QSPR modeling of selected classes of organoleptic agents

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

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

Doelima Bera

(B. Pharm)

Registration No.: **163660** of **2022-2023**

Examination Roll. No.: **M4PHC24007**

Class Roll No.: **002211402016**

Under the Guidance of

DR. KUNAL ROY

Professor

Drug Theoretics & Cheminformatics Laboratory

Department of Pharmaceutical Technology

Jadavpur University

Kolkata – 700032

India

2024

DECLARATION

I hereby declare that as per of my knowledge this thesis contains literature study and original research as a part of my thesis on **"Intelligent Consensus Predictions of the Retention Index of Flavor and Fragrance Compounds Using 2D Descriptors".**

All information in this thesis has been documented and represented in accordance with academic rules and ethical conduct under the guidance of **Dr. Kunal Roy.**

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

NAME: DOELIMA BERA

EXAMINATION ROLL NUMBER: **M4PHC24007**

REGISTRATION NUMBER: **163660** of **2022-2023**

THESIS TITLE: **Intelligent Consensus Predictions of the Retention Index of Flavor and Fragrance Compounds Using 2D Descriptors**

SIGNATURE WITH DATE: Doelima Bera . 28.08.24

CERTIFICATE

Department of Pharmaceutical Technology

Jadavpur University

Kolkata - 700 032

This is to certify that **Mrs. Doelima Bera**, B. Pharm. (School of Pharmaceutical Sciences, Siksha "O" Anushandhan deemed to be University), has carried out the research work on the subject entitled "**Intelligent Consensus Predictions of the Retention Index of Flavor and Fragrance Compounds Using 2D Descriptors**" under my supervision in Drug Theoretics & Cheminformatics Laboratory in the Department of Pharmaceutical Technology of this university. He has incorporated his findings into this thesis of the same title, being submitted by him, in partial fulfillment of the requirements for the degree of Master of Pharmacy of Jadavpur University. He has carried out this research work independently and with proper care and attention to my entire satisfaction.

28.8.24.

Kunal Roy, PhD, FRSC
Professor & Ex-Head
Department of Pharmaceutical Technology,
JADAVPUR UNIVERSITY,
Kolkata 700 032 (INDIA)
EiC: Molecular Diversity (Springer Nature)

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

28/8/24.

(Prof. Dr. Amalesh Samanta)
Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

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

Dipak Laha 28.8.24

(Prof. Dipak Laha)
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata

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

Acknowledgments

I deem it a pleasure and privilege to work under the guidance of Dr. Kunal Roy, Professor, Drug Theoretics & Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata-32. I express my gratitude and regard to my revered mentor for suggesting the subject of this thesis and rendering me his thoughtful suggestions and rational approaches to this thesis work. I am greatly indebted to Dr. Kunal Roy for his valuable guidance throughout the work that enabled me to complete the work. With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, and perpetual assistance of my seniors Joyita Roy, Ankur Kumar, Mainak Chatterjee, Arkaprava Banerjee, Sapna Kamari Panday, Vinay Kumar, Shilpayan Ghosh and Souvik Pore. Apart from this, MD Mobarak Hossain, Akash Chandra and Sagnik Sarkar are my juniors who all have extended their helping hands and friendly cooperation in my work.

I am thankful to the authority of Jadavpur University and Head of the Department Prof. Dr. Amalesh Samanta for providing all the facilities to carry out this work.

A word of thanks to all those people for the incredible help and motivation whose names I have been unable to mention here. Finally, I would like to thank my parents Mr. Dipak Kumar Bera and Mrs. Mousumi Bera for all the love and the incredible support without which my dissertation work would remain incomplete.

Doelima Bera 28.08.24

Doelima Bera

Examination Roll No.: **M4PHC24007**

Department of Pharmaceutical Technology,

Jadavpur University,

Kolkata-700032

Kolkata-700032

Preface

This dissertation is presented for the partial fulfillment of the degree of Master in Pharmacy in Pharmaceutical Chemistry. This research work spans around two years. This present study has been explored the development of the predictive in-silico chemometric models for properties of some organoleptic compounds by the 2D QSAR statistical approach. This predictive approach mainly considers the numerical data of the structural and physiochemical properties of the chemical compounds as the descriptors. Apart from that, for the first work the sweet and bitter taste and for the second work, the retention index was considered as the endpoint value for further in-silico prediction. However, wide-spread use of various chemical compounds in day-to-day life insist that chemical-based industries to reevaluate the toxicity, activity, or property-based studies for those chemical compounds before marketing. An experimental process with manual testing for the same is quite expensive, time-consuming, and needs a lot of hard work. In this scenario, QSAR and CAAD had come into the field as an alternative method of study. This new approach regarding the toxicity, activity, and study of the chemical compounds is not only effective but also eco-friendly. Chemistry plays a significant role in our day-to-day life. The application of chemistry extended to food, pharmaceutical, cosmetics, agriculture, biochemistry, and many other different industrial fields. Some industries deal with its physical features like solubility, partition coefficient, melting point, boiling point, and surface tension for formulation purposes whereas the chemical and therapeutic properties of the chemical were used to mitigate and control a range of disorders and diseases. However, beyond the physical, chemical, and therapeutic properties, the organoleptic properties of the chemical compounds can be used rigorously as a colorant in the cosmetic industries, pigment

production, fragrance, and flavor compounds for enhancing the sensation of taste and smell. In this recent work organoleptic compounds and their properties are investigated. The discussion and estimation of the properties of the organoleptic compounds by a synthetic method is a tedious job rather the in-silico approach has its enhanced acceptability in the industry, regulatory agencies, and different chemical data banks. The numerical collective chemical information from which the QSAR prediction is done is known as feature or descriptors. Now the descriptors that are calculated from the simplest 2-dimensional chemical structure representation are called 2D- descriptors. This 2D descriptor may be calculated from experimental information or theatrical expression. However, the descriptor or features in the QSAR-based prediction play a role as an independent variable, and the corresponding activity, property, and toxicity act as a dependent variable. The recent works are based on the application of the predictive ability of the 2-Dimentional descriptors. Apart from that RASAR is a collective mechanistic approach of QSAR and read-across is used in the later phase of the investigation. The data of the respective investigation was not only used for the predictive model development but also statistically validated with different statistical metrics. however simple 2-D QSAR model and its validation are based on pure statistics but the read across is the concept of similarity. The compounds having similar chemical structure, and biological responses are used for the further external data set prediction and for the well demarcation of the chemical spaces of the predictive investigation. RASAR is a club concept of both the QSAR and read across for and reassured prediction purposes. Thus an Insilco approach can be helpful to determine the initial data investigation and screening of some potential compounds for further synthetic experimentation if it is necessary. The regulatory agencies and data banks demand information about a large number of compounds regarding their property, activity, and toxicity characteristics. In that case, the initial Insilco screening with reassured synthetic experimentation of potential compounds can be a better approach. In our recent study, we have

investigated and developed the Insilco model of the sweet and bitter organoleptic activity of the chemical compounds as well as the estimation of the retention index for the flavor and the fragrance compounds. The properties of the organoleptic compounds have their corresponding significance throughout the food, flavor, beverage, cosmetics, and fragrance industries. Different statistical validation with their core statistical concept was used rigorously to validate the developed model in the real-world scenario. The following studies have been done in this dissertation:

Study 1: The first application of machine learning-based classification read-across structure-property relationship (c-RASPR) modelling for sweet and bitter.

Study 2: Intelligent Consensus Predictions of the Retention Index of Flavour and Fragrance Compounds Using 2D Descriptors.

The accomplished work has been presented in this dissertation in the following segments:

Chapter 1: Introduction

Chapter 2: Present work

Chapter 3: Material and Methods.

Chapter 4: Result and discussion.

Chapter 5: Conclusion

Chapter 6: References

Appendix: Reprint

ABBREVIATIONS

0D QSAR	Zero dimensional QSAR	DModX	Distance to model in the X-space
1D QSAR	One dimensional QSAR	ED	Euclidean distance
2D QSAR	Two-dimensional QSAR	E-state	Electrotopological state
3D QSAR	Three dimensional QSAR	et al.	“et alia” means “and others.”
3Rs	Replacement, Refinement and Reduction	F&F	Fragrance and flavor compounds.
4D QSAR	Four-dimensional QSAR	GA	Genetic algorithm
5D QSAR	Five-dimensional QSAR	HASL	Hypothetical active site lattice
6D QSAR	Six-dimensional QSAR	HBA	Hydrogen bond acceptor
7D QSAR	Seven dimensional QSAR	HBD	Hydrogen bond donor
AD	Applicability domain	HOMO	Highest occupied molecular orbital
ANN	Artificial neural network	ICP	Intelligent Consensus Prediction
API	Active Pharmaceutical Ingredient	KNIME	Konstanz Information Miner
AUC	Area under the curve	kNN	k-nearest neighbor
CADD	Computer-aided drug design	LDA	Linear discriminant analysis
CM	Consensus model	LMO	Leave-many-out
COMBINE	Comparative Binding Energy Analysis	LOG/Log	Logarithm
CoMFA	Comparative molecular field analysis	LOO	Leave-one-out
CoMMA	Comparative Molecular Moment Analysis	LR	Linear Regression
CoMSIA	Comparative molecular similarity indices analysis	LV	Latent variable

df	Degrees of freedom	MAE	Mean absolute error
MLR	Multiple Linear Regression	QSTR	Quantitative structure-toxicity relationship
MCC	Mathew's correlation coefficient	RF	Random forest
ML	Machine learning	RI	Refractive index
PCR	Principal Component Regression	RMSEP	Root mean square error of prediction
PLS	Partial Least Squares	ROC	Receiver operating characteristics
QSAR	Quantitative structure-activity relationship	SAR	Structure-Activity Relationship
QSPR	Quantitative structure-property relationship	SD	Standard deviation

Index

Chapter-1

Introduction

1. Introduction

The greater advancement in the chemical industries and the wide range of chemical applications in our day-to-day lives often indicate the exponential growth of the chemical industries and their market worldwide. [1] However, this huge application of the chemicals comes with the huge responsibility to properly regulate the safety, efficacy, and effective use of the chemical compounds to comply with the environment, human body, and innocent lives. [2] The core motivation behind the application of chemicals and chemical industries is to minimize the hazardous effects of chemicals on common people and simplify their lives as much as possible [3]. However, the chemical compounds have widespread applications for estimating physiochemical properties as well as the significance of the application of therapeutic characteristics [4]. Apart from that organoleptic chemicals and their uses in chemical-based industries like food, pharmaceuticals, cosmetics, and fragrance are raising concern for several regulatory agencies and research and development segments to study the characteristics of a particular chemical compound [5] The applicability study of several chemical compounds in the higher dimension merges the concepts of mathematics, biology, agriculture, physics, and all the fundamental concepts of different discipline [6] This focuses on enhancing the reliability of the concerned research and application. However, the sustainable use of exact data about a chemical compound whether it is activity, property, or toxicity can be used for the modification and change of the structural, physical, chemical, therapeutic, and organoleptic behaviour of a particular chemical compound. While there is a self-diversity of chemistry the core subject often clubs with other fundamental sciences and results such as Biochemistry, and the cheminformatics field to be explored. The logical and more rational concept of mathematics and statistics when hybridized with the concept of chemistry. it generates a class of chemical informative study known as cheminformatics or

chemostatistics that can be further used to study the entire behaviour of a chemical compound [7]. This informative study can help develop an industrial product according to its desirability. The chemical compounds that are widely used throughout the industries. They are either organic compounds or inorganic compounds. Carbon (C) plays a central role in forming any organic compounds. The remaining valences are fulfilled by the hydrogen (H) or formation of carbon-carbon, carbon-nitrogen, carbon-hydrogen, or carbon-oxygen bonds. Apart from that, there can be an inorganic salt form of a compound in the data set to be analyzed. The data source that is generally used to study in-silico prediction uses both inorganic and organic compounds. The rigorous uses of this chemical product (both organic and inorganic) in day-to-day life enable us to study its organoleptic properties. Excessive use of flavouring agents, colorants, and sweeteners can potentially degrade the taste of packaged food, artificial sweeteners, and several masking agents and supplementary food [8]. So in this scenario, an estimation of the self-property of the chemical compounds to regulate auto degradation is quite essential. Excipients used to mask the bitterness of active pharmaceutical ingredients (APIs) are often sweeteners, derived from sugars and starches. Additionally, they are used to enhance patient compliance as taking an unpalatable medication can be difficult. The use of artificial sweeteners has also become popular among patients suffering from diabetes, and metabolic disorders [9]. The use of fragrance and favour (F&F) is widespread in various consumer products. Fragrance compounds create pleasant smells, while favour compounds contribute to taste sensation [10]. Apart from that the flavour and fragrance industries largely depend on the properties of the organic compounds. These compounds have specific structures and activities that determine their sensory effects. They include alcohols, aldehydes, ketone esters, and lactones [11]. Several experimental techniques are generally used to estimate the qualitative standard of a chemical product before marketing. Retention index, elution time, resolution of the chemical compound in the process of quality assurance, and quality control all together can

be considered as some parameters to ensure the qualitative standard of an industrial product The retention time is crucial for formulating new fragrance compounds in the perfume industry. It helps identify the chemical structure of a compound and allows comparison of its retention data across different GC systems [12]. Chromatography is an important tool in various industries for ensuring the production of high-quality products, and it plays a crucial role in quality control. This method involves measuring the retention time or retention index of a compound as it passes through a gas chromatographic column's glass capillary. However, several qualitative parameters are responsible for quality assurance. Another application of the predictive quality suggests to necessary modification of a chemical structure for their desired purity in a chromatographic column depending on the nature and the polarity of the chromatographic column. As a result, the proper identification and accurate classification of the chemical compound is possible. However, more accurate Insilco prediction before a traditional synthetic approach and predefined chemical space helps for chemical categorization of a chemical even before experimentation or synthesis. The property estimation of thousands of compounds whether it is food, pharmaceutical, flavouring agent, masking agent, or fragrance compounds needs to be under study. This investigation somehow helps to maintain the optimal and desired taste, and quality of any chemical compounds that will be marketed as products. For our recent dissertation for the first study, we have used machine learning approaches such as incorporating the concept of RASAR, and for the second work the intelligent consensus prediction using simple QSAR. However, the idea of RASAR gives the view of QSAR while extending its prediction quality using the core concept of read across [13]. The utilization of c-RASPR in this inquiry will revolutionize the concept of QSPR and demonstrate how the fundamental principle of read-across can also be incorporated into a classification-based modeling framework. By choosing the best-fit classification algorithms of ML like RFC, SVC, LC, and LDA, one can predict the model more accurately. To our knowledge, this is the first

c-RASPR work with sweet and bitter compounds. However, the intelligent consensus prediction gives the idea about the aggregate judgment from several PLS models using the very same initial dataset. Thus a robust, reliable prediction and detailed information of applicability domain can be justified by following this methodology. Apart from that an appropriate chemical categorization is one of the significant applications of the ICP (intelligent consensus prediction) methodology [14].

1.1 QSAR (Quantitative Structure-Activity Relationship) as an in-silico chemometric approach

1.1.1 Basic principle

The core concept of QSAR relies on the structural, chemical, and physical information of any chemical compound in terms of numerical entity or descriptors. The response value of the compounds largely depends on this numerical information.

$$Chemical\ Responce = f(Chemical\ attribute) = f(\ Structure, property) \qquad (1.1)$$

The responses of this methodology are considered as either activity, toxicity, or the property of the chemical compounds. However, those are the dependent identity of the equation, and numerical information or descriptors act as the experimental or theoretical entity and the independent one of this process. These quantitative structure-based chemometric studies can be further classified based on the categorization of the responses. Activity-based quantitative structure studies are known as Quantitative Structural Activity Relationship studies (QSAR), property-based quantitative structure-based studies are known as (QSPR), and toxicity-based quantitative structure-based studies are known as Quantitative Structure Toxicity Relationship (QSTR). Apart from that the QSAR-based studies can be regression-based or classification based on the type of response value. The graded responses (True or False or 0/1) are responsible for classification-based statistical model development while the continuous response values are

responsible for regression-based model development. The concept of regression stands on the pillar of determining the correlation between the X variable and the Y variable mentioned above. Here regression or correlation is a term where we determine a mutual relation of a dependent variable based on the previously known variable. This mutual dependency or correlation follows the equation of linearity or straight line. However, if the correlation or regression is estimated between multiple independent variables concerning a singular endpoint or dependent variable then it is called multiple linear regression or MLR.

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

The classification-based model is nothing but again a correlation estimating and QSAR model development approach. The main difference between classification and regression-based approaches is to take a graded response value instead of a continuous value and the process largely follows the concept of linear discriminating analysis or LDA. The graded value of 1 or 0 is logistically discriminated into 2 sets of classes depending on the numerical information or descriptors of the compounds. In accordance with that QSTR studies include mutagenicity, cellular toxicity, developmental toxicity, and carcinogenicity, while QSPR includes Partition coefficient, permeability, melting point, boiling point, vapour pressure, refractive index, and retention index. Apart from that QSAR includes several biological activities of the chemical compounds like anticancer activity, antibacterial activity, and anti-malaria. Quantitative structure-based studies are principally statistical studies which also include the validation segments for further validating the model based on the different statistical parameters.

1.1.2 History of QSAR

With the advancement in chemical science, QSAR as a non-traditional synthetic approach has bloomed in the field of science. [15]. discovered the inverse proportionality relationship between water solubility and toxicity of chemicals. It was noticed that the toxic potential of

alcohol was increased in mammals due to the reduced toxicity. Later in 1868, Crum Brown and Fraser [16] discovered that different chemical and structural elements have their impact on different physiochemical properties. Again in 1890, Hans Horst Meyer noticed that the toxic potential of a chemical was largely affected by the lipophilicity of the organic compound [17]. However, after that, the impact of lipophilicity concerning biological activity was studied. Louis Hammett [18] estimates the relationship between the electronic characteristics of any acid or base impacting their reactivity and equilibrium. This is the fundamental step for the development of the mechanistic approach of QSAR. Apart from that, it also gives the idea of how the numerical information or chemical features can be correlated and influence the chemical, biological, or physiochemical responses of a chemical compound. Later in 1962 Corwin H. Hansch and co-workers [19] formally introduced the concept of QSAR. The fundamental studies regarding QSAR were to study the structure-activity relationship (SAR) of verities of natural products and pesticides and its dependence on the Hammett constant [20] as well as lipophilicity. Fundamentally Free-Wilson model [21] is a simplistic approach to quantitatively describe Structure Activity Relationship (SAR). It describes the differentiability between two compounds based on the presence or absence of any functional groups. It is a mathematical expression that gives a correlated relationship between different physiochemical descriptors and the response value according to the response value following Hansch law [22] Both of the models are interconnected both theoretically and practically. In different studies, both of the concepts combined to estimate the contribution of different structural influences as well as different physiochemical responses according to the Wilson free crick model [22]. Some dissimilarities found in the free Wilson model have been recently established and found suitable to apply in fragment-based drug design. The concept of QSAR gradually progresses following two methodologies.

1. The data set progresses from classical to non-classical QSAR [23]. While the initial QSAR investigation fundamentally dealt with generally short and congeneric compounds (Which have generally similar mechanisms of action) the progress of QSAR methodology insisted it towards developing a predictive model with a diverse set of chemical compounds (having diverse mechanisms of action) with a bigger size of data sets.
2. Evolution of the study of chemical compounds concerning the structure-activity relationship (SAR) of the compound and employing the analytical study of a compound regarding SAR to target a biological receptor with a chemical compound. QSAR developed more precisely using the same structural activity relation against verities of the receptor.

1.1.3 Core QSAR and its Objectives

1. For optimization of lead chemical compound according to the necessity and desirability.
2. Do the chemical categorization of the chemical compound based on the chemical space.
3. Find out a more reliable and potent chemical compound with the least toxicity.
4. Understand the mechanism of action of any chemical compound and select the less toxic compound accordingly.
5. Predict the Activity/Property/Toxicity of the desired chemical compounds before the synthetic approach

   In the QSAR study, the data of a large no of the chemical compounds are collected. Mechanisms of actions, toxicity, activity, and properties of the chemical compounds are vividly analyzed and further processed for Quantitative structural studies. Thus the QSAR Insilco study can give the forum to study and analyse the lead chemical compounds or desired chemical compounds before processing for a synthetic approach.

1.1.4 Molecular descriptors

Molecular descriptors are the integrated structural information of a chemical compound presented in a numerical form. However, the biological responses (Activity/ property/ Toxicity) of any chemical compound can be defined as the function of the structural or chemical features of chemical compounds [24] The concept of QSAR study relies on the concept of similarity of a defined chemical space. The chemical compounds that exist within this defined range can be further applicable to the developed predictive model. The importance of a defined chemical space not only limits up to that but also allows a new molecule to be predicted by the developed chemical space. However, the chemical space of a predictive model fundamentally depends on the numerical entity of the structural or chemical information of the chemical compounds or descriptors.

$$Biological\ Responce\ (Activity, Property\ or\ Toxicity) = f(Chemical\ structure\ or\ property\ information\ or\ descriptor) \quad (1.3)$$

The nature of the descriptors as the numerical information of the structural attributes plays a significant role in a predictive biological response. The descriptors may be structural (dependent on the occurrence frequency of a substructure), Functional group count descriptors (dependent on the number of functional groups present in a chemical compound), Geometric (dependent on the calculation of the molecular surface area), Physiochemical (electronic, steric and hydrophobic), topological or simple indicator variable (replicated parameters), electronic (based on the calculation of molecular orbital)[25] The significance of a particular descriptor can be estimated according to the correlation of the descriptor concerning the response value. The most significant descriptor to develop a QSAR model can be estimated by considering some characteristics of the features like:

1. The descriptor should have easily interpretable characteristics. However, the physiochemical interpretation of a chemical compound depends on its structural attributes but in a few cases temperature or surrounding environments can be responsible for exceptional responses.
2. A descriptor should be highly correlated with the respective endpoint along with a minor dependency on other descriptors. In the case of descriptor dependency, the major contributors were taken for predictive model development and the minor contributor or the dependent descriptors were removed along with the process of pre-treatment.
3. The descriptor should have covered the largest area of the chemical space or should have the largest domain of applicability.
4. The descriptor should be able to represent the minor structural change of a chemical compound and detect a minor error for the slightly structurally diverse compounds.
5. The descriptors should be easily calculated without depending on the experimental value. A numerical feature is a characteristic of an Insilco approach. The logic and statistics behind the descriptor computation should have its ultimate role rather than a dependency upon an experimental value.

1.1.4.1 Categorisation of Descriptors

A descriptor is a numerical entity of the structural information of a chemical compound. A feature or descriptor for a chemical compound can be classified in the following manner [25] such as physiochemical descriptor (electronic, steric, and hydrophobic), structural (based on the occurrence of a functional group, or sub-structural part), electronic (based on the calculation of each molecular orbital), geometrical (based on the calculation of molecular surface area), topological, or simple indicator variable (replicated parameter). However, the descriptors can widely be classified into 1. Whole molecular descriptor and 2. Substituent constant.

1. Whole molecular descriptor: Extension of the substituent constant method.
2. Substituent constants: Physiochemical descriptors that are established depending on the physiochemical property of the chemical compound.

1.1.4.2 2D descriptors

1.1.4.2.1 Physiochemical descriptor

These are the numerical entities that are responsible for informing about the physiochemical attributes of chemical compounds. The physiochemical alteration of a compound can greatly impact the pharmacokinetic parameters of a chemical compound in any biological system which includes absorption, distribution, metabolism, and excretion. Other than that the electronic phenomenon, steric influence, partition coefficient, and structural and functional group attributes of any chemical compound have a significant role in changing the biological response against the system.

1.1.4.2.1.1 Partition coefficient

The relative affinity of a molecule in a polar medium or a non-polar medium is important. The solubility of a drug molecule in a biological system in the presence of several biological membranes decides its potential to work in the biological system or its pharmacokinetic property. Other than that partition coefficient indicates the polarity of a particular compound for rigorous analysis in the process of quality assurance before marketing the product of interest. The generalization and representation of partition coefficient is done by logarithmic partition coefficient (log P) between n-Octanol and water.

$$P = [C]_{Octanol}/[C]_{aquous} \tag{1.4}$$

The $[C]_{Octanol}$ indicates the concentration of a compound in the lipid or non-polar phase whereas $[C]_{Aquous}$ indicates the concentration of a compound in the polar medium. The P > 1

indicates the concentration of the compound is greater in the non-polar medium or the compound is nonpolar in nature. The value of P<1 indicates the concentration of the compound is greater in the polar medium. Thus the chemical compound is polar in nature. The polarity of a compound is a key regulating authority behind its pharmacokinetic effectivity. It is estimated by distribution in a biphasic medium whether it is liquid-liquid (partition coefficient) or solid-liquid (the polarity of a compound with respect to the chromatographic stationary phase). The descriptor that describes the lipophilic parameter as log p was calculated by Ghosh and Crippen's parameter [26].

1.1.4.2.1.2 Hydrophobic substitution constant (π)

Hydrophobicity is a phenomenon of non-polar compound exerted in the aqueous medium. The tendency of aqueous solution to discard the non-polar compound by not participating in the solvation process is the core concept of hydrophobicity. The relativity of hydrophobicity regarding any particular compound with the hydrophobic substituent called π. π as the value of substituent X can be described

$$logP_{R-X} = logP_{R-H}\,\pi_X \tag{1.5}$$

$\log P_{R-X}$ and $\log P_{R-H}$ represent the partition coefficient of substituted and unsubstituted compounds respectively. The π_X is the difference between the lipophilicity of the substituted compound and the unsubstituted compound. The substitution can be described as the replacement of "H" in "RH" by the substitute 'X".

1.1.4.2.1.3 Hammett electronic constant (σ)

The electronic constant can be further classified into two different types σ_m and σ_p. The electronic effects were studied for the meta and para position rather than the position. The electronic effect in the ortho position is not considered for further studies because of the steric

effect at the ortho position with respect to the origin of the substitute. It is described in the following equation

$$\log k_X = \rho\sigma + \log k_h \quad (1.6)$$

In this equation k_x and k_h are the reaction rate constant for substitution x and h respectively. The term σ is a constant and ρ represents the analogue being studied. However, a positive value indicates the electron-withdrawing effect and a negative value denotes the electron-donating impact.

1.1.4.2.1.4 Steric parameter

The steric parameter or steric effect of a compound is often related to the higher degree of molecular weight or bulkiness. Compounds of the homologous series often show different biological activity. However, the steric activity resists intermolecular reactions rather it positively contributes to the intramolecular reaction. The quantitative indication of the steric influence of a compound is estimated by several steric parameters.

1.1.4.2.1.4.1 STERIMOL parameters

Verloop and coworkers [27] developed a multipara metric method to characterize the steric influence of a substituent in more complex biological systems to go beyond the Taft parameter employed for simple homogenous organic reactions. Verloop and their coworkers developed a collection of five descriptors (L, B1, B2, B3, B4) to describe the shape or structural phenomenon of the substituent (Verloop, 1987). L representative descriptor of the length of the substituent along the axis of the bond between the first atom of the substituent and the parent molecule. B1-B4 all of these descriptors are the width representator. However, this descriptor is all orthogonal to the length denoted or L and forms a 90-degree angle with each other. The huge number of descriptors needed to categorize the substituted elements and the huge number of compounds should be including those parameters in the final QSAR model. This finally

results in thinning of the descriptors to L, B1, and B4 where B1 has the smallest and B5 has the highest width parameter which does not have any directional relationship with L [27]

1.1.4.2.1.4.2 Molar refractivity (MR)

The molecular refractive index is a molar volume adjusted by the refractive index parameter [28]. The molecular refractive index gives the idea of the size and polarity of a compound.

$$MR = (n^2\text{-}1)/(n^2\text{+}2)\text{X (MW)/d} \quad (1.7)$$

Where n denotes refractive index, MW denotes molecular weight and d denotes the density.

1.1.4.2.2. Topological descriptors

Topological descriptors mainly depend on the graphical representation of structural phenomena. So they do not depend on the physiochemical properties or a computational result to be showcased as quantum chemical descriptors. A topological descriptor is all about the graphical representation from the 2D topological information which is the information about the existing atoms and their adjacent bonds.

1.1.4.2.2.1 Wiener index (W)

It is the collective information about the bonds present between each heavy atom that exists in a molecule. However, in graph-theoretical terms, it can be elaborated as the summation of the minimal path length between each pair of heavy molecular atoms represented in the graph. It can be determined as follows:

$$W = 1/2\sum_i\sum_j \delta_{ij} \quad (1.8)$$

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

1.1.4.2.2.2 Zagreb index (Zagreb)

It is represented as the summation of the square of the vertex degree δ_i^2 [29].

$$zagreb = \sum_i \delta_i^2 \tag{1.9}$$

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

1.1.4.2.2.3 Balaban index (J)

The balaban index is followed by the following equation

$$J = M/(\mu + 1)\sum_{edges}(\delta_i \delta_j)^{-0.5} \tag{1.10}$$

Where M represents the no of edges, μ represents the cyclometric number, δ_i , δ_j are the vertex distance degree of the adjacent vertices. This index is calculated from the matrix of the molecular graph.

1.1.4.2.2.4 Molecular Connectivity Indices

Molecular connectivity indices can be calculated employing the atomic vertex degree in H suppressed molecular graph. This is presented as the geek symbol χ (chi).

1.1.4.2.2.4.1 Randict connectivity index

This is also called as branching index or connectivity index. This is also the very first introduced connectivity index. The following equation expressed it [30].

$$\chi R = 1_\chi = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij(\delta_i \delta_j)^{-0.5}} \tag{1.11}$$

Here 'n' represents the total no of vertices present in the molecular graph, a_{ij} is the adjacency matrix element, δ_i , δ_j denotes vertex degree, and the no of other vertices joined with the vertex i and j respectively. The element $(\delta_i\ \delta_j)^{-0.5}$ can be applied for each pair of adjacent edges or vertices of the first order and is termed edge connectivity. Apart from that it can be applied for

more than two adjacent vertices. This connectivity phenomenon is mainly related to molecular branching.

1.1.4.2.2.4.2 Kier and Hall's connectivity index

It is based on Randic's principle, developed by a general concept for calculating zero-order and higher-order connectivity descriptors. Kier and Hall's connectivity index is also named as molecular connectivity [31]. The following equations are responsible for describing zero-order, first-order, and higher-order connectivity expressions.

$$\chi^0 = \sum_{i=1}^{n} \delta_i^{-0.5} \tag{1.12}$$

$$\chi^1 = \sum_{b=1}^{B} (\delta_i \ \delta_j)_b^{-0.5} \tag{1.13}$$

$$\chi^2 = \sum_{K=1}^{2P} (\delta_{i.} \ \delta_{l.} \delta_j)_k^{-0.5} \tag{1.14}$$

$$^{m}_{t}\chi = \ \sum_{K=1}^{K} (\prod_{i=1}^{n} \delta_i)\ _K^{-0.5} \tag{1.15}$$

The last equation shows a generalised equation for the higher-order indices where k runs over m^{th} order subgraphs containing n vertices and B edges. The total no of appearing m-th order is K." χ" represents the product of simple vertex degrees (δ). Theχ_t represents the continuous type of specific subgraph. The term 2_P defines the 2^{nd} order index. 2 χ denotes a path length of 2 containing 3 vertices. Likewise, for higher order, it will be m_P added with the specific graph fragment type t.

1.1.5 Classification QSAR analysis

The chemometric QSAR study can be further subdivided depending on the endpoint (graded or continuous numerical endpoint), and type of the dimension (based on the 2-dimensional or 3-dimensional descriptor). Moreover, based on the types of biological responses (Activity/

Toxicity/ Physiochemical Properties), the classification-based model can be subdivided into QASR / QSTR or QSPR model. However, considering the endpoints like adsorption, distribution, metabolism and excretion like pharmacokinetic parameters can also be taken as biological endpoints. Apart from that the dimensionality of the predictive variables (0D, 1D, 2D, 3D) can be the preliminary criteria to further categorize the classification-based QSAR model.

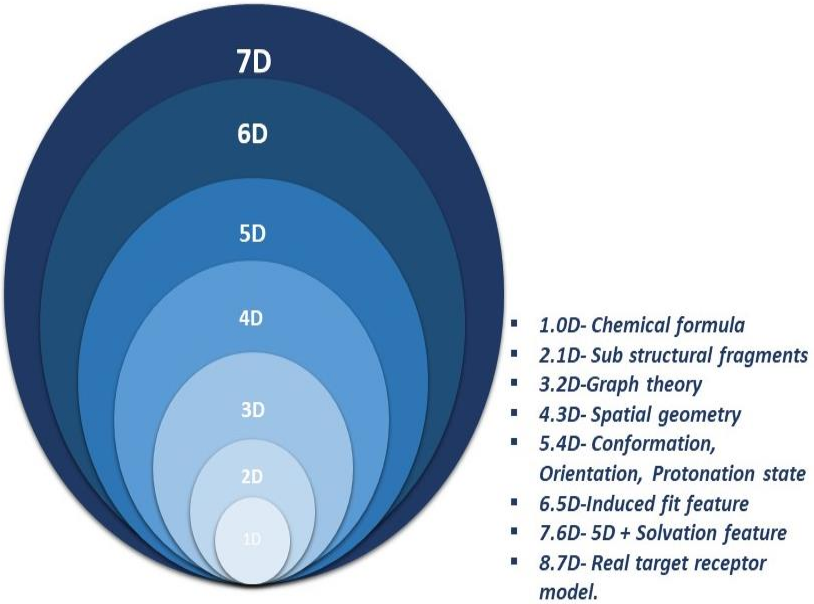


Fig.1.1 Classification of QSAR based on dimensionality

Fig.1.1 represents several QSAR methods classified based on dimensionality. Apart from that many authors have also reported QSAR studies based on the chemical nature of the molecules employed for modeling.

1.1.5.1 Classification based on the type of employed methods

Classification-based QSAR can be subdivided into the following types such as, Linear method (Linear regression), MLR (multiple linear regression), Partial least square (PLS) and (PCA/PCR) Principle component analysis or regression, and some nonlinear methods (Artificial linear network (ANN), k-nearest neighbour (kNN) and Bayesian neural network [32]

1.1.6 Classical QSAR model

1.1.6.1 Thermodynamic approach of Hansch analysis

Hansch first published the mutual dependency and correlation between biological responses and phenoxyacetic acid as well as and Hammett substituent constant and the partition coefficient [33] 1962. Hansch's analysis can apply to linear, nonlinear, and multiple linear analysis. So Hansch's analysis mainly focuses on the establishment of the property relationship. All parameters of Hansch are mainly linear free energy-related values (derived from the rate constant or equilibrium constant). The linear free energy-related approach [34] is also named as Hansch analysis and can be described as follows

$log\frac{1}{c} = k_1$ (partition parameter)+k_2 (electronic parameter)+k_3 (steric parameter)+k_4 (1.16)

Where c is the minimum effective dose responsible for any biological action. k_1, k_2, k_3, k_4 are the constant term. Hansch's model was again modified by the application of bilinear and parabolic terms extended by the term log p.

1.1.6.2 Additivity model or free Wilson analysis

The free Wilson model [35] is the true structure-activity relationship model." Mathematical model", "De novo approach" and "additivity model" all of three are often used synonymously to describe the free Wilson model. A pointer or denote variable is created for every substituted structure that exists in the parent moiety. The resultant correlation coefficient or regression coefficient represents the biological activity contributed by the corresponding elements. The free Wilson model can be elaborated by the following equation

$$BA = \sum a_j X_{ij} + \mu \quad (1.17)$$

Where BA represents the biological response,X_j is the j th substituent which is considered with a value of 1, and being absent the value is regarded as 0. a_j is the contribution of the jth substitution of the biological activity. μ is the overall biological activity of the parent moiety.

1.1.6.3. Fujita ban analysis

Fujita Ban had worked with the application part of the free Wilson's model [36]. The biological activity is here represented in the logarithmic scale. It is also a free energy-related term and fundamentally additive in nature.

$$\log A/A_0 = \sum G_i X_i \tag{1.18}$$

Here A and A_0 are the magnitude of the activity regarding substituted and unsubstituted entities. G_i is the contributed activity expressed in the logarithmic scale for the ith substitution corresponding to the substituent present in the parent molecule (denoted as H). X_i is considered with the value 1 when it is present as substituent otherwise the value is taken as 0 when it is absent.

1.1.7 Brief description of 3D QSAR methods

The 3D QSAR descriptors are comparatively more complex than the simple 2-dimensional descriptors. The calculation of mathematical 3D descriptors involves several steps. Initially, the conformation of the molecular entity is done from the structural or molecular mechanics or the available experimental data and then filtered by minimizing the energy level [37]. Thereafter after the available conformers of the data set were uniformly aligned in the space. Then the space containing the conformers was exposed to different types of descriptors. Apart from that many more independent molecular alignments have also been developed.

1.1.7.1 CoMFA

It is also known as comparative molecular field analysis. The application of CoMFA mainly focuses on the electrostatic (columbic) steric (Van der Wall) energy expressed by the molecule of interest. The aligned molecule is positioned in the 3D grid. At each point of the grid, a probe atom with unit charge is placed and the subsequent potential (Coulomb and Lennard Jones) of the energy field is determined. Then the resultants act as mathematical descriptors and are mainly used for the application of the PLS (partial least square model) based regression model. This study enables us to determine the positive and negative substitutional impact on the activity of the molecule of interest. Now CoMFA is introduced as the part of 3D QSAR approach (Podlogar and Ferguson, 2000). The application of the CoMFA method is generally expressed in the software "Sybyl software" (https://mgm.ku.edu/molecular-modeling-tutorial) from Tripos Inc.

1.1.7.2 CoMSIA

The comparative molecular similarity indices (CoMSIA) are identical to CoMFA as a part of the 3D molecular descriptor. The atom probing technique of CoMSIA is similar to CoMFA. In the Gaussian type function, in CoMSIA, molecular similarity indices are computed from the improved SEAL similarity field and used as descriptors to consider electrostatic, steric, hydrogen bonding, and hydrophobic properties. CoMSIA considers that the probe atom has a radius of 1 A^0, charge of +1, and hydrophobicity of +1 are positioned at the intersection of the surrounding lattice. Moreover, the application of the Gaussian function over the Lenard-Jones and Columbic function enables the gathering of perfect information in the grid points placed in the molecule.

1.1.7.3 SOMFA

It is also known as self–organizing molecular field analysis (SOMFA) [38]. This method has also resemblances with CoMFA and CoMSIA. Apart from that hypothetical Active Site Lattice (HASL) introduced by Doweyko et al (Doweyko, 1988) has a conceptual similarity with SOMFA. The mean-centered activity is decisive in SOMFA.

1.11.7.4 MFA

The mechanistic approach of MFA is to quantify the energy of interaction between a probe and a set of aligned [39]. This study is effective for the analysis of the data sets in which the activity information is present but the receptor or structure of the aligned molecule is unknown. The study of MFA tries to make a hypothesis and characterize the significant features of the receptor site from the common energy level and molecular features that bind to it.

1.1.7.5 GRID

The concept of GRID resembles the CoMFA and it was the first suitable method designed and developed for medicinal scientists as the substitution of the original CoMFA method. The mechanistic approach of this method determines the energy of interaction fields in molecular field analysis and calculates the suitable energetically binding sites on a known molecular structure [40].

1.1.7.6 VFA

It is also known as Voronoi Field Analysis (VFA) [41], voronary polyhedral is formed by the division of a superimposed set of molecules into subspace. For each Voronoi polyhedral there is a single atomic reference point. A cuboid with six tangential sides is divided into a three-dimensional (3D) lattice with a space of 0.3 A^0, neighbouring the union volume of the

superimposed set of molecules is built. The potential and electrostatic energy indices at each lattice point are calculated following the hard-sphere potential model and Coulomb's law.

1.1.7.7. RSA

It is also known as Receptor Surface Analysis which is a suitable method in conditions where the receptor's 3D structure is not known, [42] since one can create the receptor site's imaginary model. The RSA study focuses on capturing essential information about the receptor, unlike pharmacophore. The former captures information about the resemblance of molecules that bind to a receptor.

1.1.7.8 MQSM

It is also known as Molecular Similarity Measures (MQSM) are computed by the integration of volume between the corresponding density function (DF) of the two compared objects, weighted by the non-differential positive definite operator, known as Quantum Similarity Operator [43].

1.1.7.9.1 Alignment independent methods

The effectivity and significance of alignment-independent descriptors are greater because they offer 3D descriptors that are constant to molecule rotation and transformation in space. The study suggests there is no requirement for the superposition of the molecule.

1.1.7.9.1 CoMMA

The Comparative Molecular Moment Analysis (CoMMA) [44] enables second-order moment of charge and mass distributions. The moments correlated to dipole as well as mass centre. The CoMMA descriptors comprise principal quadrupole moment magnitudes of dipole moment and principal moments of inertia. Moreover, descriptors correlating charge to mass distribution

are described, i.e., the magnitude of the projection of dipole upon principle moments of inertia and displacement between centre of mass and centre of the dipole.

1.1.7.9.2 WHIM

The weighted Holistic Invariant Molecular (WHIM) [45] and Molecular Surface [45] descriptors afford the unaltered information using the Principle Component Analysis (PCA) on the cantered coordinates of the atoms constituting the molecule. This changes the molecule into the space that captures the most alteration. In this space, various statistics are computed and act as directional descriptors, containing proportion, variance, kurtosis, and symmetry. By merging the directional descriptors, non-directional descriptors are also described. The atoms can be weighted by mass, atomic electronegativity, atomic polarizability, van-der-walls volume, Kier and Hall's eletrotopological index, and electrostatic potential of a molecule.

1.1.7.9.3 VolSurf

The VolSurf [46] method depends on probing the grid around a molecule with specific probes, for instance, hydrogen bond donor and acceptor groups or hydrophobic interactions. The resultant lattice boxes are employed to calculate the descriptors depending on surfaces or volumes of 3D contours, described by the same probe molecular interaction energy value. By applying different probes and cut-off energy values, various molecular properties can be measured.

1.1.7.9.4 Compass

Compass was developed by Jain and co-workers [47] and is dissimilar from other alignment-independent methods in the respect that it automatically selects alignments and conformations of molecules. In the compass, every molecule is signified by a dissimilar set of feature values.

1.1.8 Receptor-based 3-D QSAR

Receptor-based methods were implemented after the crystal structure of a receptor was available. Protein or receptor-based approaches depend on the information extracted from the structure from the X-ray crystallographic and homology protein structures.

1.1.8.1 Molecular docking

It is a study of how two or more molecular structure ligands or active chemical compounds, drug molecules, and receptors or enzymes of protein bind together [48]. The capacity of interaction of a protein with small molecules performs a key role in protein dynamics which may modify the biological activity. The capacity of large molecules such as nucleic acids and proteins to bind and produce supra-molecular complexes plays a significant part in regulating biological activity. The capacity of large molecules such as nucleic acid and proteins to bind and to produce supra molecular complex plays an important part in regulating biological function. The behaviour of small molecules in binding pockets of target proteins can be defined by molecular docking. The docking methodology aims to recognize the exact poses of ligands in the binding pocket of a protein and to forecast the affinity between the ligand and the protein molecules. Molecular docking can be categorized as (a) protein-nucleic acid docking (b) protein-small molecule docking and (3) protein-protein docking. Protein-ligand docking signifies a simpler end of the complexity spectrum and there are several programs available that can be executed to predict molecules that may potentially prevent proteins. Protein-protein docking is usually much more complicated. The cause is proteins are flexible and their conformational space is fairly huge.

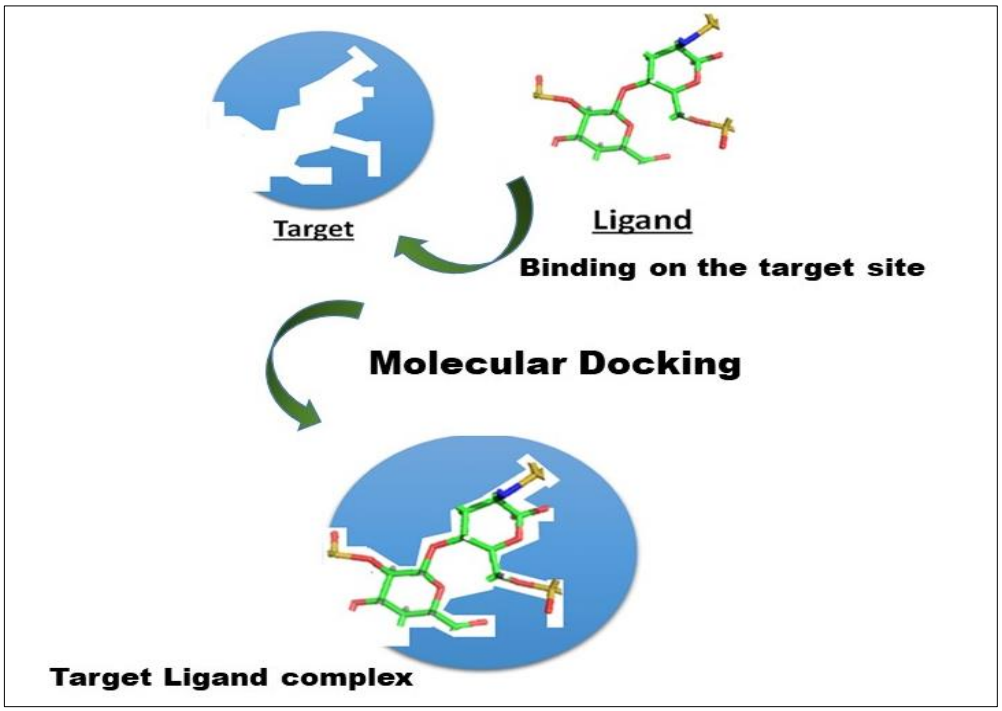


Fig.1.2 Mechanics of Molecular Docking

Docking can be studied by positioning rigid fragments or molecules into the active site of protein employing several methods like geometric hashing, pose clustering, clique search, etc. The performing ability of docking is dependent on the search algorithms (like Genetic algorithms, Monte Carlo methods, Tabu searches, Distance geometry methods, Fragment-based methods, etc.) and the scoring function (i.e., Empirical free energy scoring functions, Knowledge-based potential of mean force or Force field method). First, the constitution of all probable conformations and orientation of the protein binds with the ligand. The scoring function receives input and yields a number that shows favourable interaction. The most vital use of docking software is a virtual screening of the most promising and interesting molecules that are chosen from an available database of auxiliary investigation [48].

1.1.9 Methodology of QSAR

There are four fundamental steps of QSAR analysis includes – (1) Data preparation, (2) Data processing, (3) Data validation, and (4) Data interpretation (Roy et al., 2015). The step can be briefly described by the following:

1.1.9.1 Data preparation

Initially, to maintain the uniformity of data, the endpoint is transformed into the obligatory unit (micromolar or millimolar). Then the chemical structures are drawn by employing several popular software like Marvin Sketch, Chem Sketch, Chem Draw, etc or the structures can be downloaded from online public databases like PubChem, ChemSpider, etc. The energy minimization and conformational analysis are done if necessary. Next, the file containing the structures is employed for descriptor calculation and then the data pre-treatment can be performed to eliminate noisy data, constants, etc. Finally, the descriptors comprise dissimilar variables and a single worksheet which is called a QSAR data matrix. An extra column representing the name or serial numbers of the molecules can be included for fast and easy identification of any molecule or compound.

1.1.9.2. Data processing

a. Data division

A robust, well-predicted, and overall validated QSAR model generation is the core objective of a QSAR study. In that context a proper division of the dataset into a training set (employed to develop a model) and a test set (employed for validation of the developed model). Apart from that the most comprehensible technique to select a training set is dependent on an important physiochemical descriptor or a cluster of chemical similarity. A large number of compounds are selected for the training set which is employed in model development. Generally, it is the ratio of 80:20 for considering the chemical compounds as the part of train

and test set respectively. The fundamental algorithm is based on the principle that a structurally similar molecule to the training set molecules can be predicted confidently because the model has learned the features that are shared by the training set molecules and is capable of searching them in the new compound. The selection of the training set molecules and test set will be in such a way that the test set compounds will fall within the structural domain of the training set molecules. The structural alteration in the test set compounds will result in below-quality prediction and generation of outliers. Different types of data division procedures can be applicable to divide the data set into training and test sets like the Kennard Stone method, Activity / Property-based division, Principle Component Analysis (PCA), Kohonen's Self Organizing Map (SOM), D-optimal design, Sphere exclusion, etc [49].

b. Feature selection

A feature selection process can also be named as a dimensionality reduction procedure because it reduces the feature space of the dataset to the more reliable and significant descriptor. The process follows to directly eliminating the noise and non-significant input features [50] which helps for enhanced interpretability in QSAR modelling as well as the predictive capability of the model [51]. Several feature selection algorithms can be integrated with one or more model development approaches under a similar interface so that the best possible combination of descriptors can able to develop a robust and quality predictive model. Several feature selection methods employed in the QSAR study include stepwise variable selection, Genetic Algorithm (GA), Best Subset Selection (BSS), Variable Subset Selection, Factor analysis, and Most Discriminating Feature selection (MDF). Generally, few are noticeably interested in the endpoint or response. However, descriptors being inter-correlated have negative influences on a QSAR study. A fundamental requirement of several statistical techniques is that the number of data points data points should be higher than the number of descriptors/variables.

c. Model development

This process indicates that the best-selected structural features are to be collected in a single model using an explicit formalism. However, after completing the descriptor calculation the rest of the QSAR study was done by the feature mapping method. The core objective of the QSAR study is to establish a correlated mathematical equation between the descriptors and the response or endpoint being studied. Different techniques like Multiple Linear Regression (MLR), Partial Least Squares (PLS), etc. are applied to develop regression-based models. However, Linear Discriminating Analysis (LDA) is employed for the development of a classification-based model. The feature selection process is done by statistical assessment of the resultant QSAR model and the above-mentioned feature selection procedures were employed to conduct the process. Finally, the best model is selected based on quality prediction and various validation metrics [52].

1.1.9.3 Model validation

The robustness, quality prediction, and statistical significance of the QSAR models are determined depending on the quality of models, as demonstrated by different globally accepted internal and external validation metrics. The developed model for the corresponding endpoint values is validated, utilizing several internal and external validation metrics. The training set is validated using the validation criteria and the responsible validation metrics like the determination coefficient (R^2), leave one out (LOO), cross-validation (Q^2_{LOO}), $r^2_{m(train)}$,$\Delta r^2_{m(train)}$(Roy and Mitra, 2011), root mean square error of calibration (RMSEC), standard deviation (SD) of 100% data of training set, mean absolute error at 5% high residual data points ($MAE_{train\ 95\%}$). The test set predictions are evaluated by several external statistical metrics like R^2_{Pred},Q^2_{F1}, Q^2_{F2}, $r^2_{m(test)}$(i.e $r^2_{m(test0}$, $\Delta r^2_{m(test)}$, standard deviation (SD) of 100%

of data of the test set, root means square error of prediction (RMSEP), 95% mean absolute error of the test set ($MAE_{95\%}$), concordance correlation coefficient (CCC), etc. [53].

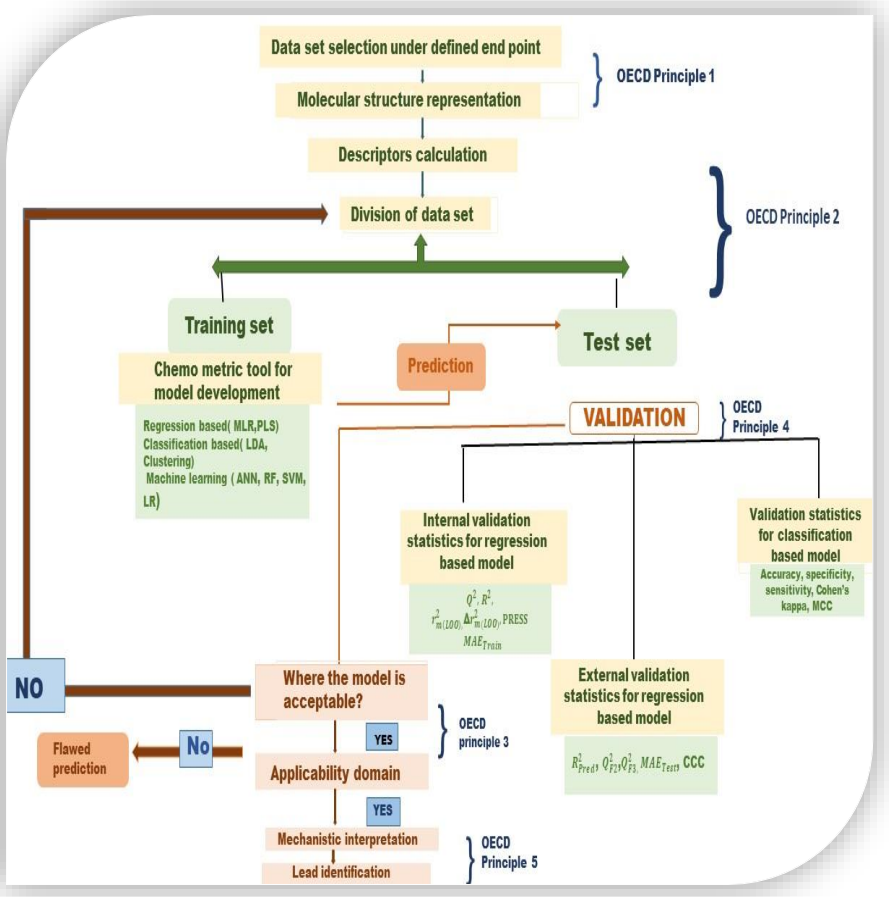


Fig 1.3 General workflow of QSAR

1.1.9.4 Model interpretation

The QSAR study enables the molecular features to be interpreted rigorously. The correlation relationship between the structural attributes and the corresponding response variable contributes to understanding the mechanism of action. Subsequently collecting the observation and experimental results from the developed and validated model indicates the behavioural characteristic of molecules of interest. This information is significant for the further modification of the structural attributes of the molecule of interest to achieve the expected goal.

1.1.10 Application of QSAR Studies

QSAR is an in-silico approach that is effectively used to monitor the activity/ property/toxicity of chemicals while combining the chemical as well as statistical concepts. The consecutive behavioural interpretation of the molecule of study and a fine-tuning with its corresponding biological response can be significantly applicable to a large set of chemical compounds such as (1) Pharmaceuticals, (2) Food and Nutraceuticals, (3) Flavour and Fragrance compounds, (4) Analytical reagents, (5) Solvent, (6) Cosmetic product (7) Surface modifying agents, (8) Toxins, Xenobiotic and different biological products, (9) Agricultural products. Apart from modelling biological activity and toxicity endpoints, the applicability of QSAR spread for ADME study involves the pharmacokinetic profile of potential drug candidates before its synthesis as well as efficacy in the biological system.

1.1.11 QSAR and OECD

The respective OECD encourages the application of QSAR modelling by the financial assistance of the European Union (EU) with the core objective of enriching QSAR as the tool for risk assessment of the compound of interest. The member countries of the OECD have implemented a collective protocol to employ its real use in the ethical background. The OECD QSAR venture the QSAR toolbox (https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm) the validation principles of evolved models and regulation article objects to advance the application of QSAR modelling by industry and governments to simplify the assessment of chemical hazards. In the 1990s the OECD investigated the impact of QSAR modelling for the assessment of aquatic hazards and pollution caused by some contamination of chemicals in a workshop. This is the progression after the employment of SAR models on the exposure assessment of biodegradability and environment-friendly properties of the chemical compounds by the member countries of the OECD. The next major

discussion was held on the Regulatory use of QSAR for Human Health and Environmental Endpoints. Again there was a discussion about Chemicals, Pesticides, and Biotechnology. Finally, it leads to the conclusion to develop an obvious method to evaluate and rigorously validate QSAR models for the constitution of a clear base for their further application.

The OECD decided on the following five principles to enable the regulatory application of QSAR modelling:

Principle 1: About a defined endpoint: The endpoints/ responses modelled in the current study applied to three different data sets. A definite endpoint means a biological, physiochemical, and therapeutic response as an endpoint. Both the continuous and graded values are considered as a definite endpoint.

Principle 2: About an unambiguous algorithm: Different computational statics based on different algorithms were used to compute different classes of descriptors and successive QSAR model development employing particular software tools.

Principle 3: A defined domain of applicability: The applicability domain (AD) for all the statistically relevant developed models. The implementation of the applicability domain is to select the outliers of the definite prediction and further chemical categorization based on the prediction.

Principle 4: Appropriate measures of goodness of fit, robustness, and predictive ability. Several validation statistics and statistical plots are used to thoroughly validate the quality of the prediction and ensure the goodness fit and robustness of the model.

Principle 5: A mechanistic interpretation, if possible: In our present work all the descriptors responsible for the developed model were recognized, correlated with the corresponding endpoint and the mathematical relation is established. However, it is helpful to interpret the structural attributes as well as the physiochemical attributes of molecules of interest with the

respective endpoints. Apart from that molecular interpretation not only indicates the necessary optimization mechanism to achieve the desired goal but also gives the idea of chemical characterization.

1.1.12. Read-Across

Read across is an Insilco chemometric method but while we categorize it in detail it comes under a non-statistical algorithm. It is predominantly based on the similarity whether it is structural, chemical, or biological activity based on the defined kernel (Euclidean, Gaussian, and Laplacian kernel) based similarity [54]. Initially, we considered 10 number of close source compounds. Based on the similarity pattern they gave their prediction opinion for a particular estimated compound. The resulting outputs are generally taken as weighted average prediction value, weighted average standard deviation, and weighted average standard error.

1.1.13 RASAR

RASAR is an amalgamated method of both the QSAR and read-across. While QSAR is a statistical method read across showcased as a non-statistical method. For the development of the RASAR model, the descriptors are calculated from the training set into two segments as 2D QSAR descriptors in supervised form and the read across based measured from unsupervised form[55]. After that, the two types of descriptors are clubbed and further proceeds for the RASAR model development. The prediction of the RASAR model is carried out following the same mechanistic approach as read across depending on the predictive opinion of 10 close source compounds. The validation parameters are considered in terms of R^2 , Q^2 mimicking the validation process of simple QSAR. That's how RASAR stands as a hybrid approach of both QSAR and read-across. The above process is applicable for q-RASAR model development. However, the c-RASAR model can be used only using RASAR

descriptors. For this scenario accuracy, AUC, MCC, sensitivity, and specificity can be considered as the validation statistics.

1.1.14 ML model development

Machine Learning is an artificial intelligence-derived computational program, which was used here for the purpose, of enhancing the accuracy and prediction quality from the previous c-RASAR model [56].

1.1.14.1 Random forest- It is a supervised machine learning algorithm based on some decision tree. The protagonist's role in this decision tree is to decide the best-fit rule to classify the input data based on the features. In the hierarchical arrangement of the decision tree data crosses through each event and each event has some probability. However, after completion of the whole process, the total probability of that event should be 1. The hierarchical nodes present on the decision tree are the root node (does not have any incoming branch), the internal node (has one incoming and two or more outgoing branches), terminal branch (one incoming and one outgoing branch). In this ensemble method, the decision trees in the forest are protagonists. The final decision taken on the majority voting came out from each node. Terminal ends of the nodes are connected to the target and non-terminal nodes are the descriptors. Each tree is constructed with a training set that has compressed size from the original data by random replacement of the original descriptors. Now the new capsized data set is being trained. The remaining descriptors are used for external validation or error detection.

1.1.14.2 Support vector machine- SVM is a labeled or supervised machine learning algorithm. It tries to analyze different classes of the compound constructing a hyperplane. The advantage of SVM is the efficiency of this algorithm can be shown in a higher dimensional data set where no of descriptors is more than the no of samples. But for an input where the no of descriptors is much more than the no of observations, the SVM failed to show a good result.

Thus from the above algorithm, it can be expected SVM will show an enhancement in accuracy when the algorithm is used for such a data set where the no of observations is much more than the no of descriptors. The algorithm of SVM is specialized to differentiate between the class of compounds that perfectly suit for the classification-based model development.

1.1.14.3 Logistic Regression-It is a statistical classification-based model that measures the correlation between the categorical dependent variable and one or more than that independent variable but the classes or category is one for this case. It does not necessarily have to linear relationship between the dependent and independent variables. The independent variable need neither be normally distributed nor linearly related even nor for equal variance for each group. Logistic regression can be stated as follows

$$logit[p(x)] = \log\left[\frac{p(x)}{\{1 - P(X)\}}\right] = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots$$

Here the logit represents the log with base €, P represents the feature which ranges from 0-1 as an intercept b_1, b_2 are the coefficient values related to the corresponding value related to corresponding descriptors. The value of 0 of the corresponding coefficients denotes the null contribution of the descriptors towards its interpretation while thee with a "+" sign denotes the null contribution of the descriptors towards its interpretation. The descriptors with "-"coefficients negatively towards the interpretation of the model.

Chapter-2

Present work

Present work

A chemical compound has its particular physiochemical behavior, therapeutic potency as well as organoleptic characteristics. However, the huge applicability of its particular phenomenon suggests studying their qualitative, structural studies to the biological response. In the present work, the investigation is predominantly based on the organoleptic properties of different chemical compounds. The goal is to achieve the desired product while minimizing the hazards and the negative influence of those chemical products in our day-to-day lives. Other than that one of the core focuses of this investigation is to achieve a more accurate result while simplifying the methodology. However, analysis of the applicability domain helps chemical categorization of the chemical compounds that are still not synthesized or yet to be synthesized. Structural modification and chemical categorization somehow contribute to the synthesis of potentially new safer chemical compounds whether it is organic or inorganic. Therefore, in the current study, the possibility of predicting reliable data was checked by the rigorous validation of the QSAR model. Apart from that, we used RASAR a concept of both the QSAR and Read across. The methodology is predominantly dependent on the prediction ability of the ten close source compounds of a chemical of interest. The similarity whether it is chemical, structural, or biological contributes to a prediction opinion. Of late several regulatory agencies consider the chemometric Insilco (e.g. QSAR, read across) method as one of the significant tools for risk assessment even for determining the property or the biological activity of a potential chemical. In our present study, the 2D QSAR approach has been used to develop both the classification and regression-based model. Apart from that RASAR has also been used as the clubbing concept of QSAR and Read Across. Before finalizing the model development a variable selection strategy (MDF for classification-based model, MLR genetic pool, Best

subset for regression-based model) was applied to select some significant and manageable number of descriptors to minimize the noise and correlated descriptors in the data set. Development of a validated, predictive model as QSPR / RASAR provides rational estimation for property determination of organoleptic compounds.

2.1. Study 1 Dataset 1

The data set of the first study mainly deals with 2370 sweet taste compounds and 2431 bitter taste compounds. The compounds that show '1' have a taste (sweet or bitter), while the compounds that show '0' are non-sweet or non-bitter. The data sets contain diverse compounds, including carbohydrates and sweeteners such as D-Xylose, Amylose, D-Mannitol, D-Mannose, and Aspartame, as well as some other natural products such as Quinine and xanthotoxins. The details of the datasets (both sweet and bitter) are discussed later.

2.2 Study 2 Dataset 2

It is essential to have consistent and reliable data for the development of QSPR models. In this second study, 1208 data points for aromatic substances were collected which describes the experimental property as the Kováts retention index (RI) in a non-polar stationary capillary column (0.28 mm ×50 m). They used methyl silicone OV-101 as coating material admixed with 1% Carbowax 20 M, and the column was programmed to increase from 80 to 200 °C at a rate of 2 °C/min. The RI values used as an endpoint ranged from 350 to 2180. The Kováts retention index is independent of individual chromatographic system specifications and allows comparing values measured by different analytical laboratories and analysis times. The fragrance ingredients are often obtained from commercial suppliers as mixtures of isomers (e.g., cis-trans), which the supplier does not separate.

Chapter-3

Method and materials.

3. Materials and method

The present dissertation was performed with the core objective to showcase the applicability of a transparent methodological framework to develop a predictive QSAR as well as RASAR model while using simply interpretable two-dimensional (2D) molecular descriptors as well as RASAR descriptors. The necessary strategies are taken to be granted for descriptor calculation, descriptor pretreatment, or descriptor thinning for the entire data set following the predictive judgment and robustness of the models. A details explanation of the working data set, principal, and methodology of the recent studies and a precise vivid discussion of the mechanism and algorithm of each study have been done.

3.1 Study1

3.1.1 Dataset

Developing an *in-silico* model requires careful consideration of the data set. In this case, we confidently focused on sweet and bitter taste-related compounds to develop a classification-based model. We extensively validated the model using the estimated required data. To obtain the necessary data, we conducted a thorough search on GitHub repositories [57] for Sweet-DB (Sweet database) and Bitter-DB (Bitter database). We successfully extracted 2370 compounds for the sweet taste and 2431 compounds for the bitter taste from the given data sets. The compounds that show '1' have a taste (sweet or bitter), while the compounds that show '0' are non-sweet or non-bitter. The data sets contain diverse compounds, including carbohydrates and sweeteners such as D-Xylose, Amylose, D-Mannitol, D-Mannose, and Aspartame, as well as some other natural products such as Quinine and xanthotoxins.

3.1.2 Molecular representation and data curation

We used Marvin Sketch software (https://chemaxon.com/marvin) to create a structural representation of the sweet and bitter compounds data from their SMILES. To ensure accuracy, we curated both data sets using a KNIME workflow (http://teqip.jdvu.ac.in/QSAR_Tools/), removing any salt forms related to the chemical structures. We also conducted duplicate analysis and removed mixture compounds from both databases (Sweet-DB and Bitter-DB). For mixture compounds, we selected the most important fragment for further analysis and used it for the *in-silico* QSPR classification model development. As a result, we reduced the number of compounds in the case of Sweet-DB from 2370 to 2311, and in the case of Bitter-DB from 2431 to 2370.

3.1.3 Descriptor calculation and pre-treatment

In the first step of our analysis, the chemical structures of the compounds were considered for corresponding descriptors calculation followed by a curation step. 2D structural and physicochemical descriptors were calculated using alvaDesc software (https://www.alvascience.com/alvadesc/). Constitutional, Ring, Connectivity index, Functional group count, Atom centered fragment, Atom type E-state, 2D-atom pair, and molecular properties were considered for the descriptor calculation This software not only calculates the descriptors of the chemical compounds but also removes the missing, less significant and inter-correlated descriptors as the method of pre-treatment. As a result, 573 descriptors for the chemical compounds were obtained. In order to filter out the most contributing features (descriptors), first the pre-treatment was done to remove the inter-correlated descriptors with less significance toward model development.

3.1.4 Data division

To develop a classification-based QSPR model, it is necessary to divide the dataset into a training set and a test set. The training set is used for model development, while the test set is used to evaluate the model's predictive [58] In this study, we divided the data sets into a 50-50

ratio for ease of model development. As a result, both the training and test sets contain 50% of the corresponding entire datasets. Therefore, the sweet dataset's training set contains 1156 compounds, and its test set contains 1155 compounds. Similarly, the bitter dataset's training set contains 1186 compounds, and its test set contains 1184 compounds respectively.

3.1.5 Feature selection

Chemical compounds have unique features or descriptors that define their characteristics. In QSPR analysis, selecting the most important features is crucial to identify the contributing factors towards the response. There are several techniques available for feature selection in QSPR studies [59] but in this particular study, we have used the most discriminating features selection algorithm (MDF) analysis [60]. for stepwise linear discriminant analysis (LDA). In this method, the training set is normalized from 0-1, and the compounds are divided into two groups - active and inactive - with responses of 1 and 0, respectively. The mean of each descriptor for each class is then calculated, and the absolute difference is determined by subtracting the mean inactive part from the mean of the active part. The features with the highest absolute differences are identified as the most discriminating features and are used in the QSPR analysis. Furthermore, for read-across (RA) analysis and RASPR descriptor calculation, features from the stepwise LDA model were selected, for further calculation.

3.1.6 Analysis of unbalanced set

When performing QSPR modelling based on classification, it is important to balance an unbalanced set before developing any model. This means that the number of active compounds should be similar to the number of inactive ones. This step is necessary to avoid any bias toward any one class of compounds. In the case of the sweet dataset, the number of active and inactive compounds was approximately equal, so no balancing was required. However, for the bitter dataset, the training set was initially biased towards inactive compounds, with the ratio of inactive to active compounds being approximately 2:1. Therefore, balancing the training set

was necessary. To balance the dataset, we oversampled the active compounds by duplicating them, so that the overall ratio between inactive and active compounds was close to 1:1. The modified training set for the bitter dataset was then used for model development.

3.1.7 Conventional Classical QSPR model

A linear discriminant analysis (LDA) model [61] was created using the most discriminating features (MDF) for both the Sweet-DB and Bitter-DB datasets, using STATISTICA 7.1 (STATSOFT Inc. USA http://www.statsoft.com). LDA is a statistical method that classifies input data into two linear classes. Unlike multiple linear regression (MLR), LDA provides a predictive correlation equation that determines the positive and negative influence of a descriptor based on the discriminant function. One of the primary principles of LDA is to differentiate between classes. The Discriminant Function equation describes the influence of each descriptor for the LDA model. The Discriminating Function can be described from the equation

$$DF = c_0 + c_1x_1 + c_2x_2 + c_3x_3 + \cdots + c_nx_n$$

1. Here, in equation (1), the DF stands for the Discriminating Function, X_1, X_2,…, X_n is predictor scores for the total n variables, and C_1, C_2…C_n are the corresponding weights. Here, for the current work, while going for the linear discriminating analysis, the tolerance limit is set for 0.0001, *F* to enter for 4.0, and *F* to remove for 3.9. Later on, the developed model was validated using internationally accepted validation metrics like accuracy, balanced accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), Cohen's κ, and area under the ROC-curve (AUC) [62].

3.1.9 Development of Read across (RA) based prediction

The selected descriptors from the conventional QSPR model generated through stepwise regression were used for read-across (RA) analysis [63]. The selected descriptors from both the datasets (training and test sets) were utilized for read-across predictions using the tool Read-Across-v4.2.1 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.7hxieb6k5y4b). This approach utilizes the supervised learning method and generates similarity-based predictions based on Euclidean distance-based similarity, the Gaussian Kernel similarity, and the Laplacian Kernel similarity. The default settings for the Read-Across-based predictions are σ=1, γ=1, No. of close source neighbours = 10. This tool utilizes a set of "n" close source compounds for every query or test set compound. To derive the optimum setting for the RA predictions, a hyperparameter optimization was also performed.

3.1.9.1 RASPAR descriptors calculation

In addition to the 2D descriptors, we also calculated RASPR descriptors calculated using RASPR-Desc-Calc-v3.0.1, which is available from the DTC Lab tools supplementary site (https://sites.google.com/jadavpuruniversity.in/dtc-labsoftware/home). We used the default setting of the Read-Across hyperparameters to calculate RASAR descriptors for both datasets. The standard for RASPR descriptor calculation was based on the suggested Euclidean distance, where the number of the closest source compounds was set to ten and the threshold for the distance was set to one. The calculation of RASPR descriptors [64]. considered the structural and physicochemical features or descriptors from the previously developed QSPR LDA model. The calculated c-RASPR descriptors were then used to perform LDA models in conjunction with the forward stepwise regression method of variable selection, with the criteria F = 4 for inclusion and F = 3.9 for variable exclusion. Unlike QSPR descriptors, the calculated RASPR descriptors encode information related to the close source congeners of a particular query compound, rather than the query compound itself. The derived descriptors are similar to the latent variables that feature all related information of structural and physicochemical

descriptors obtained from QSPR models and generate models with a reduced number of descriptors that contain all chemical information.

3.1.8.2 Machine learning-based model development

1. In our work, we used various machine learning (ML) approaches to test the predictive ability of the developed c-RASPR models for both Sweet-DB and Bitter-DB. We compared the prediction quality of the developed ML models developed using a support vector classifier (SVC) [65].
1. logistic regression (LR) [66]
2. and random forest classifier (RFC) [67]
3. with the default settings corresponding to hyperparameters using the ML classifier tool available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis, with the selected descriptors derived from the c-RASPR (LDA-RASPR) model. SHAP (Shapley Additive explanation) analysis plot was also performed to explain the supervised model (SVC) and assign the importance of the modelled descriptors for specific prediction [68].

3.1.9 Applicability Domain (AD)

The concept of applicability domain can be defined based on the molecular descriptor space. The reliability of predictions for objects outside the training set chemical space can be determined by evaluating the performance of the model on unseen objects during validation. However, it is important to note that objects that are further away from the molecular descriptor space covered by the training set may result in larger error rates [69]. AD (Applicability Domain) aims to identify objects, anomalies, or outliers in the molecular descriptor space. To predict the properties of a new or unknown compound, it must fall within the theoretical chemical space known as the Applicability Domain (AD) of the model. There are various techniques to determine the AD of a model, but we have used the leverage approach for both

the training and test sets to determine the structural outliers. We used Hi_Calculator-v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) to perform the analysis.

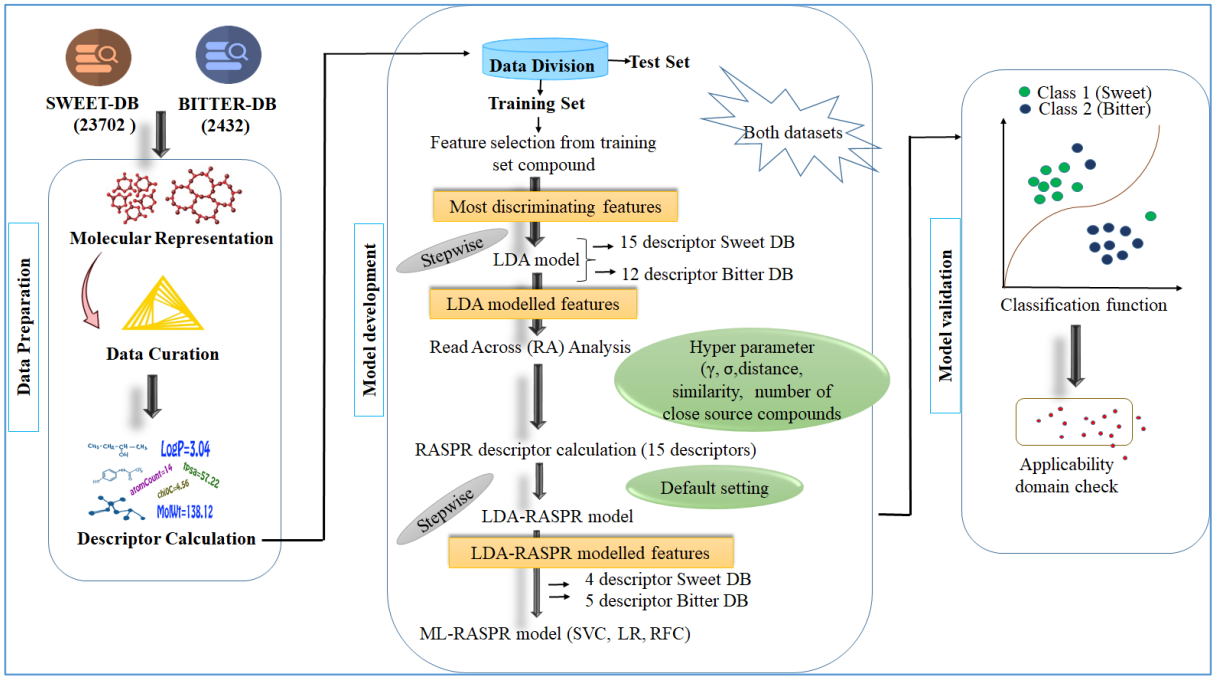


Fig 3.1 Workflow of Study 1

3.2 Study 2

3.2.1 Dataset Collection

It is essential to have consistent and reliable data for the development of QSPR models. In this study, 1208 data points for aromatic substances were collected from the literature [70] for model development. The researchers [70] reported the experimental property as the Kovar's retention index (RI) in a non-polar stationary capillary column (0.28 mm ×50 m). They used methyl silicone OV-101 as coating material admixed with 1% Carbowax 20 M, and the column was programmed to increase from 80 to 200 °C at a rate of 2 °C/min. The RI values used as an endpoint ranged from 350 to 2180 [71]. Kovart's retention index is independent of individual chromatographic system specification and allows comparing values measured by different analytical laboratories and analysis times. The fragrance ingredients are often obtained from commercial suppliers as mixtures of isomers (e.g., cis-trans), which the supplier does not

separate. However, we cannot neglect the effect of temperature, pH, and surrounding environment for transforming a particular isomeric form of a chemical compound to another (like a transformation of a cis compound to trans and trans compound to cis form) for a mixture of compounds while it was supplied for experimentation. This consequence may result in exceptional responses where a single compound represents two different isomeric mixtures with the same molecular weight. In this scenario the compounds like Allyl anthranilate 1 and Allyl anthranilate 2 may not represent the pure cis or trans isomeric form of a compound rather they were represented as a mixture of both the geometrical isomers. In the present study, it was interpreted as a single compound with an isomeric mixture while considering the impact of other external factors as well. In that case, collecting the average retention index value (compounds with quite similar chromatographic peaks) of Rojas et al. is justified for further development of an accurate and interpretable model. This kind of approximation is very common in any 2D-QSPR analysis.

3.2.2 Molecular representation and data curation

A total of 1208 flavour and fragrance compounds, each with its corresponding SMILES, chemical names, and retention index, were initially compiled (provided in supplementary information 1). To ensure accuracy, for compounds with more than one reported retention index value, the average value was calculated, and duplicate entries were removed, resulting in a final curated dataset of 1194 compounds. The structural representation of the compounds was done using Marvin Sketch software (https://chemaxon.com/marvin). Additionally, a curated SDF file of the flavour and fragrance compounds was obtained after incorporating explicit hydrogen, ring aromatization, and 2D form cleaning for the descriptor calculations.

3.2.3 Descriptor calculation

In this study, we used the Alvadesc software (https://www. alvascience.com/alvadesc) to calculate descriptors for flavour and fragrance compounds. These descriptors are numerical

values that define the physiochemical properties of a compound. We have used only simple, direct mathematical algorithms nature, reproducible, and easily interpretable 2D descriptors [72] to avoid the complexity of 3D analysis and energy [73] 2D descriptors have a deluge of contributions in extracting chemical attributes and some are capable of representing 3D features to some extent [74] 74 However, it is not possible to differentiate between the isomers (cis, trans, etc.) of compounds completely using 2D-QSPR models. The work of Rojas et al. had already concluded that 3D descriptors did not significantly improve the quality parameters of the QSPR model. From the previous conclusion, we have decided to develop simpler 2D-QSPR models while using the concept of intelligent consensus predictions. Lastly, the redirection toward the source data, the unseparated mixture of both the geometrical isomers of a particular compound, and their response values indicate an inseparable form of cis and trans isomers even after the application of 3D descriptors. In the present study, the isomers were recognized as a single compound. In that case, collecting the average retention index value (compounds with quite similar chromatographic peaks) of Rojas et al. is justified for further development of an accurate and interpretable model. This kind of approximation is very common in any 2D-QSPR analysis. A total of 2400 2D descriptors were calculated, including constitutional descriptors (molecular composition of a referenced compound), ETA indices (extended top chemical atom), ring (information related to the presence of ring descriptors), functional group count, atom-centered fragment, connectivity index, atomtype E-state (description related to the electronic state of the atoms), 2D atom pair, and molecular properties [72]Additionally, data pre-treatment was performed using the DataPreTreatmentGUI_1.2 (http://teqip.jdvu.ac.in/QSAR_ Tools/) software to eliminate correlated (correlation cut-off of 0.95) and descriptors having low variance cut-of (less than or equal to 0.01), resulting in a total of 309 curated descriptors for further modelling.

3.2.4 Dataset division

Partitioning the dataset is an essential step in developing the QSAR model. A chemometric statistical model requires two independent datasets: a training set for developing the model and a test set for validating the [75] Generally, the whole dataset was divided into the training and test set in the ratio of 70:30 (approx.). In the present investigation, the dataset of fragrance and favour compounds was divided into four clusters based on their properties (sorted responsebased method.) using the Dataset Division 1.2 tool (http:// teqip.jdvu.ac.in/QSAR_Tools/). This property-based data division resulted in a training set of 896 compounds and a test set of 298 favoured and fragrance compounds.

3.2.5 Test training pre-treatment

The training and test set data may contain correlated and noisy descriptors that are not relevant to the data modelling purpose. Therefore, pre-treating both the training and test sets is necessary. In our study, we utilized the Data Pre-treatment tool (http://teqip.jdvu.ac.in/QSAR_Tools/) to pre-treat the training and test sets after division, using a variance cut of 0.01 and a correlation cut-off of 0.95. This process resulted in 162 less correlated descriptors, ultimately minimizing the error in model development.

3.2.6 Feature selection and model development

The selected features after pre-treatment were utilized for the feature selection process. Genetic algorithm (GA) followed by BSS (Best Subset Selection) (http://teqip.jdvu.ac. in/QSAR_Tools/) was used for feature selection [76] Initially, some features were also selected using the stepwise selection method. Stepwise regression can be defined as a multiple linear regression which was evolved with the step-by-step mechanism. After removing the selected features from the first stepwise run, the stepwise method was again performed with the remaining pool of descriptors. Besides stepwise feature selection, GA was also performed for

the feature selection procedure. GA tool has many advantages over other feature selection methods. It is based on fitness function on mean absolute error (MAE)-based pick-up criteria. We have employed our in-house tool "Genetic Algorithm_v4.1_Train" (http://teqip.jdvu.ac.in/QSAR_Tools/) to find the most relevant descriptors with the RI endpoint. The best subset selection (BSS) approach was used to find the optimal combination of descriptors for a robust prediction model. After selecting the best descriptors from both feature selection methods, we performed partial least squares (PLS) regression to build the preliminary QSPR models. PLS methods were employed to develop the final robust models to avoid any chances of inter-correlation among descriptors. The PLS regression method is a generalized technique of the "Multiple Linear Regression (MLR)" method, where we can examine strongly collinear, correlated, noisy data and many X variables. The PLS regression has been carried out with a Java-based software tool "PLS_SingleY_version" (http:// teqip.jdvu.ac.in/QSAR_Tools/). The PLS model was further utilized for best subset selection (BSS). The best subset selection was performed with the in-house tool developed in our laboratory (http://teqip.jdvu.ac.in/QSAR_Tools/). Six descriptor models (five PLS models) were generated based on MAE-based [77]

3.2.7 Model validation criteria

The developed QSTR models were rigorously validated via various internationally accepted metrics to ensure the robustness, predictability, goodness of fit, and quality of the models. For training set compounds, internal validation metrics such as cross-validated correlation coefficient Q2 (LOO) (leave one out), rm^2_{loo}, MAE_{train} (mean absolute error),$RMSD_{train}$ (root mean square standard deviation error), and coefficient of determination R2 were calculated to measure the robustness and goodness of ft. of the model. For test set compounds, we have predicted external set compounds using globally accepted different validation metrics

like predictive MAE_{test}, $RMSD_{test}$, R^2(R^2 pred), or Q_{F1}^2 and Q_{F2}^2 to judge the predictability of the model [78]

3.2.8 Applicability Domain Assessment

The applicability domain is the biological, chemical, or physiochemical hypothetical space of the training set chemicals through the recently created QSPR model. The main use of this domain is to predict the toxicity value of compounds that fall in this domain and have unknown values. We have used the DModX (distance to mean X) approach to predict the AD of the PLS models (OECD principle 3) using SIMCA-P software [78-80]. The DModX uses Y and X residuals as diagnostic values to ensure model quality. If the DModX value is greater than the critical value, it means that the query compound is outside the domain of the model [77, 78–80]:

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

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

3.2.9 Intelligent Consensus Prediction

This method evaluates the performance of the consensus models in comparison to the individual models based on MAE-based criteria (i.e., 95%). It is recognized that a single model may not be able to accurately predict all of the test compounds. This implies that one QSPR model may be more suitable for one test compound, while another model may be better for a different test compound [73,82,83]. A specific QSPR model may not be equally effective in predicting all query compounds in the query list. To get the best prediction results, we need to consider the consensus of all the predictions made by these four models. For this, consensus

prediction should be made intelligently, i.e., in a query compound-specific way, using all or most of the valid models. This is different from doing a simple average of predictions from all available models. Consensus prediction is better than individual model predictions since it combines all the good characteristics of each model. Thus, the drawbacks of one individual model are taken care of by other models (s). This makes the predictions less biased, more reliable, and more precise. The individual models may have differently defined applicability domains, while the consensus method combines the ADs of the individual models, thus providing a greater chemical space coverage as well. Moreover, the consensus method does not affect the quality of the internal statistical parameters of the individual models [84]. In the present study, we have chosen five models (M1–M5) to conduct a consensus prediction using the "Intelligent Consensus Predictor" tool that is available on our laboratory website (http://teqip.jdvu.ac.in/QSAR_Tools/. The steps involved in developing the models are depicted in Fig.

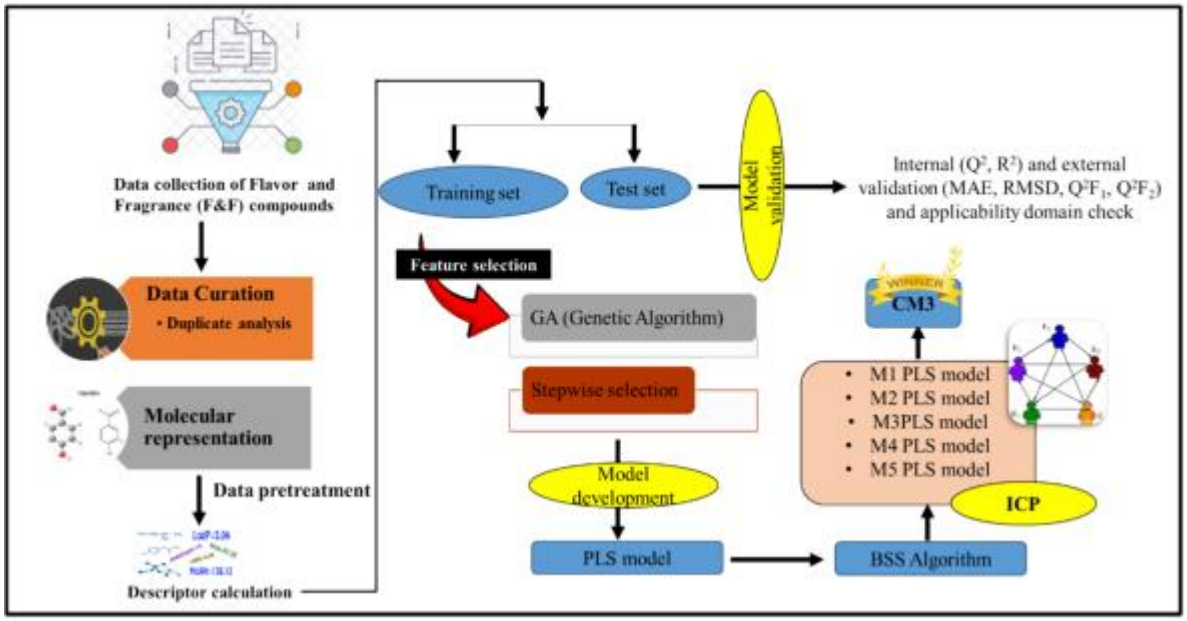


Fig 3.2 Workflow of study 2

Chapter-4

Result and discussion

Result and Discussion

4.1. Study 1: The first application of machine learning-based classification read-across structure-property relationship (c-RASPR) modelling for sweet and bitter

The aim of the present study was to effectively investigate the applicability of RASAR (a hybrid algorithm) along with machine learning over simple QSAR for a quality prediction of a given data set. Here we have explored the mechanism of a classification-based model. Finally, the obtained results according to the investigation were documented while the detailed interpretation of the results from RASAR descriptors was done. This represents the deep understanding of statistical phenomenon while considering biological response corresponding to the mechanism of RASAR.

4.1.1. Machine learning-based classification read across structure-property relationship (c-RASAR) model:

In a recent investigation, LDA QSPR models (denoted by equation (2) and equation (3)) were developed that included sweet and bitter compounds, respectively, through a stepwise selection of descriptors using STATISTICA software (STATISTICA 7.1 STATSOFT Inc., 2023) as discussed in the materials and methods section. A similarity-based approach was then used to ensure accuracy and avoid complexity given the high number of descriptors in the model. Thus, to improve model interpretability and transferability for each data set, classification LDA models were developed using similarity-based measures computed in the c-RASPR approach from the selected LDA QSPR descriptors (15 for sweet and 12 for bitter data sets (**Supplementary file of study 1)**. The predicted response of an unknown compound was determined by using the known response of similar structural analogs. The computation was performed based on the basic settings with ED (Euclidean distance-based) similarity

computation for sweet and bitter databases. To calculate the RASPR descriptors based on the similarity and error-based measures of the close source compounds for each target compound obtained from the QSPR model, we used the 15 RASPR features obtained from physicochemical variables based on similarity measures of RA (Read- Across) prediction. These features extract the information of 2D descriptors based on similarity and error measures with the close source chemical as per user-defined input. The obtained models with the respective descriptors (LDA-RASPR) show better predictivity compared to the classical LDA QSPR models with a smaller number of features

The developed classical LDA models in this study demonstrated a commendable accuracy of 0.69-0.73 in predicting the quality; other metrics like precision (0.60-0.77), F1 score (0.69-0.75), Matthews correlation coefficient (MCC) (0.38-0.69), Cohen's kappa (0.38-0.41), and AUC (0.65-0.74) for both bitter and sweet datasets for training and test sets are represented in **Table 4.1**. Furthermore, the incorporation of c-RASPR descriptors resulted in further enhancement of the results for both the sweet and bitter data sets (denoted by equations (4) and (5)). Among the machine learning models tested, SVC exhibited the best performance, and hence, only the SVC results were reported in this paper. While the results of other models are provided in the supplementary section for reference (**supplementary file of study 1**), the present study offers a comparison of the validation results obtained through different methodologies, namely LDA-QSPR, LDA-RASPR, and SVC-RASPR, presented in **Table 4. 1** below. The study utilizes sweet and bitter compounds for the first time with a new modelling algorithm (c-RASPR) for classification-based modelling. The results of this study demonstrate a significant enhancement in the prediction quality for the query set compared to the classical QSPR model in terms of different validation metrics. We have also calculated the applicability domain, according to the OECD principle for both the datasets and found that 0.031% and 0.043% compounds were outside the AD for Sweet-DB and Bitter-DB datasets, respectively.

4.1.1.1. Result for the classification-based LDA-QSPR model (M 1.1 and M 1.2).

Model M 1.1 (QSPR model for sweet compounds)

$$df(sweet) = -0.74777 - 0.10182 * \max_{conj_{path}} ** - 1.00822 * NsssN + 1.93896 * B03[C-O] - 0.00291 * MW + 0.79705 * B02[O-O] - 1.3022 * 0 - 062 - 0.11899 * LOGPcon + 0.50642 * nRCOOH - 0.962906 * C - 018 - 0.4332 * F01[N-O] - 1.2918 * nArCOOH - 1.35446 * nNq - 0.8674 * C - 033 - 1.0149 * nRNHR + 0.35148 * NaaaC.$$

Model M 1.2 (QSPR model for bitter compounds)

$$df(bitter) = -0.3624 + 0.1069 * \max_{conj_{path}} * + 0.1462 * F01[C-N] - 1.0398 * nRCOOH + 0.0037 * MW - 0.9880 * B01\{C-O] + 1.1514 * O - 062 - 0.8308 * B02[O-O] - 1.3619 * C - 036 + 0.9478 * B01[O-S] + 0.7348 * F02[N-S\} + 1.0808 * nRNHR + 0.5057 * C - 019$$

The QSPR-LDA models (denoted by equations (2) and (3)) for both sweet and bitter compounds data have good statistical metric values including Sensitivity, Specificity, Accuracy, and Precision, all of which are above 0.5. For **model 1.1** (the QSPR model for sweet compounds), descriptors like B03[C-O], B02[O-O], nRCOOH, and NaaaC positively contribute to the endpoint. These descriptors suggest that sweetness is influenced by the polarity of the compound due to the presence of oxygens and carboxylic acid fragments in the aliphatic chain, and electron-richness in the form of aromatic fused carbons in the sweet compounds. On the other hand, descriptors like max_conj_path, NsssN, MW, O-062, LOGPcon, C-018, F01[N-O], nArCOOH, nNq, C-033, nRNHR contribute negatively to the respective endpoint. These descriptors indicate the hydrophobic nature of the compounds. The descriptors C-018, nArCOOH, and C-033 represent the presence of electronegativity. Again, the presence of hydrogen bonding atoms is represented by NsssN, nNq, and O-062 descriptors.

In **model 1.2**, the core concept of QSPR is applied, and it has been validated similarly to model 1. It has been found that max_conj_path, F01[C-N], MW, O-062, B01[O-S], F02[N-S], nRNHR, and C-019 descriptors showed positive contributions. These features are quite similar to the features obtained from **model 1.1**, which are negatively correlated. These descriptors represent hydrophobicity, electronegative nature, hydrogen acceptor, and the presence of a more aromatic nature of the compounds. All of these properties are seen to influence the generation of a bitter taste. On the other hand, features like nRCOOH, B01 [C-O], B02 [O-O], and C-036 are more closely associated with **model 1.1** and are positively correlated. It is worth noticing that features like polarity, i.e., the presence of oxygen or aliphatic carboxylic acids, are negatively correlated to the corresponding endpoint (bitter taste). To obtain better classification metrics, the descriptors obtained from both model 1.1 and model 1.2 of the QSPR-LDA are highly relevant and have been confidently utilized for the classification-based LDA using RASPR descriptors.

4.1.2. Result for the classification-based LDA RASPR models (M 1.3 and M 1.4).

Two classification-based LDA RASPR models (denoted by equations (4) and (5)) were created using different sets of similarity descriptors for sweet and bitter compounds. The sweet compounds were evaluated with 15 QSPR descriptors, while the bitter compounds were evaluated with 12 QSPR descriptors. Then, the selected features were used to calculate the RASPR descriptors, which resulted in 15 descriptors being generated for both sweet and bitter compounds, respectively. Next, the generated 15 descriptors (RASPR) were used to create an LDA RASPR model for each of the sweet and bitter compounds. The obtained LDA RASPR models (using RASPR descriptors) for sweet and bitter compounds are presented in **Equations 4** and **5**, respectively. Further details of statistical parameters can be found in **Table 4.1**

Model M 1.3 (LDA-RASPR model for sweet taste compounds)

$$df(sweet) = -8.914 + 3.334 * RA\ function + 7.445 * MaxPos + 8.105 * g_m * AvgSim - 7.408 * g_m$$

Model M 1.4 (LDA-RASPR model for bitter taste compounds

$$df(bitter) = 0.729 + 3.908 * RA\ function + 0.718 * Pos.Avg > Sim - 8.423 * Avg.Sim + 6.92 * MaxPos - 11.485 * g_m * SDsimilarity$$

Table 4.1. Comparative quality QSPR, LDA-RASPR and SVC-RASPR (ML) models for the sweet and bitter data sets

Data	**Model**	**Division**	**AUC**	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-measure**	**G-mean**	**MCC**	**Cohen's k**
Sweet	LDA QSPR **(M1)**	Training	0.674	0.747	0.674	0.708	0.674	0.709	0.709	0.693	0.418
		Test	0.651	0.737	0.651	0.692	0.656	0.694	0.693	0.389	0.386
	LDA -RASPR **(M3)**	Training	0.705	0.780	0.705	0.740	0.705	0.740	0.741	0.485	0.482
		Test	0.666	0.777	0.666	0.719	0.677	0.723	0.719	0.444	0.440
	SVC -RASPR (ML) **(M5)**	Training	**0.797**	**0.783**	**0.708**	**0.744**	**0.708**	**0.744**	**0.744**	**0.492**	**0.489**
		Test	**0.776**	**0.780**	**0.677**	**0.720**	**0.678**	**0.726**	**0.726**	**0.448**	**0.443**
Bitter	LDA QSPR **(M2)**	Training	0.744	0.729	0.744	0.736	0.773	0.750	0.737	0.472	0.471
		Test	0.717	0.690	0.717	0.706	0.601	0.643	0.704	0.399	0.396
	LDA -RASPR **(M4)**	Training	0.744	0.725	0.744	0.734	0.772	0.772	0.734	0.467	0.466
		Test	0.749	0.728	0.749	0.741	0.641	0.682	0.738	0.467	0.465
	SVC -RASPR (ML) (**M6)**	Training	**0.870**	**0.804**	**0.811**	**0.807**	**0.835**	**0.819**	**0.808**	**0.614**	**0.613**
		Test	**0.732**	**0.622**	**0.793**	**0.728**	**0.651**	**0.636**	**0.702**	**0.419**	**0.419**

Models 3 and 4 represent the LDA-RASPR models for sweet and bitter taste compounds, respectively. **Equation 1.4 of (M 1.4)** shows that the *RA function, Max.Pos,* and g_m**Avg.Sim* are positively correlated with the response values, while *gm* contributes negatively. The LDA-RASPR model provides better predictions than the corresponding previous LDA QSPR model, as confirmed by the significant increase in Cohen's kappa values for both the training and test sets. The same methodology was followed for the bitter taste compounds, as indicated in Equation 5. The LDA -RASPR model for bitter taste compounds consisted of five descriptors, including *RA function, Pos.Avg.Sim,* and *Max.Pos,* which showed a positive contribution to the bitter taste. *Avg.Sim* and g_m**SD Similarity*, on the other hand, negatively contributes to the response. Interestingly, the use of RASPR descriptors reduced the number of descriptors from 15 to 4 for sweet compounds and from 12 to 5 for bitter compounds, while enhancing the validation metric values of the previously developed LDA QSPR models of the sweet data set **(**Model 1.1) and also the bitter data set (Model 1.2). In case of the interpretation of models 3 and 4, first of all we have to consider the constituent descriptors of the model as listed in the **Table 4.2**.

Table 4.2. Representation of RASPR Descriptors with their respective meaning (for both M 1.3 and M 1.4).

RASPR Descriptors	**Meaning**
RA function	Read across derived composite feature describing all the structural and physiochemical features as a single function
Pos.Avg.Sim	Average similarity value of the positive close source compounds.
Max Pos	Similarity value to the closest positive source compound.
Avg. Sim	Average similarity value of the close source congeners

g_m*Avg.Sim	The product of g_m and *Avg similarity* of close source compounds.
g_m	A novel concordance measure (Banerjee-Roy Coefficient).
g_m*SD Similarity	Product of g_m and standard deviation of similarity values of close source compounds.

4.1.3 Result for the classification-based ML- based models (M 1.5 and M 1.6)

For the development of the ML model, the features for the sweet data set and bitter data set were separately taken from the previous LDA RASPR models (M 1.3 and M 1.4). Model M 1.5 was built using the *RA function, Max Pos,* g_m*Avg.Sim,* and g_m descriptors. On the other hand, the M6 model was developed using *Avg. Sim, RA function, Pos.Avg.Sim*, and g_m*SD* similarity descriptors. Although RF, SVC, and LR analyses were performed, the SVC algorithm was found to be the best-performing one for both datasets in terms of both internal and external predictions. The results of the SVC-RASPR models are shown in **Table 4.1** showing better quality than the corresponding LDA-RASPR models. The ROC curves of the developed SVC-RASPR models are shown in Figures **4.1** and **4.2** and the SHAP (**Figures 4.3 and 4.4**) analysis for SVC-RASPR provides insight into each individual RASPR descriptor and its corresponding significance.

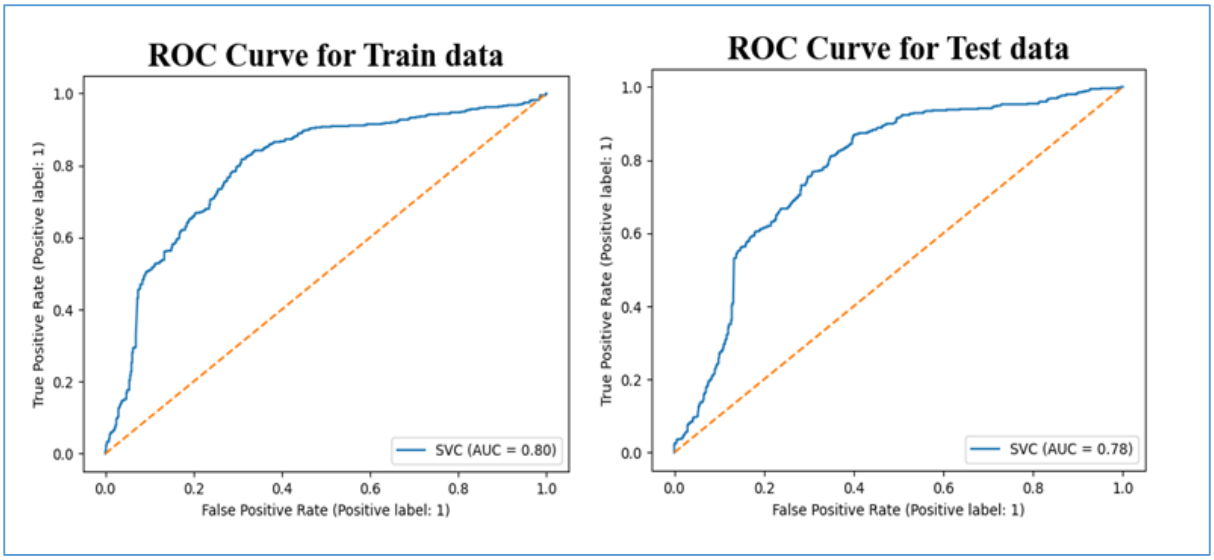


Figure 4.1. ROC curves for Sweet DB compounds (**M5**) (for both the training set and test set)

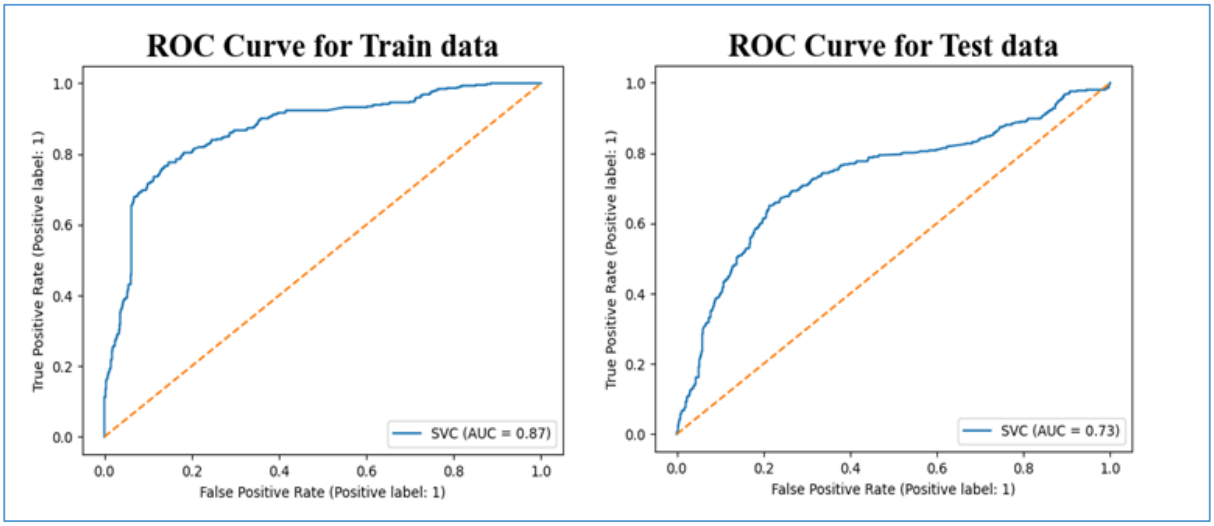


Figure 4.2 ROC curve for Bitter DB compounds (**M 1.6**) (for both the training set and test set)

4.1.3.1 Interpretation for the ML-RASPR model of sweet data set-related compounds

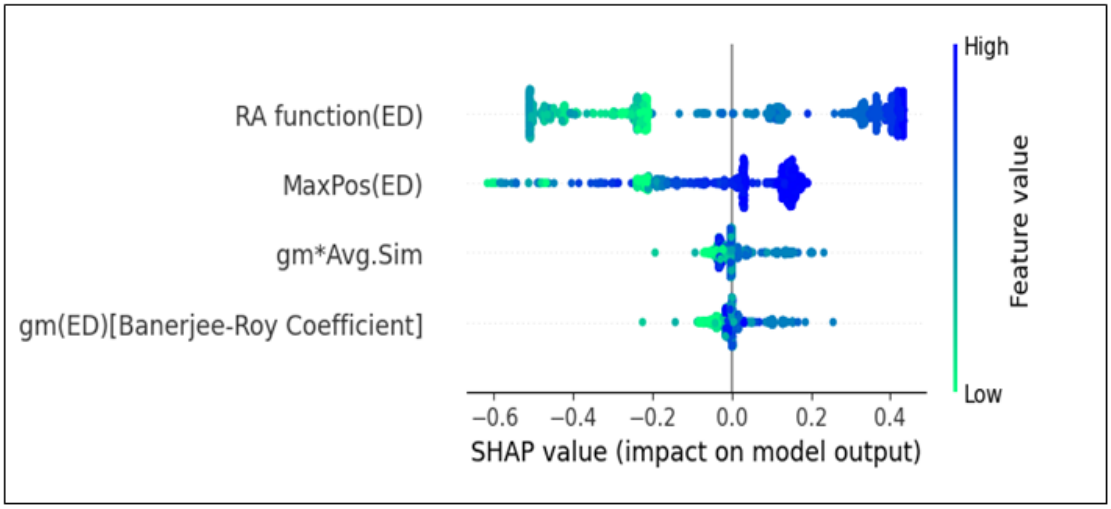


Figure 4.3 SHAP analysis for the sweet compounds (model 1.5).

Figure 4.3 here

The core motive of a SHAP analysis is to determine the individual contribution of the descriptors that are responsible for model development. The plots obtained with the SHAP value denote the same. The impact of a particular descriptor may vary from model to model. The role of the SHAP analysis is like the t-test of statistics to determine the individual contribution of the descriptors from the developed model.

Based on the SHAP (SHapley Additive exPlanation) analysis plot [85] it was evident that the feature with the highest significance value is the ***RA function***. This function is a composite score of all the individual 2D descriptors that were used to build the model, and it is derived from the Euclidean distance-based similarity algorithm. Therefore, it encodes information on various structural and physiochemical descriptors and shows a positive contribution to the specific endpoint or response. For instance, compound no. **11** with an *RA function* value of 0.90 has a higher sweetness activity compared to compound no. **109**, which has an *RA function* value of 0.30 and less sweetness activity.

The SHAP analysis identifies ***MaxPos*** as the second-most significant descriptor. *MaxPos* represents the similarity value between the query compound and the closest positive source compound. Compounds with higher *MaxPos* values are expected to have higher response

values. For example, compound no. 35 has a *MaxPos* value of 0.98 and shows a high response value. Compound no. 59, with a *MaxPos* value of 0.99, exhibits sweetness characteristics.

The next significant descriptor in the order of importance is $\boldsymbol{g_m}$***Avg.Sim***. It is a product of two primary RASAR descriptors g_m and *Avg.Sim.*, which has a significant positive impact on model development. This is observed in compound no. **21** (g_m**Avg Sim* value is 1, indicating the presence of sweetness) and compound no. **31** (g_m**Avg Sim* value is -0.142, indicating the absence of sweetness). Thus, these observations help to interpret this secondary cross-product.

In this series, the last descriptor is $\boldsymbol{g_m}$, also known as the concordance coefficient. As per equation 3, the product, g_m**Avg.Sim*, represents a huge positive contribution, while g_m shows a negative contribution probably as a penalty factor. For instance, compound no. **15** (with a g_m value of -0.4, indicating sweetness characteristics) and compound no. **156** (with a g_m value of 0.4, showing no sweetness characteristics) can explain the scenario where the contribution of g_m is negative.

4.1.3.2 Interpretation related to ML-RASPR model of bitter taste related compounds

The ***Avg. Sim*** is a significant descriptor in RASPR that is based on similarity. When we examine the role of *Avg. Sim* in the c-RASPR model, it indicates that the similarity value between compounds in the bitter data set decreases as the distance between the compounds increases. This suggests that as the distance among the ten closest source compounds for the query molecule increases, the similarity value decreases. The obtained result shows that the distance is inversely proportional to the similarity of the compounds. The negative correlation suggests that an increase in the *Avg. Sim* descriptor indicates the absence of bitter taste, as observed in Compound no. **9** with a value of *Avg. Sim* of 0.986. On the other hand, Compound no. **1491** with an *Avg. Sim* of 0.722 indicates the presence of a bitter taste.

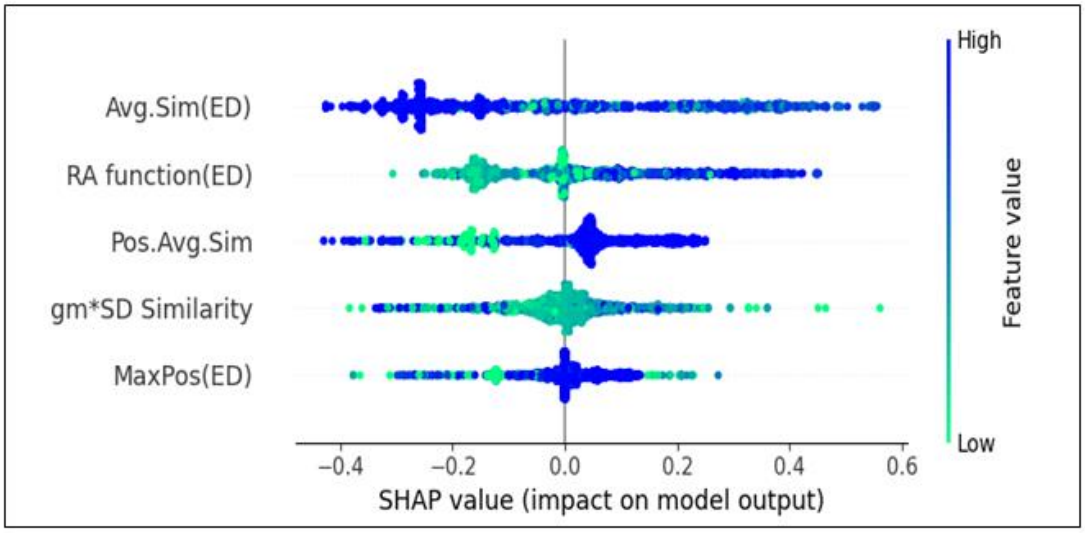


Figure. SHAP analysis for the bitter compounds (model 6).

Figure 4. 4 here

One of the most significant RASPR descriptors for the model is the ***RA function***. The equation shows a positive correlation of this descriptor. *RA function* is a composite score of all the 2D descriptors used to construct the corresponding QSPR model. Therefore, when we assess the contribution of the *RA function* to the model, it indicates a positive effect of all the collective descriptors. For instance, compound number **41** (with an *RA function* value of 0.902) indicates the presence of bitter taste, whereas compound number **25** (with an *RA function* value of 0.20) indicates the absence of bitter taste. This can explain the direct proportionality of this descriptor to the bitter taste endpoint.

The third most significant descriptor is ***Pos. Avg. Sim***, and it is positively correlated with the bitter taste endpoint or response. *Pos. Avg. Sim* refers to the average similarity value among the positive close-source compounds. This indicates a greater tendency towards positive class predictions and less towards negative class predictions. Compound no. **381** with a *Pos.Avg.Sim* value of 0.97 undoubtedly has a bitter taste. In contrast, compound no. **1** with a *Pos.Avg.Sim* value of 0.646 has no taste sensation, making it an excellent and indisputable example of this feature.

In this series, the final feature to consider is ***MaxPos***, which is the maximum similarity value to the closest positive close congener. The higher the *MaxPos* value, the more likely it is for a compound to be positively predicted. For instance, compound no **35** has a *Max.Pos* value of 1, which confidently leads to a positive activity prediction of the compound. On the contrary, compound no 1 has a *Max.Pos* value of 0.721 which confidently leads to a negative activity prediction of the model. Therefore, this descriptor holds a great influence in accurately predicting the activity of the compounds.

4.1.5 Comparison with other work

1.In the previous study, Rojas et al [71]. Conducted a QSPR modelling analysis for both sweet and bitter compounds on a total of 566 compounds in a Sweet–Tasteless dataset and 508 compounds in a Sweet-Bitter dataset. The authors employed sensitivity and specificity as classification-based validation parameters to evaluate the quality of their models. However, they did not perform dataset balancing; therefore, the computed sensitivity and specificity which are the possible indicators of the true positive and true negative predictions may have been affected by the nature of the dataset. In the current work, we have applied balancing to the imbalanced bitter data set following the method of oversampling. We have also reported the values of metrics like MCC, Cohen's kappa, and AUC_ROC. Although a direct comparison between our study and Rojas et al [71]. was not feasible as a result of variations in the methodology and the number of compounds used, we endeavored to assess the difference in prediction quality between their models and ours (**Table 4. 3**). As we know, the study conducted by Rojas et al [71] utilized 566 compounds for the sweet dataset and 508 compounds for the bitter dataset. In comparison, our study consisted of 2311 compounds for the sweet dataset and 2370 compounds for the bitter dataset, which is almost four times larger than the previous datasets. The size and diversity of the compounds in the dataset can influence the validation parameters, but we were able to generate decent values for the validation parameters,

including sensitivity, specificity, MCC, Cohen's kappa, AUC-ROC, precision, F-measure, and G- mean. In a recent work, Tuwani et al [86] applied dimensionality reduction techniques like the Boruta algorithm and principal component analysis before applying the final machine learning classification methods to sweet and bitter data sets but they considered only a very limited number of compounds in the test sets (in the order of 1/8 times of our test sets). Thus, the quality of our predictions is not directly comparable to their models (however, we have shown their best 2D descriptor models based on ROC values of the test sets in **Table 4.3).** When compared to Xiu et [87] work on identifying novel umami molecules using QSAR and molecular docking results, our work is reassuring in terms of the controlling features of sweet and bitter activity, such as the presence of polar groups (C-018, nArCOOH, C-033, F01[C-N], B01[O-S], F02[N-S], nRNHR) and hydrophobicity parameters (max_conj_path, NsssN, MW, O-062, C-018, F01[N-O], nArCOOH, nNq, C-033, nRNHR) that are essential factors for sweet/bitter molecules for ligand binding.

Table 4.3 Comparison with the previous work (Rojas et al. and Tuwani et al.).[70,71]

Model	No. of compounds (n)	Division (n_{Train} or n_{Test})	AUC	Sensitivity	Specificity	Accuracy	Precision	F-measure	G-mean	MCC	Cohen's k
Present Work											
Sweet-Nonsweet . (SVC-RASPR)	2311	Train (1156)	0.797	0.783	0.708	0.744	0.708	0.744	0.744	0.492	0.489
		Test (1155)	0.776	0.780	0.677	0.720	0.678	0.726	0.726	0.448	0.443
Bitter-Nonbitter . (SVC-RASPR)	2370	Train (1186)	0.870	0.804	0.811	0.807	0.835	0.819	0.808	0.614	0.613
		Test (1184)	0.732	0.622	0.793	0.728	0.651	0.636	0.702	0.419	0.419
Rojas et al.[70]											
Sweet -Tasteless	566	Train (396)	-	0.89	0.78	-	-	-	-	-	-
		Test (170)	-	0.96	0.77	-	-	-	-	-	-
Sweet-Bitter	508	Train (356)	-	0.75	0.75	-	-	-	-	-	-
		Test (152)	-	0.95	0.63	-	-	-	-	-	-
Tuwani et al.[71]											
Sweet-Non-sweet (2D RF-Boruta)	2366	Train (2205)	0.923	0.835	0.867	-	-	0.847	-	-	-
		Test (161)	0.863	0.683	0.943	-	-	0.798	-	-	-
Bitter-Non-bitter (2D AB-PCA)	2411	Train (2257)	0.863	0.723	0.860	.-	-	0.737	-	-	-
		Test (154)	0.868	0.793	0.874	-	-	0.849	-	-	-

Study 2

Intelligent Consensus Predictions of the Retention Index of Flavour and Fragrance Compounds Using 2D Descriptors

The goal of our second study is to predict a dataset effectively following the concept of regression while simplifying the overall mechanism and algorithm of the QSAR study. The second study is majorly developed on the pillar of 2D descriptors by using Intelligent consensus prediction. The necessary strategies like stepwise MLR along with best subset selection were taken into consideration for the descriptor thinning procedure. Finally, a consensus model results as a collective prediction of individual five PLS models. Later on, a detailed study of the corresponding applicability domain was performed. Thus, as a result, a greater area of chemical space was well demarcated where performing the chemical categorization was much easier.

4.2.1. Intelligent consensus prediction of the QSAR model while using five independent PLS model

The goal of this study is to create statistical models using simple and easily interpretable 2D descriptors. We have established various QSPR (PLS) models and validated them with different internationally accepted validation metrics. From the statistical results (summarized in Table 4.3), it was concluded that the developed models were accurate, predictive robust, and reproducible. Additionally, we have also conducted the applicability domain assessments (compounds situated outside the applicability domain criteria were considered outliers) and Y-randomization tests (to check whether models did not come by any chance) of developed models. We have also provided the probable mechanistic interpretation of the modelled descriptors that play a key role in determining the retention index of favor and fragrance compounds The scatter plots (given in Fig. 4.5) of the established models (M 1–M 5) show that the observed and predicted responses are quite similar and exhibit a good correlation.

4.2.2. Developed QSPR model for retention index

We have developed multiple regression-based QSPR models using the retention index (RI) of the favor and fragrance compounds as the endpoint. Intelligent consensus prediction was also employed to enhance the external prediction of the developed PLS models. The details of the modeled descriptors (models (M 1–M 5)) (provided in Supplementary Information 1) along with their meaning, contribution, and mechanistic interpretation of modelled descriptors are provided in Tables 4.4 and 4.5. Various PLS plots [88] (VIP plots (given in the supplementary file for study 2), loading plots (given in Figs. S6–S10 in Supplementary Information 2), score plots (given in the supplementary file of study 2), DModX plots (supplementary file of study 2), and Y-randomization plots (supplementary file of the study 2) were developed employing using SIMCA software (https://www. umetrics.com). The insights obtained from the developed models (M 2.1–M 2.5) for the retention index are explained in the Mechanistic interpretation section. The Y-randomization test and applicability domain (AD) assessment of the established models (M 2.1–M 2.5) were provided in the Y-randomization and Applicability domain section.

4.2.3 Y randomization of the PLS model

The Y-randomization test acts as a checkpoint whether the developed model is a result of a chance correlation or not. The X columns were fixed and the Y column was randomized with a different permutation and combination multiple times (here it is 100 times). The resulting randomized models were compared with the best-fitted model to analyze the significance of the developed models. The randomized model's fundamental validation statistics (R2 and Q2) should be poor when comparing it with the best fit model. The poor quality of the randomized models assures that the recently developed model is not a result of a chance correlation [88, 89]. Thus, the poor result of the randomized models indicates the acceptability of the developed model. The intercept value of R2 Y (within 0.3) and the intercept value of Q2 Y (within 0.05)

as validation statistics of the randomized models make the best-ftted model acceptable [88,89]. The Y randomized plots for each PLS model (model M 1–M 5) were given in (Supplementary file of study 2)

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

Model No.	Equation	Training Set						Test Set			
		R^2	Q^2	$r^2_{m\_loo}$	Δr^2_m	MAE_{train}	$RMSD_{train}$	Q^2F_1	Q^2F_2	$RMSD_{test}$	MAE_{test}
M1 **(LV-4)**	$RI = 157.448 +$ 6.555 ×MW+16.207× nAA-50.76× nR=Cp+94.838× nHDon-42.202× C-001+52.159× SdssC	0.909	0.907	0.866	0.080	57.126	96.168	0.945	0.945	73.756	52.250
M2 **(LV-4)**	$RI = -139.993 +$ 6.52 ×MW+9.966× C%-83.309× nR=Cp+87.463× nHDon-45.335× C-001+35.648× SdssC	0.918	0.916	0.879	0.073	52.648	91.246	0.945	0.945	73.679	49.835
M3 **(LV-4)**	$RI = 175.381 +$ 6.746 ×MW-86.038× nR=Cp+83.855× nHDon-60.406× C-001+58.776× SdssC+48.568× SaasC	0.915	0.914	0.875	0.075	54.593	92.718	0.943	0.943	74.928	52.039
M4 **(LV-4)**	$RI = 176.5111 +$ 6.551 ×MW+14.932×	0.908	0.907	0.865	0.082	57.196	96.479	0.943	0.943	75.463	53.577

	nAA-39.752× nROR+64.807× nHDon-46.849× C-001+43.92× SdssC										
M5 **(LV-4)**	$RI = -45.8096 +$ 6.5409 ×MW+6.76× C%+78.8267× nHDon-48.8858× C-001+36.6962× SdssC+27.4712× SaasC	0.913	0.911	0.872	0.077	54.648	94.188	0.943	0.943	75.372	51.420
CM0	Cumulative prediction from all input individual models.	-						0.948	0.948	-	41.053
CM1	Cumulative prediction from all individual qualified models.	-						0.948	0.948	-	41.053
CM2	Weighted average prediction from all qualified individual models.	-						0.949	0.949	-	39.930
CM3	**Best selection of prediction (compound-wise) from all qualified individual models.**	**-**						**0.950**	**0.950**	**-**	**38.447**

Here, **LV** represents the latent variables, **MAE** represents the mean absolute error, $\boldsymbol{R^2}$ is the determination coefficient, $\boldsymbol{Q^2}$ is the leave one out, whereas **RMSD** represents the root mean square standard deviation error. **CM0** = Ordinary consensus predictions. **CM1** = Average of predictions from individual models IM1 through IM5. **CM2** = Weighted average predictions from individual models IM1 through IM5. **CM3** = Best selection of predictions (compound-wise) from individual models IM1 through IM5. *Note that we have run the "Intelligent consensus predictor tool" using the options, AD: No; Dixon Q-test: No; Euclidean distance: No

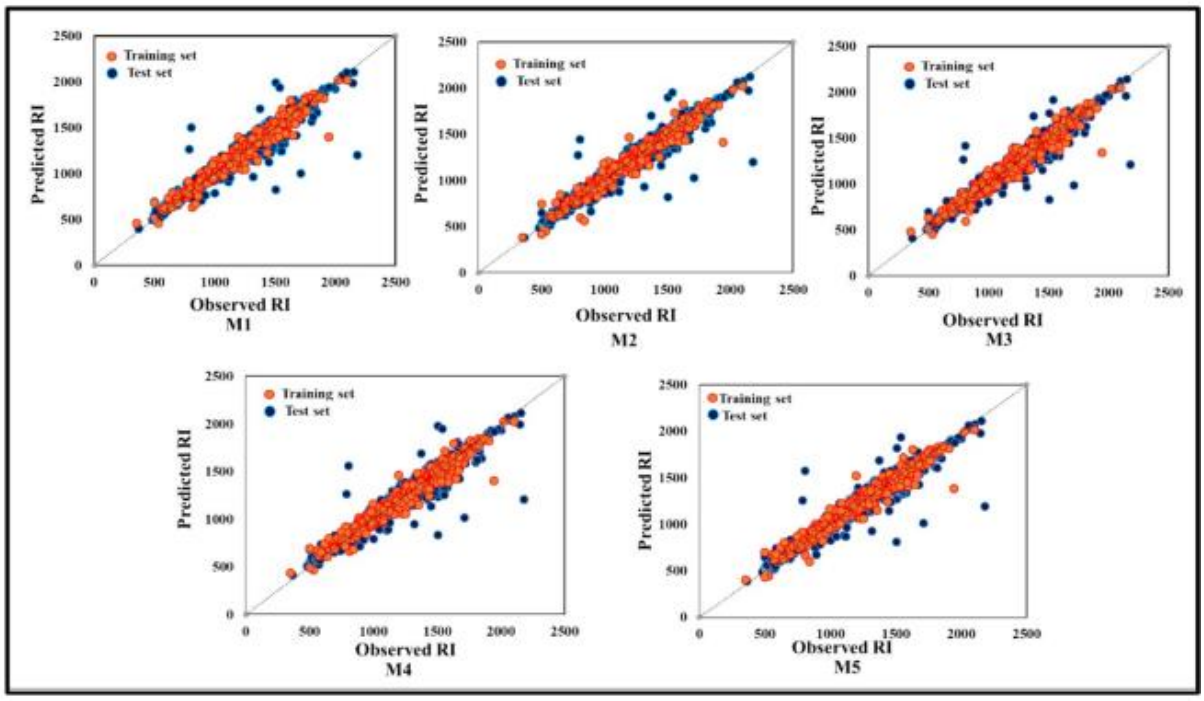


Fig.4.5 Statistical Plots of Study 2

4.2.4 Applicability Domain Assessment

The domain of applicability [90] was analyzed with the DModX approach using the SIMCA-P software (https:// www.umetrics.com). DModX plots of developed models (M1–M5) were provided (given in the supplementary file of study 2). From this assessment, it was observed that test set compounds 128, 661,745, 1002, and 1027 from Model 1; 361, 448, 745, 1002, and 1086 from Model 2; 10, 128,661,745 and 1027 from Model 3; 224, 425, 489, 594, 661, 1159, and 1170 from Model 4; 10, 128, 361, 656, 766, 1002, 1027, 1086, and 1184 from Model 5 are situated outside the domain of applicability (structural out lire).

4.2.5 Mechanistic interpretation of modelled descriptors

We have provided a probable mechanistic interpretation of the modelled descriptors, as per OECD guidelines 5. The type, meaning, contribution, and probable mechanistic interpretation of modelled descriptors are provided in Table.

4.2.6 PLS model interpretation

The first latent variable represents the geometrical property (in the form of MW, C%, nAA) and represents the size of molecules which is directly related to lipophilicity and leads h RI values (+ve contribution). Bulkiness and Partition coefficient (LOG P) are also dependent on molecular weight, leading to high lipophilicity in respective compounds (justified by structures of molecules too). The next significant latent variable is contributed by the descriptors SdssC, SaasC, nROR nR=Cp, and C-001 descriptors, and all of them together contribute to the electronic effect. nROR nR=Cp and C-001 have negative contributions but SdssC and SaasC have positive effects with low contribution; therefore, the overall contribution of this latent variable is negative toward the property endpoint which is also justified by the structures of molecules (presence of such features).

4.2.7 Comparison of the Recent Work

It is not possible to provide a strict comparison between the present study with related work due to the different composition of training and test set, total number of compounds used, number of variables used, etc., but we have tried to provide a possible comparison. Rojas et al. (2015) [70]and Rojas et al. (2015) [71] reported an in silico model using the retention index (RI) of 1184 flavour and fragrance compounds as an endpoint. The statistical results showed that the RMSD values for both the training and test sets were higher compared to the present work (the lower the RMSD value, the better the model quality). However, some of the previous studies lacked the reporting of exhaustive validation results in the form of different internationally accepted validation metrics, the use of simple and reproducible descriptors, specific findings (features responsible for the design and development of novel and suitable F&F compound), consensus prediction, as well as a wide domain of applicability. We have developed PLSICP models to assess the retention index (RI) of flavour and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D

descriptors and retention index (RI) as endpoints. The present work demonstrates better robustness, quality, reliability, and predictivity than the previously developed models. Our models were developed using a comparatively lower number of variables. Consensus predictions (in our case, the winner model is CM3) were also employed to improve the predictivity of the models. Our developed models have a wide domain of applicability and consist of simple, robust, reproducible, and easily interpretable 2D descriptors. Models were rigorously validated using internationally accepted validation metrics which show reliability, predictivity, and robustness. Some important features are reported in our study which will help design a novel and suitable F&F and related compounds. The comparison of the previous work (Rojas et al. (2015[70] and Rojas et al. (2015) [71] with the present study along with different validation metrics and ICP results is provided in Table 4.6.

Table 4.5. Type, meaning, contribution, and mechanistic interpretation of modeled descriptors.

Sl. No	Descriptors with contribution	Presence in developed PLS Model	Meaning of the descriptors	Type of descriptors	Mechanistic interpretation
1.	MW (+ve)	M1, M2, M3, M4, M5	Molecular weight	Constitutional index	This descriptor is directly related to the hydrophobicity (lipophilicity). Generally, lipophilic compounds may take more time for elution from the chromatographic column. Thus, a higher numerical value of this descriptor leads to a high RI value as shown in compound **2** (MW= 172.3, RI value = 1095) and inversely, it occurs in compound **675** (MW= 44.06, RI value=363) (given in **Fig.3**).
2.	nHDon (+ve)	M1, M2, M3, M4, M5	The number of donor atoms for H bonds	Functional group count	It was observed from the present dataset that compounds containing a higher number of hydrogen bond donors have also high molecular weight (MW) which is directly correlated with lipophilicity, resulting in high RI values as shown in compound **1191** (nHDon=2, RI value=2091, MW=241.31) and the absence of such atoms in any compounds leads to low RI value as shown in compound **674** (nHDon=0, RI value=350, MW=46.08) (given in **Fig.3**).
3.	C-001(-ve)	M1, M2, M3, M4, M5	The presence of CH3R/CH4 group	Atom-centred fragment	This descriptor signifies the branching in any compound that is inversely correlated with hydrophobicity and directly related to hydrophilicity. This phenomenon is demonstrated in compound **1192** (C-001=0, RI value=2147), and vice-versa occurs in compound **687** (C-001=4, RI value=590) (given in **Fig.3**).
4.	SdssC (+ve)	M1, M2, M3, M4, M5	The sum of dssC E-state	Atom-type E-state index	The positive correlation of this descriptor indicates that the presence of such fragments in any compound increases the RI value as shown in compound **1056** (SdssC=3.03, RI value=1802) and the absence of such fragments in any compound leads to a low RI value as shown in compound **894** (SdssC=0, RI value=1224). The presence of this fragment (=C<) reduces the polarity (hydrophilicity) of molecules. Thus, polarity and hydrophobicity are inversely related to each other (given in **Fig.3**).

5.	nR=Cp (-ve)	M1, M2, M3	Number of terminal sp^2 carbons	Functional group count	The presence of terminal sp^2 carbon indicates a significant enhancement in branching in any molecules which reduces the hydrophobic (lipophilic) character of the molecules and ultimately reduces the RI value of the organic flavor and fragrance compounds. This phenomenon is demonstrated in compound **1046** (nR=Cp=0, RI value=2180) and oppositely occurs in compound **192** (nR=Cp=1, RI value=1228) (given in **Fig.4.6**).

5.	nR=Cp (-ve)	M1, M2, M3	Number of terminal sp^2 carbons	Functional group count	The presence of terminal sp^2 carbon indicates a significant enhancement in branching in any molecules which reduces the hydrophobic (lipophilic) character of the molecules and ultimately reduces the RI value of the organic flavor and fragrance compounds. This phenomenon is demonstrated in compound **1046** (nR=Cp=0, RI value=2180) and oppositely occurs in compound **192** (nR=Cp=1, RI value=1228) (given in **Fig.4.6**).

6.	C% (+ve)	M2, M5	The percentage of C atoms	Constitutional index	A high percentage of carbon atoms (large-carbon skeleton molecules) in any compound leads to enhancement in hydrophobicity (lipophilicity) which leads to a high RI value as shown in compound **1192** (C%=48.57, RI value=2147), and the inverse phenomenon occurs in compound **674** (C%=22.22, RI value=350) (given in **Fig.4.6**).
7.	nAA (+ve)	M1, M4	The number of aromatic atoms	Constitutional index	Aromatic compounds contain a hydrophobic nucleus which contributes towards non-polarity. Non-polar compounds are hydrophobic (a high RI value) in nature. Thus, the presence of more such fragments (aromatic atoms) in compounds leads to high RI values as shown in compound **1192** (nAA=12, RI value=2147), and vice-versa occurs in compound **276** (nAA=5, RI value=1217) (given in **Fig.4.6**). Presence of aromatic ring lead to increase in size of molecules, ultimately enhancing the lophilicity.
8.	SaasC (+ve)	M3, M5	The sum of aaaC E-states	Atom-type E-state index	This descriptor signifies the presence of an aromatic substitution in any compound. Aromaticity is inversely related to polarity [51] and, consequently directly related to hydrophobicity. Thus, the presence of such a structure

					fragment reduces the RI value as demonstrated in compound **360** (SaasC=6.039, RI value=1837) and vice-versa occurs in compound **219** (SaasC=2.019, RI value=1227) (given in **Fig.4.6**).
9.	nROR (-ve)	M4	The number of aliphatic ethers	Functional group count	Generally, ethers (C-O bond of ether) are polar in nature [52]. Therefore, the presence of such fragments (aliphatic ethers) in any molecule enhances the polarity and consequently hydrophilicity of the compound. Hydrophilicity and RI are inversely related to each other. Therefore, the presence of such a fragment reduces the RI value as shown in compound **559** (nROR=3, RI value=763) and an inverse phenomenon occurs in compound **900** (nROR=1, RI value=1235) (given in **Fig.4.6**).

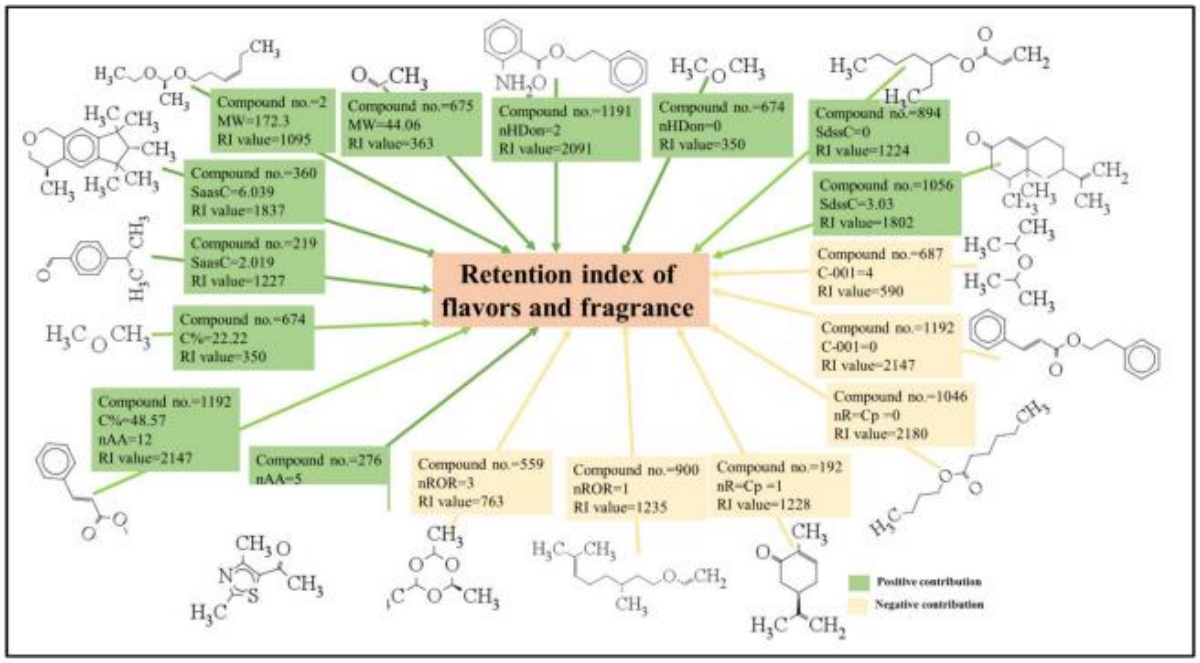


Fig 4.6 Structural correlation of chemical compounds with retention index

4.2.8 Advantages and implementation of the present work

We have developed regression-based QSPR models using 2D descriptors and the GA-PLS method (avoid any chances of inter-correlation among descriptors) to assess the retention index of flavour and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D descriptors and rigorously validated with various internationally accepted validation metrics (both external and internal validation metrics) in compliance with the OECD guidelines to check the robustness, reliability, predictivity, and domain of applicability. Consensus predictions were also employed to improve the external predictivity and domain of applicability of the developed models (in our case, CM3 is the winner model). Some important findings regarding RI of F& F compounds were observed from this study: hydrophobicity, the presence of larger fragments, high molecular weight, and aromaticity were responsible for the high RI value (+ve contribution) of the flavour and fragrance compounds, while polarity and hydrophilicity reduce (−ve contribution) the retention index of the flavour and fragrance compounds. Hence, this information can be used for the selection and optimization of the stationary phase according to the available organic compounds (flavour and

fragrance compounds) and for achieving the desired retention index. Finally, developed models can be used for data gap fling (prediction of RI value of untested and new compounds within the domain of applicability); consequently, this information (with known calculated RI values) can be used in the flavour and fragrance industry to identify unknown compounds (by comparing with RI values) in complex mixtures by reducing time, cost, the need of highly skilled labour, costly instrumentation, and complexity of experimentation. Thus, developed models will help design and develop suitable and novel flavours and fragrances that fulfill the product's requirement before experimental verification.

Table 4.6. Comparison with the previous work by Rojas et al. (2015a) and Rojas et al. (2015b).

Developed model	Total number of compounds used	No. of compounds on the training set and test set.	No. of features in the initial pool	Type of the features	No. of features in the final model	R^2_{train}	R^2_{test}	$RMSD_{train}$	$RMSD_{test}$
Present work									
Model 1	**Initially 1208, and after curation 1194.**	**894 in the training set and 298 in the test set.**	**309**	**2D**	**6 (LV-4)**	**0.909**	**-**	**96.168**	**73.756**
Model 2					**6 (LV-4)**	**0.918**	**-**	**91.246**	**73.756**
Model 3					**6 (LV -4)**	**0.915**	**-**	**92.718**	**74.928**
Model 4					**6 (LV -4)**	**0.908**	**-**	**96.479**	**75.463**
Model 5					**6 (LV -4)**	**0.913**	**-**	**94.188**	**75.372**
Previous Rojas et al. (2015) [70]	1206	N_{train}=400, N_{val}=405, N_{test} =403	1815 conformational descriptors.	2D	4	0.910	0.93	100.94	82.99
Rojas et al. (2015)[71]	Initially 1206 and after curation 1184	N_{train}=395,N_{val}=396, N_{test}=393	1815 non-conformational descriptors.	2D	7	0.902	0.904	137.60	121.978

Chapter-5

Conclusion

5.Conclusion

A proper investigation of some organoleptic compounds and their impactful properties often helps to understand the upcoming effects on the area of their implication. This research work was done with the objective of understanding the influence of organoleptic compounds and their properties while focusing on the implication of some conceptual, simplest algorithms that are predominant for quality prediction. The proper analysis of the organoleptic chemicals whether it is physiochemical, or potential helps to better understand their ultimate fate considering their utility. Restructure and design of the chemicals according to the resulting feature, interpretation, and chemical categorization often help to meet the desired goal of risk management and effective use of the chemicals.

In the first study of the recent dissertation, we have developed a classification-based c-RASPR model while employing the concept of machine learning. With the progress of the study, the responsible significant features were identified for further detailed interpretation regarding the explanation of statistical concepts concerning the corresponding biological responses. Moreover, in our second work, we have developed QSPR models with an implication of intelligent consensus prediction. For this study also the resulting 2D descriptors were recognized and vividly interpreted with the physiochemical and structural phenomenon following the respective biological responses. In these two studies, we have explored the predominant concept of QSAR. Intelligent consensus model prediction and RASPR are nothing but the extended implications of QSAR. While Intelligent consensus prediction is restricted within the idea of QSAR while giving an aggregate judgment of several validated results, the concept of RASPR extended towards read across. It's often taken into account that RASAR gives its prediction opinion on the similarity parameter (chemical, biological, and structural) of ten close source compounds. In these two studies, we have tried a detailed understanding of

the QSAR and RASAR algorithms for estimating the effective use of the chemical compounds as well as the necessary optimization of the chemical structures to achieve the desired goal. The core objective is to develop a reliable, simplified technique.

5.1 The first application of machine learning-based classification read-across structure-property relationship (c-RASPR) modeling for sweet and bitter

The identification of contributing features for both sweet and bitter compounds is vital for taste-sensing mechanisms. In the investigation, two large and diverse data sets were used to develop classification-based predictive models. Initially, preliminary QSPR models were developed using the most discriminating features (MDF), which provided moderate prediction results but left suggestions for further improvement. Although the prediction results of the models using this QSPR method were of moderate quality, they provided suggestions for improving the prediction quality. These QSPR models also give information about the important features that regulate the properties of the sweet and bitter tastes of the organic compounds.

In the next segment, the LDA-RASPR model, which combines QSPR and Read-across techniques, showed better prediction quality for both sweet and bitter data sets than the corresponding LDA-QSPR models. Additionally, machine learning algorithms (ML) were applied to both sweet and bitter data sets with RASPR descriptors, and the Support Vector Classification (SVC) algorithm provided the best results. The comparison of simple LDA-QSPR and ML-RASPR methods showed that the latter outperforms the former in terms of predictive quality for both data sets. This suggests that the concept of ML itself enhances the learning experience and can be used along with the methodology of RASPR for enhanced model prediction quality. In general, we reconciled the laboratory studies and developed predictive models that suggest that the presence of polar groups (C-018, nArCOOH, C-033, F01[C-N], B01[O-S], F02[N-S], nRNHR) and hydrophobicity parameters (max_conj_path,

NsssN, MW, O-062, C-018, F01[N-O], nArCOOH, nNq, C-033, nRNHR) are essential factors for sweet/bitter compounds molecules for ligand binding for both sweet and bitter activities. In conclusion, the hybrid method of the RASPR algorithm, along with ML, provides a more authentic and reliable methodology for chemometric model development. The increasing prediction quality trend suggests that a hybrid method (ML-RASPR) is preferable over the QSPR methodology for model prediction quality enhancement. The developed simple classification-based models with a limited number of RASPR descriptors could be an efficient alternative approach for the identification of sweet/bitter compounds with a low number of regressing variables.

5.2 Intelligent Consensus Predictions of the Retention Index of Flavour and Fragrance Compounds Using 2D Descriptors.

In the current study, regression-based QSPR models were developed using the PLS method to assess the retention index of flavour and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D descriptors and retention index (RI) as endpoints. Feature selection was performed using different strategies (such as the stepwise selection method and the Genetic Algorithm (GA)) to extract the most significant descriptors contributing to the property endpoint (retention index). We have rigorously validated the developed models using various globally accepted validation metrics (both external and internal validation metrics) in compliance with the OECD (Organization for Economic Cooperation and Development) principles. Consensus predictions were also employed to improve the external predictiveness of the developed models (in our case, CM3 is the winner model). From the statistical results, it was concluded the developed models are robust, reliable, predictive, and wide domain of applicability. From the mechanistic interpretation, it was observed that hydrophobicity, the presence of larger fragments, high molecular weight, and aromaticity enhance the retention index (RI) of the flavour and fragrance compounds. In

contrast, polarity and hydrophilicity reduce the retention index of the flavour and fragrance compounds. Hence, this information can be used for the selection and optimization of the stationary phase according to the available organic compounds (flavour and fragrance compounds) and for achieving the desired retention index. Finally, developed models can be used to predict the RI values for any new or unknown compound (data gap fling), consequently, this information (with known calculated RI values) can be used in the flavour and fragrance industry to identify unknown compounds (by comparing with RI values) in complex mixtures by reducing the time, cost, and complexity of experimentation. Thus, developed models will be helpful in designing suitable and novel flavours and fragrances that fulfill the product's requirements before experimental verification.

Chapter-6

References

References

1. Sari I, Sinaga P, Hernani H. The impact of industrial revolution 4.0 on basic chemistry learning. InAIP Conference Proceedings 2021 Apr 2 (Vol. 2331, No. 1). AIP Publishing.
2. Hurley DW. Lead chemicals—compliance with environmental regulations. Journal of Vinyl Technology. 1982 Mar;4(1):10-5.
3. Hopkins A. Risk-management and rule-compliance: Decision-making in hazardous industries. Safety science. 2011 Feb 1;49(2):110-20.
4. Scruggs CE, Ortolano L, Wilson MP, Schwarzman MR. Effect of company size on potential for REACH compliance and selection of safer chemicals. Environmental science & policy. 2015 Jan 1;45:79-91.
5. Canter DA. Role of the regulatory agencies in the activities of the National Toxicology Program. Regulatory Toxicology and Pharmacology. 1981 Jun 1;1(1):8-18.
6. ((https://books.google.com/books?hl=en&lr=&id=89BIAwAAQBAJ&oi=fnd&pg=PP1&dq=chemistry+clubed+with+mathematics+and+chemistry&ots=R7Qi5ltf7q&sig=GkagsY1xwfrWrZL5uyWVHNu_UR0
7. Wishart DS. Introduction to cheminformatics. Current protocols in bioinformatics. 2016 Mar;53(1):14-.
8. Silva MM, Reboredo FH, Lidon FC. Food colour additives: A synoptical overview on their chemical properties, applications in food products, and health side effects. Foods. 2022 Jan 28;11(3):379.
9. Iizuka K. Is the use of artificial sweeteners beneficial for patients with diabetes mellitus? The advantages and disadvantages of artificial sweeteners. Nutrients. 2022 Oct 22;14(21):4446.

10. Rastogi SC, Heydorn S, Johansen JD, Basketter DA. Fragrance chemicals in domestic and occupational products. Contact dermatitis. 2001 Oct;45(4):221-5.

11. https://www.google.co.in/books/edition/Common_Fragrance_and_Flavor_Materials/0jFdJAooDL0C?hl=en&gbpv=1&dq=chemical+nature+of+the+flavor+and+fragrance+compound&pg=PP2&printsec=frontcover)

12. Soboleva E, Ambrus Á. Application of a system suitability test for quality assurance and performance optimisation of a gas chromatographic system for pesticide residue analysis. Journal of Chromatography A. 2004 Feb 20;1027(1-2):55-65.

13. Roy K, Banerjee A. Tools, Applications, and Case Studies (q-RA and q-RASAR). Inq-RASAR: A Path to Predictive Cheminformatics 2024 Jan 26 (pp. 51-88). Cham: Springer Nature Switzerland.

14. Lancaster WA, Praissman JL, Poole FL, Cvetkovic A, Menon AL, Scott JW, Jenney FE, Thorgersen MP, Kalisiak E, Apon JV, Trauger SA. A computational framework for proteome-wide pursuit and prediction of metalloproteins using ICP-MS and MS/MS data. BMC bioinformatics. 2011 Dec;12:1-2.

15. Cros AF. *Action de l'alcool amylique sur l'organisme* (Doctoral dissertation, Faculté de médecine de Strasbourg).

16. Brown AC, Fraser TR. V.—On the connection between chemical constitution and physiological action. part. i.—on the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. Earth and Environmental Science Transactions of The Royal Society of Edinburgh. 1868 Jan;25(1):151-203.

17. Borman S. New QSAR techniques eyed for environmental assessments. Chem. Eng. News. 1990 Feb 19;68(8):20-3.

18. Hansch C, Leo A, Taft RW. A survey of Hammett substituent constants and resonance and field parameters. Chemical reviews. 1991 Mar 1;91(2):165-95.

19. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature. 1962 Apr 14;194(4824):178-80.

20. Hammett LP. The effect of structure upon the reactions of organic compounds. Benzene derivatives. Journal of the American Chemical Society. 1937 Jan;59(1):96-103.

21. Free SM, Wilson JW. A mathematical contribution to structure-activity studies. Journal of medicinal chemistry. 1964 Jul;7(4):395-9.

22. Kubinyi H. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. Quantitative Structure-Activity Relationships. 1988;7(3):121-33.

23. Fujita T, Winkler DA. Understanding the roles of the "two QSARs". Journal of chemical information and modeling. 2016 Feb 22;56(2):269-74.

24. Ming D. https://scholar. google. com/citations? user= 0O8Oka0AAAAJ&hl= enhttps://scholar. google. com/citations? user= 0O8Oka0AAAAJ&hl= en.

25. Mauri A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs. 2020:801-20.

26. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. The Journal of Physical Chemistry A. 1998 May 21;102(21):3762-72.

27. Verloop A. The STERIMOL approach to drug design. New York: Marcell Decker. 1987.

28. Mannhold R, Rekker RF. The hydrophobic fragmental constant approach forcalculating log P in octanol/water and aliphatic hydrocarbon/water systems. Perspectives in Drug Discovery and Design. 2000 Jun;18:1-8.

29. Gutman I, Trinajstić N. Graph theory and molecular orbitals. Total φ-electron energy of alternant hydrocarbons. Chemical physics letters. 1972 Dec 15;17(4):535-8.

30. Randic M. Characterization of molecular branching. Journal of the American Chemical Society. 1975 Nov;97(23):6609-15.

31. Kier LB, LH H. The nature of structure-activity relationships and their relation to molecular connectivity.

32. Kier LB, LH H. The nature of structure-activity relationships and their relation to molecular connectivity.

33. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature. 1962 Apr 14;194(4824):178-80.

34. Hansch C. Quantitative approach to biochemical structure-activity relationships. Accounts of chemical research. 1969 Aug 1;2(8):232-9.

35. Free SM, Wilson JW. A mathematical contribution to structure-activity studies. Journal of medicinal chemistry. 1964 Jul;7(4):395-9.

36. Fujita T, Ban T. Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. Journal of Medicinal Chemistry. 1971 Feb;14(2):148-52.

37. Guner OF. History and evolution of the pharmacophore concept in computer-aided drug design. Current topics in medicinal chemistry. 2002 Dec 1;2(12):1321-32.

38. Robinson DD, Winn PJ, Lyne PD, Richards WG. Self-organizing molecular field analysis: A tool for structure− activity studies. Journal of medicinal chemistry. 1999 Feb 25;42(4):573-83.

39. Macchiarulo A, Gioiello A, Thomas C, Massarotti A, Nuti R, Rosatelli E, Sabbatini P, Schoonjans K, Auwerx J, Pellicciari R. Molecular field analysis and 3D-quantitative structure− activity relationship study (MFA 3D-QSAR) unveil novel features of bile acid recognition at TGR5. Journal of chemical information and modeling. 2008 Sep 22;48(9):1792-801.

40. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of medicinal chemistry. 1985 Jul;28(7):849-57.

41. Chuman H, Karasawa M, Fujita T. A novel three-dimensional QSAR procedure: Voronoi field analysis. Quantitative Structure-Activity Relationships. 1998 Aug;17(04):313-26.

42. Hahn M. Receptor surface models. 1. Definition and construction. Journal of medicinal chemistry. 1995 Jun;38(12):2080-90.

43. Amat L, Besalú E, Carbó-Dorca R, Ponec R. Identification of active molecular sites using quantum-self-similarity measures. Journal of Chemical Information and Computer Sciences. 2001 Jul 23;41(4):978-91.

44. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. Journal of medicinal chemistry. 1994 Nov;37(24):4130-46.

45. Todeschini R, Gramatica P, Provenzani R, Marengo E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling

physicochemical properties of polyaromatic hydrocarbons. Chemometrics and Intelligent Laboratory Systems. 1995 Feb 1;27(2):221-9.

46. Crivori P, Cruciani G, Carrupt PA, Testa B. Predicting blood− brain barrier permeation from three-dimensional molecular structure. Journal of medicinal chemistry. 2000 Jun 1;43(11):2204-16.

47. Jain AN, Koile K, Chapman D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. Journal of Medicinal Chemistry. 1994 Jul;37(15):2315-27.

48. Kirkpatrick P. Gliding to success. Nature Reviews Drug Discovery. 2004 Apr 1;3(4):299-.

49. Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. QSAR & Combinatorial Science. 2006 Mar;25(3):235-51.

50. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018 Jul 26;300:70-9.

51. Saxena AK, Prathipati P. Comparison of mlr, pls and ga-mlr in qsar analysis. SAR and QSAR in Environmental Research. 2003 Oct 1;14(5-6):433-45.

52. Roy K. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert opinion on drug discovery. 2007 Dec 1;2(12):1567-77.

53. Chirico N, Gramatica P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. Journal of chemical information and modeling. 2012 Aug 27;52(8):2044-58.

54. Schultz TW, Amcoff P, Berggren E, Gautier F, Klaric M, Knight DJ, Mahony C, Schwarz M, White A, Cronin MT. A strategy for structuring and reporting a read-across

prediction of toxicity. Regulatory Toxicology and Pharmacology. 2015 Aug 1;72(3):586-601.

55. Roy K, Banerjee A. Tools, Applications, and Case Studies (q-RA and q-RASAR). Inq-RASAR: A Path to Predictive Cheminformatics 2024 Jan 26 (pp. 51-88). Cham: Springer Nature Switzerland.

56. Das K, Behera RN. A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering. 2017 Feb;5(2):1301-9.

57. https://github.com/cosylabiiit/bittersweet/

58. Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. QSAR & Combinatorial Science. 2006 Mar;25(3):235-51.

59. Goodarzi M, Dejaegher B, Heyden YV. Feature selection methods in QSAR studies. Journal of AOAC International. 2012 May 1;95(3):636-51.

60. Banerjee A, Roy K. Prediction-inspired intelligent training for the development of classification read-across structure–activity relationship (c-RASAR) models for organic skin sensitizers: assessment of classification error rate from novel similarity coefficients. Chemical Research in Toxicology. 2023 Aug 16;36(9):1518-31.

61. Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. Journal of the American statistical association. 2011 Dec 1;106(496):1566-77.

62. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS medicine. 2008 Aug;5(8):e165.

63. Hewitt M, Ellison CM, Enoch SJ, Madden JC, Cronin MT. Integrating (Q) SAR models, expert systems and read-across approaches for the prediction of developmental toxicity. Reproductive toxicology. 2010 Aug 1;30(1):147-60.

64. Banerjee A, Roy K. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. Chemical Research in Toxicology. 2023 Feb 22;36(3):446-64.

65. Pisner DA, Schnyer DM. Support vector machine. InMachine learning 2020 Jan 1 (pp. 101-121). Academic Press.

66. Stoltzfus JC. Logistic regression: a brief primer. Academic emergency medicine. 2011 Oct;18(10):1099-104.

67. Rigatti SJ. Random forest. Journal of Insurance Medicine. 2017 Jan 1;47(1):31-9.

68. Jeon J, Seo N, Son SB, Lee SJ, Jung M. Application of machine learning algorithms and shap for prediction and feature analysis of tempered martensite hardness in low-alloy steels. Metals. 2021 Jul 22;11(8):1159.

69. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. Journal of Molecular Graphics and Modelling. 2008 Jun 1;26(8):1315-26.

70. Rojas C, Duchowicz PR, Tripaldi P, Diez RP. QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemometrics and Intelligent Laboratory Systems. 2015 Jan 15;140:126-32.

71. Rojas C, Ballabio D, Consonni V, Tripaldi P, Mauri A, Todeschini R. Quantitative structure–activity relationships to predict sweet and non-sweet tastes. Theoretical Chemistry Accounts. 2016 Mar;135:1-3.

72. Todeschini R, Consonni V. Handbook of molecular descriptors. John Wiley & Sons; 2008 Jul 11.

73. . Kumar A, Ojha PK, Roy K. The first report on the assessment of maximum acceptable daily intake (MADI) of pesticides for humans using intelligent consensus predictions. Environmental Science: Processes & Impacts. 2024;26(5):870-81.

74. Roy K, Narayan Das R. A review on principles, theory and practices of 2D-QSAR. Current drug metabolism. 2014 May 1;15(4):346-79.

75. Kumar A, Ojha PK, Roy K. QSAR modeling of chronic rat toxicity of diverse organic chemicals. Computational Toxicology. 2023 May 1;26:100270.

76. De P, Bhattacharyya D, Roy K. Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling. Structural Chemistry. 2020 Jun;31(3):1043-55.

77. Roy K, Das RN, Ambure P, Aher RB. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems. 2016 Mar 15;152:18-33.

78. Roy K, Kar S, Das RN. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press; 2015 Mar 3.

79. Kumar A, Kumar V, Ojha PK, Roy K. Chronic aquatic toxicity assessment of diverse chemicals on Daphnia magna using QSAR and chemical read-across. Regulatory Toxicology and Pharmacology. 2024 Mar 1;148:105572.

80. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems. 2001 Oct 28;58(2):109-30.

81. Jancy S, Preetha R. Application of Multivariate Statistical Analysis for Food Safety and Quality Assurance. InMathematical and Statistical Applications in Food Engineering 2020 Jan 30 (pp. 263-275). CRC Press.

82. Kumar A, Podder T, Kumar V, Ojha PK. Risk assessment of aromatic organic chemicals to T. pyriformis in environmental protection using regression-based QSTR

and Read-Across algorithm. Process Safety and Environmental Protection. 2023 Feb 1;170:842-54.

83. Khan K, Jillella GK, Gajewicz-Skretna A. Integrated Modeling of Organic Chemicals in Tadpole Ecotoxicological Assessment: Exploring Qstr, Q-Rasar, and Intelligent Consensus Prediction Techniques. Q-Rasar, and Intelligent Consensus Prediction Techniques.

84. Roy K, Ambure P, Kar S, Ojha PK. Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models. Journal of Chemometrics. 2018 Apr;32(4):e2992.

85. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. Journal of medicinal chemistry. 2019 Sep 12;63(16):8761-77.

86. Tuwani R, Wadhwa S, Bagler G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. Scientific reports. 2019 May 9;9(1):7155.

87. Xiu H, Liu Y, Yang H, Ren H, Luo B, Wang Z, Shao H, Wang F, Zhang J, Wang Y. Identification of novel umami molecules via QSAR models and molecular docking. Food & Function. 2022;13(14):7529-39.

88. Kumar A, Ojha PK, Roy K. QSAR modeling of chronic rat toxicity of diverse organic chemicals. Computational Toxicology. 2023 May 1;26:100270.

89. Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. Journal of chemical information and modeling. 2007 Nov 26;47(6):2345-57.

90. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems. 2015 Jul 15;145:22-9.

Supplementary file

Supplementary for Study 1

https://drive.google.com/drive/folders/1h5Esa0zFE85XdNf8O4AGo_VrzVyjAud0?usp=drive_link

Supplementary for Study 2

The online version contains supplementary material available at https://doi.org/10.1007/s10 .

APPENDIX

Reprints

Chromatographia
https://doi.org/10.1007/s10337-024-04349-5

ORIGINAL

Intelligent Consensus Predictions of the Retention Index of Flavor and Fragrance Compounds Using 2D Descriptors

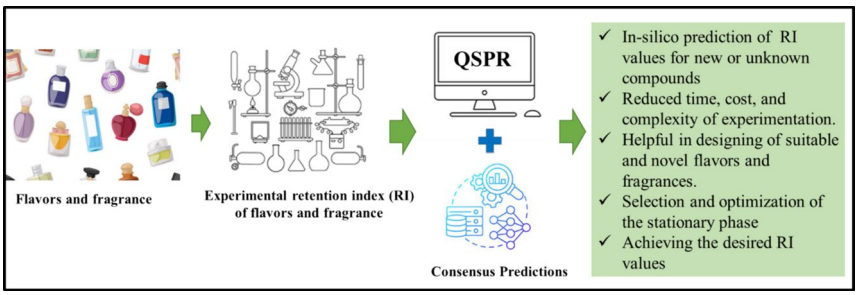
Doelima Bera[1] · Ankur Kumar[2] · Joyita Roy[1] · Kunal Roy[1]

Received: 13 June 2024 / Revised: 6 July 2024 / Accepted: 10 July 2024

Abstract

The demand for novel flavors and fragrance (F&F) compounds has increased, highlighting the need for a systematic design approach. Currently, the F&F industry relies heavily on experimental approaches without considering the potential consequences of altering the features that contribute to the fragrance of the compound. In silico approaches have great potential to identify the necessary features for creating novel F&F compounds. In the present study, Quantitative Structure–Property Relationship (QSPR) models were developed using 1208 compounds and simple 2D descriptors, focusing on the RI (retention index) as the endpoint to predict the olfactory properties of molecules. Feature selection was initially carried out by multi-layered stepwise regression followed by feature thinning using the Genetic Algorithm (GA) and optimal feature combination selection using the BSS (best subset selection) method. Final models were developed using the Partial Least Squares (PLS) method. Additionally, internal and external validation of the models was performed using different validation metrics suggesting that the developed models are reliable, predictive, reproducible, and robust. To enhance the external prediction of the developed models, an Intelligent Consensus Prediction (ICP) method was employed and **CM3** (consensus model 3) (best selection of predictions (compound-wise) from individual models) was found to provide the best predictivity. The modeling descriptors suggested that the hydrophobicity, high molecular weight, aromaticity, and presence of large-size fragments (high percentage of carbon) enhance the RI values. Conversely, polarity and hydrophilicity decrease the RI values. This study can be used to optimize the stationary phase according to the flavor and fragrance compounds to obtain the desired retention index (RI values).

Graphical abstract



Keywords Flavor and fragrance (F&F) molecules · RI (retention Index) · 2D descriptors · QSAR · ICP (intelligent consensus prediction)

Introduction

The use of fragrance and flavor (F&F) is widespread in various consumer products. Fragrance compounds create pleasant smells, while flavor compounds contribute to taste

Extended author information available on the last page of the article

Published online: 18 July 2024

sensations [1]. These compounds have specific structures and activities that determine their sensory effects. They include alcohols, aldehydes, ketone esters, and lactones [2]. Industries such as food and pharmaceuticals use these compounds to mask unpleasant tastes [3]. Fragrances are essential in perfume, beverages, cosmetics, food, and pharmaceuticals. Most synthetic chemical compounds mimicking natural products are used in F&F compound industries [4]. The global F&F market is expected to reach USD 36.49 billion by 2029, with a compound annual growth rate of 4.7% [5]. The demand for novel F&F compounds is driven by safety and environmental regulations. However, designing F&F compounds still relies on empirical techniques, which can be tedious and time consuming, leading to limited exploration of potential candidates [6]. The traditional approaches of trial and error are tedious, resource intensive, and time consuming [7]. This method also leads to a limited exploration of the potential candidates. Thus, there remains a high chance of missing a potent candidate for F&F to be incorporated into consumer products. The launching of such products to the market can also be costly. The retention time is crucial for formulating new fragrance compounds in the perfume industry. It helps identify the chemical structure of a compound and allows comparison of its retention data across different GC systems. Chromatography is an important tool in various industries for ensuring the production of high-quality products, and it plays a crucial role in quality control. This method involves measuring the retention time or retention index of a compound as it passes through a gas chromatographic column's glass capillary. There is a growing need and interest in developing structure–odor relationship models using the structure of fragrance compounds. A recent study utilized the retention index to develop in silico chemometric models for these compounds in the chromatographic column [8]. Manual sniffing and recording can be an inefficient and complicated process resulting in numerous errors. For example, the ambiguity of gas chromatography or gas chromatography–mass spectrometry test values cannot alter the real fragrance retention index [9]. Other factors, such as environmental conditions and differences in individual olfactory sensitivity, can also affect the reported reading. Therefore, to address the challenges involved in the design of fragrance molecules, a systematic framework should be developed for designing and screening suitable fragrances that fulfill the product's requirement before experimental verification.

Several researchers have tried to develop computational techniques to explain the perceptual and physicochemical space of fragrance molecules. Rojas et al. (2015) also analyzed the retention index of 1184 fragrance-like compounds on a stationary phase using QSPR methods [8]. Rojas and colleagues [10] researched flavor and fragrance compounds to develop Quantitative Structure–Property Relationships (QSPR) models. Keller et al. (2017) performed a machine learning (ML) algorithm to predict intensity, pleasantness, and semantic descriptors from the structural information of odor compounds [11]. Dua et al. (2008) worked on retention time by taking 43 aromatic constituents of saffron [12]. Furthermore, Sharma et al. (2020) conducted QSPR studies to predict the retention indices of fragrance compounds in stationary phases with three different polarities [13]. In addition, Villa et al. (2017) conducted QSPR studies to predict the retention indices of fragrance compounds in stationary phases with three different polarities [14]. In 2022, Kumar et al. (2022) reported QSPR modeling of fragrance compounds on the carbowax glass capillary using gas chromatography using 1179 flavor and fragrance compounds for model development [15]. Noorizadeh et al. (2011) also analyzed the retention index of essential oils using QSPR methods [16]. Pourbasheer et al. (2015) reported QSPR models to calculate the GC retention indices of essential oils [17]. Liu et al. (2021) reported a QSPR model for the assessment of fragrance retention grades for monomer flavors [18]. Ahmadi et al. (2024) predict the retention indices of volatile organic compounds using the QSPR model [19]. Riahi et al. (2008) assessed the retention indices of essential oil compounds using GA-MLR methods [20]. Kumar et al. (2022) reported QSRR models of flavors and fragrance compounds studied on the stationary phase methyl silicone OV-101 column in gas chromatography using correlation intensity index and consensus modeling [21].

Several machine learning methods such as neural networks and SVR (support vector machine) have been also used to develop QSPR models for the assessment of RI indices of the various compounds. Keller et al. (2017) applied a machine learning (ML) algorithm to predict intensity, pleasantness, and semantic descriptors from the structural information of odor compounds [11]. Maulana et al. (2020) employed an artificial neural network to assess the Kovats retention indices for fragrance and flavor [22]. Matyushin et al. (2020) used multimodal machine learning for the calculation of the gas chromatographic retention index [23]. Wang et al. (2021) reported machine learning models for the assessment of RI of compounds in beers [24]. Agustia et al. (2022) employed Support Vector Regression to calculate the Kovats retention indices of flavors and fragrances [25]. Matyushin et al. (2019) estimated the gas chromatographic retention indices employing deep convolutional neural networks [26]. K et al. (2019) reported machine learning models for GC–MS fingerprint profiling of food flavor prediction [27]. Vrzal et al. (2021) reported a Deep learning-based gas chromatographic retention index predictor (DeepReI) [28]. Matyushin et al. (2021) also reported deep learning-based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases [29]. Vigneau et al. (2018) employed Random forests

(a machine learning methodology) to highlight the volatile organic compounds involved in olfactory perception [30]. However, some of the previous studies lacked the reporting of exhaustive validation results in the form of different internationally accepted validation metrics, the use of simple and reproducible descriptors, specific findings (features responsible for the design and development of novel and suitable F&F compound), consensus prediction, as well as a wide domain of applicability.

The primary objective of the current work is to predict the quality and retention index of various compounds with unknown retention indexes (new and untested F&F compounds) to avoid the time, complexity of the experimental process, high cost, highly skilled labor, and expensive experimental equipment. In this work, we have developed QSPR models only using 2D descriptors which are simple due to a direct mathematical algorithm for calculation, reproducible, and easily interpretable, in order to avoid the complexity of 3D analysis and energy minimization. 2D descriptors have a deluge of contributions in extracting chemical attributes, and some are also capable of representing 3D features to some extent [31]. However, it is not possible to differentiate between the isomers (cis, trans, etc.) of compounds completely using 2D-QSPR models. We used only 2D descriptors and did not use 3D descriptors. This means we did not account for the geometric isomers and considered them as a pure and individual form. We also did not consider the effect of the cis or trans isomeric form. Therefore, it is reasonable for us to gather the average retention index value from Rojas et al.[10] to develop a more accurate and understandable model. The developed QSPR models will help in identifying and distinguishing features of chemical compounds, ultimately aiding in determining their retention index for both polar and non-polar stationary phases. Thus, developed QSPR models can also be used to optimize stationary phases. The present QSPR models were established using 1208 data points (significantly more than previously reported) which will provide a wider domain of applicability (can calculate the RI of a wide range of F&F and related compounds). The models were developed by considering only simple, reproducible, and easily interpretable 2D descriptors, making them simpler, more reliable, robust, and more accurate when dealing with medium to large datasets with retention index as the endpoint. The developed Partial Least Squares (PLS) models were further used for consensus modeling to enhance the predictivity of the test set fragrance compounds, thus showing higher predictivity and a wide domain of applicability. An applicability domain (AD) was defined to increase the reliability of the prediction model. This work will provide a reliable model for predicting Retention Index (RI) values for unevaluated and un-synthesized flavors and fragrances and related compounds, making it a valuable asset for professionals in the field of aroma, flavor chemical synthesis, and perfume blending. This study also provides detailed and advanced knowledge about some important features responsible for the RI of F&F and related compounds: hydrophobicity, the presence of larger fragments, hydrogen donor groups, and aromaticity were responsible for the high RI value (+ ve contribution) of the flavor and fragrance compounds, while polarity and hydrophilicity reduce (-ve contribution) the retention index of the flavor and fragrance compounds. Thus, the present study aims to develop and design suitable and novel flavors and fragrances as per the product's requirement, data gap filling (related to the RI value of new and untested compounds), and an alternative to complex, time-consuming, and costly analytical testing techniques.

Materials and Methods

Dataset Collection

It is essential to have consistent and reliable data for the development of QSPR models. In this study, 1208 data points for aromatic substances were collected from the literature [10] for model development. The researchers [10] reported the experimental property as the Kováts retention index (RI) in a non-polar stationary capillary column (0.28 mm × 50 m). They used methyl silicone OV-101 as coating material admixed with 1% Carbowax 20 M, and the column was programmed to increase from 80 to 200 °C at a rate of 2 °C/min. The RI values used as an endpoint ranged from 350 to 2180 [8, 10]. The Kováts retention index is independent of individual chromatographic system specifications and allows comparing values measured by different analytical laboratories and analysis times. The fragrance ingredients are often obtained from commercial suppliers as mixtures of isomers (e.g., cis–trans), which the supplier does not separate. However, we cannot neglect the effect of temperature, pH, and surrounding environment for transforming a particular isomeric form of a chemical compound to another (like a transformation of a cis compound to trans and trans compound to cis form) for a mixture of compounds while it was supplied for experimentation. This consequence may result in exceptional responses where a single compound represents two different isomeric mixtures with the same molecular weight. In this scenario the compounds like Allyl anthranilate 1 and Allyl anthranilate 2 may not represent the pure cis or trans isomeric form of a compound rather they were represented as a mixture of both the geometrical isomers. In the present study, it was interpreted as a single compound with an isomeric mixture while considering the impact of other external factors as well. In that case, collecting the average retention index value (compounds with quite similar chromatographic peaks) of Rojas et al. is justified

for further development of an accurate and interpretable model. This kind of approximation is very common in any 2D-QSPR analysis.

Molecular Representation and Data Curation

A total of 1208 flavor and fragrance compounds, each with its corresponding SMILES, chemical names, and retention index, were initially compiled (provided in supplementary information 1). To ensure accuracy, for compounds with more than one reported retention index value, the average value was calculated, and duplicate entries were removed, resulting in a final curated dataset of 1194 compounds. The structural representation of the compounds was done using Marvin Sketch software (https://chemaxon.com/marvin). Additionally, a curated SDF file of the flavor and fragrance compounds was obtained after incorporating explicit hydrogen, ring aromatization, and 2D form cleaning for the descriptor calculations.

Descriptor Calculation

In this study, we used the Alvadesc software (https://www.alvascience.com/alvadesc) to calculate descriptors for flavor and fragrance compounds. These descriptors are numerical values that define the physiochemical properties of a compound. We have used only simple, direct mathematical algorithms nature, reproducible, and easily interpretable 2D descriptors [32] to avoid the complexity of 3D analysis and energy minimization [33]. 2D descriptors have a deluge of contributions in extracting chemical attributes and some are capable of representing 3D features to some extent [31]. However, it is not possible to differentiate between the isomers (cis, trans, etc.) of compounds completely using 2D-QSPR models. The work of Rojas et al. had already concluded that 3D descriptors did not significantly improve the quality parameters of the QSPR model. From the previous conclusion, we have decided to develop simpler 2D-QSPR models while using the concept of intelligent consensus predictions. Lastly, the redirection toward the source data, the unseparated mixture of both the geometrical isomers of a particular compound, and their response values indicate an inseparable form of cis and trans isomers even after the application of 3D descriptors. In the present study, the isomers were recognized as a single compound. In that case, collecting the average retention index value (compounds with quite similar chromatographic peaks) of Rojas et al. is justified for further development of an accurate and interpretable model. This kind of approximation is very common in any 2D-QSPR analysis. A total of 2400 2D descriptors were calculated, including constitutional descriptors (molecular composition of a referenced compound), ETA indices (extended topochemical atom), ring (information related to the presence of ring descriptors), functional group count, atom-centered fragment, connectivity index, atom-type E-state (description related to the electronic state of the atoms), 2D atom pair, and molecular properties [32, 33]. Additionally, data pre-treatment was performed using the DataPreTreatmentGUI_1.2 (http://teqip.jdvu.ac.in/QSAR_Tools/) software to eliminate correlated (correlation cut-off of 0.95) and descriptors having low variance cut-off (less than or equal to 0.01), resulting in a total of 309 curated descriptors for further modeling.

Dataset Division

Partitioning the dataset is an essential step in developing the QSAR model. A chemometric statistical model requires two independent datasets: a training set for developing the model and a test set for validating the model [34]. Generally, the whole dataset was divided into the training and test set in the ratio of 70:30 (approx.). In the present investigation, the dataset of fragrance and flavor compounds was divided into four clusters based on their properties (sorted response-based method.) using the Dataset Division 1.2 tool (http://teqip.jdvu.ac.in/QSAR_Tools/). This property-based data division resulted in a training set of 896 compounds and a test set of 298 flavored and fragrance compounds.

Test-Training Pre-Treatment

The training and test set data may contain correlated and noisy descriptors that are not relevant to the data modeling purpose. Therefore, pre-treating both the training and test sets is necessary. In our study, we utilized the Data Pre-treatment tool (http://teqip.jdvu.ac.in/QSAR_Tools/) to pre-treat the training and test sets after division, using a variance cut of 0.01 and a correlation cut-off of 0.95. This process resulted in 162 less correlated descriptors, ultimately minimizing the error in model development.

Feature Selection and Model Development

The selected features after pre-treatment were utilized for the feature selection process. Genetic algorithm (GA) followed by BSS (Best Subset Selection) (http://teqip.jdvu.ac.in/QSAR_Tools/) was used for feature selection [35]. Initially, some features were also selected using the stepwise selection method. Stepwise regression can be defined as a multiple linear regression which was evolved with the step-by-step mechanism. After removing the selected features from the first stepwise run, the stepwise method was again performed with the remaining pool of descriptors. Besides stepwise feature selection, GA was also performed for the feature selection procedure. GA tool has many advantages over other feature selection methods. It is based on fitness

function on mean absolute error (MAE)-based pick-up criteria. We have employed our in-house tool "GeneticAlgorithm_v4.1_Train" (http://teqip.jdvu.ac.in/QSAR_Tools/) to find the most relevant descriptors with the RI endpoint. The best subset selection (BSS) approach was used to find the optimal combination of descriptors for a robust prediction model. After selecting the best descriptors from both feature selection methods, we performed partial least squares (PLS) regression to build the preliminary QSPR models. PLS methods were employed to develop the final robust models to avoid any chances of inter-correlation among descriptors. The PLS regression method is a generalized technique of the "Multiple Linear Regression (MLR)" method, where we can examine strongly collinear, correlated, noisy data and many X variables. The PLS regression has been carried out with a Java-based software tool "PLS_SingleY_version" (http://teqip.jdvu.ac.in/QSAR_Tools/). The PLS model was further utilized for best subset selection (BSS). The best subset selection was performed with the in-house tool developed in our laboratory (http://teqip.jdvu.ac.in/QSAR_Tools/). Six descriptor models (five PLS models) were generated based on MAE-based criteria [36].

Model Validation Criteria

The developed QSTR models were rigorously validated via various internationally accepted metrics to ensure the robustness, predictivity, goodness of fit, and quality of the models. For training set compounds, internal validation metrics such as cross-validated correlation coefficient $Q^2_{(LOO)}$ (leave one out), $r^2_{m\_loo,}$ MAE_{train} (mean absolute error), $RMSD_{train}$ (root mean square standard deviation error), and coefficient of determination R^2 were calculated to measure the robustness and goodness of fit of the model. For test set compounds, we have predicted external set compounds using globally accepted different validation metrics like predictive MAE_{test}, $RMSD_{test}$, R^2 (R^2_{pred}), or Q^2_{F1} and Q^2_{F2} to judge the predictability of the model [37].

Applicability Domain Assessment

The applicability domain is the biological, chemical, or physiochemical hypothetical space of the training set chemicals through the recently created QSPR model. The main use of this domain is to predict the toxicity value of compounds that fall in this domain and have unknown values. We have used the DModX (distance to mean X) approach to predict the AD of the PLS models (OECD principle 3) using SIMCA-P software [38–40]. The DModX uses Y and X residuals as diagnostic values to ensure model quality. If the DModX value is greater than the critical value, it means that the query compound is outside the domain of the model [36, 38–40]:

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

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

Intelligent Consensus Predictor (ICP)

This method evaluates the performance of the consensus models in comparison to the individual models based on MAE-based criteria (i.e., 95%). It is recognized that a single model may not be able to accurately predict all of the test compounds. This implies that one QSPR model may be more suitable for one test compound, while another model may be better for a different test compound [33, 41, 42]. A specific QSPR model may not be equally effective in predicting all query compounds in the query list. To get the best prediction results, we need to consider the consensus of all the predictions made by these four models. For this, consensus prediction should be made intelligently, i.e., in a query compound-specific way, using all or most of the valid models. This is different from doing a simple average of predictions from all available models. Consensus prediction is better than individual model predictions since it combines all the good characteristics of each model. Thus, the drawbacks of one individual model are taken care of by other models (s). This makes the predictions less biased, more reliable, and more precise. The individual models may have differently defined applicability domains, while the consensus method combines the ADs of the individual models, thus providing a greater chemical space coverage as well. Moreover, the consensus method does not affect the quality of the internal statistical parameters of the individual models [43]. In the present study, we have chosen five models (M1–M5) to conduct a consensus prediction using the "Intelligent Consensus Predictor" tool that is available on our laboratory website (http://teqip.jdvu.ac.in/QSAR_Tools/. The steps involved in developing the models are depicted in Fig. 1.

Results and Discussion

The goal of this study is to create statistical models using simple and easily interpretable 2D descriptors. We have established various QSPR (PLS) models and validated them with different internationally accepted validation metrics**.** From the statistical results (summarized in Table 1),

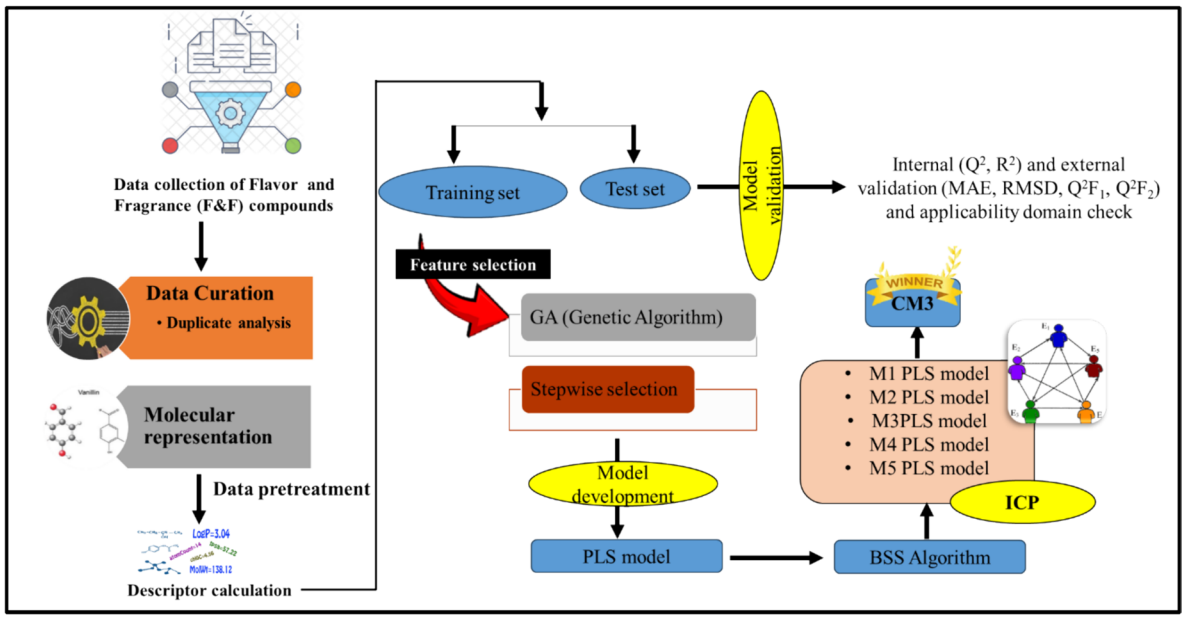


Fig. 1 Schematic representation of the present study

it was concluded that the developed models were accurate, predictive robust, and reproducible. Additionally, we have also conducted the applicability domain assessments (compounds situated outside the applicability domain criteria were considered outliers) and Y-randomization tests (to check whether models did not come by any chance) of developed models. We have also provided the probable mechanistic interpretation of the modeled descriptors that play a key role in determining the retention index of flavor and fragrance compounds The scatter plots (given in Fig. 2) of the established models (M1–M5) show that the observed and predicted responses are quite similar and exhibit a good correlation.

Developed QSPR Models for the Retention Index (RI)

We have developed multiple regression-based QSPR models using the retention index (RI) of the flavor and fragrance compounds as the endpoint. Intelligent consensus prediction was also employed to enhance the external prediction of the developed PLS models. The details of the modeled descriptors (models (M1–M5)) (provided in Supplementary Information 1) along with their meaning, contribution, and mechanistic interpretation of modeled descriptors are provided in Table 2. Various PLS plots [34, 35] (VIP plots (given in Figs. S1–S5 in Supplementary Information 2), loading plots (given in Figs. S6–S10 in Supplementary Information 2), score plots (given in Figs. S11–S15 in Supplementary Information 2), DModX plots (given in Figs. S16–S25 in Supplementary Information 2), and Y-randomization plots (given in Figs. S26–S30 in Supplementary Information 2) were developed employing using SIMCA software (https://www.umetrics.com). The insights obtained from the developed models (M1–M5) for the retention index are explained in the Mechanistic interpretation section. The Y-randomization test and applicability domain (AD) assessment of the established models (M1–M5) were provided in the Y-randomization and Applicability domain section.

Y Randomization of the PLS Models

The Y-randomization test acts as a checkpoint whether the developed model is a result of a chance correlation or not. The X columns were fixed and the Y column was randomized with a different permutation and combination multiple times (here it is 100 times). The resulting randomized models were compared with the best-fitted model to analyze the significance of the developed models. The randomized model's fundamental validation statistics (R^2 and Q^2) should be poor while comparing it with the best-fitted model. The poor quality of the randomized models assures that the recently developed model is not a result of a chance correlation [34, 44]. Thus, the poor result of the randomized models indicates the acceptability of the developed model. The intercept value of R^2Y (within 0.3) and the intercept value of Q^2Y (within 0.05) as validation statistics of the randomized models make the best-fitted model acceptable [34, 44]. The Y randomized plots for each PLS model (model M1–M5) were given in (given in Figs. S26–S30 in Supplementary Information 2).

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

Model No	Equation	Training set						Test set			
		R^2	Q^2	$r^2_{m\_loo}$	Δr^2_m	MAE_{train}	$RMSD_{train}$	Q^2F_1	Q^2F_2	$RMSD_{test}$	MAE_{test}
M1 (LV-3)	$RI = 157.448 + 6.555\times MW + 16.207\times nAA\text{-}50.76\times(nR=Cp) + 94.838\times nHDon\text{-}42.202\times C\text{-}001 + 52.159\times SdssC$	0.909	0.907	0.866	0.080	57.126	96.168	0.945	0.945	73.756	52.250
M2 (LV-4)	$RI = -139.993 + 6.52\times MW + 9.966\times C\%\text{-}83.309\times(nR=Cp) + 87.463\times nHDon\text{-}45.335\times C\text{-}001 + 35.648\times SdssC$	0.918	0.916	0.879	0.073	52.648	91.246	0.945	0.945	73.679	49.835
M3 (LV-4)	$RI = 175.381 + 6.746\times MW\text{-}86.038\times(nR=Cp) + 83.855\times nHDon\text{-}60.406\times C\text{-}001 + 58.776\times SdssC + 48.568\times SaasC$	0.915	0.914	0.875	0.075	54.593	92.718	0.943	0.943	74.928	52.039
M4 (LV-4)	$RI = 176.5111 + 6.551\times MW + 14.932\times nAA\text{-}39.752\times nROR + 64.807\times nHDon\text{-}46.849\times C\text{-}001 + 43.92\times SdssC$	0.908	0.907	0.865	0.082	57.196	96.479	0.943	0.943	75.463	53.577
M5 (LV-4)	$RI = -45.8096 + 6.5409\times MW + 6.76\times C\% + 78.8267\times nHDon\text{-}48.8858\times C\text{-}001 + 36.6962\times SdssC + 27.4712\times SaasC$	0.913	0.911	0.872	0.077	54.648	94.188	0.943	0.943	75.372	51.420
CM0	Cumulative prediction from all input individual models	–						0.948	0.948	–	41.053
CM1	Cumulative prediction from all individual qualified models	–						0.948	0.948	–	41.053
CM2	Weighted average prediction from all qualified individual models	–						0.949	0.949	–	39.930
CM3	Best selection of prediction (compound-wise) from all qualified individual models	–						0.950	0.950	–	38.447

Here, LV represents the latent variables, MAE represents the mean absolute error, R^2 is the determination coefficient, Q^2 is the leave one out, whereas RMSD represents the root mean square standard deviation error. CM0 = Ordinary consensus predictions. CM1 = Average of predictions from individual models IM1 through IM5. CM2 = Weighted average predictions from individual models IM1 through IM5. CM3 = Best selection of predictions (compound-wise) from individual models IM1 through IM5. *Note that we have run the "Intelligent consensus predictor tool" using the options, AD: No; Dixon Q-test: No; Euclidean distance: No.

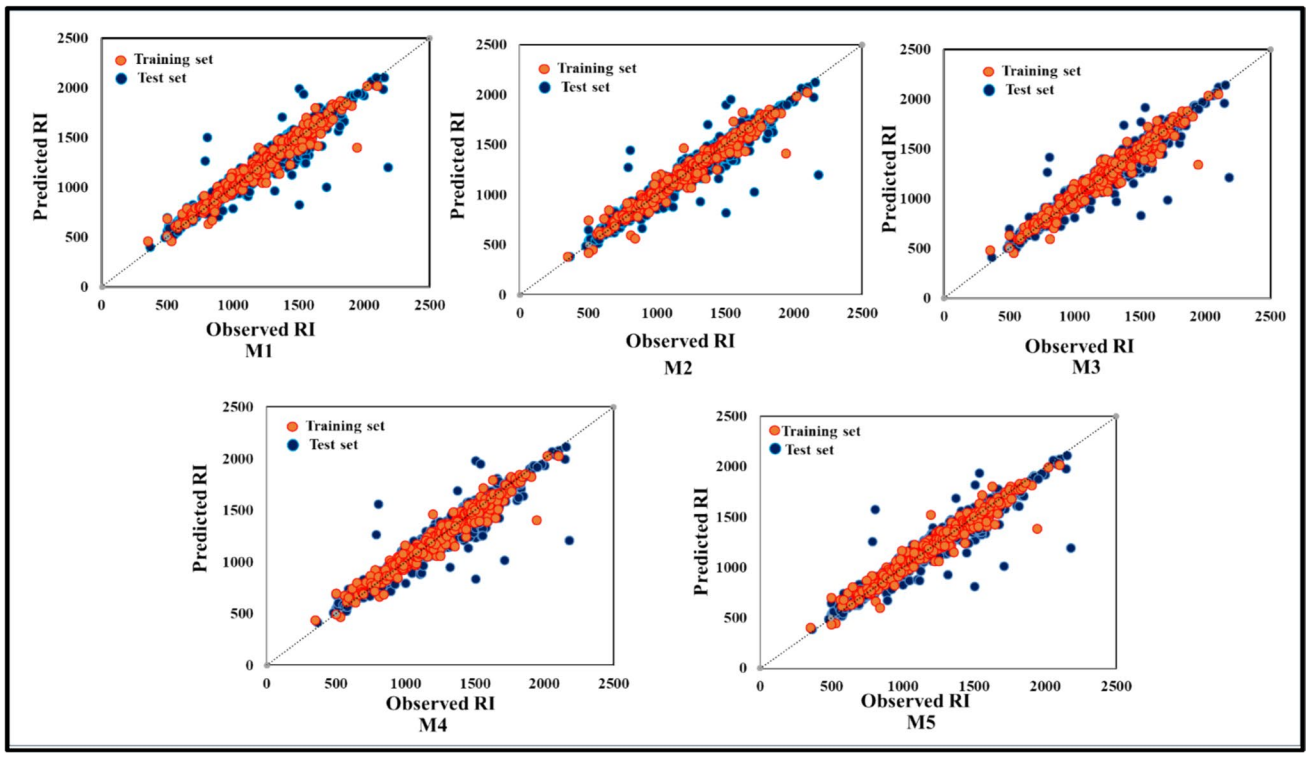


Fig. 2 Scatter plots of the developed models

Applicability Domain Assessment

The domain of applicability [45] was analyzed with the DModX approach using the SIMCA-P software (https://www.umetrics.com). DModX plots of developed models (M1–M5) were provided (given in Figs. S16–S25 in supplementary information 2). From this assessment, it was observed that test set compounds 128, 661,745, 1002, and 1027 from Model 1; 361, 448, 745, 1002, and 1086 from Model 2; 10, 128,661,745 and 1027 from Model 3; 224, 425, 489, 594, 661, 1159, and 1170 from Model 4; 10, 128, 361, 656, 766, 1002, 1027, 1086, and 1184 from Model 5 are situated outside the domain of applicability (structural outliers).

Mechanistic Interpretation of the Modeled Descriptors

We have provided a probable mechanistic interpretation of the modeled descriptors, as per OECD guidelines 5. The type, meaning, contribution, and probable mechanistic interpretation of modeled descriptors are provided in Table 2**.**

PLS Model Interpretation

The first latent variable represents the geometrical property (in the form of MW, C%, nAA) and represents the size of molecules which is directly related to lipophilicity and leads to high RI values (+ve contribution). Bulkiness and Partition coefficient (LOG P) are also dependent on molecular weight, leading to high lipophilicity in respective compounds (justified by structures of molecules too). The next significant latent variable is contributed by the descriptors SdssC, SaasC, nROR nR = Cp, and C-001 descriptors, and all of them together contribute to the electronic effect. nROR nR = Cp and C-001 have negative contributions but SdssC and SaasC have positive effects with low contribution; therefore, the overall contribution of this latent variable is negative toward the property endpoint which is also justified by the structures of molecules (presence of such features).

Comparison of the Recent Work

It is not possible to provide a strict comparison between the present study with related work due to the different composition of training and test set, total number of compounds used, number of variables used, etc., but we have tried to provide a possible comparison. Rojas et al. (2015) [8] and Rojas et al. (2015) [10] reported an in silico model using the retention index (RI) of 1184 flavor and fragrance compounds as an endpoint. The statistical results showed that the RMSD values for both the training and test sets were higher compared to the present work (the lower the RMSD value, the better the model quality). However, some of the previous studies lacked the reporting of

Table 2 Type, meaning, contribution, and mechanistic interpretation of modeled descriptors

Sl. No	Descriptors with contribution	Presence in developed PLS model	Meaning of the descriptors	Type of descriptors	Mechanistic interpretation
1	MW (+ ve)	M1, M2, M3, M4, M5	Molecular weight	Constitutional index	This descriptor is directly related to the hydrophobicity (lipophilicity) [46]. Generally, lipophilic compounds may take more time for elution from the chromatographic column. Thus, a higher numerical value of this descriptor leads to a high RI value as shown in compound 2 (MW = 172.3, RI value = 1095) and inversely, it occurs in compound 675 (MW = 44.06, RI value = 363) (given in Fig. 3)
2	nHDon (+ ve)	M1, M2, M3, M4, M5	The number of donor atoms for H bonds	Functional group count	It was observed from the present dataset that compounds containing a higher number of hydrogen bond donors have also high molecular weight (MW) which is directly correlated with lipophilicity, resulting in high RI values [46, 47] as shown in compound 1191 (nHDon = 2, RI value = 2091, MW = 241.31), and the absence of such atoms in any compounds leads to low RI value as shown in compound 674 (nHDon = 0, RI value = 350, MW = 46.08) (given in Fig. 3)
3	C-001(− ve)	M1, M2, M3, M4, M5	The presence of CH3R/CH4 group	Atom-centered fragment	This descriptor signifies the branching in any compound that is inversely correlated with hydrophobicity and directly related to hydrophilicity [48, 49]. This phenomenon is demonstrated in compound 1192 (C-001 = 0, RI value = 2147), and vice versa occurs in compound 687 (C-001 = 4, RI value = 590) (given in Fig. 3)
4	SdssC (+ ve)	M1, M2, M3, M4, M5	The sum of dssC E-state	Atom-type E-state index	The positive correlation of this descriptor indicates that the presence of such fragments in any compound increases the RI value as shown in compound 1056 (SdssC = 3.03, RI value = 1802) and the absence of such fragments in any compound leads to a low RI value as shown in compound 894 (SdssC = 0, RI value = 1224). The presence of this fragment (= C <) reduces the polarity (hydrophilicity) of molecules. Thus, polarity and hydrophobicity are inversely related to each other (given in Fig. 3)
5	nR = Cp (− ve)	M1, M2, M3	Number of terminal sp^2 carbons	Functional group count	The presence of terminal sp^2 carbon indicates a significant enhancement in branching in any molecules which reduces the hydrophobic (lipophilic) character of the molecules and ultimately reduces the RI value [49] of the organic flavor and fragrance compounds. This phenomenon is demonstrated in compound 1046 (nR = Cp = 0, RI value = 2180) and oppositely occurs in compound 192 (nR = Cp = 1, RI value = 1228) (given in Fig. 3)

Table 2 (continued)

Sl. No	Descriptors with contribution	Presence in developed PLS model	Meaning of the descriptors	Type of descriptors	Mechanistic interpretation
6	C% (+ve)	M2, M5	The percentage of C atoms	Constitutional index	A high percentage of carbon atoms (large-carbon skeleton molecules) [34] in any compound leads to enhancement in hydrophobicity (lipophilicity) which leads to a high RI value as shown in compound 1192 (C% = 48.57, RI value = 2147), and the inverse phenomenon occurs in compound 674 (C% = 22.22, RI value = 350) (given in Fig. 3)
7	nAA (+ve)	M1, M4	The number of aromatic atoms	Constitutional index	Aromatic compounds contain a hydrophobic nucleus which contributes toward non-polarity. Non-polar compounds are hydrophobic (a high RI value) in nature [50]. Thus, the presence of more such fragments (aromatic atoms) in compounds leads to high RI values as shown in compound 1192 (nAA = 12, RI value = 2147), and vice versa occurs in compound 276 (nAA = 5, RI value = 1217) (given in Fig. 3). The presence of aromatic ring leads to increase in size of molecules, ultimately enhancing the lipophilicity
8	SaasC (+ve)	M3, M5	The sum of aaaC E-states	Atom-type E-state index	This descriptor signifies the presence of an aromatic substitution in any compound. Aromaticity is inversely related to polarity [51] and, consequently directly related to hydrophobicity. Thus, the presence of such a structure fragment reduces the RI value as demonstrated in compound 360 (SaasC = 6.039, RI value = 1837) and vice versa occurs in compound 219 (SaasC = 2.019, RI value = 1227) (given in Fig. 3)
9	nROR (−ve)	M4	The number of aliphatic ethers	Functional group count	Generally, ethers (C-O bond of ether) are polar in nature [52]. Therefore, the presence of such fragments (aliphatic ethers) in any molecule enhances the polarity and consequently hydrophilicity of the compound. Hydrophilicity and RI are inversely related to each other. Therefore, the presence of such a fragment reduces the RI value as shown in compound 559 (nROR = 3, RI value = 763) and an inverse phenomenon occurs in compound 900 (nROR = 1, RI value = 1235) (given in Fig. 3)

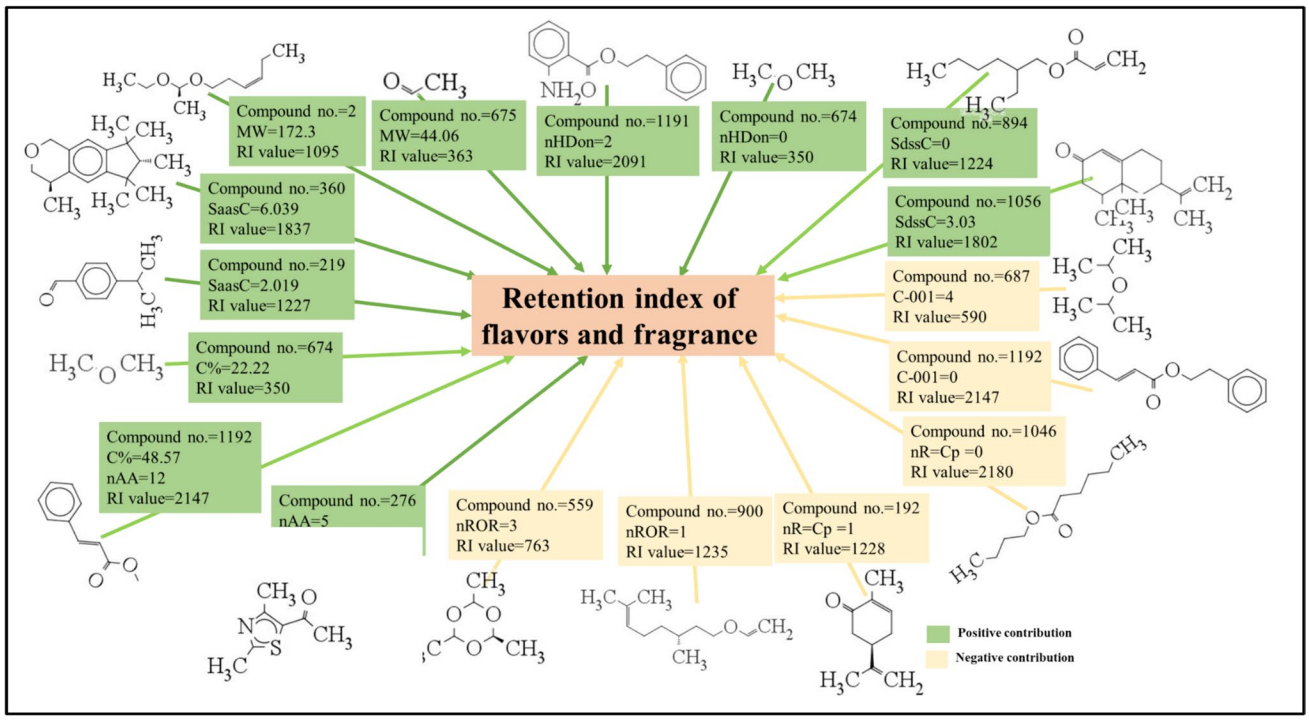


Fig. 3 Mechanistic interpretation of the developed models

exhaustive validation results in the form of different internationally accepted validation metrics, the use of simple and reproducible descriptors, specific findings (features responsible for the design and development of novel and suitable F&F compound), consensus prediction, as well as a wide domain of applicability. We have developed PLS-ICP models to assess the retention index (RI) of flavor and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D descriptors and retention index (RI) as endpoints. The present work demonstrates better robustness, quality, reliability, and predictivity than the previously developed models. Our models were developed using a comparatively lower number of variables. Consensus predictions (in our case, the winner model is CM3) were also employed to improve the predictivity of the models. Our developed models have a wide domain of applicability and consist of simple, robust, reproducible, and easily interpretable 2D descriptors. Models were rigorously validated using internationally accepted validation metrics which show reliability, predictivity, and robustness. Some important features are reported in our study which will help design a novel and suitable F&F and related compounds. The comparison of the previous work (Rojas et al. (2015) [8] and Rojas et al. (2015) [10]) with the present study along with different validation metrics and ICP results is provided in Table 3.

Advantages and Implication of the Present Work

We have developed regression-based QSPR models using 2D descriptors and the GA-PLS method (avoid any chances of inter-correlation among descriptors) to assess the retention index of flavor and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D descriptors and rigorously validated with various internationally accepted validation metrics (both external and internal validation metrics) in compliance with the OECD guidelines to check the robustness, reliability, predictivity, and domain of applicability. Consensus predictions were also employed to improve the external predictivity and domain of applicability of the developed models (in our case, CM3 is the winner model). Some important findings regarding RI of F& F compounds were observed from this study: hydrophobicity, the presence of larger fragments, high molecular weight, and aromaticity were responsible for the high RI value (+ ve contribution) of the flavor and fragrance compounds, while polarity and hydrophilicity reduce (−ve contribution) the retention index of the flavor and fragrance compounds. Hence, this information can be used for the selection and optimization of the stationary phase according to the available organic compounds (flavor and fragrance compounds) and for achieving the desired retention index. Finally, developed models can be used for data gap filling (prediction of RI value of untested and new compounds

Table 3 Comparison with the previous work by [8, 10]

Developed model	Total number of compounds used	No. of compounds on the training set and test set	No. of features in the initial pool	Type of the features	No. of features in the final model	R^2_{train}	R^2_{test}	$RMSD_{train}$	$RMSD_{test}$
Present work	Initially 1208, and after curation 1194	894 in the training set and 298 in the test set	309	2D	6 (LV-3)	0.909	–	96.168	73.756
Model 1									
Model 2					6 (LV-4)	0.918	–	91.246	73.756
Model 3					6 (LV -4)	0.915	–	92.718	74.928
Model 4					6 (LV -4)	0.908	–	96.479	75.463
Model 5					6 (LV -4)	0.913	–	94.188	75.372
Previous Rojas et al. (2015) [8]	1206	$N_{train} = 400$, $N_{val} = 405$, $N_{test} = 403$	1815 conformational descriptors	2D	4	0.910	0.93	100.94	82.99
Rojas et al. (2015)[10]	Initially 1206 and after curation 1184	$N_{train} = 395$, $N_{val} = 396$, $N_{test} = 393$	1815 non-conformational descriptors	2D	7	0.902	0.904	137.60	121.978

within the domain of applicability); consequently, this information (with known calculated RI values) can be used in the flavor and fragrance industry to identify unknown compounds (by comparing with RI values) in complex mixtures by reducing time, cost, the need of highly skilled labor, costly instrumentation, and complexity of experimentation. Thus, developed models will help design and develop suitable and novel flavors and fragrances that fulfill the product's requirement before experimental verification.

Conclusion

In the current study, regression-based QSPR models were developed using the PLS method to assess the retention index of flavor and fragrance compounds. Models were developed using simple, reproducible, and easily interpretable 2D descriptors and retention index (RI) as endpoints. Feature selection was performed using different strategies (such as the stepwise selection method and the Genetic Algorithm (GA)) to extract the most significant descriptors contributing to the property endpoint (retention index). We have rigorously validated the developed models using various globally accepted validation metrics (both external and internal validation metrics) in compliance with the OECD (Organization for Economic Cooperation and Development) principles. Consensus predictions were also employed to improve the external predictivity of the developed models (in our case, CM3 is the winner model). From the statistical results, it was concluded the developed models are robust, reliable, predictive, and wide domain of applicability. From the mechanistic interpretation, it was observed that hydrophobicity, the presence of larger fragments, high molecular weight, and aromaticity enhance the retention index (RI) of the flavor and fragrance compounds. In contrast, polarity and hydrophilicity reduce the retention index of the flavor and fragrance compounds. Hence, this information can be used for the selection and optimization of the stationary phase according to the available organic compounds (flavor and fragrance compounds) and for achieving the desired retention index. Finally, developed models can be used to predict the RI values for any new or unknown compound (data gap filling), consequently, this information (with known calculated RI values) can be used in the flavor and fragrance industry to identify unknown compounds (by comparing with RI values) in complex mixtures by reducing the time, cost, and complexity of experimentation. Thus, developed models will be helpful in designing suitable and novel flavors and fragrances that fulfill the product's requirement before experimental verification.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10337-024-04349-5.

Acknowledgements DB thanks the All India Council for Technical Education (AICTE), New Delhi for providing support in the form of a fellowship. AK thanks GPC Regulatory India Private Limited for financial support in the form of a project assistant (GPC Regulatory India Private Limited sponsored research, Ref No-P-1/RS/171/22, date-07-09.2022).

Author Contributions DB contributed to data curation, formal analysis, validation, and writing – initial draft. AK and JR were involved in validation and writing – initial draft. KR performed conceptualization, supervision, and writing – editing.

Funding This study was supported by All India Council for Technical Education, New Delhi.

Data Availability The DTC Lab software tools are available from http://teqip.jdvu.ac.in/QSAR_Tools/ (DATA CURATION, KNIME workflow). The chemical's name, SMILES value, and Retention Index

value of the entire datasets and model descriptors for QSPR (M1–M5) models (training and test set) have been made available in Supporting Information 1.

Declarations

Conflict of Interest The authors declare no competing interests.

Ethical Approval Not applicable.

Consent to Participate This article does not contain any studies with human participants or animals, clinical trial registration, or plant reproducibility performed by any author.

Consent for Publication All authors have approved this paper and agree with its publication.

References

1. Rastogi SC, Heydorn S, Johansen JD, Basketter DA (2001) Fragrance chemicals in domestic and occupational products. Contact Dermat 45(4):221–225. https://doi.org/10.1034/j.1600-0536.2001.450406.x
2. https://www.google.co.in/books/edition/Common_Fragrance_and_Flavor_Materials/0jFdJAooDL0C?hl=en&gbpv=1&dq=chemical+nature+of+the+flavor+and+fragrance+compound&pg=PP2&printsec=frontcover
3. Hu S, Liu X, Zhang S, Quan D (2023) An overview of taste-masking technologies: approaches, application, and assessment methods. AAPS PharmSciTech 24(2):67. https://doi.org/10.1208/s12249-023-02520-z
4. Babushok VI (2015) Chromatographic retention indices in identification of chemical compounds. TrAC Trends Anal Chem 1(69):98–104. https://doi.org/10.1016/j.trac.2015.04.001
5. Fortune Business Insights. Flavors and fragrances market size, share report (2021–2028) (2021). https: //www.fortunebusinessinsights.com/flavors-and-fragrances-market-102329 Accessed 13 Jun 2024
6. Sell CS (2014) Chemistry and the sense of smell. John Wiley & Sons. https://books.google.co.in/books?id=Mpc6AwAAQBAJ.
7. Zhang L, Mao H, Liu L, Du J, Gani R (2018) A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. Comput Chem Eng 115:295–308. https://doi.org/10.1016/j.compchemeng.2018.04.018
8. Rojas Villa CX, Duchowicz PR, Tripaldi P, Pis Diez R. Quantitative structure-property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. https://ri.conicet.gov.ar/handle/11336/48829.
9. Ahmad Dar A, Sangwan PL, Kumar A (2020) Chromatography: an important tool for drug discovery. J Sep Sci 43(1):105–119. https://doi.org/10.1002/jssc.201900656
10. Rojas C, Duchowicz PR, Tripaldi P, Diez RP (2015) QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemom Intell Lab Syst 140:126–132. https://doi.org/10.1016/j.chemolab.2014.09.020
11. Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, Szalai B, Mainland JD, Ihara Y, Yu CW, Wolfinger R, Vens C (2017) Predicting human olfactory perception from chemical features of odor molecules. Science 355(6327):820–826. https://doi.org/10.1126/science.aal2014
12. Du H, Wang J, Hu Z, Yao X (2008) Quantitative structure-retention relationship study of the constituents of saffron aroma in SPME-GC–MS based on the projection pursuit regression method. Talanta 77(1):360–365. https://doi.org/10.1016/j.talanta.2008.06.038
13. Sharma A, Kumar R, Semwal R, Aier I, Tyagi P, Varadwaj PK (2020) DeepOlf: deep neural network-based architecture for predicting odorants and their interacting olfactory receptors. IEEE/ACM transactions on computational biology and bioinformatics. 19(1):418–28. https://ieeexplore.ieee.org/abstract/document/9115844.
14. Rojas Villa CX, Duchowicz PR, Tripaldi P, Pis Diez R. Quantitative structure-property relationships for predicting the retention indices of fragrances on stationary phases of different polarity. https://ri.conicet.gov.ar/handle/11336/63796.
15. Kumar A, Kumar P, Singh D (2022) QSRR modelling for the investigation of gas chromatography retention indices of flavour and fragrance compounds on Carbowax 20 M glass capillary column with the index of ideality of correlation and the consensus modelling. Chemom Intell Lab Syst 224:104552. https://doi.org/10.1016/j.chemolab.2022.104552
16. Noorizadeh H, Farmany A, Noorizadeh M (2011) Quantitative structure-retention relationships analysis of retention index of essential oils. Quim Nova 34:242–249. https://doi.org/10.1590/S0100-40422011000200014
17. Pourbasheer E, Beheshti A, Vahdani S, Nekoei M, Danandeh M, Abbasghorbani M, Ganjali MR (2015) Simple QSPR modeling for prediction of the GC retention indices of essential oil compounds. J Essent Oil Bear Plants 18(6):1298–1309. https://doi.org/10.1080/0972060X.2014.884768
18. Liu Q, Luo D, Wen T, GholamHosseini H, Li J (2021) In silico prediction of fragrance retention grades for monomer flavors using QSPR models. Chemom Intell Lab Syst 15(218):104424. https://doi.org/10.1016/j.chemolab.2021.104424
19. Ahmadi S, Lotfi S, Hamzehali H, Kumar P (2024) A simple and reliable QSPR model for prediction of chromatography retention indices of volatile organic compounds in peppers. RSC Adv 14(5):3186–3201
20. Riahi S, Ganjali MR, Pourbasheer E et al (2008) QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. Chroma 67:917–922. https://doi.org/10.1365/s10337-008-0608-4
21. Kumar P, Kumar A, Lal S, Singh D, Lotfi S, Ahmadi S (2022) CORAL: quantitative structure retention relationship (QSRR) of flavors and fragrances compounds studied on the stationary phase methyl silicone OV-101 column in gas chromatography using correlation intensity index and consensus modelling. J Mol Struct 5(1265):133437
22. Maulana A, Noviandy TR, Idroes R, Sasmita NR, Suhendra R, Irvanizam I (2020) Prediction of kovats retention indices for fragrance and flavor using artificial neural network. IEEE, New York, pp 1–5
23. Matyushin DD, Buryak AK (2020) Gas chromatographic retention index prediction using multimodal machine learning. IEEE Access 8:223140–223155. https://doi.org/10.1109/ACCESS.2020.3045047
24. Wang YT, Yang ZX, Piao ZH, Xu XJ, Yu JH, Zhang YH (2021) Prediction of flavor and retention index for compounds in beer depending on molecular structure using a machine learning method. RSC Adv 11(58):36942–36950. https://doi.org/10.1039/D1RA06551C
25. Agustia M et al (2022) Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances. IEEE, New York, pp 13–18
26. Matyushin DD, Sholokhova AY, Buryak AK (2019) A deep convolutional neural network for the estimation of gas chromatographic retention indices. J Chromatogr A 6(1607):460395. https://doi.org/10.1016/j.chroma.2019.460395

27. Bi K, Zhang D, Qiu T, Huang Y (2019) GC-MS fingerprints profiling using machine learning models for food flavor prediction. Processes 8(1):23. https://doi.org/10.3390/pr8010023
28. Vrzal T, Malečková M, Olšovská J (2021) DeepReI: deep learning-based gas chromatographic retention index predictor. Anal Chim Acta 22(1147):64–71. https://doi.org/10.1016/j.aca.2020.12.043
29. Matyushin DD, Sholokhova AY, Buryak AK (2021) Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases. Int J Mol Sci 22(17):9194. https://doi.org/10.3390/ijms22179194
30. Vigneau E, Courcoux P, Symoneaux R, Guérin L, Villière A (2018) Random forests: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. Food Qual Prefer 1(68):135–145. https://doi.org/10.1016/j.foodqual.2018.02.008
31. Roy K, Narayan Das R (2014) A review on principles, theory and practices of 2D-QSAR. Curr Drug Metab 15(4):346–379
32. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley
33. Kumar A, Ojha PK, Roy K (2024) The first report on the assessment of maximum acceptable daily intake (MADI) of pesticides for humans using intelligent consensus predictions. Environ Sci Process Impacts. https://doi.org/10.1039/D4EM00059E
34. Kumar A, Ojha PK, Roy K (2023) QSAR modeling of chronic rat toxicity of diverse organic chemicals. Comput Toxicol 26:100270. https://doi.org/10.1016/j.comtox.2023.100270
35. De P, Bhattacharyya D, Roy K (2020) Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling. Struct Chem 31:1043–1055. https://doi.org/10.1007/s11224-019-01481-z
36 Roy K, Das RN, Ambure P, Aher RB (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
37. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press, Cambridge
38. Kumar A, Kumar V, Ojha PK, Roy K (2024) Chronic aquatic toxicity assessment of diverse chemicals on Daphnia magna using QSAR and chemical read-across. Regul Toxicol Pharmacol 1(148):105572. https://doi.org/10.1016/j.yrtph.2024.105572
39. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58(2):109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
40. SIMCA-P U.M.E.T.R.I.C.S. (2002) 10.0, info@ umetrics. com: www. umetrics. com, Umea.
41. Kumar A, Podder T, Kumar V, Ojha PK (2023) Risk assessment of aromatic organic chemicals to T. pyriformis in environmental protection using regression-based QSTR and read-across algorithm. Process Saf Environ Prot 170:842–854. https://doi.org/10.1016/j.psep.2022.12.067
42 Khan K, Jillella GK, Gajewicz-Skretna A (2024) Integrated modeling of organic chemicals in tadpole ecotoxicological assessment: exploring Qstr, Q-Rasar, and intelligent consensus prediction techniques. Q-Rasar Intell Consens Predict Tech. https://doi.org/10.2139/ssrn.4724872
43. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? J Chemom 32(4):e2992. https://doi.org/10.1002/cem.2992
44. Rücker C, Rücker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47(6):2345–2357. https://doi.org/10.1021/ci700157b
45. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 145:22–29. https://doi.org/10.1016/j.chemolab.2015.04.013
46. Zapadka M, Kaczmarek M, Kupcewicz B, Dekowski P, Walkowiak A, Kokotkiewicz A, Łuczkiewicz M, Buciński A (2019) An application of QSRR approach and multiple linear regression method for lipophilicity assessment of flavonoids. J Pharm Biomed Anal 164:681–689. https://doi.org/10.1016/j.jpba.2018.11.024
47. Ciura K, Belka M, Kawczak P, Bączek T, Nowakowska J (2018) The comparative study of micellar TLC and RP-TLC as potential tools for lipophilicity assessment based on QSRR approach. J Pharm Biomed Anal 149:70–79. https://doi.org/10.1016/j.jpba.2017.10.034
48. Kumar A, Ojha PK, Roy K (2024) First report on pesticide subchronic and chronic toxicities against dogs using QSAR and chemical read-across. SAR QSAR Environ Res 35(3):241–263. https://doi.org/10.1080/1062936X.2024.2320143
49. Hall LM, Hill DW, Bugden K, Cawley S, Hall LH, Chen MH, Grant DF (2018) Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds. J Chem Inf Model 58(3):591–604. https://doi.org/10.1021/acs.jcim.7b00496
50. Braibanti A, Fisicaro E, Compari C (2000) Hydrophobic effect: solubility of non-polar substances in water, protein denaturation, and micelle formation. J Therm Anal Calorim 61(2):461–481. https://doi.org/10.1023/a:1010169417937
51. Xing B, McGill WB, Dudas MJ (1994) Sorption of α-naphthol onto organic sorbents varying in polarity and aromaticity. Chemosphere 28(1):145–153. https://doi.org/10.1016/0045-6535(94)90208-9
52. Mandal S, Mandal S, Ghosh SK, Sar P, Ghosh A, Saha R, Saha B (2016) A review on the advancement of ether synthesis from organic solvent to water. RSC Adv 6(73):69605–69614. https://doi.org/10.1039/C6RA12914E

Authors and Affiliations

Doelima Bera[1] · Ankur Kumar[2] · Joyita Roy[1] · Kunal Roy[1]

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

[1] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

[2] Drug Discovery and Development Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India