, G.J., Webster, T.F., 2009. Generalized concentration addition: A method for examining mixtures containing partial agonists. J. Theor. Biol. 259, 469–477. https://doi.org/10.1016/j.jtbi.2009.03.030

Hübner, Y., Hoettges, K.F., Hughes, M.P., 2003. Water quality test based on dielectrophoretic measurements of fresh water algae Selenastrum capricornutum. J. Environ. Monit. 5, 861–864. https://doi.org/10.1039/b309131g

Ighalo, J.O., Adeniyi, A.G., Adeniran, J.A., Ogunniyi, S., 2021. A systematic literature analysis of the nature and regional distribution of water pollution sources in Nigeria. J. Clean. Prod. 283, 124566. https://doi.org/10.1016/J.JCLEPRO.2020.124566

Iwasaki, Y., Gauthier, P., 2016. Concentration addition and response addition to analyze mixture toxicity: Is it worth testing? Environ. Toxicol. Chem. 35, 526–527. https://doi.org/10.1002/etc.3263

Jacobs, N.F.B., Towle, K.M., Finley, B.L., Gaffney, S.H., 2019. An updated evaluation of potential health hazards associated with exposures to asbestos-containing drywall accessory products. Crit. Rev. Toxicol. 49, 430–444. https://doi.org/10.1080/10408444.2019.1639612

Jeliazkova, N., Jaworska, J., Worth, P., 2010. Open Source Tools for Read-Across and Category Formation, in: Cronin, M.T.., Madden, J.. (Eds.), In Silico Toxicology: Principles and Applications. The Royal Society of Chemistry, UK.

Jiang, L., Hu, X., Yin, D., Zhang, H., Yu, Z., 2011. Occurrence, distribution and seasonal variation of antibiotics in the Huangpu River, Shanghai, China. Chemosphere 82, 822–828. https://doi.org/10.1016/J.CHEMOSPHERE.2010.11.028

Jiang, L., Lin, Z., Hu, X., Yin, D., 2010. Toxicity prediction of antibiotics on luminescent bacteria, Photobacterium phosphoreum, based on their Quantitative structure-activity relationship models. Bull. Environ. Contam. Toxicol. 85, 550–555. https://doi.org/10.1007/s00128-010-0157-z

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science (80-.). 349, 255–260. https://doi.org/10.1126/science.aaa8415

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

Kar, S., Leszczynski, J., 2019. Exploration of computational approaches to predict the toxicity of chemical mixtures. Toxics 7, 1–18. https://doi.org/10.3390/TOXICS7010015

Kar, S., Sanderson, H., Roy, K., Benfenati, E., Leszczynski, J., 2020. Ecotoxicological

assessment of pharmaceuticals and personal care products using predictive toxicology approaches. Green Chem. 22, 1458–1516. https://doi.org/10.1039/c9gc03265g

Katritzky, A.R., Fara, D.C., Petrukhin, R.O., Tatham, D.B., Maran, U., Lomaka, A., Karelson, M., 2002. The present utility and future potential for medicinal chemistry of QSAR / QSPR with whole molecule descriptors. Curr. Top. Med. Chem. 2, 1333–1356.

Kemper, N., 2008. Veterinary antibiotics in the aquatic and terrestrial environment. Ecol. Indic. 8, 1–13. https://doi.org/10.1016/j.ecolind.2007.06.002

Kennedy, J., Eberhart, R., 1994. Particle swarm optimization, in: IEEE International Conference on Neural Network. New Jersey, p. 1942.

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

Khan, K.I., Roy, K., 2021. Chemometric Modeling of Toxicity of Chemicals to Marine Bacteria, in: Roy, K. (Ed.), Chemometrics and Cheminformatics in Aquatic Toxicology. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119681397.ch18

Khan, P.M., Rasulev, B., Roy, K., 2018. QSPR Modeling of the Refractive Index for Diverse Polymers Using 2D Descriptors. ACS Omega 3, 13374–13386. https://doi.org/10.1021/acsomega.8b01834

Kier, L.B., Hall, L.H., 1990. An electrotopological-state index for atoms in molecules. Pharm. Res. 7, 801–807. https://doi.org/10.1023/A:1015952613760

Kim, D., Cui, R., Moon, J., Kwak, J. Il, An, Y.J., 2019. Soil ecotoxicity study of DEHP with respect to multiple soil species. Chemosphere 216, 387–395. https://doi.org/10.1016/J.CHEMOSPHERE.2018.10.163

Kim, J., Kim, S., Schaumann, G.E., 2013. Reliable predictive computational toxicology methods for mixture toxicity: Toward the development of innovative integrated models for environmental risk assessment. Rev. Environ. Sci. Biotechnol. 12, 235–256. https://doi.org/10.1007/S11157-012-9286-7

Klatte, S., Schaefer, H.C., Hempel, M., 2017. Pharmaceuticals in the environment – A short review on options to minimize the exposure of humans, animals and ecosystems. Sustain. Chem. Pharm. 5, 61–66. https://doi.org/10.1016/j.scp.2016.07.001

Klopman, G., 1984. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J. Am. Chem. Soc. 106, 7315–7321. https://doi.org/https://doi.org/10.1021/ja00336a004

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., Ivanov, P., 2016. Jupyter Notebooks - A publishing format for reproducible computational workflows, in: Loizides, F., Scmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Kolpin, D.W., Furlong, E.T., Meyer, M.T., Thurman, E.M., Zaugg, S.D., Barber, L.B.,

Buxton, H.T., 2002. Pharmaceuticals, Hormones, and Other Organic Wastewater Contaminants in U.S. Streams, 1999−2000: A National Reconnaissance. Environ. Sci. Technol. 36, 1202–1211. https://doi.org/10.1021/es011055J

Kononenko, I., Kukar, M., 2007. Machine learning and data mining. Horwood Publishing Limited, Chichester.

Kortenkamp, A., Backhaus, T., Faust, M., 2009. State of art report on mixture toxicity.

Kortenkamp, A., Faust, M., Backhaus, T., Altenburger, R., Scholze, M., Müller, C., Ermler, S., Posthuma, L., Brack, W., 2019. Mixture risks threaten water quality: the European Collaborative Project SOLUTIONS recommends changes to the WFD and better coordination across all pieces of European chemicals legislation to improve protection from exposure of the aquatic environment to. Environ. Sci. Eur. 31, 1–4. https://doi.org/10.1186/s12302-019-0245-6

Kovshov, A.A., 2019. Assessment of the risk of polychlorinated biphenyls exposure in the indigenous population of coastal Chukotka. IOP Conf. Ser. Earth Environ. Sci. 263, 012033. https://doi.org/10.1088/1755-1315/263/1/012033

Kubinyi, H., 1988. Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis. Quant. Struct. Relationships 7, 121–133. https://doi.org/10.1002/QSAR.19880070303

Kubinyi, H., 1976. Quantitative structure-activity relationships. 2. A mixed approach, based on Hansch and Free-Wilson analysis. J. Med. Chem. 19, 587–600. https://doi.org/https://doi.org/10.1021/jm00227a004

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, 26th ed. Springer.

Kutter, E., Hansch, C., 1969. Steric parameters in drug design. Monoamine oxidase inhibitors and antihistamines. J. Med. Chem. 12, 647–652.

Lamastra, L., Balderacchi, M., Trevisan, M., 2016. Inclusion of emerging organic contaminants in groundwater monitoring plans. MethodsX 3, 459–476. https://doi.org/10.1016/j.mex.2016.05.008

Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K. V, McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Jairam Ramesh, M., Johan Rockström, P., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., Schayck, O.C.P. van, Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. The Lancet Commision 391, 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0

Leardi, R., 2001. Genetic algorithms in chemometrics and chemistry: A review. J. Chemom. https://doi.org/10.1002/cem.651

Leardi, R., 1996. Genetic algorithms in feature selection, in: Devillers, J. (Ed.), Genetic Algorithms in Molecular Modeling. Academic Press (Elsevier), Great Britain, pp. 67–

86. https://doi.org/https://doi.org/10.1016/B978-012213810-2/50004-9

Li, C., Chen, J., Wang, J., Ma, Z., Han, P., Luan, Y., Lu, A., 2015. Occurrence of antibiotics in soils and manures from greenhouse vegetable production bases of Beijing, China and an associated risk assessment. Sci. Total Environ. 521–522, 101–107. https://doi.org/10.1016/j.scitotenv.2015.03.070

Li, D., Yang, M., Hu, J., Ren, L., Zhang, Y., Li, K., 2008. Determination and fate of oxytetracycline and related compounds in oxytetracycline production wastewater and the receiving river. Environ. Toxicol. Chem. 27, 80–86. https://doi.org/10.1897/07-080.1

Li, F., Fan, T., Sun, G., Zhao, L., Zhong, R., Peng, Y., 2022. Systematic QSAR and iQCCR modelling of fused/non-fused aromatic hydrocarbons (FNFAHs) carcinogenicity to rodents: reducing unnecessary chemical synthesis and animal testing. Green Chem. 24, 5304–5319. https://doi.org/10.1039/D2GC00986B

Li, S., Zhu, H., Xia, Y., Yu, M., Liu, K., Ye, Z., Chen, Y., 1959. Mass culture of unicellular green alga. Acta Hydrobiol. Sin 4, 121–129.

Lifnick, R.L., 1989. Hans Horst Meyer and the lipoid theory of narcosis. Trends Pharmacol. Sci. 10, 265–269. https://doi.org/10.1016/0165-6147(89)90025-4

Lin, Z., Zhong, P., Yin, K., Wang, L., Yu, H., 2003. Quantification of joint effect for hydrogen bond and development of QSARs for predicting mixture toxicity. Chemosphere 52, 1199–1208. https://doi.org/10.1016/S0045-6535(03)00329-1

Lipnick, R.L., 1991. Charles Ernest Overton: narcosis studies and a contribution to general pharmacology, in: Lipnick, R.L. (Ed.), Studies of Narcosis. Springer, Dordrecht, pp. 14–22. https://doi.org/10.1007/978-94-011-3096-7_3

Loewe, S., Muischnek, H., 1926. Über Kombinationswirkungen. Naunyn. Schmiedebergs. Arch. Pharmacol. 114, 313–326. https://doi.org/doi.org/10.1007/BF01952257

Logan, D.T., Wilson, H.T., 1995. An ecological risk assessment method for species exposed to contaminant mixtures. Environ. Toxicol. Chem. 14, 351–359. https://doi.org/10.1002/etc.5620140222

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

Luo, Y., Guo, W., Ngo, H.H., Nghiem, L.D., Hai, F.I., Zhang, J., Liang, S., Wang, X.C., 2014. A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. Sci. Total Environ. 473–474, 619–641. https://doi.org/10.1016/j.scitotenv.2013.12.065

Ma, J., Lin, F., Wang, S., Xu, L., 2003. Toxicity of 21 herbicides to the green alga Scenedesmus quadricauda. Bull. Environ. Contam. Toxicol. 71, 594–601. https://doi.org/10.1007/s00128-003-8521-x

Ma, J., Wang, S., Wang, P., Ma, L., Chen, X., Xu, R., 2006. Toxicity assessment of 40 herbicides to the green alga Raphidocelis subcapitata. Ecotoxicol. Environ. Saf. 63, 456–

462. https://doi.org/10.1016/j.ecoenv.2004.12.001

Malik, L.A., Bashir, A., Qureashi, A., Pandith, A.H., 2019. Detection and removal of heavy metal ions: a review. Environ. Chem. Lett. 17, 1495–1521. https://doi.org/10.1007/s10311-019-00891-z

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: A review. Front. Public Heal. 8, 1–13. https://doi.org/10.3389/fpubh.2020.00014

Mansour, F., Al-Hindi, M., Saad, W., Salam, D., 2016. Environmental risk analysis and prioritization of pharmaceuticals in a developing world context. Sci. Total Environ. 557–558, 31–43. https://doi.org/10.1016/j.scitotenv.2016.03.023

Martin, Y.C., Holland, J.B., Jarboe, C.H., Plotnikoff, N., 1974. Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans. J. Med. Chem. 17, 409–413. https://doi.org/10.1021/jm00250a008

Martin, Y.C., Lynn, K.R., 1971. Quantitative structure-activity relationships in leucomycin and lincomycin antibiotics. J. Med. Chem. 14, 1162–1166. https://doi.org/10.1021/jm00294a005

Martinson, J.N.V., Pinkham, N. V., Peters, G.W., Cho, H., Heng, J., Rauch, M., Broadaway, S.C., Walk, S.T., 2019. Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. ISME J. 13, 2306–2318. https://doi.org/10.1038/s41396-019-0435-7

Martinson, J.N. V., Walk, S.T., 2020. Escherichia coli residency in the gut of healthy human adults. EcoSal Plus 9, 10–1128. https://doi.org/10.1128/ecosalplus.esp-0003-2020

Marvin|ChemAxon [WWW Document], 2013. URL https://chemaxon.com/products/marvin (accessed 11.14.21).

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

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. MATCH Commun. Math. Comput. Chem. 56, 237–248.

McDonald, G.C., 2009. Ridge regression. Wiley Interdiscip. Rev. Comput. Stat. 1, 93–100. https://doi.org/10.1002/wics.14

Mendoza, A., Aceña, J., Pérez, S., López de Alda, M., Barceló, D., Gil, A., Valcárcel, Y., 2015. Pharmaceuticals and iodinated contrast media in a hospital wastewater: A case study to analyse their presence and characterise their environmental risk and hazard. Environ. Res. 140, 225–241. https://doi.org/10.1016/j.envres.2015.04.003

Minitab Statistical Software [WWW Document], 2022. URL https://www.minitab.com/en-us/products/minitab/ (accessed 3.1.22).

Modi, S., Hughes, M., Garrow, A., White, A., 2012. The value of in silico chemistry in the

safety assessment of chemicals in the consumer goods and pharmaceutical industries. Drug Discov. Today 17, 135–142.

Mukherjee, R.K., Kumar, V., Roy, K., 2022. Ecotoxicological QSTR and QSTTR modeling for the prediction of acute oral toxicity of pesticides against multiple avian species. Environ. Sci. Technol. 56, 335–348. https://doi.org/10.1021/acs.est.1c05732

Muratov, E., Tropsha, A., Varlamova, E. V, Kuzmin, V.E., Artemenko, A.G., Muratov, N.N., Mileyko, S., Fourches, D., 2014. Everything out validation approach for QSAR models of chemical mixtures. JSM Clin. Pharm 1, 1005.

Mwense, M., Wang, X.Z., Buontempo, F. V., Horan, N., Young, A., Osborn, D., 2004. Prediction of noninteractive mixture toxicity of organic compounds based on a fuzzy set method. J. Chem. Inf. Comput. Sci. 44, 1763–1773.

Nations, U., 2017. United Nations Globally harmonized system of classification and labelling of chemicals (GHS). New York and Geneva.

Nazemzadeh, N., Nielsen, R.F., Gernaey, K. V., Mansouri, S.S., Andersson, M.P., 2021. Implementation of first-principles surface interactions in a hybrid machine learning assisted modelling of flocculation. Comput. Aided Chem. Eng. 50, 845–850. https://doi.org/10.1016/B978-0-323-88506-5.50132-7

Nováková, Z., Novák, J., Kitanovski, Z., Kukučka, P., Smutná, M., Wietzoreck, M., Lammel, G., Hilscherová, K., 2020. Toxic potentials of particulate and gaseous air pollutant mixtures and the role of PAHs and their derivatives. Environ. Int. 139, 105634. https://doi.org/10.1016/j.envint.2020.105634

Nowell, L.H., Norman, J.E., Moran, P.W., Martin, J.D., Stone, W.W., 2014. Pesticide toxicity index-a tool for assessing potential toxicity of pesticide mixtures to freshwater aquatic organisms. Sci. Total Environ. 476–477, 144–157. https://doi.org/10.1016/j.scitotenv.2013.12.088

OECD, 2021. The OECD QSAR Toolbox.

OECD, 2014. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, OECD Publishing, Paris, OECD Series on Testing and Assessment. https://doi.org/10.1787/9789264085442-en

OECD, 2007. Guidance documents on the validation of (Quantitative) structure-activity relationships [(Q)SAR] models [WWW Document]. OECD Environ. Heal. Saf. Publ. URL https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2 (accessed 12.22.22).

OECD, 2004. Validation of (Q)SAR Models - OECD [WWW Document]. URL https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed 9.30.23).

Ogungbemi, A.O., van Gestel, C.A.M., 2018. Extrapolation of imidacloprid toxicity between soils by exposing Folsomia candida in soil pore water. Ecotoxicology 27, 1107–1115. https://doi.org/10.1007/s10646-018-1965-x

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. https://doi.org/10.1016/j.chemolab.2011.03.011

Olah, M., Mracec, M., Ostopovici, L., Rad, R., Bora, A., Hadaruga, N., Olah, I., Banda, M., Simon, Z., Oprea, T.I., 2005. WOMBAT: World of Molecular Bioactivity, in: Opera, T.I. (Ed.), Chemoinformatics in Drug Discovery. Wiley-VCH: New York, pp. 221–239. https://doi.org/10.1002/3527603743.ch9

Olah, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga, N., Hadaruga, D., Moldovan, R., Fulias, A., Mractc, M., Oprea, T.I., 2008. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery, in: Schreiber, S.L., Kapoor, T.M., Weiss, G. (Eds.), Chemical Biology: From Small Molecules to Systems Biology and Drug Design. John Wiley & Sons, Ltd, pp. 760–786. https://doi.org/10.1002/9783527619375.CH13B

Olah, M.M., Bologa, C.G., Oprea, T.I., 2004. Strategies for compound selection. Curr. Drug Discov. Technol. 1, 211–220. https://doi.org/10.2174/1570163043334965

Olmstead, A.W., LeBlanc, G.A., 2005. Toxicity assessment of environmentally relevant pollutant mixtures using a heuristic model. Integr. Environ. Assess. Manag. 1, 114–122. https://doi.org/10.1897/IEAM_2004-005R.1

Onuzulu, C.D., Rotimi, O.A., Rotimi, S.O., 2019. Epigenetic modifications associated with in utero exposure to endocrine disrupting chemicals BPA, DDT and Pb. Rev. Environ. Health 34, 309–325. https://doi.org/10.1515/reveh-2018-0059

Oprisiu, I., Varlamova, E., Muratov, E., Artemenko, A., Marcou, G., Polishchuk, P., Kuz'Min, V., Varnek, A., 2012. QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. Mol. Inform. 31, 491–502. https://doi.org/10.1002/minf.201200006

Ou, H. se, Ye, J. shao, Ma, S., Wei, C. hai, Gao, N. yun, He, J. zhao, 2016. Degradation of ciprofloxacin by UV and UV/H2O2 via multiple-wavelength ultraviolet light-emitting diodes: Effectiveness, intermediates and antibacterial activity. Chem. Eng. J. 289, 391–401. https://doi.org/10.1016/j.cej.2016.01.006

Pandey, S.K., Ojha, P.K., Roy, K., 2020. Exploring QSAR models for assessment of acute fish toxicity of environmental transformation products of pesticides (ETPPs). Chemosphere 252, 126508. https://doi.org/10.1016/j.chemosphere.2020.126508

Paul, R., 2021. Personal Communication: paulrahul1997@gmail.com.

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

Pereira, A.S., Dâmaso-Rodrigues, M.L., Amorim, A., Daam, M.A., Cerejeira, M.J., 2018. Aquatic community structure in Mediterranean edge-of-field waterbodies as explained by environmental factors and the presence of pesticide mixtures. Ecotoxicology 27, 661–674. https://doi.org/10.1007/s10646-018-1944-2

Pliska, V., Testa, B., Waterbeemd, H. van de, 2008. Lipophilicity in drug action and

toxicology, in: Testa, B., Waterbeemd, H. van de (Eds.), Methods and Principles in Medicinal Chemistry. VCH, Weinheim, Germany, pp. 311–337.

Polanski, J., 2009. Receptor dependent multidimensional QSAR for modeling drug - receptor interactions. Curr. Med. Chem. 16, 3243–3257. https://doi.org/10.2174/092986709788803286

Pore, S., Banerjee, A., Roy, K., 2023. Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells. Sustain. Energy Fuels 7, 3412–3431. https://doi.org/10.1039/d3se00457k

Posthuma, L., De Zwart, D., 2006. Predicted effects of toxicant mixtures are confirmed by changes in fish species assemblages in Ohio, USA, rivers. Environ. Toxicol. Chem. 25, 1094–1105. https://doi.org/10.1897/05-305R.1

Preston, S., Coad, N., Townend, J., Killham, K., Paton, G.I., 2000. Biosensing the acute toxicity of metal interactions: Are they additive, synergistic, or antagonistic? Environ. Toxicol. Chem. 19, 775–780. https://doi.org/10.1002/etc.5620190332

Prucell, W.P., Bass, G.P., Clayton, J.M., 1973. Strategy of drug design. A molecular guide to biological activity. Wiley and Sons, New York.

PubChem [WWW Document], 2023. URL https://pubchem.ncbi.nlm.nih.gov/ (accessed 7.7.23).

Pullman, A., Pullman, B., 1955. Electronic structure and carcinogenic activity of aromatic molecules new developments. Adv. Cancer Res. 3, 117–169. https://doi.org/10.1016/S0065-230X(08)60919-7

Qin, L., Chen, Y.H., Zhang, X., Mo, L.Y., Zeng, H.H., Liang, Y.P., 2018. QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. Chemosphere 198, 122–129. https://doi.org/10.1016/j.chemosphere.2018.01.142

Qin, L.T., Liu, M., Zhang, X., Mo, L.Y., Zeng, H.H., Liang, Y.P., 2021. Concentration Addition, Independent Action, and Quantitative Structure–Activity Relationships for Chemical Mixture Toxicities of the Disinfection By products of Haloacetic Acids on the Green Alga Raphidocelis subcapitata. Environ. Toxicol. Chem. 40, 1431–1442. https://doi.org/10.1002/etc.4995

Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. WIRS Comput. Mol. Sci. 6, 147–172. https://doi.org/10.1002/wcms.1240

Randić, M., 1991. Generalized molecular descriptors. J. Math. Chem. 7, 155–168. https://doi.org/10.1007/BF01200821

Rasheed, T., Bilal, M., Nabeel, F., Iqbal, H.M.N., Li, C., Zhou, Y., 2018. Fluorescent sensor based models for the detection of environmentally-related toxic heavy metals. Sci. Total Environ. 615, 476–485. https://doi.org/10.1016/j.scitotenv.2017.09.126

Rasulev, B., Jabeen, F., Stafslien, S., Chisholm, B.J., Bahr, J., Ossowski, M., Boudjouk, P., 2017. Polymer coating materials and their fouling release activity: A cheminformatics approach to predict properties. ACS Appl. Mater. Interfaces 9, 1781–1792.

https://doi.org/10.1021/acsami.6b12766

Rekker, R.F., 1992. The history of drug research: From Overton to Hansch. Quant. Struct. Relationships 11, 195–199. https://doi.org/10.1002/QSAR.19920110214

Ribo, J.M., Kaiser, K.L.E., 1987. Photobacterium phosphoreum toxicity bioassay. I. Test procedures and applications. Environ. Toxicol. Water Qual. 2, 305–323. https://doi.org/10.1002/tox.2540020307

Richardson, B.J., 1869. Medical Times and Gezette 2. John Churchil and Sons, London.

Richet, M.C., 1893. Note sur le rapport Entre la Toxicite et les Propriretes Physiques des Corps. Compt. Rend. Soc. Biol. 45, 775–776.

Rider, C. V., LeBlanc, G.A., 2005. An integrated addition and interaction model for assessing toxicity of chemical mixtures. Toxicol. Sci. 87, 520–528. https://doi.org/10.1093/toxsci/kfi247

Ritskes-Hoitinga, M., Luijk, J. Van, 2019. How can systematic reviews teach us more about the implementation of the 3Rs and animal welfare? Animals 9, 1163. https://doi.org/10.3390/ani9121163

Rodríguez-Eugenio, N., McLaughlin, M., Pennock, D., 2018. Soil pollution: A hidden reality, FAO. Rome, Italy.

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. https://doi.org/10.1021/ci00020a020

Roy, J., Kumar Ojha, P., Carnesecchi, E., Lombardo, A., Roy, K., Benfenati, E., 2020. First report on a classification-based QSAR model for chemical toxicity to earthworm. J. Hazard. Mater. 386, 121660. https://doi.org/10.1016/j.jhazmat.2019.121660

Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin. Drug Discov. 2, 1567–1577. https://doi.org/10.1517/17460441.2.12.1567

Roy, K., Ambure, P., 2016. The “double cross-validation” software tool for MLR QSAR model development. Chemom. Intell. Lab. Syst. 159, 108–126. https://doi.org/10.1016/j.chemolab.2016.10.009

Roy, K., Ambure, P., Kar, S., 2018a. Prediction Reliability Indicator: A new tool to judge the quality of predictions from QSAR models for new query compounds, in: MOL2NET, International Conference Series on Multidisciplinary Sciences. https://doi.org/10.3390/mol2net-04-05265

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

Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., 2013. Some case studies on application of “rm2” metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data. J. Comput. Chem. 34,

1071–1082. https://doi.org/10.1002/JCC.23231

Roy, K., Das, R.N., 2015. The "ETA" indices in QSAR/QSPR/QSTR research, in: Roy, K. (Ed.), Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. IGI Global, Hershey PA, USA, pp. 48–83. https://doi.org/10.4018/978-1-4666-8136-1.ch015

Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom. Intell. Lab. Syst. 152, 18–33. https://doi.org/10.1016/j.chemolab.2016.01.008

Roy, K., Ghosh, G., 2010. Exploring QSARs with Extended Topochemical Atom (ETA) Indices for Modeling Chemical and Drug Toxicity. Curr. Pharm. Des. 16, 2625–2639. https://doi.org/10.2174/138161210792389270

Roy, K., Ghosh, G., 2004. QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes. J. Chem. Inf. Comput. Sci. 44, 559–567. https://doi.org/10.1021/ci0342066

Roy, K., Ghosh, G., 2003. Introduction of extended topochemical atom (ETA) indices in the valance electron mobile (VEM) environment as tools for QSAR/QSPR studies. Internet Electron. J. Mol. Des. 2, 599–620.

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

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., Kar, S., Ojha, P.K., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model. 52, 396–408. https://doi.org/10.1021/CI200520G

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

RSC, 2022. ChemSpider, Search and share chemistry [WWW Document]. URL http://www.chemspider.com/ (accessed 11.10.22).

Rucker, C., Rucker, G., Meringer, M., 2007. y-Randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–57. https://doi.org/10.1021/ci700157b

Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 27, 2345–2357. https://doi.org/10.1021/ci700157b

Rusch, M., Spielmeyer, A., Zorn, H., Hamscher, G., 2019. Degradation and transformation of fluoroquinolones by microorganisms with special emphasis on ciprofloxacin. Appl. Microbiol. Biotechnol. 103, 6933–6948. https://doi.org/10.1007/s00253-019-10017-8

Russell, W.M.S., Burch, R.L., 1959. The Principles of Humane Experimental Technique. London: Methuen [WWW Document]. URL https://caat.jhsph.edu/principles/the-principles-of-humane-experimental-technique (accessed 12.20.19).

Saldana, D.A., Starck, L., Mougin, P., Rousseau, B., Creton, B., 2013. Prediction of flash points for fuel mixtures using machine learning and a novel equation. Energy and Fuels 27, 3811–3820. https://doi.org/10.1021/EF4005362

Sarker, I.H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. 2, 1–21. https://doi.org/10.1007/s42979-021-00592-x

Schultz, T.W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D.J., Mahony, C., Schwarz, M., White, A., Cronin, M.T.D., 2015. A strategy for structuring and reporting a read-across prediction of toxicity. Regul. Toxicol. Pharmacol. 72, 586–601. https://doi.org/10.1016/j.yrtph.2015.05.016

Seth, A., Ojha, P.K., Roy, K., 2020. QSAR modeling with ETA indices for cytotoxicity and enzymatic activity of diverse chemicals. J. Hazard. Mater. 394, 122498. https://doi.org/10.1016/j.jhazmat.2020.122498

Seth, A., Roy, K., 2020. QSAR modeling of algal low level toxicity values of different phenol and aniline derivatives using 2D descriptors. Aquat. Toxicol. 228, 105627. https://doi.org/10.1016/J.AQUATOX.2020.105627

Sigurnjak Bureš, M., Cvetnić, M., Miloloža, M., Kučić Grgić, D., Markić, M., Kušić, H., Bolanča, T., Rogošić, M., Ukić, 2021. Modeling the toxicity of pollutants mixtures for risk assessment: a review. Environ. Chem. Lett. 19, 1629–1655. https://doi.org/10.1007/s10311-020-01107-5

Silva, P., Rocha, M.J., Cruzeiro, C., Malhão, F., Reis, B., Urbatzka, R., Monteiro, R.A.F., Rocha, E., 2012. Testing the effects of ethinylestradiol and of an environmentally relevant mixture of xenoestrogens as found in the Douro River (Portugal) on the maturation of fish gonads—A stereological study using the zebrafish (Danio rerio) as model. Aquat. Toxicol. 124–125, 1–10. https://doi.org/10.1016/J.AQUATOX.2012.07.002

Singh, S., Kumar, V., Chauhan, A., Datta, S., Wani, A.B., Singh, N., Singh, J., 2018. Toxicity, degradation and analysis of the herbicide atrazine. Environ. Chem. Lett. 16, 211–237. https://doi.org/10.1007/s10311-017-0665-8

Singh, S., Singh, N., Kumar, V., Datta, S., Wani, A.B., Singh, D., Singh, K., Singh, J., 2016. Toxicity, monitoring and biodegradation of the fungicide carbendazim. Environ. Chem. Lett. 14, 317–329. https://doi.org/10.1007/s10311-016-0566-2

Snarey, M., Terrett, N.K., Willett, P., Wilton, D.J., 1997. Comparison of algorithms for dissimilarity-based compound selection. J. Mol. Graph. Model. 15, 372–385. https://doi.org/10.1016/S1093-3263(98)00008-4

Snedecor, G.W., Cochran, W.G., 1967. Statistical Methods. McGraw-Hill and IBH, New Delhi.

So, S.S., Richards, W.G., 1992. Application of neural networks: Quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR Inhibitors. J. Med. Chem. 35, 3201–3207. https://doi.org/10.1021/jm00095a016

Sobati, M.A., Abooali, D., Maghbooli, B., Najafi, H., 2016. A new structure-based model for

estimation of true critical volume of multi-component mixtures. Chemom. Intell. Lab. Syst. 155, 109–119. https://doi.org/10.1016/J.CHEMOLAB.2016.04.007

Sturini, M., Speltini, A., Maraschi, F., Pretali, L., Ferri, E.N., Profumo, A., 2015. Sunlight-induced degradation of fluoroquinolones in wastewater effluent: Photoproducts identification and toxicity. Chemosphere 134, 313–318. https://doi.org/10.1016/j.chemosphere.2015.04.081

Sun, G., Zhang, Y., Pei, L., Lou, Y., Mu, Y., Yun, J., Li, F., Wang, Y., Hao, Z., Xi, S., Li, C., Chen, C., Zhao, L., Zhang, N., Zhong, R., Peng, Y., 2021. Chemometric QSAR modeling of acute oral toxicity of polycyclic aromatic hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse. Ecotoxicol. Environ. Saf. 222, 112525. https://doi.org/10.1016/j.ecoenv.2021.112525

Syberg, K., Jensen, T.S., Cedergreen, N., Rank, J., 2009. On the use of mixture toxicity assessment in reach and the water framework directive: A review. Hum. Ecol. Risk Assess. 15, 1257–1272. https://doi.org/10.1080/10807030903304922

Tian, D., Lin, Z., Yin, D., Zhang, Y., Kong, D., 2012. Atomic charges of individual reactive chemicals in binary mixtures determine their joint effects: An example of cyanogenic toxicants and aldehydes. Environ. Toxicol. Chem. 31, 270–278. https://doi.org/10.1002/etc.1701

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemiformatics. Wiley-VCH, New York.

Topliss, J.G., Edwards, R.P., 1979. Chance factors in studies of quantitative structure-activity relationships. J. Med. Chem. 22, 1238–1244. https://doi.org/10.1021/JM00196A017

Topp, E., Renaud, J., Sumarah, M., Sabourin, L., 2016. Reduced persistence of the macrolide antibiotics erythromycin, clarithromycin and azithromycin in agricultural soil following several years of exposure in the field. Sci. Total Environ. 562, 136–144. https://doi.org/10.1016/j.scitotenv.2016.03.210

Torrens, F., Castellano, G., 2019. Chlorofluorocarbons, ozone levels evaluation, people actions, tropospheric O3, and air quality, in: Vakhrushev, A. V., Mukbaniani, O. V., Susanto, H. (Eds.), Chemical Technology and Informatics in Chemistry with Applications. Apple Academic Press.

Tosi, S., Sfeir, C., Carnesecchi, E., vanEngelsdorp, D., Chauzat, M.P., 2022. Lethal, sublethal, and combined effects of pesticides on bees: A meta-analysis and new risk assessment tools. Sci. Total Environ. 844, 156857. https://doi.org/10.1016/J.SCITOTENV.2022.156857

Truhaut, R., 1977. Ecotoxicology: Objectives, Principles and Perspectives. Ecotoxicol. Environ. Saf. 1, 151–173. https://doi.org/https://doi.org/10.1016/0147-6513(77)90033-1

Tute, M.S., 1990. History and objectives of quantitative drug design, in: Ramsden, C.A. (Ed.), Quantitative Drug Design. Pergamon Press, Oxford, pp. 1–31.

U. Simca-P, 10.0 [WWW Document], 2002. URL www.umetrics.com

US-EPA, 1986. Guidelines for the health risk assessment of chemical mixtures, Federal

Register.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 18, 463–477. https://doi.org/10.1038/s41573-019-0024-5

van De Waterbeemd, H., 1992. The history of drug research: From Hansch to the present. Quant. Struct. Relationships 11, 200–204. https://doi.org/10.1002/QSAR.19920110215

Van den Brink, P.J., Boxall, A.B.A., Maltby, L., Brooks, B.W., Rudd, M.A., Backhaus, T., Spurgeon, D., Verougstraete, V., Ajao, C., Ankley, G.T., Apitz, S.E., Arnold, K., Brodin, T., Cañedo-Argüelles, M., Chapman, J., Corrales, J., Coutellec, M.A., Fernandes, T.F., Fick, J., Ford, A.T., Giménez Papiol, G., Groh, K.J., Hutchinson, T.H., Kruger, H., Kukkonen, J.V.K., Loutseti, S., Marshall, S., Muir, D., Ortiz-Santaliestra, M.E., Paul, K.B., Rico, A., Rodea-Palomares, I., Römbke, J., Rydberg, T., Segner, H., Smit, M., van Gestel, C.A.M., Vighi, M., Werner, I., Zimmer, E.I., van Wensem, J., 2018. Toward sustainable environmental quality: Priority research questions for Europe. Environ. Toxicol. Chem. 37, 2281–2295. https://doi.org/10.1002/etc.4205

Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley & Sons, New York.

Vedani, A., Dobler, M., 2002. 5D-QSAR: The key for simulating induced fit? J. Med. Chem. 45, 2139–2149. https://doi.org/10.1021/jm011005p

Vedani, A., Dobler, M., Lill, M.A., 2005. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. J. Med. Chem. 48, 3700–3703. https://doi.org/10.1021/jm050185q

Veith, G.D., Broderius, S.J., 1990. Rules for Distinguishing Toxicants That Cause Type I and Type II Narcosis Syndromes. Environ. Health Perspect. 87, 207–211. https://doi.org/https://doi.org/10.1289/ehp.9087207

Venkatapathy, R., Wang, N.C.Y., 2013. Developmental toxicity prediction. Methods Mol. Biol. 930, 305–340.

Verdisson, S., Couderchet, M., Vernet, G., 2001. Effects of procymidone, fludioxonil and pyrimethanil on two non-target aquatic plants. Chemosphere 44, 467–474. https://doi.org/10.1016/S0045-6535(00)00468-9

Verhaar, H.J.M., Leeuwen, C.J. Van, Hermens, J.L.M., 1992. Classifying Environmental-Pollutants .1. Structure-Activity-Relationships for Prediction of Aquatic Toxicity. Chemosphere 25, 471–491. https://doi.org/10.1016/0045-6535(92)90280-5

Verloop, A., Hoogenstraaten, W., Tipker, J., 1976. Development and application of new steric substituent parameters in drug design, in: Ariens, E.J. (Ed.), Drug Design. Academic Press, New York, pp. 165–207.

Vink, S.R., Mikkers, J., Bouwman, T., Marquart, H., Kroese, E.D., 2010. Use of read-across and tiered exposure assessment in risk assessment under REACH - A case study on a phase-in substance. Regul. Toxicol. Pharmacol. 58, 64–71.

Wagner, J.M., Shimshak, D.G., 2007. Stepwise selection of variables in data envelopment

analysis: Procedures and managerial perspectives. Eur. J. Oper. Res. 180, 57–67. https://doi.org/10.1016/j.ejor.2006.02.048

Wang, D., Ning, Q., Dong, J., Brooks, B.W., You, J., 2020. Predicting mixture toxicity and antibiotic resistance of fluoroquinolones and their photodegradation products in Escherichia coli. Environ. Pollut. 262, 114275. https://doi.org/10.1016/j.envpol.2020.114275

Wang, D., Wu, X., Lin, Z., Ding, Y., 2018. A comparative study on the binary and ternary mixture toxicity of antibiotics towards three bacteria based on QSAR investigation. Environ. Res. 162, 127–134. https://doi.org/10.1016/j.envres.2017.12.015

Wang, T., Tang, L., Luan, F., Nat, M., Cordeiro, D.S., 2018. Prediction of the toxicity of binary mixtures by QSAR approach using the hypothetical descriptors. Int. J. Mol. Sci. 19, 3423. https://doi.org/10.3390/ijms19113423

Ward, T.J., Rausina, G.A., Stonebraker, P.M., Robinson, W.E., 2002. Apparent toxicity resulting from the sequestering of nutrient trace metals during standard Selenastrum capricornutum toxicity tests. Aquat. Toxicol. 60, 1–16. https://doi.org/10.1016/S0166-445X(01)00259-4

Wold, S., Ruhe, A., Wold, H., W. J. Dunn, I., 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735–743. https://doi.org/10.1137/0905052

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

Wu, Z., Li, D., Meng, J., Wang, H., 2010. Introduction to SIMCA-P and its application, in: Vinzi, V.E., Chin, W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares. Berlin, Germany, pp. 757–774.

Xiong, W., Sun, Y., Zhang, T., Ding, X., Li, Y., Wang, M., Zeng, Z., 2015. Antibiotics, antibiotic resistance genes, and bacterial community composition in fresh water aquaculture environment in China. Microb. Ecol. 70, 425–432. https://doi.org/10.1007/s00248-015-0583-x

Yang, R.S.H., Thomas, R.S., Gustafson, D.L., Campain, J., Benjamin, S.A., Verhaar, H.J.M., Mumtaz, M.M., 1998. Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling, in: Environmental Health Perspectives. Public Health Services, US Dept of Health and Human Services, pp. 1385–1393. https://doi.org/10.1289/ehp.98106s61385

Yao, Z., Lin, Z., Wang, T., Tian, D., Zou, X., Gao, Y., Yin, D., 2013. Using molecular docking-based binding energy to predict toxicity of binary mixture with different binding sites. Chemosphere 92, 1169–1176. https://doi.org/10.1016/j.chemosphere.2013.01.081

Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474. https://doi.org/10.1002/jcc.21707

Ye, J. shao, Liu, J., Ou, H. se, Wang, L. lin, 2016. Degradation of ciprofloxacin by 280 nm ultraviolet-activated persulfate: Degradation pathway and intermediate impact on proteome of Escherichia coli. Chemosphere 165, 311–319. https://doi.org/10.1016/j.chemosphere.2016.09.031

Yi, H., Yan, M., Huang, D., Zeng, G., Lai, C., Li, M., Huo, X., Qin, L., Liu, S., Liu, X., Li, B., Wang, H., Shen, M., Fu, Y., Guo, X., 2019. Synergistic effect of artificial enzyme and 2D nano-structured Bi2WO6 for eco-friendly and efficient biomimetic photocatalysis. Appl. Catal. B Environ. 250, 52–62. https://doi.org/10.1016/j.apcatb.2019.03.008

Young, D., Martin, T., Venkatapathy, R., Harten, P., 2008. Are the chemical structures in your QSAR correct? QSAR Comb. Sci. 27, 1337–1345. https://doi.org/10.1002/QSAR.200810084

Zahradnik, R., 1962. Influence of the structure of aliphatic substituents on the magnitude of the biological effect of substances. Arch. Int. Pharmacodyn. Ther. 135, 311–329.

Zahradník, R., 1962. Correlation of the biological activity of organic compounds by means of the linear free energy relationships. Experientia 18, 534–536. https://doi.org/10.1007/BF02151616

Zeng, J., Chen, F., Li, M., Wu, L., Zhang, H., Zou, X., 2019. The mixture toxicity of heavy metals on Photobacterium phosphoreum and its modeling by ion characteristics-based QSAR. PLoS One 14, 1–13.

Zhang, H., Huang, C.H., 2005. Oxidative transformation of fluoroquinolone antibacterial agents and structurally related amines by manganese oxide. Environ. Sci. Technol. 39, 4474–4483. https://doi.org/10.1021/es048166d

Zhang, J., Ding, T.T., Dong, X.Q., Bian, Z.Q., 2018. Time-dependent and Pb-dependent antagonism and synergism towards: Vibrio qinghaiensis sp.-Q67 within heavy metal mixtures. RSC Adv. 8, 26089. https://doi.org/10.1039/c8ra04191a

Zheng, W., Tropsha, A., 2000. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. J. Chem. Inf. Comput. Sci. 40, 185–194. https://doi.org/10.1021/ci980033m

Zou, X., Lin, Z., Deng, Z., Yin, D., Zhang, Y., 2012. The joint effects of sulfonamides and their potentiator on Photobacterium phosphoreum: Differences between the acute and chronic mixture toxicity mechanisms. Chemosphere 86, 30–35. https://doi.org/10.1016/j.chemosphere.2011.08.046

APPENDIX

REPRINTS

Journal of Hazardous Materials 408 (2021) 124936

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors

Mainak Chatterjee, Kunal Roy *

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

Editor: Dr. S Nan

Keywords:
Chemical mixtures
Aquatic toxicity
QSAR models
Partial least squares (PLS)

ABSTRACT

The rapid industrialization has led to the generation of various organic chemicals and multi-component mixtures which affect the environment adversely. Although organic chemicals are often exposed to the environment as a form of chemical mixtures rather than individual compounds, there is insufficient toxicity data available for the chemical mixtures due to the associated complexities. Most importantly, the nature of toxicity of mixtures is completely different from the individual chemicals, which makes the evaluation more difficult and challenging. In this paper, we have developed QSAR models for various individual and mixture data sets for the prediction of the aquatic toxicity. We have used Partial Least Squares (PLS) regression as a statistical tool to build the models. The various structural features of the individual chemicals and the mixture components have been modeled against the toxicity end point pEC_{50} (negative logarithm of median effective concentration in molar scale) of the aquatic organisms *Photobacterium phosphoreum* (marine bacterium) and *Selenastrum capricornutum* (freshwater algae). The mixture descriptors have been calculated by the weighted descriptor generation approach. The models were developed in accordance with OECD guidelines, and the quality of each model has been adjudged by strict validation parameters. The final models are robust, extremely predictive and interpretable mechanistically which can be used for the prediction of toxicity of untested chemical mixtures under the domain of applicability of the developed models.

1. Introduction

Aquatic toxicity is a major concern to the scientific world in this modern era of rapid industrialization and chemical advances. It is the unwanted and harmful effects on the aquatic organisms and ecosystems in various trophic levels due to the exposure of various manufactured chemicals, anthropogenic and natural materials (Aquatic Toxicity, 2016). Various water bodies (both fresh water and sea water) are being contaminated with the chemical and industrial wastes due to the improper waste management in industries, overuse or misuse and improper discharge of pharmaceuticals, pesticides, organic chemicals, or heavy metals (in the form of metal oxide nano-materials) from the various sectors of society (Schwarzenbach et al., 2006). As a result of this, the aquatic species and indirectly the human being are getting affected adversely due to the toxicity and bioaccumulation of toxic substances in the body. It is widely accepted that the entire living world including these aquatic eco-systems are continuously being exposed to the chemicals as the form of multi-component mixtures and the probable combinations of such chemical mixtures are increasing exponentially in the environment (Evans et al., 2016). According to Yang et al. (1998), "there is no such object as a single chemical exposure"; however the existing regulations of human risk assessment (HRA) predominantly consider the exposure and evaluation of the toxicity of individual compounds, instead of focusing on the toxicity of chemical mixtures. The toxicity data obtained from pure individual compounds may not resemble the toxicity of chemical mixtures. Interactions among the components in these mixtures can result in significant and complex changes to the apparent properties of pure compounds, which lead to additive, synergistic, or antagonistic effects, and thus pose a potential threat to the environmental safety and human health (Yang et al., 1998). So to get rid of these issues, proper risk assessment and toxicity data generation of the chemical mixtures are very essential, and hence it has grabbed sufficient attention of the scientists since past few decades (Escher et al., 2017; Villa et al., 2014; Wang et al., 2012).

Most of the chemical species occurring in the environment are at concentration far beneath their individual median effective concentrations (EC_{50}) or sometimes it may be lower than their individual no observed effective concentrations (NOEC); however they can still

* Corresponding author.
E-mail addresses: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).

https://doi.org/10.1016/j.jhazmat.2020.124936
Received 8 October 2020; Received in revised form 17 December 2020; Accepted 20 December 2020
Available online 22 December 2020
0304-3894/© 2020 Elsevier B.V. All rights reserved.

produce some effects to the living species of different compartments of environment due to the co-contamination with other chemical components of mixtures (Yang et al., 1998; Khan et al., 2020). The significance of the combined effects of co-contaminants have long been acknowledged by the regulatory authorities (Martens et al., 1984), and there are some commonly used approaches available for the toxicity prediction and risk assessment of chemical mixtures (Khan et al., 2020; Logan and Wilson, 1995). The experimental toxicity tests of whole mixtures are time consuming, laborious, costly, and ethically complicated processes. The major hurdle in the application of whole mixture toxicity test is that the mixture can be comprised of an unspecified number of individual components and in different concentration ratios; therefore, it is practically not feasible to perform the toxicity experimentation of all possible combinations of mixtures. Concentration addition (CA) and Independent action (IA) are the two most widely used component based approaches for the toxicity prediction of chemical mixtures (Khan et al., 2020). These approaches mainly depend on the effects of individual components of a mixture to predict the joint action of the same. German pharmacologist Loewe proposed the Concentration Addition (CA) hypothesis for the 1st time in 1926 (Loewe, 1953; Loewe and Muischnek, 1926). The CA models assume that the components of a mixture act via similar mode of action by binding to the same target site but exert different biological responses and thus, one chemical acts as a dilution of the other and can be replaced at a persistent quantity for the other component of the mixture. The CA model can be expressed mathematically by Loewe additivity Eq. (1) as follows:

$$C_{eff} = \sum C_a/EC_a \quad (1)$$

where, C_{eff} is the effective concentration of a mixture; C_a is the actual concentration of compound a; EC_a is the effective concentration of compound 'a' which can be sufficient for the mortality of 50% studied animal.

The IA models assume statistically independent responses through different binding sites and dissimilar modes of action from the components of mixtures (Kar and Leszczynski, 2019). According to this model, the mixture components do not interfere in the pharmacokinetic events such as uptake, distribution, and metabolism. The independent action of binary mixture can be estimated by the joint probability of statistically independent events by using the following Eq. (2) (Altenburger et al., 2003):

$$E(C_{mix}) = 1 - \prod_{a=1}^{n}[1 - E(C_a)] \quad (2)$$

where $E(C_{mix})$ stands for the combined effect of a mixture; $E(C_a)$ is the effective concentration of mixture component 'a' present at the concentration (C_a). Both of these strategies are experimentation dependent mathematical calculations and do not consider the possible interactions (synergism and antagonism) among the components of the mixtures. So, for the prediction of mixture toxicity and the risk assessment using the above mentioned hypothesis, toxicity tests of individual components are necessary which need sufficient time, manpower, money and moreover the animal experimentations as the prerequisite of the procedure. The computational toxicity prediction is another promising approach for the toxicity assessment of chemical mixtures using multiple algorithms. It is important to mention that the computational approaches are not the absolute substitution or alternative of the in-vivo or in-vitro experimentations. Rather, they aim to complement the toxicity tests by minimizing the animal experimentations, decreasing the time and cost of the experiments, followed by the improved toxicity prediction and risk assessment (Kar and Leszczynski, 2019).

Quantitative structure activity relationship (QSAR) is one of the widely used computational approaches which correlates the toxicity/activity/property of compounds with the structural and physicochemical features (Roy et al., 2015a). It has been extensively applied for the prediction of toxicity, activity, and property of single chemicals. QSAR has also been used to predict the concentration effects of the component in mixtures from the combined effects and defined concentration ratios of mixture components. Several QSAR modelers throughout the globe have proposed many QSAR models for the combined effects of chemical mixtures from the molecular descriptors that were computed as composite properties on the basis of the concentration ratios of mixture components (Hoover et al., 2019; Wang et al., 2018). Although the use of QSAR models on joint effects of mixtures is limited to the additive toxicity prediction and risk assessment of binary mixtures, a few modelers have successfully applied the concept to the toxicity prediction of multi-component mixtures also (Lu et al., 2009). This approach is also recommended by regulatory agencies like United States Environmental Protection Agency (US EPA), the Agency for Toxic Substances and Disease Registry (ATSDR), European regulations like the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), and the Organization of Economic Co-operation and Development (OECD).

In the current study, various QSAR models were developed for the future prediction of aquatic toxicity and risk assessment of various pharmaceuticals, organic chemicals and their corresponding binary mixtures. Two aquatic organisms, *Photobacterium phosphoreum* and *Selenastrum capricornutum* were chosen against which the toxicity data are collected. *P. phosphoreum* is a gram negative and bioluminescent bacterium living in symbiosis with various marine organisms and a classic marker of the presence of marine toxicants (De Zwart and Slooff, 1983; Ribo and Kaiser, 1987), whereas *S. capricornutum* is a type of microalgae of freshwater environment and is commonly used as a sensitive biomarker of the presence of freshwater toxicants (Hübner et al., 2003; Ward et al., 2002). Negative logarithms of median effective concentration (pEC50) of the chemical species and mixtures were used as the response variable (Y-variable in the data matrix) of the data sets. The weighted descriptor generation approach of binary mixtures has been used to calculate the mixture descriptors in this work. Here, the objectives of the current work are (1) to identify the maximum possible structural features of the collected data sets of individual chemicals and mixtures responsible for the toxicity; (2) to propose general QSAR models based on the structural features of the individual compounds and the corresponding mixtures; (3) to show the significant structural features in the individual compounds as well as in the mixture components and their effects towards modeled responses (pEC50).

2. Materials and methods

2.1. Toxicity data

2.1.1. Series 1 compounds and type 1 mixture data sets

The toxicity endpoint values of Series 1 compounds [13 aldehydes (AHs), 11 cyanogenics (CGs), 10 sulfonamides (SAs), and Trimethoprim (TMP)] (Fig. F1 and Fig. F2 of Supplementary Material 1), measured through the bioluminescence assay of a bioluminescent marine bacterium *P. phosphoreum*, were collected from the literature (Chung-Yuan and Ch'i-Feng, 1996; Jiang et al., 2010; Lin et al., 2003; Tian et al., 2012; Zou et al., 2012). The bioluminescence assay was performed using the diluted bacteria (maintained at agar slants at 4 °C) that had been cultured at 20 °C in yeast-tryptone-salts-glycerol broth for 12–14 h. Toxicity was measured with the Microtox test instrument (Model toxicity analyzer DXY-2) by quantifying the decrease in light emission from the bacteria after exposure to 3% NaCl solution of the test chemical for 15 min. Based on the decrease in light emission, EC_{50} (median effective concentration) was calculated by using the Probit model (Finney, 1971) and reported as negative logarithm of EC_{50} (pEC50). Toxicity data of series 1 compounds are enlisted in Table T1 of the Supplementary Material 1. The toxicity test of type 1 mixtures (obtained from the Series 1 compounds), performed in same manner as the individual chemicals, is discussed in the Eq. (3) as follows (Preston et al., 2000):

$$EC_{50M} = \frac{C_M}{\frac{C_A}{EC_{50A}} + \frac{C_B}{EC_{50B}}} \quad (3)$$

where, $EC_{50\ M}$ is the median effective concentration of sample mixture; C_A, C_B, and C_M are the concentrations of mixture component A, mixture component B, and the whole mixture respectively; EC_{50A} and EC_{50B} are the median effective concentration of components A and B respectively. Toxicity data of type 1 mixtures were obtained from the literature by (Yao et al., 2013) and presented in Table T3 of the Supplementary Material 1.

2.1.2. Series 2 compounds and type 2 mixture data sets

The toxicity endpoint values of Series 2 compounds [10 triazines (TAs) and 10 urea herbicides (UEs)] (Fig. F3 of Supplementary Material 1), measured through the toxicity assay of the freshwater microalgae *S. capricornutum*, were collected from the literature (Fairchild et al., 1997; Gaggi et al., 1995; Ma et al., 2006). The toxicity assay was performed using the diluted bacteria (maintained at agar slants at 4°C) that had been photo-autotrophically incubated in HB-4 medium (Li et al., 1959) in a rotator shaker (100 rpm) at 25 °C with illumination by cool white fluorescent light (light intensity = 450 E m^{-1} s^{-1}) in continuous mode (Ma et al., 2003; Verdisson et al., 2001). For toxicity test, 15-mL of aliquots of HB-4 medium containing 5×10^4 mL^{-1} algal cells was inoculated in 50-mL Erlenmeyer flask. Different concentrations of the test samples were added in different flasks and incubated for 96 h at the previously mentioned condition. Finally the optical density of the culture was measured with the spectrophotometer at 680 nm and compared with control sample (without any test sample). The growth inhibition was estimated from this comparative study and reported. Toxicity data of Series 2 compounds are enlisted in Table T2 of the Supplementary Material 1. The toxicity data of Type 2 mixtures (obtained from the Series 2 compounds), estimated in the same manner as the Type 1 mixtures, were obtained from the literature by (Yao et al., 2013) and presented in Table T4 of the Supplementary Material 1.

2.2. Molecular descriptors

Molecular descriptors are the quantitative values of structural features of molecules which are computed and used to correlate the molecular structure with the biological response (activity or toxicity) or chemical properties in a QSAR study. Various classes of descriptors based on dimensionality (e.g., 0 Dimensional- 7 Dimensional descriptors) can be used for the QSAR analysis. 2D descriptors are easy to calculate, easily interpretable, which can correlate structural features with the response. These have been extensively used in many published QSAR literature (Hoover et al., 2019; Seth et al., 2020). Hence, in this work, only 2D descriptors were used for the individual and mixture datasets. The 2D structures of 55 individual compounds (Series 1: 35 compounds; Series 2: 20 compounds) have been drawn using "MarvinSketch" (version 14.10.27) software (Marvin ChemAxon, 2013) with explicit hydrogen addition and proper aromatization for descriptor calculation. These structures have been used without energy minimization and conformational analysis, as we have used only 2D descriptors in our study. Total nine classes of 2D descriptors namely Constitutional indices, Ring descriptors, Functional group counts, Atom centered fragments, Atom type E-state indices, 2D atom pairs, Molecular properties, and Extended Topochemical Atom (ETA) indices have been calculated. All the classes of descriptors are easy to calculate and easily interpretable. The ETA descriptors (Roy and Ghosh, 2010) have been calculated using the "PaDEL-Descriptor" software tool (Yap, 2011) whereas the other eight classes of descriptors have been computed using "DRAGON 6″ (2011) software (Mauri et al., 2006). Descriptor generation of mixtures is not as simple as the individual chemicals. One way to calculate the descriptors for mixtures is to estimate the values of variables from individual components, and then, the obtained descriptor values from individual components are multiplied by their concentration ratio present in the mixture and finally added together to get the final descriptor value for mixtures. This strategy is known as weighted descriptor generation approach and has been used in many published QSAR analysis of mixtures (Khan et al., 2018; Rasulev et al., 2017). In the present work, the descriptors of mixtures have been calculated as following-

$$\text{Descriptors_of_molecule_M} : D_1{}^M, D_2{}^M, D_3{}^M, \ldots\ldots, D_n{}^M$$

$$\text{Descriptors_of_molecule_N} : D_1{}^N, D_2{}^N, D_3{}^N, \ldots\ldots, D_n{}^N$$

If the molecules M and N are mixed in a toxic ratio X:Y, then i^{th} descriptor values of the mixture MN will be calculated as per the following Eq. (4):

$$D_i{}^{MN} = D_i{}^M \times X + D_i^N \times Y \quad (4)$$

Here, Di^{MN} is the i^{th} descriptor of mixture MN, D_i^M is the i^{th} descriptor of component M, D_i^N is the i^{th} descriptor of component N, and $X + Y = 1$.

2.3. Dataset division

For the QSAR model development and external validation, each of the four datasets was divided into training and test sets. The training set compounds were used for model building whereas the test set compounds (marked with *in the datasets) were used for the rigorous external validation of the developed QSAR models. Four division methods and two statistical tools have been employed to improve the precision of QSAR models and to reduce the selection bias of the division techniques. Each dataset was divided using 1) the sorted activity based division; 2) Euclidean distance based division; 3) Kennard-Stone selection; and 4) Modified k-Medoids clustering technique (Roy et al., 2015b). The sorted activity based division method is a Y-response based strategy (activity based) whereas the other three are X-response based (structural features) strategies. From the 4 division methods, total 13 training-test set pairs (1 seed from Kennard-Stone selection, 4 seeds from Activity sorted response based divisions, 4 seeds from Euclidean distance based divisions, 4 seeds from k-Medoids clustering techniques) were obtained in each dataset. The first three approaches were executed using 'Dataset Division GUI 1.2′ tool and the last one was performed using 'Modified k-Medoids GUI 1.3′ tool available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. Preliminary models were developed from each division by using stepwise multiple linear regression (S-MLR) and examined through rigorous validation approaches. The sorted activity based division and Euclidean distance based division were chosen for the individual data sets and the mixture data sets, respectively based on the validation results from the four datasets and were taken forward for the further improvement of the quality of models.

2.4. Model development

Initially we reduced the descriptor pool from the best division of each dataset by using multistage Stepwise Multiple Linear Regression (S-MLR) method in 'MINITAB' software in multiple iterations. After the size reduction of the descriptor matrix, two statistical tools i.e., Genetic Algorithm (GA) (Leardi, 2001), and Best Subset Selection (BSS) (Roy et al., 2015c) were employed to select the best descriptor combinations. GA and BSS were carried out using 'DCVToolv2.0′ and 'BestSubsetSelection_v2.1′ tool respectively. Various combinations of descriptors were obtained from these tools and the final combinations were chosen on the basis of internal and external validation metrics values (see Section 2.5. Statistical Validation Metrics for details). The best combinations of descriptors from the 4 datasets were obtained from BSS which show acceptable values of validation parameters. Finally four partial least squares (PLS) models were constructed from the best

descriptor combinations by using 'PLS_SingleY_1.0' software tool. After development of the models, the predictivity of the models was checked by rigorous external validation using 'XternalValidationPlus_1.2' tool. The whole model development procedure is schematically represented in Fig. 1 and the size reduction of descriptor matrices of four datasets in each step of model development has been shown in Fig. 2.

The best subset selection is a strategy to optimize the best descriptor combinations in order to obtain the important structural information responsible for the toxicity (pEC50). This method can be used for comparatively smaller descriptor pools. For larger descriptor pools, a pool reduction is required as a prerequisite of BSS to save the computation time. BSS usually explores all possible descriptor combinations among which one is taken as the best on the basis of the statistical parameters. We performed BSS by using 'MLR Best Subset Selection' tool available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/.

Partial Least Squares regression is a generalization of MLR (Wold et al., 2001). It can be employed for strongly collinear, correlated, noisy, and numerous descriptors (X-variables). It can also model several response variables (Y-variables) simultaneously. The basic idea behind PLS regression is to extract the latent variables (LVs) considering for maximum of the manifest factor variation during modeling the response. This can be achieved by extracting the latent variables T (X-scores) and U (Y-scores) from descriptors and response, respectively. The X-scores are employed to predict the Y-scores which are used for constructing the response predictions.

2.5. Statistical validation metrics

Statistical validation metrics are employed for the evaluation of model quality, measurement of the fitness, robustness, stability and predictivity of the constructed models. The determination coefficient (R^2) and adjusted R^2 are the measures of goodness-of-fit (Roy et al., 2016). For an ideal model, the R^2 value should be equal to 1, and as it deviates from the ideal value, the quality of the model gets deteriorated. If we go on increasing the number of descriptors in a model with a constant number of observations, then R^2 will keep increasing as well but this will lead to a low statistical reliability and a decrease of degree of freedom. Calculation of adjusted R^2 considers the number of descriptors into account and it penalizes the models with large descriptor numbers. Internal validation parameters (based on the training set compounds) are computed to check the robustness of constructed models. It is determined by a cross-validated correlation coefficient Q^2_{LOO} (Leave-one-out). But the internal validation parameters are not sufficient to claim on the performance of a model for a new set of compounds. The external validation deals with the predictive ability of the model based on test set compounds. It can be determined by checking parameters like predictive R^2 (R^2_{Pred}) or Q^2_{F1} and Q^2_{F2} (Chirico and Gramatica, 2012). Additionally, the r^2_m metrics like average $r^2_{m(LOO)}$, $\Delta r^2_{m(LOO)}$,average $r^2_{m(test)}$, and $\Delta r^2_{m(test)}$were evaluated to check the internal, external, and overall predictivity (Roy et al., 2015b).

2.6. Applicability domain

The applicability domain (AD) is the theoretical region in chemical

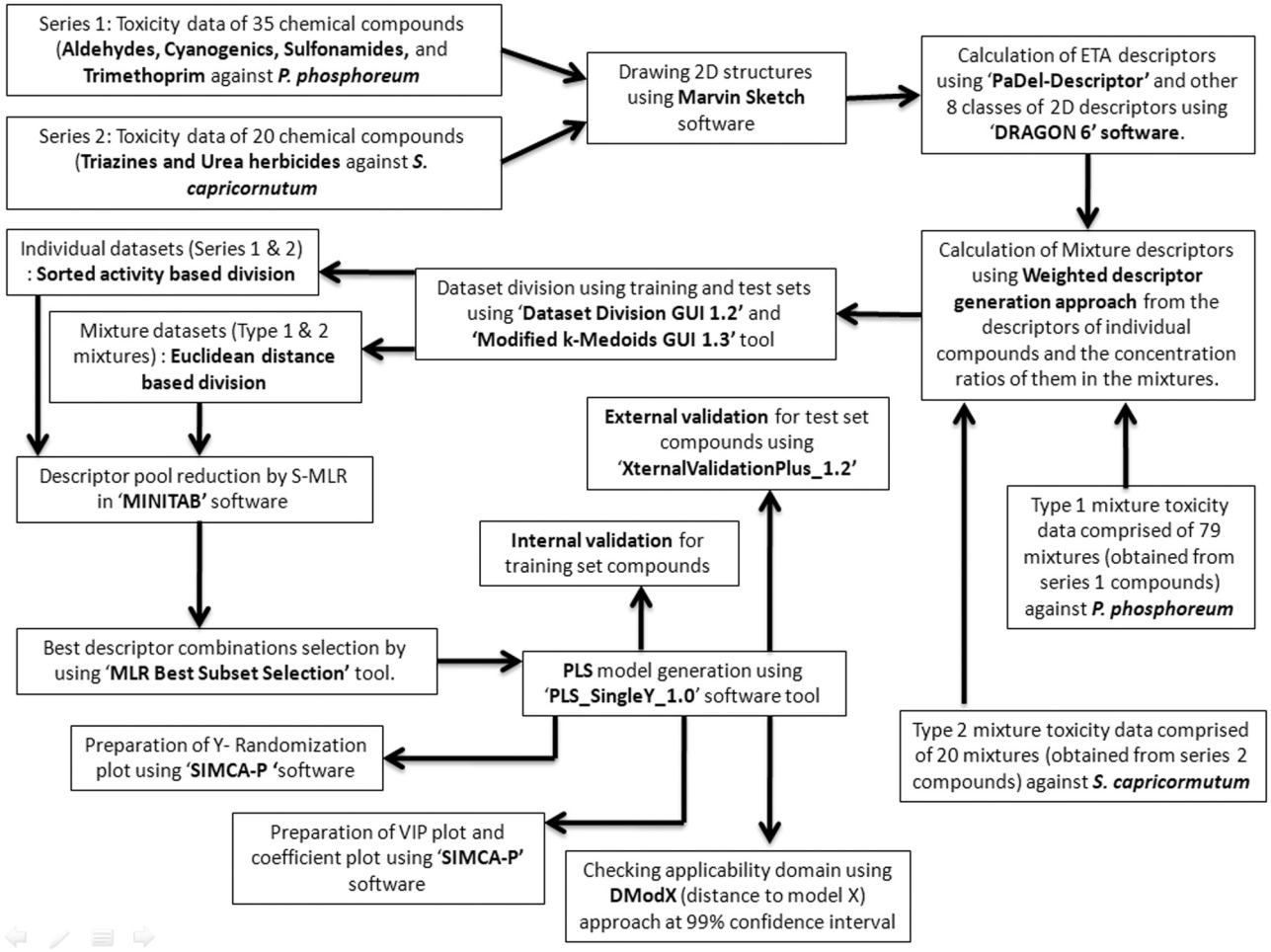


Fig. 1. Schematic representation of the model development process.

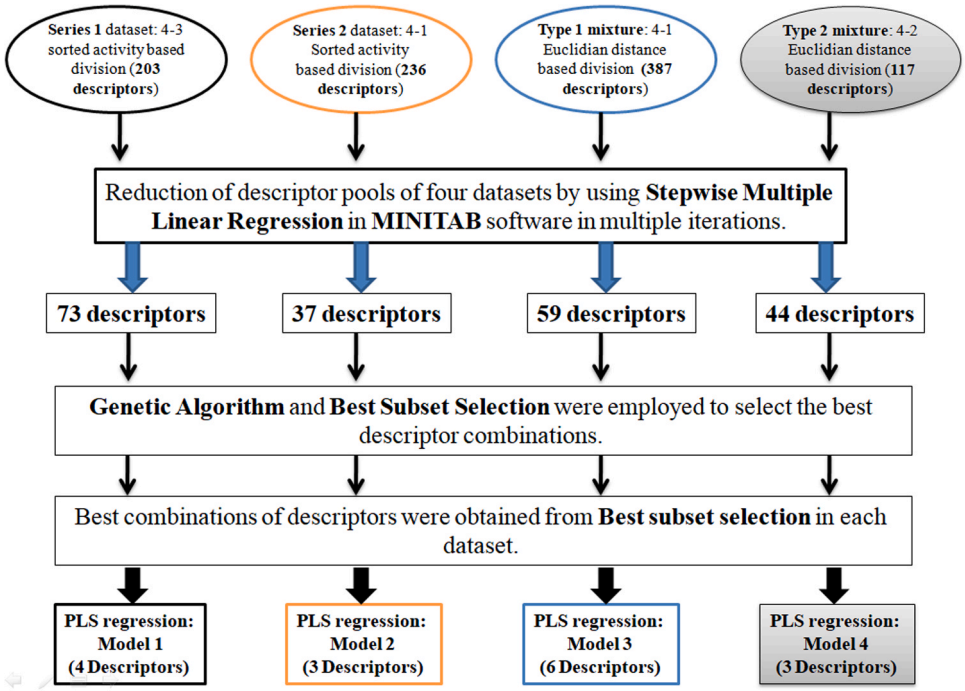


Fig. 2. Schematic representation of the size reduction of descriptor matrices of all data sets in each step of model development.

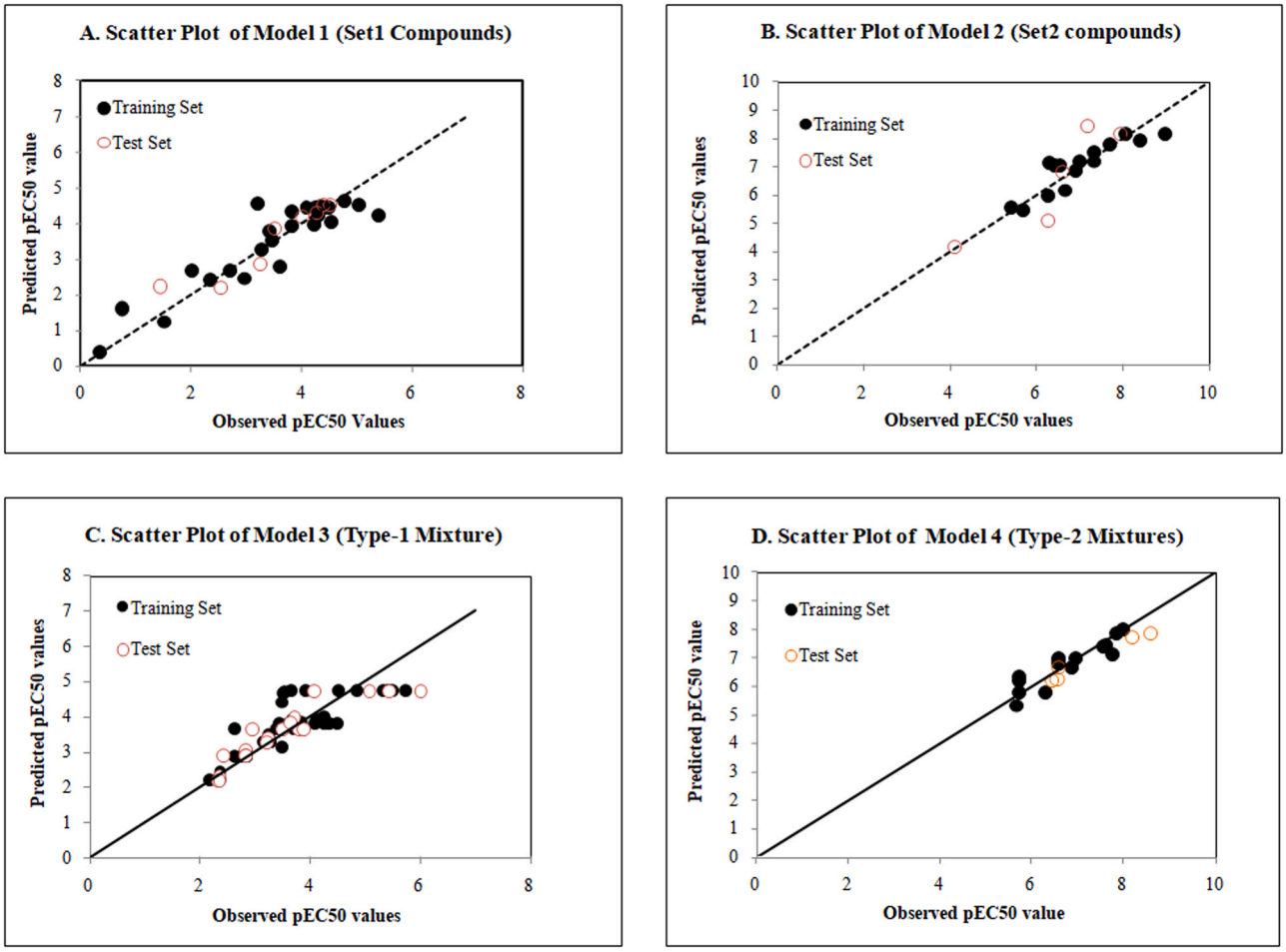


Fig. 3. Scatter plots of four PLS models; [A: Scattered plot of Model 1 (Set1 compounds); B: Scatter plot of Model 2 (Set2 compounds); C: Scatter plot of Model 3 (Type-1 mixtures); D: Scatter plot of Model 4 (Type-2 mixtures)].

space, defined by the model descriptors and the modeled response where the predictions are reliable (Gadaleta et al., 2016). According to the OECD (Organization of Economic Cooperation and Development) principle 3, the applicability domains of all QSAR models should be identified so that one can predict the untested compounds reliably with the developed model if they fall in the AD space of training set compounds. To comply with the OECD principle 3, we performed the AD assessment of all the studied models in this work. DModX approach is usually used to predict the applicability domain of PLS models and is defined in the Eq. (5) as follows (Wold et al., 2001)-

$$DModX = \frac{\sqrt{\frac{SSE_i}{K-A}}}{\sqrt{\frac{SSE}{(N-A-A_0)(K-A)}}} \quad (5)$$

For observation *i*, in a model with *A* components, *K* descriptors, and *N* observations, SSE is the squared sum of the residuals of X space. A_0 is 1 if the model is centered and 0 otherwise. If the DModX value is larger than around 2.5 times of the standard deviation (SD) of the X residuals (correspondent to an F value of 6.25), it signifies that the observation is outside the AD of the model considering a normally distributed data. The DModX has been performed at a confidence level of 99% using SIMCA-P software (Simca-P, 2002) for the applicability domain analysis. A threshold line is attributed as D-critical line and in this study, a confidence level of 99% has been used as the standard (D-Crit = 0.009999).

3. Result and discussion

Four PLS models were constructed from the four datasets mentioned in the materials and methods section. The first two models were for the individual chemical datasets (Models 1 and 2) whereas the next two were for the mixtures (Models 3 and 4). The details of latent variables and descriptors of the models are enlisted in Table T5 of Supplementary Material 1. All the models were evaluated rigorously by internationally accepted internal and external validation criteria. The models were robust and statistically sound considering the stipulated threshold values of the validation parameters (Table T6 of Supplementary Material 1). The scatter plots of observed vs. predicted response (pEC50) of the models are given in Fig. 3. The values of all descriptors appearing in Models 1–4 for training and test set compounds are given in Supplementary Material 2.

3.1. QSAR study of toxicity of individual chemicals (set 1) to P. Phosphoreum

Model 1 (Set 1) has been developed with four descriptors using three latent variables (LVs) and can be illustrated in the following Eq. (6).

$$pEC50 = -1.00784 + 2.08722 \times B04[C-C] - 7.3775 \times X5A - 2.36678 \times ETA_dBetaP + 6.70342 \times Mv \quad (6)$$

$$N_{train} = 27;\ R^2 = 0.84; R^2_{adj} = 0.81; Q^2_{LOO} = 0.78\ (3LV);\ avgr^2_{m\ (LOO)} = 0.68; \Delta r^2_{m\ (LOO)} = 0.18; MAE_{Train} = 0.42 N_{test} = 8; Q^2{}_{F1} = 0.87; Q^2{}_{F2} = 0.8;\ avgr^2_{m\ (test)} = 0.79; \Delta r^2_{m\ (test)} = 0.11; MAE_{Test} = 0.27$$

Model (1) showed acceptable values of validation metrics (both internal and external). The R^2, R^2_{adj}, and LOO-Q^2 values are 0.84, 0.81, and 0.78 (3 LVs),respectively signifying good statistical reliability and robustness of the model. The predictive ability of the model was judged by means of predictive R^2 (R^2_{pred}) or Q^2_{F1}, which shows good predictive ability of the model ($Q^2_{F1} = 0.87$).

A regression coefficient plot (Wold et al., 2001) (Fig. F4 of Supplementary Material 1) suggests the positive and negative contribution of descriptors towards the modeled response (pEC50). The descriptors with positive contributions (B04[C-C], Mv) suggest an increase in the response with an increase in the values of descriptors whereas descriptors with negative contributions (X5A, ETA_dBetaP) suggest the inverse relationship. The variable importance plot (VIP) (Fig. F4 of Supplementary Material 1) signifies the importance of each descriptor obtained in the final PLS model to the modeled response (pEC50). The descriptors having VIP score > 1 show higher statistical significance towards toxicity of the compounds (Akarachantachote et al., 2014); and the variables are arranged in a descending order of importance from left to right of the plot. The significance of the descriptors towards modeled response (Model 1) is in following order: B04[C-C] > X5A >ETA_dBetaP>Mv. B04[C-C] is highly significant descriptor of this model. The rest three descriptors are having almost similar level of significance towards the response.

The most contributing descriptor towards the toxicity of compounds is B04[C-C] which is basically a 2D atom pair descriptor, which checks whether two carbon atoms present at topological distance 4 or not. It has a positive correlation with the response which signifies an increase of the response in presence of C…C at the topological distance 4 and vice-versa. Fig. F5 of Supplementary Material 1 depicts that propionaldehyde (Comp. No. 2) and succinonitrile (Comp. No. 22), which are devoid of C…C at the topological distance 4, have lower pEC50 values than valeraldehyde (Comp. No. 4) and phthalonitrile (Comp. No. 23) with C…C at the topological distance 4.

The next important descriptor X5A is the average connectivity index of order five which has a negative contribution towards the modeled response. In Fig. F6 of Supplementary Material 1, it has been clearly depicted that pEC50 value gets lowered with an increase in the X5A value from α-Hydroxyisobutyronitrile (Comp.No. 16) to succinonitrile (Comp. No. 22).

ETA_dBetaP ($\Delta\beta'$) is an important feature of Model 1 (Set1), which is a 2nd generation Extended Topochemical Atom (ETA) descriptor. It is a measure of unsaturation content relative to the molecular size. It has a negative contribution towards the modeled response. The mathematical form of this descriptor is depicted in Fig. F7 of Supplementary Material 1,which clearly suggests the proportional relationship of ETA_dBetaP with the unsaturation content ($\Delta\beta$) and inversely proportional relationship of this particular descriptor with molecular size (N*v*). Among the Set1 compounds, sulfapyridine (Comp. no. 26) and sulfachloropyridazine (Comp. No. 31) have similar unsaturation count, but the latter compound has a higher molecular size and correspondingly lower ETA_dBetaP value resulting in a higher pEC50 value. A drastic decrease of pEC50 value has been observed from lactonitrile (Comp.No. 24) to succinonitrile (Comp.No. 22) with an increase in the unsaturation content of succinonitrile which is clearly depicted in Fig. F7 of Supplementary Material 1.

Mv (Mean atomic van der Waals volume) is the least important descriptor, which is a measure of molar volume. The positive contribution of this descriptor suggests an importance of molar bulk to the response as seen in more bulky compounds like *p*-bromobenzaldehyde (Comp. No. 9) or benzyl cyanide (Comp.No. 19). In contrast, less bulky compounds like *p*-chlorobenzaldehyde (Comp. No. 8) and benzonitrile (Comp. No. 18) have lower Mv values and hence lower pEC50 values also (Fig. F8 of Supplementary Material 1).

3.2. QSAR study of toxicity of individual chemicals (set 2) to S. Capricornutum

The developed Model 2 (Set2 Compounds) is composed of three descriptors and 2 latent variables (LV). The PLS model (Model 2) is shown in the Eq. (7):

$$pEC50 = 5.8289 - 2.62934 \times Hy + 0.4797 \times F07[C-C] + 0.73373 \times B04[N-Cl] \quad (7)$$

$$N_{Train} = 15; R^2 = 0.80; R^2_{adj} = 0.75;\ Q^2LOO = 0.65(2LV);\ avgr^2_{m\ (LOO)} = 0.56; \Delta r^2_{m\ (LOO)} = 0.10; MAE_{Train} = 0.46$$

$$N_{Test} = 5; Q^2_{F1} = 0.68; Q^2_{F2} = 0.61; avg\ r^2_{m\ (test)} = 0.64;\ \Delta r^2_{m(test)} = 0.17; MAE_{Test} = 0.58$$

This model showed acceptable values of validation metrics (both internal and external). The coefficient of determination R^2, adjusted coefficient of determination R^2_a, and the cross-validated correlation coefficient LOO-Q^2values are 0.80, 0.75, and 0.65 (2 LVs), respectively signifying good statistical reliability and robustness of the model. The predictive ability of the model was judged by means of predictive R^2 (R^2_{pred}) or Q^2_{F1}, which shows good predictive ability of the model (Q^2_{F1} = 0.68). The regression coefficient plot of Model 2 (Fig. F9 of Supplementary Material 1) suggests the positive contributions of F07[C-C] and B04[N-Cl] descriptors whereas negative contribution of the Hy descriptor to the response. Among the modeled descriptors, F07[C-C] is the most important one towards the response, showing a VIP score higher than 1. The next two important descriptors are Hy and B04[N-Cl] respectively (Fig. F9 of Supplementary Material 1).

F07[C-C] is the frequency of presence of two carbon atoms at the topological distance seven. It is positively correlated with the modeled response which means that an increase inpEC50 value will be observed if the F07[C-C] descriptor value increases and vice-versa. Simazine (Comp. No. 40) and atrazine (Comp. No. 36) of Set2 are similar to each other in structure. However, as simazine has a less F07[C-C] value than that of atrazine, the toxicity of the former is also less. The occurrences of two carbons at the topological distance 7 in both the compounds have been depicted in Fig. F10 of Supplementary Material 1.

Toxicity of chemicals to microorganisms mainly depends on the penetration into the cell and the effective binding to the target protein. Hy is the next significant descriptor of Model 2; this is a measure of hydrophilic factor. More hydrophilic nature is detrimental to the penetration of organic molecules into cell through the hydrophobic cell membrane and hence Hy is negatively correlated with the modeled response. This negative correlation can be justified if we compare the descriptor values of prometon (Comp. No. 39) and simazine (Comp. No. 40). We can see the presence of two H-bond donor groups (-NH) in both of the compounds. However, due to the presence of more sterically hindering groups (-isopropyl) in prometon, the chances of H-bond formation are less in prometon and hence the corresponding hydrophilic factor will also be less. In contrast, because of the presence of less sterically hindering substituent (-ethyl), simazine has more chances to form H-bond and hence be more hydrophilic. In another pair of compounds, chlorotoluron (Comp. No. 47) and chlorosulfuron (Comp. No. 54), an increase of hydrophilic factor as observed with an increase in the number of H-bond donor groups in chlorosulfuron results in a corresponding decrease in the toxicity (Fig. F11 of Supplementary Material 1).

B04[N-Cl] is the least important descriptor which is negatively correlated with the modeled response. This descriptor identifies the presence of nitrogen and chlorine atoms at the topological distance four. The B04[N-Cl] values, e.g., 1 and 0, suggest the presence and absence respectively of the N and Cl at the topological distance four. In the dataset, compounds like metribuzine (Comp. No. 41) and monuron (Comp. No. 49) do not have this feature, and hence are more toxic. On contrary, compounds like terbuthylazine (Comp. No. 44) and chlorotoluron (Comp. No. 47) having nonzero descriptor values are less toxic. Therefore, the negative correlation of B04[N-Cl] with pEC50 value is demonstrated (Fig. F12 of Supplementary Material 1).

3.3. QSAR study of toxicity of mixtures (type-1) to P. Phosphoreum

For the Type-1 mixture dataset, we have developed a PLS regression model (Model 3) with 2 latent variables. This model is composed of six descriptors and can be illustrated in the following Eq. (8):

$$pEC50 = 4.75139 - 0.00012 \times B03[N-O] - 1.58803 \times nR = Cp - 0.16606 \times nROH - 0.94445 \times N-074 - 0.35592 \times C-006 + 0.10023 \times F07[O-O] \quad (8)$$

$$N_{Train} = 60; R^2 = 0.76; R^2_{adj} = 0.73; Q^2LOO = 0.70(2LV); avgr^2_{m(LOO)} = 0.60; \Delta r^2_{m\ (LOO)} = 0.13; MAE_{Train} = 0.35$$

$$N_{Test} = 19; Q^2_{F1} = 0.82; Q^2_{F2} = 0.82; avgr^2_{m\ (test)} = 0.70;\ \Delta r^2_{m\ (test)} = 0.13;\ MAE_{Test} = 0.31$$

The well accepted internal validation parameter values (R^2 = 0.76, R^2_{adj}= 0.73, Q^2_{LOO}= 0.70) suggest the statistical reliability and robustness of this model. Model 3 has good predictive ability also as it shows acceptable values of external validation parameters (Q^2_{F1}= 0.82, Q^2_{F2}= 0.82). Among the modeled descriptors, B03[N-O], nR=Cp, nROH, N-074, and C-006 have negative contributions whereas only F07 [O-O] has a positive contribution towards the modeled response as suggested by the regression coefficient plot of Model 3 (Fig. F13 of Supplementary Material 1). The variable importance plot (Fig. F13 of Supplementary Material 1) of this model suggests the order of importance of the modeled descriptors and the order is: N-074 > nR=Cp> B03 [N-O]> C-006 > F07[O-O]> nROH.

N-074 is an atom centered fragment descriptor and the most significant one to demonstrate the role of structural features of mixture components concerning the toxicity end point (pEC50) in this PLS model. This descriptor computes the number of –R#N/-R=N- fragments into the molecule (#: triple bond; =: double bond; R: any group linked through carbon atom). Mixture components like malononitrile (Comp. No. 14) and glyconitrile (Comp. No. 15) have this structural feature, whereas propionaldehyde (Comp. No. **2**) and benzaldehyde (Comp. No. 5) have no such fragment in their structures (Fig. 4). Hence mixtures containing malononotrile or glyconitrile have non-zero descriptor values (e.g., Mix. No. 2, 14). It has been observed that a mixture (e.g. Mix. No. 2) with a higher descriptor value is less toxic than the mixture (e.g., Sr. No. 14) with a lower descriptor value which clearly justifies the negative correlation of N-074 with the modeled response (Table T7 of Supplementary Material 1).

nR=Cp is another important descriptor of Model 3. It is a functional group count descriptor which computes the number of terminal sp^2-hybridized carbon atom in the molecular structure of corresponding compounds. The negative contribution of this descriptor towards the response suggests that a mixture with lower descriptor value is more toxic than a mixture with higher descriptor value. Type-1 mixtures like 38, 40, 41, and 15 are the equimolar (1:1) mixtures of the Set 1 compounds (Table T8 of Supplementary Material 1). Among the components of above mentioned mixtures, allyl cyanide (Comp. No. 17) contains terminal C(sp^2) while the rest have no such terminal carbon (Fig. 4). It has been noted that only the allyl cyanide containing mixtures (e.g. Sr. No. 38, 40, 41) have non-zero descriptor values and they show similar and less toxicity than the mixtures (e.g. Sr. No. 15) with a lower (zero) descriptor value.

B03[N-O] is a 2D atom pair descriptor which describes the presence or absence of nitrogen and oxygen atoms at the topological distance 3. This descriptor is negatively correlated with the modeled response which means that the pEC50 value rises with an increase in the descriptor value and vice-verse. Type-1 mixtures such as 67, 72, 75, and 78 are the mixtures of Set 1 compounds sulfamethoxazole (Comp. No. 27) and trimethoprim (Comp. No. 35) in different toxic ratios (Table T9 of Supplementary Material 1). Sulfamethoxazole contains a pair of nitrogen and oxygen at the topological distance 3 in its molecular structure, whereas trimethoprim is completely devoid of this feature (Fig. 4). It has been observed that the B03[N-O] values get increased with an increase in the percentage of sulfamethoxazole in the mixtures and the

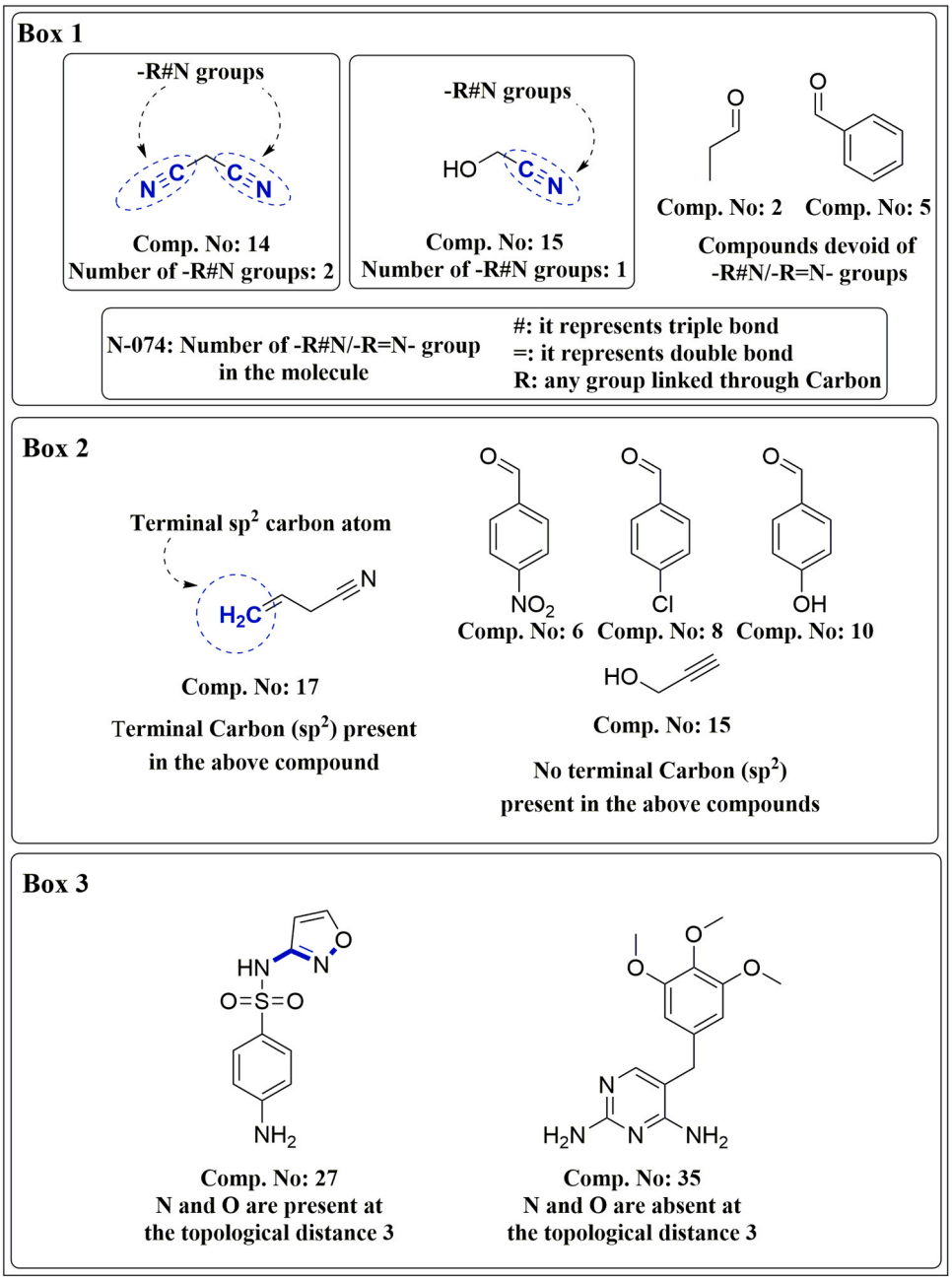


Fig. 4. Depiction of feature N-074, nR=Cp, and B03[N-O] in the mixture components (Box 1 represents N-074; Box 2 represents nR=Cp; Box 3 represents B03[N-O]).

corresponding toxicity values get decreased (e.g.,Mix. No. 67, 72, and 75). However, the presence of higher percentage of trimethoprim than sulfamethoxazole in the mixture does not affect the descriptor value as seen in mixture 78.

C-006 is the next important descriptor of Model 3 which is an atom centered fragments count, and it particularly suggests the number of $–CH_2RX$ fragment (R: any group linked through C; X: heteroatoms such as O, N, P, S, Se, halogens) in the molecule. The Set 1 compounds except glycolonitrile (Comp. No. 15) are devoid of this feature (Fig. 5). C-006 is negatively correlated with the response which can be clearly demonstrated by the data enlisted in Table T10 of Supplementary Material 1. The mixtures with higher descriptor values contain glycolonitrile as a component as seen Mixture no. 15 and 18. It has been observed that the mixtures with higher descriptor values (e.g., Mix. No. 15 and 18) showed less toxicity than the mixtures with lower descriptor values (Mix. No. 27 and 52).

F07[O-O] is again a 2D atom pair descriptor of the developed Model 3. It is the frequency count of oxygen atom pairs at the topological distance 7. Among the type-1 mixture components, *p*-nitrobenzaldehyde (Comp. No. **6**) and *p*-terephthaldehyde (Comp. No. 7) have this feature in their structures (Fig. 5) and hence the mixtures containing these two components (e.g., Mix. No. 27 and 54) have non-zero descriptor values. F07[O-O] is the only descriptor of Model 3 which have positive contribution towards modeled response. It has been observed that the toxicity increases with an increase in the descriptor value (Table T11 of Supplementary Material 1) which clearly justifies the positive contribution of this descriptor.

The least important descriptor of Model 3 is nROH which is a functional group count descriptor. It computes the number of hydroxyl groups present in the chemical structures of compounds. It is a negatively contributing descriptor, i.e., the mixtures with lower descriptor values are more toxic than the mixtures with lower descriptor values.

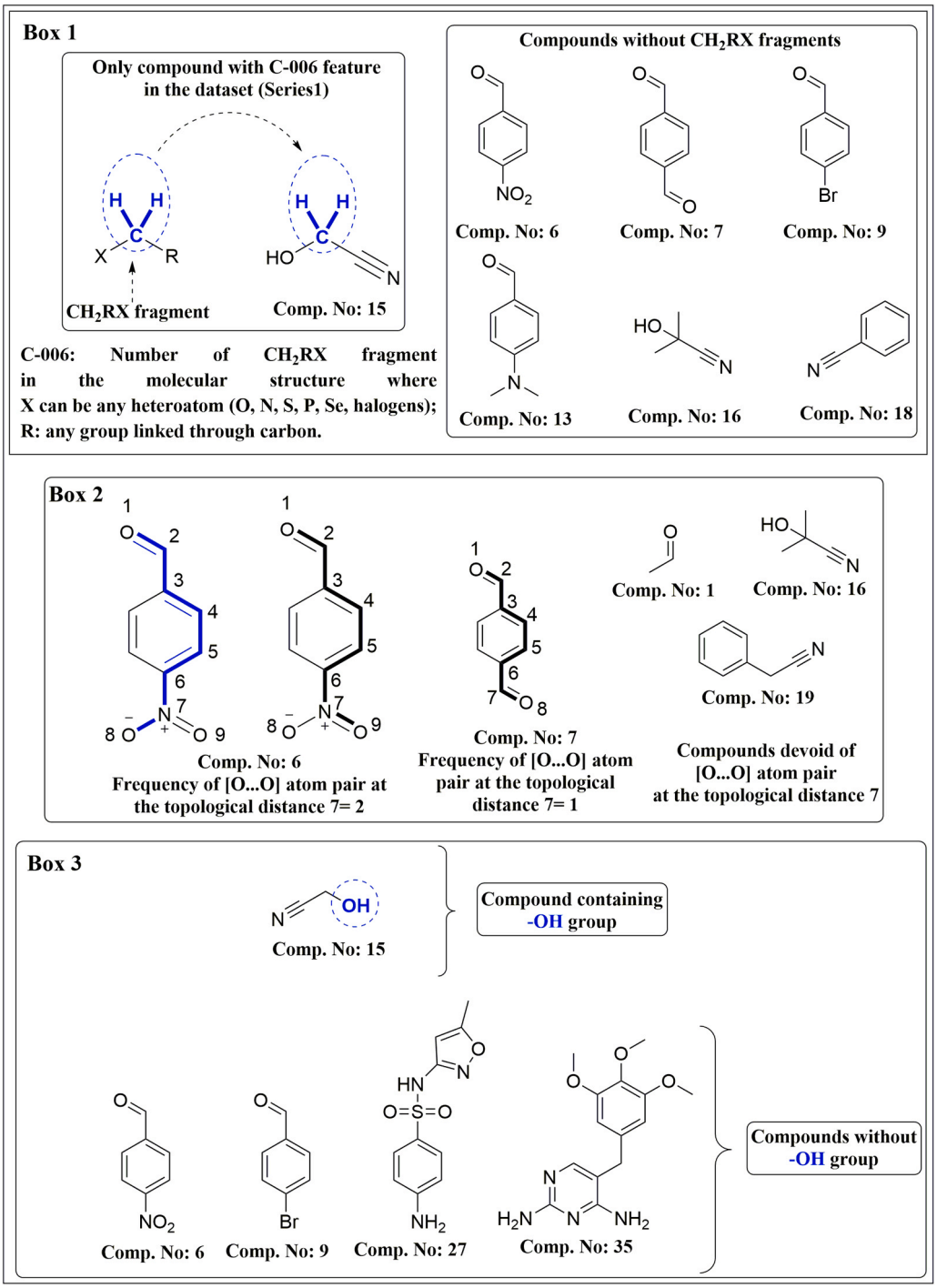


Fig. 5. Depiction of feature C-006, F07[O-O], and nROH in the mixture components (Box 1 represents C-006; Box 2 represents F07[O-O]; Box 3 represents nROH).

This feature can be identified in type-1 mixture component like glycolonitrile (Comp. No. 15) (Fig. 5). Table T12 of Supplementary Material 1 demonstrates that the mixtures (e.g., Mix No. 15, 18) with higher nROH values have lower toxicity as compared to the mixtures (e.g., Mix. No. 69) with lower descriptor values, and these observations clearly address the negative correlation of nROH with pEC50.

3.4. QSAR study of toxicity of mixtures (type-2) to S. Capricornutum

The 4th and final model has been composed of 3 descriptors with 2 latent variables and is shown in Eq. (9):

$$pEC50 = -12.38041 + 1.34717 \times ETA_Alpha - 0.25978 \times F06[C-S] + 0.13686 \times nCsp3 \quad (9)$$

$N_{Train} = 15; R^2 = 0.82; R^2adj = 0.72; Q^2LOO = 0.73\ (2LV); avgr^2_{m\ (LOO)} = 0.64; \Delta r^2_{m\ (LOO)} = 0.09; MAE_{Train} = 0.35$

$N_{Test} = 5; Q^2_{F1} = 0.82; Q^2_{F2} = 0.77; avgr^2_{m\ (test)} = 0.79; \Delta r^2_{m\ (test)} = 0.04; MAE_{Test} = 0.37$

This model was constructed for the Type-2 mixture dataset. The acceptable values of validation parameters (R^2= *0.82;* R^2adj= *0.72;* Q^2LOO= *0.73*) confirm the statistical reliability and robustness of Model

4. Among the three descriptors, ETA_Alpha and $nCsp^3$ have positive contributions whereas F06[C-S] has a negative contribution towards the toxicity of the mixtures (Fig. F14 of Supplementary Material 1). As per the variable importance plot of this model (Fig. F14 of Supplementary Material 1), the significance level of the model descriptors are found to be in the following order: ETA_Alpha> F06[C-S]> $nCsp^3$.

ETA_Alpha (∑α), a 1st generation Extended Topochemical Atom (ETA) descriptor, is a measure of molecular bulk of compounds. Molecular bulk is an important feature of chemical compounds for penetrating the cell wall of microorganisms and at the same time for binding to the receptor also. ∑α has a maximum and positive contribution to the response which suggests a proportional relationship between ∑α and pEC50. Mixture No. 82 and 98 are two equimolar Type-2 mixtures containing monolinuron (Comp. No. 48) as a constant component. It has been observed that Mix. No. 82 has a higher ∑α value and correspondingly a higher pEC50 value due to the presence of more bulky 2nd component (Comp. No. 36) than Mix. No. 98 (Fig. 6). Another pair of mixtures (Mix. No. 84 and 87) have also shown the proportional relationship of ETA_Alpha and the toxicity.

The next important descriptor of Model 4 is F06[C-S]. It is a 2D atom pair descriptor which calculates the number of occurrences of carbon-sulphur atom pair at the topological distance six. It is the only negative contributing descriptor of Model 4; so, it has an inversely proportional relationship to the modeled response. The feature has been identified in the mixture component like symmetry (Comp. No. 38) and is depicted in Fig. 7. Components having [C-S] atom pair at the topological distance 6 have the key effect to raise the descriptor values as seen in Table T13 of Supplementary Material 1. We have seen that more toxic mixtures (e.g., Mix. No. 92) have lower descriptor values than less toxic mixtures (e.g., Mix. No. 88).

The last significant descriptor of Model 4 is $nCsp^3$ which computes the number of sp^3- hybridized carbon atoms. It is a constitutional descriptor which has a positive contribution to the toxicity of type-2 mixtures. Two observations have been depicted in Fig. 7 where the sp^3- hybridized carbon atoms are marked in the mixture components. An increase in pEC50 value has been observed with an increase in the descriptor value (e.g., Mix. nos. 99 and 95).

3.5. Applicability domain of the developed PLS models

From Fig. F15 of Supplementary Material 1, we can see that the test set compounds of each developed model have fallen into the applicability domain of corresponding training sets at a 99% confidence interval (D-critical = 0.009999).

3.6. Randomization of the developed PLS models

We have also analyzed the models by randomization plots in order to confirm that the developed models are not the results of any chance correlation. In the randomization approach, a number of models are constructed randomly by permuting different combinations of descriptors (X) or response (Y) variables based on the fit of already

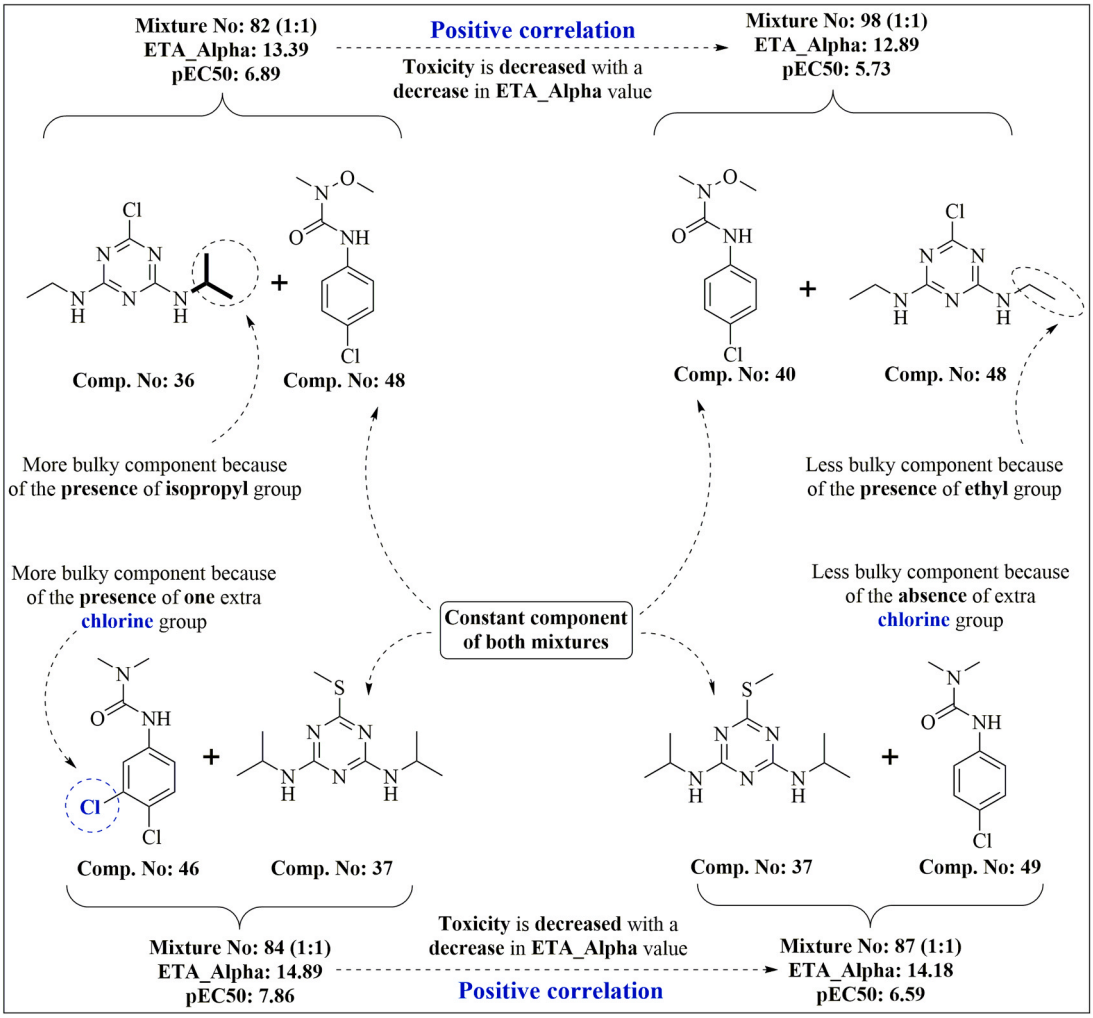


Fig. 6. Contribution of ETA_Alpha to pEC50 of Type-2 mixtures.

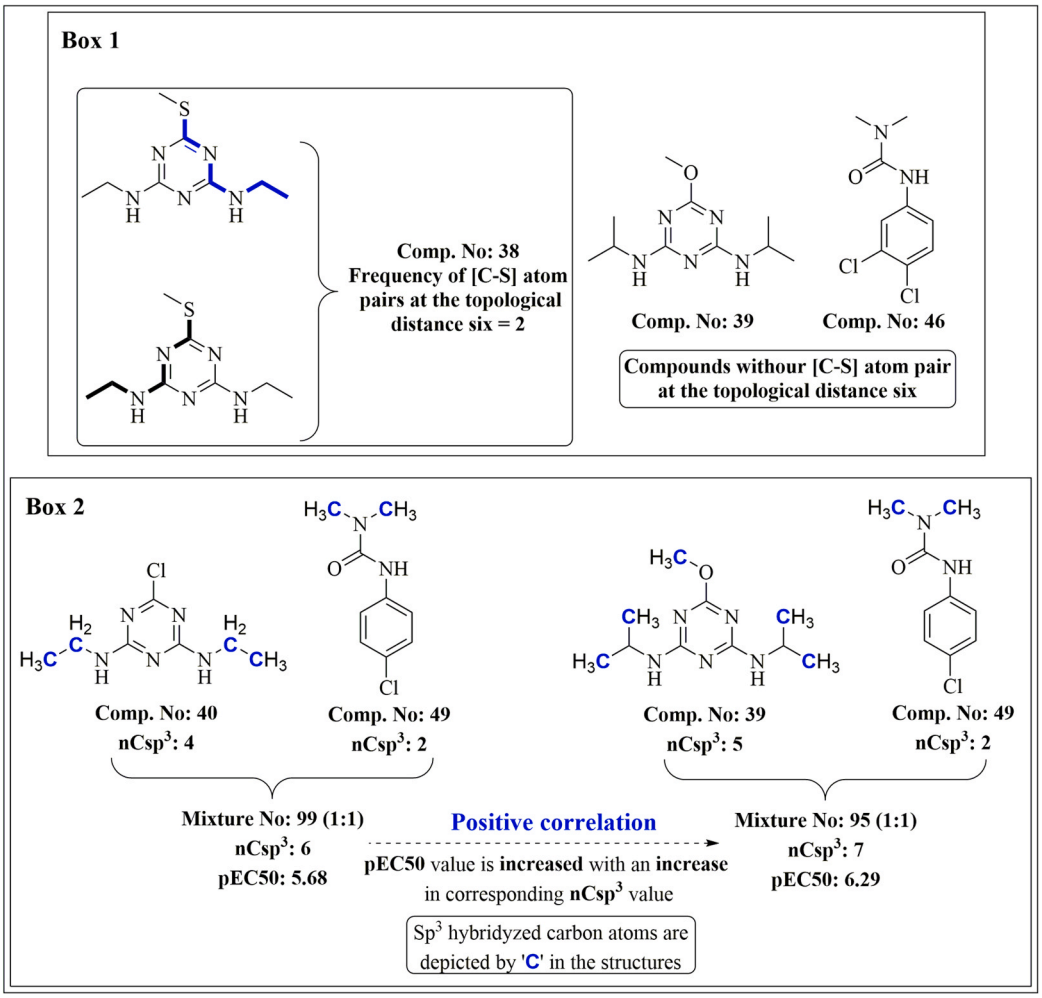


Fig. 7. Depiction of features F06[C-S] and nCsp^3 in type 2 mixture components (Box 1 represents F06[C-S]; Box 2 represents nCsp^3).

recorded models (Topliss and Edwards, 1979). In this work, for the training set, the X variable data remained unaltered and the Y data were shuffled randomly (Y-randomization). The recorded model was fitted to the permuted data and compared with the best fit. The R^2 and Q^2 values of the random models (Y-axis) are plotted against the correlation coefficient between the original Y values and the permuted Y values (X axis). The statistical metrics (R^2 and Q^2) of the randomization model should be poor (R_Y^2 intercept <0.3; Q_Y^2 intercept <0.05) and not within the acceptable range of those for regression models. Otherwise the constructed models may be considered as the results of chance correlation (Rücker et al., 2007). Here we have used 100 permutations for the analysis and the obtained plots of four models have been shown in Fig. F16 of Supplementary Material 1. The R_Y^2 intercept and Q_Y^2 intercept values have been mentioned in Table T14 of Supplementary Material 1, which suggest that all the models are statistically significant and not the results of any chance correlation.

4. Conclusion

Prediction of toxicity of various individual chemical data sets and corresponding mixture data sets were accomplished by PLS (partial least squares) regression method. The 2D descriptors were successfully used to measure the contribution of structural features of individual chemicals as well as the components of mixtures to overall toxicity. In addition, the weighted descriptor generation strategy was successfully employed to calculate the mixture descriptors in this work. It has been found out that the structural features associated with the toxicity for individual chemicals have no significant contribution for the toxicity of mixtures. Presence of multiple bonded Nitrogen and terminal sp^2 hybridized carbon atom in mixture components can help to reduce the toxicity of binary mixtures against marine bacterium. In case of fresh water organisms, the molecular bulk of mixture components can enhance the toxicity. The statistical validation results show that the developed models are robust and highly predictive and not the results of any chance correlation as suggested by the Y-randomization test. We should mention that we could not include the data from all types of functional groups in this study due to the scarcity of toxicity data; however, these models can be used efficiently for the prediction of aquatic toxicity and risk assessment of the chemical mixtures under their domain of applicability.

CRediT authorship contribution statement

MC: Data curation, Methodology, Validation, Investigation, Writing - original draft. **KR:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

There are no conflicts to declare.

Acknowledgement

MC thanks to All India Council of Technical Education (AICTE), New

Delhi, India for financial support in the form of National Doctoral Fellowship (NDF). KR thanks SERB, Govt of India for financial assistance under the MATRICS scheme (MTR/2019/000008).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2020.124936.

References

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

Altenburger, R., Nendza, M., Schüürmann, G., 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ. Toxicol. Chem. 22, 1900. https://doi.org/10.1897/01-386.

Aquatic Toxicity [WWW Document], 2016. URL ⟨https://www.chemsafetypro.com/Topics/CRA/ecotox_aquatic_toxicity.html⟩ (accessed 11.23.20).

Chirico, N., 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. J. Chem. Inf. Model. 52, 2044–2058. https://doi.org/10.1021/ci300084j.

Chung-Yuan, C., Ch'i-Feng, H., 1996. Toxicity of organic mixtures containing cyanogenic toxicants. Environ. Toxicol. Chem. 15, 1464–1469. https://doi.org/10.1002/etc.5620150906.

De Zwart, D., Slooff, W., 1983. The Microtox as an alternative assay in the acute toxicity assessment of water pollutants. Aquat. Toxicol. 4, 129–138. https://doi.org/10.1016/0166-445X(83)90050-4.

Escher, B.I., Baumer, A., Bittermann, K., Henneberger, L., König, M., Kühnert, C., Klüver, N., 2017. General baseline toxicity QSAR for nonpolar, polar and ionisable chemicals and their mixtures in the bioluminescence inhibition assay with Aliivibrio fischeri. Environ. Sci. Process. Impacts 19, 414–428. https://doi.org/10.1039/c6em00692b.

Evans, R.M., Martin, O.V., Faust, M., Kortenkamp, A., 2016. Should the scope of human mixture risk assessment span legislative/regulatory silos for chemicals? Sci. Total Environ. 543, 757–764. https://doi.org/10.1016/j.scitotenv.2015.10.162.

Fairchild, J.F., Ruessler, D.S., Haverland, P.S., Carlson, A.R., 1997. Comparative sensitivity of Selenastrum capricornutum and Lemna minor to sixteen herbicides. Arch. Environ. Contam. Toxicol. 32, 353–357. https://doi.org/10.1007/s002449900196.

Finney, D., 1971. Probit Analysis. Chembridge University Press,, London, UK.

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

Gaggi, C., Duccini, M., Bacci, E., Sbrilli, G., Bucci, M., Naby, A.M.H. El, 1995. Toxicity and hazard ranking of *s* -triazine herbicides using microtox® two green algal species and a marine crustacean. Environ. Toxicol. Chem. 14, 1065–1069. https://doi.org/10.1002/etc.5620140618.

Hoover, G., Kar, S., Guffey, S., Leszczynski, J., Sepúlveda, M.S., 2019. Chemosphere In vitro and in silico modeling of per fl uoroalkyl substances mixture toxicity in an amphibian fi broblast cell line. Chemosphere 233, 25–33. https://doi.org/10.1016/j.chemosphere.2019.05.065.

Hübner, Y., Hoettges, K.F., Hughes, M.P., 2003. Water quality test based on dielectrophoretic measurements of fresh water algae Selenastrum capricornutum. J. Environ. Monit. 5, 861–864. https://doi.org/10.1039/b309131g.

Jiang, L., Lin, Z., Hu, X., Yin, D., 2010. Toxicity prediction of antibiotics on luminescent bacteria, photobacterium phosphoreum, based on their quantitative structure-activity relationship models. Bull. Environ. Contam. Toxicol. 85, 550–555. https://doi.org/10.1007/s00128-010-0157-z.

Kar, S., Leszczynski, J., 2019. Exploration of computational approaches to predict the toxicity of chemical mixtures. Toxics 7, 15. https://doi.org/10.3390/toxics7010015.

Khan, P.M., Kar, S., Roy, K., 2020. Ecotoxicological QSARs of mixtures. In: Methods in Pharmacology and Toxicology. Humana Press Inc., pp. 437–475. https://doi.org/10.1007/978-1-0716-0150-1_19

Khan, P.M., Rasulev, B., Roy, K., 2018. QSPR modeling of the refractive index for diverse polymers using 2D descriptors. ACS Omega 3, 13374–13386. https://doi.org/10.1021/acsomega.8b01834.

Leardi, R., 2001. Genetic algorithms in chemometrics and chemistry: a review. J. Chemom. 15, 559–569. https://doi.org/10.1002/cem.651.

Li, S., Zhu, H., Xia, Y., Yu, M., Liu, K., Ye, Z., Chen, Y., 1959. Mass culture of unicellular green alga. Acta Hydrobiol. Sin. 4, 121–129.

Lin, Z., Zhong, P., Yin, K., Wang, L., Yu, H., 2003. Quantification of joint effect for hydrogen bond and development of QSARs for predicting mixture toxicity. Chemosphere 52, 1199–1208. https://doi.org/10.1016/S0045-6535(03)00329-1.

Loewe, S., 1953. The problem of synergism and antagonism of combined drugs. Arzneimittelforschung 3, 285–290.

Loewe, S., Muischnek, H., 1926. Uber kombinationswirkungen. Naunyn Schmiedeberg's Arch. Pharm. 114, 313–326.

Logan, D.T., Wilson, H.T., 1995. An ecological risk assessment method for species exposed to contaminant mixtures. Environ. Toxicol. Chem. 14, 351–359. https://doi.org/10.1002/etc.5620140222.

Lu, G.H., Wang, C., Wang, P.F., Chen, Z.Y., 2009. Joint toxicity evaluation and QSAR modeling of aromatic amines and phenols to bacteria. Bull. Environ. Contam. Toxicol. 83, 8–14. https://doi.org/10.1007/s00128-009-9694-8.

Ma, J., Lin, F., Wang, S., Xu, L., 2003. Toxicity of 21 herbicides to the green alga Scenedesmus quadricauda. Bull. Environ. Contam. Toxicol. 71, 594–601. https://doi.org/10.1007/s00128-003-8521-x.

Ma, J., Wang, S., Wang, P., Ma, L., Chen, X., Xu, R., 2006. Toxicity assessment of 40 herbicides to the green alga Raphidocelis subcapitata. Ecotoxicol. Environ. Saf. 63, 456–462. https://doi.org/10.1016/j.ecoenv.2004.12.001.

Martens, M., Mosselmans, G., Fmero, S., Jacobs, G., Lafontaine, A., 1984. Some thoughts on a possible regulatory approach at EEC level on the classification and labeling of dangerous preparations. Regul. Toxicol. Pharmacol. 4, 145–156. https://doi.org/10.1016/0273-2300(84)90037-0.

Marvin ChemAxon [WWW Document], 2013. URL ⟨https://chemaxon.com/products/marvin⟩ (accessed 6.16.20).

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: an easy approach to molecular descriptor calculations. MATCH Commun. Math. Comput. Chem. 56, 237–248.

Preston, S., Coad, N., Townend, J., Killham, K., Paton, G.I., 2000. Biosensing the acute toxicity of metal interactions: are they additive, synergistic, or antagonistic? Environ. Toxicol. Chem. 19, 775–780. https://doi.org/10.1002/etc.5620190332.

Rasulev, B., Jabeen, F., Stafslien, S., Chisholm, B.J., Bahr, J., Ossowski, M., Boudjouk, P., 2017. Polymer coating materials and their fouling release activity: a cheminformatics approach to predict properties. ACS Appl. Mater. Interfaces 9, 1781–1792. https://doi.org/10.1021/acsami.6b12766.

Ribo, J.M., Kaiser, K.L.E., 1987. Photobacterium phosphoreum toxicity bioassay. I. Test procedures and applications. Environ. Toxicol. Water Qual. 2, 305–323. https://doi.org/10.1002/tox.2540020307.

Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom. Intell. Lab. Syst. 152, 18–33. https://doi.org/10.1016/j.chemolab.2016.01.008.

Roy, K., Ghosh, G., 2010. Exploring QSARs with extended topochemical atom (ETA) indices for modeling chemical and drug toxicity. Curr. Pharm. Des. 16, 2625–2639. https://doi.org/10.2174/138161210792389270.

Roy, K., Kar, S., Das, R.N., 2015a. Backgr. QSAR Hist. Dev., Underst. Basics QSAR Appl. Pharm. Sci. Risk Assess. https://doi.org/10.1016/b978-0-12-801505-6.00001-6.

Roy, K., Kar, S., Das, R.N., 2015b. Valid. QSAR Models, : Underst. Basics QSAR Appl. Pharm. Sci. Risk Assess. 231–289. https://doi.org/10.1016/B978-0-12-801505-6.00007-7.

Roy, K., Kar, S., Das, R.N., 2015c. Sel. Stat. Methods QSAR, : Underst. Basics QSAR Appl. Pharm. Sci. Risk Assess. 191–229. https://doi.org/10.1016/b978-0-12-801505-6.00006-5.

Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–2357. https://doi.org/10.1021/ci700157b.

Schwarzenbach, R.P., Escher, B.I., Fenner, K., Hofstetter, T.B., Johnson, C.A., Von Gunten, U., Wehrli, B., 2006. The challenge of micropollutants in aquatic systems. Science 313, 1072–1077. https://doi.org/10.1126/science.1127291.

Seth, A., Ojha, P.K., Roy, K., 2020. QSAR modeling with ETA indices for cytotoxicity and enzymatic activity of diverse chemicals. J. Hazard. Mater. 394, 122498 https://doi.org/10.1016/j.jhazmat.2020.122498.

Tian, D., Lin, Z., Yin, D., Zhang, Y., Kong, D., 2012. Atomic charges of individual reactive chemicals in binary mixtures determine their joint effects: an example of cyanogenic toxicants and aldehydes. Environ. Toxicol. Chem. 31, 270–278. https://doi.org/10.1002/etc.1701.

Topliss, J.G., Edwards, R.P., 1979. Chance factors in studies of quantitative structure-activity relationships. J. Med. Chem. 22, 1238–1244. https://doi.org/10.1021/jm00196a017.

U. Simca-P, 10.0 [WWW Document], 2002. URL ⟨www.umetrics.com⟩.

Verdisson, S., Couderchet, M., Vernet, G., 2001. Effects of procymidone, fludioxonil and pyrimethanil on two non-target aquatic plants. Chemosphere 44, 467–474. https://doi.org/10.1016/S0045-6535(00)00468-9.

Villa, S., Vighi, M., Finizio, A., 2014. Experimental and predicted acute toxicity of antibacterial compounds and their mixtures using the luminescent bacterium Vibrio fischeri. Chemosphere 108, 239–244. https://doi.org/10.1016/j.chemosphere.2014.01.046.

Wang, T., Tang, L., Luan, F., Nat, M., Cordeiro, D.S., 2018. Prediction of the toxicity of binary mixtures by QSAR approach using the hypothetical descriptors. IJMS 19, 3423. https://doi.org/10.3390/ijms19113423.

Wang, T., Zhou, X., Wang, D., Yin, D., Lin, Z., 2012. Using molecular docking between organic chemicals and lipid membrane to revise the well known octanol-water partition coefficient of the mixture. Environ. Toxicol. Pharmacol. 34, 59–66. https://doi.org/10.1016/j.etap.2012.02.008.

Ward, T.J., Rausina, G.A., Stonebraker, P.M., Robinson, W.E., 2002. Apparent toxicity resulting from the sequestering of nutrient trace metals during standard Selenastrum capricornutum toxicity tests. Aquat. Toxicol. 60, 1–16. https://doi.org/10.1016/S0166-445X(01)00259-4.

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.

Yang, R., Thomas, R., Gustafson, D., Campain, J., Benjamin, S., Verhaar, H., Mumtaz, M., 1998. Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling. Environ. Health Perspect. 106, 1385–1393.

Yao, Z., Lin, Z., Wang, T., Tian, D., Zou, X., Gao, Y., Yin, D., 2013. Using molecular docking-based binding energy to predict toxicity of binary mixture with different

binding sites. Chemosphere 92, 1169–1176. https://doi.org/10.1016/j.chemosphere.2013.01.081.
Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474. https://doi.org/10.1002/jcc.21707.
Zou, X., Lin, Z., Deng, Z., Yin, D., Zhang, Y., 2012. The joint effects of sulfonamides and their potentiator on photobacterium phosphoreum: differences between the acute and chronic mixture toxicity mechanisms. Chemosphere 86, 30–35. https://doi.org/10.1016/j.chemosphere.2011.08.046.



SAR and QSAR in Environmental Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gsar20

Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

M. Chatterjee & K. Roy

To cite this article: M. Chatterjee & K. Roy (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 and QSAR in Environmental Research, 33:6, 463-484, DOI: 10.1080/1062936X.2022.2081255

To link to this article: https://doi.org/10.1080/1062936X.2022.2081255

View supplementary material

Published online: 31 May 2022.

Submit your article to this journal

Article views: 429

View related articles

View Crossmark data

Citing articles: 3 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gsar20

SAR AND QSAR IN ENVIRONMENTAL RESEARCH
2022, VOL. 33, NO. 6, 463–484
https://doi.org/10.1080/1062936X.2022.2081255

Taylor & Francis
Taylor & Francis Group

Check for updates

Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures

M. Chatterjee and K. Roy

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

ABSTRACT

The quantitative structure–activity relationship (QSAR) modelling of mixtures is not as simple as that for individual chemicals, and it needs additional care to avoid overestimation of the performance. In this research, we have developed a 2D-QSAR model using only 2D interpretable and reproducible descriptors to predict the aquatic toxicity of mixtures of polar and non-polar narcotic substances present in the environment. Partial least squares (PLS) regression has been used to model the response variable (log 1/EC50 against *Photobacterium phosphoreum*) and the structural features of 84 binary mixtures of polar and nonpolar narcotic toxicants complying with the Organization of Economic Co-operation and Development (OECD) protocols. The model was cross-validated by mixtures-out and compounds-out cross-validation to nullify the developmental bias. The reliability of prediction of the model has been judged by the Prediction Reliability Indicator (PRI) tool using a newly designed set. The new model is robust, reproducible, extremely predictive, easily interpretable, and can be used for reliable prediction of aquatic toxicity of any untested chemical mixtures within the applicability domain. We have additionally used a machine learning-based chemical read-across algorithm in this study to improve the quality of predictions for the toxicity of the mixtures with the modelled descriptors.

ARTICLE HISTORY
Received 31 March 2022
Accepted 19 May 2022

KEYWORDS
Aquatic toxicity; chemical mixtures; QSAR model; cross-validation

Introduction

Aquatic toxicity is the unwanted and harmful effects on the aquatic organisms and ecosystems at different trophic levels due to the exposure of various manufactured chemicals, anthropogenic, and natural materials [1]. It is one of the burning issues of the 21st century due to rapid chemical advancement and the industrial revolution. The waterbodies are being polluted from various sources that affect the aquatic ecosystems adversely. The chemical and industrial wastes contaminate both the freshwater and marine waterbodies. The various major sources of contamination are inappropriate waste management in industries, overuse or misuse of agrochemicals, pharmaceuticals, nanomaterials, heavy metals in various sectors of society, and discharge of various bio-

CONTACT K. Roy kunal.roy@jadavpuruniversity.in; kunalroy_in@yahoo.com
Supplemental data for this article can be accessed at: https://doi.org/10.1080/1062936X.2022.2081255
© 2022 Informa UK Limited, trading as Taylor & Francis Group

wastes, etc., [2]. Recently, Yang et al. have claimed that micropollutant mixtures are the leading cause of aquatic toxicity instead of a single chemical [3]. In the open environment, single chemical exposure is a myth [4]; as per the report published by Evans et al., the environmental ecosystems and human beings are always being exposed to multicomponent chemical mixtures, and the number and variety of such combinations are ever-changing [5]. The nature of toxicity poses by the chemical mixtures is not similar to the individual toxicants due to the probable additivity and/or interactions among mixture components [4]. It has also been found that the chemicals present in their unobserved effective concentration (NOEC) can produce toxicity in presence of co-contaminants [6]. Therefore, the environmental mixtures pose a threat to the environmental ecosystems and mankind also if not assessed properly.

The regulatory authorities have acknowledged the significance of toxicity assessment of chemical mixtures for the past few decades [7] and formulated some guidelines for risk assessment of chemical mixtures [8,9]; however, the scarcity of sufficient toxicity data of chemical mixtures is still evident [6]. There are some conventional experimental techniques for toxicity estimation of mixtures such as concentration addition (CA) and independent action (IA). These methods essentially consider the additive effect of the mixture components and neglect the chances of interactions among components. Apart from these, CA and IA models need preliminary experimentations to get the toxicity values of the component chemicals that also limits the applications of these methods to some extent for associated complexities [10]. To overcome these issues, the regulatory agencies and the research institutions across academia and industries have focused on various new approach methodologies (NAM) in recent years [11]. NAMs by definition include in vitro, in chemico, or in silico methods that aim to support hazard assessment, improve understanding of toxic effects and reduce animal experimentations or replace them entirely with advanced animal-free testing approaches [11–13]. Among various alternative approaches for the toxicity assessment of chemicals (including mixtures), in silico modelling NAMs have shown promising results. The computer-aided techniques include quantitative structure–activity relationships (QSARs), read-acrosses, structural alerts, etc. The QSAR is a statistical modelling approach that reveals the relationship between toxicity (molecular property in general) and molecular structure. It is one of the most popular and mostly used in silico approaches in predictive toxicology. The regulatory agencies like the United States Environmental Protection Agency (US-EPA), the Agency for Toxic Substance and Disease Registry (ATDSR), Organization of Economic Co-operation and Development (OECD), and European regulations like Registration, Evaluation, Authorization, and Restrictions of Chemicals (REACH) legislation, etc., have recommended the use of few QSARs for environmental toxicology and risk assessment. Read-across is another in silico approach, which is increasingly used nowadays for toxicity prediction and data gap filling [14]. It is a similarity-based approach that can be used for toxicity prediction of mixtures as well.

Though QSAR is a leading alternative for predicting the toxicity of a single chemical, it is still in the developing stage for the mixture toxicity assessment. The generation of mixture QSAR is more challenging than the conventional QSARs of individual chemicals. In QSAR of pure chemicals, structural exclusivity of the test set (necessary for external validation) is attained by random/similarity-based positioning of data points into training and test sets. In case of mixture QSAR, the same component of a test mixture may be

present in another mixture, because mixtures are essentially the combinations of more than one pure chemical with different concentrations; therefore, structural exclusivity cannot be achieved just by random positioning them between training and test sets. As a result, the developed model can be biased and the model's predictive ability can be overestimated [15]. One may argue about the utilization of the mixture descriptors here, but the same component chemical in both the sets as different mixtures may influence the quality measures of the model. These considerations inspire us to start applying more rigorous strategies for the validation of QSAR models of mixtures. In this current study, we have developed a properly validated mixture QSAR model for the prediction of aquatic toxicity from the dataset of 84 binary mixtures of polar and non-polar narcotic toxicants [16]. We have taken the negative logarithm of median effective concentration (log $1/EC_{50}$) of binary mixtures against marine bacteria *Photobacterium phosphoreum* as the endpoint (Y-variable) of this study. *P. phosphoreum* is a type of gram-negative bioluminescence bacteria that lives in symbiosis with marine organisms and is a classical marker of marine toxicants [17,18]. The objectives of the current research are- (1) to identify the important structural features of the mixtures of polar and non-polar narcotic toxicants responsible for aquatic toxicity, (2) to develop and propose a new mixture QSAR model following the OECD protocols for the prediction of aquatic toxicity, (3) to cross-validate the model for more rigorous validation and for avoiding the development bias of mixture QSAR [15,19], (4) to interpret the modelled descriptors mechanistically for better understanding, (5) to improve the quality of predictions of binary mixtures using a machine learning-based chemical read-across algorithm.

Materials and methods

Toxicity data

The toxicity data were collected from published literature [16], where log $1/EC_{50}$ (negative logarithm of median effective concentration in the mol/l scale) of 84 binary mixtures were reported against the bioluminescent marine bacterium *Photobacterium phosphoreum*. The binary mixtures were constructed from four environmental non-polar narcotic toxicants (chlorobenzene: #1; 1,2,3-trichlorobenzene: #2; bromobenzene: #3; 1-bromo-2,3-dichlorobenzene: #4) and 4 environmental polar narcotic toxicants (phenol: #5; 2,4-dichlorophenol: #6; aniline: #7; 3,4-dichloroaniline: #8) as classified by [20]. Each combination of mixtures was reported thrice in three specific ratios (0.5:1; 1:1; 1:0.5). Among the 84 binary mixtures, 18 mixtures (6 combinations) were constructed from only the non-polar narcotic toxicants (#1 to #4), 18 mixtures (6 combinations) were constructed from only polar narcotic toxicants (#5 to #8), and the rest 48 mixtures (16 combinations) were constructed from both the non-polar and polar narcotic toxicants (#1 to #8). The toxicity data of all mixtures have been enlisted in Table S1 in the Supplementary Material.

The toxicity study was reported to be performed [16] in Microtox test instrument (Model toxicity analyser DXY-2) using the freeze-dried *Photobacterium phosphoreum* (T_3 mutation). The organism was reconstituted in agar slants at 4°C and used in the bioluminescence assays. The assays were performed in yeast-tryptone-salts-glycerol broth culture, maintained at 20°C for 12–14 hours. Three per cent NaCl solutions of test samples were applied to the bacteria culture and the intensity of the luminance was measured

after 15 min of exposure. Each mixture was tested thrice to check its reproducibility [16]. The EC_{50} was computed using the Probit model [21] from the measured luminance and reported as log $1/EC_{50}$ (mol/l).

Molecular descriptors

Molecular descriptors are defined as the numerical values of the structural features of molecules that are essentially used for correlating the molecular structure with different molecular properties (activity/toxicity/property) in the QSAR study [22,23]. There are various classes of descriptors available which can be used for QSAR analysis. Based on dimensionality, descriptors can be classified into eight classes (0 dimensional to 7 dimensional) [18]. The lower-dimensional descriptors (0D and 1D) represent various constitutional features of molecules such as molecular weight, number of atoms, number of bonds, etc. Two-dimensional descriptors need 2D molecular graph representation for calculations, and they include topological descriptors, 2D atom pairs, connectivity index, etc. Three-dimensional descriptors are of higher-order and need the pretreatment of 3D structures (energy minimization and conformational analysis) for calculations. Electronic parameters, spatial parameters, WHIM descriptors, molecular surface area parameters, etc., are examples of 3D descriptors. Four-dimensional to seven-dimensional descriptors usually represent various interactive parameters of ligand and receptors, solvation circumstances, induced-fit parameters, etc. These are computationally more exhaustive, tough to interpret, and time-consuming for computation, thus they are used in much less extent in conventional QSAR. Among all classes, 0D to 2D descriptors are the most simple, easy to calculate, and easily interpretable descriptors that can efficiently correlate the structural properties with the molecular properties. These descriptors have been used in various previously published QSAR studies also [24,25]. Therefore, for the convenience of work, different subclasses of 0D to 2D descriptors (constitutional indices, ring descriptors, molecular properties, connectivity index, 2D atom pairs, functional group counts, atom type-E-state indices, atom-centred fragments, and extended topochemical atom indices) have been utilized in the present study.

Computation of mixture descriptors is not an easy task and cannot be obtained directly as software output. Mixture descriptors were calculated in two steps in this work. Firstly, the previously mentioned descriptors of individual mixture components were calculated using alvaDesc software [26]. Before descriptor calculation, 2D structures (including explicit hydrogen) of mixture components were drawn and properly aromatized using Marvin Sketch software [27]. The 2D structures were taken into the descriptor calculation step without prior energy minimization or conformational analysis (which is required only for higher dimensional descriptors). After getting the descriptors of individual components, mixture descriptors have been obtained by using the weighted descriptor generation approach using a Python code available in laboratory [28]. According to this principle, one can get the mixture descriptors by summing up the multiplications of descriptor value of mixture components and the mass ratio of the corresponding components in the mixture. For simplification, the weighted descriptor generation approach has been mathematically defined in the following equation (1) [6,24]:

$$D_iMix = \sum^{D_i} Component \times X_{Component} \quad (1)$$

where D_iMix is i^{th} the descriptor of a mixture; $D_iComponent$ is the i^{th} descriptor of a specific component, $X_{Component}$ is the mass ratio of the corresponding component, and $\sum X_{Component}$ is always equal to 1 for a mixture.

Dataset division

Dataset division is an integral part of QSAR modelling. A dataset is usually divided into training and test sets for model development and validation. The training set is used for model building, whereas the test set is used for rigorous external validation. The compositions of both sets are kept mutually exclusive in classical QSAR analysis so that the developed model can be properly validated by using a set of molecules (test set) completely different from the training set. In mixture QSARs, the values of mixture descriptors are unique and that is why the test and training sets of mixture data attain mutual exclusiveness in quantitative terms. In the present study, the mixture data have been divided into training and test sets by utilizing four statistical data division tools namely, (1) Euclidean distance-based division, (2) Kennard-Stone division, (3) Sorted response-based division, (4) Modified k-Medoids clustering. Among these four techniques, the sorted response-based division is a Y-variable-based strategy, whereas the rest are X–variable-based strategies. The modified k-medoids clustering was performed in the 'Modified k-Medoid GUI 1.3' tool and the rest techniques were executed by the 'Dataset Division GUI 1.2' tool available at https://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. From these four distinct methods, 13 types of division (Kennard-Stone division: 70% training set and 30% test set; 4 seeds of Euclidean distance-based divisions: 75% training set and 25% test set; 4 seeds of sorted response-based divisions: 75% training set and 25% test set; and 4 seeds of modified k-medoids divisions: 75% of training set and 25% test set) were obtained. Based on the validation metrics (Q^2_{LOO}, Q^2_{F1}) obtained from the preliminary stepwise multiple linear regression (S-MLR) models, the fourth seed—four clustered (every fourth data point of each cluster is taken into the test set) sorted response-based division was chosen for further analysis.

Feature selection and model development

After getting the best division among all, the descriptor pool was initially reduced by a multilayered (multiple iterations of) stepwise regression algorithm in the MINITAB software. Two feature selection techniques namely, Genetic algorithm [29] and Best subset selection [30] were used afterwards for identifying the most important structural features of the descriptor matrix responsible for mixture toxicity. The Genetic algorithm and Best subset selection were executed by 'DCVTool_v2.0' and 'BestSubsetSelection_v2.1', respectively, and various descriptor combinations were obtained. The best combination was chosen based on the statistical metrics values (section 'Internal and external validation'). The best combination of descriptors was obtained from Best subset selection and at the end, a partial least square (PLS) regression model was developed by 'PLSSingleY_v1.0' tool from the selected descriptors. Another tool 'XternalValidationPlus_v1.2' was used for the rigorous external validation of the

constructed model with the test set of mixtures. various PLS plots such as variable importance plot, regression coefficient plot, applicability domain plot, Y-randomization plot, etc., were developed in 'SIMCA-P' software for visual interpretation of the developed model. The complete model development process has been schematically represented in Figure 1.

The best subset selection is one of the feature selection strategies available for the QSAR study. It finds out the best descriptor combination to identify the important structural features responsible for toxicity (biological activity in general). It explores all possible descriptor combinations and that is why it is a computationally exhaustive process. Therefore, to minimize the computation time, a reduced pool of descriptors should be given as input instead of a whole descriptor matrix. The best combination of descriptors is selected based on the computed values of internal and external validation metrics provided by the tool itself. We performed this process in the 'BestSubsetSelection_v2.1' tool available at http://dtclab.webs.com/software-tools/ and http://teqip.jdvu.ac.in/QSAR_Tools/.

Partial least squares (PLS) regression is the generalized form of the multiple linear regression method [31]. It is more advanced than MLR and can be used for correlated, strongly collinear, and noisy data. PLS can handle numerous X variables (descriptors) and various Y variables (responses) simultaneously. As an algorithm, PLS extracts the latent variables (LVs) considering maximum variation in the descriptor matrix during modelling the response. It can be achieved by calculating the X-scores (latent variable T) and Y-score (latent variable U) from the X and Y variables respectively. The T vectors ate utilized to get the U vectors whereas the U vectors are further employed to get the

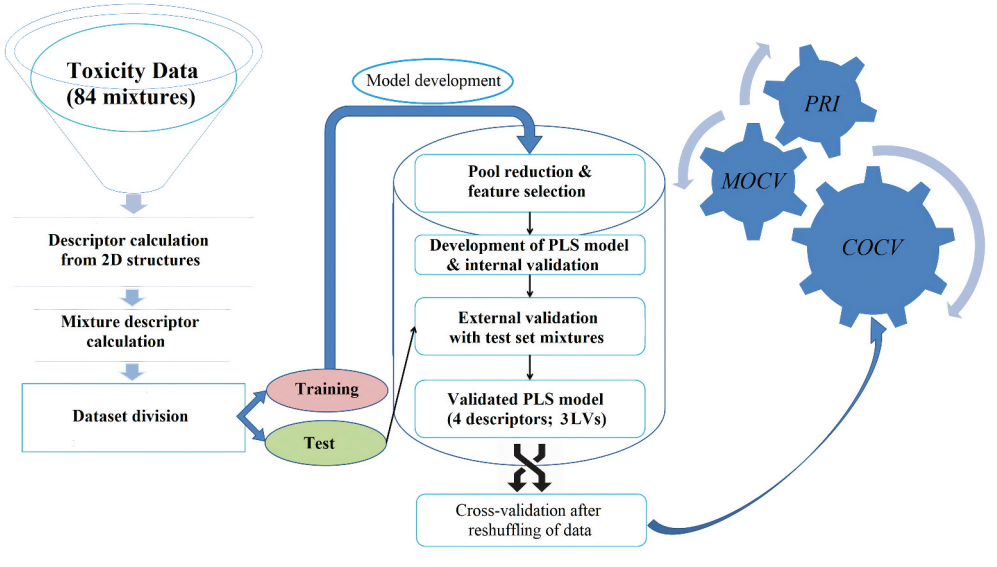


Figure 1. Schematic representation of the model development process and validation. (COCV: compounds out cross-validation; MOCV: mixtures out cross-validation; PRI: prediction reliability indicator).

response prediction. During reporting a PLS model, the number of LVs is very important and it should not be equal to or more than the number of descriptors in the model.

Internal and external validation

Internal and external validation tests are executed for the judgment of model quality, measurement of the fitness, robustness, reproducibility, and predictive ability of the developed models. The internal validation metrics (r^2, r^2_{adj}, Q^2_{LOO} etc.) are employed to check the goodness-of-fit and robustness of the model, whereas the external validation metrics (Q^2_{F1}, Q^2_{F2}, $RMSE_P$ etc.) are employed to judge the overall predictivity. The fitness of the model is judged by measuring the determination coefficient (r^2) and adjusted r^2 [32]. The quality of a model gets deteriorated with the deviation of r^2 from its ideal value 1. The value of r^2 depends upon the number of descriptors used in the model; hence, it increases with increasing the number of descriptors. This creates the problem of overfitting and low statistical reliability. Therefore, to avoid overfitting, the metric r^2_{adj} is employed. The computation of r^2_{adj} considers the number of X–variables, and hence it penalizes the model with large number of descriptors. The robustness of a model is adjudged by another internal validation metric namely, cross-validated correlation coefficient Q^2_{LOO} (Leave-one-out). The ideal value of Q^2_{LOO} is also 1 and any deviation from the ideal value suggests the deterioration of the model quality. Various error-based metrics such as, mean average error (MAE) and root mean square error of prediction ($RMSE_P$) are also used to check the error of predictions of a model and these metric values should be near 0. The internal validation metrics and the error-based metrics cannot suggest the predictive performance of the model on a new set of compounds; thus, external validation is used. External validation metrics are the measures of the predictive ability of a model and are computed from an external set of compounds called the test set. The external correlation coefficients like Q^2_{F1} (also termed as r^2_{Pred}) and Q^2_{F2} are employed to determine the predictivity of the model [33]. Some additional r_m^2 rank metrics (e.g. average $r_m^2{}_{(LOO)}$, $\Delta r_m^2{}_{(LOO)}$, average $r_m^2{}_{(test)}$, and $\Delta r_m^2{}_{(test)}$) are also calculated to evaluate the internal, external, and overall predictivity [32].

The Variable importance plot (VIP) is a 'SIMCA-P' plot that signifies the order of importance of descriptors towards modelled response. The descriptors having a VIP score of more than 1 are statistically more significant and they are arranged in descending order of VIP score in the graph from left to right [34]. Coefficient plots show the positive and negative relationships of the descriptors with the modelled response. This is essentially a bar diagram obtained from 'SIMCA-P' where positive coefficients can be identified by the upward bars and negative coefficients can be identified by the downward bars. The positive and negative coefficients signify the proportional and inverse relationships of the descriptors, respectively, with the response. Y-randomization is a process to check the developed model, whether it is a result of any chance correlation or not [35]. For that, the Y variables of the training set are shuffled randomly and X–variables remain unaltered. The newly formed model is then fitted with the permuted random models and compared with the best fit. The resulting r^2 and Q^2 values of the random models are plotted (Y-axis) against the correlation coefficient among original

Y values and permuted Y values (X-axis). Very poor r^2 and Q^2 values (<0.05) signify the statistical credibility of the developed model [35]. This Y-randomization plot was obtained in the present study from 'SIMCA-P' software and discussed in the Result and Discussion section.

The applicability domain is a theoretical portion of the chemical space defined by the modelled descriptors and response, in which the predictions are reliable [36]. The toxicity of a complete spectrum of chemical substances cannot be predicted with sufficient confidence with a single QSAR model due to their structural diversity, and that is why the QSAR modellers have to provide the applicability domain plot of the developed models so that the toxicity can be predicted with reliability. OECD (Organization of Economic Co-operation and Development) also suggests analyzing the applicability domain of QSAR models in principle 3 [37]; thus to comply with the guideline, we have carried out the AD analysis in this research. Among various methods of applicability domain analysis, DModX is used for the PLS models, and we used 'SIMCA-P' software to obtain the DModX plot. The DModX approach is defined in the following equation (2) [31]:

$$DModX = \frac{\sqrt{\frac{SSE_i}{K-A}}}{\sqrt{\frac{SSE}{(N-A-A_0)(K-A)}}} \quad (2)$$

Here, SSE is the squared sum of the residuals of X space, A_0 is 1 if the model is centred and 0 otherwise for observation *i*, *A* is the number of latent variables, *K* is the number of descriptors, and *N* is the number of data points. If the DModX score is higher than 2.5 times of the standard deviation of the X residuals (F value 6.25), it indicates that the observation is beyond the applicability domain of the particular model considering a normal data distribution. A threshold line is termed as a D-critical line and in this study, a confidence level of 99% has been applied as the standard (D-Crit = 0.00999898).

Cross-validation

Mixture QSAR is different from the conventional ones in various aspects, and one of the important aspects is the dataset division. As seen in the 'Molecular descriptors' section, each mixture descriptor is composed of the descriptors of two different chemicals; therefore, there is structural information of more than one pure compound in each data point of the mixture. This now leads to a question on structural exclusiveness between training and test sets during model development, because a single mixture component may be present in both test and training sets simultaneously as different mixtures. It may lead us to compute overestimated validation metrics and hence the quality of the developed model may be misleading. To overcome this problem, Oprisiu et al. suggested 3 different cross-validation (CV) strategies, namely, the Points-out cross-validation, Mixtures-out cross-validation, and the Compounds-out cross-validation [19] whereas, Muratov et al suggested the Everything out cross-validation exclusively for mixture QSARs [15].

To obtain the cross-validated metrics, the whole data set is again divided manually by following specific patterns of different cross-validations. These new training sets are used to develop new PLS models with the same combination of descriptors and the same number of latent variables as in the original PLS model; then these resultant models are

validated with the new test sets and the metrics are calculated. In mixtures out CV, the mixtures (those specific combinations of component chemicals) are unique among themselves in training and test sets. In other words, the test mixtures must not be present in the training set. Although in this cross-validation, the individual components may be present in both sets as separate mixtures, the mixtures are unique among themselves with respect to the new training and test sets. The new training-test pair for Mixtures-out CV has been tabulated in Table S2 in the Supplementary Material. The compounds-out CV suggests the partial structural exclusiveness between the training and test sets. The binary test mixtures must contain one component that is unique with respect to the training set. To perform this cross-validation, the data set was divided again and the new training-test pair has been tabulated in Table S3 in the Supplementary Material. The compounds-out CV has been achieved after keeping out the common mixture component of the test set (#5: phenol) from the training set. PLS regression has been carried out again twice with the same set of descriptors and same number of LVs with the new training-test pairs (which are shown in Tables S2 and S3 in the Supplementary Material) by 'PLSSingleY_v1.0' tool (available at https://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/.) to get the cross-validated metrics. These cross-validations have been pictorially represented in Figure 2 for better understanding.

The Everything-out CV is the complete structural exclusiveness between the training and test sets which can be achieved by placing completely different mixture components in the test set from the training set. In other words, all the components of the mixtures of the test set should be structurally different from the training set mixtures. Due to the scarcity of toxicity data, everything-out CV could not be performed in this research; instead of this, a set of mixtures has been designed (containing unique mixture components from the original modelling set), and the quality of prediction of the developed model has been judged using this designed set by a Prediction reliability indicator [38] (Sec. '*Prediction reliability indicator and designing of a hypothetical dataset*').

Chemical read-across

Read-across prediction is a similarity-based in silico approach where the toxicity or biological response of untested chemicals (target chemical) is predicted from the known toxicity or biological response values of chemical alanog/s (source chemical/s) [14]. Two types of read-across algorithms are mostly seen, such as the analogue approach (the toxicity/response of an untested chemical is predicted from a single known analog) and the category approach (the toxicity/response of an untested chemical is predicted from more than one known analogs). The category approach is more reliable in terms of quantitative prediction because the analogue approach can provide the toxicity value of itself as the prediction. Recently, the present authors' group has reported a machine learning-based chemical read-across tool- 'Read-across v4.1' which can predict the toxicity (biological activity in general) based on the structural similarity of the source and target chemicals [39,40]. This tool calculates the similarity (structural similarity or the similarity of descriptors) by three distinct approaches, namely, Euclidean distance-based approach, Gaussian kernel function similarity, and Laplacian kernel function similarity, and gives the prediction by computing the weighted average of the similar source compounds based on their similarity [39,40].

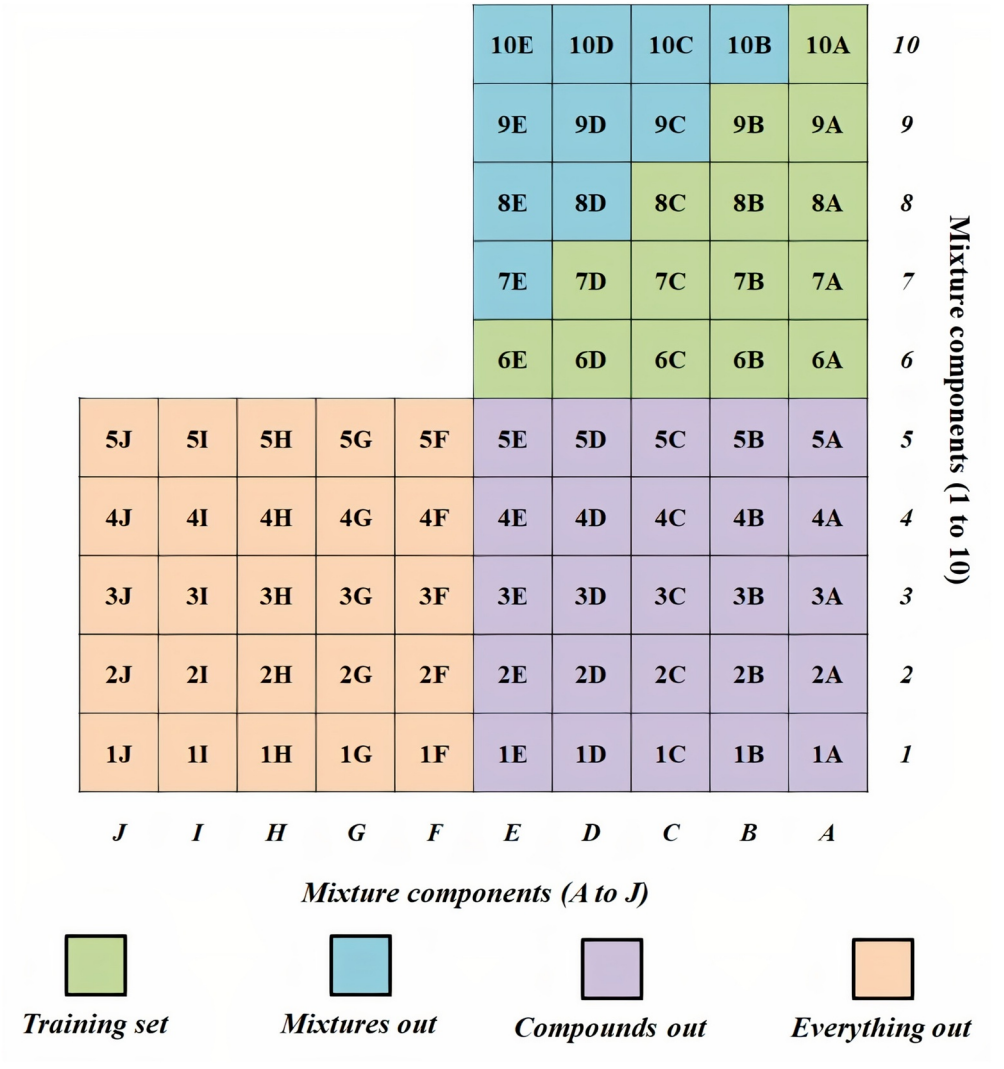


Figure 2. Division patterns of mixtures out, compounds out, and everything out cross-validations (the boxes represent the mixtures).

In this study, we have used the machine learning-based chemical read-across algorithm for the prediction of toxicity of the test mixtures with the modelled descriptors. For a successful prediction with Gaussian ad Laplacian kernel function similarity estimations, we have optimized the hyperparameter sigma (σ) and gamma (γ) respectively. For the optimization, the original training set (used for QSAR modelling) is randomly divided into subtrain and subtest sets in 3:1 proportion. The subtrain and subtest sets were then subjected to 'Read-across v4.1' with different σ and γ values (we have used 0.25 to 2 as σ and γ with 0.25 interval). The other determinants of the tool such as the number of close training compounds (10), distance threshold (1), and similarity threshold (0) were kept constant during optimization. The optimized setting has been chosen by checking the quality metrics Q^2_{F1} (calculated

Table 1. Designed query mixture dataset.

Mixture No.	*Component A*	*Component B*	*Mass ratio*
135	Chlorobenzene (#1)	2-bromo-4-chlorophenol (#11)	0.5:0.5
136	1,2,3-Trichlorobenzene (#2)	3,4-dibromoaniline (#12)	0.5:0.5
137	Bromobenzene(#3)	*p*-Chloroaniline (#13)	0.5:0.5
138	1-Bromo-2,3-dichlorobenzene (#4)	2,4-Dibromo-1-chlorobenzene (#9)	0.5:0.5
139	Phenol (#5)	2,4-Dibromo-1-chlorobenzene (#9)	0.5:0.5
140	2,4-Dichlorophenol (#6)	Fluorobenzene (#10)	0.5:0.5
141	Aniline (#7)	*p*-Chloroaniline	0.5:0.5
142	3,4-Dichloroaniline (#8)	Fluorobenzene (#10)	0.5:0.5
143	2,4-Dibromo-1-chlorobenzene (#9)	2-bromo-4-chlorophenol (#11)	0.5:0.5
144	2,4-Dibromo-1-chlorobenzene (#9)	3,4-dibromoaniline (#12)	0.5:0.5
145	2,4-Dibromo-1-chlorobenzene (#9)	*p*-Chloroaniline (#13)	0.5:0.5
146	Fluorobenzene (#10)	2-bromo-4-chlorophenol (#11)	0.5:0.5
147	Fluorobenzene (#10)	3,4-dibromoaniline (#12)	0.5:0.5
148	Fluorobenzene (#10)	*p*-Chloroaniline (#13)	0.5:0.5
149	2,4-Dibromo-1-chlorobenzene (#9)	Fluorobenzene (#10)	0.5:0.5
150	2-bromo-4-chlorophenol (#11)	*p*-Chloroaniline (#13)	0.5:0.5
151	2-bromo-4-chlorophenol (#11)	3,4-dibromoaniline (#12)	0.5:0.5
152	*p*-Chloroaniline (#13)	3,4-dibromoaniline (#12)	0.5:0.5

#: number of mixture components; #1 to #8: mixture components present in the studied dataset; #9 to #13: components of designed mixtures

by the tool itself with the predictions). The optimized setting was used finally with the original training and test set to get the final prediction. We have gradually decreased the number of close training compounds (10 to 2) to get the best predictions.

Prediction reliability indicator and designing of a hypothetical dataset

To develop a good QSAR model, proper validation is necessary, and to comply with the strict validation criteria, we performed external validation with the test set. Due to the scarcity of available mixture data, we could not perform the true external validation with an external dataset. Thus, we have designed a hypothetical mixture dataset (18 data points) to check the reliability of the prediction of the untested mixtures by the developed QSAR model. The designed dataset has been shown in the following Table 1. This dataset contains 18 hypothetical mixtures with equal mass (0.5:0.5) of two different components. Mixture No. 135 to 148 (8 combinations) are designed from a compound of the training set (#1 to #8) [11] and a completely new compound (#9 to #13) with respect to the original training set. On the contrary, mixture no. 143 to 152 (10 combinations) were designed from completely new compounds (#9 to #13) which were not present in the training set. The new compounds are the close congeners of the mixture components of this study and that is the reason behind the selection. The structures of the new pure compounds were collected from the PubChem chemical database maintained by National Centre for Biotechnology Information, USA (https://pubchem.ncbi.nlm.nih.gov/).

The mixture descriptors were calculated using the method discussed in the section '*Molecular Descriptors*'. The reliability of the prediction of the query dataset was adjudged by the statistical tool called 'Prediction Reliability Indicator' [38] which is available from https://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. This tool categorizes the reliability of prediction of the query chemicals of the designed dataset into 'Good', 'Moderate', and 'Bad', depending upon the composite score. This tool

essentially uses three sets of principles for the categorization of the quality of predictions and calculate the composite score; these are- (1) mean absolute error of prediction (leave-one-out) of a query compound from 10 most close training compounds; (2) applicability domain in terms of similarity, by using the standardization approach; (3) closeness of predicted values to the mean of the observed response of training set compounds.

Result and discussion

The developed model and its validation

A PLS model (with three latent variables) was developed from 63 data points (training set), reported in the materials and methods section. The model essentially reports four important structural features, responsible for the mixture toxicity. The model has been represented in the following equation (3):

$$\log 1/EC_{50} = 2.34406 - 0.16478 * N\% + 0.22706 * T(N..Cl) + 0.98204 * B03[O - Cl] + 0.1175 * MLOGP2 \quad (3)$$

The predicted values along with the descriptors of the model have been presented in Table S4 in the Supplementary Material. The reported model was validated rigorously following the OECD Principle 4. We have computed both the statistical validation metrics and reported them in the following Table 2. All the computed metric values are in the acceptable range (as per the OECD protocol), and it suggests the statistical credibility of the model. Statistical quality metrics like the r^2 (0.82) and r^2_{adj} (0.81) values indicate the goodness-of-fit, whereas the Q^2_{LOO} (0.78) value indicates the robustness of the developed model. The good predictive ability of the model can be justified by the computed external validation metric values such as Q^2_{F1} (0.87). The model shows low prediction errors that can be justified by the small values of $MAE_{Train;\ LOO}$ (0.24) and $RMSE_P$ (0.26).

Table 2. Computed values of statistical validation metrics.

Internal validation metrics	*Computed value*
Data points in the training set	63
r^2_{Train}	0.82
r^2_{adj}	0.81
Q^2_{LOO}	0.78
Scaled Average $r^2_{m\ (Train;\ LOO)}$	0.70
Scaled $\Delta r^2_{m\ (Train;\ LOO)}$	0.14
MAE (TRAIN; LOO)	0.24
External validation metrics	*Computed values*
Data points in the test set	21
Q^2_{F1}	0.87
Q^2_{F2}	0.87
Scaled average $r^2_{m\ (Test)}$	0.83
Scaled $\Delta r^2_{m\ (Test)}$	0.005
MAE (Test)	0.22
$RMSE_P$	0.26

LOO: Leave-one-out; MAE: Mean absolute error; $RMSE_P$: Root mean square error of prediction.

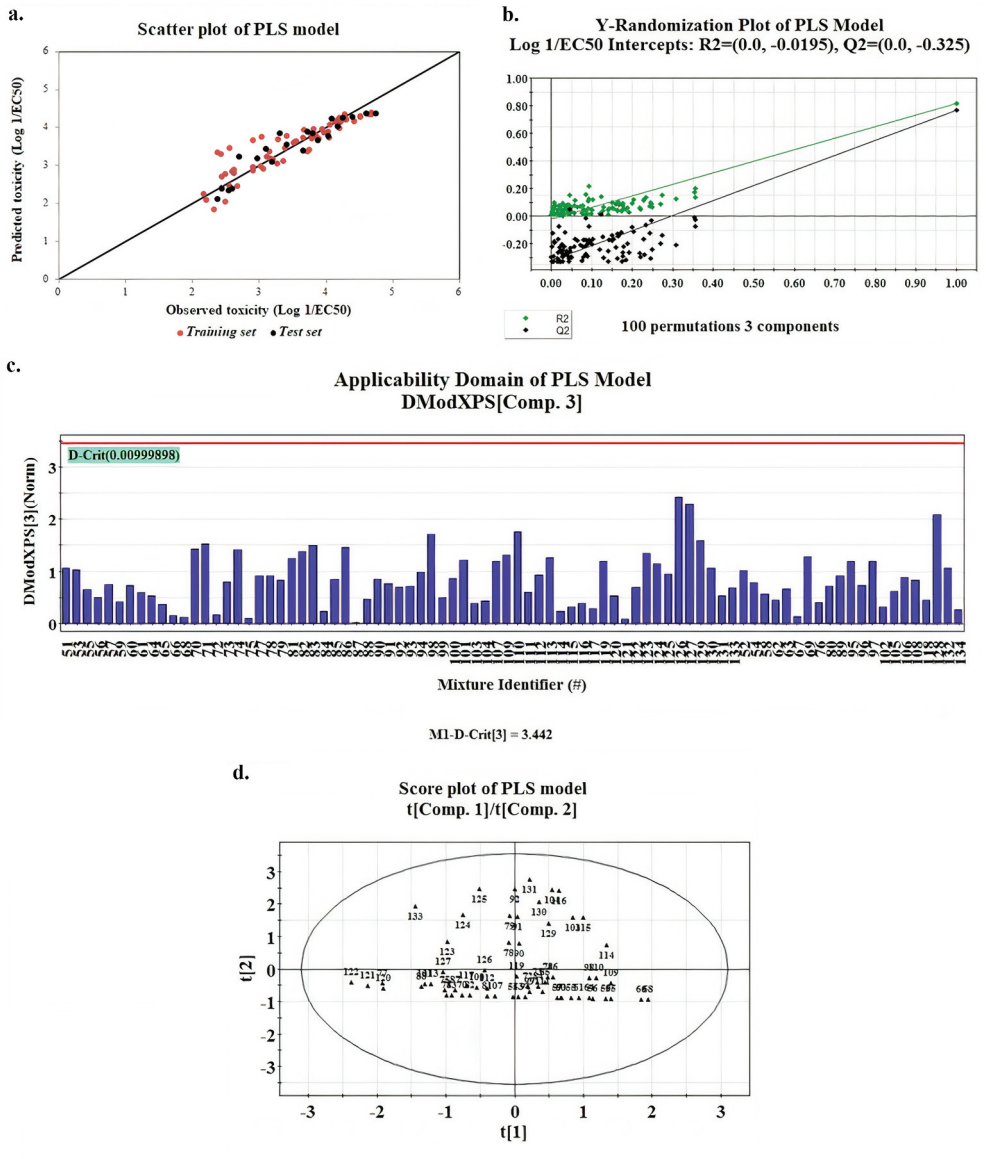


Figure 3. Various plots of developed PLS model; (A) Scatter plot (observed vs. predicted toxicity); (B) Y-randomization plot; (C) Applicability domain column plot (D) Score plot of PLS model.

Various plots have been incorporated in this study to graphically represent various aspects of the developed model. In Figure 3, the scatter plot (A), the Y-randomization plot (B), the applicability domain column plot (C), and the score plot (D) have been reported. An observed vs. predicted response scatter plot indicates the overall correlation between the observed and predicted response values. A good correlation can be identified by the scattering of plotted observed vs. predicted data at a particular region of the chemical space. It has been observed in Figure 3A that the plotted observed vs. predicted response data are scattered (for both the training and test sets) within a limited region along the

axis passing through the origin. It justifies a good correlation between the observed and predicted response and hence the goodness-of-fit of the model. We have performed the Y-randomization test, computed the resulting r^2 and Q^2, and shown the corresponding plot in Figure 3B. The poor r^2 (- 0.0195) and Q^2 (- 0.325) intercepts suggest that the developed model is not the result of any chance correlation. Finally, we have also identified the applicability domain of the developed model and represented in Figure 3C, to comply with the OECD principle 3. It has been observed from the applicability domain column plot that all the test mixtures are inside the applicability domain of the training mixtures of the developed model. Therefore, the predicted responses of the test set compounds are reliable enough to decide the predictive ability of the developed model. We have also reported a score plot [31] of the PLS model (Figure 3D) where the training data points are projected in 1st 2LV space of the developed model. It has been observed that the data points are scattered throughout the LV space and they are not clustered as such. This observation explains the structural diversity of the data points.

Result of cross-validation and prediction reliability indicator

We have performed two rigorous cross-validations in this study as discussed in the section '*Cross-validation*'. To obtain the cross-validated metrics, the new divisions (which are shown in Tables S2 and S3 in the Supplementary Material) have been used. We have computed the r^2, r^2_{adj}, and Q^2_{LOO}, and Q^2_{F1} as the quality measures. All cross-validated metric values have been tabulated in Table 3 along with the validation parameters of the original PLS model (Equation 3). It has been found that the values of cross-validated metrics are very similar to the original PLS model. Although the Q^2_{F1} of compounds out cross-validation is shown to be inferior (0.73) to the original PLS model but still in the acceptable range as per the standard protocol. Therefore, the statistical credibility, robustness, and predictive ability of the developed model have been confirmed again by these results of cross-validations.

We did not get enough mixture toxicity data for this particular endpoint for true external validation. Therefore, we have used the PRI tool to check the reliability of predictions of the developed PLS model with the newly designed mixture dataset (Sec. '*Prediction reliability indicator and designing of a hypothetical dataset*'). It has been found that all the designed mixtures can be predicted well with the developed model. As far as the composite score is concerned, all the predictions have shown 'Good' reliability. Apart from that, the new mixtures were inside the applicability domain of the developed model; therefore, the predicted results obtained by the PRI tool are reliable.

Table 3. Cross-validated metric values.

Cross-validated metrics	*Mixtures out cross-validation*	*Compounds out cross-validation*	*Developed PLS model*
n_{Train}	60	63	63
r^2	0.82	0.85	0.82
r^2_{adj}	0.81	0.84	0.81
Q^2_{LOO}	0.78	0.83	0.78
n_{Test}	24	21	21
Q^2_{F1}	0.87	0.73	0.87

LOO: Leave-one-out; adj: adjusted.

Mechanistic interpretation of modelled descriptors

OECD suggested five main guidelines to the QSAR modelers for developing unbiased, properly validated and mechanistically interpretable models. OECD Principle 5 suggests interpreting the modelled descriptors as far as possible and identifying the effects of structural features towards the response being modelled (molecular properties in general). Therefore, to comply with the 5th guideline, we have identified the modelled descriptors in the dataset and mechanistically shown how the structural features affect the toxicity. Four descriptors were identified during modelling and they are MLOGP2, T(N.Cl), N%, and B03[O-Cl]. It has been seen from the coefficient plot (Figure 4A) that among the modelled descriptors- MLOGP2, T(N.Cl), and B03[O-Cl] are positively contributing, whereas only N% is negatively contributing to the toxicity. We have also got the knowledge about the relative importance of the descriptors from the variable importance plot (Figure 4B) and the relative order of importance is MLOGP2 > T(N.Cl) > N% > B03[O-Cl].

MLOGP2 is a molecular property descriptor that is defined as the value of squared Moriguchi octanol-water partition coefficient of the chemical compounds (chemical mixture in this study), and it essentially suggests the lipophilicity of the mixtures. The value of this descriptor increases with the increment of the lipophilic property of test substances. It has been identified as a positively correlated and the most important descriptor of the developed model (Figure 5) and hence affects the toxicity to a large extent. In Figure 5, it has been observed that the MLOGP2 and log 1/EC_{50} (toxicity) values decrease simultaneously from mixture number 96 to 98 which justifies the positive correlation of the descriptor with the toxicity. Mixture 96 contains a higher percentage of more lipophilic component 1,2,3-trichlorobenzene (#2) and on the contrary, mixture 98 contains a higher percentage of more hydrophilic component 2,4-dichlorophenol (#6). Therefore, it can be concluded from these observations that the lipophilicity of the mixture is an important feature that has a proportional relationship with the toxicity and the presence of more lipophilic components can enhance the aquatic toxicity of mixtures.

The next important descriptor in this model is T(N.Cl) (2D atom pairs descriptor), which signifies the collective topological distance between nitrogen and chlorine. Among various mixture components of this study, only 3,4-dichloroaniline (#8) contains this feature and has been depicted in Figure 6. Nitrogen and chlorine are H-bond donor and H-bond acceptor groups respectively; hence, T(N.Cl) represents the hydrogen bonding propensity of molecules or mixture components. The value of this descriptor increases due to the presence of more than one nitrogen or chlorine atom in the structure (as seen in 3,4-dichloroaniline) and

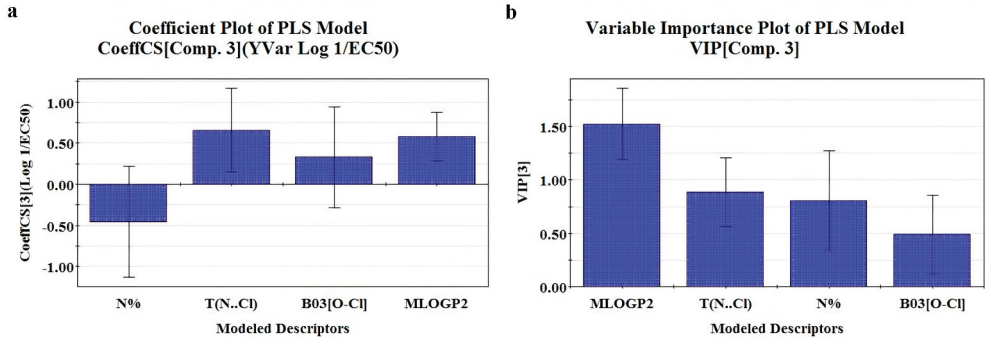


Figure 4. Coefficient plot and variable importance plot of the developed PLS model.

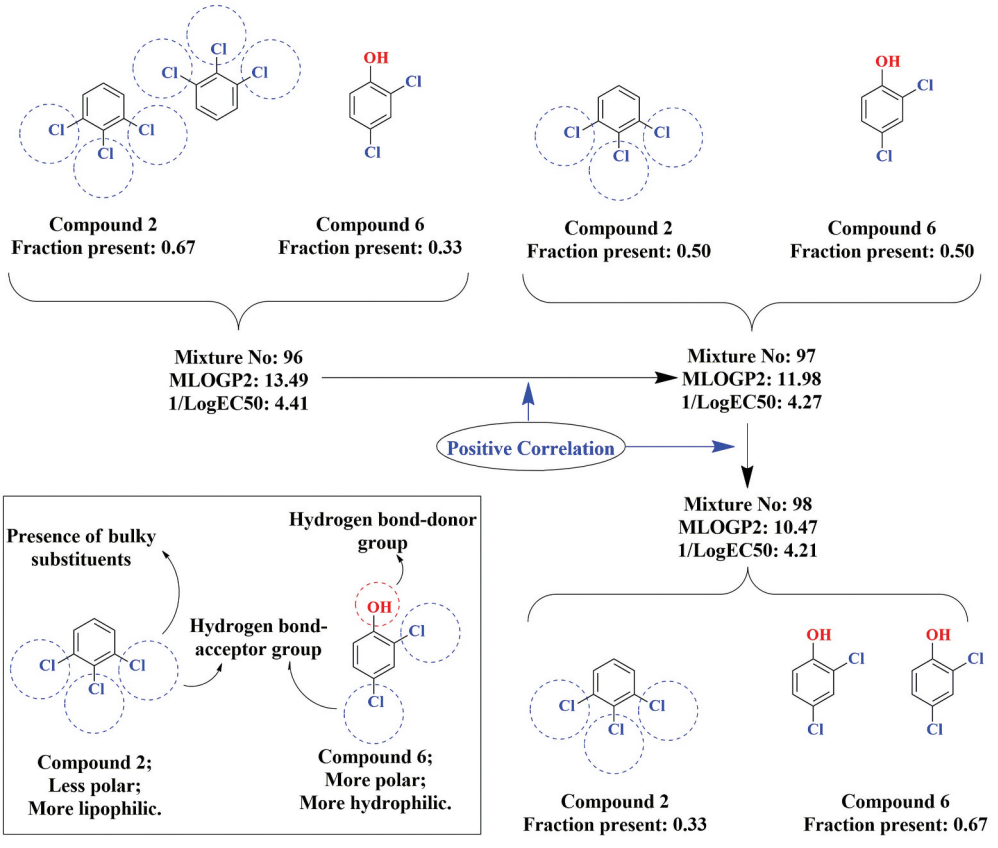


Figure 5. Demonstration of the effect of descriptor MLOGP2 towards aquatic toxicity of the mixtures.

this may increase the hydrogen bonding propensity and hence the solubility in the aquatic medium. It has been observed that the mixtures containing the component 3,4-dichloroaniline (#8) only have non-zero values of T(N.Cl) (e.g. mixture 78, 79, 80). The simultaneous increase of the observed toxicity of mixtures with the increment of the descriptor value justifies the positive contributing effect of the descriptor. Therefore, it can be said that the presence of components devoid of the feature T(N.Cl) or with its small value can help to reduce the aquatic toxicity of mixtures by reducing the solubility in the aquatic environment.

N% is a constitutional descriptor, and it contributes negatively to the aquatic toxicity of mixtures. The descriptor is defined as the percentage of nitrogen atoms present in the molecules or mixtures components. Nitrogen is a hydrogen bond donor group, and it can enhance the hydrophilic nature of a molecule by forming H-bonds with water. Hydrophilic compounds are dissolved in aquatic medium but cannot cross the hydrophobic plasma membrane of the organisms easily and hence show less toxicity. In Figure 7, we have depicted some mixture components like chlorobenzene (#1), 2,4-dichlorophenol (#6), etc. which are devoid of any nitrogen atom, whereas a component like aniline (#7) contains nitrogen. Mixtures containing nitrogen-enriched components (e.g. mixture 76, 77) have shown less toxicity than the mixture without nitrogen-enriched components (e.g. mixture 73). This observation clearly shows the inverse relationship of the mentioned descriptor with toxicity.

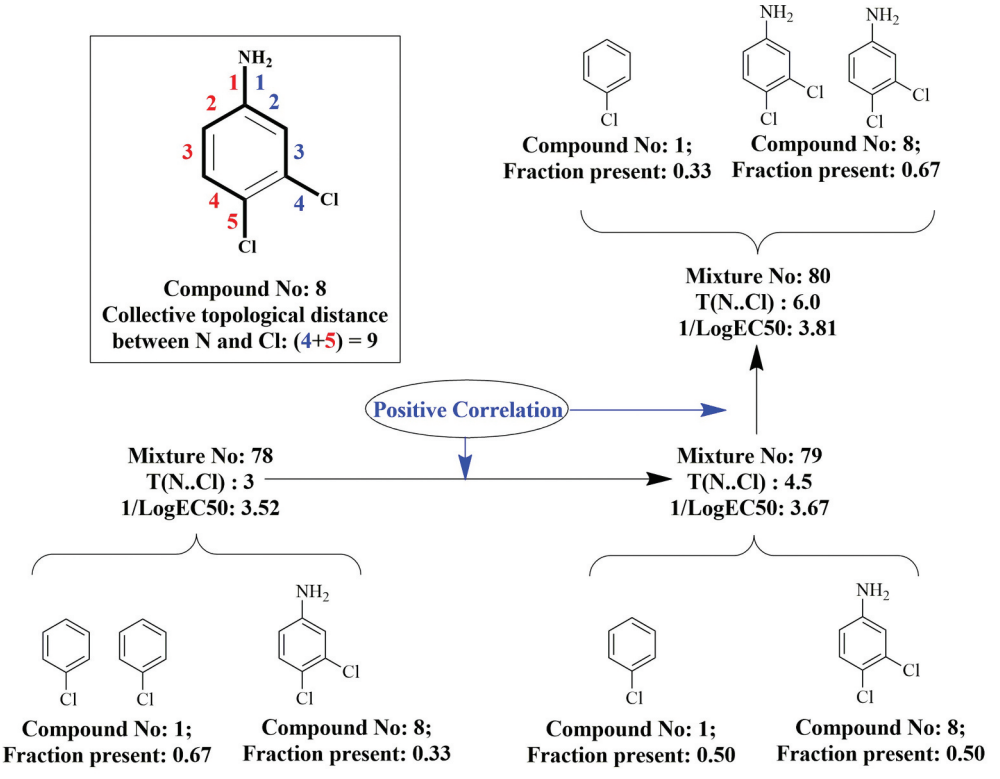


Figure 6. Demonstration of the effect of descriptor T(N.Cl) towards aquatic toxicity of the mixtures.

The least important descriptor of the developed model is B03[O-Cl] which is a 2D atom pairs descriptor. It is a binary descriptor that indicates the presence (descriptor value 1) or absence (descriptor value 0) of O-Cl atom pairs at topological distance 3. As per the coefficient plot of the PLS model (Figure 3A), B03[O-Cl] is a positively contributing descriptor. This feature can be seen in the mixture component 2,4-Dichlorophenol (#6) among the studied dataset (Figure 8). We have depicted the individual components of mixture 106 and 109 in Figure 8. Oxygen and chlorine can form hydrogen bonds and can be dissolved easily in the aquatic medium. It has been observed that this feature is absent in the components of mixture 106 and that is why it shows mixture descriptor value 0; on the contrary, mixture 109 contains the feature and shows higher toxicity than mixture 106. This observation indicates the positive correlation of the descriptor with the studied response and shows the importance of solubility to show toxicity against aquatic organisms.

Result of chemical read-across

The result of the chemical read-across has been presented in two steps in this study. The optimization of hyperparameters σ and γ for Gaussian kernel and Laplacian kernel similarity estimations, respectively, has been performed as the first step of machine learning as discussed

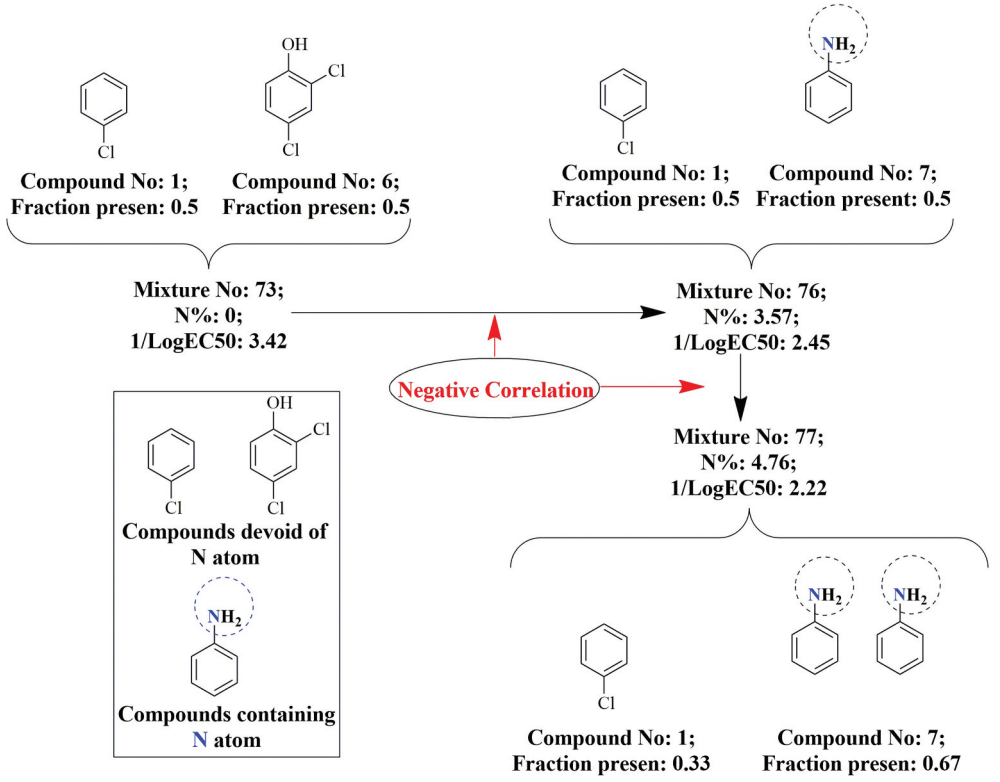


Figure 7. Demonstration of the effect of descriptor N% towards aquatic toxicity of the mixtures.

in the section 'Chemical read-across'. The result of the optimization has been presented in Table S5 in the Supplementary Material. From that table, the optimized values of σ and γ have been obtained and those are 0.5 and 2 respectively.

The optimized setting has been used in the next step to get the predictions of the actual test set. We have applied the number of close training compounds as a determinant of prediction; thus, nine predictions have been obtained with 10 to 2 close training compounds. The best prediction has been obtained with three close training compounds by Gaussian kernel function similarity and is given in Table S4 in the Supplementary Material. The calculated values of Q^2_{F1} and Q^2_{F2} were 0.94 and 0.94, respectively, which suggest good predictivity. The error measures, such as RMSEp and MAE, have also been computed and the values were 0.18 and 0.15, respectively, which are on the lower side. We have given the predicted values of each test mixture along with the standard error of predictions in Table S4 in the Supplementary Material. Based on the results of chemical read-across predictions, an improvement in the prediction quality of binary mixture has been clearly observed in this study.

Comparison with a previously published model

Lin et al. reported an MLR model in 2003 on the mentioned dataset [16]. The model is based on the experimentally derived octanol-water partition coefficient-based hydrophobicity descriptors. Experimental estimation of octanol-water partition coefficient needs

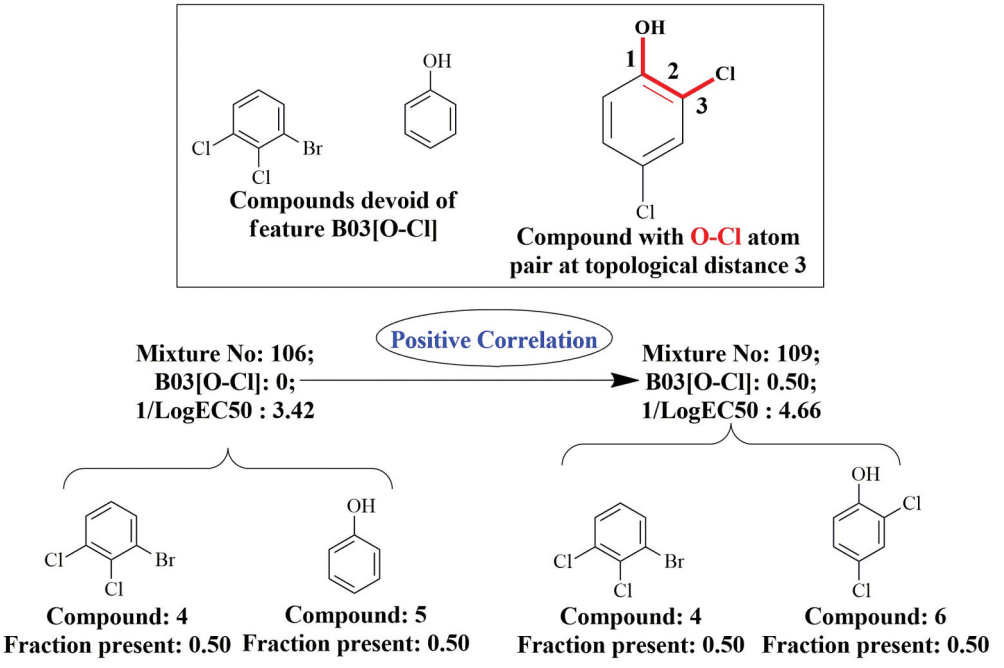


Figure 8. Demonstration of the effect of descriptor B03[O-Cl] towards aquatic toxicity of the mixtures.

more resources, time, and manpower. To predict the toxicity of untested mixtures, a laboratory setup is also needed. Although the model quality was good in terms of the r^2 (0.948) and standard error (0.166), the model was not externally validated. Therefore, the predictivity of the mentioned model for new untested mixtures has not been sufficiently established. In the present work, our objective was to use solely computational descriptors that can be derived in no time given the structure of query compounds. The new PLS model has been developed and validated following the OECD protocol. Various internal and external validation metrics have been computed and reported in the present study (Sec. '*The developed model and its validation*'). Acceptable r^2 (0.82) and Q^2_{LOO} (0.8) values prove the robustness of the model, whereas the acceptable Q^2_{F1} (0.87) value suggests its predictive ability. The predictive ability of the binary mixtures has been further improved (Q^2_{F1} = 0.94, MAE = 0.15) by the use of chemical read-across with the modelled descriptors. To avoid the complexity associated with the experimentally derived descriptors, only 2D structural descriptors have been utilized in this study. So to check the toxicity of untested mixtures, the prior formulation is not needed.

Conclusion

Prediction of aquatic toxicity of 84 binary mixtures has been accomplished by the PLS regression method. Various 2D descriptors have been successfully used to identify the important structural features of mixture components and model the overall mixture toxicity. The weighted descriptor generation approach has also been utilized for the calculation of mixture descriptors. The Y-randomization test has also been successfully

executed in this study which confirms the model as not a result of chance correlation. It has been found out that the lipophilicity, hydrogen bonding propensity, and solubility of the mixture components are the important features that affect the aquatic toxicity of mixtures of polar and non-polar narcotic toxicants. Overall lipophilicity of mixture components enhances the toxicity of the mixtures; at the same time, the solubility of mixture components is important for the dissolution into the aquatic medium that indirectly promotes the toxicity. The presence of nitrogen atoms in the mixture component can reduce the toxicity, whereas the simultaneous occurrence of nitrogen and chlorine can enhance the toxicity. Various newly devised cross-validation strategies have also been employed to avoid the developmental bias and overestimation of the predictive ability of the model in this study. We should mention that the true external validation and everything-out cross-validation could not be performed in this study due to the non-availability of toxicity data against the particular endpoint; instead, the reliability of prediction of the newly designed theoretical mixtures has been adjudged by the PRI tool and proved to be reliable. Finally, the prediction quality of the binary mixtures has been successfully improved by using a machine learning-based chemical read-across algorithm with the identified structural features. In a nutshell, the use of similarity-based QSAR with different cross-validation strategies and the machine learning-based chemical read-across algorithm as alternative approaches for mixture risk assessment has been successfully established in this study.

Acknowledgements

MC sincerely acknowledges to All India Council for Technical Education (AICTE, New Delhi) for the financial assistance in the form of the National Doctoral Fellowship (NDF). Financial assistance for SERB, New Delhi under the MATRICS scheme is thankfully acknowledged.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the All India Council for Technical Education.

References

[1] Aquatic toxicity, (2016). https://www.chemsafetypro.com/Topics/CRA/ecotox_aquatic_toxicity.html (accessed November 25, 2021).

[2] R. Schwarzenbach, B. Escher, K. Fenner, T. Hofstetter, A. Johnson, U. Gunten, and B. Wehrli, *The challenge of micropollutants in aquatic systems*, Science 313 (2006), pp. 1072–1077. doi:10.1126/science.1127291.

[3] Y. Yang, Z. Chen, J. Zhang, S. Wu, L. Yang, L. Chen, and Y. Shao, *The challenge of micropollutants in surface water of the Yangtze river*, Sci. Total Environ. 780 (2021), pp. 146537. doi:10.1016/J.SCITOTENV.2021.146537.

[4] R. Yang, R. Thomas, D. Gustafson, J. Campain, S. Benjamin, H. Verhaar, and M. Mumtaz, *Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling*, Environ. Health Perspect. 106 (1998), pp. 1385–1393. doi:10.1289/ehp.98106s61385.

[5] R.M. Evans, O.V. Martin, M. Faust, and A. Kortenkamp, *Should the scope of human mixture risk assessment span legislative/regulatory silos for chemicals?*, Sci. Total Environ. 543 (2016), pp. 757–764. doi:10.1016/j.scitotenv.2015.10.162.

[6] S. Kar and J. Leszczynski, *Exploration of computational approaches to predict the toxicity of chemical mixtures*, Toxics 7 (2019), pp. 15. doi:10.3390/toxics7010015.

[7] D.W. Kolpin, E.T. Furlong, M.T. Meyer, E.M. Thurman, S.D. Zaugg, L.B. Barber, and H.T. Buxton, *Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999 −2000: A national reconnaissance*, Environ. Sci. Technol. 36 (2002), pp. 1202–1211. doi:10.1021/ES011055J.

[8] A. Kienzler, E. Berggren, J. Bessems, S. Bopp, S. van der Linden, and A. Worth, *Assessment of Mixtures - Review of Regulatory Requirements and Guidance*, European Commission, Brussels. 2014. doi:10.2788/84264.

[9] S.K. Bopp, A. Kienzler, A.N. Richarz, S.C. van der Linden, A. Paini, N. Parissis, and A.P. Worth, *Regulatory assessment and risk management of chemical mixtures: Challenges and ways forward*, Crit. Rev. Toxicol. 49 (2019), pp. 174–189. doi:10.1080/10408444.2019.1579169.

[10] J. Kim, S. Kim, and G.E. Schaumann, *Reliable predictive computational toxicology methods for mixture toxicity: toward the development of innovative integrated models for environmental risk assessment*, Rev. Environ. Sci. Biotechnol. 12 (2013), pp. 235–256. doi:10.1007/S11157-012-9286-7.

[11] *New approach methodologies in regulatory science: proceedings of a scientific workshop: Helsinki, 19-20 April 2016*, European Chemical Agency, Helsinki, Finland. 2016. https://echa.europa.eu/documents/10162/21838212/scientific_ws_proceedings_en.pdf/a2087434-0407-4705-9057-95d9c2c2cc57. (accessed November 25, 2021).

[12] P. Nymark, M. Bakker, S. Dekkers, R. Franken, W. Fransman, A. García-Bilbao, D. Greco, M. Gulumian, N. Hadrup, S. Halappanavar, V. Hongisto, K.S. Hougaard, K.A. Jensen, P. Kohonen, A.J. Koivisto, M. Dal Maso, T. Oosterwijk, M. Poikkimäki, I. Rodriguez-Llopis, R. Stierum, J.B. Sørli, and R. Grafström, *Toward rigorous materials production: New approach methodologies have extensive potential to improve current safety assessment practices*, Small 16 (2020), pp. 1904749. doi:10.1002/SMLL.201904749.

[13] J.C. Madden, S.J. Enoch, A. Paini, and M.T.D. Cronin, *A review of in silico tools as alternatives to animal testing: principles, resources and applications*, Altern. Lab. Anim. 48 (2020), pp. 146–172. doi:10.1177/0261192920965977.

[14] A.B. Raies and V.B. Bajic, *In silico toxicology: Computational methods for the prediction of chemical toxicity*, WIREs Comput. Mol. Sci. 6 (2016), pp. 147–172. doi:10.1002/wcms.1240.

[15] E.N. Muratov, A. Tropsha, E.V. Varlamova, V.E. Kuzmin, A.G. Artemenko, N.N. Muratov, S. Mileyko, and D. Fourches, *Everything out validation approach for QSAR models of chemical mixtures*, JSM Clin. Pharm. 1 (2014), pp. 1005.

[16] Z. Lin, L. Wang, and H. Yu, *Quantification of joint effect for hydrogen bond and development of QSARs for predicting mixture toxicity*, Chemosphere 52 (2003), pp. 1199–1208. doi:10.1016/S0045-6535(03)00329-1.

[17] D. De Zwart and W. Slooff, *The Microtox as an alternative assay in the acute toxicity assessment of water pollutants*, Aquat. Toxicol. 4 (1983), pp. 129–138. doi:10.1016/0166-445X(83)90050-4.

[18] J.M. Ribo and K.L.E. Kaiser, Photobacterium phosphoreum *toxicity bioassay. I. test procedures and applications*, Environ. Toxicol. Water Qual. 2 (1987), pp. 305–323. doi:10.1002/tox.2540020307.

[19] I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polishchuk, V. Kuz'Min, and A. Varnek, *QSPR approach to predict nonadditive properties of mixtures. application to bubble point temperatures of binary mixtures of liquids*, Mol. Inform. 31 (2012), pp. 491–502. doi:10.1002/minf.201200006.

[20] G.D. Veith and S.J. Broderius, *Rules for distinguishing toxicants that cause type I and type II narcosis syndromes*, Environ. Health Perspect. 87 (1990), pp. 207–211. doi:10.1289/ehp.9087207.
[21] D. Finney, *Probit Analysis*, Cambridge University Press, London, UK, 1971.
[22] K. Roy, S. Kar, and R.N. Das, *Chemical information and descriptors*, in *Understanding the Basics of QSAR for Application in Pharmaceutical Science and Risk Assessment*, K. Roy, S. Kar, and R.N. Das, eds., Academic Press, Elsevier, London, 2015, pp. 47–80.
[23] P. Gramatica, *Principles of QSAR modeling*, Int. J. Quant. Struct. Relat. 5 (2020), pp. 61–97. doi:10.4018/ijqspr.20200701.oa1.
[24] G. Hoover, S. Kar, S. Guffey, J. Leszczynski, and M.S. Sepúlveda, *In vitro and in silico modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line*, Chemosphere 233 (2019), pp. 25–33. doi:10.1016/J.CHEMOSPHERE.2019.05.065.
[25] M. Chatterjee and K. Roy, *Prediction of aquatic toxicity of chemical mixtures by the QSAR approach using 2D structural descriptors*, J. Hazard. Mater. 408 (2021), pp. 124936. doi:10.1016/j.jhazmat.2020.124936.
[26] A. Mauri, *alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints*, in *Ecotoxicological QSARs*, K. Roy, ed., Humana Press Inc, New York, 2020, pp. 801–820. doi:10.1007/978-1-0716-0150-1_32.
[27] Marvin, ChemAxon, San Diego, USA, 2018; software available at https://chemaxon.com/products/marvin (accessed October 25, 2021).
[28] R. Paul, Personal communication: paulrahul1997@gmail.com, (2021).
[29] R. Leardi, *Genetic algorithms in chemometrics and chemistry: A review*, J. Chemom. 15 (2001), pp. 559–569. doi:10.1002/cem.651.
[30] K. Roy, S. Kar, and R.N. Das, *Selected statistical methods in QSAR*, in *Understanding the Basics of QSAR for Application in Pharmaceutical Science and Risk Assessment*, K. Roy, S. Kar, R.N. Das, eds., Academic Press, Elsevier, London, 2015, pp. 191–229.
[31] S. Wold, M. Sjöström, and L. Eriksson, *PLS-regression: A basic tool of chemometrics*, Chemom. Intell. Lab. Syst. 58 (2001), pp. 109–130. doi:10.1016/S0169-7439(01)00155-1.
[32] K. Roy, S. Kar, and R.N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, Elsevier, London, 2015.
[33] N. Chirico and P. Gramatica, *Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection*, J. Chem. Inf. Model. 52 (2012), pp. 2044–2058. doi:10.1021/CI300084J.
[34] N. Akarachantachote, S. Chadcham, and K. Saithanu, *Cutoff threshold of variable importance in projection for variable selection*, Int. J. Pure Appl. Math. 94 (2014), pp. 307–322. doi:10.12732/IJPAM.V94I3.2.
[35] J.G. Topliss and R.P. Edwards, *Chance factors in studies of quantitative structure-activity relationships*, J. Med. Chem. 22 (1979), pp. 1238–1244. doi:10.1021/JM00196A017.
[36] D. Gadaleta, G.F. Mangiatordi, M. Catto, A. Carotti, and O. Nicolotti, *Applicability domain for QSAR models: Where theory meets reality*, Int. J. Quant. Struct. Relat. 1 (2016), pp. 45–63. doi:10.4018/IJQSPR.2016010102.
[37] Validation of (Q)SAR Models - OECD, (2004). https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed April 20, 2022).
[38] K. Roy, P. Ambure, and S. Kar. *How precise are our quantitative structure–activity relationship derived predictions for new query chemicals?,* ACS Omega 3 (2018), pp. 11392-11406. doi:10.1021/acsomega.8b01647.
[39] M. Chatterjee, A. Banerjee, P. De, A. Gajewicz-Skretna, and K. Roy, *A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data*, Environ. Sci. Nano 9 (2022), pp. 189–203. doi:10.1039/d1en00725d.
[40] A. Banerjee and K. Roy, *First report of q-RASAR modeling towards an approach of easy interpretability and efficient transferability*, ChemRxiv Cambridge Open Engag. (2022), doi:10.26434/chemrxiv-2022-0qclt.

Chemosphere 308 (2022) 136463

Contents lists available at ScienceDirect

Chemosphere

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

Chemical similarity and machine learning-based approaches for the prediction of aquatic toxicity of binary and multicomponent pharmaceutical and pesticide mixtures against *Aliivibrio fischeri*

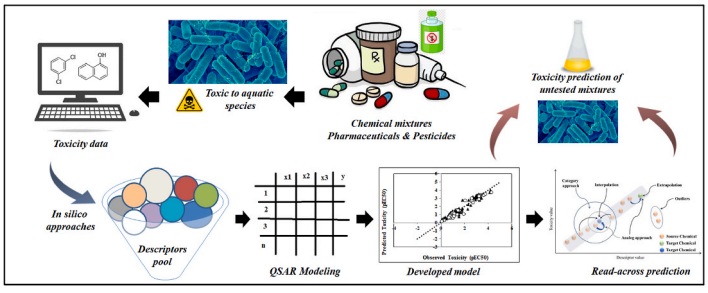
Mainak Chatterjee, Kunal Roy *

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

HIGHLIGHTS

- We apply QSAR and chemical read-across for the toxicity prediction of binary and multi-component mixtures against *Aliivibrio fischeri.*
- We apply mixtures out, compounds out, and everything out cross-validation.
- We have used simple 2D structural descriptors for fast prediction, easy interpretability and transferability of the model.
- These methods are the easy, fast, cheap, and ethically acceptable alternatives to the mixture's risk assessment.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Nynke Kramer

Keywords:
Pharmaceutical mixtures
Pesticides
Mixture toxicity
QSAR
Read-across

ABSTRACT

Different classes of chemicals are present in the environment as mixtures. Among them, pharmaceuticals and pesticides are of major concern due to their improper use and disposal, and subsequent additive and non-additive effects. To assess the environmental risk posed by the mixtures of pharmaceuticals and pesticides, a quantitative structure-activity relationship (QSAR) model has been developed in this study using the pEC_{50} values of 198 binary and multi-component mixtures against the marine bacterium *Aliivibrio fischeri.* The developed partial least squares (PLS) model has been rigorously validated and proved to be a robust and extremely predictive one. To address the chances of overestimation of validation metrics, three cross-validation tests (mixtures out, compounds out, and everything out) have been applied, and the results were satisfactory. The use of simple 2-dimensional descriptors makes the prediction much quick, and also makes the model easily interpretable. A machine learning-based chemical read-across prediction has also been performed to justify the effectiveness of selected structural features in this study. In a nutshell, this study proves QSAR and chemical read-across as effective alternative approaches for the toxicity prediction of pharmaceutical and pesticide mixtures and also approves the use of mixture descriptors for modelling mixtures successfully.

* Corresponding author.
E-mail addresses: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).

https://doi.org/10.1016/j.chemosphere.2022.136463
Received 23 June 2022; Received in revised form 10 September 2022; Accepted 12 September 2022
Available online 16 September 2022
0045-6535/© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

The continued population growth, rapid urbanization, and simultaneous chemical advancement greatly increase the chemical pollution worldwide (Carley and Christie, 2019). The chemical industries are flourishing in different parts of the world (mostly in the developing countries) to fulfill the growing demand of mankind (Landrigan et al., 2018), and these socio-economical and industrial growths are creating a serious threat not only to groundwater prevention but also to the environmental safety (Daughton, 2003). A majority of chemical pollutants starting from plant protection products, pharmaceuticals, and industrial chemicals to household chemicals have been found in water resources like groundwater layers and wastewater streams in ng/L to mg/L ranges (Lamastra et al., 2016; Luo et al., 2014). These chemical species are usually found in water bodies as multi-component chemical mixtures; however, mixture toxicity is one of the neglected domains which should gain sufficient attention for obvious reasons (Kolpin et al., 2002; Kortenkamp and Altenburger, 1999). In the past decades, contamination of the aquatic environment with various pharmaceuticals and pesticides has grabbed the attention of global scientific communities (Gao et al., 2012; Kemper, 2008; Topp et al., 2016) and it is becoming an important topic of discussion after finding their simultaneous presence in water bodies (Elliott et al., 2018; Kolpin et al., 2002). Scientists are now working on the risk assessment of chemical mixtures; however, mixture risk assessment by physical experimentation is not a good idea due to the unending possibilities of combinations. There are two largely accepted and widely used mixture toxicity models, namely concentration addition (CA) (Loewe and Muischnek, 1926) and independent action (IA) (Bliss, 1939) which assume the additive nature of mixture components, but they never consider the probable interactions (especially for the biologically active chemicals like pharmaceuticals and pesticides) among the components (Kim et al., 2013). Apart from this, the CA and IA models need preliminary experimentation, which limits the application of these modelling to some extent. To get rid of these above-mentioned issues, the scientists and regulatory agencies are working together; however, a universally acceptable toxicity protocol for mixture toxicity study is still rare (Iwasaki and Gauthier, 2016). Therefore, an efficient method for the identification and estimation of interactive effects (synergism and/or antagonism) along with the additive effect of mixture components is necessary for better risk assessment of chemicals in a real-world scenario.

The concept of similarity stands upon a relative analysis among different points with respect to others. In the context of ecotoxicology and drug design, it is necessary to have an objective measure of similarity to compare various chemicals concerning their toxicity or activity (Gallegos Saliner, 2006). Machine learning algorithms are now gaining sufficient attention among scientific communities for predictive modelling of toxicity and biological activity based on chemical similarity (Nazemzadeh et al., 2021). The hybrid modelling approaches can be very useful in this regard due to their ability to combine the detailed process understanding with machine learning; thus, they can provide an opportunity to integrate prior process knowledge with various measurement data for efficient modelling of toxicity of hazardous mixtures (Nazemzadeh et al., 2021). Quantitative structure-activity relationships (QSARs) and chemical read-across algorithms are the most widely used similarity-based computational approaches which can efficiently predict the toxicity of individual chemicals by relative similarity analysis with respect to the chemical structures and properties (Hao et al., 2020; Li et al., 2022); hence, the use of these *in silico* methods is widely accepted and approved by the concerning authorities like the Organization of Economic Co-operation and Development (OECD), European Chemical Agency (ECHA), United States Environmental Protection Agency (US-EPA), and regulations like Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) etc. (Kar et al., 2020; Klatte et al., 2017; Mansour et al., 2016; Raies and Bajic, 2016). Therefore, these methods, especially QSAR can be a good alternative for mixture toxicity assessment, which can give predictions based on various structural features of the mixture components. Various research groups have applied QSAR for the mixture toxicity assessment in the recent past and successfully proposed numerous mathematical models for different environmental endpoints (Carnesecchi et al., 2020; Chatterjee and Roy, 2021; Hoover et al., 2019; Qin et al., 2021), but the number of such studies is still very less in comparison with individual component toxicity predictions. QSAR can also help to identify the interactions among structural features which may be essential for the toxicological effect (biological properties in general). It should be noted here that the QSAR modelling of mixtures is not as simple as the individual toxicants; thus, various additional steps should be followed for successful modelling. Among the key steps, mixture descriptor calculation, and validation of the developed model are the most important. There are very few approaches in the literature, where the cross-validation strategies have been properly used for the mixture toxicity predictions, and hence the chance of overestimation of quality metrics is persisting. Therefore, to study the mixture toxicity of pharmaceuticals and pesticides by QSAR rigorously, more attempts should be made by following the appropriate procedures and without compromising the guidelines of QSAR model development for single chemicals given by OECD.

The objective of this study was to apply the in-silico approaches for the prediction of aquatic toxicity of binary and multicomponent mixtures effectively. Keeping the objective in mind, we developed a QSAR model for the prediction of aquatic toxicity of binary and multicomponent chemical mixtures in this communication. To build the model, a toxicity dataset of 198 hazardous mixtures against *Aliivibrio fischeri* was taken. The mixtures were composed of 12 pharmaceuticals and 8 pesticides which are abundantly used either for the treatment of various diseases, or the agricultural purposes and hence can be present in wastewater streams and can flow through the river and freshwater resources (Gao et al., 2012; Kemper, 2008; Topp et al., 2016). 2D structural descriptors have been used in this study for the ease of calculation and interpretation (Chatterjee and Roy, 2021; Mukherjee et al., 2022; Sun et al., 2021). Mixture descriptor calculation is one of the tricky steps of mixture-QSAR. Qin et al. reported various mixing rules for the computation of mixture descriptors in 2018 among which the theory of additivity was important one (Qin et al., 2018). In 2019, Hoover et al., and Kar et al. have separately mentioned about the theory of additivity as one of the promising approaches for mixture descriptor calculation (Hoover et al., 2019; Kar and Leszczynski, 2019). Thus, we have also applied the theory of additivity for obtaining the mixture descriptors in this research; the details of the computation strategy have been discussed in "*2.2. Mixture descriptor calculation*" section. Partial least squares (PLS) regression has been applied to develop the model, which has been cross-validated using mixtures out, compounds out, and everything out cross-validation strategies afterwards (Muratov et al., 2014; Oprisiu et al., 2012). The selected descriptors have been thoroughly studied in the light of this model. We have also reported the applicability domain (Gadaleta et al., 2016) of the developed model which is essential to claim the reliability of prediction with the new model. Besides all these, a machine learning-based chemical read-across prediction tool has been used to justify the relevance of the selected structural features in this study, and it is one of the instances where a chemical read-across algorithm has been applied for mixture toxicity prediction. The complete study has been performed without any animal experimentation and in the environmental friendly condition to comply with the 3R concept (Replacement, Reduction, and Refinement) of REACH legislation(Ritskes-Hoitinga and Van Luijk, 2019).

2. Materials and methods

2.1. Toxicity data

Good quality toxicity data (biological data) is one of the most important prerequisites of QSAR model development, and the quality of

such dataset is an important factor along with the number of data points for the performance of the model. Therefore, the dataset should be chosen and curated carefully for obtaining the optimum prediction from the developed model. In this research, the toxicity dataset of 198 hazardous mixtures (178 binary mixtures, 15 ternary mixtures, and 5 quaternary mixtures) has been utilized for model development (Table S3 in **Supplementary material 1**). We have taken the toxicity data of 153 binary mixtures and 45 multi-component mixtures separately against *Aliivibrio fischeri* from two published articles by Bures et al. (Sigurnjak Bureš et al., 2021) and Qin et al. (2018) respectively. The test organism (also called *Vibrio fischeri*) is a gram-negative, rod-shaped, bioluminescent marine bacterium that lives globally in symbiosis with various sea creatures such as Hawaiian bobtail squid, and it is a classic marker of aquatic toxicity (Abbas et al., 2018). The data reported by Bures et al. contains 51 binary combinations of 3 different molar concentration ratios (0.75:0.25; 0.5:0.5; 0.25:0.75) whereas, the data reported by Qin et al. contains 5 binary combinations, 3 ternary combinations, and 1 quaternary combination with 5 variable molar concentration ratios each. The studied mixtures were composed of 12 marketed pharmaceuticals and 8 marketed pesticides. All the mixture components along with their 2D structures have been displayed in Table S2 in **supplementary material 1**. The testing protocol and data generation have also been elaborated in **supplementary material 1**.

2.2. Mixture descriptor calculation

Todeschini and Consonni define molecular descriptor as "*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*" (Consonni and Todeschini, 2010). It has been observed to use the software-derived descriptors of individual chemicals directly for the classical QSARs (Ghanem et al., 2017; Mukherjee et al., 2022). But in the case of mixture's toxicity modelling, one has to consider the influence of at least 2 different chemicals (binary mixture) simultaneously during descriptor calculation. The theory of additivity is mostly employed during the calculation of mixture descriptors from the descriptors of component chemicals (the process of 2D descriptor calculation of mixture component is elaborated in **supplementary material 1**), and it has been successfully used in various instances (Chatterjee and Roy, 2021; Hoover et al., 2019; Wang et al., 2018). Therefore, we have also applied the theory of additivity to get the corresponding mixture descriptors (weighted descriptors) in this study. The descriptors of mixture components were obtained from the alvaDesc software (251 descriptors of the individual components were obtained after internal pretreatment of alvaDesc software) (Mauri, 2020). The mixture descriptors have been obtained by employing an in-house tool 'Mixture Descriptor Calculator v1.0' (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) from the 251 descriptors of each mixture components. If a mixture is composed of *m* chemicals (1 to m) with molar concentration ratios M_1, M_2, …., M_m ($M_1+M_2+ \ldots +M_m = 1$), and the descriptors of each chemical are D^1 (D_1^1, D_2^1, …., D_m^1), D^2 (D_1^2, D_2^2, …., D_m^2), D^n (D_1^n, D_2^n, …., D_m^n) respectively, then according to the theory of additivity, the corresponding weighted descriptors (D_{Mix}) can be calculated by the following equations (1)–(3):

$$D^1_{Mix} = \left(D^1_1 \times M_1\right) + \left(D^1_2 \times M_2\right) + \ldots + \left(D^1_m \times M_m\right) \quad (1)$$

$$D^2_{Mix} = \left(D^2_1 \times M_1\right) + \left(D^2_2 \times M_2\right) + \ldots + \left(D^2_n \times M_m\right) \quad (2)$$

$$D^n_{Mix} = \left(D^n_1 \times M_1\right) + \left(D^n_2 \times M_2\right) + \ldots + \left(D^n_m \times M_m\right) \quad (3)$$

After calculating the descriptors, a pretreatment operation was carried out by using the 'DataPreTreatmentGUI' tool (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) where the software removes the intercorrelated descriptors (intercorrelation cutoff 0.95), descriptors with the least variability among compounds (variability cutoff 0.0001), descriptors with constant values or near-constant values, and descriptors with at least one missing value. Finally, a pool of 247 mixture descriptors was obtained for modelling.

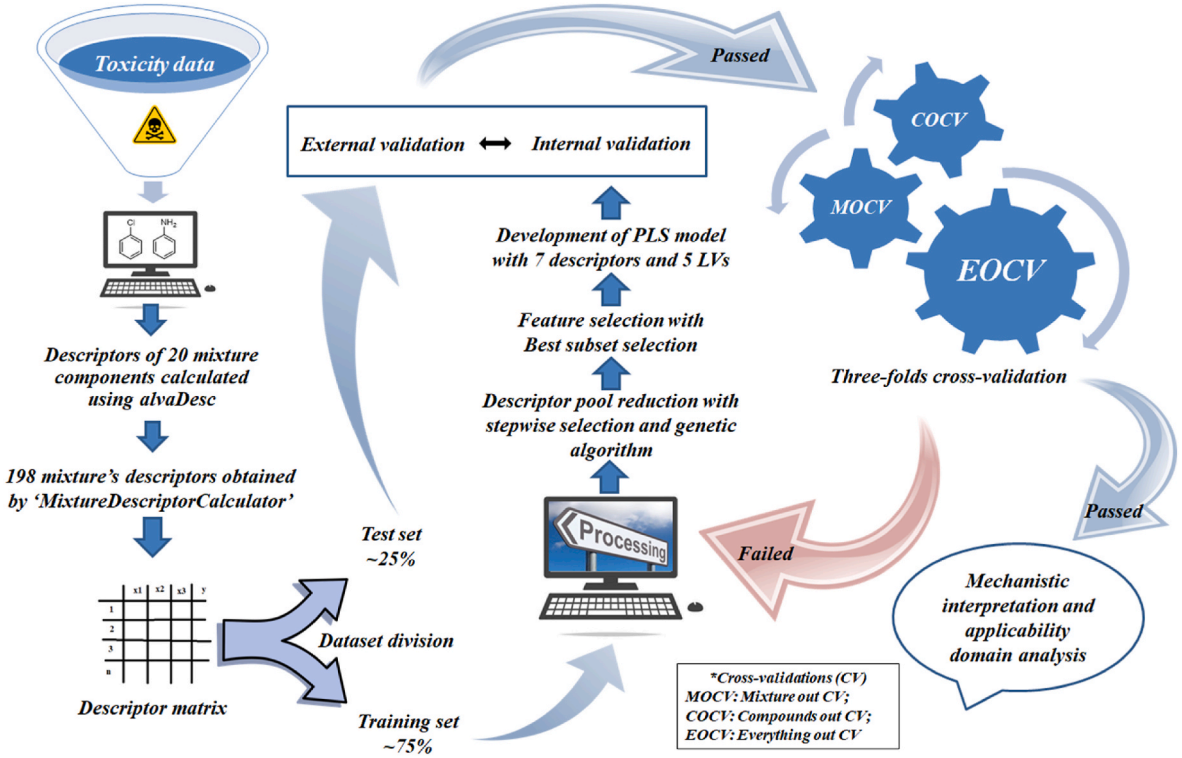


Fig. 1. Workflow of QSAR model development.

2.3. Dataset division

The dataset was divided into training and test sets for model development and external validation respectively, by applying Kennard-Stone selection, Euclidean distance-based division, and sorted activity-based division with the freely available tool 'Dataset Division GUI 1.2' (https://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools)(Roy et al., 2015). Preliminary regression models were generated using stepwise selection using the Minitab software (Minitab Statistical Software, 2022) and based on the internal cross-validated predicted variance (Q^2_{LOO}) and external predicted variance Q^2_{F1} (Q^2_{LOO} and Q^2_{F1} values of 9 preliminary model obtained from 9 divisions have been incorporated in Table S1 in **supplementary material 1**), only an Euclidean distance-based division (4th seed) was selected (149 training data, 49 test data with 247 descriptors) for further processing. The final 4-4 euclidean distance-based division has been displayed in Table S3 in **supplementary material 1**. The dataset division is further elaborated in **supplementary material 1**.

2.4. Feature selection and model development

In this communication, the data matrix of the final Euclidean distance-based division (4th seed) was taken forward to further steps for model development which consists of 247 descriptors (X variables) and the pEC50 values (Y variable). The pool of 247 descriptors was reduced to 40 by using both stepwise selection (Wagner and Shimshak, 2007) and genetic algorithm (elaborated in **supplementary material 1**) (Leardi, 1996; Rogers and Hopfinger, 1994). The stepwise selections have been carried out with the F-to-enter 4 and F-to-remove 3.9 setting in multiple iterations using Minitab software (Minitab Statistical Software, 2022). This F feature is the stepping criterion which is calculated by squaring the t values of the incoming descriptors. For a more detailed explanation of the stepwise selection strategy, one can refer to the literature (Wagner and Shimshak, 2007). The train-test pair with the reduced pool of 40 descriptors along with the dependent variable pEC50 were then subjected to the Best subset selection algorithm (Roy et al., 2015) (elaborated in **supplementary material 1**) for obtaining various descriptor combinations (subsets), and the best combination of features was selected (7 descriptors) based on the internal and external validation parameters such as Q^2_{LOO} and Q^2_{F1} respectively (Roy et al., 2015) (elaborated in section **'Statistical validation metrics'** in **supplementary material 1**). Finally, partial least squares (PLS) regression (elaborated in **supplementary material 1**) (Cramer, 1993; Wold et al., 2001) was carried out to obtain the desired model with 7 selected descriptors of the training set, and it was rigorously validated with the test set of mixtures. We have also constructed various PLS plots such as coefficient plot, variable importance plot, score plot, loading plot, Y-randomization plot, and applicability domain (DModX) plot using the SIMCA-P software (Wu et al., 2010). To avoid the chances of overestimating validation metrics, we have performed 3 cross-validation tests namely mixtures out (Oprisiu et al., 2012), compounds out (Oprisiu et al., 2012) and everything out cross-validation (Muratov et al., 2014) in this study. A detailed explanation and procedures of these cross-validations have been given in **supplementary material 1**. According to Muratov et al., everything out cross-validation penalizes all possible interferences among test and training set compounds and thus provides realistic metric values of the model (Muratov et al., 2014). The complete process of model development along with the data division and descriptor calculation have been showcased schematically in Fig. 1.

2.5. Chemical read-across

Read-across is a similarity-based *in silico* method which predicts the toxicity (biological activity in general) of the untested chemicals from the known toxicity data of chemical analogues from the same chemical category (Raies and Bajic, 2016). It is one of the most important alternative approaches in chemical hazard assessment aimed to fill data gaps. Recently, we have developed and reported a similarity-based machine learning algorithm 'Read-across v4.1' (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) that predicts the toxicity of unknown substances (target chemicals) from the structural information of known analogues (source chemicals) (Banerjee and Roy, 2022; Chatterjee et al., 2022; Banerjee et al., 2022). Three different similarity estimation strategies such as Euclidean distance-based similarity, Gaussian kernel function similarity (GK), and Laplacian kernel function similarity (LK) have been utilized in this machine learning algorithm. The success of this method depends on the relevance of provided structural features (descriptors) towards toxicity based on which the similarity is calculated (Schultz et al., 2015). The software selects up to 10 most similar source compounds for each target chemical and predicts the responses by taking the weighted average of toxicity values of them. In the current study, we have applied chemical read-across for mixture toxicity prediction in one of the initial attempts. It has been also used as an additional tool to cross-check the model quality as well as the relevance of the modelled descriptors (See **supplementary material 1** for further explanation).

There are two hyperparameters namely sigma (for GK) and gamma (for LK) used in this tool, which should be optimized first for obtaining the best prediction. To optimize the hyperparameters, we randomly divided the model training set into 70% sub-training set and 30% calibration set and used them as source and target mixture respectively, by changing the values of sigma and gamma from 0.25 to 2 at 0.25 intervals. The optimized setting has been found based on the external validation metric Q^2_{F1} (provided by the tool itself). Finally, we subjected the training and test mixtures to 'Read-across v4.1' with gradual changes in the number of close training compounds from 10 to 2 to get the best prediction. One can also apply the distance (1–0) and similarity thresholds (0–1) instead of providing the number of close training compounds (10–2); however, both the determinants (thresholds of distance & similarity, and the number of close training compounds) cannot be used simultaneously. The workflow for read-across predictions is shown in Fig. S1 in **supplementary material 1**.

3. Results and discussions

3.1. The developed model and it's validation

In the present communication, we have developed a PLS regression model with the 2D structural features selected by the BSS algorithm from the reduced pool of descriptors after using GA and stepwise selection, as discussed in section '*2.4. Feature selection and model development*'. The statistical soundness of the model is proven by the computed validation metric values ("Validation of (Q)SAR Models," OECD, 2004). The final PLS model was developed with seven selected descriptors using five LVs, as depicted in the following equation (4) along with the computed internal and external validation metrics:

$$\begin{aligned} pEC50 = & -1.06168 + 3.21035 \times B03[N-0] + 0.42675 \times LOGP99 \\ & - 0.6651 \times B04[N-N] - 8.67482 \times Eta_D_epsiC \\ & + 0.49428 \times B07[N-Cl] - 1.44239 \times B08[C-Cl] \\ & - 1.65409 \times H - 054 \end{aligned} \quad (4)$$

$n_{Train} = 149$, $LV_s = 5$, $R^2 = 0.942$, $R^2_{adj} = 0.939$, $Q^2_{LOO} = 0.932$, $\overline{r^2_{m(LOO)}} = 0.904$, $\Delta r^2_{m(Train)} = 0.042$, $S_{Train} = 0.254$, $MAE_{Train} = 0.172$, $RMSE_C = 0.247$, Training set prediction quality (based on *MAE-based criteria)* = Good

$n_{Test} = 49, Q^2_{F1} = 0.95, Q^2_{F2} = 0.95, \overline{r^2_{m(Test)}} = 0.932, \Delta r^2_{m(Test)} = 0.014$, $CCC = 0.975$, $S_{Test} = 0.269, MAE_{Test} = 0.183$, $RMSE_P = 0.246$,Test set prediction quality *(based on MAE-based criteria)* = Good

The model showed a high coefficient of determination ($R^2 = 0.942$) with statistically significant R^2_{adj} (0.939) values which suggest the

Table 1
Modelled descriptors: definition, class, and type of contribution to toxicity.

Descriptor	Definition	Class	Contribution to toxicity
B03[N-O]	Presence/absence of N–O atom pair at topological distance 3	2D-atom pairs	Positive
LOGP99	Wildmann-Crippen octanol-water partition coefficient (logP)	Molecular properties	Positive
B04[N–N]	Presence/absence of N–N atom pair at topological distance 4	2D-atom pairs	Negative
Eta-D-epsiC	A measure of the contribution of electronegativity	ETA indices	Negative
B07 [N–Cl]	Presence/absence of N–Cl atom pair at topological distance 4	2D-atom pairs	Positive
B08[C–Cl]	Presence/absence of C–Cl atom pair at topological distance 8	2D-atom pairs	Negative
H-054	H-attached sp^3 hybridised carbon with 3 electronegative atoms at the next carbon/ carbons	Atom centred fragments	Negative

goodness-of-fit of the model. The high value of the cross-validated correlation coefficient ($Q^2_{LOO} = 0.932$), and a low standard error of estimate ($S_{Train} = 0.254$) signifies the robustness of the developed model. The reliability of predictions of the model was adjudged by the computed external correlation coefficient values ($Q^2_{F1} = 0.95$ and $Q^2_{F2} = 0.95$), which show good predictivity. The training and test set prediction qualities of the developed model based on the MAE-based criteria were proved to be "Good". Acceptable values of other validation parameters such as r^2_m metrics (Ojha et al., 2011) ($\overline{r^2_{m(LOO)}} = 0.904, \Delta r^2_{m(Train)} = 0.042$, $\overline{r^2_{m(Test)}} = 0.932, \Delta r^2_{m(Test)} = 0.014$), error measures ($RMSE_C = 0.247$, $RMSE_P = 0.247$, $MAE_{Train} = 0.172$, $MAE_{Test} = 0.183$), and concordance correlation coefficient ($CCC = 0.975$) etc. suggest the statistical quality and reliability of prediction of the studied model.

B03[N–O], LOGP99, B04[N–N], Eta_D_epsiC, B07[N–Cl], B08[C–Cl], and H-054 are the descriptors of interest in this research. The descriptor values of all the mixtures have been given in **supplementary material 2**. The definition, class, and the type of contribution of these modelled descriptors (See the regression coefficient plot in Fig. S2-A in **supplementary material 1**) have been tabulated in Table 1.

The additional loading and score plots (Ergon, 2004) of the PLS model were drawn using SIMCA-P software (Wu et al., 2010) and depicted in Fig. S2-B and S2-C respectively, in **supplementary material 1**. The score vectors of 1st two components t1 and t2 are plotted in the Score plot. The ellipse signifies the applicability domain of the model with 1st two latent variables; thus four mixtures (**161**, **166**, **167**, and **173**) are found to be the outliers with respect to LV1 and LV2. The loading plot provides a relationship between descriptors (X-variables) and toxicity (Y-variable) in the 1st two latent variable spaces (w1 and w2). Here in the loading plot, pEC50 along with seven descriptors have been drawn. If we look closely at the loading plot, then it is observed that the Eta_D_epsiC and B07[N–Cl] are placed at the farthest and nearest points from the origin, hence they are signified as the most important and least important descriptors respectively.

We performed the Y-randomization test (Rucker et al., 2007) with the developed model using SIMCA-P software (Wu et al., 2010) to confirm whether the model is the result of any chance correlation or not (elaborated result in **supplementary material 1**). The small values of R^2 and Q^2 intercepts (($R^2_Y = -0.01$ and $Q^2_Y = -0.33$)) justify the robustness of the reported model, and also cancel out the possibilities of chance correlation.

We have also analysed the applicability domain (Gadaleta et al., 2016) of the developed PLS model by DModX (distance to model X) approach (De et al., 2018) using SIMCA-P software (Wu et al., 2010) with a 99% confidence interval. Three training mixtures (a 0.25:0.75 mixture of omeprazole and imatinib, a 0.75:0.25 mixture of imatinib and tobramycin, and a 0.149:0.851 mixture of tetracycline hydrochloride and linuron) were found outside the standard 99% D-critical line (D-Crit = 0.00999898) of Fig. S3 in **supplementary material 1** whereas all the test mixtures were inside the threshold. The applicability domain analysis has been elaborated in **supplementary material 1** along with the derived plots.

Table 2
Computed metrics for mixtures out, compounds out, and everything out cross-validation.

Internal validation metrics	Original PLS model	Mixture out cross-validation	Compounds out cross-validation	Everything out cross-validation
R^2	0.942	0.958	0.912	0.807
Q^2_{LOO}	0.932	0.950	0.883	0.777
MAE-LOO (Train)	0.205	0.188	0.185	0.172
External validation metrics				
Q^2_{F1}	0.950	0.913	0.968	0.859
MAE_{Test}	0.183	0.208	0.234	**0.653**

3.2. Results of external cross-validation

The developed model is proved to be robust and extremely predictive as per the OECD guidelines for QSAR modelling of individual data points. However, we have also performed the mixtures out, compounds out, and everything out cross-validations to penalize the overestimation of validation metrics if any. To perform the cross-validation, we shuffled the data points of the original training-test pairs (4-4 Euclidean distance-based division) as discussed in section **'Cross-validation'** in **supplementary material 1** (mixtures out, compounds out, and everything out validation sets have been shown as Table S4 to **S6** respectively in **supplementary material 1**). The internal and external validation parameters of the cross-validated models were computed with the shuffled training-test pairs and given in Table 2.

All three cross-validated models have shown acceptable validation metric values (except MAE_{Test} in everything out CV) which signifies the statistical reliability of the developed model and thus counters the chances of misjudgement of model quality. An expected deterioration of the model quality has been observed in everything out cross-validation (more profound in the case of the MAE_{Test} value (0.653)) which justifies the complete structural exclusiveness among training and test set mixtures. The observed vs. predicted scatter plots of cross-validated models along with the original PLS model have been given in Fig. 2. A uniform scattering has been observed around the lines passing through the origins in all the cases which suggests the goodness-of-fit and robustness of the developed model.

3.3. Study of modelled descriptors

In this present research, the toxicity of binary and multi-component mixtures has been modelled with seven 2D descriptors as seen in section '***3.1. The developed model and it's validation***'. These modelled descriptors essentially represent seven unique structural features which are responsible for the toxic action. These descriptors are depicted in Fig. 3 as the variable importance plot (Wold et al., 2001) with their relative importance towards the toxic endpoint (pEC50). A descriptor with VIP score >1 suggests a significant impact on toxicity. In this study, Eta_D_epsiC and B03[N–O] have VIP scores of more than 1, among which *Eta_D_epsiC* is the most important. The importance of the rest descriptors such as H-054, B08[C–Cl], LOGP9, B04[N–N], and B07 [N–Cl] is gradually declining. The relationship of the descriptors with toxicity is also depicted in Fig. 3. B03[N–O], LOGP99, B07[N–Cl] show a positive contribution to the modelled response whereas, Eta_D_epsiC, H-054, and B08[C–Cl], and B04[N–N] have negative contributions towards the toxicity. We have depicted all the defined structural fragments as 2D diagrams in the different studied mixture components in the

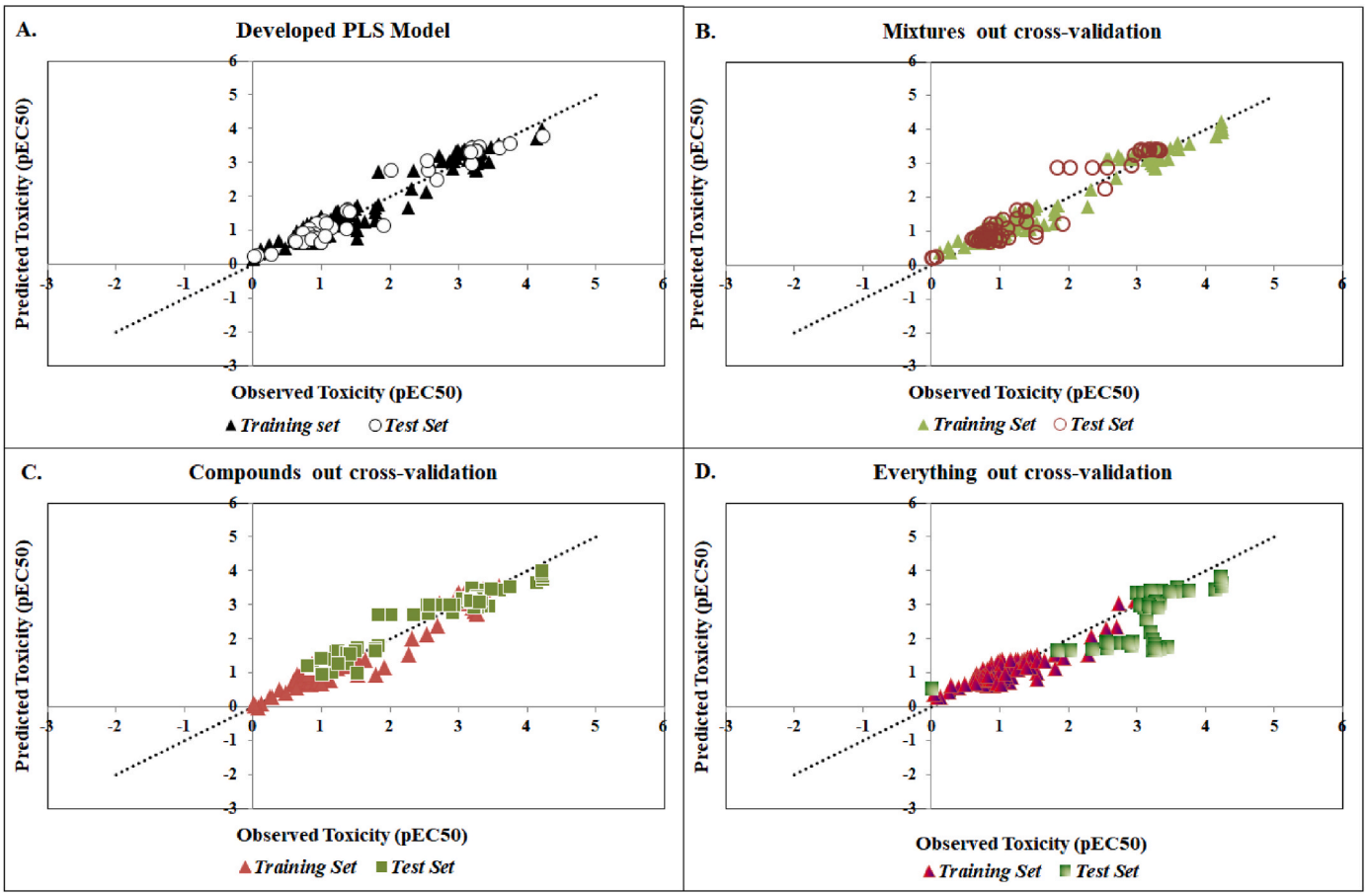


Fig. 2. Observed vs. predicted scatter plots (A. Original PLS model, B. Mixtures out cross-validation, C. Compounds out cross-validation, D. Everything out cross-validation).

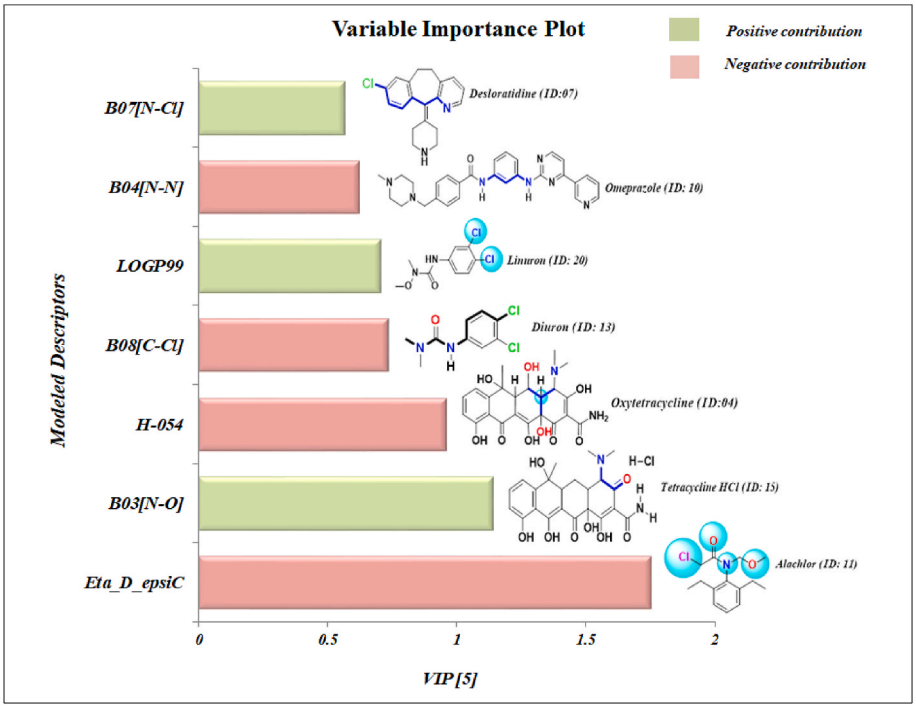


Fig. 3. Variable importance plot with the 2D representation of the modelled descriptors.

following Fig. 3.

Eta_D_epsiC ($\Delta\varepsilon_C$) is an extended topochemical atom descriptor which measures the effect of electronegativity (Roy and Das, 2015). The descriptor can be mathematically expressed as $\Delta\boldsymbol{\varepsilon_C} = \boldsymbol{\varepsilon_3} - \boldsymbol{\varepsilon_4}$, where $\boldsymbol{\varepsilon_3}$ and $\boldsymbol{\varepsilon_4}$ are defind in equations (5) and (6) respectively.

$$\varepsilon_3 = \frac{[\sum \varepsilon]_R}{N_R} \tag{5}$$

$$\varepsilon_4 = \frac{[\sum \varepsilon]_{SS}}{N_{SS}} \tag{6}$$

According to Roy et al., ε_3 is "*the sum of epsilon (ε) relative to the total number of atoms present in the molecule(N_R) including hydrogen in the connected molecular graph of reference alkane*", whereas, ε_4 is "*the sum of epsilon (ε) relative to the total number of atoms present in the molecule (N_{SS}) including hydrogen for a saturated carbon skeleton of the molecule*" (Roy and Ghosh, 2004). The reference alkane can be defined as the molecular graph where all the heteroatoms and covalent multiple bonds are replaced with carbons and single bonds respectively; on the contrary, the saturated carbon skeleton is the saturated molecular graph of the corresponding molecule without changing the heteroatoms with carbon (De et al., 2018). From this definition, it is understood that the ε_3 reduces with an increasing number of electronegative atoms in the molecule (due to the replacement of heteroatoms (electronegative) with carbon), however ε_4 remains unaltered. Thus, the resultant $\Delta\varepsilon_C$ becomes more negative with an increasing number of electronegative atoms. It has been observed in the studied mixtures also. We have compared the change of toxicity from mixture **67** to mixture **80** in Fig. S4 in **Supplementary material 1**. Mixture 67 (Eta_D_epsiC = −0.075) is an equimolar combination of isoproturon (#**12**) and alachlor (#**11**) whereas mixture 80 (Eta_D_epsiC = −0.153) is a 0.25:0.75 combination of alachlor (#**11**) and chlorfenvinphos (#**14**). Chlorfenvinphos has more electronegative atoms in its structure than isoproturon and this structural difference makes the value of Eta_D_epsiC more negative in mixture **80** as compared to mixture **67** (as alachlor is a constant component between the two mixtures). An increament of the toxicity has also been observed from mixture **67** (pEC50 = 0.88) to mixture **80** (pEC50 = 1.03), which justifies the negative contribution of the descriptor Eta_D_epsiC to the toxicity.

The second most important descriptor is the B03[N–O] which belongs to the 2D atom pairs class of descriptors. It can have binary values 0 and 1 which essentially signifies the absence and presence of the [nitrogen-oxygen] atom pair at topological distance 3 respectively. The presence of this feature in a mixture increases the toxicity as shown in Fig. S5 in **supplementary material 1**. This observation justifies the positive correlation of B03[N–O] with the toxic endpoint. A mixture (#**169**) of imatinib (#**8**) and desloratadine (#**7**) does not have this feature and hence shows less toxicity (pEC50 = 0.49) than another mixture **163** (pEC50 = 2.55) with the B03[N–O] feature in the component's structures desloratadine (#**7**) and omeprazole (#**10**). Omeprazole contains this particular feature and hence makes the key difference here (desloratadine is the constant component between two mixtures).

H-054 has grabbed the third place in the variable importance plot (Fig. 3) of the current model. It is an atom-centred fragment descriptor which is defined by a hydrogen-attached sp^3 hybridized carbon-containing three electronegative atoms in the adjacent carbon/carbons. This descriptor has shown a negative contribution to the toxicity of binary and multi-component mixtures. Oxytetracycline (#**4**) is a mixture component containing the specified fragment that is depicted in Fig. S6 in **supplementary material 1**. It has been observed that the mixture (#**105**) containing oxytetracycline has a greater H-054 value and a smaller toxicity value (pEC50 = 1.65) in comparison with the mixture (#**186:** H-054 = 0, pEC50 = 4.25) devoid of oxytetracycline. Therefore, the inverse relationship between H-054 and the mixtures' toxicity can be explained by these observations.

B08[C–Cl] is another descriptor of the model, which signifies the presence or absence of carbon-chlorine atom pair at topological distance 8. It belongs to the 2D atom pairs class of descriptors. In this study, B08[C–Cl] was negatively correlated to the modelled response. It has been observed that a less toxic mixture **173** (pEC50 = 0.02) has a higher B08[C–Cl] value (0.75). The opposite situation has been found in a more toxic mixture **180** (pEC50 = 3.35) where B08[C–Cl] is zero (Fig. S7 in **Supplementary material 1**). Thus, the negative contribution of this descriptor is well explained by these observations. From a close look at the components of mixtures **173** and **180**, it has been found that desloratadine (#**7**) is responsible for the high B08[C–Cl] value of mixture **173** (Fig. S7 in **Supplementary material 1**). It may be concluded from these observations that the presence of chlorine atom with a large and fused ring system is beneficial from the regulatory point of view because this feature will help to reduce toxicity.

LOGP99 is a molecular property descriptor which represents the Wildmann-Crippen octanol-water partition coefficient (logP). It is a measure of lipophilicity and contributes positively to the mixtures' toxicity in this research. It has been observed from the studied mixtures that the chlorine or other halogen-containing mixtures are more lipophilic. The halogen atoms increase the molecular size and London dispersion forces to some extent which indirectly increases the lipophilicity(Pliska et al., 2008). More lipophilic compounds can cross the lipopolysaccharide-containing cell wall of *Aliivibrio fischeri*, and hence increase the toxicity. Mixture **186** contains tetracycline hydrochloride (#15) and a chlorine-containing component linuron (#20), which have shown more LOGP99 (1.63) and pEC50 (4.25) values (Fig. S8 in **Supplementary material 1**). On the other hand, a combination (mixture **167**) of imatinib (#**8**) and tobramycin (#**9**) has shown less LOGP99 (−3.57) and pEC50 (0.13) values, which justify the proportional relationship between them.

B04[N–N] is a 2D atom pair descriptor which is found to be inversely correlated with the toxic endpoint of this study. It is defined by the presence of nitrogen atom pairs in topological distance 4. This feature has been observed in mixture components such as imatinib (#**8**) and tobramycin (#**9**) as seen in Fig. S9 in **Supplementary material 1**. In both cases, two amino (-NH2) or substituted amino (–NH–) groups are found at alternative carbons concerning a benzene or cyclohexane ring. In the above-mentioned figure, it has been observed that the presence of this feature reduces toxicity to a large extent. This negative relationship is also evident in a less toxic mixture **166** with nitrogen atom pair at topological distance 4 (B04[N–N] = 1, pEC50 = 0.26) and a more toxic mixture **186** without this feature (B04[N–N] = 0, pEC50 = 4.25).

The least important descriptor in this research to explain the toxicity is B07[N–C]. It is again a 2D atom pair descriptor which signifies the nitrogen and chlorine atom pair at the topological distance 7. This feature has been shown in the structure of desloratadine (#**7**) in Fig. 3 and Fig. S10 in **supplementary material 1**. The coefficient plot (Fig. S2-A) of this model suggests a positive correlation of B07[N–N] with the toxicity. To explain this, we have taken up two mixtures **122** and **162** (Fig. S10 in **supplementary material 1**). Mixture **162** is a 0.75: 0.25 mass combination of omeprazole (#**10**) and desloratadine (#**7**), whereas mixture **122** is a 0.25: 0.75 mass combination of dexamethasone (#**5**) and isoproturon (#**12**). Among these four components, only desloratadine possesses N–Cl atom pair at a topological distance of 7 and hence, mixture **162** shows a positive descriptor value (B07[N–N] = 0.25). As per the expectation, mixture **162** has imparted a greater toxic response (pEC50 = 2.98) than mixture **122** (pEC50 = 0.63) which does not have this feature. Therefore, the proportional relationship between pEC50 and B07[N–N] is justified.

3.4. Result of chemical read-across

The machine learning-based chemical read-across has been performed in this study as an additional tool to cross-check the relevance of identified descriptors. To do so, the hyperparameters σ and γ were

Table 3
Comparative assessment of the proposed model with previous reports of toxicity (pEC50) of binary and multi-component mixtures against *Aliivibrio fischeri*.

Variables	Model reported by Qin et al., in 2018	Model reported by Bures et al., in 2021	The proposed QSAR model in this study	Similarity-based read-across in this study
Data points	45	153	198	198
Type of model	GA-MLR	GA-MLR	PLS	Weighted average prediction (Laplacian kernel)
No. of descriptors	3	4	**7 (LVs = 5)**	**7**
Internal validation				
n_{Train}	31	123	149	149
R^2	0.936	0.901	**0.941**	-
Q^2_{LOO}	0.908	0.863	**0.931**	-
External Validation				
n_{Test}	14	30	49	49
Q^2_{F1}	0.748	0.863	**0.950**	**0.919**

GA-MLR: genetic algorithm-multiple linear regression; PLS: partial least squares; LVs: latent variables.

optimized first as discussed in the section '***2.5. Chemical read-across***'. The optimization of the parameters is given in Table S7 in **Supplementary material 1**. From the previously mentioned table, σ = 2 and γ = 0.5 have been identified as the optimized values as in these two conditions, the best Q^2_{F1} value (0.644 for both GK and LK) is obtained without changing the other determinants. In the next step, these optimized settings are used to get the best prediction of the test set from the training set. The best prediction has been obtained in the setting σ = 2, γ = 0.5, number of close training compounds = 2, distance threshold = 1, and similarity threshold = 0. The computed external validation metrics are given in the following Table S8 in **supplementary material 1** as the quality measure of prediction. It has been observed that the external predictive variance values obtained by the three methods are high enough to justify the relevance of the selected descriptors. The error measures RMSEp and MAE are also on the lower side which shows the reliability of this justification. Among the three similarity estimation techniques of read-across, the best prediction has been obtained by the Laplacian kernel function similarity ($\mathbf{Q^2_{F1} = 0.919}$).

4. A comparative assessment of the model with previous reports

To justify the merit of this study, we have compared the proposed QSAR model with two previously published literature reports (Qin et al., 2018; Sigurnjak Bureš et al., 2021) for the aquatic toxicity of binary and multi-component mixtures of pharmaceuticals and pesticides against *Aliivibrio fischeri* in terms of the quality measures. The validation metrics (R^2, Q^2_{LOO}, and Q^2_{F1}) of these three models are summarised in the following Table 3. The table clearly depicts the better quality of the proposed QSAR model than the previous reports in terms of goodness-of-fit, robustness, and external predictivity. The chemical read-across also provides better prediction than the previous results but is inferior to the QSAR model. The current model covers a comparatively large spectrum of molecular diversity than the previous models which is evident from the number of training data (n_{Train} = 153, 45, 123 in the proposed model, model of Qin et al., and model of Bures et al. respectively). The previous models were essentially composed of 3D features along with 2D descriptors; hence calculation and interpretation were not so easy. On the contrary, the newly developed model has been constructed with simple 2D structural descriptors, which makes the model easily interpretable and useable for quick predictions. Therefore, this comparative assessment strongly suggests better quality, interpretability, and transferability of the proposed model with fewer and easier necessary prerequisites for prediction.

5. Overview and conclusion

This research mainly focuses on the use of alternative *in silico* approaches (especially QSAR and chemical read-across) to the toxicity prediction of binary and multi-component mixtures. Chemical mixtures are no longer an extended reality in the perspective of environmental components in this era of chemical advancement. After confirming the presence of pharmaceuticals and pesticides in wastewater streams, and by keeping in mind their probable effect on the environmental species, the study was designed with pharmaceutical and pesticide mixtures against the aquatic bacterium *Aliivibrio fischeri*. A PLS regression model was developed in this study with the 2D structural descriptors, which is proved as an efficient and reliable model. The use of 2D descriptors has sufficiently reduced the computation time and made the model easily transferable and interpretable. The weighted descriptor generation approach was used to impart the mixture effect in this study, and three cross-validation strategies were applied to counter the chances of overestimating validation metrics. The results of these cross-validations clearly suggested the efficiency of the developed model. The external validation has been performed with a multicomponent mixture-containing test set, and good model quality has been obtained which suggests the efficiency of the model to predict multicomponent mixtures. This model successfully identified seven structural features which are responsible for the mixture's toxicity, and also demonstrated their roles on toxicity enhancement. We compared the quality of the developed QSAR model with previous reports and found the better quality, fast applicability, and easy interpretability of our model. We applied a newly designed tool based on chemical read-across prediction in this study to check the relevance of modelled descriptors. The read-across predictions were found to be good with respect to quality measures and the quality of predictions was also reliable. In a nutshell, a mixture QSAR has been reported in this study with better quality of prediction, machine learning-based quantitative read-across predictions have been successfully used to mixture toxicity assessment with the descriptors obtained in the QSAR, and the relevance of the selected descriptors has been successfully established by the result of machine learning-based chemical read-across algorithm.

Author contributions

MC: Data curation, Model development and validation, Analysis, First draft; KR: Supervision, Conceptualization, Editing, Funding.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

MC sincerely thanks All India Council for Technical Education (AICTE), New Delhi for financial support in the form of National Doctoral Fellowship (NDF). KR thanks SERB, Govt of India for financial assistance under the MATRICS scheme (MTR/2019/000008).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.136463.

References

Abbas, M., Adil, M., Ehtisham-ul-Haque, S., Munir, B., Yameen, M., Ghaffar, A., Shar, G. A., Asif Tahir, M., Iqbal, M., 2018. Vibrio fischeri bioluminescence inhibition assay for ecotoxicity assessment: a review. Sci. Total Environ. 626, 1295–1309. https://doi.org/10.1016/j.scitotenv.2018.01.066.

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.

Banerjee, A., Chaterjee, M., De, P., Roy, K., 2022. Quantitative predictions from chemical read-across and their confidence measures. Chemometr. Intell. Lab. Syst. 227, 104613 https://doi.org/10.1016/j.chemolab.2022.104613.

Bliss, C.I., 1939. The toxicity of poison applied jointly. Ann. Appl. Biol. 26, 585–615. https://doi.org/10.1111/j.1744-7348.1939.tb06990.x.

Carley, M., Christie, I., 2019. The ecology of an industrial planet. In: Managing Sustainable Development. Routledge, London, pp. 3–24.

Carnesecchi, E., Toropov, A.A., Toropova, A.P., Kramer, N., Svendsen, C., Dorne, J., Lou Benfenati, E., 2020. Predicting acute contact toxicity of organic binary mixtures in honey bees (A. mellifera) through innovative QSAR models. Sci. Total Environ. 704, 135302 https://doi.org/10.1016/j.scitotenv.2019.135302.

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., 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.

Consonni, V., Todeschini, R., 2010. Molecular descriptors. In: Puzyn, T., Leszczynski, J., Cronin, M. (Eds.), Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9783-6_3.

Cramer, R.D., 1993. Partial least squares (PLS): its strength and limitations. Perspect. Drug Discov. Des. 1, 169–178. https://doi.org/10.1007/bf02174528.

Daughton, C.G., 2003. Cradle-to-cradle stewardship of drugs for minimizing their environmental disposition while promoting human health. I. Rational for and avenues toward a green pharmacy. Environ. Health Perspect. 111, 757–774. https://doi.org/10.1289/ehp.5947.

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 Adv. 8, 4662–4670. https://doi.org/10.1039/c7ra13159c.

Elliott, S.M., Brigham, M.E., Kiesling, R.L., Schoenfuss, H.L., Jorgenson, Z.G., 2018. Environmentally relevant chemical mixtures of concern in waters of United States tributaries to the Great Lakes. Integrated Environ. Assess. Manag. 14, 509–518. https://doi.org/10.1002/ieam.4041.

Ergon, R., 2004. Informative PLS score-loading plots for process understanding and monitoring. J. Process Control 14, 889–897. https://doi.org/10.1016/j.jprocont.2004.02.004.

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

Gallegos Saliner, A.B.S.P., 2006. Mini-Review on chemical similarity and prediction of toxicity. Curr. Comput. Aided Drug Des. 2, 105–122. https://doi.org/10.2174/157340906777441681.

Gao, L., Shi, Y., Li, W., Liu, J., Cai, Y., 2012. Occurrence, distribution and bioaccumulation of antibiotics in the Haihe River in China. J. Environ. Monit. 14, 1248–1255. https://doi.org/10.1039/c2em10916f.

Ghanem, O. Ben, Mutalib, M.I.A., Lévêque, J.M., El-Harbawi, M., 2017. Development of QSAR model to predict the ecotoxicity of Vibrio fischeri using COSMO-RS descriptors. Chemosphere 170, 242–250. https://doi.org/10.1016/j.chemosphere.2016.12.003.

Hao, Y., Sun, G., Fan, T., Tang, X., Zhang, J., Liu, Y., Zhang, N., Zhao, L., Zhong, R., Peng, Y., 2020. In vivo toxicity of nitroaromatic compounds to rats: QSTR modelling and interspecies toxicity relationship with mouse. J. Hazard Mater. 399, 122981 https://doi.org/10.1016/j.jhazmat.2020.122981.

Hoover, G., Kar, S., Guffey, S., Leszczynski, J., Sepúlveda, M.S., 2019. In vitro and in silico modeling of per fl uoroalkyl substances mixture toxicity in an amphibian fi broblast cell line. Chemosphere 233, 25–33. https://doi.org/10.1016/j.chemosphere.2019.05.065.

Iwasaki, Y., Gauthier, P., 2016. Concentration addition and response addition to analyze mixture toxicity: is it worth testing? Environ. Toxicol. Chem. 35, 526–527. https://doi.org/10.1002/etc.3263.

Kar, S., Leszczynski, J., 2019. Exploration of computational approaches to predict the toxicity of chemical mixtures. Toxics 7, 15. https://doi.org/10.3390/TOXICS7010015.

Kar, S., Sanderson, H., Roy, K., Benfenati, E., Leszczynski, J., 2020. Ecotoxicological assessment of pharmaceuticals and personal care products using predictive toxicology approaches. Green Chem. 22, 1458–1516. https://doi.org/10.1039/c9gc03265g.

Kemper, N., 2008. Veterinary antibiotics in the aquatic and terrestrial environment. Ecol. Indicat. 8, 1–13. https://doi.org/10.1016/j.ecolind.2007.06.002.

Kim, J., Kim, S., Schaumann, G.E., 2013. Reliable predictive computational toxicology methods for mixture toxicity: toward the development of innovative integrated models for environmental risk assessment. Rev. Environ. Sci. Biotechnol. 12, 235–256. https://doi.org/10.1007/S11157-012-9286-7.

Klatte, S., Schaefer, H.C., Hempel, M., 2017. Pharmaceuticals in the environment – a short review on options to minimize the exposure of humans, animals and ecosystems. Sustain. Chem. Pharm. 5, 61–66. https://doi.org/10.1016/j.scp.2016.07.001.

Kolpin, D.W., Furlong, E.T., Meyer, M.T., Thurman, E.M., Zaugg, S.D., Barber, L.B., Buxton, H.T., 2002. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. Streams, 1999−2000: a national reconnaissance. Environ. Sci. Technol. 36, 1202–1211. https://doi.org/10.1021/es011055J.

Kortenkamp, A., Altenburger, R., 1999. Approaches to assessing combination effects of oestrogenic environmental pollutants. Sci. Total Environ. 233, 131–140. https://doi.org/10.1016/S0048-9697(99)00228-4.

Lamastra, L., Balderacchi, M., Trevisan, M., 2016. Inclusion of emerging organic contaminants in groundwater monitoring plans. MethodsX 3, 459–476. https://doi.org/10.1016/j.mex.2016.05.008.

Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N., Nil), Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., van Schayck, O.C.P., Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. Lancet 391, 462–512. https://doi.org/10.1016/s0140-6736(17)32345-0.

Leardi, R., 1996. Genetic algorithms in feature selection. In: Devillers, J. (Ed.), Genetic Algorithms in Molecular Modeling. Academic Press (Elsevier), Great Britain, pp. 67–86. https://doi.org/10.1016/B978-012213810-2/50004-9.

Li, F., Fan, T., Sun, G., Zhao, L., Zhong, R., Peng, Y., 2022. Systematic QSAR and iQCCR modelling of fused/non-fused aromatic hydrocarbons (FNFAHs) carcinogenicity to rodents: reducing unnecessary chemical synthesis and animal testing. Green Chem. 24, 5304–5319. https://doi.org/10.1039/D2GC00986B.

Loewe, S., Muischnek, H., 1926. Effect of combinations: mathematical basis of problem. N-S Arch. Exp. Pathol. Pharmakology 114, 313–326.

Luo, Y., Guo, W., Ngo, H.H., Nghiem, L.D., Hai, F.I., Zhang, J., Liang, S., Wang, X.C., 2014. A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. Sci. Total Environ. 473, 619–641. https://doi.org/10.1016/j.scitotenv.2013.12.065. –474.

Mansour, F., Al-Hindi, M., Saad, W., Salam, D., 2016. Environmental risk analysis and prioritization of pharmaceuticals in a developing world context. Sci. Total Environ. 557 (558), 31–43. https://doi.org/10.1016/j.scitotenv.2016.03.023.

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.

URL Minitab Statistical Software [WWW Document], 2022. accessed 3.1.22. https://www.minitab.com/en-us/products/minitab/.

Mukherjee, R.K., Kumar, V., Roy, K., 2022. Ecotoxicological QSTR and QSTTR modeling for the prediction of acute oral toxicity of pesticides against multiple avian species. Environ. Sci. Technol. 56, 335–348. https://doi.org/10.1021/acs.est.1c05732.

Muratov, E.N., Tropsha, A., Varlamova, E.V., Kuzmin, V.E., Artemenko, A.G., Muratov, N.N., Mileyko, S., Fourches, D., 2014. Everything out" validation approach for qsar models of chemical mixtures. JSM Clin. Pharm. 1, 1005.

Nazemzadeh, N., Nielsen, R.F., Gernaey, K.V., Mansouri, S.S., Andersson, M.P., 2021. Implementation of first-principles surface interactions in a hybrid machine learning assisted modelling of flocculation. Comput. Aided Chem. Eng. 50, 845–850. https://doi.org/10.1016/B978-0-323-88506-5.50132-7.

URL Validation of (Q)SAR Models - OECD [WWW Document], 2004. accessed 4.20.22. https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm.

Ojha, P.K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring rm2 metrics for validation of QSPR models. Chemometr. Intell. Lab. Syst. 107, 194–205. https://doi.org/10.1016/j.chemolab.2011.03.011.

Oprisiu, I., Varlamova, E., Muratov, E., Artemenko, A., Marcou, G., Polishchuk, P., Kuz'Min, V., Varnek, A., 2012. QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. Mol. Inform. 31, 491–502. https://doi.org/10.1002/minf.201200006.

Pliska, V., Testa, B., Waterbeemd, H. van de, 2008. Lipophilicity in drug action and toxicology. In: Testa, B., Waterbeemd, H. van de (Eds.), Methods and Principles in Medicinal Chemistry. VCH, Weinheim, Germany, pp. 311–337.

Qin, L., Chen, Y., Zhang, X., Mo, L., Zeng, H., 2018. QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. Chemosphere 198, 122–129. https://doi.org/10.1016/j.chemosphere.2018.01.142.

Qin, L.T., Liu, M., Zhang, X., Mo, L.Y., Zeng, H.H., Liang, Y.P., 2021. Concentration addition, independent action, and quantitative structure-activity relationships for chemical mixture toxicities of the disinfection by products of haloacetic acids on the green alga raphidocelis subcapitata. Environ. Toxicol. Chem. 40, 1431–1442. https://doi.org/10.1002/etc.4995.

Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. WIREs Comput. Mol. Sci. 6, 147–172. https://doi.org/10.1002/wcms.1240.

Ritskes-Hoitinga, M., Van Luijk, J., 2019. How can systematic reviews teach us more about the implementation of the 3Rs and animal welfare? Animals 9, 1163. https://doi.org/10.3390/ANI9121163.

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. https://doi.org/10.1021/ci00020a020.

Roy, K., Das, R.N., 2015. The “ETA” indices in QSAR/QSPR/QSTR research. In: Roy, K. (Ed.), Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. IGI Global, Hershey PA, USA, pp. 48–83. https://doi.org/10.4018/978-1-4666-8136-1.ch015.

Roy, K., Ghosh, G., 2004. QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. J. Chem. Inf. Comput. Sci. 44, 559–567. https://doi.org/10.1021/ci0342066.

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

Rucker, C., Rucker, G., Meringer, M., 2007. y-Randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–2357. https://doi.org/10.1021/ci700157b.

Schultz, T.W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D.J., Mahony, C., Schwarz, M., White, A., Cronin, M.T.D., 2015. A strategy for structuring and reporting a read-across prediction of toxicity. Regul. Toxicol. Pharmacol. 72, 586–601. https://doi.org/10.1016/j.yrtph.2015.05.016.

Sigurnjak Bureš, M., Ukić Cvetnić, M., Prevarić, V., Markić, M., Rogošić, M., Kušić, H., Bolanča, T., 2021. Toxicity of binary mixtures of pesticides and pharmaceuticals toward Vibrio fischeri: assessment by quantitative structure-activity relationships. Environ. Pollut. 275, 115885 https://doi.org/10.1016/j.envpol.2020.115885.

Sun, G., Zhang, Y., Pei, L., Lou, Y., Mu, Y., Yun, J., Li, F., Wang, Y., Hao, Z., Xi, S., Li, C., Chen, C., Zhao, L., Zhang, N., Zhong, R., Peng, Y., 2021. Chemometric QSAR modeling of acute oral toxicity of polycyclic aromatic hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse. Ecotoxicol. Environ. Saf. 222, 112525 https://doi.org/10.1016/j.ecoenv.2021.112525.

Topp, E., Renaud, J., Sumarah, M., Sabourin, L., 2016. Reduced persistence of the macrolide antibiotics erythromycin, clarithromycin and azithromycin in agricultural soil following several years of exposure in the field. Sci. Total Environ. 562, 136–144. https://doi.org/10.1016/j.scitotenv.2016.03.210.

Wagner, J.M., Shimshak, D.G., 2007. Stepwise selection of variables in data envelopment analysis: procedures and managerial perspectives. Eur. J. Oper. Res. 180, 57–67. https://doi.org/10.1016/j.ejor.2006.02.048.

Wang, T., Tang, L., Luan, F., Cordeiro, M.N.D.S., 2018. Prediction of the toxicity of binary mixtures by QSAR approach using the hypothetical descriptors. Int. J. Mol. Sci. 19, 3423. https://doi.org/10.3390/ijms19113423.

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

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

Journal of Hazardous Materials 460 (2023) 132358

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Research Paper

Machine learning - based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

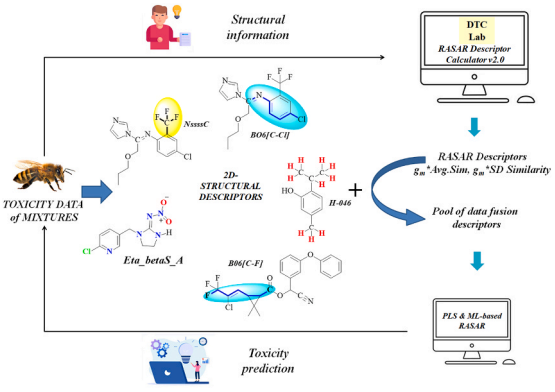
Mainak Chatterjee [a], Arkaprava Banerjee [a], Simone Tosi [b], Edoardo Carnesecchi [c], Emilio Benfenati [d], Kunal Roy [a,*]

[a] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
[b] Department of Agricultural, Forest, and Food Sciences, University of Turin, Turin, Italy
[c] Institute for Risk Assessment Sciences, Utrecht, the Netherlands
[d] Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCSS, via Mario Negri 2, 20156 Milano, Italy

HIGHLIGHTS

- A novel q-RASAR model has been developed for the prediction of toxicity of organic mixtures in honey bees.
- Three different mixing rules have been used to calculate the mixture descriptors.
- The developed model has been validated following the strict OECD guidelines.
- The use of machine learning-based algorithms further enhanced the predictability of the q-RASAR model.
- The toxicity of environmentally relevant untested organic mixtures has been predicted by this new model.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Youn-Joo An

Keywords:
Mixing rules
Ecotoxicity
Predictability
Q-RASAR
Artificial intelligence
Bees

ABSTRACT

We have reported here a quantitative read-across structure-activity relationship (q-RASAR) model for the prediction of binary mixture toxicity (acute contact toxicity) in honey bees. Both the quantitative structure-activity relationship (QSAR) and the similarity-based read-across algorithms are used simultaneously for enhancing the predictability of the model. Several similarity and error-based parameters, obtained from the read-across prediction tool, have been put together with the structural and physicochemical descriptors to develop the final q-RASAR model. The calculated statistical and validation metrics indicate the goodness-of-fit, robustness, and good predictability of the partial least squares (PLS) regression model. Machine learning algorithms like ridge regression, linear support vector machine (SVM), and non-linear SVM have been used to further enhance the predictability of the q-RASAR model. The prediction quality of the q-RASAR models outperforms the previously reported quasi-SMILEs-based QSAR model in terms of external correlation coefficient (Q^2_{F1} SVM q-RASAR: 0.935 vs. Q^2_{VLD} QSAR: 0.89). In this research, the toxicity values of several new untested binary mixtures have been predicted with the new models, and the reliability of the PLS predictions has been validated by the prediction

* Corresponding author.
E-mail addresses: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).

https://doi.org/10.1016/j.jhazmat.2023.132358
Received 22 December 2022; Received in revised form 2 August 2023; Accepted 20 August 2023
Available online 22 August 2023
0304-3894/© 2023 Elsevier B.V. All rights reserved.

reliability indicator tool. The q-RASAR approach can be used as reliable, complementary, and integrative to the conventional experimental approaches of pesticide mixture risk assessment.

1. Introduction

The honey bee (*Apis mellifera*), a major pollinator species, plays a key role in food production and biodiversity protection, and environmental sustainability [1,2]. As per the Food and Agriculture Organization (FAO) of the United Nations, "a third of world's food production depends on bees" [3], which establishes our dependency on honey bees for food security. They are also an important source of income for beekeepers and a unique source of products (such as honey, royal jelly, pollen, etc) for the consumers worldwide. The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) has stated - "Sacred passages about bees in all the worlds' major religions highlight their significance to human societies over millennia" [4]. It thus raises concern about the continuous decline of bee health, generally linked to human activities [5], including habitat loss, extensive farming, and excessive use of agrochemicals (i.e., plant protection products such as pesticides) [2]. Because a wide range of chemicals is used as plant protection products in the crop fields, honey bees are simultaneously exposed to a wide range of active substances used as ingredients of plant protection products rather than single chemical entities [6]. Mixture toxicity estimation is thus a more realistic approach for the risk assessment of agrochemicals in honey bees. Environmental scientists and scientific advisory bodies of governmental agencies (e.g., The European Food Safety Agency, European Chemicals Agency, U.S. Environmental Protection Agency, etc.) are aware of this issue, and they use the honey bee as surrogate species for individual pesticide testing [7]. However, it is practically impossible to test (e.g., in vivo) all possible chemical combinations which honey bees may be exposed to under real condition at crop fields. *In silico* predictive approaches become thus instrumental in such situations. They are cost-effective, fast, and efficient methods for the toxicity/activity/property predictions. The use of *in silico* predictions (quantitative structure-activity relationships or QSARs and read-across) for regulatory purposes is already approved, and the corresponding scientific advisory bodies are trying to adopt the use of computer-based new approach methodologies for environmental risk assessment of individual toxicants [8]. One may find a good number of studies on individual toxicity modeling in the literature; but mixture toxicity modeling in honey bees is still rare. Carnesecchi et al. reported an innovative QSAR modeling study in 2020 for the toxicity prediction of binary mixtures in honey bees [9]. They incorporated the toxic unit of individual components of mixtures as an important determinant of mixture toxicity prediction which enhance the predictability impressively. Tosi et al. used innovative approaches (i.e., toxic units) and demonstrated a concerning major lack of data on the sublethal (71% of pesticides with missing sublethal data) and combined (about 99% of pesticides with missing mixture data) toxicity of pesticides in bees [10]; this knowledge gap is largely unfulfilled by regulatory studies performed for pesticide authorisation and risk assessment. This leads to a serious concern, especially when the number of combinations and their compositions are poorly investigated [6,10]. Therefore, the objective of this research is to develop a novel and efficient *in silico* model with enhanced external predictability that can reliably predict the exerted toxicity of new or untested binary chemical mixtures against honey bees without using additional in vivo toxicity data. Besides this, we have also tried to find a suitable mixing rule for computing more predictive mixture descriptors in this study.

With these objectives, we have developed a quantitative read-across structure-activity relationship model (q-RASAR) for the prediction of binary mixture toxicity in honey bees in this research. RASAR is the combination of read-across and QSAR, which is developed for the improvement of predictability from both read-across and QSAR approaches [11]. It incorporates the similarity and error-based parameters obtained from the read-across prediction tool with the pool of structural and physicochemical features of the dataset and then finalizes the combination of descriptors [12]. Ridge regression, linear and non-linear support vector machine algorithms are also used in this study to further enhance the predictive ability of the q-RASAR model. Apart from the modelling, the best mixing rule to calculate the mixture descriptors has been identified in this research. The calculated validation metrics indicate the goodness-of-fit, robustness, and good predictability of the model. Here, the toxicity values of several new untested binary mixtures (obtained from the real-world data recently published by Tosi et al. [10]) have also been predicted with the new model, and the reliability of the predictions has been checked by the prediction reliability indicator (PRI) tool. This computational research is in consonance with the 3Rs strategy (Replacement, Reduction, and Refinement of animal experimentation) of the REACH legislation [13].

2. Materials and methods

2.1. Toxicity data

For developing any *in silico* model, the availability of good quality response (activity/toxicity/property) data is of utmost importance. The number of data points is also a major concern while developing a statistically sound model. In 2019, Caenesecchi et al. reported several ecotoxicologically important mixtures with corresponding acute contact/oral toxicity data against honey bees in a meta-analysis [14]. The same research group further clarified the experimental data and reported in a separate modeling study in 2020 [9]. We have collected the toxicity data of 123 organic chemical mixtures (Table S1 in Supplementary Material 1) from the latter publication of Carnesecchi et al. [9] and have used it in our modeling. These mixtures were mostly composed of two (binary mixtures) or more (multicomponent mixtures on very few occurrences) plant protection products (pesticides, insecticides, acaricides, defoliants, plant growth retardants, etc.), that have been tabulated in Table S2 of Supplementary Material 1. Carnesecchi et al. [9] reported the acute contact toxicity data (pLD50 values) as μg active substance/bee unit (the endpoint of this research), which represents the concentration of the mixtures that is enough to produce a lethal effect on 50% of the total population of honey bees. There is no data on mass or concentration ratio in the source article; therefore, we have used a mass ratio 50:50 as the default value for binary mixtures as an approximation (0.33:0.67 mass ratio has been taken for mixtures SM116, SM119, and SM 122 where 1 component is present twice of the other) in this research. Among 123 data points, 24 different sets of duplicates (69 duplicate entries) have been found. Among these 24 sets, 11 sets (24 entries) have response values in a very close range (variance <0.02; standard deviation <0.15). Thus, the response means of these 11 sets of duplicate data have been taken as the observed response and used in modeling. Therefore, a total of 65 mixtures (initial 54 mixtures + 11 duplicate mixtures) have been employed as the modeling data set. The rest 13 sets (44 entries) of mixtures have response values with higher ranges (variance >0.05 and/or standard deviation > 0.2) and thus have been used as true external set mixtures. These true external set mixtures have been employed for true external validation in this research. The modeling set mixtures and true external set mixtures have been tabulated in Table S3 and Table S4 respectively of Supplementary Material 1.

2.2. Mixture descriptors

The calculation of mixture descriptors is not as simple as the de-

scriptors of individual chemicals. The theory of additivity is mostly used to calculate mixture descriptors where a linear combination of the molecular descriptors of mixture components weighted by their respective masses are employed to consider the effects of each component [15–18]. But the bioactive mixture components may interact with each other and develop synergistic or antagonistic effects. If we thoroughly review the scientific literature, then several mixing rules apart from the theory of additivity can be found which are used to consider the interactive effects of mixture components [15,19–22]. In this study, we have employed three different mixing rules namely the linear combination of molecular contributions, the squared molecular contributions, and the norm of molecular contributions, to optimize the descriptor calculation. The mathematical forms of the three mixing rules have been given in the following Eqs. 1 to 3 respectively [23]-

$$Mixture \quad Descriptor1 = \sum p_i \times x_i \tag{1}$$

$$Mixure \quad Descriptor2 = \left(\sum p_i \times x_i\right)^2 \tag{2}$$

$$Mixture \quad Descriptor3 = \sqrt{\sum (p_i \times x_i)^2} \tag{3}$$

where p_i represents the mass fraction i^{th} component, x_i represents the descriptor of i^{th} component.

Several classes of molecular descriptors can be obtained for modeling. The most common classification of descriptors is based on dimensionality; 0-dimensional to 7-dimensional descriptors can be computed among which 2D and 3D descriptors are mostly used. In the present study, up to 2D descriptors are employed for the ease of computation and interpretation of descriptors, and reproducibility and transferability of the developed models. 2D molecular structures, obtained from Marvin sketch software [24], have been used for descriptor calculation without any energy minimization and conformational analysis (prerequisites of 3D descriptor calculation). "AlvaDesc" software [25] has been utilized to calculate 9 types of 0D–2D descriptors (molecular properties, constitutional indices, ring descriptors, connectivity index, functional group count, 2D-atom pairs, atom type-E state indices, atom-centred fragments, and electro-topochemical atom indices) in this study. After obtaining the descriptors of individual components, the mixture descriptors have been computed with the above-mentioned Eqs. 1 – 3 by using our in-house tool "Mixture_Desc_Calc v1.0″.

2.3. Dataset division and feature selection for QSAR and subsequent modelling

Dataset division is one of the integral parts of any statistical modeling (especially QSARs) when the external data points are not available in sufficient numbers or are practically unavailable. In that case, the modeling set is divided into two parts namely training and test sets that are used respectively for model development and external validation. There is no fixed ratio of the number of data points in training and test sets; although there must be a justification for the number of data points selected in both sets. By keeping in mind the possibility of the bias-variance dilemma [26], the QSAR experts suggest a 70:30–80:20 division of training-test data points which have been used in many previous studies [16–18,27–29]. In this research, we have used 75% of the data for model development and internal validation (training set), and 25% of data for external validation (test set). A dataset can be divided either randomly or following some protocols based on the descriptor values (X-based division) or exerted response values (Y-based division) [30]. We have utilized both X-based (Euclidean distance-based division) and Y-based divisions (sorted activity-based division) for each class of mixture descriptors (Mixture Descriptor 1, Mixture Descriptor 2, and Mixture Descriptor 3) with the help of our in-house "DatasetDivision GUI 1.2″ tool. Eight types of divisions have been obtained (4 seeds of Euclidean distance-based divisions and 4 seeds of sorted activity-based divisions) from each mixture descriptor class, that has been used for the selection of the proper division set. Preliminary stepwise multiple linear regression (SMLR) models have been developed from all the training sets using MINITAB software with F to enter = 4 and F to remove = 3.9 criteria in multiple iteations. For a standard MLR equation, each descriptor should be significant in 95% confidence level ($p < 0.05$) which corresponds to an approximate t-value more than 2 ($F = t^2 = (2)^2 = 4$). Thus, we used F to enter = 4 and F to remove = 3.9. The best divisions from each class of mixture descriptor have been selected based on the proper balancing between results of both internal and external validation results of the preliminary SMLR models in terms of cross-validated correlation coefficient (Q^2_{LOO}) and external correlation coefficient (Q^2_{F1}). The quality metrics R^2, Q^2_{LOO}, and Q^2_{F1} of all preliminary models have been tabulated in Table S5 of Supplementary Material 1 as the indicators of goodness of fit, robustness, and external predictability respectively.

The results of preliminary SMLR models suggest the 4th seed among four sorted activity-based divisions from each descriptor class as the best division. Thus, we went ahead with the 4–4 sorted activity-based divisions (4–4 ABD) for feature selection. The feature selection was performed parallelly for each class of mixture descriptors. We have used all possible subset selection strategy for identifying the most important descriptors responsible for toxicity in honey bees (using a freely available in-house tool "BestSubsetSelectionModified_v2.1″). It is simply based on the multiple linear regression (MLR) algorithm and explores all possible combinations of the descriptors from a provided pool [30]. All possible subset selection is a computationally exhaustive process which needs greater time and computer proficiency than the other feature selection approaches. Therefore it is always accompanied by a descriptor pool reduction step. Here in this study, we have used Genetic algorithm (GA) for pool reduction. The GA is another efficient feature selection technique and was first used by Rogers and Hopfingers [31]. The simultaneous exploration of many possible solutions in the various regions of descriptor space makes it more powerful [32]. It is a two-step process where N number of descriptor population (parents) are selected from the whole pool first, and the fitness of each combination is determined. In the second step, crossover and mutation take place to get a new generation of descriptor combinations from the parents based on their scaled fitness scores. It is a process of genetic movement (movement of features) from parents to children; thus selection, crossover, and mutation are repeated until all the parent descriptors are replaced with the new ones. The fitness scores of the new combinations are again evaluated and the process continues until 80% of the new combinations attain the target score. We have performed GA using a freely available tool "GeneticAlgorithm_v4.1″ with the training set of mixtures.

Finally, three different combinations of six descriptors have been obtained from three types of mixture descriptors (linear combination of molecular contributions, squared molecular contributions, and norm of molecular contributions) by all possible subset selections. From these three sets of descriptors, we have developed three different QSAR models by partial least squares (PLS) regression. PLS is the generalization of multiple linear regression and is used to model strongly correlated, collinear, and noisy data with numerous descriptors [33]. It can also handle numerous Y variables (biological responses/molecular properties) simultaneously. The problem of overfitting with a large number of descriptors (close to the number of data points) in MLR can be successfully resolved using PLS [34]. In this method, latent variables (LVs) such as X scores (T) and Y scores (U) are calculated. Y scores are calculated from the X scores and response variables and then used for the final prediction. The LVs account for the maximum variation of the descriptor matrix to calculate the response. In this research, we have performed PLS using the freely available tool "PLS_Single Y_version 1.0″ and reported the resultant model with the optimum number of LVs (the number of LVs must be less than the number of descriptors) in Table S6 of Supplementary Material 1. These selected descriptors were taken

forward to the next step i.e., RASAR descriptor calculation, which has been discussed afterwards.

2.4. Read-across and RASAR descriptors calculation

The read-across structure-activity relationship (RASAR) approach combines two predictive techniques namely read-across and QSAR, for better predictability. The RASAR descriptors include similarity, error, and concordance measures obtained from the read-across prediction tool [12]. To develop RASAR models, these read-across-derived parameters may be clubbed together with significant structural and physicochemical descriptors that are correlated with the exerted molecular property in general. The RASAR descriptor calculation is accompanied by the read-across predictions and subsequent hyperparameter optimization. Read-across is a similarity-based prediction method where the properties/toxicities of new/unknown compounds are predicted from their structural analogues. These structural analogues are selected based on the structural similarity between query and source compounds, which can be obtained by several similarity estimation techniques. In one of our previous works, we have explained the similarity estimation by the Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity in detail along with the hyperparameter optimization (sigma: Gaussian kernel similarity estimation; gamma: Laplacian kernel similarity estimation, and the number of closed source analogues for optimum prediction: all the three similarity estimation techniques), which is necessary for obtaining the best quality of predictions [35]. In this study, different sigma values (0.25–2 with 0.25 interval), different gamma values (0.25–2 with 0.25 interval), and 2–10 no. of close source analogues have been used for optimizing these hyperparameters (based on the quality of prediction in terms of Q^2_{F1} value for a validation set). The hyperparameters have been optimized by employing an in-house tool "Auto RA Optimize v1.0″ and the optimized settings have been reported in Table S7 of Supplementary Material 1. Finally, the RASAR descriptors were calculated with the optimized hyperparameters by employing a freely available in-house tool "RASAR Descriptor Calculator v2.0″. One point should be noted here: the training set is used as both source and query mixture set for the calculation of RASAR descriptors of the training set; therefore, it will select the query compound itself as the most similar compound along with the others for each training data point. To avoid this issue, we have followed the "leave-same-out" strategy where the same compound (query compound present in the training set) is not taken into consideration based on the similarity and corresponding response value during RASAR descriptor calculation. However, the RASAR descriptors of the test set can be calculated conventionally by considering training and test set compounds as the source and query chemicals respectively.

2.5. q-RASAR modelling

Three different q-RASAR models have been developed in this research from three different classes of descriptors. To do so, we have performed another round of feature selection with all possible subset selection strategy. In each descriptor class, a pool of 21 combined descriptors (physicochemical, structural, and RASAR) has been used. All possible subset selection has been performed using the 'BestSubsetSelectionModified_v2.1′ tool. Although the models were developed solely based on training set information, we have shown the model with the highest external predictivity as evidenced from the external correlation coefficient values (Q^2_{F1}). Finally, three different PLS q-RASAR models have been developed with the selected descriptors by using the 'PLS_Single Y_version 1.0′ tool. From the quality of predictions of the developed PLS q-RASAR models, the most important mixing rule has been identified and reported as the final model of this study. To enhance the predictability of the final model, we have used machine learning (ML)-based linear and non-linear support vector machines (SVM) algorithms [36], and ridge regression [37]. We have used Python-based codes using the Jupyter Notebook web application [38] in the Anaconda Navigator version 2022.05 with Python version 3.10.4 for ML-based model development. The associated hyperparameters were optimized by the Grid search technique using 5-fold cross-validation mean squared error as the objective function in each of the ML approaches, and the best model was selected for the prediction of test set compounds for which the quality was judged in terms of Q^2_{F1} and $MAE_{(Test)}$ values. A detailed flow diagram of this study has been given in Fig. 1.

2.6. Validation protocols

The acceptance of any predictive model depends on proper validation. The Organization for Economic Cooperation and Development (OECD) proposed five different principles which are obeyed worldwide for the characterization of QSAR models for regulatory purposes [39]. Statistical validation metrics are extensively used to evaluate the quality and predictability of the developed QSAR models, based on which the models are accepted or rejected. The QSAR modellers must perform proper statistical validation during modelling to comply with the OECD guideline 4 [40]. Two types of metrics are calculated to comply with the guideline. The statistical quality and internal validation metrics such as determination coefficient (R^2), cross-validated (*leave-one-out*) determination coefficient (Q^2_{LOO}), mean absolute error (MAE), etc. are computed with the training data points [30]. R^2, Q^2_{LOO}, and MAE are the measures of goodness-of-fit, robustness, and errors of prediction respectively. There are high chances of overestimation of R^2 where the number of descriptors is close to the number of data points. To penalize the overestimated R^2, adjusted R^2 is calculated which considers the number of descriptors also along with the number of data points. The value of R^2, R^2_{adj}, and Q^2_{LOO} should be near 1, and a higher deviation from the ideal value suggests the inferiority of the model. An optimum value of $Q^2_{LOO} > 0.6$ is mostly used [30]. The error metric (MAE) should be near zero. The Akaike information criteria (AIC) values were also computed for comparing the q-RASAR models obtained from three different mixture descriptors in this study. This metric estimates the prediction error of the models developed from a single dataset and thereby judges the relative efficiency of the models. For a regression model, AIC is computed by employing the following Eq. (4) [41]:

$$AIC = n \times \log(\sigma^2) + 2 \times K \tag{4}$$

where, n = number of training data, K = number of descriptors (LVs for PLS), and σ^2 = mean sum of squared residuals. The least AIC value of a model is the indicator of higher relative efficiency among various models developed from a dataset.

To check the possibility of any chance correlation during model development, Y-randomization analysis [42] has been performed using the "SIMCA-P" software [43] and the results are reported. The response values of the training set are randomly shuffled (100 times in this study) without altering the descriptors and then refitted with the original model for computing the R^2 and Q^2 (computed metrics of randomized models) values. Finally, the computed R^2 and Q^2 values are plotted (Y-axis) against the correlation coefficient between the original Y values and the permuted Y values (X-axis). The intercept of R^2 and Q^2 (randomized) plots are recorded. Poor values of these intercepts (R^2 intercept < 0.3; Q^2 intercept < 0.05) are the indicators of a statistically significant model without any chance correlation [44]. To comply with the OECD principle 3 (a defined domain of applicability), an applicability domain analysis [45] has also been performed by the DModX approach using "SIMCA-P" software.

The external validation is performed to check the predictability of the developed model and the test set compounds are exclusively used for this purpose. External correlation coefficients such as Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}

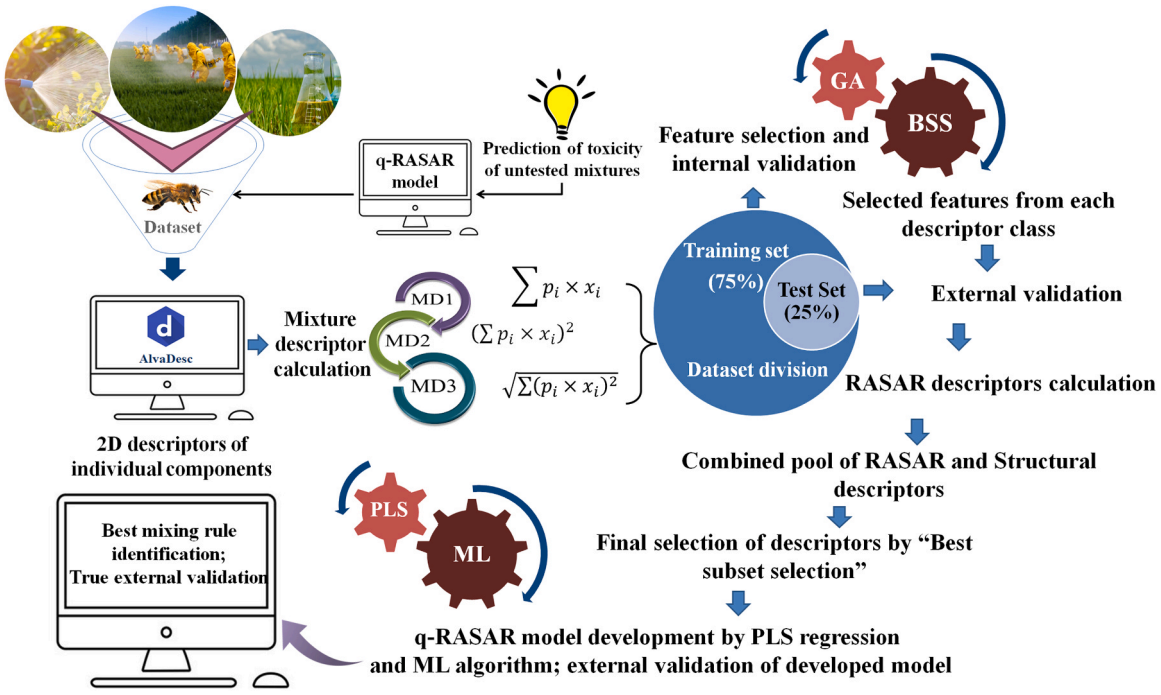


Fig. 1. Flow diagram of the q-RASAR modeling approach (*BSS: Best subset selection; GA: Genetic algorithm; ML: Machine learning; PLS: Partial least squares).

are the classic markers of predictability and their values should be near 1 [30]. In common practice, the optimum value of these three metrics for the selection of model should be > 0.6. Error metrics like mean absolute error (MAE_{Test}) is also calculated to check the error of predicted results. Besides these classical metrics, r^2_m metrics (*average* $r^2_{m(LOO)}$, $\Delta r^2_{m(LOO)}$, *average* $r^2_{m(Test)}$, $\Delta r^2_{m(Test)}$) are calculated for internal validation as well as external validation [46]. The optimum value of average $r^2_{m(LOO)}$ and average $r^2_{m(Test)}$ is > 0.5, whereas the expected value for $\Delta r^2_{m(LOO)}$ and $\Delta r^2_{m(Test)}$ should be < 0.2. A true external validation has also been performed in this study with 13 true external set mixtures, which are showcased in Table S4 of Supplementary Material 1. As discussed in the section "2.1. Toxicity data", these 13 sets of duplicate mixtures correspond to 44 data points and respective 44 observed toxicity values. The most predictive model (reported in the section "3.2. Best model and result of the validation") has been used for the prediction of toxicity of these 13 different true external set mixtures and classifying them as toxic or non-toxic based on a threshold value (response mean of the training set = −0.127 μg active substance/bee). Among the observed toxicity values, the most similar one among duplicate data points with the predicted response is selected for the calculation of classification-based metrics. The classification-based metrics have been calculated and reported subsequently.

2.7. Prediction of unknown mixtures with PLS and ML-based q-RASAR models and reliability analysis with PRI tool

To validate the performance of the new q-RASAR model against unknown mixtures (different from the modeling set), we have used a relevant and reliable dataset of 361 binary mixtures composed of organic chemicals used as plant protection products published by Tosi et al., [10]. After manual checking of these new binary mixtures, 143 data points have been excluded (reasons: same as modeling mixtures, duplicates, ternary mixtures, or metallic mixtures). The remaining 218 data points have been included in our analysis for toxicity prediction and further reliability analysis. These 218 binary mixtures are composed of 64 new components along with the previous 27 components (Table S2 in Supplementary Material 1) that are present in the modeling set. The unique mixture components and the composition of untested mixtures have been represented in Tables S8 and S9 respectively in Supplementary Material 1. The 2D structures of the new components have been retrieved from the PubChem database [47] and also cross-checked with the ChemSpider database [48]. The 2D descriptors of new components have been calculated by "AlvaDesc" software; then the mixture descriptors have been computed with the most relevant mixing rule (identified by the PLS q-RASAR model) discussed in the section "3.1. Developed q-RASAR models and best mixing rule identification". The RASAR descriptors of true external mixtures have been computed by the methodology discussed in the section "Read-across and RASAR descriptors calculation". Finally, the predicted response from PLS q-RASAR and different ML-based q-RASAR models (linear SVM, non-linear SVM, ridge regression) have been reported.

As an additive measure, the reliability of PLS q-RASAR predictions has been checked by an in-house java-based tool "Prediction reliability indicator" (PRI) [49]. This tool can categorize the reliability of the predicted values into 'Good', 'Moderate', and 'Unreliable', depending upon the composite score. PRI essentially uses three sets of principles for the reliability analysis of predicted responses and calculates the composite scores; these are- (1) mean absolute error of prediction (leave-one-out) of a query compound from 10 closest training compounds; (2) applicability domain in terms of similarity, by using standardization approach; (3) closeness of predicted values to the mean of the observed response of training set compounds.

3. Result and discussion

3.1. Developed q-RASAR models and best mixing rule identification

Three distinct q-RASAR models have been developed by partial least squares approach with the optimum number of latent variables (LVs) in this study from the mixture descriptors obtained by three different mixing rules. The optimum number of latent variables has been selected based on the X-variance during PLS regression and those variance values have been showcased in Table S10 of Supplementary Material 1. The developed models have also been represented in Table S10 of Supplementary Material 1. Each model has been rigorously validated following the OECD protocols. The computed internal and external validation

metrics along with the identified descriptors and the optimum number of latent variables have been shown in the following Table 1.

The computed validation metrics (as shown in Table 1) confirm the fitness, robustness, and predictive performance of the developed models. Q^2_{F1} value is an exclusive indicator of predictability as discussed in the section "2.6. Validation protocols". All three models have good predictability, but the model obtained from "Mixture descriptor 3″ is superior in terms of Q^2_{F1} value (0.91). Besides this, the least AIC value has been observed in the model obtained from "Mixture descriptor 3″ ($AIC_{Mix.\ Desc.\ 3} = -31.86$), which also supports the supremacy of the model. This superiority of predictability indicates the efficiency of mixing rule 3 (norm of molecular contributions) for calculating the mixture descriptors. Therefore, the 3rd model has been identified as the most predictive one for mixture toxicity and has been reported as the ultimate model of this study.

3.2. Best model and result of the validation

The best model is developed with the mixture descriptors obtained by mixing rule 3, as discussed in the previous section "3.1. Developed q-RASAR models and best mixing rule identification". This PLS q-RASAR model is composed of seven descriptors (two RASAR descriptors and five 2D structural descriptors) in the 4 latent variable space. The mathematical form of the final model has been shown in Eq. (5)-

$$pLD_{50} = -16.9846 - 1.252 \times B06[C-Cl] + 2.6563 \times B06[C-F] + 36.9044 \times Eta_betaS_A + 0.1754 \times H-046 + 0.9453 \times NssssC + 0.3549 \times g_m * Avg.Sim - 3.2626 \times g_m * SDSimilarity \quad (5)$$

Rigorous internal and external validation tests have been performed, and the computed validation metric values have been showcased in Table 1. The goodness-of-fit, robustness, and predictability of the model have been confirmed by the determination coefficient ($R^2 = 0.89$), cross-validated correlation coefficient ($Q^2_{LOO} = 0.84$), and external correlation coefficient ($Q^2_{F1} = 0.91$) values respectively. The mean absolute errors of both training and test sets are also on the lower side which signify the correctness of predictions obtained by the model ($MAE_{Train} = 0.31$, $MAE_{Test} = 0.32$). Besides these, acceptable values of additional r^2_m metrics ($\overline{r^2_m(Train)} = 0.78, \Delta r^2_m(Train) = 0.08, \overline{r^2_m(Test)} = 0.87, \Delta r^2_m(Test) = 0.06$) and CCC (0.95) indicate the statistical soundness of the model. We have also performed the Y-randomization test with 100 times randomization of response values by employing "SIMCA-P" software; the resultant plot has been given in the following Fig. 2**D**. The least R^2 and Q^2 intercept values (R^2 intercept = 0.0596; Q^2 intercept = −0.524) obtained by resultant plot discard the possibility of chance correlation. The observed vs. predicted toxicity scatter plot, variable importance plot, and score plot with 2 latent variable spaces have been given in Figs. 2**A**, 2**B**, and 2**C** respectively. A uniform scattering has been observed around the line passing through the origin of Cartesian coordinates (Fig. 2**A**) by plotting the observed and predicted toxicity values in the X and Y axis respectively, which suggests good fit and predictability of the model.

A true external validation has also been performed as discussed in "2.6. Validation protocols" with the true external set mixtures. Several classification-based metrics have been computed from the predicted responses obtained by the PLS q-RASAR model and reported in the following Table 2. Good sensitivity, specificity, and accuracy values indicate good classification ability of the model. The computed G-means value (0.89) indicates the balanced performance of the model between toxic and non-toxic classes. The computed values of Matthews correlation coefficient and Cohen's κ are 0.69 and 0.65 respectively which suggest an acceptable prediction and an agreement between observed and predicted classification.

To further enhance the predictability of the developed q-RASAR model, ML algorithms have been used in this study. Linear ML approaches such as ridge regression and linear SVM and non-linear ML approaches like non-linear SVM were employed to develop models using the identified descriptors by the previous best subset selection algorithm. After obtaining the predictions, we have computed the Q^2_{F1} value as an identifier of external predictability. The computed Q^2_{F1} values of linear SVM, ridge regression, and non-linear SVM were 0.913, 0.905, and 0.935 respectively. The predictability of non-linear SVM was slightly better than the other ML approaches and the previous PLS q-RASAR model (PLS q-RASAR: $Q^2_{F1} = 0.91$) as well. The mean absolute error of prediction of test set (MAE_{Test}) has also been computed for three ML methods. The least error has been observed in non-linear SVM (MAE_{Test}: non-linear SVM = 0.195, linear SVM = 0.256, ridge regression = 0.265). The computed validation metrics of these three ML methods were computed from the predicted responses and reported in the following Table 3.

3.3. Applicability domain analysis of the PLS q-RASAR model

The applicability domain of a statistical model is the theoretical region in the chemical space (defined by the modeled descriptors and

Table 1
Modeled descriptors, and computed validation metrics of three PLS q-RASAR models obtained from three different classes of mixture descriptors.

Validation metrics	Mixture descriptor 1	Mixture descriptor 2	Mixture descriptor 3
Identified descriptors	**Seven descriptors:** B05[C-O], D/Dtr03, NaaCH, nRNR2, D/Dtr05, F05[C-F], Abs MaxPos-MaxNeg	**Seven descriptors:** B06[O-Cl], D/Dtr03, Eta_betaS_A, mindsN, nCsp2, RDCHI, SD similarity(GK)	**Seven descriptors:** B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC, gm*Avg.Sim, gm*SD Similarity
No. LVs	6	3	4
R^2	0.90	0.88	0.89
Q^2_{LOO}	0.85	0.84	0.84
MAE_{Train}	0.31	0.33	0.31
$\overline{r^2_m(Train)}$	0.80	0.78	0.78
$\Delta r^2_m(Train)$	0.07	0.09	0.08
AIC	-29.72	-31.72	**-31.86**
Q^2_{F1}	0.89	0.88	**0.91**
Q^2_{F2}	0.88	0.88	**0.91**
Q^2_{F3}	0.87	0.87	**0.90**
MAE_{Test}	0.32	0.34	0.32
CCC	0.93	0.93	**0.95**
$\overline{r^2_m(Test)}$	0.78	0.85	0.87
$\Delta r^2_m(Test)$	0.08	0.06	0.06
MAE-based prediction quality	Moderate	Good	**Good**

LVs: Latent variables; MAE: mean absolute error; AIC: Akaike information criteria; CCC: concordance correlation coefficient.

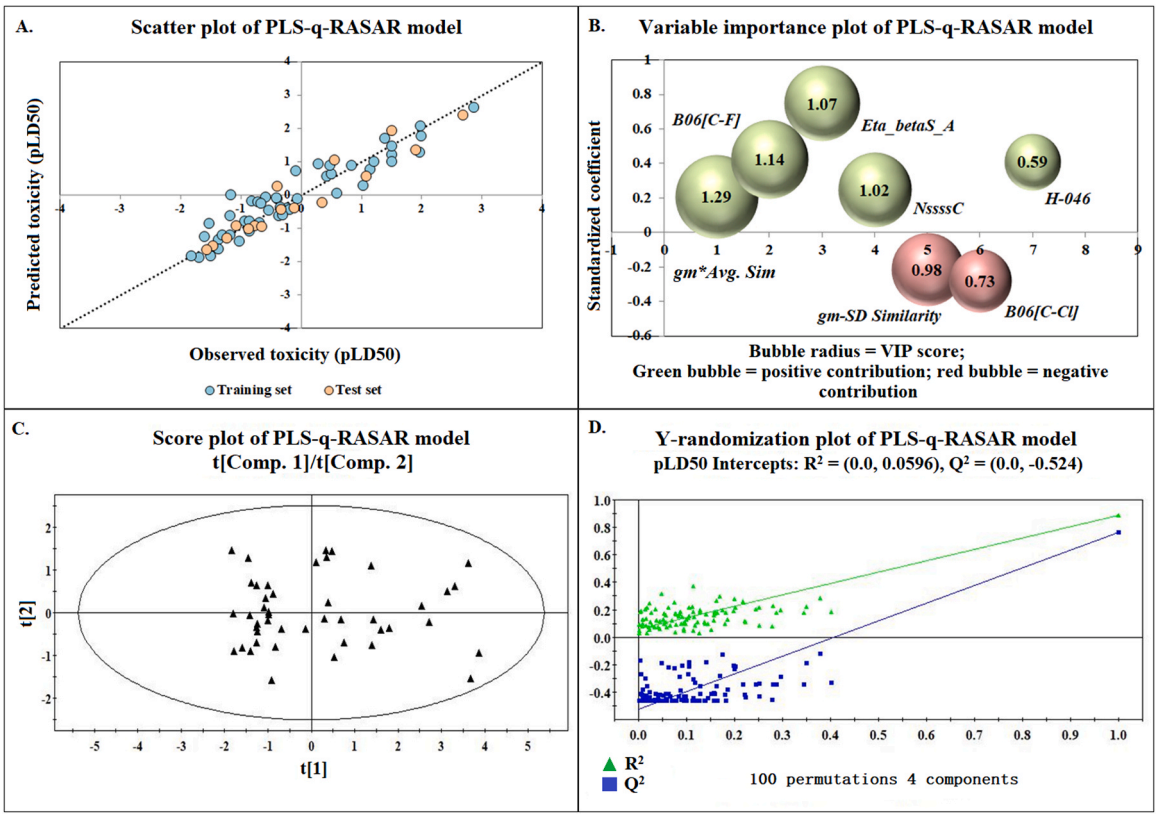


Fig. 2. A. Scatter plot, B. Variable importance plot, C. Score plot, and D. Y-randomization plot of the developed PLS q-RASAR model.

Table 2
Classification-based metrics* obtained from true external validation.

Classification-based metrics	Computed values
Sensitivity	1.00
Specificity	0.80
Accuracy	0.85
Precision	0.60
F-measure	0.75
G-means	0.89
MCC	0.69
Cohen's κ	0.65

MCC: Matthews correlation coefficient
*The threshold for the classification is the response mean of the training set (−0.127 μg active substance/bee)

Table 3
Computed validation metrics from the predictions obtained by the ML-based q-RASAR models.

Quality and validation metrics	Linear SVM	Ridge regression	Non-linear SVM
Internal validation metrics			
R^2_{Train}	0.874	0.888	**0.959**
Q^2_{LOO}	0.809	0.831	0.735
MAE_{Train}	0.251	0.260	**0.147**
External validation metrics			
R^2_{Test}	0.919	0.914	**0.947**
Q^2_{F1}	0.913	0.905	**0.935**
Q^2_{F2}	0.912	0.904	**0.934**
Q^2_{F3}	0.902	0.894	**0.927**
CCC	0.952	0.950	**0.962**
$\overline{r^2_m(Test)}$	0.884	0.883	0.836
$\Delta r^2_m(Test)$	0.061	0.055	0.051
MAE_{Test}	0.256	0.265	**0.195**
$RMSE_P$	0.310	0.323	**0.268**

$RMSE_P$: root mean squared error of predictions

response values) in which the predictions obtained by the model are reliable [45]. The prediction of toxicity/activity/property of all possible molecules in the universe with a single statistical model is practically impossible; the query compounds must have optimum structural similarity with the training compounds for obtaining a reliable prediction. Therefore, checking the applicability domain is a mandatory requirement of any statistical model, and it is also suggested by OECD principle 3 [40]. To comply with the OECD protocol, applicability domain analysis of the newly developed PLS q-RASAR model has been performed with the SIMCA-P software by employing the DModX approach with 99% confidence level. The DModX (distance to model in descriptor space) plots for both the training and test set mixtures have been showcased in Fig. 3. A 99% threshold line (as standard; D-Crit = 0.00999898) with respect to 4 latent variable space of the training data is considered as the D-critical line, and it indicates the applicability domain of the corresponding model beyond which the predictions are unreliable. In this study, all the training set mixtures (Fig. 3**A**) and test set mixtures (Fig. 3**B**) are inside the D-Critical line which indicates the reliability of predictions by the model. A score plot of the training set has also been constructed (Fig. 2**C**) in this study by using the SIMCA-P software which tells us about the outlier mixture in the 2 latent variable space. However, there is no mixture beyond the threshold line defined by the ellipse.

3.4. Insights on modeled descriptors

In the section "3.1. Developed q-RASAR models and best mixing rule identification", seven descriptors of class "Mixture Descriptors 3″ have been identified as the important features which affect the toxicity to a great extent; thus, B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC, g_m*Avg.Sim*, g_m*SD Similarity* are the descriptors of interest in this study.

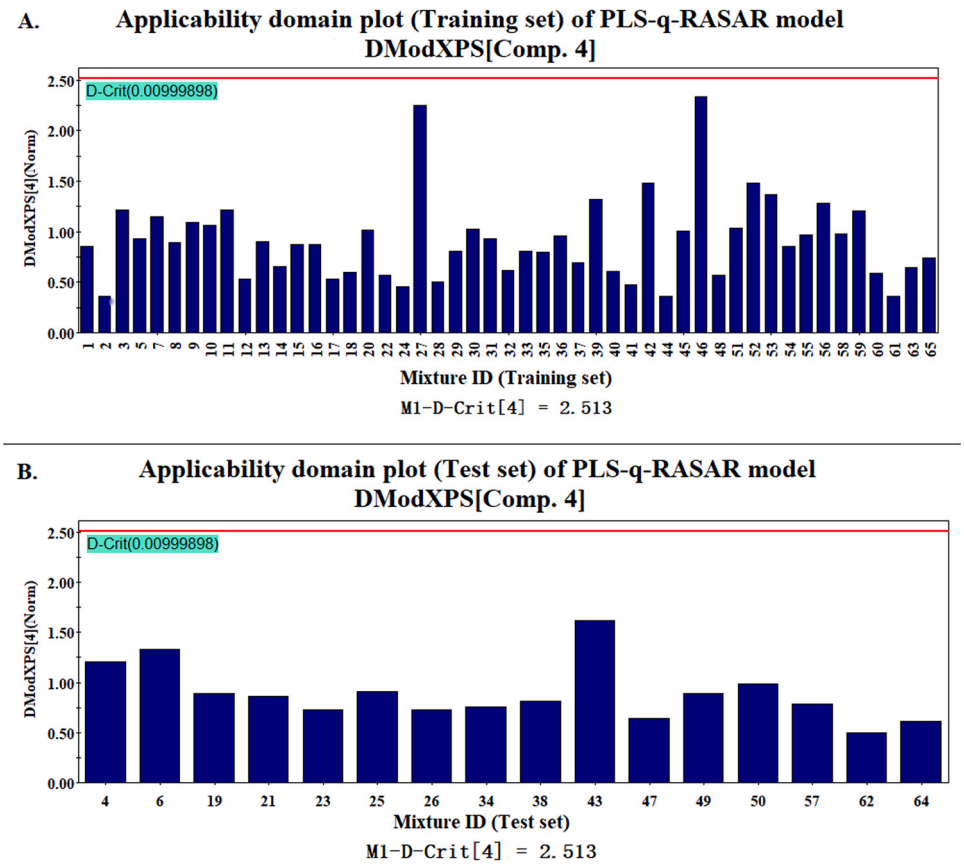


Fig. 3. Applicability domain (DModX) plots of the PLS q-RASAR model obtained by "SIMCA-P" software (A = Training set; B = Test set).

Among these seven descriptors, B06[C-Cl], B06[C-F], Eta_betaS_A, H-046, NssssC are the 2D-structural descriptors computed by "AlvaDesc" software whereas, g_m*Avg.Sim* and g_m*SD Similarity* are the RASAR descriptors computed by the "RASAR Descriptor Calculator v2.0″ tool. The algebraic sign (positive/negative) of a regression coefficient is a major diagnostic feature of the type of descriptor contribution. A positive regression coefficient generally suggests a proportional relationship between the descriptor and the endpoint (provided other contributing descriptor values remain the same), whereas a negative regression coefficient suggests an inversely proportional relationship (given that other contributing descriptor values remain the same). This gives a general idea about the contribution of the modeled descriptors toward the endpoint. However, with changes in the structural features, values of multiple descriptors (with positive and negative contributions) change simultaneously, and the observed endpoint value is a function of all (positive and negative) contributing features. Thus, to demonstrate the effect of a particular descriptor toward the endpoint, only representative and most suitable examples might be given which may not be universally applicable for all possible examples. The definitions, type, and the type of contribution towards toxicity (positive or negative) of modeled descriptors have been given in the following Table 4.

The relative importance of modeled descriptors towards toxicity has been checked with the SIMCA-P software and represented as a bubble plot in Fig. 2B. Each descriptor has been depicted as the bubbles with variable diameters where the diameter of the bubbles correspond to the VIP score of individual descriptors. According to Wold et al., descriptors with VIP scores greater than 1 represent higher importance towards toxicity prediction by the developed model [34]. Therefore, g_m*Avg.Sim*, B06[C-F], Eta_betaS_A, and NssssC affect the toxicity prediction of organic mixtures against honey bees to a greater extent in this newly developed q-RASAR model. In Fig. 2B, the relative importance of modeled descriptors has been plotted along X axis sequentially in descending order from left to right; therefore, the order of importance towards toxicity is g_m*Avg.Sim* > B06[C-F] > Eta_betaS_A > NssssC > g_m*SD Similarity* > B06[C-Cl] > H-046. Among these seven descriptors, B06[C-Cl] and g_m*SD Similarity* have negative contributions towards toxicity prediction; the rest are positively contributing descriptors. One can analyze the dataset mixtures for understanding the

Table 4
Definitions, types, and direction of the coefficient (positive or negative) of modeled descriptors.

Descriptors	Definition	Type	Type of contribution toward toxicity
B06[C-Cl]	Presence or absence of C-Cl atom pair at topological distance 6	2D-atom pairs descriptors	Negative (-)
B06[C-F]	Presence or absence of C-F atom pair at topological distance 6	2D-atom pairs descriptors	Positive (+)
Eta_betaS_A	Eta Sigma average VEM count	Electro topochemical atom index descriptors	Positive (+)
H-046	H attached to CO (sp^3) no X attached to next C	Atom-centred fragments	Positive (+)
NssssC	Number of atom/s of type ssssC	Atom-type E-state indices	Positive (+)
g_m*Avg.Sim*	g_m × average similarity of close source compounds where gm is a concordance measure	Concordance measure: RASAR descriptors	Positive (+)
g_m*SD Similarity*	g_m × standard deviation of close source compounds where g_m is a concordance measure	Concordance measure: RASAR descriptors	Negative (-)

type of contribution which further justifies the appropriateness of the model. Ideally, to know a more comprehensive contributing trend of a specific descriptor, a comparative analysis should be performed among dataset mixtures by keeping the values of other descriptors constant. But the prediction of the endpoint (here pLD50) is derived from all the modeled descriptors simultaneously and with changes in the structural features, different descriptor values change simultaneously; therefore, the other descriptor values cannot be kept constant to study the effect of a particular descriptor. Thus, the general relationship between the respective descriptor and endpoint identified by the model may not be followed simultaneously by all the mixtures in a dataset. Although the general trend can be identified by a comparative analysis of dataset mixtures which has been reported in Table S11 **of** Supplementary Material 1).

The g_m*Avg.Sim* is a RASAR descriptor representing the product of concordance measure g_m and the average of similarity values of close source mixtures against each query mixture [50]. A sharp decrease in descriptor values has been observed (Mixture **41**: g_m*Avg.Sim* = 0.687; Mixture **44**: g_m*Avg.Sim* = -0.967) from the most toxic mixture to the least toxic mixture in the dataset (Mixture **41**: pLD_{50} = 2.886; Mixture **44**: pLD_{50} = −1.806). There are several other examples in the dataset (See Table S11 in Supplementary Material 1) which justify the proportional relationship (positive contribution) of the most important descriptors of this study.

B06[C-F] is another descriptor of greater relative importance. It is a 2D atom pair descriptor that indicates the presence or absence of C-F atom pairs at the topological distance 6. The presence of the specified fragment has been depicted in the 2D structure of the mixture component cyhalothrin (**C7**) in Fig. 4. A proportional relationship between the B06[C-F] descriptor and toxicity has been observed in this study. There are several examples represented in Table S11 of Supplementary Material 1 which agree to this proportional relationship. It has also been observed that the least toxic mixture **44** does not have a non-zero descriptor value (mixture **44**: composed of thymol and diethyl but-2-enedioate; pLD_{50} = −1.806; B06[C-F] = 0) whereas the most toxic mixture **41** (mixture of cyhalothrin and piperonyl butoxide; pLD_{50} = 2.886; B06[C-F] = 0.5) has a positive descriptor value. This descriptor indicates a positive contribution as well as the presence of lipophilic F as a structural component. Therefore, lipophilicity may enhance the mixture toxicity to honeybee.

Eta_betaS_A ($\sum \beta_s^i$) is an electro-topochemical atom descriptor which is the next important identified structural feature of this study. Roy and Das defined this descriptor as "*summed β values for all the sigma bonds (VEM sigma contribution) relative to the number of vertices*" [51]. It represents the electron richness of any chemical relative to the molecular size. We have depicted the presence of electronegative atoms with

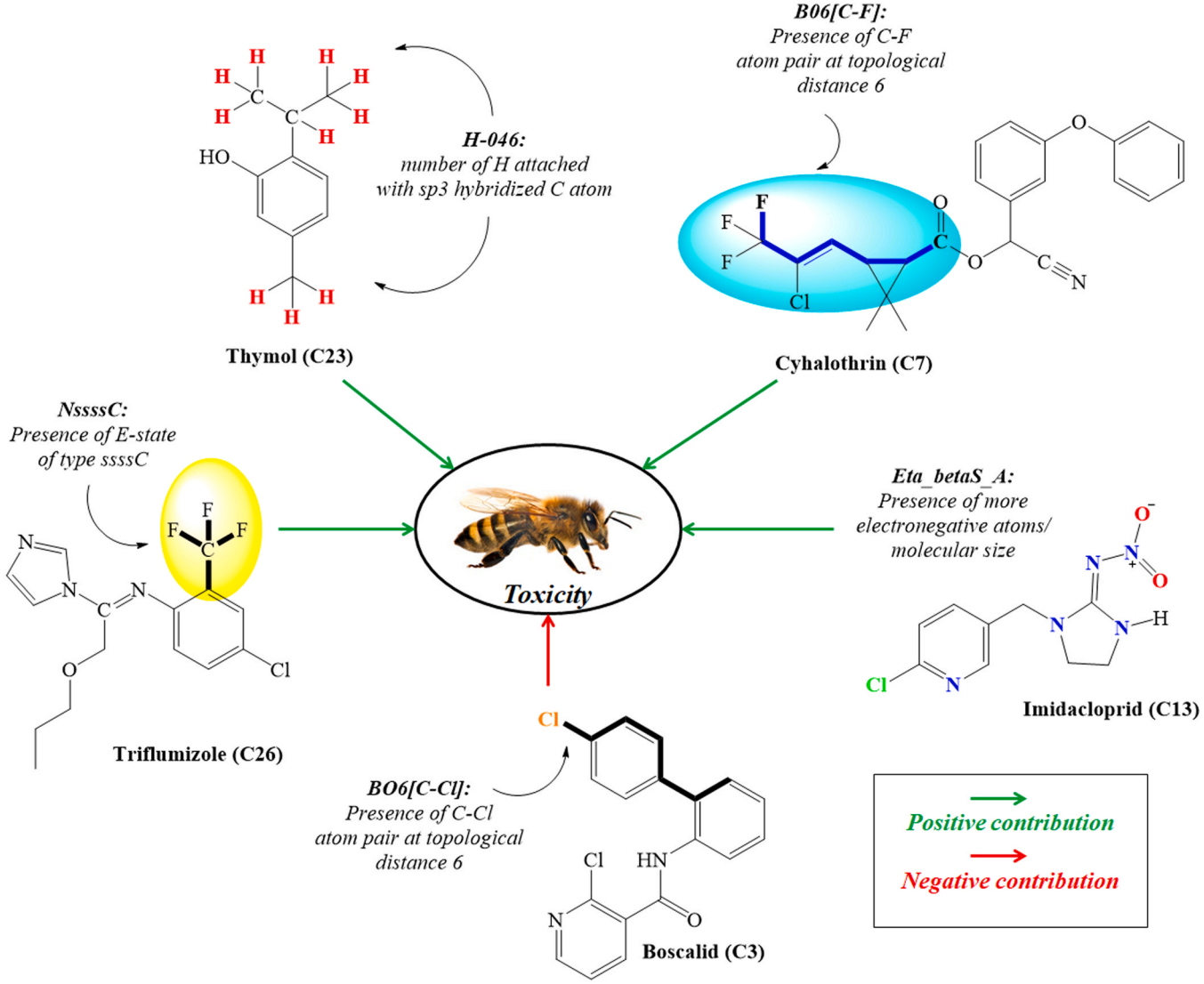


Fig. 4. Depiction of the presence of 2D-structural descriptor fragments in 2D structure of mixture components.

coloured symbols in the structure of a mixture component imidacloprid (**C13**) in Fig. 4. It has been observed that a mixture of chemicals with less electron richness (e.g., mixture **44**: composed with thymol and diethyl but-2-enedioate; $pLD_{50} = -1.806$; Eta_btaS_A = 0.392) has less toxicity than a mixture of chemicals with greater electron richness (e.g., mixture **56**: composed of imidacloprid and triflumizole; pLD_{50} =2.03; Eta_betaS_A = 0.479). We have presented several other pairs of mixtures (see Table S11 of Supplementary Material 1) which signify this proportional relationship. Therefore the presence of electronegative atoms in mixture components as identified by Eta_betaS_A descriptor positively contributes towards mixture toxicity against honey bees.

The next important descriptor is NssssC (an atom type E-state indices) that represents the number of atoms of type ssssC; where "s" indicates the attachment of single-bonded atoms apart from H. In mixture component triflumizole (**C26**), the ssssC type fragment (C-CF_3) has been pointed and has been depicted in Fig. 4. The descriptor also has a positive contribution towards toxicity as identified by the standardized coefficient plotted in Fig. 2**B**, and is evident from the comparative analysis between 3 pairs of studied mixtures represented in Table S11 of Supplementary Material 1. The most toxic mixture has shown a greater NssssC value (e.g., mixture **41**, composed of cyhalothrin and piperonyl butoxide; $pLD_{50} = 2.886$; NssssC = 1) than the least toxic mixture (e.g., mixture **44**, composed of thymol and diethyl but-2-enedioate; $pLD_{50} = -1.806$; NssssC = 0).

The *g_m*SD Similarity* is another RASAR descriptor with comparatively lower importance (than the other RASAR descriptor). It is the product of concordance measure g_m and the standard deviation of similarity values of close source mixtures against each query mixture of read-across prediction [50]. Unlike the other RASAR descriptor of this study (*g_m*Avg.Sim*), *g_m*SD Similarity* has a negative contribution towards toxicity prediction. This inverse relationship can be seen in several mixtures; we have represented 3 such mixture pairs in Table S11 of Supplementary Material 1. It has also been observed that the most toxic mixture of the dataset has a lower *g_m* SD Similarity* value (mixture **41**, composed of cyhalothrin and piperonyl butoxide; $pLD_{50} = 2.886$; *g_m* SD Similarity* = 0.094) than a comparatively less toxic mixture (mixture **40**, composed of cyhalothrin and tribufos; $pLD_{50} = 1.523$; *gm* SD Similarity* = 0.144).

B06[C-Cl] is a less important structural descriptor in this study which indicates the presence or absence of C-Cl atom pair at the topological distance 6. In Fig. 4, the specified fragment has been depicted in the structure of the mixture component boscalid (**C3**). Mixtures containing this fragment have higher mixture descriptor values and thus lower toxicity as observed in mixtures represented in Table S11 of Supplementary Material 1. There is also no C-Cl atom pair at the topological distance 6 in the structural components of the most toxic mixture in the dataset (mixture **41** which is composed of cyhalothrin and piperonyl butoxide), which again justifies the negative correlation between B06[C-Cl] and exerted toxicity.

Among the modeled descriptors, H-046 has the least importance. It is an atom-centred fragment descriptor that indicates the number of H attached to sp^3 hybridized carbon atoms with no attachment of heteroatom-connected adjacent carbon. Such 'H's have been marked in the structure of a mixture component thymol (**C23**) in Fig. 4. It possesses a positive coefficient, and thus has a positive contribution towards toxicity. A marked decrement in both H-046 and toxicity values has been observed in several mixtures noted in Table S11 of Supplementary Material 1 (e.g., mixture **41**, composed of cyhalothrin and piperonyl butoxide; H-046 =6.946; $pLD_{50} = 2.886$; and mixture **3**, composed of acetamiprid and diethyl but-2-enedioate; H-046 = 0, $pLD_{50} = -0.841$), which justify the proportional relationship between H-046 and toxicity (pLD_{50}).

3.5. Prediction of toxicity of unknown mixtures and reliability analysis

In this communication, the toxicity of 218 untested binary mixtures against honey bees was predicted by the developed PLS q-RASAR model and also with the ML-based q-RASAR models. We got four sets of predictions from PLS q-RASAR, linear SVM q-RASAR, ridge regression, and non-linear SVM q-RASAR models, respectively. The averages of these four sets of predictions have also been computed. All these predictions along with their averages have been reported in Table S12 of Supplementary Material 1.

The reliability of PLS q-RASAR predictions of these untested mixtures has been checked with the prediction reliability indicator (PRI) tool as discussed in the section "2.7. Prediction of unknown mixtures with PLS and ML-based q-RASAR models and reliability analysis with PRI tool". The PRI tool also provides information about the applicability domain of the untested mixtures/compounds along with the reliability of predictions. Both the reliability of predictions and applicability domain status of untested mixtures have been given in Table S12 of Supplementary Material 1. It has been observed that 91.28% of untested mixtures were inside the domain of applicability of the developed PLS q-RASAR model; therefore, the prediction of those 91.28% mixtures is reliable. The rest 8.72% of the mixtures were outside the applicability domain of the new model; thus unreliable predictions were obtained (Fig. 5). The reliability data obtained from the PRI tool follows the same pattern; "unreliable" predictions have been obtained from 8.72% of query mixtures. Predictions of 87.61% of total mixtures have been obtained with "Good" reliability score, whereas, predictions of 3.67% untested mixtures among total have shown "Moderate" reliability score (Fig. 5).

4. Comparative assessment of the quality of q-RASAR model with the previous work

To justify the efficiency of q-RASAR modeling in mixture toxicity predictions, we have compared the quality of the developed models in this study with a previously reported quasi-SMILES-based innovative QSAR model. Carnesecchi et al. reported two different approaches of toxicity predictions of binary mixtures using innovative QSAR modeling by considering the toxic unit data into account (Approach A: without considering the toxic unit; Approach B: considering the toxic unit) [9]. Both of these studies (current and previous) used the same mixture dataset against honey bees and thus, they are comparable in terms of goodness-of-fit, robustness, and predictability. In the following Table 5, the quality metrics (such as R^2, Q^2_{LOO}, Q^2_{F1}, and CCC) of RASAR and the previously reported models have been shown. Both the RASAR and previous QSAR models have comparable goodness-of-fit in terms of R^2 value (R^2 of SVM q-RASAR = 0.96; R^2 of the previous QSAR model-Approach B = 0.97); but, with respect to the robustness and external predictability, both PLS q-RASAR and non-linear SVM q-RASAR models outperform the previous QSAR model. The computed Q^2_{LOO} value of PLS q-RASAR model is 0.84 whereas, the Q^2_{CLB} value of the previous QSAR model (Approach B) is 0.81. The SVM q-RASAR model shows a greater Q^2_{F1} value (0.935) than the previously reported QSAR model (Approach B: R^2_{VLD}=0.89), which indicates better predictability. The PLS q-RASAR model also has better predictability (Q^2_{F1} = 0.91) than the previous models (both Approaches A and B). The RASAR models are more precise as indicated by the MAE_{Test} values in comparison to the previous QSAR model (MAE_{Test} of PLS q-RASAR and SVM q-RASAR are 0.32 and 0.23 respectively; MAE_{VLD} of approach A and approach B of previous QSAR models are 0.34 and 0.53 respectively). Besides the quality measures, the q-RASAR models have been developed following the OECD protocols and have also shown reliability during predictions of untested mixtures. The current study has showcased the good classifying ability of PLS q-RASAR model making it a more appropriate approach for the risk assessment of binary and multicomponent mixtures against honey bees.

We may add here that mixture toxicity can also be determined by experiment dependent models like Concentration Addition (CA). In such cases, experimental toxicity data of mixture components is one of the

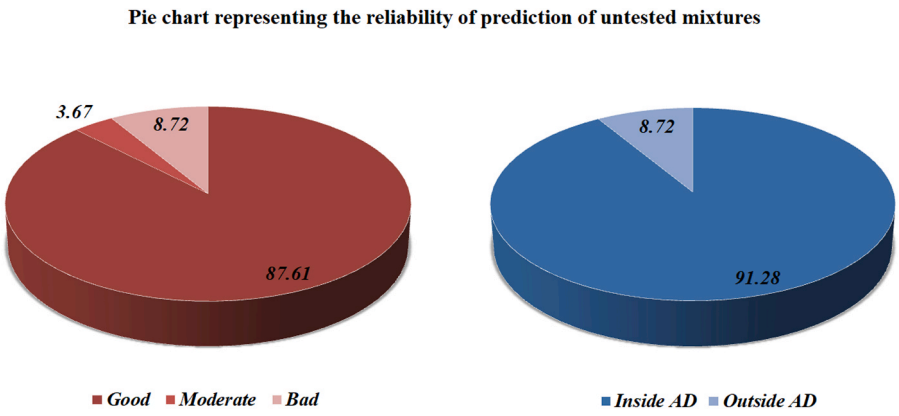


Fig. 5. Reliability analysis of the predictions of untested mixtures obtained by the PLS q-RASAR model (*All numbers represent the percentage (%) of predictions).

Table 5
Comparative assessment of quality metrics between q-RASAR and previously reported quasi-SMILES-based innovative QSAR models.

Model	Mix. Desc, Formula	R^2	Q^2_{LOO}	Q^2_{F1}	CCC	MAE_{Test}	No. of mixtures
PLS q-RASAR-	$Mix.Desc.3 = \sqrt{\sum (p_i \times x_i)^2}$	0.89	0.84	**0.91**	**0.95**	0.32	65 (Training = 49; Test = 16) Excluding duplicates.
Non-linear SVM q-RASAR	$Mix.Desc.3 = \sqrt{\sum (p_i \times x_i)^2}$	0.96	0.74	**0.935**	**0.96**	0.20	65 (Training = 49; Test = 16) Excluding duplicates
Quasi-SMILES-based QSAR model[9] **Approach A***	Quasi-SMILES fragments	TRN= 0.62 iTRN= 0.62 **CLB= 0.79**	TRN= 0.57 iTRN= 0.58 **CLB= 0.75**	**VLD= 0.87**	TRN= 0.76 iTRN= 0.75 **CLB= 0.89**	MAE_{VLD}= 0.34	123 (including duplicate entries)
Quasi-SMILES-based QSAR model[9] **Approach B* ***	Quasi-SMILES fragments	**TRN= 0.97**(Split 1); **0.94** (Split 4)	**CLB= 0.81** (Split 1); **0.84** (Split 4)	**VLD= 0.89** (Split 1); **0.78** (Split 4)	**CLB= 0.92** (both in Split 1 and Split 4)	MAE_{VLD}= 0.53 (Split 1); 0.51 (Split 4)	97 (Split 1: TRN=25 iTRN=24 CLB=24 VLD=24 Split 4: TRN=24 iTRN=24 CLB=24 VLD=25)

* Approach A: without considering the toxic unit data; * *Approach B: considering the toxic unit data; p_i = descriptor value of i^{th} component; x_i = mass ratio of i_{th} component; *LV*= latent variables; Mix. Desc. = mixture descriptors; TRN: training set; iTRN: invisible training set; CLB: calibration set; VLD: validation set.

necessary prerequisites. But QSARs or q-RASARs are the mathematical models developed from the existing experimental data and the structural properties of the mixture components. Therefore, one can predict the toxicity of the new/untested mixtures using the q-RASAR model without any additional experimentation. Besides this, the CA model always consider the theory of additivity to obtain the mixture effect; however, there are chances of synergism and antagonism due to interactive nature of chemicals. In this study, we have considered two other mixing rules apart from the simple addition of toxicity. Therefore, more precise predictions are expected from the newly developed q-RASAR model.

5. Limitations of the developed q-RASAR models

We have computed the mixture descriptors using the assumed mass ratio (50:50) and based on that, we have developed the new PLS/LSVM-based q-RASAR models. A very good correlation between experimental and predicted responses (along with a very low level of model error) has been observed which has been confirmed by the enhanced determination coefficient R^2 (0.89) value and low MAE_{Train} (0.31) value. This justifies the validity of our assumption. Different other validation metrics also suggest the goodness-of-fit, robustness, and good predictability of the developed PLS model. Therefore, we have used this model as the optimum one to predict the toxicity of untested pesticide mixtures to honey bees. The developed model should also be applicable for mixtures with different mass ratios as the descriptors will be computed accordingly reflecting the actual mass ratios. In our understanding, it is better to fill data gaps of mixture toxicity using a model with some approximations and assumptions than having no precise quantitative toxicity data for different untested mixtures.

6. Overview and conclusion

This research has explored the application of q-RASAR modeling approach for the prediction of acute contact toxicity of binary mixtures in honey bees for the first time. 2D-structural descriptors of the mixture components have been used as the preliminary input. The mixture descriptors have been calculated by employing three different mixing rules from the descriptors of individual mixture components and their corresponding mass ratio. Based on the computed quality metrics of three PLS q-RASAR models (obtained with three classes of mixture descriptors), the 3rd model has been identified as the most predictive one. Therefore, mixing rule 3, i.e., norm of the molecular contributions of mixture components has been selected as the best rule for calculating the mixture descriptors. The PLS q-RASAR model obtained with mixture descriptor 3 is a good fit with the training data, robust, and an extremely predictive one. Good classifiability of the model's predictions in toxic

and non-toxic classes has also been established by performing a classification-based external validation with a true-external dataset in this study. Besides these, linear and non-linear ML approaches have been used to enhance the predictability of the developed model. A non-linear SVM-based q-RASAR was the most predictive model of this study in terms of Q^2_{F1} value ($Q^2_{F1} = 0.935$). We have compared the performance of q-RASAR model with previously reported innovative QSAR, and the q-RASAR has been proven as a more robust and predictive approach for mixture toxicity assessment of organic mixtures in honey bees. Among the modeled descriptors, RASAR descriptors have a major contribution for enhancing the predictability of the model. On the other hand, the structural descriptors have effectively pointed out a specific feature i.e., the presence of electronegative atoms that increases the toxicity of organic mixtures. The descriptors namely B06[C-F], Eta_betaS_A, and NssssC mostly indicate the presence of electronegative atoms and all of them contribute positively. Furthermore, the PLS and ML-based q-RASAR models have been employed to predict the acute contact toxicity of 218 untested binary mixtures in honey bees and the predicted responses have been reported along with their consensus (average) prediction. Here we thus attempted to fill the major data gap on binary mixture toxicity [52]. In a nutshell, we proposed this approach as a new, efficient, reproducible, and easily transferable technique that can predict the acute contact toxicity of untested binary mixtures within the applicability domain with sufficient reliability. It can thus be used as a reliable integrative approach to the standard empirical approaches on pesticide mixture toxicity risk assessment.

CRediT authorship contribution statement

MC: Investigation, Formal analysis, Data curation, Writing – original draft. **AB:** Software, Validation, Writing – review & editing. **ST:** Resources, Writing – review & editing. **EC:** Resources, Writing – review & editing. **EB:** Resources, Writing – review & editing. **KR:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Acknowledgement

Mainak Chatterjee sincerely thanks the All India Council for Technical Education (AICTE), New Delhi for financial assistance in the form of National Doctoral Fellowship (2019–2023). Arkaprava Banerjee acknowledges the financial assistance of the Life Sciences Research Board, Defence Research & Development Organization (LSRB-DRDO), New Delhi.

List of software tools and websites used

"AlvaDesc" software (available from https://www.alvascience.com/alvadesc/, Date of access: 10–10-2022).

"Auto RA Optimizer" v1.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, Date of access: 20–10-2022).

"Best Subset Selection Modified" v2.1(available from: http://teqip.jdvu.ac.in/QSAR_Tools/, Date of access: 15–10-2022).

"Dataset Division GUI"v1.2 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/, Date of access: 10–10-2022).

"Genetic Algorithm" v4.1 (available from: http://teqip.jdvu.ac.in/QSAR_Tools/, Date of access: 15–10-2022).

"Minitab" v14 (available from: https://www.minitab.com/en-us/, Date of access: 10–10-2022).

"Mixture_Desc_Calc" v1.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, Date of access: 10–10-2022).

PLS_Single Y" v1.0 (available from: http://teqip.jdvu.ac.in/QSAR_-Tools/, Date of access: 15–10-2022).

"Prediction Reliability Indicator" (available from: http://teqip.jdvu.ac.in/QSAR_Tools/, Date of access: 25–10-2022).

"RASAR Descriptor Calculator" v2.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, Date of access: 20–10-2022).

"SIMCA-P" (available from: https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca, Date of access: 30–10-2022).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.132358.

References

[1] Klein, A.M., Vaissière, B.E., Cane, J.H., Steffan-Dewenter, I., Cunningham, S.A., Kremen, C., et al., 2006. Importance of pollinators in changing landscapes for world crops. Proc R Soc B Biol Sci 274, 303–313. https://doi.org/10.1098/RSPB.2006.3721.

[2] M. Sakalian, The UN environment programme: Why bees are essential to people and planet, (2022). https://www.unep.org/news-and-stories/story/why-bees-are-essential-people-and-planet (accessed November 18, 2022).

[3] FAO-UN, Food and Agriculture Organization of the United Nations. World Bee Day, 20 May, (2022). https://www.fao.org/world-bee-day/en/ (accessed November 18, 2022).

[4] E.S. Brondizio, J. Settele, S. Díaz, H.T. Ngo, Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services Bonn, Ger 2019 doi: 10.5281/zenodo.3831673.

[5] I. Utkina, Food and Agriculture Organization of the United Nations: Declining bee populations pose threat to global food security and nutrition, (2022). ⟨http://www.fao.org/documents/card/en/c/cc2323en⟩ (accessed November 18, 2022).

[6] Tosi, S., Costa, C., Vesco, U., Quaglia, G., Guido, G., 2018. A 3-year survey of Italian honey bee-collected pollen reveals widespread contamination by agricultural pesticides. Sci Total Environ 615, 208–218. https://doi.org/10.1016/J.SCITOTENV.2017.09.226.

[7] Backhaus, T., Faust, M., 2012. Predictive environmental risk assessment of chemical mixtures: a conceptual framework. Environ Sci Technol 46, 2564–2573. https://doi.org/10.1021/ES2034125.

[8] Thomas, P.C., Bicherel, P., Bauer, F.J., 2019. How in silico and QSAR approaches can increase confidence in environmental hazard and risk assessment. Integr Environ Assess Manag 15, 40–50. https://doi.org/10.1002/IEAM.4108.

[9] Carnesecchi, E., Toropov, A.A., Toropova, A.P., Kramer, N., Svendsen, C., Lou Dorne, J., et al., 2020. Predicting acute contact toxicity of organic binary mixtures in honey bees (A. mellifera) through innovative QSAR models. Sci Total Environ 704, 135302. https://doi.org/10.1016/j.scitotenv.2019.135302.

[10] Tosi, S., Sfeir, C., Carnesecchi, E., vanEngelsdorp, D., Chauzat, M.P., 2022. Lethal, sublethal, and combined effects of pesticides on bees: a meta-analysis and new risk assessment tools. Sci Total Environ 844, 156857. https://doi.org/10.1016/J.SCITOTENV.2022.156857.

[11] 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.

[12] 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.

[13] Ritskes-Hoitinga, M., Luijk, J.Van, 2019. How can systematic reviews teach us more about the implementation of the 3Rs and animal welfare. Animals 9, 1163. https://doi.org/10.3390/ANI9121163.

[14] Carnesecchi, E., Svendsen, C., Lasagni, S., Grech, A., Quignot, N., Amzal, B., et al., 2019. Investigating combined toxicity of binary mixtures in bees: meta-analysis of laboratory tests, modelling, mechanistic basis and implications for risk assessment. Environ Int 133, 105256. https://doi.org/10.1016/j.envint.2019.105256.

[15] Qin, L.T., Chen, Y.H., Zhang, X., Mo, L.Y., Zeng, H.H., Liang, Y.P., 2018. QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. Chemosphere 198, 122–129. https://doi.org/10.1016/J.CHEMOSPHERE.2018.01.142.

[16] 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.
[17] Chatterjee, M., 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.
[18] 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.
[19] Altenburger, R., Nendza, M., Schüürmann, G., Schüu, S., Schüürmann, S., 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ Toxicol Chem 22, 1900–1915. https://doi.org/10.1897/01-386.
[20] Sobati, M.A., Abooali, D., Maghbooli, B., Najafi, H., 2016. A new structure-based model for estimation of true critical volume of multi-component mixtures. Chemom Intell Lab Syst 155, 109–119. https://doi.org/10.1016/J.CHEMOLAB.2016.04.007.
[21] Gaudin, T., Rotureau, P., Fayet, G., 2015. Mixture descriptors toward the development of quantitative structure-property relationship models for the flash points of organic mixtures. Ind Eng Chem Res 54, 6596–6604. https://doi.org/10.1021/ACS.IECR.5B01457.
[22] Chatterjee, M., Roy, K., 2022. Computational modeling of mixture toxicity. In: Emilio Benfenati (Ed.), Silico Methods Predict. Drug Toxic. Humana Press Inc., New York, pp. 561–587. https://doi.org/10.1007/978-1-0716-1960-5_22.
[23] Saldana, D.A., Starck, L., Mougin, P., Rousseau, B., Creton, B., 2013. Prediction of flash points for fuel mixtures using machine learning and a novel equation. Energy Fuels 27, 3811–3820. https://doi.org/10.1021/EF4005362.
[24] ChemAxon, Marvin, (2018). https://chemaxon.com/products/marvin (accessed October 10, 2022).
[25] Mauri, A., 2020. alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. In: Roy, K. (Ed.), Methods Pharmacol. Toxicol. Humana Press Inc., New York, 801–820. doi:10.1007/978-1-0716-0150-1_32.
[26] Z.A. Almaliki, Bias-variance dilemma?, Towar. Data Sci. (2018). https://towardsdatascience.com/bias-variance-dilemma-74e5f1f52b12 (accessed October 19, 2022).
[27] Wang, T., Tang, L., Luan, F., Nat, M., Cordeiro, D.S., 2018. Prediction of the toxicity of binary mixtures by QSAR approach using the hypothetical descriptors. Int J Mol Sci 19, 3423. https://doi.org/10.3390/ijms19113423.
[28] Hoover, G., Kar, S., Guffey, S., Leszczynski, J., Sepúlveda, M.S., 2019. In vitro and in silico modeling of per fluoroalkyl substances mixture toxicity in an amphibian fi broblast cell line. Chemosphere 233, 25–33. https://doi.org/10.1016/j.chemosphere.2019.05.065.
[29] Chatterjee, M., Roy, K., 2022. Recent advances on modelling the toxicity of environmental pollutants for risk assessment: from single pollutants to mixtures. Curr Pollut Rep 8, 81–97. https://doi.org/10.1007/s40726-022-00219-6.
[30] Roy, K., Kar, S., Das, R.N., 2015. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press (Elsevier). https://doi.org/10.1016/C2014-0-00286-9.
[31] 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. https://doi.org/10.1021/ci00020a020.
[32] Leardi, R., 1996. Genetic algorithms in feature selection. In: Devillers, J. (Ed.), Genet. Algorithms Mol. Model. Academic Press (Elsevier), Great Britain, pp. 67–86. https://doi.org/10.1016/B978-012213810-2/50004-9.
[33] Cramer, R.D., 1993. Partial least squares (PLS): its strength and limitations. Perspect Drug Discov Des 1, 169–178. https://doi.org/10.1007/bf02174528.
[34] 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.
[35] 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.
[36] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell Syst Their Appl 13, 18–21. https://doi.org/10.1109/5254.708428.
[37] Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634.
[38] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., et al., 2016. Jupyter notebooks - a publishing format for reproducible computational workflows. In: Loizides, F., Scmidt, B. (Eds.), Position. Power Acad. Publ. Play. Agents Agendas. IOS Press, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.
[39] OECD, Guidance documents on the validation of (Quantitative) structure-activity relationships [(Q)SAR] models, OECD Environ. Heal. Saf. Publ. (2007). https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2 (accessed December 22, 2022).
[40] OECD, Validation of (Q)SAR Models - OECD, (2004). https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed November 30, 2022).
[41] Burnham, K.P., Anderson, D.A., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second ed. Springer-Verlag, New York. https://doi.org/10.1007/b97636.
[42] Topliss, J.G., Edwards, R.P., 1979. Chance factors in studies of quantitative structure-activity relationships. J Med Chem 22, 1238–1244. https://doi.org/10.1021/JM00196A017.
[43] 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.), Handb. Partial Least Squares. Springer, Berlin, Heidelberg, pp. 757–774. https://doi.org/10.1007/978-3-540-32827-8_33.
[44] Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in QSPR/QSAR. J Chem Inf Model 47, 2345–2357. https://doi.org/10.1021/ci700157b.
[45] Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct Relatsh 1, 45–63. https://doi.org/10.4018/IJQSPR.2016010102.
[46] 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. https://doi.org/10.1016/j.chemolab.2011.03.011.
[47] NCBI, PubChem, (2022). https://pubchem.ncbi.nlm.nih.gov/ (accessed November 10, 2022).
[48] RSC, ChemSpider, Search and share chemistry, (2022). http://www.chemspider.com/ (accessed November 10, 2022).
[49] K. Roy, P. Ambure, S. Kar, Prediction Reliability Indicator: A new tool to judge the quality of predictions from QSAR models for new query compounds, in: MOL2NET, Int. Conf. Ser. Multidiscip. Sci., 2018. doi:10.3390/mol2net-04–05265.
[50] Banerjee, A., Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. Chem Res Toxicol 36, 446–464. https://doi.org/10.1021/acs.chemrestox.2c00374.
[51] K. Roy, R.N. Das, The “ETA” indices in QSAR/QSPR/QSTR research, in: K. Roy (Ed.), Quant. Struct. Relationships Drug Des. Predict. Toxicol. Risk Assess., IGI Global, Hershey PA, USA, 2015: pp. 48–83. doi:10.4018/978–1-4666–8136-1.ch015.
[52] Traynor, K.S., Tosi, S., Rennich, K., Steinhauer, N., Forsgren, E., Rose, R., et al., 2021. Pesticides in honey bee colonies: Establishing a baseline for real world exposure over seven years in the USA. Environmental Pollution 279, 116566. https://doi.org/10.1016/j.envpol.2021.116566.

Journal of Hazardous Materials 459 (2023) 132129

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

"Data fusion" quantitative read-across structure-activity-activity relationships (q-RASAARs) for the prediction of toxicities of binary and ternary antibiotic mixtures toward three bacterial species

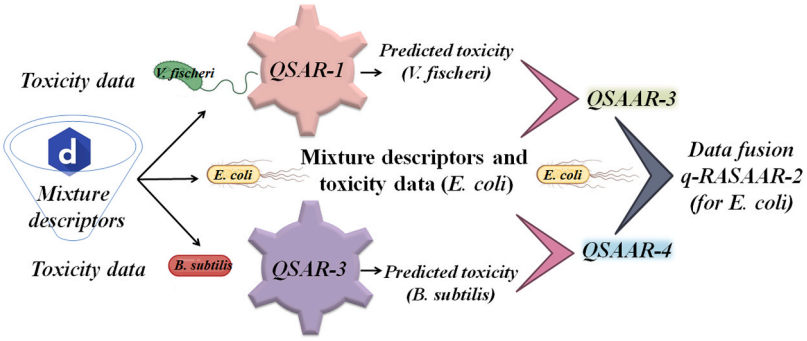
Mainak Chatterjee, Kunal Roy *

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

HIGHLIGHTS

- Simple 2D-QSAR models have been developed for the mixture toxicity against *V. fischeri, E. coli,* and *B. subtilis*.
- Three mixing rules have been applied to compute the optimum mixture descriptors.
- The q-RASAR algorithm is found as an efficient strategy to develop interspecies QSAAR (q-RASAAR) models.
- Mixture toxicity against *E. coli* is additive, whereas that against *V. fischeri* and *B. subtilis* involves interaction terms.
- The mixture toxicity patterns against *V. fischeri* and *B. subtilis* show similarities unlike the *E. coli* toxicity.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Debora Rodrigues

Keywords:
QSAR
QSAAR
Antibiotics
Mixture toxicity
Q-RASAAR

ABSTRACT

Antibiotics are often found in the environment as pollutants. They are usually found as mixtures in the environment and may produce toxicity against different ecological species due to joint exposure in the sub-optimal range. Sometimes the degradation products of parent chemicals also interact with it and cause mixture toxicity. In this study, we have developed three different mixture-Quantitative Structure-Activity Relationship (mixture-QSAR) models for three different bacterial species (*Vibrio fischeri, Escherichia coli,* and *Bacillus subtilis*). The toxicity data were collected from a previous experimental report in the literature, which comprised binary and ternary mixtures of sulfonamides (SAs), sulfonamide potentiators (SAPs), and tetracyclines (TCs). We have also explored the interspecies modeling to find inter-correlation among the toxicity of these studied organisms and have developed quantitative structure activity-activity relationship (QSAAR) models by employing the "data fusion" quantitative read-across structure-activity–activity relationship (q-RASAAR) and partial least squares (PLS) regression algorithms. All the models are strictly validated using both internal and external validation tests as suggested in the OECD guidelines. Three different mixing rules have been used in this study for descriptor computations to incorporate the additive and interaction effects among the mixture components. To the best of our knowledge, this is the first report of interspecies mixture toxicity models which can predict the cellular toxicity of binary and ternary mixtures against any of the three above-mentioned organisms.

* Corresponding author.
E-mail addresses: kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).

https://doi.org/10.1016/j.jhazmat.2023.132129
Received 10 May 2023; Received in revised form 28 June 2023; Accepted 21 July 2023
Available online 23 July 2023
0304-3894/© 2023 Elsevier B.V. All rights reserved.

1. Introduction

The existence of antibiotics and other pharmaceuticals in various environmental compartments is a proven truth [1,2]. Antibiotics are one of the most used pharmaceuticals in diverse fields and are discharged through different sewage systems such as hospitals, cattle fields, aquacultures, etc. [3,4]; thus, these are becoming one of the major pollutants. Due to their bioactive nature, antibiotics inevitably impact the lower environmental organisms (mostly aquatic and soil species) [5]. They mostly target the microorganism, especially the bacteria in the environment. Moreover, the bacterial species come in contact with multiple chemicals (here antibiotics) simultaneously, which may further potentiates the toxic effects [6,7]. In some previous research articles, deviations of toxicity from the standard were reported due to significant interactions (additive response, synergism, antagonism) among antibiotics [8,9]. Besides this, the degradation products also play an essential role in mixture toxicity as reported by Wang and co-workers [10]. Therefore, a mixture toxicity assessment of antibiotics is a significant and necessary task which should be carried out for a proper risk assessment of antibiotics against different bacterial species; however, the majority of current research is based on a single bacterium.

Mixture toxicity assessment is not as simple as individual pollutants due to several reasons. i.e., no fixed number of combinations, no fixed composition of mixtures, no fixed physical state of component chemicals, chances of interactions among mixture components, no fixed end point of interest, etc.[11]. Experimental toxicity tests are costly, time-consuming, and hazardous procedures which are often associated with the unsustainable use of biological organisms and animals; on the contrary, computational methods are cheap, fast, and greener approaches that may be used for the toxicity prediction of both mixtures and individual chemicals without compromising efficiency [12]. The quantitative structure-activity/property/toxicity relationships (QSAR/QSPR/QSTR) are one of the major in silico predictive approaches that are being used for a long time. On top of that, they are approved by several international agencies like the US Environmental Protection Agency (USEPA), European Chemical Agency (ECHA), European Food Safety Authority (EFSA), World Health Organization (WHO), Organization of Economic Cooperation and Development (OECD), etc. [13]. We would use the more general term "QSAR" in rest of this document to include QSTR also. In this context, interspecies toxicity models or quantitative structure-activity-activity relationships (QSAAR) are emerging as important applications of QSARs. QSAAR is a specific form of predictive approach in which the toxicity of a specific organism is used with several other additional structural and physicochemical descriptors as predictors for developing the toxicity model of other organisms [8,14]. It has been observed in several previously published reports that the interspecies toxicities are often correlated with each other and that is why the toxicity of certain species can be an important predictor for other species [15,16]. One of the major advantages of QSAAR over QSAR is the use of experimental toxicity as a predictor variable which incorporates the biological perspective into account to some extent during the toxicity modeling of other endpoints. Thus, it helps to develop better quality models with improved external predictability than conventional QSAR/QSPR/QSTRs.

In the current study, we have developed three QSAR models from three different mixture datasets against three common bacterial endpoints i.e., *Vibrio fischeri*, *Escherichia coli*, and *Bacillus subtilis*. The data comprised binary and ternary mixtures of commonly used antibiotics such as sulfonamides (SAs), sulfonamide potentiators (SAPs), and tetracyclines (TCs). Simple ^{0}D–^{2}D descriptors have been employed in this study to model the mixture toxicity for avoiding more complex conformational analysis and energy minimization (important prerequisites for 3D and higher-order descriptor calculations) without compromising the predictability of developed models. For mixture descriptor calculation, three different mixing rules (linear combination of molecular contributions, the squared molecular contributions, and norm of molecular contributions) [11] have been used and the best one has been identified. We have also developed three interspecies QSAAR models using the predicted responses as X-variable with additional descriptors. The quantitative Read-across Structure-Activity Relationship (q-RASAR) [17] and partial least squares (PLS) [18] algorithms have been employed for developing the QSAARs. The RASAR is a thoughtful merging of read-across (RA) predictions and QSAR for using the non-linearly derived prediction functions obtained by the read-across algorithm in a linear regression framework of QSAR [17]. There are several instances in the literature where RASAR has been applied and proved to be effective for obtaining better external predictability than both QSAR and read-across predictions [19,17,20]. In mixture toxicity predictions, RASAR has not yet been used much; hence, we have incorporated the RASAR algorithm for developing more efficient interspecies mixture-QSAARs in this work, and the final models may thus more appropriately be termed as quantitative Read-across Structure-Activity-Acivity Relationships (q-RASAARs). In these models, the predicted toxicity to other bacterial species are used as descriptor(s) after an appropriate "data fusion". The important structural features, responsible for mixture toxicity against each organism have been identified and discussed here in this study. We have also emphasized the consideration of intercorrelated toxic endpoints for better toxicity predictions from the insights of developed QSAAR models.

In the present study, the toxicity data was obtained by simple assay where the concentration of sample broth is the determinant of lethality. Thus, the exact toxicity mechanism is uncertain and thus, the use of specific receptor –based descriptors is inappropriate. Apart from that, the targets for the toxicity may vary depending on the chemical classes and the organism being considered; therefore, the use of a fixed target receptor or protein for docking-based descriptor calculation potentially limits the efficiency of predictions. To avoid these issues, ^{0}D–^{2}D descriptors have been used in this study. The predicted toxicity values of other bacteria have also been used as descriptors for improving the predictive performance further. To the best of our knowledge, this is the first attempt where interspecies mixture toxicity models have been developed for the prediction of cellular toxicity of binary and ternary mixtures of SAs, SAPs, and TCs towards any of the three above-mentioned organisms. In a nutshell, this study helps the readers to understand the basics of mixture toxicity predictions, the use of predicted toxicity of certain species for improving the predictability, and the importance of q-RASAAR for developing statistically sound models for mixture toxicity predictions.

2. Materials and methods

2.1. Mixture toxicity data collection and dataset preparation

The toxicity data for individual antibiotics (SAs, SAPs, TCs) and their binary and ternary mixtures were collected against *Vibrio fischeri*, *Escherichia coli*, and *Bacillus subtilis* separately from the published work by Wang et al. [21]. We compiled all the data and represented them into three different mixture datasets (**MD1**, **MD2**, and **MD3**). The MD1 contains the toxicity data of 51 binary and 25 ternary mixtures against *V. fischeri*, the MD2 contains the toxicity data of 43 binary and 24 ternary mixtures against *E. coli*, and the MD3 contains the toxicity data of 29 binary and 14 ternary mixtures against *B. subtilis*. All the mixtures from these three mixture datasets (MD1, MD2, and MD3) were composed of different compositions of 8 SAs, 4 SAPs, and 4 TCs which have been showcased along with their 2D structures and individual toxicity values (pEC50 (Mol/L)) against the studied bacteria (*V. fischeri*, *E. coli*, *B. subtilis*) in Table S1 of Supplementary Material 1. All the mixtures were composed of an equitoxic quantity of the component antibiotics (amount required to produce one toxic unit for individual components) based on their individual toxicity towards the corresponding bacterium. The mixtures of MD1, MD2, and MD3 were tabulated along with the composition and observed toxicity values in

Table S2, S3, and **S4** respectively of Supplementary Material 1.

2.2. Mixture descriptor calculation considering three different mixing rules

In this study, 9 classes of ^{0}D–^{2}D descriptors (total count = 2400) such as constitutional indices, ring descriptors, connectivity index, electro-topochemical atom indices, functional group counts, atom-centred fragments, atom-type E-state indices, 2D atom pairs, and molecular properties were used. Molecular descriptors of individual mixture components (presented in Table S1 in Supplementary Material 1) were computed first using "AlvaDesc" software [22] from the properly aromatized and explicit H-containing two-dimensional structure inputs (in the *.sdf format). All the 2D structures were drawn using "Marvin Sketch" software [23] without energy minimization and conformational analysis (energy minimization and conformational analysis are important prerequisites for 3D and higher-order descriptor calculation). These classes of descriptors were previously used for modeling in various instances for both single and mixture toxicity modeling [24-26,16,27]. In a successive step, a pretreatment was performed to the obtained descriptor pool (2400 descriptors) of 16 mixture components with the in-built tool of "AlvaDesc" software. In this step, the redundant descriptors (descriptors with a constant value, descriptors with high inter-correlation, descriptors with very low diversity, etc) were removed. After the pretreatment, we computed the mixture descriptors from the pretreated pool by employing 3 different mixing rules i.e., (i) linear combination of molecular contributions, (ii) the squared molecular contributions, and (iii) norm of molecular contributions with a freely available Java-based tool namely "Mix_Desc_Calc" version 1.0 (available from: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1). The linear combination of molecular contributions was used to consider the probable additive effects of mixture components and is computed by the following Eq. (1). whereas, the other two mixing rules i.e., squared molecular contributions and norm of molecular contributions were specified for accounting interactions and are computed by the following Eqs. (2) and (3) respectively [28].

$$Mixture \quad Descriptor = \sum p_i \times x_i \tag{1}$$

$$Mixure \quad Descriptor = \left(\sum p_i \times x_i \right)^2 \tag{2}$$

$$Mixture \quad Descriptor = \sqrt{\sum (p_i \times x_i)^2} \tag{3}$$

Here, p_i represents the mass fraction of i^{th} component and x_i represents the descriptor of i^{th} component. Finally, three classes of mixture descriptors for each dataset were obtained by this method. The single molecular descriptors of 16 antibiotics and the mixture descriptors of MD1 (endpoint organism: *V. fischeri*), MD2 (endpoint organism: *E. coli*), and MD3 (endpoint organism: *B. subtilis*) have been provided in the Supplementary Material 1.

2.3. Dataset division and proper mixing rules identification

The dataset division was performed in 3 different techniques i.e., Sorted activity-based division, Euclidean distance-based division, and Kennard-Stone division using the "Dataset Division GUI" version 1.2 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). To avoid the probable selection bias of training and test mixtures, both descriptor-based (Euclidean distance-based division and Kennard-Stone division) and response-based (Sorted activity-based division) techniques were utilized. Nine different training-test division sets were obtained (4 seeds of sorted activity-based divisions, 4 seeds of Euclidean distance-based divisions, and 1 Kennard-Stone division) for each descriptor matrix; therefore a total of 27 division sets have been generated from each mixture dataset (from each dataset, we can get 3 different descriptor matrix using 3 mixing rules). In this study, we divided the mixture data into training and test sets of 75% and 25% respectively of the whole data set size; the training sets were used for model development and internal validation and the test sets were exclusively used for external validation of developed models. Before the initial modeling, all the training and test set pairs were subjected to a data pretreatment using the "Data-PreTreatmentTrainTest" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) to further reduce the redundant data generated after division. We selected the variance cutoff of 0.001 and the intercorrelation cutoff of 0.95 for the pretreatment.

The selection of the best division and best mixing rules were performed based on the statistical quality of preliminary Genetic algorithm–based multiple linear regression (GA-MLR) models. The "Double Cross-Validation" Version 2.0 tool [29] (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) was used for developing the preliminary GA-MLRs and the quality metrics (Q^2_{LOO} and Q^2_{F1}) have been tabulated in Table S5 of Supplementary material 1. In Double cross-validation, the training set is further divided into calibration and validation sets; the important descriptors are extracted by Genetic algorithm from the calibration set and modeled using MLR and then further validated with the validation set. Finally, those extracted descriptors are used for modeling the actual training set and then validated with the test set. For finding the best mixing rule, we calculated the mean of cross-validated correlation coefficients (Q^2_{LOO}) (which are based on solely the training sets) obtained from models of 9 different divisions. It was observed that mixing rule 2 (the squared molecular contributions), mixing rule 1 (linear combination of molecular contributions), and mixing rule 3 (norm of molecular contributions) produce the highest mean Q^2_{LOO} values for MD1 ($\overline{Q^2_{LOO}}$=0.920), MD2 ($\overline{Q^2_{LOO}}$=0.919), and MD3 ($\overline{Q^2_{LOO}}$=0.810), respectively; therefore these were selected as best mixing rules for MD1, MD2, and MD3 for further analysis. Besides the mixing rules identification, we also selected the best division based on the overall performance of preliminary GA-MLR in terms of balanced Q^2_{LOO} and Q^2_{F1} values. Two different seeds of Euclidean distance-based divisions were selected for MD1 (4–2 DBD) and MD3 (4–3 DBD), whereas, a sorted response-based division was selected for MD2 (4–1 ABD), and we proceeded further with the selected divisions.

2.4. Multilayered descriptor selection and QSAR/QSAAR model development

In this work, we performed a multi-layered feature selection technique to extract valuable descriptors for modeling. The initial pool reduction was commenced by stepwise selection (using "MINITAB" version 14 software [30] in multiple iterations) and genetic algorithm (GA) (using the "Genetic algorithm: version 4.1 tool; available from http://teqip.jdvu.ac.in/QSAR_Tools/). Finally, the most significant descriptors were extracted by checking all possible subsets from the reduced pool of descriptors by employing the "BestSubsetSelection" version 2.1 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/). From the extracted descriptors of MD1, MD2, and MD3, three partial least squares (PLS) regression models (QSAR) were developed by employing the "PLS_single_Y" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) in a successive step.

To evaluate the impact of environmental degradation if any on the mixture toxicity, we incorporated the dissociation constant (pKa) of antibiotics for moedling. The pKa (strongly acidic and strongly basic) data of 13 antibiotics were collected from Drug Bank Database (available from: https://go.drugbank.com/), and pKa (strongly acidic and strongly basic) data of 3 antibiotics (sulfamonomethoxine, diaveridine, and ormetoprim) were computed from "MarvinSketch" software [23]. The pKa data used for modeling has been provided in Supplementary Material 1. From the pKa data of individual components, we computed the mixture descriptors using three mixing rules discussed in Section 2.2

("Mixture descriptor calculation considering three different mixing rules") . Finally, another three QSAR models were developed using PLS regression with the identified descriptors in previous steps and the pKa (strongly acidic and strongly basic) descriptors from three datasets.

For the development of interspecies QSAAR models, common data points for two toxicity endpoints are needed. But in this study, there were no common existing mixture data points which can be used for this purpose. Hence, we employed the predicted toxicity values obtained from the developed QSAR models in this study as predictors along with additional structural descriptors. Two QSAAR models were developed for each dataset; therefore, a total of 6 interspecies QSAARs, i.e., (i) QSAAR of *V. fischeri* with the predicted toxicity of *E. coli* as a predictor, (ii) QSAAR of *V. fischeri* with the predicted toxicity of *B. subtilis* as a predictor, (iii) QSAAR of *E. coli* with the predicted toxicity of *V. fischeri* as a predictor, (iv) QSAAR of *E. coli* with the predicted toxicity of *B. subtilis* as a predictor, (v) QSAAR of *B. subtilis* with the predicted toxicity of *V. fischeri* as a predictor, and (vi) QSAAR of *B. subtilis* with the predicted toxicity of *E. coli* as a predictor were obtained. Feature selection was performed in a manner similar to the previous QSAR modeling. The extracted descriptors along with the predicted toxicity of other species were modeled against the corresponding endpoint using PLS regression by employing the "PLS_single_Y" version 1.0 tool (available from: http://teqip.jdvu.ac.in/QSAR_Tools/) in all 6 cases. The general workflow has been schematically represented in the following Fig. 1.

2.5. Application of q-RASAR in QSAAR development

We applied the concept of q-RASAR in this study to incorporate the biological descriptors in terms of the predicted toxicity of other organisms during QSAAR modeling. Three different RASAR-based QSAAR models were developed in this study from three mixture datasets. The descriptors obtained during preliminary PLS-QSAAR models were clubbed together into 3 sets (unique descriptors of QSAAR 1 and 2 are used for the RASAR descriptor calculation of MD1; unique descriptors of QSAAR 3 and 4 are used for the RASAR descriptor calculation of MD2; unique descriptors of QSAAR 5 and 6 are used for the RASAR descriptor calculation of MD3) for the RASAR descriptor calculation (Table S6 in Supplementary material 1). The RASAR descriptors are the read-across derived similarity, error, concordance, and RA-based prediction functions which may be clubbed together with additional structural and physicochemical descriptors to form the RASAR model [17]. Read-across predictions and corresponding hyperparameter optimization are the prerequisites of RASAR descriptor calculation. Read-across is a similarity-based prediction technique where the exerted response of any unknown compound (target) can be predicted from the known response values of structural analogues (source). Chatterjee et al. have introduced a novel read-across prediction workflow in a previous study where they used Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity estimation techniques for identifying the close structural analogues against query compounds [31]. The hyperparameters sigma (σ) and gamma (γ) are the radial basis functions used in Gaussian and Laplacian kernel similarity estimations respectively. The number of close-training compounds is another hyperparameter which also affects the prediction quality of all three techniques. In this study, we optimized the value of sigma, gamma, and the number of closed training compounds using 5-fold cross-validation by keeping 25% of training data out in the validation sets (different from the test sets). Different sigma values (0.25–2), different gamma values (0.25–2), and 2–10 no. of closed training compounds were used for this optimization (based on the quality of predictions in terms of Q^2_{F1} value for a validation set). This computation was performed using an in-house tool "Auto RA Optimizer" version 1.0 and the optimized settings for three mixture datasets have been reported in Table S7 of Supplementary Material 1.

Finally, the RASAR descriptors were calculated using the optimized hyperparameters by employing "RASAR Descriptor Calculator" version 2.0 (available from: https://dtclab.webs.com/software-tools). The training set itself was used as both source and target data for the RASAR descriptor calculation of the training sets, whereas, the training and test sets were used as source and target data respectively for the RASAR descriptor calculation of the test sets. There are chances of overfitting during the RASAR descriptor calculation of the training set because the same training set is used as both source and target data. To avoid this, the "leave-same-out" strategy was used where the same training datapoint has been removed from the close training compounds' list. After the RASAR descriptor calculation, these were clubbed together with structural and physicochemical descriptors of corresponding datasets

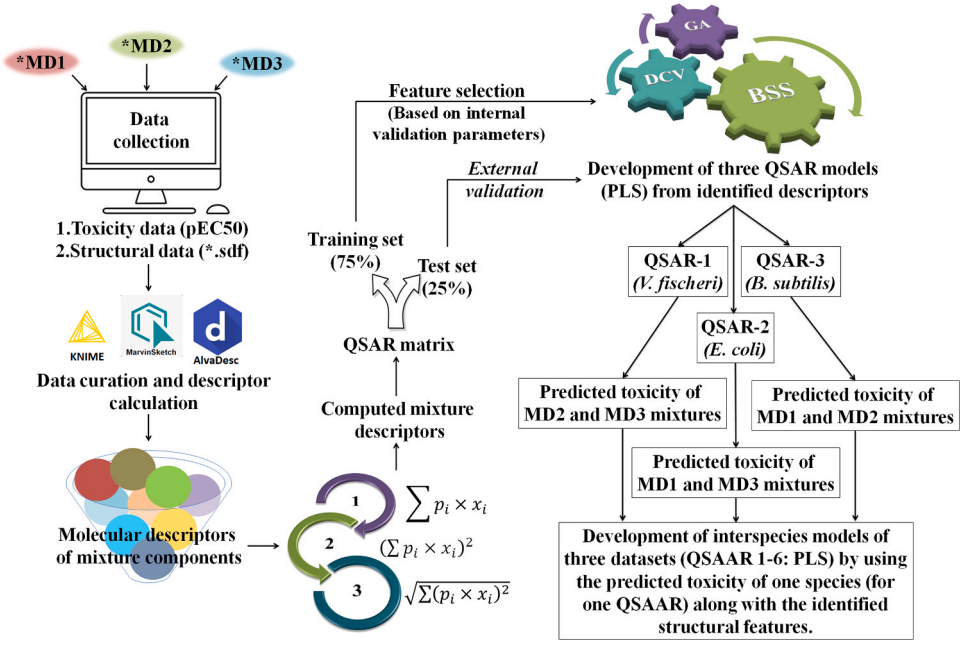


Fig. 1. General QSAR development procedures used in this study; (*MD1: mixture dataset 1, comprised of toxicity data 76 mixtures against *V. fischeri*; MD2: mixture dataset 2, comprised of toxicity data 67 mixtures against *E. coli*; MD3: mixture dataset 3, comprised of toxicity data 42 mixtures against *B. subtilis*).

(MD1, MD2, and MD3) and then the best descriptor combinations were obtained by using the "BestSubsetSelection" version 2.1 tool. In a successive step, the final RASAR models for MD1 and MD3 were developed using single descriptors by MLR, whereas, the RASAR model of MD2 was developed with 4 descriptors and 3 latent variables using PLS regression. In the current work, the RASAR decsriptors were derived by "data fusion" between structural/physicochemical descriptors with toxicity values for other endpoints. Thus, here, the q-RASAR models may more appropriately be termed as q-RASAAR models. We have schematically represented the specific RASAR development procedure (q-RASAAR-2 has been used as an example for the scheme) in Fig. 2.

2.6. Statistical model validation

Statistical validation is one of the essential aspects of any predictive operation because it establishes the reliability and relevance of the predicted response obtained by the developed model [32]. The quality of QSAR models should be accurately measured in terms of the goodness of fit, robustness, and predictability, as suggested by OECD's Principle 4 [33,34]. In the conventional practice of QSAR modeling, rigorous internal and external validations are carried out with training and test datasets respectively, for evaluating the performance of the model. In this study, we computed different internal quality and validation metrics such as determination coefficient (R^2), adjusted determination coefficient (R^2_{adj}), variance ratio (F with dregrees of freedom), leave-one-out (LOO) cross-validated correlation coefficient (Q^2_{LOO}), mean absolute error of training set predictions (MAE_{Train}), etc. and external validation metrics such as external correlation coefficients (Q^2_{F1}, Q^2_{F2}), concordance correlation coefficient (CCC), mean absolute error of test set predictions (MAE_{Test}) to adjudge whether the developed models are satisfying the acceptability criteria or not. There are certain thresholds for both the internal and external validation parameters; i.e., R^2 and R^2_{adj} should have computed values more than 0.6, whereas Q^2_{LOO}, Q^2_{F1}, and Q^2_{F2} should have computed values more than 0.5, and CCC should have a computed value more than 0.85 for acceptable models. The valiance ratio should be statistically significant in 95% confidence level at the corresponding degrees of freedom. The values of Q^2_{LOO} and Q^2_{F1} or Q^2_{F2} close to near 1 may not be the prominent indicators of a good model; therefore, the r^2_m metrics ($\overline{r^2_{m(Train)}}$, $\Delta r^2_{m(Train)}$, $\overline{r^2_{m(Test)}}$, $\Delta r^2_{m(Test)}$) were additionally calculated to better reflect the quality of predictions [35]. Roy et al. have suggested the thresholds for r^2_m metrics; the $\overline{r^2_{m(Train)}}$ and $\overline{r^2_{m(Test)}}$ should have computed values greater than 0.5, whereas, for $\Delta r^2_{m(Train)}$ and $\Delta r^2_{m(Test)}$ should have computed values less than 0.2 for acceptable models [35]. Besides these, all the error measures should have minimum computed values. The Y-randomization test [36] was another validation test which confirm whether the model is built by any chance correlation or not. It is carried out by 100 times random shuffling of the response variable (pEC50) of the training data using "SIMCA-P software" [37]. 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. The Akaike information criteria (AIC) values were also computed for the QSARs and q-RASAARs in this study and compared. This metric estimates the prediction error of the models developed from a single dataset and thereby, judges the relative efficiency of the models. The AIC was named after a Japanese scientist Hirotsugu Akaike who formulated it. For a regression model AIC is computed by employing the following Eq. (4) [38]:

$$AIC = n \times \log(\sigma^2) + 2 \times K \quad (4)$$

where, n = number of training data, K = number of descriptors, and σ^2 = mean sum squared residual. The least AIC value of a model is the indicator of more relative efficiency among various models developed from a dataset.

3. Results and discussions

In the present work, we have developed three QSAR models from MD1, MD2, and MD3 datasets against *V. fischeri*, *E. coli*, and *B. subtilis*, respectively, by employing PLS regression with the optimum number of latent variables. As discussed earlier in the "Dataset division and best mixing rule identification" section, mixing rule 2 (the squared molecular contributions), mixing rule 1 (linear combination of molecular contributions), and mixing rule 3 (norm of molecular contributions) were selected as the best mixing rules for QSAR-1, QSAR-2, and QSAR-3 respectively. The developed QSAR models have been represented in the following Eqs. (5), (6), and (7).

QSAR-1

$$pEC50(Mol/L) = 6.49686 - 0.01966 \times ALOGP2 - 0.00004 \times SAacc + 0.53033 \times nCL - 0.31642 \times C-002 - 0.55166 \times C-032 - 1.09605 \times nR05 \quad (5)$$

QSAR-2

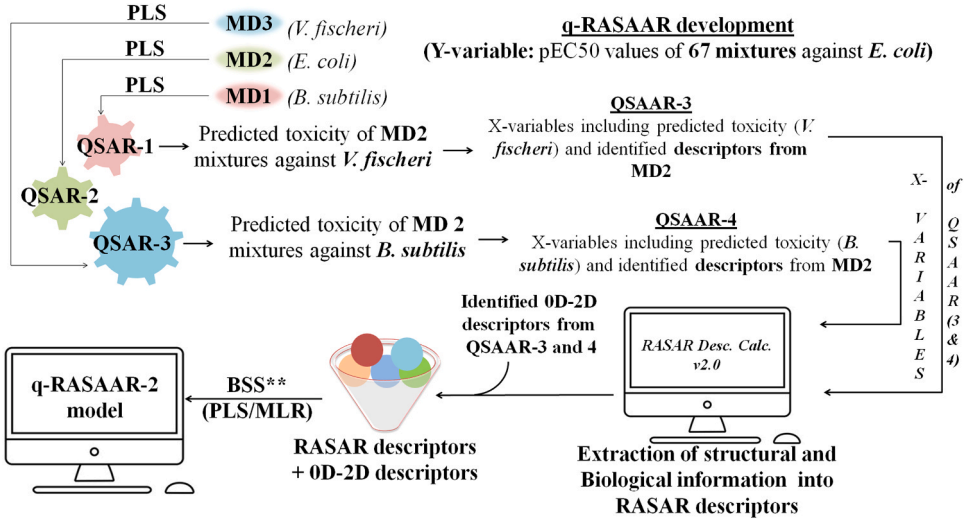


Fig. 2. Specific q-RASAAR development procedures used in this study (q-RASAAR-2 has been taken as an example for creating the scheme)(**BSS: best subset selection).

$$pEC50(Mol/L) = 8.43353 - 0.75722 \times F02[N-O] - 0.28182 \times F04[C-N] + 0.04325 \times MLOGP2 + 0.17719 \times C-028 + 0.1223 \times C-029 - 0.22751 \times B03[N-N] \quad (6)$$

QSAR-3

$$pEC50(Mol/L) = 9.36117 - 2.93858 \times Eta_{betaP_A} - 0.99832 \times B05[C-S] - 0.46666 \times NsCH3 - 8.3358 \times minaasC + 0.06123 \times T(N..Cl) \quad (7)$$

All these models were thoroughly validated and proved to be acceptable in terms of robustness and predictability. The computed validation metrics of these three models have been reported in Table 1. We also performed the Y-randomization test for the developed QSAR models; the Y-randomization plots have been given in Fig. S1 of Supplementary Material 1. The intercept values of QSAR-1, QSAR-2, and QSAR-3 respectively are R^2: 0.0192, 0.0315, 0.0307, and Q^2: − 0.628, − 0.651, − 0.589 which indicate that the models are not developed by any chance correlation. We have checked the applicability domain of these developed QSARs (PLS) by the DModX approach in 99% confidence level [39] using SIMCA-P software [37] and represented the resultant plots in Fig. S2 of Supplementary Material 1. All the mixtures (both training and test sets) of these three datasets are falling within the applicability domain (represented by the D-critical line) of corresponding training sets.

To check the impact of environmental degradation of antibiotics on the mixture toxicity predictions, three different QSAR models were developed incorporating the pKa (strong acidic and strong basic) values with the previously identified features in QSAR-1, QSAR-2, and QSAR-3 (Eqs. 5, 6, and 7 respectively). The new models along with the computed validation metrics have been tabulated in Table S8 of Supplementary Material 1. The computed validation metrics of all these three QSAR models are in the acceptable range as per OECD criteria. To check the models' performance, we computed Akaike information criteria (AIC) for both QSARs and QSARs with pKa descriptors. It has been observed in each of the three datasets that AICs of QSARs (AIC: QSAR-1 = −71.238, QSAR-2 = −82.315, QSAR-3 −41.726) are less than the AICs of QSARs with pKa (AIC: MD1 = −65.995, MD2 = −76.827, MD3 = −38.388) which indicates the better relative efficiency of QSARs in terms of performance and descriptor combination. Therefore, it infers less importance of acidic dissociation constant values in the predictive modeling of mixture toxicity of antibiotics.

Table 1
Computed validation metrics of the developed PLS-QSTR models.

Validation metrics	QSAR-1 (Organism: *V. fischeri*)	QSAR-2 (Organism: *E. coli*)	QSAR-3 (Organism: *B. subtilis*)
n_{Train}	57	51	32
n_{Test}	19	16	10
No. of descriptors (No. of LVs)	6 (5 LVs)	6 (5 LVs)	5 (4 LVs)
Internal validation metrics			
R^2	0.942	0.937	0.894
R^2_{adj}	0.935	0.928	0.873
Q^2_{LOO}	0.925	0.922	0.851
MAE_{Train}	0.097	0.089	0.123
F value(*df1, df2)	167.856 (5, 51)	133.590 (5, 45)	56.920 (4, 27)
$\overline{r^2_{m(Train)}}$	0.895	0.891	0.791
$\Delta r^2_{m(Train)}$	0.027	0.050	0.089
AIC	-71.238	-82.315	-41.726
External validation metrics			
Q^2_{F1}	0.944	0.919	0.841
Q^2_{F2}	0.944	0.915	0.745
CCC	0.969	0.959	0.845
MAE_{Test}	0.112	0.098	0.196
$\overline{r^2_{m(Test)}}$	0.899	0.868	0.569
$\Delta r^2_{m(Test)}$	0.037	0.062	0.180

*df = degree of freedom

As we discussed in the materials and method section, QSAAR models were also developed in this study to find whether there are any intercorrelations among interspecies toxicity or not. First, 6 QSAAR models were developed using PLS regression and then, the RASAR algorithm was used to develop three final RASAR-based QSAAR (q-RASAAR) models from 3 datasets. The preliminary PLS-QSAAR models have been tabulated along with their validation metrics in Table S9 of Supplementary Material 1. All these models are acceptable in terms of goodness of fit, robustness, and predictability. The final RASAR-based QSAARs (i.e., q-RASAARs) were developed using both MLR (q-RASAAR 1 and q-RASAAR-3) and PLS regression (q-RASAAR-2) with an optimum number of LVs and were validated thoroughly. The three models along with computed validation metrics have been reported in the following Table 2. The acceptable values of computed validation metrics indicate the statistical soundness and predictability of these three q-RASAAR models. One point should be noted here that both q-RASAAR-1 and 3 are univariate MLR models with excellent statistical quality, especially in terms of external predictability. We have plotted the observed vs. predicted toxicity scatter plots of both QSAR and q-RASAAR models and showcased those in the following Fig. 3. A uniform scattering surrounding the trend line has been observed in all the plots.

To comply with the OECD principle 3 (a defined domain of applicability), we have analysed the applicability domain of the final q-RASAAR models. The applicability domain is the theoretical region in descriptor space defined by the structural/biological/physicochemical features of training set compounds in which the predictions obtained by the developed model is reliable [39]. In this study, we have used the William's plot to define the domain of applicability of q-RASAAR models. The William's plots are obtained by plotting the cross-validated standardized residual in Y-axis and computed leverages (hat value) in X-axis of all training and test set compounds (Fig. S3 of Supplementary Material 1). The leverage values have been computed using "Hi-Calc" version 2.0 (available from https://dtclab.webs.com/software-tools). The applicability domain is represented by the leverage threshold h* line (h* = 3(p´+1)/n: p´=number of predictor variable, n = number of training compounds) drawn in the William's plot [40]. There is no outlier noted in q-RASAAR-1 and q-RASAAR-3 models. In q-RASAAR-2, the test mixture **EC 5** (binary mixture of diaveridine and sulfamonomethoxine) has been noted outside the AD (Fig. S3 of Supplementary Material 1).

3.1. Insights of modeled descriptors identified in the QSAR analysis

The important descriptors for three QSAR models were identified by multilayered feature selection strategy as discussed in the "Multilayered descriptor selection and QSAR/QSAAR model development" section. The identified descriptors have been tabulated along with their class and definition in Table S10 of Supplementary Material 1. For further analysis of modeled descriptors, variable importance plots (VIP) and coefficient plots of the developed PLS-QSAR models were obtained using "SIMCA-P" software [37] and represented as bubble plots in Fig. 4. The VIP scores of modeled descriptors have been represented as bubble radius whereas, the type of contributions (positive/negative) towards toxicity have been indicated by the colour of bubbles (green bubble = positive contribution; red bubble = negative contribution). The VIP score is the indicator of the relative importance of descriptors; in standard practice, descriptors with VIP scores > 1 show significant importance in toxicity prediction.

3.1.1. Features responsible for toxicity against V. fischeri identified by the QSAR-1 model

The descriptors of interest in QSAR-1 are SAacc, C-032, nR05, C-002,

Table 2
The developed q-RASAAR models and the computed validation metrics.

Models	Equation	Computed validation metrics
q-RASAAR-1 (Univariate MLR)	$pEC50 = -0.02009 + 1.00347 \times RAfunction(LK)$ (8)	$nTrain = 57$, $R^2 = 0.929$, $Q^2_{LOO} = 0.925$, $MAE_{Train} = 0.110$, $F\ value = 722.68$ (df: 1, 55), $\overline{r^2_{m(Train)}} = 0.894$, $\Delta r^2_{m(Train)} = 0.058$, $Q^2_{F1} = 0.975$, $Q^2_{F2} = 0.975$, $CCC = 0.988$, $MAE_{Test} = 0.080$, $\overline{r^2_{m(Test)}} = 0.963$, $\Delta r^2_{m(Test)} = 0.019$, $AIC = -94.858$
q-RASAAR-2 (PLS with 3 LVs)	$pEC50 = 8.62151 - 0.77254 \times FO2[N-O] - 0.26976 \times F04[C-N] - 0.1759 \times B05[N-O] + 0.04919 \times Neg.Avg.Sim$ (9)	$nTrain = 51$, $R^2 = 0.925$, $Q^2_{LOO} = 0.906$, $MAE_{Train} = 0.091$, $F\ value = 192.55$ (df: 3, 47 f), $\overline{r^2_{m(Train)}} = 0.869$, $\Delta r^2_{m(Train)} = 0.056$, $Q^2_{F1} = 0.944$, $Q^2_{F2} = 0.944$, $CCC = 0.970$, $MAE_{Test} = 0.080$, $\overline{r^2_{m(Test)}} = 0.918$, $\Delta r^2_{m(Test)} = 0.010$, $AIC = -82.424$
q-RASAAR-3 (Univariate MLR)	$pEC50 = 0.19879 + 0.96379 \times RAfunction(LK)$ (10)	$nTrain = 32$, $R^2 = 0.847$, $Q^2_{LOO} = 0.830$, $MAE_{Train} = 0.129$, $F\ value = 165.49$ (df: 1, 30), $\overline{r^2_{m(Train)}} = 0.763$, $\Delta r^2_{m(Train)} = 0.118$, $Q^2_{F1} = 0.965$, $Q^2_{F2} = 0.944$, $CCC = 0.969$, $MAE_{Test} = 0.093$, $\overline{r^2_{m(Test)}} = 0.836$, $\Delta r^2_{m(Test)} = 0.042$, $AIC = -44.586$

ALOGP2, and nCl. Among these 6 descriptors, all except nCl (positive contribution) have negative contributions toward cellular toxicity against *V. fischeri*. According to the VIP represented in the "Bubble plot" of Fig. 4**A**, the order of relative importance of modeled descriptors is SAacc > C-032 > nR05 > C-002 > ALOGP2 > nCl. We have also depicted the fragment/s responsible for the non-zero value of descriptors (QSAR-1) in Fig. 4**A**. To understand the relationship between individual descriptors and toxicity, we have systematically compared 3 pairs of mixtures in Table S11 of Supplementary Material 1. The SAacc is the most important descriptor of QSAR-1 and is categorized as a molecular property descriptor which represents the surface area of H-bond acceptor atoms. It has shown a negative contribution towards toxicity; thus, higher availability of electron acceptor atoms decreases the mixture toxicity against *V. fischeri*. If we compare the extreme points of the datasets, then it can be observed that the most toxic mixture has a lower SAacc value than the least toxic mixtures due to the presence of components with more H-bond acceptor atoms (oxytetracycline present in **VF 34** as shown in Fig. 4**A**). C-032 is another important descriptor of QSAR-1, which reduces the toxic response. The presence of X-CX-X fragments (X = electronegative atoms; i.e., O, N, S, Halogens, etc.) in the 2D structure of mixture components are identified by the C-032 descriptor. The presence of X-CX-X fragment has been marked in a component (ormetoprim) of a comparatively less toxic mixture **VF 67** in Fig. 4**A**. nR05 is the next important descriptor of QSAR-1, which gives the number of 5-membered rings in a molecular structure. The negative coefficient of nR05 signifies an inversely proportional relationship with toxicity against *V. fischeri*. The presence of 5-membered ring has been observed in the mixture component of the least toxic mixture **VF 34** (sulfamethoxazole is a 5-membered ring-containing component; see Fig. 4**A**). The C-002 is the 4th important descriptor of QSAR-1, which signifies the presence/absence of CH2R2 fragment (R is any group connected with carbon) in a molecular structure (this feature has been marked on the structural representation of pyrimethamine: component of a comparatively less toxic mixture **VF 45**; see Fig. 4**A**). It has a negative contribution toward the toxicity endpoint which can be justified by comparing the mixtures presented in Table S11 of Supplementary Material 1. The next important structural feature of QSAR-1 is the molecular property descriptor ALOGP2 which computes the value of the squared Ghose-Crippen octanol-water partition coefficient (log P^2) of chemical substances. It also has a negative contribution towards toxicity against *V. fischeri*. A systematic decrease in toxicity has been observed with an increase in the ALOGP2 values (see Table S11 in Supplementary Material 1), which supports the negative contribution of this descriptor. The only positive contributing descriptor of QSAR-1 is nCl which denotes the number of chlorine atoms present in a molecule. Among the mixture components in this study, pyrimethamine and chlortetracycline contain Cl in their structures. Therefore, Cl-containing mixtures like **VF 70**, **VF 51**, **VF 31** etc. have comparatively higher nCl values and toxicity than the mixtures without Cl atoms (e.g., **VF 34**: least toxic mixture; see Table S11 in Supplementary Material 1).

3.1.2. Features responsible for toxicity against E. coli identified by the QSAR-2 model

Just like the previous QSAR-1, the important structural features of MD2 were identified by multilayered descriptor selection and were modeled by PLS regression. Thus, the descriptors of interest in QSAR-2 are F02[N-O], F02[C-N], MLOGP2, C-028, C-029, B03[N-N]. An equal distribution between positive (C-028, MLOGP2, and C-029) and negative (F02[N-O], F04[C-N], and B03[N-N]) contributing descriptors has been observed in QSAR-2. The order of relative importance of the modeled descriptors is F02[N-O] > F02[C-N] > C-028 > MLOGP2 > B03[N-N] > C-029, according to the VIP represented in the "Bubble plot" of Fig. 4**B**. Among these, F02[N-O] and F04[C-N] have VIP scores greater than 1; therefore, their relative importance is more significant. Both of these negatively contributing descriptors are from the class of 2D atom pairs, and they compute the frequency of atom pairs N-O (former) and C-N (latter) at topological distances 2 and 4 respectively. We have systematically compared 3 pairs of mixtures for each descriptor (in Table S12 of Supplementary Material 1) to understand the relationship between structural features and toxicity. F02[N-O] is the most important descriptor of QSAR-2 which inversely impacts the toxicity against *E. coli*. We have depicted the N-O atom pair at topological distance 2 within the structure of the mixture component chlortetracycline in Fig. 4**B**. It has been observed that the mixtures with greater F02[N-O] values are less toxic (e.g., **EC 18**, **EC 48**, **EC 29**) than the mixture with less value (**EC 36**: most toxic mixture of MD2) which justifies the inversely proportional relationship between F02[N-O] and toxicity. A similar observation has been found in the case of descriptor F04[C-N] also. The C-N atom pair at topological distance 4 has been depicted in the structural representation of sulfadoxine (see Fig. 4**B**) which is a component of the least toxic mixture of MD2 (**EC 18**: chlortetracycline + sulfadoxine). In Table S12 of Supplementary Material 1,

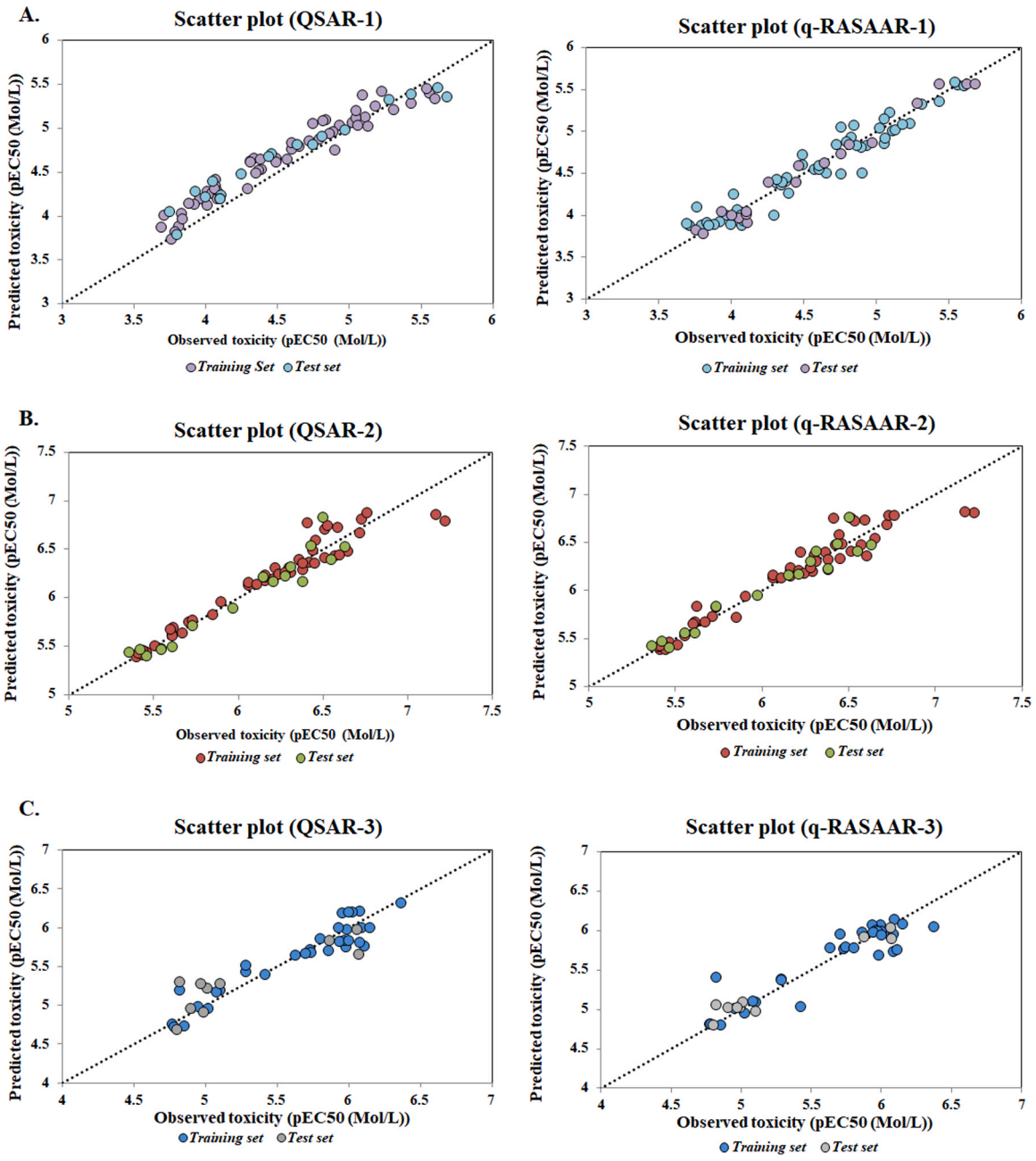


Fig. 3. Observed vs. predicted toxicity scatter plots; A. Scatter plots of QSAR-1 (left) and q-RASAAR-1 (right) models developed from MD1 dataset against toxicity of *V. fischeri*, B. Scatter plots of QSAR-2 (left) and q-RASAAR-2 (right) models developed from MD2 dataset against toxicity of *E. coli*, C. Scatter plots of QSAR-3 (left) and q-RASAAR-3 (right) models developed from MD3 dataset against toxicity of *B. subtilis*.

we have seen a systematic decrease in mixture toxicity with an increase in F04[C-N] values which agrees with the negative contribution of this descriptor. The next important descriptor is C-028 which belongs to the class 2D atom pair and indicates the presence of R-CR-X fragment in the molecular structure (R = any group attached via C atom; X = electronegative atoms i.e., O, N, S, Halogens, etc.). It has a positive contribution toward the toxic endpoint; therefore, it increases the mixture toxicity against *E. coli*. We have shown the presence of R-CR-X fragment in the structural representation of sulfamethazine (see Fig. 4**B**) which is present in different toxic mixtures as a component (e.g., **EC 10**, **EC 57**). Three pairs of mixtures have been tabulated along with their toxicity and descriptor values in Table S12 of Supplementary Material 1 for a systematic comparison among mixtures. Increased toxicity has been observed with an increase in the C-028 value in mixtures which signifies a positive contribution. MLOGP2 is a molecular property and the next positive contributing descriptor of QSAR-2. It computes the squared Moriguchi octanol-water partition coeff. (log P^2) which is a measure of the lipophilicity of the mixture. Its positive contribution shows the impact of lipophilicity toward mixture toxicity against *E. coli*. A distinct increase of both MLOGP2 and toxicity has been observed from the least toxic mixture (**EC 18**) to the most toxic mixture (**EC 36**) which aligns with their proportional relationship. B03[N-N] is again a negative contributing descriptor which indicates the presence or absence of N-N atom pair at topological distance 3. The most toxic mixture (**EC 36**: pEC50 = 7.22) does not contain components with N-N atom pair at the topological distance 3; however, we have seen the presence of N-N atom (3 topological distance apart) pairs in various less toxic mixtures (**EC 66**, **EC 34**, **EC 13**; see Table S12 in Supplementary Material 1). We have

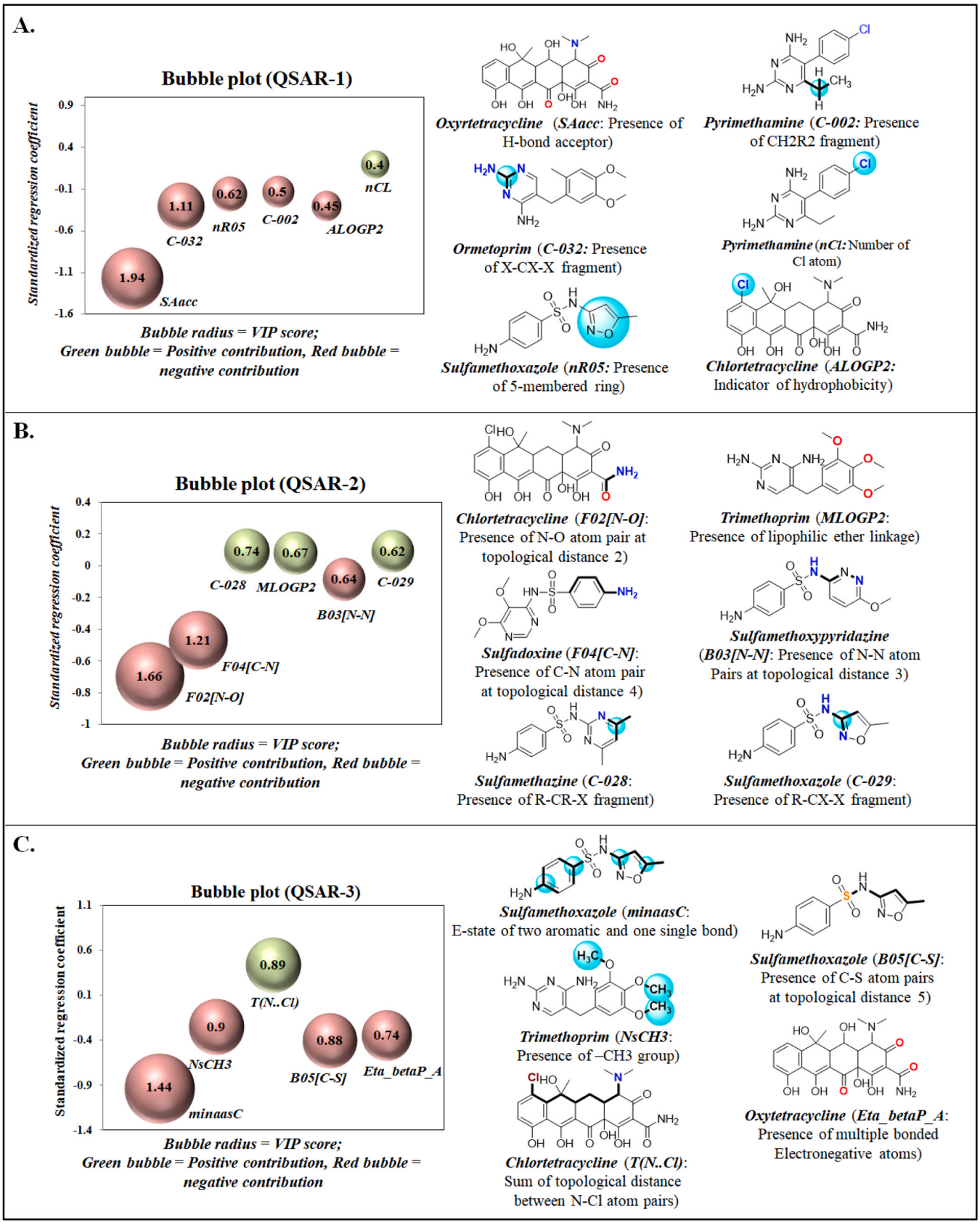


Fig. 4. Variable importance plots of QSAR models along with the structural representation of identified descriptors; A. VIP of QSAR-1, B. VIP of QSAR-2, C. VIP of QSAR-3 (Bubble radius = VIP Score; Colour coding: Green = positive contribution, Red = negative contribution).

depicted the B03[N-N] fragment in the mixture component (sulfamethoxypyridazine) of one of the least toxic mixtures (**EC 13**) in Fig. 4**B**. The final descriptor of QSAR-2 is the C-029 which positively contributes toward the toxicity against E. coli. It indicates the presence/absence of the R-CX-X fragment (R = any group attached by a C atom; X = electronegative atoms e.g., O, N, S, Halogens, etc.) in the molecule. The presence of the R-CX-X fragment has been observed in more toxic mixtures like **EC 6** (pEC50 = 6.72) and has been depicted in the molecular structure of sulfamethoxazole (a component of **EC 6**) in Fig. 4**B**. A marked decrease in toxicity and C-029 values has been observed in the mixtures tabulated in Table S12 of Supplementary Material 1 which justifies the positive contribution of C-029 towards toxicity against *E. coli*.

3.1.3. Features responsible for toxicity against B. subtilis identified by the QSAR-3 model

Five distinct descriptors namely Eta_betaP_A, B05[C-S], NsCH3, minaasC, T(N..Cl) were identified in QSAR-3. Among these descriptors, all except T(N..Cl) are negatively contributing toward the modeled endpoint (pEC50). According to the "Bubble plot" presented in Fig. 4**C**, the order of relative importance of modeled descriptors is minaasC > NsCH3 > T(N..Cl) > B05[C-S] > Eta_betaP_A; therefore, minaasC (VIP score = 1.44) and Eta_betaP_A (VIP score = 0.74) are identified as the most important and least important descriptors respectively. The minaasC is a negative contributing descriptor of QSAR-3 which indicates the electronic and topological state of fragments containing 2 aromatic and 1 single bond [41]. We have depicted 4 such fragments inside a mixture component sulfamethoxazole (a component of the least toxic mixture **BS 15** of MD3) in Fig. 4**C**. It has been observed in the studied mixtures that the toxicity decreases with an increase in minaasC value which indicates the inverse relationship. We have systematically compared 3 pairs of mixtures in terms of their pEC50 and minaasC values in Table S13 of Supplementary Material 1**;** the result agrees with the negative contribution of minaasC toward the toxicity endpoint. NsCH3 is the next important descriptor with a VIP score of 0.9, which computes the number of single-bonded CH_3 groups in a molecular structure (see Fig. 4**C**). In the comparative analysis represented in Table S13 of Supplementary Material 1, various mixtures such as **BS 35**, **BS 21**, and **BS 26** have shown more NsCH3 values and thus, less toxicity than comparatively more toxic mixtures like **BS 32**. These observations justify the negative contribution of the mentioned descriptor. T(N..Cl) is the only positive contributing descriptor of QSAR-3 which indicates the total topological distance between the N-Cl atom pair in a molecule. In Fig. 4**C**, we have depicted the presence of N-Cl atom pair in the structure of chlortetracycline which is a component of a comparatively more toxic mixture (**BS 24**; pEC50 = 5.98, T(N..Cl) = 9.60) of MD3. The mixtures without the N-Cl atom pair (e.g., **BS 2**, **BS 21**, **BS 15**) have shown less toxicity (see Table S13 in Supplementary Material 1) which signifies the positive contribution of T(N..Cl). The B05[C-S] and Eta_betaP_A are the final two descriptors of QSAR-3 which negatively affect the toxic endpoint. B05[C-S] is a 2D atom pair descriptor which indicates the presence/absence of C-S atom pair at topological distance 5, whereas the Eta_betaP_S is an electrotopochemical atom index descriptor which is a measure of the multiple bonded electronegative atoms in the molecule. In both cases, a marked increase in the corresponding descriptor values has been observed from the most toxic mixture (**BS 32**: pEC50 = 6.37, B05[C-S] = 0, Eta_betaP_S = 0.40) to the least toxic mixture (**BS 15**: pEC50 = 4.77, B05[C-S] = 0.52, Eta_betaP_S = 0.48) of MD3 dataset which indicates the inverse relationship with toxicity against *B. subtilis*. The comparative analysis among 3 pairs of mixtures represented in Table S13 of Supplementary Material 1 also agrees with the negative contribution of these two descriptors.

3.2. Insights of modeled descriptors identified in the QSAAR analysis

To understand the correlation of toxicity among different bacterial species, six preliminary QSAAR models (PLS) have been developed here. The impact of the predicted toxicity of a species on the toxicity prediction of another has been analyzed by the VIPs and coefficient plots of QSAAR models (developed using SIMCA-P software). The VIP and coefficient plots of QSAAR-1 or − 2, QSAAR-3 or − 4, and QSAAR-5 or − 6 have been showcased in Fig. S4, S5, and **S6** respectively in Supplementary Material 1. From the VIPs of QSAAR-1 and 2 (Fig. S4 of Supplementary Material 1), it has been found that the predicted toxicity (*B. subtilis*) possesses a positive contribution and significant relative importance towards the toxicity prediction against *V. fischeri*. In the VIP of QSAAR-5 (Fig. S6 of Supplementary Material 1), the predicted toxicity (*V. fischeri*) has shown maximum relative importance among other predictors for the toxicity prediction against *B. subtilis*. It is also a positively contributing descriptor that suggests that a toxic mixture against *V. fischeri* may also be toxic against *B. subtilis*. Therefore, it may be inferred from these observations that *V. fischeri* and *B. subtilis* have similar toxicity mechanisms. The predicted toxicity (*B. subtilis*) does not possess significant importance toward the toxicity prediction against *E. coli*. Among the predicted toxicities, *B. subtilis* have shown a negative contribution to the toxicity prediction against *E. coli* in QSAAR-4. Therefore, a more toxic mixture against *B. subtilis* may be less toxic towards *E. coli*. The predicted toxicity (*E. coli*) has not shown much importance in the toxicity prediction against *V. fischeri* also, and it has been confirmed by the VIP of QSAAR-3 (Fig. S5 of Supplementary Material 1). Therefore, it may be inferred that *E. coli* does not have a similar mechanism of toxicity with the other two species (*V. fischeri* and *B. subtilis*).

By analysing the descriptors of QSAAR-1 and 2, it has been observed that QSAAR-1 and 2 have two (C-032 and nCL) and three (SAacc, C-032, and ALOGP2) common descriptors respectively with QSAR-1. These common descriptors have similar impact toward toxicity as seen in QSAR-1. The QSAAR-3 and 4 share two common descriptors (F02[N-O] and F04[C-N]) with QSAR-2. Both of these descriptors are negatively contributing in QSAR-2, QSAAR-3, and QSAAR-4. Among the descriptors of QSAAR-5 and 6, only T(N.Cl) is seen as the common predictor with a positive coefficient in QSAR-3 and QSAAR-6.

3.3. Importance of q-RASAR algorithm in interspecies mixture toxicity modeling

The q-RASAR algorithm has played a crucial role in developing the QSAAR models for toxicity prediction against three bacterial endpoints in this study. As discussed earlier that the q-RASAR is a mixed approach of non-linear read-across predictions and a linear QSAR framework. Therefore, the non-linearity obtained during the similarity estimation by Gaussian kernel function/Laplacian kernel function are applied to form a linear QSAR equation. In this study, we have used the optimized Laplacian kernel function [31] for the RASAR descriptor calculation of MD1 and MD3, and the optimized Euclidean distance-based similarity [31] for the MD2 (mathematical representations of Laplacian kernel function similarity and Euclidean distance-based similarity have been tabulated in Table S14 of Supplementary Material 1). The RASAR descriptors are derived by extracting structural information obtained from the dataset; here both structural and biological information (in the form of the predicted toxicity of other bacterial species) are used to develop the final RASAR-based QSAARs or q-RASAARs (see Fig. 5). We got two univariate MLR models (q-RASAAR-1 and −3) with *RA function (LK)* as the predictor, and 1 PLS model (q-RASAAR-2) with 4 descriptors (including RASAR descriptor *Neg. Avg. Sim*) and 3 LVs. The RASAR descriptor *RA function (LK)* is the read-across derived prediction function obtained by weighted averaging of response values of identified similar structural analogues (source compounds) during non-linear similarity estimation (Laplacian kernel function), whereas, the Neg. Avg. Sim. is the mean of similarity values of non-toxic (with respect to the training set response mean) structural analogues identified by Euclidean distance-based similarity estimation [42]. Both of these

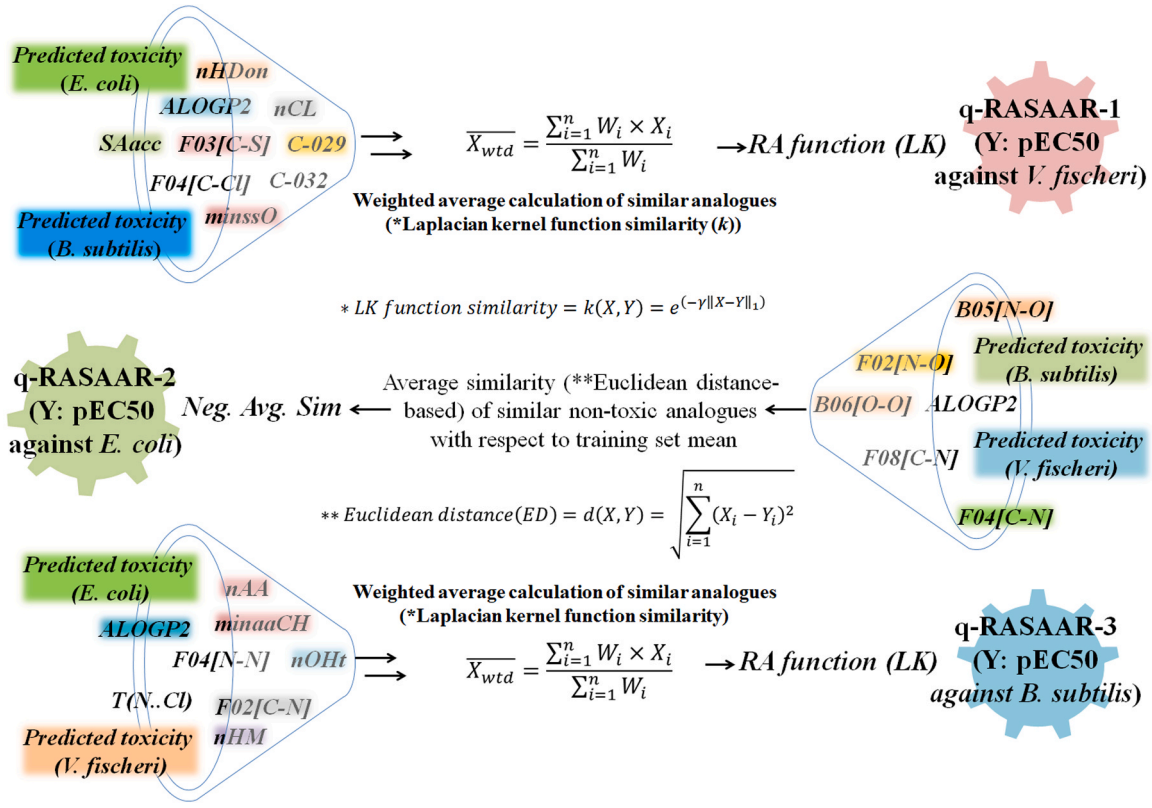


Fig. 5. Representation of the features involved to obtain relevant RASAR descriptors (for q-RASAAR modeling).

descriptors have a proportional relationship with toxic endpoints as evidenced from the positive regression coefficients (see Table 2). The statistical quality we got is comparable with QSAR models in terms of internal validation metrics and superior in terms of external validation metrics (Fig. 6). The variance ratio (F value) has been improved from QSAR (F value: QSAR-1 = 167.856, df: 5, 51; QSAR-2 = 133.59, df: 5, 45; QSAR-3 = 56.92, dfs: 4, 27) to RASAR (F value: q-RASAAR-1 = 722.68, df: 1, 55; q-RASAAR-2 = 192.55, df: 3, 47; q-RASAAR-3 = 165.49, df: 1, 30) which can be a marker of better statistical significance of RASAR models. For better understanding, we have plotted the F values of QSARs and q-RASAARs in Fig. 6. The MAE_{Test} values of QSAR, preliminary QSAARs, and q-RASAAR models have also been plotted in Y-axis of column plots (3 column plots for three datasets) and have given in Fig. 6. An increase in prediction error in terms of MAE_{Test} has been observed from QSAR-1 to QSAAR-1/− 2 and QASR-3 to QSAAR-5/− 6 which indicates inferior prediction by PLS-based interspecies modeling. But during the q-RASAAR model development, structural (descriptors) and biological features (predicted endpoint values against other two bacterial species) of two QSAARs (QSAAR-1/−2 to q-RASAAR-1; QSAAR-3/−4 to q-RASAAR-2; QSAAR-5/−6 to q-RASAAR-3) are extracted into the corresponding RASAR descriptors which significantly reduce the prediction errors. It is clearly seen that the MAE_{Test} is minimum in q-RASAAR-1 (MAE_{Test} = 0.080), q-RASAAR-2 (MAE_{Test} = 0.080), and q-RASAAR-3 (MAE_{Test} = 0.093) compared to the other QSARs (MAE_{Test}: QSAR-1 = 0.112; QSAR-2 = 0.098; QSAR-3 = 0.198) and preliminary QSAARs (MAE_{Test}: QSAAR-1 = 0.121; QSAAR-2 = 0.124; QSAAR-3 = 0.088; QSAAR-4 = 0.081; QSAAR-5 = 0.198; QSAAR-6 = 0.253). We have also noticed a decrease in CCC values from QSAR-3 (CCC = 0.845: marginal to the threshold) to QSAAR-6 (CCC = 0.807: inferior) and then a marked increase in q-RASAAR-3 (CCC = 0.969: superior) which indicates the efficiency of RASAR algorithm in interspecies modeling. In each of the three datasets, AICs of q-RASAAR (AIC: q-RASAAR-1 = −94.858, q-RASAAR-2 = −82.424, q-RASAAR-3 = −44.586) were less than the AICs of QSARs (AIC: QSAR-1 = −71.238, QSAR-2 = −82.315, QSAR-3 = −41.726) which indicates the better relative efficiency of q-RASAARs in terms of performance and descriptor combination.

4. Comparative analysis with the previous report

Comparative analysis with previous reports can be useful to judge the merit of any predictive model. The use of the interspecies model in mixture toxicity prediction is a new approach and hence, we did not get any previous QSAAR models of mixtures for the studied endpoints. However, we have compared the quality of newly developed QSAR and q-RASAAR models with previously published QSAR models by Wang *et. al.* [21]. A total of 12 MLR models (4 models against each dataset) were reported by Wang et. al. where they used molecular docking-derived E_{bind} descriptors as the predictor variables. They separately developed more specific toxicity models for binary (SA + SAP, SA + TC, SAP + TC) and ternary mixtures (SA + SAP + TC) against each bacterial species with comparatively smaller data points (see Table 3). However, a more general strategy has been taken for the development of QSAR and QSAAR in the present study, and therefore, we have used both binary and ternary mixtures together for the model development. Comparatively greater number of data points have been used for model development in this study (QSAR-1/q-RASAAR-1: n_{Train} = 57, QSAR-2/q-RASAAR-2: n_{Train} = 51, QSAR-3/q-RASAAR-3: n_{Train} = 32) which is desirable for any statistical modeling. We have also compared the robustness and predictability in terms of computed R^2, Q^2_{LOO}, Q^2_{F1} values and tabulated them into Table 3. The QSAR-1 and q-RASAAR-1 clearly exceed (QSAR-1: R^2 = 0.942, Q^2_{LOO} = 0.925, Q^2_{F1}=0.944; q-RASAAR-1: R^2 = 0.929, Q^2_{LOO} = 0.924, Q^2_{F1}=0.975) the validation results of all 4 previous QSARs for predicting toxicity to *V. fischeri*. Similar observations have been obtained for the rest two dataset also. However, we have also found two previous models with greater R^2 and Q^2_{LOO} from MD2 and MD3 than the new models of corresponding datasets. The previous QSAR model developed for the prediction of toxicity

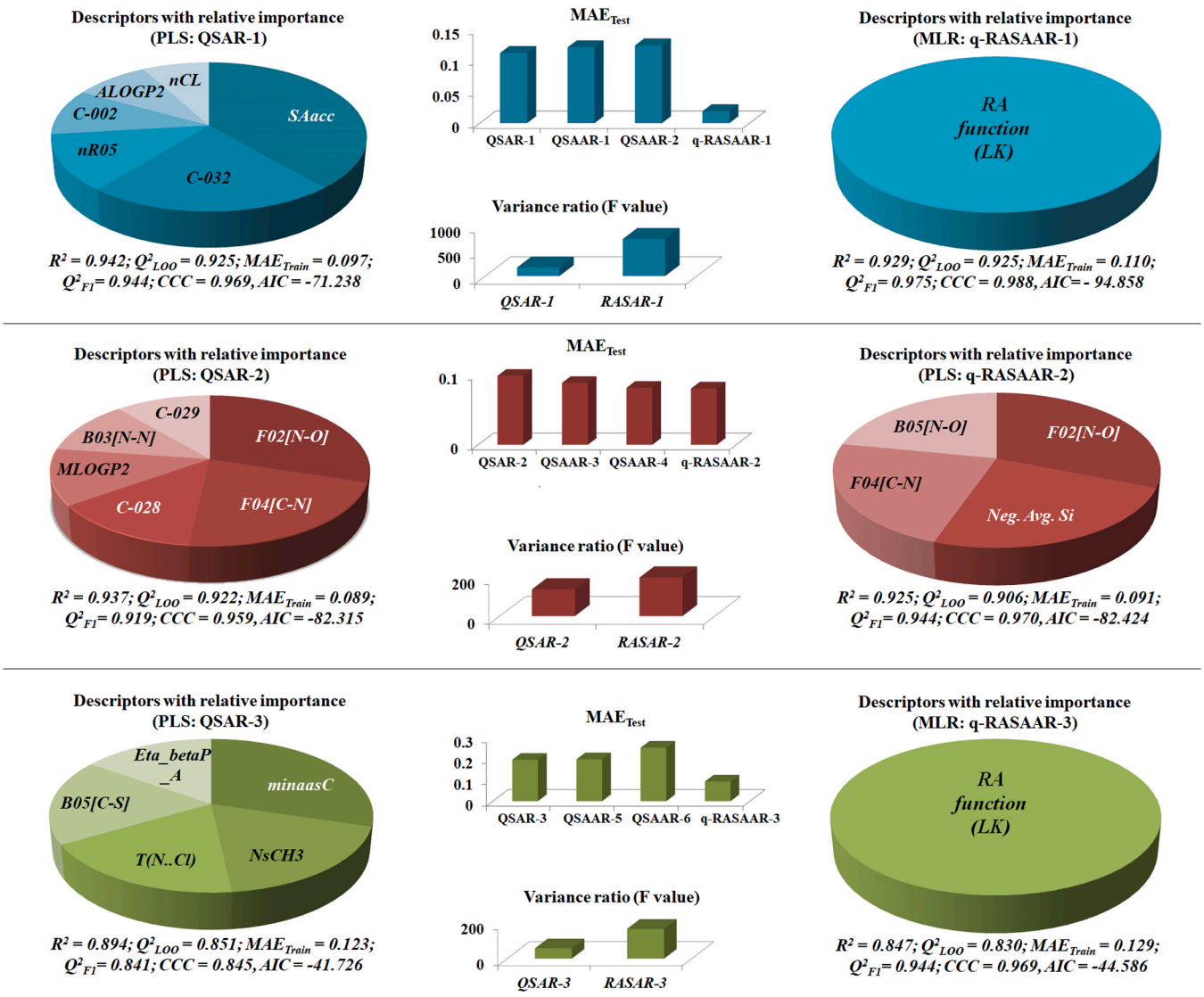


Fig. 6. Comparison of (i) importance of descriptors used in QSAR and q-RASAAR models, (ii) MAETest among QSAR, QSAAR, and q-RASAAR models, (iii) Variance ratio (F value) between QSAR and q-RASAAR models.

of mixtures (comprised of SAP and TC) against *E. coli* showed an R^2 value (0.951) greater than the corresponding QSAR-2 ($R^2 = 0.936$) and q-RASAAR-2 ($R^2 = 0.924$). However, the previous model was developed with 6 data points and thus, it is statistically less significant than the QSAR-2 and q-RASAAR-2 ($n_{Train} = 51$). Another previous model with greater R^2 (0.999) and Q^2_{LOO} (0.945) values was developed for the prediction of binary mixtures (SA-SAP) against *B. subtilis* from only 5 datapoints; thus statistically less significant. On the contrary, the newly developed QSAR-3 ($R^2 = 0.893$, $Q^2_{LOO} = 0.850$) and q-RASAAR-3 ($R^2 = 0.846$, $Q^2_{LOO} = 0.830$) models have been developed with a higher number of training data points ($n_{Train} = 32$), thus having higher statistical significance. This study aimed to develop simple and transferable prediction models for the efficient toxicity prediction of antibiotic mixtures. In this direction, ^0D–^2D descriptors have been used due to their simplicity, reproducibility, and interpretability; the developed models have been proven to be good-fit, robust, and highly predictive also in comparison with the previous report (Table 3). Although the use of ^0D–^2D descriptors limit the ability of model to explain the toxicity mechanism to some extent, but it enhances the reproducibility and transferability of the models. Molecular docking-based descriptors have been used for modeling the mixture toxicity in some of the previous studies; however, those molecular docking-based descriptors do not appear appropriate for modeling nonspecific interactions as no particular target is being considered here for the toxicity. Apart from that, the exact toxicity mechanism is also uncertain, because the toxicity data was obtained by simple assay where the concentration of sample broth is the determinant of lethality. The targets for the toxicity may vary depending on the chemical classes and the organism being considered; therefore, the use of a fixed target receptor or protein for docking-based descriptor calculation potentially limits the efficiency of predictions. The predicted toxicity values of other bacteria have also been incorporated in QSAAR/q-RASAAR models for improving the predictive performance further. We have used the available toxicity data of one bacterial species with other structural descriptors for more efficient prediction of other bacterial endpoints. Therefore, from the perspective of statistical significance, robustness, external predictability, and better usability, the new models (both QSARs and RASARs) perform well and are also able to produce more realistic predictions than the previous models.

5. Conclusion

The present study involves toxicity modeling of chemical mixtures (both binary and ternary) for three bacterial endpoints (*V. fischeri, E. coli, and B. subtilis*) with QSAR and q-RASAR-based QSAAR (q-RASAAR)

Table 3
Comparison of results obtained by this study with previously published report in terms of number of descriptors, robustness, and predictability.

Work done by	Model	Type of mixtures	Bacterial species	No. of training data	R^2	Q^2_{LOO}	Q^2_{F1}	Descriptors (No. of LVs)	Type of descriptors
Wang, et al. [21]	MLR	SA-SAP	V. fischeri	18	0.699	0.6	0.77	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		15	0.813	0.714	0.601	2	
		SAP-TC		8	0.829	0.603	0.760	2	
		SA-SAP-TC		21	0.729	0.601	0.825	3	
Current study	**PLS: QSAR-1**	**Both binary and ternary**	**V. fischeri**	**57**	**0.942**	**0.925**	**0.944**	**6 (5 LVs)**	**0D to 2D**
Current study	**q-RASAAR-1 (MLR)**	**Both binary and ternary**	**V. fischeri**	**57**	**0.929**	**0.924**	**0.975**	**1**	**RASAR descriptor**
Wang, et al. [21]	MLR	SA-SAP	E. coli	9	0.788	0.505	0.500	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		21	0.740	0.649	0.629	2	
		SAP-TC		6	0.951	0.823	0.702	2	
		SA-SAP-TC		20	0.805	0.627	0.640	3	
Current study	**PLS: QSAR-2**	**Both binary and ternary**	**E. coli**	**51**	**0.936**	**0.921**	**0.919**	**6 (5 LVs)**	**0D to 2D**
Current study	**q-RASAAR-2 (PLS)**	**Both binary and ternary**	**E. coli**	**51**	**0.924**	**0.905**	**0.944**	**4 (3 LVs)**	**0D to 2D and RASAR descriptor**
Wang, et al. [21]	MLR	SA-SAP	B. subtilis	5	0.999	0.945	0.702	2	Molecular docking-based descriptor (E_{bind})
		SA-TC		9	0.893	0.812	0.500	2	
		SAP-TC		8	0.778	0.526	0.906	2	
		SA-SAP-TC		11	0.828	0.508	0.754	3	
Current study	**PLS: QSAR-3**	**Both binary and ternary**	**B. subtilis**	**32**	**0.893**	**0.850**	**0.841**	**5 (4 LVs)**	**0D to 2D**
Current study	**q-RASAAR-3 (MLR)**	**Both binary and ternary**	**B. subtilis**	**32**	**0.846**	**0.830**	**0.964**	**1**	**RASAR descriptor**

approaches. This study was designed to develop more efficient toxicity models for mixtures' risk assessment, to identify the best mixing rule for mixture descriptor calculation, to study the use of toxicity of separate species (as predictor) in mixture toxicity modeling, and to apply RASAR algorithm for refining the QSAAR models further. Three QSAR models have been developed by PLS regression and the important structural features have been identified which affect the mixtures' toxicity. Simple ^{0}D–^{2}D descriptors have been used for encoding the structural features in this study which makes it a cost-effective, less complex, reproducible, interpretable, and transferable approach. Three different mixing rules have been used in this study and the best has been identified (MD1: the squared molecular contributions, MD2: linear combination of molecular contributions, MD3: norm of molecular contributions). Six other interspecies QSAAR models (two QSAARs for each dataset) have also been developed using PLS regression with the predicted toxicity of other bacterial species as descriptors. Finally three q-RASAAR models have been developed for predicting toxicity against three bacterial species. All these models have been strictly validated and the computed validation metrics have shown the acceptance of the models as per the OECD guidelines. The interspecies toxicity has not shown much impact on the models' quality in preliminary QSAARs (PLS). However, similar toxicity patterns have been observed in *V. fischeri* and *B. subtilis*. The RASAR algorithm has incorporated the structural (identified descriptors) and biological information (in the form of predicted toxicity of two other species) during the RASAR descriptor calculation and it significantly improves the performance of QSAARs. Among the three models, q-RASAAR-1 and q-RASAAR-3 have been developed with single predictor variable *RA function (LK)* which is a prediction function obtained by weighted averaging of the response variables of similar structural analogues. Therefore, the RASAR algorithm enables the most effective incorporation of the underlined features (both biological and structural) responsible for mixtures' toxicity and thus produces a better quality prediction. In a nutshell, this is the first instance where the interspecies mixture toxicity models coupled with the q-RASAR algorithm have been developed which can predict the cellular toxicity of binary and ternary mixtures against any of the three above-mentioned organisms with sufficient efficiency.

Enviromnetal implications

Antibiotics are often found as mixtures in the environment, and they may produce toxicity against different ecological species due to joint exposure in the sub-optimal range. They mostly target multiple bacteria in the environment; however, mixture toxicity assessment of antibiotics against multiple bacteria is still very rarely found. In this study, we proposed an interspecies QSAAR modeling coupled with the q-RASAR algorithm (q-RASAAR) for better risk assessment of antibiotic mixtures against V. fischeri, E. coli, and B. subtilis. This study proves the effectiveness of q-RASAAR approach for interspecies mixture toxicity modeling which may help us to understand the commonality of toxicity patterns among different bacterial species.

CRediT authorship contribution statement

MC: Formal analysis, Investigation, Data curation, Writing – original draft. **KR:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Acknowledgement

MC sincerely acknowledges the financial assistance of All India Council for Technical Education (AICTE), New Delhi as a form of National Doctoral Fellowship (NDF).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.132129.

References

[1] Kümmerer, K., 2009. Antibiotics in the aquatic environment - a review - part I. Chemosphere 75, 417–434. https://doi.org/10.1016/J.CHEMOSPHERE.2008.11.086.

[2] Kümmerer, K., 2009. The presence of pharmaceuticals in the environment due to human use – present knowledge and future challenges. J Environ Manag 90, 2354–2366. https://doi.org/10.1016/J.JENVMAN.2009.01.023.

[3] Álvarez-Muñoz, D., Rodríguez-Mozaz, S., Maulvault, A.L., Tediosi, A., Fernández-Tejedor, M., Van den Heuvel, F., et al., 2015. Occurrence of pharmaceuticals and endocrine disrupting compounds in macroalgaes, bivalves, and fish from coastal areas in Europe. Environ Res 143, 56–64. https://doi.org/10.1016/j.envres.2015.09.018.

[4] Xiong, W., Sun, Y., Zhang, T., Ding, X., Li, Y., Wang, M., et al., 2015. Antibiotics, antibiotic resistance genes, and bacterial community composition in fresh water aquaculture environment in China. Microb Ecol 70, 425–432. https://doi.org/10.1007/s00248-015-0583-x.

[5] Mendoza, A., Aceña, J., Pérez, S., López de Alda, M., Barceló, D., Gil, A., et al., 2015. Pharmaceuticals and iodinated contrast media in a hospital wastewater: a case study to analyse their presence and characterise their environmental risk and hazard. Environ Res 140, 225–241. https://doi.org/10.1016/j.envres.2015.04.003.

[6] Jiang, L., Hu, X., Yin, D., Zhang, H., Yu, Z., 2011. Occurrence, distribution and seasonal variation of antibiotics in the Huangpu River, Shanghai, China. Chemosphere 82, 822–828. https://doi.org/10.1016/J.CHEMOSPHERE.2010.11.028.

[7] Li, C., Chen, J., Wang, J., Ma, Z., Han, P., Luan, Y., et al., 2015. Occurrence of antibiotics in soils and manures from greenhouse vegetable production bases of Beijing, China and an associated risk assessment. Sci Total Environ 521–522, 101–107. https://doi.org/10.1016/j.scitotenv.2015.03.070.

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

[9] Zou, X., Lin, Z., Deng, Z., Yin, D., Zhang, Y., 2012. The joint effects of sulfonamides and their potentiator on Photobacterium phosphoreum: Differences between the acute and chronic mixture toxicity mechanisms. Chemosphere 86, 30–35. https://doi.org/10.1016/j.chemosphere.2011.08.046.

[10] Wang, D., Ning, Q., Dong, J., Brooks, B.W., You, J., 2020. Predicting mixture toxicity and antibiotic resistance of fluoroquinolones and their photodegradation products in Escherichia coli. Environ Pollut 262, 114275. https://doi.org/10.1016/j.envpol.2020.114275.

[11] Chatterjee, M., Roy, K., 2022. Recent advances on modelling the toxicity of environmental pollutants for risk assessment: from single pollutants to mixtures. Curr Pollut Rep 8, 81–97. https://doi.org/10.1007/s40726-022-00219-6.

[12] Chatterjee, M., Roy, K., 2022. Computational modeling of mixture toxicity. In: Emilio Benfenati. In Silico Methods for Predicting Drug Toxicity. Humana Press Inc., New York, pp. 561–587. https://doi.org/10.1007/978-1-0716-1960-5_22.

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

[14] Cronin, M.T.D., 1995. QSAR in toxicology. 2. Prediction of acute mammalian toxicity and interspecies correlations. Mol Inf 14, 117–120. https://doi.org/10.1002/qsar.19950140202.

[15] Kar, S., Das, R.N., Roy, K., Leszczynski, J., 2016. Can toxicity for different species be correlated?: the concept and emerging applications of interspecies Quantitative Structure-toxicity relationship (i-QSTR) modeling. Int J Quant Struct Relatsh 1, 23–51. https://doi.org/10.4018/IJQSPR.2016070102.

[16] Mukherjee, R.K., Kumar, V., Roy, K., 2022. Ecotoxicological QSTR and QSTTR modeling for the prediction of acute oral toxicity of pesticides against multiple avian species. Environ Sci Technol 56, 335–348. https://doi.org/10.1021/acs.est.1c05732.

[17] 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.

[18] 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.

[19] Banerjee, A., Kar, S., Pore, S., Roy, K., 2023. Efficient predictions of cytotoxicity of TiO2-based multi-component nanoparticles using a machine learning-based q-RASAR approach. Nanotoxicology 17, 78–93. https://doi.org/10.1080/17435390.2023.2186280.

[20] Luechtefeld, T., Marsh, D., Rowlands, C., Hartung, T., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 165, 198–212. https://doi.org/10.1093/toxsci/kfy152.

[21] Wang, D., Wu, X., Lin, Z., Ding, Y., 2018. A comparative study on the binary and ternary mixture toxicity of antibiotics towards three bacteria based on QSAR investigation. Environ Res 162, 127–134. https://doi.org/10.1016/j.envres.2017.12.015.

[22] 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.

[23] ChemAxon, 2018. Marvin [WWW Document]. URL https://chemaxon.com/products/marvin (accessed 10.10.22).

[24] Chatterjee, M., 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.

[25] 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.

[26] 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.

[27] Paul, R., Chatterjee, M., Roy, K., 2022. First report on soil ecotoxicity prediction against Folsomia candida using intelligent consensus predictions and chemical read-across. Environ Sci Pollut Res 29, 88302–88317. https://doi.org/10.1007/S11356-022-21937-W.

[28] Saldana, D.A., Starck, L., Mougin, P., Rousseau, B., Creton, B., 2013. Prediction of flash points for fuel mixtures using machine learning and a novel equation. Energy Fuels 27, 3811–3820. https://doi.org/10.1021/EF4005362.

[29] Roy, K., Ambure, P., 2016. The "double cross-validation" software tool for MLR QSAR model development. Chemom Intell Lab Syst 159, 108–126. https://doi.org/10.1016/j.chemolab.2016.10.009.

[30] Data Analysis, Statistical & Process Improvement Tools | Minitab, 2004.

[31] 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.

[32] Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin Drug Discov 2, 1567–1577. https://doi.org/10.1517/17460441.2.12.1567.

[33] OECD, 2007. Guidance documents on the validation of (Quantitative) structure-activity relationships [(Q)SAR] models [WWW Document]. OECD Environ. Heal. Saf. Publ. URL https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2 (accessed 12.22.22).

[34] OECD, 2004. Validation of (Q)SAR Models - OECD [WWW Document]. URL https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed 11.30.22).

[35] Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52, 396–408. https://doi.org/10.1021/CI200520G.

[36] Topliss, J.G., Edwards, R.P., 1979. Chance factors in studies of quantitative structure-activity relationships. J Med Chem 22, 1238–1244. https://doi.org/10.1021/JM00196A017.

[37] Wu, Z., Li, D., Meng, J., Wang, Huiwen, 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.

[38] Burnham, K.P., Anderson, A.D., 2002. Model Selection and Inference: A Practical Information-Theoretic Approach, second ed. Springer-Verlag, New York. https://doi.org/10.1007/b97636.

[39] Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct Relatsh 1, 45–63. https://doi.org/10.4018/IJQSPR.2016010102.

[40] Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 34, 2121–2132. https://doi.org/10.1002/jcc.23361.

[41] Kier, L.B., Hall, L.H., 1990. An electrotopological-state index for atoms in molecules. Pharm Res 7, 801–807. https://doi.org/10.1023/A:1015952613760.

[42] Banerjee, A., Gajewicz-Skretna, A., Roy, K., 2023. A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites. Mol Inform 42, 2200261. https://doi.org/10.1002/minf.202200261.

Environmental Science Processes & Impacts

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Environ. Sci.: Processes Impacts*, 2024, **26**, 105

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

Predictive binary mixture toxicity modeling of fluoroquinolones (FQs) and the projection of toxicity of hypothetical binary FQ mixtures: a combination of 2D-QSAR and machine-learning approaches†

Mainak Chatterjee and Kunal Roy *

Received 12th October 2023
Accepted 21st November 2023

DOI: 10.1039/d3em00445g

rsc.li/espi

All sorts of chemicals get degraded under various environmental stresses, and the degradates coexist with the parent compounds as mixtures in the environment. Antibiotics emerge as an additional concern due to the bioactive nature of both the parent compound and degradation products and their combined exposure to the environment. Therefore, environmental risk assessment of antibiotics and their degradation products is very much necessary. In this direction, we made use of *in silico* new approach methodologies (NAMs) and machine-learning algorithms. In this study, we have developed a robust and predictive mixture-quantitative structure–activity relationship (QSAR) model with promising quality and predictability (internal: MAE_{Train} = 0.085, Q_{LOO}^2 = 0.849, external: MAE_{Test} = 0.090, and Q_{F1}^2 = 0.859) for predicting the toxicity of the mixtures of a class of antibiotics and their degradation products. To obtain the predictive model, toxicity data of 78 binary fluoroquinolone mixtures in *E. coli* (endpoint: log 1/IC50 in molar) have been utilized. We have used only 0D–2D descriptors to efficiently encode the structural features of mixture components without any additional complexities. The optimization of the class of mixture descriptors has been performed in this study by using three different mixing rules (linear combination of molecular contributions, the squared molecular contributions, and the norm of molecular contributions). Different machine-learning approaches namely, random forest (RF), ada boost, gradient boost (GB), extreme gradient boost (XGB), support vector machine (SVM), linear support vector machine (LSVM), and ridge regression (RR) have been employed here apart from the conventional partial least squares (PLS) regression to optimize the modeling approach. A rigorous validation protocol has been used for assessing the goodness-of-fit, robustness, and external predictability of the models. Finally, the toxicity of possible untested mixtures of different photodegradation products of fluoroquinolones has been predicted using the best model reported in this study.

Environmental significance

In this study, we have developed machine learning-based mixture QSAR models from the toxicity data of the binary mixtures of fluoroquinolones in *E. coli*. Fluoroquinolones are one of the major classes of antibiotics that have been used for a long time and found in different environmental compartments along with their degradation products. However, their mixture toxicity is mostly unknown. Thus, an approach of mixture toxicity assessment has been made in light of the *in silico* predictive modeling approach. This study will help us to understand the mixture effect of FQs and their degradation products and will also provide knowledge about the toxicity prediction of untested FQ mixtures.

1. Introduction

Antibiotics are one of the widely used pharmaceuticals that are found in the environment due to improper waste management in households, hospitals, industrial setups, *etc.*[1–3] This widespread presence of antibiotics in the environment is a major cause of environmental pollution and associated biohazards.[1,4] They adversely affect various environmental ecosystems and lower organisms[5,6] and also facilitate the development of bacterial resistance.[7] Besides the parent compounds, transformation products are also causes of great concern. Most of the chemicals get degraded in the environment and make different degradation products; antibiotics are also not exempted from this natural process.[8] On top of this, the antibiotics undergo

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India. E-mail: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3em00445g

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

various transformation reactions during discharge through the wastewater treatment plants in industrial setups, and thus, produce different homologous substances.[9] Therefore, a soup of multiple chemicals is released into the environment instead of a single chemical. If we carefully explore the scientific literature, many articles prove the simultaneous presence of different antibiotics and their degradates in the environment.[10,11] It should not be taken for granted that the degradation products are less active than the parent compounds; they may retain their bioactive nature instead and may produce enhanced activity in some cases.[12,13] The interactive effects such as response addition and synergism play crucial roles in the combined activity, and thus, enhanced toxic action is seen even at a suboptimal concentration of both parent antibiotics and their degradation products.[14,15] Therefore, the assessment of mixture toxicity should be a primary objective in the environmental risk assessment. The major chemical regulatory agencies thoroughly investigate the probable environmental risk of individual antibiotics as well as their mixtures,[16] but the degradation products are mostly overlooked, although they are a potential cause of environmental threat and should be accounted for in the environmental risk assessment process.

The toxicity analysis of degradation products is largely unexplored due to obvious reasons. The degradation products from different structural analogues can be formed throughout the lifecycle of any chemicals, and it is very complex to identify, isolate, and analyse these by experimentation. The assessment of the combined toxicity among different degradation products or among parent compounds and their degradation products is a far more tricky process because there are no actual traces of the number of such combinations and their compositions. The *in silico* new approach methodologies (NAMs) such as quantitative structure–activity relationship (QSAR), read-across predictions, *etc.* can be effectively used in a situation like this where the experimentation is very troublesome. The QSAR models can be developed by correlating the already existing experimental toxicity data with the computationally derived structural and physicochemical descriptors (independent variables of chemicals) and then, can be further used for toxicity prediction. One of the major advantages of such QSAR models over experimentation is the ability to predict the toxicity of an infinite number of combinations without physical formulation; the mixtures must fall under the domain of applicability of the developed model to get a reliable prediction. Both the QSAR and read-across have been employed in many previous studies for the toxicity predictions of individual toxicants[17–19] as well as chemical mixtures.[20–23] Chemical regulatory agencies such as the US-Environmental Protection Agency (US-EPA), European Chemicals Agency (ECHA), Australian Industrial Chemical Introduction Scheme (AICIS), *etc.* and legislations such as the European Union's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) also encourage the use of such computational approaches for environmental risk assessment and data gap filling.[24,25]

Among various antibiotics, fluoroquinolones are widely used, and thus, they are often detected in different environmental compartments (wastewater streams, wastewater receiving rivers, *etc.*) in the ng L^{-1} to mg L^{-1} range.[26,27] Fluoroquinolones (FQs) get degraded during wastewater treatment in industrial setups and different environmental stresses like sunlight and heat can also affect the integrity of the molecules after disposal.[28–30] Although the basic quinolone ring is stable enough to tolerate environmental stresses, the side chains can be easily broken. Therefore, there are high chances of combined effects developed from fluoroquinolone mixtures towards lower environmental species which should be identified. Many research groups have previously worked on fluoroquinolone degradates and their toxicity, but most of the time they got the combined toxic effects without sufficient knowledge about the combination of substances.[8,28,31] Wang *et al.* have published an article in the recent past where they reported an *in silico* toxicity model for predicting the toxicity of photodegradation products of fluoroquinolones.[32] The molecular docking-based energy descriptors have been used to model the mixture toxicity, and they claimed the mixture effects to be mostly additive in the case of fluoroquinolones. However, these descriptors are not simple and reproducible, and there are uncertainties about the use of molecular docking-based descriptors due to the lack of knowledge on the exact mechanism of toxicity mediation. Therefore, there is scope to explore the topic in light of the machine-learning algorithms and to develop more efficient predictive models with simple, reproducible, and easily interpretable descriptors.

In the present work, we have developed well-validated mixture-QSAR models from the existing toxicity data of fluoroquinolone mixtures in *Escherichia coli*.[32] *E. coli* is one of the resident bacteria present in the lower intestine of warm-blooded animals including human beings and it has several beneficial effects.[33,34] Therefore, it is a good marker to assess the detrimental effects of fluoroquinolone mixtures towards gut microbiota. Different machine-learning algorithms have been employed with the conventional partial least squares (PLS) regression to optimize the modeling method. We have also optimized the mixing rule for mixture descriptor calculation in this study. To avoid exhaustive calculations, simple, interpretable, and reproducible 0D–2D descriptors have been used. The models have been validated employing the strict validation protocol suggested by the Organization for Economic Co-operation and Development (OECD).[35] The probable degradation products of the fluoroquinolones have been identified from the literature. Based on this, the hypothetical binary mixtures have been generated from the degradation products, and their parent compounds and their combined toxicities have been predicted using the newly developed QSAR model. Mostly the freely available software tools have been used to model the toxicity which makes this work cost-effective.

2. Materials and methods

2.1. Toxicity data collection

We have collected a toxicity dataset of 78 binary fluoroquinolone mixtures from the recently published article of Wang *et al.*[32] and have used these for toxicity modeling. The log(1/IC50) values (on the molar scale) have been taken as the

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

endpoint of this study where IC50 is defined as the median inhibitory concentration of binary mixtures in *E. coli*. The mixtures were prepared from 13 different fluoroquinolone antibiotics namely ciprofloxacin (CIP), danofloxacin (DAN), enoxacin (ENO), enrofloxacin (ENR), levofloxacin (LEV), marbofloxacin (MAR), moxifloxacin (MOX), norfloxacin (NOR), ofloxacin (OFL), orbifloxacin (ORB), pefloxacin (PEF), sarafloxacin (SAR), and sparfloxacin (SPA). All of these component chemicals have been tabulated along with their 2D structures and experimental toxicity data [log(1/IC50)] in Table S1 of the ESI.† According to the report of Wang *et al.*,[32] equitoxic amounts of component chemicals were mixed to obtain the desired mixtures and then used for generating the toxicity data; we have tabulated the compositions (same as those collected) of these 78 binary mixtures in Table S2 of the ESI.† The IC50s of the studied mixtures were determined by using a logistically fitted concentration–inhibition curve. The growth inhibition of *E. coli* was determined by comparing the optical density of inoculated lysogeny broth (LB) before and after the addition of sample mixtures. The bacterial strain of *E. coli* was inoculated in a sterile LB medium and cultured at 37 °C for toxicity assay. The optical density of the LB medium was tested using a Multiskan™ FC Microplate Photometer (Thermo-Fisher Scientific, USA).

2.2. Mixture descriptor calculation

In any mixture-QSAR model, each mixture is considered as an individual data point represented by the unique structural features that are numerically encoded in the mixture descriptors. Generally, molecular descriptors are defined as "the final result of a logical 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";[36] however, mixture descriptors are different from the individual molecular descriptors. Computation of a mixture descriptor is a two-step process where the molecular descriptors of the mixture components are computed first. In the next step, the mixture descriptors are calculated from the structural information of the component chemicals and their corresponding concentration or mass ratio. Different mixing rules are found in the literature for calculating the mixture descriptors among which a linear combination of molecular contributions is mostly used. In a linear combination, the mixture effect is assumed to be additive; however, the mixture components may also produce supra-additive/sub-additive responses due to biological interactions namely, synergism and antagonism, respectively. Qin *et al.*[37] have enlisted numerous mixing rules apart from the linear combination of molecular contributions among which the norm of molecular contributions[38] and square molecular contributions[39] is most prominent. Therefore, to consider the probable additive, supra-additive, and sub-additive mixture effects, we have employed the linear combination of molar contributions, the square molar contributions, and the norm of molar contributions respectively in the present work. These three mixing rules have been presented in the following eqn (1)–(3):

$$\text{Mixing rule 1} = \text{linear combination of molecular contributions} = \Sigma(p_i \times x_i) \quad (1)$$

$$\text{Mixing rule 2} = \text{square molecular contributions} = (\Sigma(p_i \times x_i))^2 \quad (2)$$

$$\text{Mixing rule 3} = \text{norm of molecular contributions} = \sqrt{\sum (p_i \times x_i)^2} \quad (3)$$

where p_i and x_i are the mole fraction and descriptor values, respectively, of the ith component. The sum of mole fractions of all component chemicals should be 1. We have used only 0D–2D descriptors to make it simple and easily interpretable. Nine classes of 0D–2D descriptors namely constitutional indices, molecular properties, ring descriptors, connectivity index, electro-topochemical atom indices, functional group counts, atom-centred fragments, atom-type E-state indices, and 2D atom pairs (total descriptor count = 2400) have been computed from the AlvaDesc software[40] and accordingly pretreated to omit the redundant descriptors (descriptors with the same values and descriptors with strong inter-correlation). Finally, three classes of pretreated mixture descriptors (363 descriptors in each class) have been obtained by using a Java-based software tool "Mixture-Desc-Calc v1.0" (freely available online: **https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1**).

2.3. Dataset division and optimization of the mixing rule

Ideally, the QSAR models should be developed with the available experimental data and then validated with an external experimental dataset. However, the availability of such experimental data is very limited, and therefore, the complete dataset gets divided into training and test sets to serve the purpose. The training data points are used for model development, whereas the test data points are separately employed for external validation. Here in this study, we have developed QSAR models using 75% of the total data, and the rest have been used to test the developed model (as the test set). There are many reported research studies in the literature where well-validated QSAR models were obtained with 75 : 25 training-test divisions.[21,22,41] Both the response-based (sorted-activity-based division) and descriptor-based (Euclidean distance-based division and Kennard-Stone division) division techniques have been applied to get the optimum division. The division has been performed using the "Dataset Division GUI" version 1.2 tool (freely available online: **http://teqip.jdvu.ac.in/QSAR_Tools/**), and nine different training-test division sets have been obtained (4 seeds of sorted activity-based divisions, 4 seeds of Euclidean distance-based divisions, and 1 Kennard-Stone division) for each descriptor matrix; therefore a total of 27 division sets were generated.

For optimizing the mixing rule and selecting the best division set, we have employed the "Double Cross-Validation v2.0" (DCV) tool[42] (freely available online: **http://teqip.jdvu.ac.in/QSAR_Tools/**) and have developed preliminary genetic algorithm-based multiple linear regression (GA-MLR) models from each of the 27 training sets for the respective mixing rule.

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

Another round of data pretreatment has been carried out (using "dataPreTreatmentTrainTest v1.0" tool; freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/) before starting the preliminary GA-MLR modeling for removing the redundant descriptors generated after division. The preliminary models have been validated in the same tool ("Double Cross-Validation v2.0") and validation metrics such as R^2, $Q_{LOO}{}^2$, $MAE_{Train95\%}$, $Q_{F1}{}^2$, and $MAE_{Test95\%}$ have been calculated. To optimize the best mixing rule, the averages of the above-mentioned metrics have been calculated, and based on the average values, the optimum mixing rule has been selected. To avoid any division bias and to make the criteria more robust, we have used the mean of the validation metrics of the preliminary models obtained from 9 different training-test pairs as the determination criteria of optimum mixing rule selection. After optimization of the mixing rule, we have compared the balance of internal and external validation metrics of different division sets under the optimized mixing rule, and based on that, the best division was selected. The mixing rule-wise computed validation metrics for each division have been tabulated in Table S3 of the ESI†.

2.4. Feature selection and model development

A multi-layered algorithm has been adopted in this study for extracting valuable structural features for modeling. The most suitable descriptor combination has been finalized by comparing the determination coefficient (R^2) and cross-validated correlation coefficient ($Q_{LOO}{}^2$) values of all possible subsets of descriptors. Generally in the regression-based model, a 5 : 1 ratio is maintained between the number of training data points and the number of modeled descriptors; the same rule has been followed here too and the combinations with minimum descriptors have been prioritized. We have generated all possible subsets from the descriptor pool using "BestSubsetSelection v2.1" tool. However, all possible subset selection is computationally very exhaustive and it needs a lot of time to compute the validation metrics of the MLR models developed from the descriptor combinations. Thus, a pool reduction is carried out (from 138 descriptors to 40 descriptors) before all possible subset selections. The initial pool reduction has been commenced by a genetic algorithm (GA) (using the "Genetic algorithm v4.1" tool; available from http://teqip.jdvu.ac.in/QSAR_Tools/) in this study. The important descriptors have been taken into the reduced pool based on the R^2 and $Q_{LOO}{}^2$ values of the incoming descriptor combinations.

After selecting the best descriptor combination, we have performed the final regression analysis by using a PLS algorithm by employing the "PLS_single_Y v1.0" tool (freely available online: http://teqip.jdvu.ac.in/QSAR_Tools/). The PLS is the generalized form of MLR that can handle strongly correlated and colinear data too.[43] In the PLS algorithm, latent variables (LVs) are computed from the descriptors and they take part in the regression analysis, unlike MLR. We have optimized the number of latent variables (LVs) by comparing the calculated *x*-variances after incorporating each incoming LV. Finally, a PLS model has been obtained with the optimum number of LVs from the selected descriptors.

2.5. Machine-learning models

Machine learning (ML) is the science of programming computers so that they can learn from the existing data and use that knowledge to solve future assignments more efficiently.[44] Its applications span across different sectors including regulatory decision-making and *in silico* predictive modeling also. The computers are trained by using the significant patterns of large data against a specific learning algorithm, and then that learning experience is used for predicting the corresponding properties of query compounds. Based on the learning algorithm, ML methods are categorised into three major classes, namely supervised learning, unsupervised learning, and reinforcement. Supervised and unsupervised ML are trained by using labelled and unlabelled data, respectively.[45] Reinforcement is different from these two and it is trained by the feedback (either reward or penalty) of the performed action.[45,46] In the present study, we have applied supervised machine learning by employing algorithms like random forest (RF),[47] ada boost,[48] gradient boosting (GB),[48] extreme gradient boosting (XGB),[49] linear support vector machine (LSVM),[50] support vector machine (SVM),[47] and ridge regression (RR),[48,51] to develop a regression model.

Generally, different models are trained by a specific ML algorithm for a dataset and the best one is selected based on the performance. However, there may be some possibilities for further improvement of the model, and to do so, tuning of hyperparameters is carried out. In this study, we have employed our in-house Python-based tool "Optimization and Cross-validation v1.0" (available online: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1) for tuning the hyperparameters for all the ML algorithms. The GridSearhCV and negative mean squared error have been used as the optimizing algorithm and the scoring function, respectively, in this study. The same training set containing the scaled matrix of previously identified descriptors and endpoint values has been used for hyperparameter optimization. After obtaining the optimized setting, the ML models have been developed using the Python-based "Machine Learning Regression v1.0" tool (available online: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1). For obtaining further details of the ML method used, one may consult the article published by Pore *et al.*[52] The same standardized training and test sets have been used for all ML-based models [standardized/scaled value = (original value − mean of the respective column)/standard deviation of the same column; training set means have been used for the scaling of both training and test sets] as used for the PLS model development. The entire modeling process has been schematically represented in the following Fig. 1.

2.6. Statistical validation

Statistical validation is an inevitable part of any predictive modeling. As per OECD guideline 4, the QSAR model should be a good fit with the training data, sufficiently robust and

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

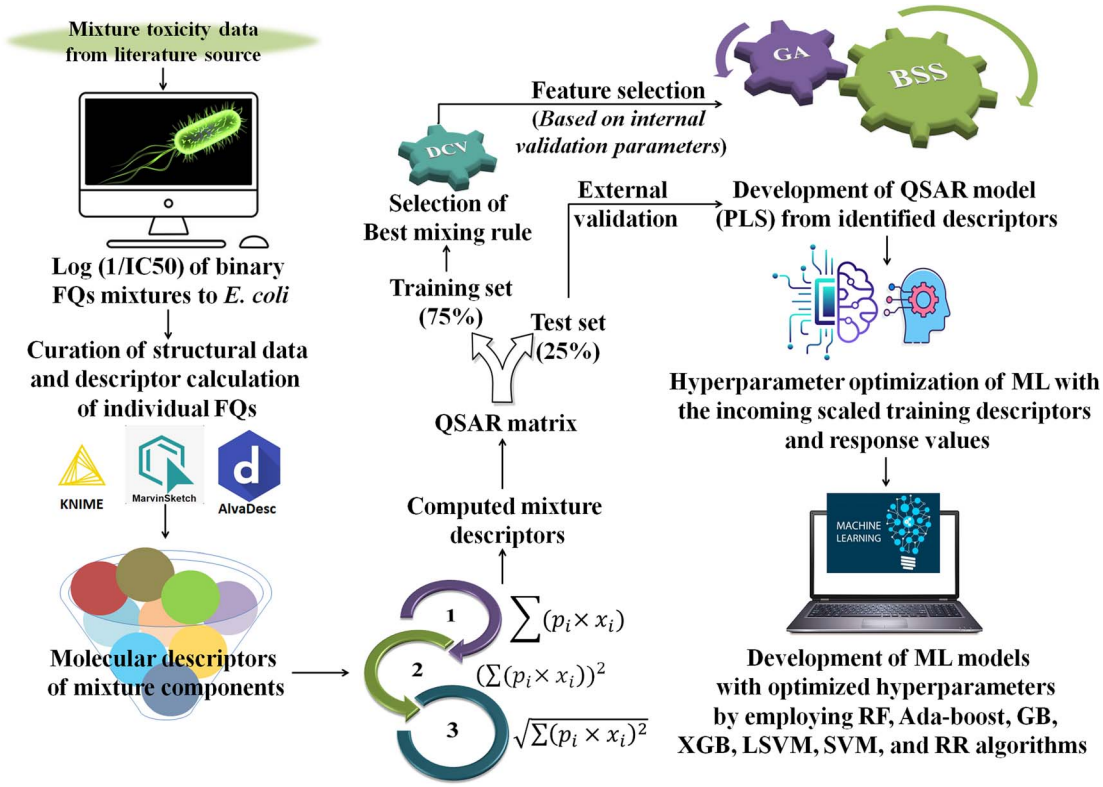


Fig. 1 Schematic representation of the entire modeling process [*FQs = fluoroquinolones, DCV = double cross-validation, GA = genetic algorithm, BSS = best subset selection, ML = machine learning, PLS = partial least squares, RF = random forest, GB = gradient boost, XGB = extreme gradient boost, LSVM = linear support vector machine, SVM = support vector machine (non-linear), and RR = ridge regression].

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 the determination coefficient (R^2), adjusted R^2 (R_{Adj}^2), cross-validated correlation coefficient (Q_{LOO}^2), and mean absolute error of the training set (MAE_{Train}) have been computed from the training data, whereas external validation metrics such as the external correlation coefficient (Q_{F1}^2, Q_{F2}^2, Q_{F3}^2), mean absolute error of the test set (MAE_{train}) and concordance correlation coefficient (CCC) have been computed with the test data. The determination coefficients (R^2 and/or R_{Adj}^2) are the measurement parameters of goodness of fit, mean absolute errors (MAE_{Train} or MAE_{Test}) are the measures of error of predictions, and the external correlation coefficients (Q_{F1}^2, Q_{F2}^2, and Q_{F3}^2) are the measures of the external predictability of the model. The cross-validated correlation coefficient (Q_{LOO}^2) is measured for adjudging the robustness of the statistical model. The difference between Q_{LOO}^2 and R^2 is a marker of robustness and this difference should ideally be a smaller one for a robust model. The threshold limits for external validation metrics ($Q_{F1}^2/Q_{F2}^2/Q_{F3}^2 \geq 0.5$) along with internal validation metrics ($R^2/R_{Adj}^2 \geq 0.6$, $Q_{LOO}^2 \geq 0.5$) have been mentioned as per the criterion of Golbaikh and Tropsha for defining the statistical quality of the regression-based models.[53,54] The concordance correlation coefficient (CCC) measures the deviation between the regression line and the concordance line (line passing through the origin); therefore, it is a measure of both precision and accuracy. Any value of CCC > 0.85 indicates the reliability of the developed model.[55] The important validation metrics which are used in this study have been computed by using the following mathematical eqn (4)–(11):

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum \left(Y_{obs} - \overline{Y}_{train}\right)^2} \quad (4)$$

$$R_{adj}^2 = \frac{\{(n-1) \times R^2\} - p}{n - p - 1} \quad (5)$$

$$Q_{Loo}^2 = 1 - \frac{\sum \left(Y_{obs(train)} - Y_{Pred(train)}\right)^2}{\sum \left(Y_{obs(train)} - \overline{Y}_{train}\right)^2} \quad (6)$$

$$MAE_{Train} = \frac{\sum \left|Y_{Obs(Train)} - Y_{Pred(Train)}\right|}{n_{Train}} \quad (7)$$

$$Q_{F1}^2 = 1 - \frac{\sum \left(Y_{obs(test)} - Y_{Pred(test)}\right)^2}{\sum \left(Y_{obs(test)} - \overline{Y}_{train}\right)^2} \quad (8)$$

$$Q_{F2}^2 = 1 - \frac{\sum \left(Y_{obs(test)} - Y_{Pred(test)}\right)^2}{\sum \left(Y_{obs(test)} - \overline{Y}_{test}\right)^2} \quad (9)$$

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

$$Q_{F3}^2 = 1 - \frac{\left[\sum \left(Y_{obs(test)} - Y_{pred(test)}\right)^2\right] / n_{test}}{\left[\sum \left(Y_{obs(train)} - \overline{Y}_{train}\right)^2\right] / n_{train}} \quad (10)$$

$$MAE_{test} = \frac{\sum \left| Y_{obs(test)} - Y_{Pred(test)} \right|}{n_{test}} \quad (11)$$

In this study, we have also performed the *Y*-randomization test to ensure that the developed QSAR models are not the result of any chance correlation. The 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"[56] 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.

2.7. Toxicity prediction of the binary mixture dataset designed from parent FQs and their degradates: use of the PRI tool

As found in the literature, the chemical species get degraded under open sunlight and other environmental stresses.[8] There are many instances where photodegradation/chemical degradation of FQs has been reported in the literature.[57,58] As we are mostly focused on the environmental aspects of FQs, we have concentrated on the photodegradation products only. Sturini *et al.*,[8] Babić *et al.*,[59] Ye *et al.*,[31] and Ou *et al.*[28] reported different photodegradation pathways of FQs which are graphically represented in Fig. 2 by taking ciprofloxacin as a prototype compound. Reductive halogenation at the 6th position, hydroxy substitution at the 6th position, and hydroxy addition at the 2nd position are the important transformation reactions that may take place under UV light in the quinolone ring whereas, H-elimination, oxidation, and photolysis are the important transformation reactions that may take place at the ring attached at 7th position of quinolone. We have not considered any reaction where the structural integrity of the core quinolone ring of FQ (important for the toxic response of FQs) is compromised. We have designed a hypothetical mixture dataset from nine parent FQs (ciprofloxacin, danofloxacin, enoxacin, enrofloxacin, levofloxacin, marbofloxacin, norfloxacin, pefloxacin, and sarafloxacin) and their corresponding degradation products as the component chemicals of 50 : 50 binary mixtures. All possible combinations of components (45 mixtures) have been considered for each FQ. The composition of all the hypothetical mixtures is tabulated in Table S4 of the ESI.†

The 2D structures of all degradation products have been made in MarvinSketch software with explicit hydrogen addition and proper aromatization and then subjected to the AlvaDesc software for descriptor calculation. The mixture descriptors have been computed from the computed descriptors (modeled descriptors) of individual components by employing the linear combination of molecular contribution in a similar manner as discussed in the previous "mixture descriptor calculation"

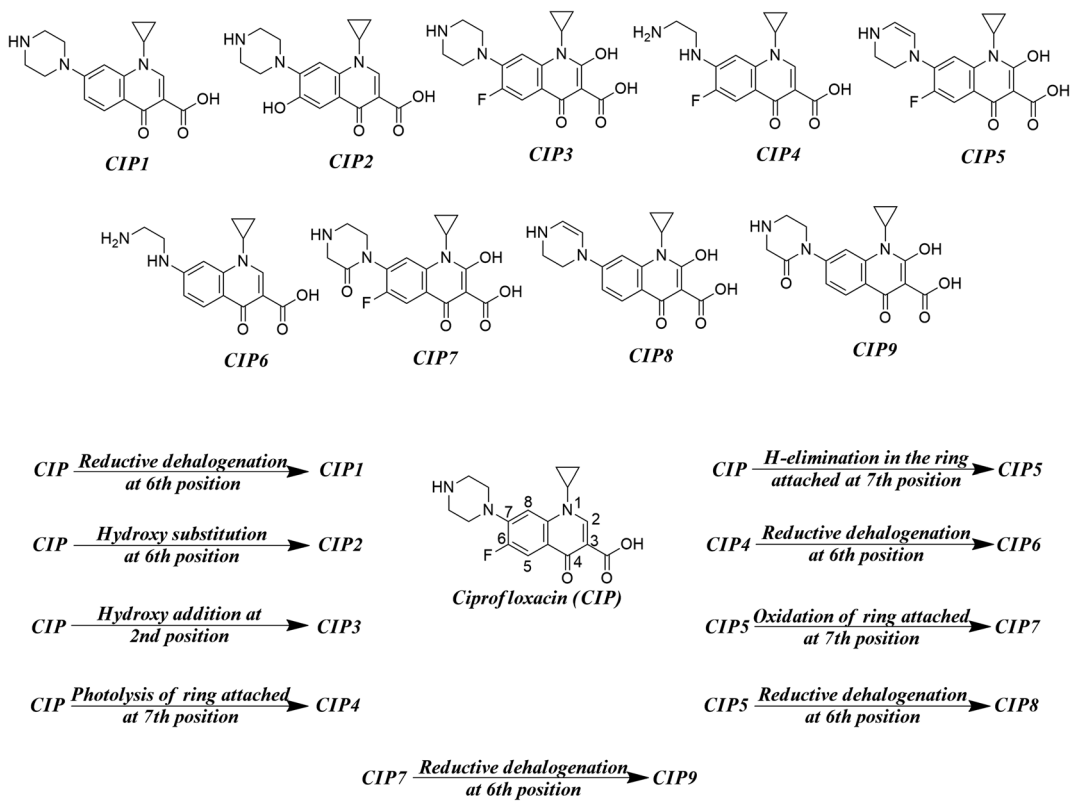


Fig. 2 Different transformation reactions (UV light-mediated) and the probable photodegradation degradation products.

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

section. Finally, the "prediction reliability indicator" (PRI) tool (available online: http://teqip.jdvu.ac.in/QSAR_Tools/) has been employed for the prediction of mixture toxicity and the reliability of the predictions as well. This tool was developed by Roy *et al.*[60] and predicts the toxicity of untested mixtures using the developed model and also categorizes the reliability of the predicted responses into 'good', 'moderate', and 'poor', depending upon the composite score. The reliability of predictions obtained from the PRI tools is essentially based on three sets of principles *i.e.*, "(1) mean absolute error of prediction (leave-one-out) of a query compound from the 10 closest training compounds; (2) applicability domain in terms of similarity, by using the standardization approach; (3) closeness of predicted values to the mean of the observed response of training set compounds". This tool also provides the applicability domain status of the untested mixtures based on the Euclidean distance analysis of query mixtures with the training set mixtures.

3. Results and discussion

3.1. Optimized mixing rule and best division

We have optimized the mixing rule based on the statistical quality of the preliminary GA-MLR models developed with 9 sets of division obtained from the descriptor matrix of each mixing rule. The computed validation metrics (R^2, $Q_{LOO}{}^2$, $MAE_{Train95\%}$, $Q_{F1}{}^2$, and $MAE_{Test95\%}$) of each model have been tabulated in Table S3 of the ESI.† We have taken the average of validation metrics from all the divisions. Among the three mixing rules, the linear combination of molecular contributions has shown the best average metric values ($R^2 = 0.883$, $Q_{LOO}{}^2 = 0.853$, $MAE_{Train95\%} = 0.085$, $Q_{F1}{}^2 = 0.827$, and $MAE_{Test95\%} = 0.091$), and therefore, it has been selected as the optimum mixing rule for mixture descriptor calculation in this study. To select the best division, we have compared the performance metrics of GA-MLR models obtained from mixing rule 1. It has been observed that the 2nd seed of 75 : 25 activity-based division (4-2 ABD) yielded a model with the most balanced performance ($R^2 = 0.874$, $Q_{LOO}{}^2 = 0.842$, $MAE_{Train95\%} = 0.087$, $Q_{F1}{}^2 = 0.837$, and $MAE_{Test95\%} = 0.088$). Thus, we have proceeded to the feature selection step with 4-2 ABD.

3.2. Partial least squares (PLS) regression model and validation

As discussed in the section "2.4. Feature selection and model development", all possible subset selection has been employed to identify the best descriptor combination from the reduced pool of 40 descriptors. Six descriptors namely H-047, Eta_F_A, nR03, F05[C-O], nBnz, and MLOGP2 have been identified as the most prominent descriptors for the prediction of toxicity of fluoroquinolone mixtures. We have moved forward with these six descriptors and developed the QSAR model using the PLS algorithm. To optimise the number of LVs, a PLS regression has been run with the MINITAB software[61] and the *X*-variances of each incoming LV have been noted (see Table S5 of the ESI†). With 5 LVs, the maximum *X*-variance has been extracted (*X*-variance = 0.948) from the descriptor matrix, and therefore, final PLS regression has been performed with 5 LVs. It should be noted that the number of LVs in PLS must be lower than the number of descriptors; otherwise, the PLS model will be converted to an MLR equation. The developed PLS model has been shown in the following eqn (12) along with the computed validation metrics-

$$\text{Log (1/IC50)} = -1.09774 + 0.1092 \times \text{H-047} + 8.10906 \times \text{Eta_F_A} + 1.50499 \times \text{nR03} - 0.11551 \times \text{F05[C-O]} - 0.33295 \times \text{nBnz} - 0.144 \times \text{MLOGP2} \quad (12)$$

$n_{Train} = 59$, $R^2 = 0.882$, $R_{Adj}{}^2 = 0.868$, $Q_{LOO}{}^2 = 0.849$, $MAE_{Train} = 0.085$, $\overline{r_m{}^2(\text{Train, LOO})} = 0.788$, $\Delta r_m^2(\text{Train, LOO}) = 0.123$, $RMSE_C = 0.108$;

$n_{Test} = 19$, $Q_{F1}{}^2 = 0.859$, $Q_{F2}{}^2 = 0.858$, $Q_{F3}{}^2 = 0.873$, $MAE_{Test} = 0.090$, $\overline{r_m{}^2(\text{Test})} = 0.811$, $\Delta r_m^2(\text{Test}) = 0.089$; CCC = 0.927, $RMSE_P = 0.112$.

The developed model is statistically sound and externally predictive as evidenced by the high computed validation metric values (above 0.8) such as R^2, $R_{Adj}{}^2$,$Q_{LOO}{}^2$, $Q_{F1}{}^2$, $Q_{F2}{}^2$, and $Q_{F3}{}^2$. The error metrics of both training and test sets (such as MAE_{Train} and $RMSE_C$ and MAE_{Test} and $RMSE_P$) are on the lower side (below 0.15) which signifies the reliability of predictions. There is a very minute difference (0.033) between the determination coefficient (R^2) and the cross-validated correlation coefficient ($Q_{LOO}{}^2$), which is an indicator of the robustness of the model. The computed values of $r_m{}^2$ metrics are also in the acceptable range. We have computed the CCC of the developed model and its higher value (CCC = 0.927) represents the proximity of the regression line and the concordance line (line passing through the origin of the observed *vs.* predicted scatter plot) of the developed model; therefore, the accuracy and precision of the predicted responses have been acknowledged. We have shown the observed *vs.* predicted scatter plot in the following Fig. 3A where the least scattering and uniform distribution of the training and test data points have been observed surrounding the line passing through the origin. The score plot obtained using the SIMCA-P software indicates the uniform representations of the training data points from all the quadrants of the ellipse (Fig. 3B). We have run the *Y*-randomization test using SIMCA-P software to cut down the possibility of chance correlation. The R^2 and Q^2 intercepts of the *Y*-randomization plot have been computed and the resultant *Y*-randomization plot is given in the following Fig. 3C. These R^2 and Q^2 intercept values ($R^2 = 0.0069$ and $Q^2 = -0.577$) have not crossed the threshold limit which indicates the statistical soundness of the model and affirms that it is not the result of any chance correlation.

To know the contributory effect and order of importance of the modeled descriptors, we have derived the standardized regression coefficients and the VIP scores using SIMCA-P software. The variable importance in projection (VIP) scores are the numerical values which portray the variable importance of individual descriptors towards the modeled endpoint. After arranging the descriptors in descending order of the VIP scores and giving them the number of importance (1 to 6), we plotted

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

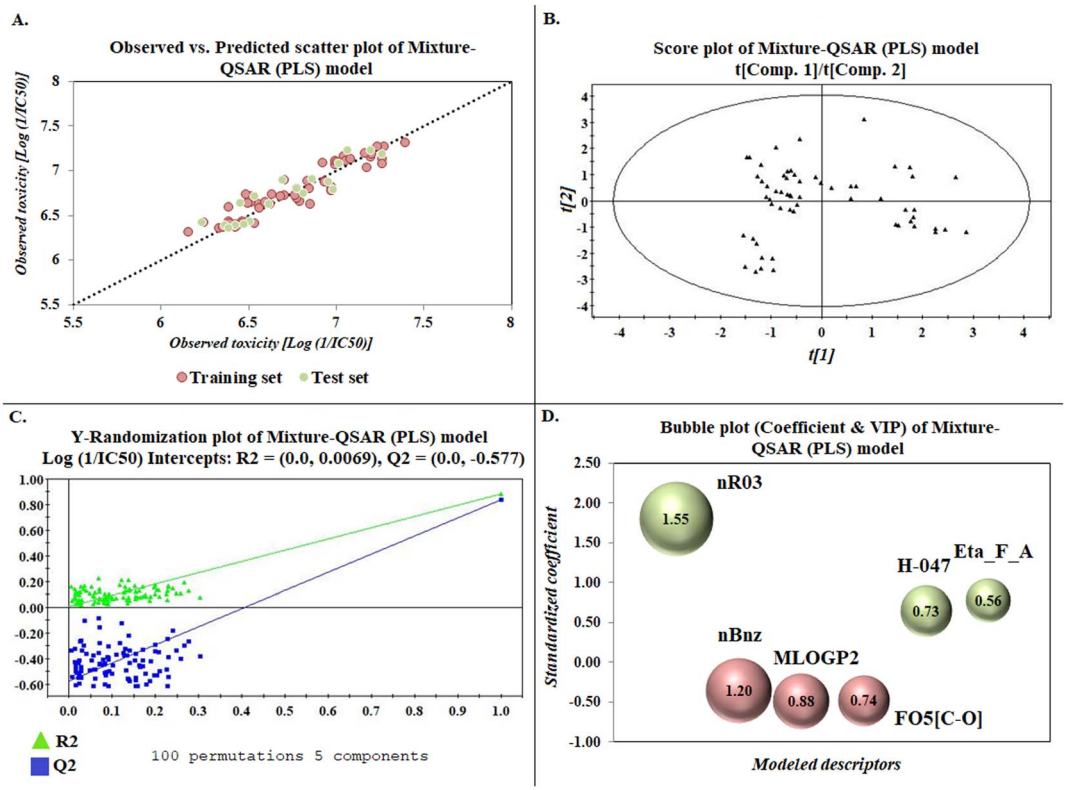


Fig. 3 Graphical representations [(A) observed *vs.* predicted toxicity scatter plot, (B) score plot, (C) *Y*-randomization plot, and (D) bubble plot (coefficient & VIP)].

them against the respective standardized coefficient in a bubble plot (Fig. 3D). The number of importance and the standardized coefficient have been plotted in the bubble plot along the *X*-axis and *Y*-axis, respectively, where the bubbles' radii represent the VIP scores of the respective descriptors. As per the bubble plot, nR03 and nBnz are the most important descriptors with VIP scores > 1. Among the 6 descriptors, nR03, H-047, and Eta_F_A are the positively contributory descriptors (represented by green colour in Fig. 3D), whereas nBnz, MLOGP2, and F05[C-O] are the negatively contributory descriptors (represented by red colour in Fig. 3D).

3.3. Machine learning models and the result of validation

To get the best possible predictions from ML, we first optimized the hyperparameter settings of different ML algorithms using the scaled training set. The optimized settings for different ML approaches are given in Table S6 of the ESI.† Finally, we got the ML-derived predictions using RF, AdaBoost, GB, XGB, SVM, LSVM, and RR algorithms. The validation metrics have been computed and tabulated in the following Table 1. Among these seven different ML algorithms, the model derived by extreme gradient boost (XGB) has the most balanced performance. The high R^2 (0.885) and $Q_{LOO}{}^2$ (0.858) values, and the minute difference between them (0.027) confirm the goodness of fit and robustness, respectively. In contrast, the high value of external correlation coefficients ($Q_{F1}{}^2 = 0.858$, $Q_{F2}{}^2 = 0.857$, and $Q_{F3}{}^2 = 0.872$) indicates the good external predictability of the model. The error metrics such as MAE_{Train} (0.267) and MAE_{Test} (0.279) are also on the lower side. Togo *et al.*, recently published an explainable artificial intelligence (XAI) study for the prediction of developmental toxicity of 585 chemicals, where they reported the XGB-based ML model to be the most efficient and interpretable model.[62] One point may be noted here that the machine-learning algorithms may cause model overfitting, and hence, rigorous cross-validation must be performed. The difference between the determination coefficient (R^2) and cross-validated correlation coefficient ($Q_{LOO}{}^2$) should be checked (the difference should be minimal) with the conventional validation metrics.

Table 1 Computed validation metrics obtained from the predictions of ML models

Metrics	RF	AdaBoost	GB	XGB	SVM	LSVM	RR
R^2	0.786	0.907	0.996	0.885	0.917	0.877	0.881
$Q_{LOO}{}^2$	0.657	0.648	0.619	0.858	0.713	0.844	0.852
MAE_{Train}	0.349	0.240	0.053	0.267	0.198	0.263	0.271
$RMSE_C$	0.458	0.302	0.064	0.336	0.286	0.348	0.343
$Q_{F1}{}^2$	0.772	0.769	0.701	0.858	0.838	0.838	0.847
$Q_{F2}{}^2$	0.770	0.767	0.698	0.857	0.837	0.837	0.845
$Q_{F3}{}^2$	0.794	0.791	0.730	0.872	0.854	0.854	0.862
$MAE_{Test}{}^a$	0.345	0.355	0.391	0.279	0.284	0.287	0.284
CCC	0.859	0.867	0.839	0.925	0.906	0.910	0.918
$RMSE_P{}^a$	0.450	0.453	0.515	0.355	0.379	0.379	0.369

[a] MAE and RMSEp values based on standardized input values.

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

We have compared the XGB-derived ML model and the PLS model based on their overall performance. Both of these models have almost similar goodness of fit, robustness, and external predictability. However, the PLS model has shown the least errors of predictions ($MAE_{Train} = 0.085$ and $MAE_{Test} = 0.090$) as compared to the XGB-derived ML model ($MAE_{Train} = 0.267$ and $MAE_{Test} = 0.279$). Therefore, in terms of the overall performance, the PLS model has been chosen as the best one.

3.4. Insights of modeled descriptors

As we have stated earlier, six 2D descriptors namely, H-047, Eta_F_A, nR03, F05[C-O], nBnz, and MLOGP2 have been identified by using the best subset selection algorithm and then modeled using PLS regression in this study. We have discussed the contributions of each descriptor in the coming section following their order of importance towards the studied endpoint. The structural features defined by the identified descriptors have been showcased in the following Fig. 4.

3.4.1. nR03 descriptor. nR03 is the most important and positive contributing descriptor of this study. It belongs to the class of 'ring descriptor' and represents the number of three-membered ring systems in the molecule. It has been observed that the presence of the 3-membered ring in the mixture component potentiates mixture toxicity. For better understanding, we have compared a set of mixtures with structural representation in Fig. S1 of the ESI.† The most toxic mixture [mixture **M12**: log(1/IC50) = 7.4] of this data set consists of the cyclopropyl ring containing the components ciprofloxacin and sparfloxacin which enhance the mixture toxicity by a considerable amount. In contrast, the mixture of enoxacin and norfloxacin (mixture **M28**) is devoid of the 3-membered ring system, and thus, it is found to be the least toxic (mixture **M28**: log(1/IC50) = 6.16) among all the training data points.

3.4.2. nBnz descriptor. nBnz is the next important descriptor, which indicates the number of benzene or benzene-like ring systems in a molecule. It is a negative contributing descriptor that seems to decrease the mixture toxicity. We have comparatively analysed two mixtures (**M56** and **M22**) with different toxicities [log(1/IC50): **M56** = 7.17; **M22** = 7.27] in Fig. S2 of the ESI.† It has been observed that the comparatively less toxic mixture **M56** contains marbofloxacin (mole fraction = 0.4) and sarafloxacin (mole fraction = 0.6) as the mixture components, and the more toxic mixture **M22** contains danofloxacin (mole fraction = 0.61) and sarafloxacin (mole fraction = 0.39) as the mixture components. Sarafloxacin has a 4-fluorobenzyl substitution on the nitrogen of the central quinolone ring, and it is a constant component of both the mixtures. However, it constitutes 60% of mixture **M56** and 39% of mixture **M22**. Among the variable components, marbofloxacin constitutes 40% of mixture **M56**, and it has a 6-membered fused ring attached to the central quinolone. In contrast, danofloxacin of mixture **M22** constitutes 61% of the mixture and contains a cyclopropyl ring in its structure (cyclopropyl ring helps to increase toxicity). Due to the presence of more benzene-like ring systems attached to the central quinolone ring, less toxic mixture **M56** has a greater nBnz value (1.6) than the more toxic counterpart mixture **M22** (nBnz = 1.39). This comparative analysis clearly justifies the negative contributory effect of the

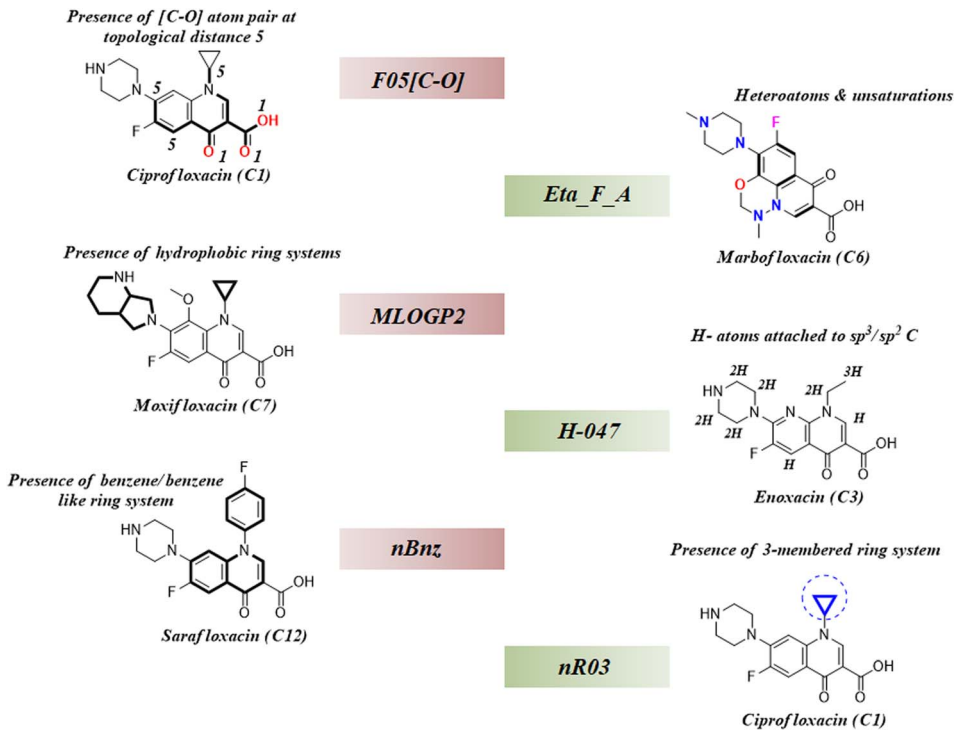


Fig. 4 Structural representation of modelled descriptors (green blocks: positive contributory descriptors and red blocks = negative contributory descriptors).

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

descriptor nBnz and also establishes the positive contributory effect of the descriptor nR03.

3.4.3. MLOGP2 descriptor. The next important descriptor is the molecular property MLOGP2, which negatively affects the endpoint prediction in this study. MLOGP2 is the squared Moriguchi octanol–water partition coefficient which indicates the lipophilicity of chemicals. Generally, the lipophilic chemicals become more toxic due to the easy transportation in the cell through the lipid bi-layers; however, the increase has a limit. Roy *et al.* reported the parabolic relationship between lipophilicity and cellular toxicity where they showed a continuous decrease in toxicity after reaching the peak level with a continuous increase in lipophilicity.[63] Besides this, Wang *et al.* identified the π–π stacking interaction and hydrogen bonding interaction as the most important mechanisms for toxic action of fluoroquinolones in a docking study.[32] Therefore, it has been established that the lipophilicity is not increasing the toxicity here. We have comparatively analysed the studied mixtures also in Fig. S2 of the ESI.† A more lipophilic mixture **M27** (MLOGP2 = 4.014) is found to be less toxic [log(1/IC50) = 6.25] than a comparatively less lipophilic mixture **M05** [MLOGP2 = 2.331; log(1/IC50) = 7.28] which justifies the negative contributory effect of MLOGP2 towards the toxic endpoint of this study.

3.4.4. F05[C-O] descriptor. F05[C-O] is a 2D-atom pair descriptor which represents the frequency of [C-O] atom pairs present at topological distance 5 in a molecule. Its negative regression coefficient indicates an inverse correlation between the toxic endpoint and the descriptor value. We have carefully analysed the components of two different mixtures (**M59** and **M12**) presented in Fig. S2 of the ESI.† It has been observed that the atom pair [C-O] at topological distance 5 mostly represents the central quinolone ring. Therefore, a greater value of this descriptor corresponds to the attachment of extra ring systems to the central quinolone as seen in Fig. S2 of the ESI,† and this structural deviation of the main quinolone ring is detrimental to the toxicity. Mixture **M59** has a greater descriptor value (F05[C-O] = 12.04) and comparatively less toxicity [log(1/IC50) = 6.69] than mixture **M12** with a lower descriptor value (F05[C-O] = 8.0) and comparatively more toxicity [log(1/IC50) = 7.4]. These observations justify the negative contributory effect of F05[C-O] towards the studied endpoint in *E. coli*.

3.4.5. H-047 descriptor. H-047 is the penultimate descriptor as per the descending order of importance (see Fig. 2D). This atom-centered fragment descriptor is a measure of H-atoms attached to the sp^3/sp^2 carbons in a molecule. Its positive regression coefficient indicates the linear correlation between the descriptor value and the toxicity. Mixtures from the studied dataset also align with this trend. For example, a mixture of enoxacin and norfloxacin (mixture **M28**) has a lower H-047 value (11.47) as well as toxicity [log(1/IC50) = 6.16] as compared to the mixture of ciprofloxacin and marbofloxacin (mixture **M05**) which has a greater H-047 value (13.08) and toxicity [log(1/IC50) = 7.28]. In Fig. S1 of the ESI,† we have analysed the structural components of both mixtures **M28** and **M05** for a better understanding of the descriptor.

3.4.6. Eta_F_A descriptor. Eta_F_A is a 1st generation electro-topochemical atom (ETA) index descriptor which positively contributes towards the prediction of the studied endpoint [log(1/IC50)] to *E. coli*. It is a measure of heteroatoms and unsaturation present relative to the overall size of a molecule. According to the published report of Roy *et al.*, Eta_F_A (η_F') is calculated by employing the following formula[64]

$$\eta_F' = \eta_F / N_v$$

where η_F is the functionality index which estimates the unsaturation and heteroatoms of a molecule, whereas N_v is the total number of non-hydrogen atoms present in a molecule. For a better understanding of the descriptor, we have compared two studied mixtures (**M7** and **M57**) in Fig. S1 of the ESI.† Mixture **M7** is composed of ciprofloxacin and norfloxacin as the component chemicals, whereas mixture **M57** contains marbofloxacin and sparfloxacin. The unsaturation count of both these mixtures is the same; however, mixture **M57** bears more heteroatoms (8 N, 3 F, and 1 O) than mixture **M7** (6 N and 2 F). Generally, the presence of electronegative atoms influences the formation of intermolecular H-bonds with the amino acid residues of the cellular receptor/protein and therefore increases the toxic potential.[32] A similar observation has been proven in this study also. The mixture toxicity is increased from lower heteroatom containing mixture **M7** [Eta_F_A = 0.938; log(1/IC50) = 6.34] to greater heteroatom containing mixture **M57** [Eta_F_A = 0.999; log(1/IC50) = 7.24].

3.5. Applicability domain analysis

The applicability domain (AD) is the theoretical region in descriptor space where the model-derived predictions become reliable.[65] The reliable predictions of all chemicals cannot be derived from a single model and that is why the knowledge of AD should be provided with a newly developed model. OECD also suggests the same as a form of OECD principle 3 (a defined domain of applicability).[35] To comply with the norm, we have also analyzed the domain of applicability of the developed PLS-based mixture QSAR model in this study. The DModX (distance to the model in X-space) approach has been employed in SIMCA-P software to derive the AD of the PLS model. We have given the AD plots in Fig. S3 of the ESI.† No outliers have been found (in both training and test sets) at a 99% confidence interval (D-Crit value = 0.00999898) which again strengthens the performance claim of the new model.

3.6. Predicted responses of untested mixtures and the result of reliability analysis

We have predicted the toxicity of 405 hypothetical mixtures (designed from the parent FQs and their probable photodegradation products) with the developed PLS model employing the PRI tool. The predicted responses are tabulated in Table S7 of the ESI† along with their respective reliability of prediction and AD status. The toxicity of 354 mixtures (87.41%), 42 mixtures (10.37%), and 9 mixtures (2.22%) has been predicted respectively with "good", "moderate", and "poor" reliability. Among these mixtures, 86.67% fall under the domain of applicability of the developed PLS model, whereas 13.33% of mixtures have been identified as outliers. This analysis once

 This journal is © The Royal Society of Chemistry 2024

Table 2 Comparative assessment of the performance of the present models (PLS and XGB) with the previous work

Criteria and metrics of comparison	Linear regression model of previous work by Wang *et al.*[32]	PLS-based regression model of the present work	XGB-based regression model of the present work
Dataset size	$n_{Train} = 63$, $n_{Test} = 15$	$n_{Train} = 59$, $n_{Test} = 19$	$n_{Train} = 59$, $n_{Test} = 19$
Modeled descriptors used	Molecular docking-based energy descriptor	2D descriptors	2D descriptors
Type of mixtures	Additive	Additive	Additive
R^2	0.824	0.882	0.885
$Q_{LOO}{}^2$	0.778	0.849	0.858
$RMSE_C$	0.135	0.108	0.106 (unstandardized)
$Q_{F1}{}^2$	0.697	0.859	0.858
$RMSE_P$	0.156	0.112	0.112 (unstandardized)

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

again proved the efficiency of the developed model for predicting the toxicity of untested mixtures. We have also graded the response as highly toxic and less toxic after keeping the observed response of prototype FQ ciprofloxacin [log(1/IC50) = 7.30 mol L^{-1}] as the threshold. According to this grading procedure, 19.75% of mixtures have been classified as highly toxic, whereas the rest have come under the less toxic category.

4. Comparison of models

We have compared the modeling approaches as well as the performance metrics of this study with those in a previously published work by Wang *et al.*[32] The same toxicity dataset was used in both of these studies; however, we have changed the training-test division in the current work. All the models have been developed with 59 training compounds in this study, unlike the 63 training compounds used in the previous work. Besides this, more efficient PLS regression and ML-based modeling approaches have been used in this study instead of the simple linear regression of the previous model. As per the reported validation metrics by Wang *et al.*, the previous linear regression was a robust model with sufficient goodness of fit and moderate external predictability (Table 2). The present study has achieved better performance in all aspects of the previous model. Both the PLS and XGB models have shown better R^2 (PLS = 0.882 ans XGB = 0.885), $Q_{LOO}{}^2$ (PLS = 0.849 and XGB = 0.858), and $Q_{F1}{}^2$ (PLS = 0.859 and XGB = 0.858) values than the previous model ($R^2 = 0.824$, $Q_{LOO}{}^2 = 0.778$ and $Q_{F1}{}^2 =$ 697). The difference between R^2 and $Q_{LOO}{}^2$ (an indicator of robustness) is less in the PLS (0.882 − 0.849 = 0.033) and XGB (0885 − 0.858 = 0.027) models of this study as compared to the linear regression model (0.824 − 0.778 = 0.046) of the previous work. Based on the unstandardized error values of the XGB model, it clearly outperforms the linear regression of the previous work (XGB-unstandardized: $RMSE_C = 0.106$, $RMSE_P =$ 0.112; linear regression: $RMSE_C = 0.135$, and $RMSE_P = 0.156$). The PLS-based regression model also outperforms the previously reported linear regression in both the errors (PLS: $RMSE_C$ = 0.108 and $RMSE_P = 0.112$). Above all, the present models (PLS and ML) have been developed with simple and easily reproducible 0D–2D descriptors which make the model easily transferable as compared to the linear regression model obtained from the molecular docking-based energy descriptors (docking energy descriptor calculation involves non-reproducible and less transferable steps *i.e.*, energy minimization and conformational analysis of the 3D structures of molecules). Finally, as a part of this work, we have also analyzed the effect of the structural features as well as the mole fraction of each component on the toxicity of *E. coli* which helps the reader to graphically understand the additive nature of the studied mixtures.

5. Conclusion

The present work represents the development of QSAR models for toxicity prediction of FQ mixtures in *E. coli* and the use of different linear and non-linear machine learning algorithms for model development. The models were made only with the binary mixtures' toxicity data. The conventional PLS algorithm and the XGB-based ML approach have shown promising efficiency in mixture toxicity prediction. We have successfully applied three different mixing rules to compute the mixture descriptors in this study and the linear combination of molecular contributions (additive effect) has been identified as the best one. Our model outperforms the previous one in terms of goodness of fit, robustness, and external predictivity. The use of 2D-structural descriptors made the model quite simple, reproducible, and easily transferable as compared to its more complicated counterparts. Six different 2D-structural features have been identified which successfully demonstrate the effect of structural features on mixture toxicity of FQs. The presence of a tricyclic ring on the 1st N of quinolone and a greater number of heteroatoms and unsaturation in the mixture component can be detrimental to environmental safety. In contrast, the lipophilicity and the presence of more benzene/benzene-like ring structures in the mixture component can reduce the greater toxic effect. The toxic effect of different photodegradation products on mixture toxicity has also been assessed in this study. We have predicted the toxicity of 405 hypothetical mixtures (binary) which are empirically designed from 9 FQs and their corresponding photodegradation products. Among these 405 mixtures, 19.75% have shown a higher risk of being toxic towards *E. coli*. This result strengthens the claim of efficient prediction of real untested/new mixtures without the physical synthesis using 2D-QSAR and ML-based models. Thus, these *in silico* modeling approaches can be efficiently used in

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

the decision-making of chemical regulatory agencies in the near future. However, this projection cannot cover the entire chemical domain of the FQs and their different degradation products which makes us bound to perform more exhaustive work towards diverse ecological species for better risk assessment. Besides the toxicological perspectives, these *in silico* approaches can be used for predicting the chances of degradation by taking physical parameters such as pK_a/pK_b (dissociation constants) as independent variables. This is another important path which needs to be explored exhaustively in the near future.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Mainak Chatterjee sincerely acknowledges the financial assistance of All India Council for Technical Education (AICTE-New Delhi) as a form of National Doctoral Fellowship.

References

1 K. Kümmerer, Antibiotics in the Aquatic Environment – A Review – Part I, *Chemosphere*, 2009, **75**, 417–434, DOI: 10.1016/J.CHEMOSPHERE.2008.11.086.
2 D. Álvarez-Muñoz, S. Rodríguez-Mozaz, A. L. Maulvault, A. Tediosi, M. Fernández-Tejedor, F. Van den Heuvel, M. Kotterman, A. Marques and D. Barceló, Occurrence of Pharmaceuticals and Endocrine Disrupting Compounds in Macroalgaes, Bivalves, and Fish from Coastal Areas in Europe, *Environ. Res.*, 2015, **143**, 56–64, DOI: 10.1016/j.envres.2015.09.018.
3 W. Xiong, Y. Sun, T. Zhang, X. Ding, Y. Li, M. Wang and Z. Zeng, Antibiotics, Antibiotic Resistance Genes, and Bacterial Community Composition in Fresh Water Aquaculture Environment in China, *Microb. Ecol.*, 2015, **70**, 425–432, DOI: 10.1007/s00248-015-0583-x.
4 K. Kümmerer, The Presence of Pharmaceuticals in the Environment Due to Human Use – Present Knowledge and Future Challenges, *J. Environ. Manage.*, 2009, **90**, 2354–2366, DOI: 10.1016/j.jenvman.2009.01.023.
5 N. Martins, R. Pereira, N. Abrantes, J. Pereira, F. Gonçalves and C. R. Marques, Ecotoxicological effects of ciprofloxacin on freshwater species: data integration and derivation of toxicity thresholds for risk assessment, *Ecotoxicology*, 2012, **21**, 1167–1176, DOI: 10.1007/s10646-012-0871-x.
6 K. Kümmerer, Antibiotics in the Aquatic Environment – A Review – Part II, *Chemosphere*, 2009, **75**, 435–441, DOI: 10.1016/j.chemosphere.2008.12.006.
7 Y. Yang, W. Song, H. Lin, W. Wang, L. Du and W. Xing, Antibiotics and Antibiotic Resistance Genes in Global Lakes: A Review and Meta-Analysis, *Environ. Int.*, 2018, **116**, 60–73, DOI: 10.1016/j.envint.2018.04.011.
8 M. Sturini, A. Speltini, F. Maraschi, L. Pretali, E. N. Ferri and A. Profumo, Sunlight-induced degradation of fluoroquinolones in wastewater effluent: photoproducts identification and toxicity, *Chemosphere*, 2015, **134**, 313–318, DOI: 10.1016/j.chemosphere.2015.04.081.
9 X. Guo, Z. Yan, Y. Zhang, X. Kong, D. Kong, Z. Shan and N. Wang, Removal mechanisms for extremely high-level fluoroquinolone antibiotics in pharmaceutical wastewater treatment plants, *Environ. Sci. Pollut. Res.*, 2017, **24**, 8769–8777, DOI: 10.1007/s11356-017-8587-3.
10 L. Jiang, X. Hu, D. Yin, H. Zhang and Z. Yu, Distribution and seasonal variation of antibiotics in the Huangpu River, Shanghai, China, *Chemosphere*, 2011, **82**, 822–828, DOI: 10.1016/j.chemosphere.2010.11.028.
11 C. Li, J. Chen, J. Wang, Z. Ma, P. Han, Y. Luan and A. Lu, Occurrence of antibiotics in soils and manures from greenhouse vegetable production bases of Beijing, China and an associated risk assessment, *Sci. Total Environ.*, 2015, **521–522**, 101–107, DOI: 10.1016/j.scitotenv.2015.03.070.
12 S. Achermann, V. Bianco, C. B. Mansfeldt, B. Vogler, B. A. Kolvenbach, P. F. X. Corvini and K. Fenner, Biotransformation of sulfonamide antibiotics in activated sludge: The formation of pterin-conjugates leads to sustained risk, *Environ. Sci. Technol.*, 2018, **52**, 6265–6274, DOI: 10.1021/acs.est.7b06716.
13 D. M. Cwiertny, S. A. Snyder, D. Schlenk and E. P. Kolodziej, Environmental designer drugs: When transformation may not eliminate risk, *Environ. Sci. Technol.*, 2014, **48**, 11737–11745, DOI: 10.1021/es503425w.
14 S. Kar and J. Leszczynski, Is intraspecies QSTR model answer to toxicity data gap filling: Ecotoxicity modeling of chemicals to avian species, *Sci. Total Environ.*, 2020, **738**, 139858, DOI: 10.1016/j.scitotenv.2020.139858.
15 X. Zou, Z. Lin, Z. Deng, D. Yin and Y. Zhang, The joint effects of sulfonamides and their potentiator on Photobacterium phosphoreum: differences between the acute and chronic mixture toxicity mechanisms, *Chemosphere*, 2012, **86**, 30–35, DOI: 10.1016/j.chemosphere.2011.08.046.
16 O. Nicolotti, E. Benfenati, A. Carotti, D. Gadaleta, A. Gissi, G. F. Mangiatordi and E. Novellino, REACH and *in silico* methods: an attractive opportunity for medicinal chemists, *Drug Discov. Today*, 2014, **19**, 1757–1768, DOI: 10.1016/j.drudis.2014.06.027.
17 Y. Hao, T. Fan, G. Sun, F. Li, N. Zhang, L. Zhao and R. Zhong, Environmental toxicity risk evaluation of nitroaromatic compounds: machine learning driven binary/multiple classification and design of safe alternatives, *Food Chem. Toxicol.*, 2023, **170**, 113461, DOI: 10.1016/j.fct.2022.113461.
18 S. Chen, G. Sun, T. Fan, F. Li, Y. Xu, N. Zhang, L. Zhao and R. Zhong, Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods, *Sci. Total Environ.*, 2023, **876**, 162736, DOI: 10.1016/j.scitotenv.2023.162736.
19 F. Li, T. Fan, G. Sun, L. Zhao, R. Zhong and Y. Peng, Systematic QSAR and IQCCR modelling of fused/non-fused aromatic hydrocarbons (FNFAHs) carcinogenicity to rodents: reducing unnecessary chemical synthesis and

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

animal testing, *Green Chem.*, 2022, **24**, 5304–5319, DOI: 10.1039/D2GC00986B.

20 M. Chatterjee and K. Roy, 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, DOI: 10.1016/j.chemosphere.2022.136463.

21 M. Chatterjee and K. Roy, "Data fusion" quantitative read-across structure–activity-activity relationships (q-rasaars) for the prediction of toxicities of binary and ternary antibiotic mixtures toward three bacterial species, *J. Hazard. Mater.*, 2023, **459**, 132129, DOI: 10.1016/j.jhazmat.2023.132129.

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

23 P. Kumar, A. Kumar and D. Singh, CORAL: development of a hybrid descriptor based QSTR model to predict the toxicity of dioxins and dioxin-like compounds with correlation intensity index and consensus modelling, *Environ. Toxicol. Pharmacol.*, 2022, **93**, 103893, DOI: 10.1016/j.etap.2022.103893.

24 A. B. Raies and V. B. Bajic, silico toxicology: Computational methods for the prediction of chemical toxicity, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2016, **6**, 147–172, DOI: 10.1002/wcms.1240.

25 S. Klatte, H. C. Schaefer and M. Hempel, Pharmaceuticals in the environment – a short review on options to minimize the exposure of humans, animals and ecosystems, *Sustainable Chem. Pharm.*, 2017, **5**, 61–66, DOI: 10.1016/j.scp.2016.07.001.

26 R. Gothwal and Shashidhar, Occurrence of high levels of fluoroquinolones in aquatic environment due to effluent discharges from bulk drug manufacturers, *J. Hazard., Toxic Radioact. Waste*, 2017, **21**, 05016003, DOI: 10.1061/(ASCE)HZ.2153-5515.0000346.

27 A. S. Oberoi, Y. Jia, H. Zhang, S. K. Khanal and H. Lu, Insights into the fate and removal of antibiotics in engineered biological treatment systems: a critical review, *Environ. Sci. Technol.*, 2019, **53**, 7234–7264, DOI: 10.1021/acs.est.9b01131.

28 H. S. Ou, J. S. Ye, S. Ma, C. H. Wei, N. Y. Gao and J. Z. He, Degradation of ciprofloxacin by UV and UV/H2O2 *via* multiple-wavelength ultraviolet light-emitting diodes: Effectiveness, intermediates and antibacterial activity, *Chem. Eng. J.*, 2016, **289**, 391–401, DOI: 10.1016/j.cej.2016.01.006.

29 M. I. Vasquez, E. Hapeshi, D. Fatta-Kassinos and K. Kümmerer, Biodegradation potential of ofloxacin and its resulting transformation products during photolytic and photocatalytic treatment, *Environ. Sci. Pollut. Res.*, 2013, **20**, 1302–1309, DOI: 10.1007/s11356-012-1096-5.

30 X. Wei, J. Chen, Q. Xie, S. Zhang, L. Ge and X. Qiao, Distinct photolytic mechanisms and products for different dissociation species of ciprofloxacin, *Environ. Sci. Technol.*, 2013, **47**, 4284–4290, DOI: 10.1021/es400425b.

31 J. S. Ye, J. Liu, H. S. Ou and L. L. Wang, Degradation of ciprofloxacin by 280 nm ultraviolet-activated persulfate: Degradation pathway and intermediate impact on proteome of Escherichia coli, *Chemosphere*, 2016, **165**, 311–319, DOI: 10.1016/j.chemosphere.2016.09.031.

32 D. Wang, Q. Ning, J. Dong, B. W. Brooks and J. You, Predicting mixture toxicity and antibiotic resistance of fluoroquinolones and their photodegradation products in Escherichia coli, *Environ. Pollut.*, 2020, **262**, 114275, DOI: 10.1016/j.envpol.2020.114275.

33 J. N. V. Martinson and S. T. Walk, Escherichia coli residency in the gut of healthy human adults, *EcoSal Plus*, 2020, **9**, 10–1128, DOI: 10.1128/ecosalplus.esp-0003-2020.

34 J. N. V. Martinson, N. V. Pinkham, G. W. Peters, H. Cho, J. Heng, M. Rauch, S. C. Broadaway and S. T. Walk, Rethinking gut microbiome residency and the enterobacteriaceae in healthy human adults, *ISME J.*, 2019, **13**, 2306–2318, DOI: 10.1038/s41396-019-0435-7.

35 OECD, *Validation of (Q)SAR Models – OECD*, https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm, accessed 30 September 2023.

36 V. Consonni and R. Todeschini, Molecular descriptors, in *Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics*, ed. T. Puzin, J. Leszczynski and M. Cronin, Springer, 2010, DOI: 10.1007/978-1-4020-9783-6_3.

37 L. T. Qin, Y. H. Chen, X. Zhang, L. Y. Mo, H. H. Zeng and Y. P. Liang, QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide, *Chemosphere*, 2018, **198**, 122–129, DOI: 10.1016/j.chemosphere.2018.01.142.

38 T. Gaudin, P. Rotureau and G. Fayet, Mixture descriptors toward the development of quantitative structure-property relationship models for the flash points of organic mixtures, *Ind. Eng. Chem. Res.*, 2015, **54**, 6596–6604, DOI: 10.1021/acs.iecr.5b01457.

39 D. A. Saldana, L. Starck, P. Mougin and B. Rousseau, Prediction of flash points for fuel mixtures using machine learning and a novel equation, *Energy Fuels*, 2013, **27**, 3811–3820, DOI: 10.1021/ef4005362.

40 A. Mauri, AlvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints, in *Methods in Pharmacology and Toxicology*, ed. K. Roy, Humana Press Inc., New York, 2020, pp. 801–820, DOI: 10.1007/978-1-0716-0150-1_32.

41 S. Ghosh, M. Chatterjee and K. Roy, Predictive quantitative read-across structure–property relationship modeling of the retention time (log tR) of pesticide residues present in foods and vegetables, *J. Agric. Food Chem.*, 2023, **71**, 9538–9548, DOI: 10.1021/acs.jafc.3c01438.

42 K. Roy and P. Ambure, The "Double Cross-Validation" software tool for MLR QSAR model development, *Chemom. Intell. Lab. Syst.*, 2016, **159**, 108–126, DOI: 10.1016/j.chemolab.2016.10.009.

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

43 S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130, DOI: 10.1016/S0169-7439(01)00155-1.
44 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, Applications of machine learning in drug discovery and development, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477, DOI: 10.1038/s41573-019-0024-5.
45 M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 2015, **349**, 255–260, DOI: 10.1126/science.aaa8415.
46 I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, *SN Comput. Sci.*, 2021, **2**, 1–21, DOI: 10.1007/s42979-021-00592-x.
47 R. J. Chase, D. R. Harrison, G. M. Lackmann and A. McGovern, *A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks and Deep Learning*, Weather Forecast, 2023, pp. 1271–1293, DOI: 10.1175/waf-d-22-0187.1.
48 A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Inc., USA, 2019.
49 T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.
50 R. Burbidge, M. Trotter, B. Buxton and S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem.*, 2001, **26**, 5–14, DOI: 10.1016/S0097-8485(01)00094-8.
51 G. C. McDonald, Ridge regression, *Wiley Interdiscip. Rev. Comput. Stat.*, 2009, **1**, 93–100, DOI: 10.1002/wics.14.
52 S. Pore, A. Banerjee and K. Roy, Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells, *Sustainable Energy Fuels*, 2023, **7**, 3412–3431, DOI: 10.1039/d3se00457k.
53 K. Roy, S. Kar and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, 2015, DOI: 10.1016/C2014-0-00286-9.
54 K. Roy and I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, *Comb. Chem. High Throughput Screening*, 2011, **14**, 450–474, DOI: 10.2174/138620711795767893.
55 N. Chirico and P. Gramatica, Real external predictivity of qsar models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.*, 2011, **51**, 2320–2335, DOI: 10.1021/ci200211n.
56 Z. Wu, D. Li, J. Meng and H. Wang, Introduction to SIMCA-P and its application, in *Handbook of Partial Least Squares*, ed. V. Esposito Vinzi, W. Chin, J. Henseler and H. Wang, Springer, Berlin, Heidelberg, 2010, pp. 757–774, DOI: 10.1007/978-3-540-32827-8_33.
57 M. Rusch, A. Spielmeyer, H. Zorn and G. Hamscher, Degradation and transformation of fluoroquinolones by microorganisms with special emphasis on ciprofloxacin, *Appl. Microbiol. Biotechnol.*, 2019, **103**, 6933–6948, DOI: 10.1007/s00253-019-10017-8.
58 H. Zhang and C. H. Huang, Oxidative transformation of fluoroquinolone antibacterial agents and structurally related amines by manganese oxide, *Environ. Sci. Technol.*, 2005, **39**, 4474–4483, DOI: 10.1021/es048166d.
59 S. Babić, M. Periša and I. Škorić, Photolytic degradation of norfloxacin, enrofloxacin and ciprofloxacin in various aqueous media, *Chemosphere*, 2013, **91**, 1635–1642, DOI: 10.1016/j.chemosphere.2012.12.072.
60 K. Roy, P. Ambure and S. Kar, How precise are our quantitative structure–activity relationship derived predictions for new query chemicals?, *ACS Omega*, 2018, **3**, 11392–11406, DOI: 10.1021/acsomega.8b01647.
61 *Data Analysis, Statistical & Process Improvement Tools | Minitab*, 2004, https://www.minitab.com/en-us/, accessed 2023-04-05.
62 M. V. Togo, F. Mastrolorito, F. Ciriaco, D. Trisciuzzi, A. R. Tondo, N. Gambacorta, L. Bellantuono, A. Monaco, F. Leonetti, R. Bellotti, C. D. Altomare, N. Amoroso and O. Nicolotti, TIRESIA: an explainable artificial intelligence platform for predicting developmental toxicity, *J. Chem. Inf. Model.*, 2023, **63**, 56–66, DOI: 10.1021/acs.jcim.2c01126.
63 J. Roy, P. Kumar Ojha, E. Carnesecchi, A. Lombardo, K. Roy and E. Benfenati, First report on a classification-based QSAR model for chemical toxicity to earthworm, *J. Hazard. Mater.*, 2020, **386**, 121660, DOI: 10.1016/j.jhazmat.2019.121660.
64 K. Roy and G. Ghosh, Introduction of extended topochemical atom (ETA) indices in the valance electron mobile (VEM) environment as tools for QSAR/QSPR studies, *Internet Electron. J. Mol. Des.*, 2003, **2**, 599–620.
65 D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti and O. Nicolotti, Applicability domain for QSAR models: Where theory meets reality, *Int. J. Quant. Struct. Relat.*, 2016, **1**, 45–63, DOI: 10.4018/IJQSPR.2016010102.

 This journal is © The Royal Society of Chemistry 2024

Published on 22 November 2023. Downloaded on 1/24/2024 11:12:09 AM.

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

44 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, Applications of machine learning in drug discovery and development, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477, DOI: **10.1038/s41573-019-0024-5**.

45 M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 2015, **349**, 255–260, DOI: **10.1126/science.aaa8415**.

46 I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, *SN Comput. Sci.*, 2021, **2**, 1–21, DOI: **10.1007/s42979-021-00592-x**.

47 R. J. Chase, D. R. Harrison, G. M. Lackmann and A. McGovern, *A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks and Deep Learning*, Weather Forecast, 2023, pp. 1271–1293, DOI: **10.1175/waf-d-22-0187.1**.

48 A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Inc., USA, 2019.

49 T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, DOI: **10.1145/2939672.2939785**.

50 R. Burbidge, M. Trotter, B. Buxton and S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem.*, 2001, **26**, 5–14, DOI: **10.1016/S0097-8485(01)00094-8**.

51 G. C. McDonald, Ridge regression, *Wiley Interdiscip. Rev. Comput. Stat.*, 2009, **1**, 93–100, DOI: **10.1002/wics.14**.

52 S. Pore, A. Banerjee and K. Roy, Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells, *Sustainable Energy Fuels*, 2023, **7**, 3412–3431, DOI: **10.1039/d3se00457k**.

53 K. Roy, S. Kar and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, 2015, DOI: **10.1016/C2014-0-00286-9**.

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

55 N. Chirico and P. Gramatica, Real external predictivity of qsar models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.*, 2011, **51**, 2320–2335, DOI: **10.1021/ci200211n**.

56 Z. Wu, D. Li, J. Meng and H. Wang, Introduction to SIMCA-P and its application, in *Handbook of Partial Least Squares*, ed. V. Esposito Vinzi, W. Chin, J. Henseler and H. Wang, Springer, Berlin, Heidelberg, 2010, pp. 757–774, DOI: **10.1007/978-3-540-32827-8_33**.

57 M. Rusch, A. Spielmeyer, H. Zorn and G. Hamscher, Degradation and transformation of fluoroquinolones by microorganisms with special emphasis on ciprofloxacin, *Appl. Microbiol. Biotechnol.*, 2019, **103**, 6933–6948, DOI: **10.1007/s00253-019-10017-8**.

58 H. Zhang and C. H. Huang, Oxidative transformation of fluoroquinolone antibacterial agents and structurally related amines by manganese oxide, *Environ. Sci. Technol.*, 2005, **39**, 4474–4483, DOI: **10.1021/es048166d**.

59 S. Babić, M. Periša and I. Škorić, Photolytic degradation of norfloxacin, enrofloxacin and ciprofloxacin in various aqueous media, *Chemosphere*, 2013, **91**, 1635–1642, DOI: **10.1016/j.chemosphere.2012.12.072**.

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

61 *Data Analysis, Statistical & Process Improvement Tools | Minitab*, 2004, **https://www.minitab.com/en-us/**, accessed 2023-04-05.

62 M. V. Togo, F. Mastrolorito, F. Ciriaco, D. Trisciuzzi, A. R. Tondo, N. Gambacorta, L. Bellantuono, A. Monaco, F. Leonetti, R. Bellotti, C. D. Altomare, N. Amoroso and O. Nicolotti, TIRESIA: an explainable artificial intelligence platform for predicting developmental toxicity, *J. Chem. Inf. Model.*, 2023, **63**, 56–66, DOI: **10.1021/acs.jcim.2c01126**.

63 J. Roy, P. Kumar Ojha, E. Carnesecchi, A. Lombardo, K. Roy and E. Benfenati, First report on a classification-based QSAR model for chemical toxicity to earthworm, *J. Hazard. Mater.*, 2020, **386**, 121660, DOI: **10.1016/j.jhazmat.2019.121660**.

64 K. Roy and G. Ghosh, Introduction of extended topochemical atom (ETA) indices in the valance electron mobile (VEM) environment as tools for QSAR/QSPR studies, *Internet Electron. J. Mol. Des.*, 2003, **2**, 599–620.

65 D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti and O. Nicolotti, Applicability domain for QSAR models: Where theory meets reality, *Int. J. Quant. Struct. Relat.*, 2016, **1**, 45–63, DOI: **10.4018/IJQSPR.2016010102**.

Manas Chatterjee. 08.11.2023

08.11.2023

This journal is © The Royal Society of Chemistry 2024