Understanding the accumulation of organic air pollutants on plant by developing a new partitioning model

BY

Mohammad Yaqoob Sharafat

00000359695

Supervisor

Dr. Deedar Nabi

Institute of Environmental Sciences and Engineering

School of Civil and Environmental Engineering

National University of Sciences and Technology

Islamabad, Pakistan

(2022)

THESIS ACCEPTANCE CERTIFICATE

It is certified that the copy of MS/MPhil thesis entitled by Mr. Mohammad Yaqoob Registration No. 00000359695 of **IESE (SCEE)** has been vetted by undersigned, found complete in all aspects as_per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfilment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members have also been incorporated in the said thesis.

Signature with stamp: ___________________

Name of the Supervisor: Dr. Deedar Nabi

Date: ________________

Signature of HoD with stamp: ________________

Date: ________________

Countersign by

Signature(Dean/Principal):_______

Date:_______

CERTIFICATE

It is certified that the contents and form of the thesis entitled

‘Understanding the accumulation of organic air pollutants on plant by developing a new partitioning model’

Submitted by:

Mohammad Yaqoob

have been found satisfactory for the requirement of the degree

Supervisor: ___________________

Dr. Deedar Nabi

Associate Professor

IESE, SCEE, NUST

Member:

____

Dr. Muhammad Ali Inam

Assistant Professor

IESE, SCEE, NUST

Member:

____

Dr. Musharib Khan

Assistant Professor

IESE, SCEE, NUST

DECLERATION

I certify that research work titled **'Understanding the accumulation of organic air pollutants on plant by developing a new partitioning model'** is my own work. The work has not been presented elsewhere for assessment. Where material has been used from other sources, it has been properly acknowledged and referred.

Mohammad Yaqoob

00000359695

Plagiarism Certificate

This thesis has been checked for plagiarism. Turnitin endorsed by supervisor is attached.

Signature of student:…… Mohammad Yaqoob

Signature of Supervisor: ……Dr.Deedar Nabi

DEDICATION

‘This Thesis is dedicated to my parents whose continuous support and prayers are always with me whenever and wherever required’

ACKNOWLEDGEMENT

First and foremost, all thanks to Allah Subhanahu wa ta'ala for granting me the ability, mindset, and perseverance to successfully complete my study.

I would like to acknowledge my supervisor, ***Dr Deedar Nabi****, whose expertise, understanding, and patience added considerably to my experience. I am grateful of his vast knowledge and skill set. The assistance he provided in writing this thesis is greatly appreciated.*

I would also like to express my gratitude to my GEC members, ***Dr. Muhammad Ali Inam****, and* ***Dr.Musharib Khan****, their guidance and expertise generously helped me during my research.*

I would like to thank my research fellows and friends, particularly Dr.Rahmatullah Wadatyar, Ahmad Aakash, Lybah Sajid, Sheerin Aftab and Muhammad Arslan for tirelessly helping with my research work.

Finally, I would like to take this opportunity to express profound gratitude from for my beloved parents and siblings, for their continuous support and undying love.

At last, I would thank ***HEC*** *of Pakistan for allowing me this opportunity to study in one of the best university of Pakistan.*

(Mohammad Yaqoob)

Table of content

List of Abbreviations

Kcw	Cuticle to water partition coefficient
ASMs	Abraham Solvation Models
Kow	Octano - water partition coefficient
Kaw	Air - water partition coefficient
PM	Partitioning model
EPI	Estimation Program Interfce
EPA	Environmental Protection Agency
RMSE	Route Mean Square Error
SP- LFER	Single Parameter Linear Free Energy Relationship
PP-LFER	Poly Parameter Linear Free Energy Relationship
QSAR	Quantitative Structure-Activity Relationship
2P-PM	Two-Parameter Partitioning Model
ASDs	Abraham Solvation Descriptors
KCMw	Cuticle Membrane-water partition coefficent
KMXw	Matrix membrane -water partition coefficient
OP	One-Parameter
PCA	Principal Component Analysis
MLR	Multi-Linear Regression

LIST OF TABLES

List of Figures

ABSTRACT

It is critical to estimate experimentally plant cuticle/water partition coefficient (Kcw), and understand the mechanisms of partition and environmental fate of organic pollutants because of the high cost of experiments. Only a hundred compounds have had their experimental Kcw values determined. As a result, computer models that predicted Kcw values based on chemical structures might be helpful when assessing new compounds.. A large dataset was used in this investigation of 279Kcw values for 117 diverse chemicals was collected from 24 different plant species. Based on this dataset, developed the Abraham solvation models (ASMs) to select the best prediction model for Kcw estimation among them. which was created and offers a high level of predictability. We reduced the complexity of ASMs by creating and analyzing a two-parameter partitioning model for estimating Kcw of neutral organic compounds. We present two-parameter partition models by (octanol/water partition coefficient) Kow and (air/water partition coefficients) Kaw. The results from partitioning model(PM) by the determination coefficient R^2=0.95 — 0.97, external validation coefficient R^2=0.93, and RMSE=0.57 — 0.63 log unit, the PM model has strong predictability and robustness. In conclusion, The proposed PM model can also be used to calculate the logKcw of prospective organic contaminants directly from their chemical structures.. Our PMs are simple to include in the popular EPI-SuiteTM screening tool. Parameter-intensive ASMs performed similarly to PMs in terms of explanatory and predictive power. Our models offer simple, alternative methods for assessing the hazard of complicated mixes of organic micropollutants

Keywords

partition coefficients, EPI-Suite™, RMSE, PM

Chapter 1

INTRODUCTION

1.1 Background

In the circulation of organic pollutants plants play a significant role and their accumulation on plants attracts attention in the field of environmental research. The cuticle of a plant serves as a barrier between the plant and its surroundings.(Qi et al., 2020a) The cuticle is the first layer that reduces transpiration(Qi et al., 2020a), abiotic and biotic stressors, dehydration, UV irradiation, insect attack, pest, pathogen(Fernández et al., 2017b; Yeats & Rose, 2013a), desiccation, covers the overhead pod of all land plant(Yeats & Rose, 2013a), controls the chemicals or gas exchange between plants and the environment, and defends plants against environmental threats(Eddula et al., 2021a; Yeats & Rose, 2013a). Cuticles cover the majority of the plant's upper components, including fruits, leaves, flowers, and less woody stems. To understand and improve the processes of cuticle and wax formation, there has been researched done for fifty years. Furthermore, several reports assess in several plants and organs the composition, formation, structure, and function of cuticles(Fernández et al., 2017b). However, Cuticles from Clivia minata, Fiscus Elastica (rubber), Agave Americana, Citrus aurantium (bitter orange), Populus canescent, and Arabidopsis thaliana leaves, as well as the fruits of Lycopersicon esculentum Mill (tomato) and Capsicum annuum L (pepper) were used in the research.)(Eddula et al., 2021b; Fernández et al., 2017b).Up-taking and accumulation of pollutants by plants cause of enhancement of these pollutants to high trophic level via the food chain(Qi et al., 2020a). Plants are exposed to insecticides and herbicides used by farmers to control pest infestations and eliminate unwanted vegetation. we should be knowing the accumulation of pollutants in plants leaves, fruits, and grasses which are bested as a food source for animals and humans(Eddula et al., 2021b). Organic pollutants accumulated in plants by a variety of mechanisms, including absorption, translocation, transpiration, metabolic degradation, sorption, and desorption between stream and biomass(Zou et al., 2011). Therefore, from the perspective of environmental safety, it is necessary to assess the interaction mechanisms and uptake of contaminant infiltrating plants. For the assessment of uptaking chemicals by plants we use the partition coefficient between plants and water(Qi et al., 2020a) The term cuticle-water partition coefficient is mathematically presented by Kcw(Eddula et al., 2021b) For the accumulation of organic pollutants on a plant, plant cuticles are considered as a distinguishable ingredient (Kpw)(Qi et al., 2020a). From the environment, the accumulation of hydrophobic organic pollutants into plant cuticles is considered a significant component(Eddula et al., 2021b). Experimenting to determine the coefficient of partition between plant cuticles -water (Kcw) is costly, time consuming, and labor intensive, therefore, it is infeasible to test (Kcw)for all pollutants. As far as we know, hundreds of chemicals have had their experimental (Kcw) values measured.. For quickly measuring solute properties and partition coefficient of organic compounds there is an urgent need for mathematical equations(Eddula et al., 2021c; Qi et al., 2020b) previously, the single parameter linear free energy relationship (SP-LFER)model was developed for the prediction of Kcw by octanol/water partition coefficient. but it was not perfect for highly polar and hydrophobic compounds. LogKow was not enough for the prediction of LogKcw.(Qi et al., 2020) In addition, for understanding the Kcw the SP-LFER cannot describe all molecular interactions to present all functions of solute(Eddula et al., 2021b; Qi et al., 2020a). In order to forecast the partition coefficient, of the diverse systems and chemicals, Abraham Poly Parameter Linear Free Energy Relationship(PP-LFER) model is an effective predictor(Eddula et al., 2021b; Khawar & Nabi, 2021; Poole et al., 2013; Qi et al., 2020). Also, some other prediction models like ABSOLV, Quantitative Structure-Activity Relationship(QSAR), and Quantum Chemical Method are used to estimate organic compound partition coefficients on plant cuticles. However, because of the small number of modeling sets, the current models may not be enough for predicting logKcw of diverse chemical compounds. For non-polar and slightly polar chemicals, these datasets are preferred, but not for very polar ones. Furthermore, Individual interactions were not previously quantified in prior models(Qi et al., 2020b). The goal of this research is to create a two-parameter partitioning model (2P-PM) for predicting LogKcw based on a huge dataset containing 279LogKcw values for 117 unique compounds tested by 24 plant species a mixture of polar and non-polar substances.

1.2 Structure and Function of Plant Cuticle

1.2.1Structure of Cuticle

Almost 540 million years ago in the middle period of Paleozoic faced to bandle challenges related to their novel terrestrial environment, like excess temperature, dehydration, high exposure of UV radiation and gravity.The plant cuticle is a composite of wax(organic chemicals soluble lipids)and cutin. However, polysaccharides and cuticles separate parts in a plant but they have some association according to physical and overlap functions. According to microscopic analysis, the cuticle is divided into two main parts: cuticle rich part(cuticular layer) which is fix with polysaccharides, and waxes rich part which is less rich of polysaccharides. The composition of wax is vary entirely in each plant species, ontogeny , and growth condition of pant. In the majority of cases, the major part of chemicals containing the cuticular wax is obtained from a very long chain of fatty acids like alkane series, acetaldehyde, primary and secondary alcohols, phenols, and esters(Yeats & Rose, 2013b). Although the contact between interior structure, function and chemical composition of plant cuticle still unknown. palnt cuticle act as a barrier in front of biotic and abiotic stresser from environment. In addition , in several report assess the plant cuticle composition, function, structure and formation in defferent plant species(Fernández et al., (Fernández et al., 2017a). since 19th-century different approaches used for the studying plant cuticles in various plant to assess their composition and structure. as a result, the researcher introduced different terminology of cuticle structure. Basically plant cuticle consist of three layers:layer of cuticle, proper cuticle, and epicuticular wax. interior layer is a layer of cuticle comprise of cellulos as well as cutin and polysaccharides which is covering the primary wall. The proper cuticle is a covering layer of cuticle that is made of intracuticular wax and cutin and it is free of polysaccharides according to the ancient explanation. Due to this information now cuticle is considered an essential component of the exterior periclinal walls rather than a separate area and cuticle considered a diverse heterogeneous exterior era of the cell wall(Skrzydeł et al., 2021).

1.2.2 Function of plant cuticle

As we know plant cuticle is a barrier between terrestrial environments and the overhead part of the plant, also it serves many duties and function(Skrzydeł et al., 2021). The cuticle is the first layer that reduces transpiration(Qi et al., 2020a), abiotic and biotic stressors, dehydration, UV irradiation, insect attack, pest, pathogen(Fernández et al., 2017b; Yeats & Rose, 2013a), desiccation, covers the overhead pod of all land plant(Yeats & Rose, 2013a), controls the chemicals or gas exchange between plants and the environment, and defends plants against environmental threats(Eddula et al., 2021a; Yeats & Rose, 2013a). Cuticles cover the majority of the plant's upper components, including fruits, leaves, flowers, and less woody stems. Also between cell development and relationship, the cuticle has an important and vital role and at the starting stage cuticle provides a merge between organ of plant. According to water permeability in cuticle relationship between the composition, structure, and function of the cuticle is complex. However, there is no relationship between the thickness of the cuticle(quantity of wax) and permeability of water, which is proven by the different researchers in different plant species. while in the tomato fruit the cuticular transpiration is affected by cuticle chemical composition(Skrzydeł et al., 2021). Occasionally, the opposite function perform by some plan organ, like the attachment of the pollinator insect to the flower surfaces(Bräuer et al., 2017)and contemporary sack-sucking insects forbidding the attachment of surface flowers.In some cases, cuticles performed a reverse function in the same organ of plant in diverse parts like in Maize plant leaves cuticle help in hardness of surface and prohibited water losses of the pod but nationally, cuticles help to increase bulliform cell or motor cell in both (surface hardness and water losses)(Matschi et al., 2020). In pine plant cuticles on the bases of needles help in water up taking, while needles are covered with cuticles which give them a hydrophobic situation and which cause the motion of water droplets through the needles. Also, cuticles do opposite functions in the diverse organs of similar plants. for instance, young growing organs emerging and simulated by plant cuticles. while plant cuticles help in pistil and pollen grain interaction(Skrzydeł et al., 2021).Here in the below table show us the function of cuticle in different flower organs within leaves and fruits.

Table-1 In diverse plants organs, the function of cuticles

Organs/Members	Functions/Tasks	Species/Kinds of Plant
Flowers/Blossom organs	conservation of merging/fusion of organs	Arabidopsis thaliana, Solanum lycopersicon
	Prevention of insect	Aristolochia fimbriata
Perigonium/ Perianth	Organ coherence controlling	A. thaliana
	Eliminating of floral odor	Clarkia breweri, Antirrhinum majus
Floral Leaf/Petal	Cleaning it self, Conservation from UV radiation	Mutisia decurrens
	Volatiles are released to draw insects and improve pollination.	Antirrhinum majus
	Assistance with pollinator affilation	
	Pollinators' color and attraction to visual effects	
Pistillode/Pital	Pistan intraction and pollen control	Brassica, Arabidopsis thaliana, Oryza sativa
Apocarpous Gynoecium/Carpel	conservation of merging/fusion of organs	Catharanthus roseus
Fruitage/Fruit	non-stomatal transpiration regulation and stomatal transpiration control	Prunus avium cv. Sam
	Conservation against dehydration	Malusdomestica (Jonagold , Jonagored, Elstar), Mangifera Indica cv. Cogshall , Capsicum, Solanum Lycopersicon, Malus domestica,
	Conservation from UV radiation	Cydonia oblonga

Organs/Members	Functions/Tasks	Species/Kinds of Plant
Fruitage/Fruit	Conservation against pathogen	Solanum lycopersicon
	Conservation against insect	Prunus domestica
Leaf	Conservation against dehydration(prevention to perspiration)	Arabidopsis thaliana, Zea mays, Olea europaea , Triticum aestivum, Citrus sinensis, Prunus laurocerasus , Vanilla planifolia, Ruellia
	Conservation against high temperature	Salvinia natans
	provision of extremely lipophilic surface	Salvinia natans, Aponogeton madagascariensis
	Conservation against insects	Arabidopsis thaliana, Chenopodium album, Nepenthes albomarginata, Prunus avium
	Conservation against pathogens	Cucumis sativus, Phaseolus vulgaris, Arabidopsis thaliana , Ilex aquifolium, Prunus avium
	Conservation from UV radiation	Triticum aestivum cv. Shango
	Cleaning itself	Mutisia decurrens, Nelumbo nucifera, Colocasia esculenta
	conservation of merging/fusion of organs	Arabidopsis thaliana
	Effect of triboelectric charging	Rhododendron

1.3 Partition Coefficient(P)

Is the ratio of compound concentration in a mixture of two immiscible phases when they are at equilibrium. Normally one phase is taken as hydrophobic such as octanol, lipid etc and the other phase is taken as hydrophilic as water. It is therefore a measure of differences in solubility of compound present in two phases.

Partitioning property of a chemical has a great influence on its environmental fate, its distribution and bioavailability. Fate models for predicting environmental behaviors and ecological impact assessment of chemicals typically involve partitioning properties, defined as follows.

$$Pxy,i = \left\{\frac{Cx,i}{Cy,i}\right\} Pxy,i = \left\{\frac{Cx,i}{Cy,i}\right\} \text{ equilibrium..............(1)}$$

Where P_{xy},i is partition coefficient between two phases x and y, and $C_{x,i}$ and $C_{y,i}$ are the concentrations of contaminant i at partitioning equilibrium present in these phases. Thus, to evaluate the chemical exposure and transport in the environment, Equilibrium partition coefficients are required (Otsuka, 2006). Nonionic organic substances are partitioned between water and natural organic phases, these phases are the measure of the hydrophobicity of compound in which is targeted compound. It is expressed as (octanol-water partition) K_{ow} coefficient or constant. Likewise, air-water partition coefficient (K_{aw}) is a function of compound volatility which is the partitioning of the compound in between air and liquid. In one-parameter Linear Free Energy Relationships (OP-LFERs), partitioning constant parameters are conducted to find out the unknown equilibrium partition coefficient between two phases. As one-parameter LFERs use only one such parameter so it has limited predictive power because as we know that a single parameter does not or has no ability to complete the molecular interactions that influence a compound's equilibrium partitioning between two phases. So, for different regression coefficients for different compound classes is required in LFERs(Goss & Schwarzenbach, 2003).

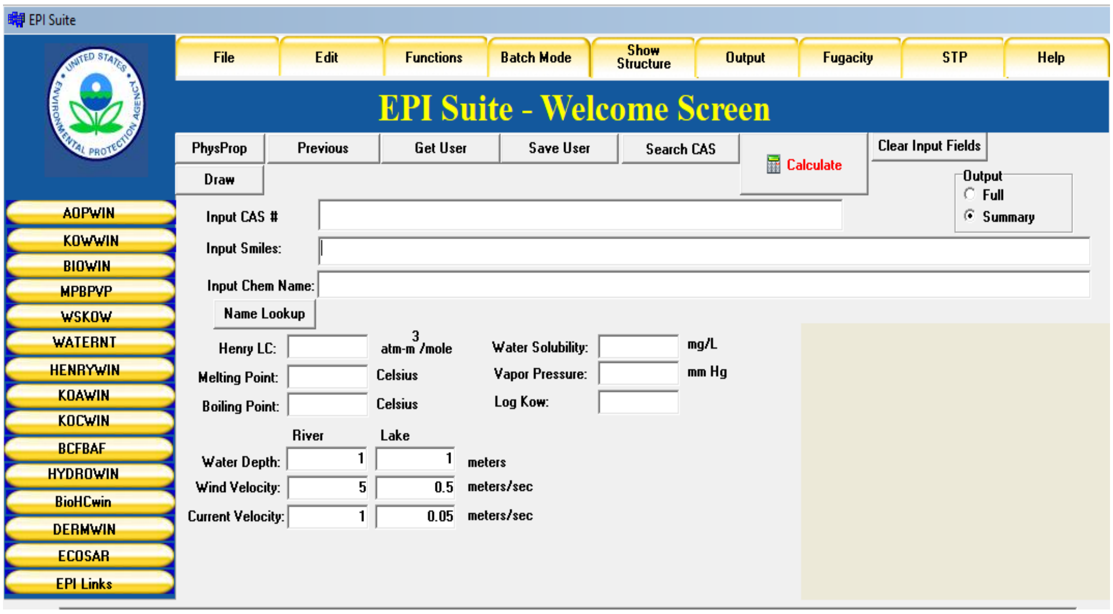
Therefore, for the handling of the variability of both compound and sorbent, there is a need of a much-refined approach than those of op-LFERs for quantification and reliable prediction of equilibrium partitioning. Using an equation known as poly-parameter linear free energy relationships (pp-LFERs), a highly effective tool for explaining partitioning data of huge and heterogeneous data sets of compounds exists. (M. H. Abraham et al., 1999).

Abraham solvation models are excellent models for the forcoasting of environmental, biochemical, and physicochemical properties. And it is based on theoretical footings of linear free energy relationships. It is based on the intermolecular interactions that are important to describe the interaction of a contaminant with environment and these intermolecular interactions they are encoded in Abraham Solvation Descriptors (ASDs) that are depicted over here such S,E that show polarity and polarizability, A and B hydrogen bonding interactions parameter, V and L are used to incorporate the energy for the cavity formation. ASDs are intermolecular interactions that are offered by chemicals when they are released in the environment. e,s,a,b,v,l, are corresponding intermolecular interaction parameters that are offered by different environmental phases (M. H. Abraham et al., 1999).

1.4 EPI Suite ^{TM} (Etimation Program Interface)

As we know in the United States and European Union almost 100000 chemicals are registered and considered that annually 1000 novel chemicals are prefabricated in the United State. For the risk assessment of chemicals which pose to humans, animals, and Environmental health many agencies and governments have to inflict. risk assessment of these chemical substances is necessary because of their physicochemical properties, exist in environmental phases, bioaccumulation, degradation, vulnerability, hazards to humans, and ecological. United State and the European Union have different experiments for new chemicals which expose to the environment like physicochemical properties, biodegradation, bioaccumulation, animal toxicities, ecotoxicity, and the Toxic Substance Control Act. Toxic Substances Control Act was developed in 1976 and modification of its done on 2016. US EPA used a predictive computational model (EPI Suite^{TM})for risk assessment of novel chemicals in the environment to find the properties and fate of chemicals in the environment. In order to calculate precise parameters models based on quantitative/structural activity relationships were created and utilized by the US EPA. In the starting, EPI Suite was by the name of EPIWIN when in 2000 by the purchase of EPI from SRC the US EPA change its name to EPI Suite. EPI Suite was available freely after purchasing through US EPI and its uses to assess and the fate of chemicals in the environment. The below figure show us EPI Suite interface(Card et al., 2017).

Figure- 1.1 EPI Suite software interface



Output: ○ Full ● Summary

Input CAS #:

Input Smiles:

Input Chem Name:

Henry LC: atm-m^3/mole

Melting Point: Celsius

Boiling Point: Celsius

Water Solubility: mg/L

Vapor Pressure: mm Hg

Log Kow:

	River	Lake	
Water Depth:	1	1	meters
Wind Velocity:	5	0.5	meters/sec
Current Velocity:	1	0.05	meters/sec

In EPI Suite software we have below models everyone use for different purposes.

Table-2 Models in EPI Suite

No	Models	Calculates/Estimate
1	AOPWIN	Potential for oxidation in the atmosphere
2	KOWWIN	Octanol/Water partitioning coefficient
3	BIOWIN	Biodegradability of compounds
4	MPBPVP	Chemical's melting point, boiling point, and vapour pressure.
5	WSKOW	Based on log K_{ow}, calculates water solubility.
6	WASTERNT	Solubility of compound
7	HENRYWIN	Partitioning of air and water.
8	KOAWIN	Octanol/water partitioning
9	KOCWIN	Organic carbon partitioning coefficient
10	BCFBAF	Bionconcentration and bioaccumulation
11	HYDROWIN	Rate of hydrolysis
12	ECOSAR	Toxicity
13	BioHCwin	Hydrocarbon biodegradation
14	DERMWIN	The amount to which organic substances are absorbed via the skin

EPI suite software is basically an excellent tool used for risk assessment of new chemical manufacturers but it has some limitations and restrictions.it has no accurate prediction and estimation for(inorganic chemicals, organometallic chemicals, some ionizable organic chemicals, chemicals with high molecular weight, perfluorinated and other highly halogenated substances, nanoparticles, compounds in the training set that don't have any functional groupings). Additionally, EPI cannot predict and forecast some environmental destiny and physicochemical properties of chemicals which help us in risk assessment of chemicals like .in organisms and specese bio-concentration, bio-accumulation, bio-transformation rate, and the end products. In 2007 this program was reviewed by US EPA in updates many models in EPI Suite which are limited in this software and also added useful tools of EPI Links. Further development will add in upcoming versions.

1.5 Problem Statement

The experimental values of Kcw are not readily available for many organic chemicals and the experimentation is expensive, laborious, and sophisticated. The existing models used to predict the K_{cw} values (i) are parameter intensive (ii) are computationally expensive (iii) have difficult chemical interpretation (iv) do not able to give a mechanistic insight (v) do not have enough available database of predictive parameters and have limited estimation power extended to few chemicals.

1.6 Objectives

Based on the problem statement, the research was created with the following goals in mind.:

1. To investigate the partitioning behavior of organic air pollutants for their accumulation on plants using salvation-based models (Recalibration of ASMs).
2. To develop a LFER model for the prediction of plant cuticle/water partitioning coefficients for organic air pollutant (two-parameter models).
3. To extract the mechanistic understanding the effects of K_{ow} and K_{aw} of air organic pollutants for their tendencies to accumulate on plant surfaces.

1.7 Study's scope

The study work was divided into three phases.

1. In the first phase, Abraham Solvation Models(ASMs) have been developed to Cuticle to water partitioning for neutral organic chemicals.
2. The 2P-partitioning model has been developed to Cuticle to water partitioning for neutral organic chemicals.

In the third phase, the model has been tested for certain criteria of internal and external validity to check its robustness.

Chapter 2

LITERATURE REVIEW

Performed a study that combined the 279Kcw experimental values of plant tissues, cuticle membrane/water partition coefficient (LogKCMw) and matric membrane/water partition coefficient(LogKMXw) for different plant species because there was no difference for many of the chemicals and plant partitioning coefficients. The difference was smaller than 0.20 Log units, suggesting that experimental uncertainty measurement was the reason. For both cuticle/water partition (logKcw) and cuticle/air (logKca) they chose to derive Abraham model correlations. the mathematical equation for LogKcw and LogKca(Eddula et al., 2021c).

$LogKca = LogKcw + LogKw$-- -(2)

Kw=gas-water partition coefficient.

A minimum of 30 to 40 experimental data points are required to develop a significant Abraham model correlation. For the plant species Lycopersicom esculentum Mill, for the plant species Capsicum annuum L, there are more than adequate experimental values,there are just enough experimental values to get a logKcw equation. By assessing the experimental data, the Abraham model correlations for Lycopersicom esculentum Mill were identified.

$LogKcw = -0.19(0.05) + 0.912(0.06)E - 0.6(0.08)S - 0.28(0.08)A - 3.84(0.09)B + 3.45(0.06)V$--(3)

(N = 114, SD = 0.21, SE = 0.22, R^2 = 0.98, Radj2 = 0.98, F = 1371)

For Capsicum annuum L

$LogKcw = -0.14(0.19) - 1.02(0.08)E - 1.24(0.16)S - 0.15(0.14)A - 3.58(0.14)B + 5.50(0.10)V$---(4)

(N = 41, SD = 0.23, SE = 0.25, R2 = 0.98, Radj2 = 0.97, F = 382.1)

To obtain the widest correlation possible, they combined all of the experimental data into a single correlation expression. The following correlation equation emerged from the combined data sets' final regression analyses:

$LogKcw = -0.15(0.04) - 0.99(0.03)E - 0.76(0.06)S - 0.08(0.05)A - 3.60(0.05)B + 3.412(0.038)V$---(5)

(N = 26, SD = 0.23, SE = 0.23, R^2 = 0.98, $Radj^2$ = 0.98, F = 3833)

Also in this study, Abraham did model development for air/cuticle partition coefficient (Kac) for the same data set the result is better than Kcw model and they used L descriptor in place of V. equation is below.

$LogKca = -0.45(0.04) - 0.30(0.04)E - 1.16(0.06)S - 3.31(0.06)A - 1.04(0.07)B + 0.77(0.01)L$--(6)

(N = 215, SD = 0.22, SE = 0.22, R^2 = 0.99, $Radj^2$ = 0.99, F = 8508)

In the performed study the ploy-parameter Linear Free Energy Relationship (pp-LFER) technique was employed according to the equations below.

$LogK = c + eE + sS + aA$+bB+vV---(7)

The logarithmic partition coefficients between a condensed phase (for example, plant cuticles and water) are denoted by LogK. The capital letters are Abraham's descriptors. Lowercase letters represent regression coefficients. The Abraham descriptors were taken from UFZ-LSER online database. In this study, they used SSPSS 22.0 (IBM USA) software for multiple linear regression(MLR). In which Abraham descriptors were independent variables and LogKcw was a dependent variable. The result of single parameter Linear Free Energy Relationship (sp-LFER) between LogKcw and LogKow developed R^2=0.79. But there was no correlation for compound LogKow<-1 between Logkcw and Logkow. As a result, using a simple correlation LogKow to predict logKcw for various chemicals is ineffective. The pp-LFER model was created for forecasting logKcw based on the 279 data points as below(Qi et al., 2020b):

$LogKcw = -0.46(\pm0.1) + 0.95(\pm0.07)E - 0.6(\pm0.13)S + 0.34(\pm0.13)A - 2.74(\pm0.1)B + 3.01(\pm0.08)V$--(8)

n_{tra}=224 $R^2_{ad,tra}$=0.93 $RMSE_{tra}$=0.52 Q^2_{boot}=0.95 $P < 0.05$

n_{ext}=224 $R^2_{ad,ext}$=0.93 $RMSE_{ext}$=0.52 Q^2_{ext}=0.95

Great predictability, resilience, and external forecast accuracy are all factors to consider, the pp-LFER model performed well. Moreover, since the range of logKcw values assessed by different plant species is small, we created a pp-LFER model employing average values for chemicals with multiple logKcw values recorded from different plants. As a consequence, the following pp-LFER model was created using 125 data points:

$$LogKcw = -0.17(\pm 0.17) + 1.16(\pm 0.14)E - 0.55(\pm 0.23)S - 2.33(\pm 0.15)B + 2.47(\pm 0.17)V \text{---} (9)$$

n_{tra}=100 $R^2_{ad,tra}$=0.90 $RMSE_{tra}$=0.66 Q^2_{boot}=0.88 $P < 0.05$

n_{ext}=25 $R^2_{ad,ext}$=0.88 $RMSE_{ext}$=0.86 Q^2_{ext}=0.86

In the performed study Quantitative Structure-Activity Relationship(QSAR) model was developed based on a support vector machine(SVM) for predicting the water/ polymer matrix membrane plant cuticle partition coefficient(Kwmx).The model was conducted on two parameters which drived from chemical structures. The conducted model was developed by using two parameters descriptors under the OECD guidelines. MLOGP (Moriguchi octanol/water partition coefficient)which is represented the hydrophobicity of a molecule and nDP which is represented the double bonds of a molecule base on the different compound and diverse plant species. TSI and K-W statistics were used to examine the variety of chemicals and data, revealing that the chemicals under consideration are different in nature and also plant species are significantly different(Gupta & Mallick, 2018).

For checking the statistical validation of the model used external and internal validation of the model result showed high level of confidence. The conducted QSAR model has a high predictability for novel compounds.The finding indicates that the SVM model is a viable and strong approach for predicting of log Kwmx structurally significant values varied compounds in various species and that it maybe put to use to screen compounds for environmental risk assessment.

The preform study developed a model for both dry deposition (matrix membrane to air partition KMXa)and wet deposition(matrix membrane to water partition coefficient KMXw). Both of these partitioning mechanisms impact the amount of a chemical that is taken up by plants(Platts & Abraham, 2000).

Result of regression analysis on KMXa based on linear free energy relationship (LFER) by using solvation descriptors.

R^2=0.994 SD=0.232 n=62

And the result of KMXw based on LFER by using solvation descriptors.

R^2=0.981 SD=0.236 n=62

We have seen those multiple kinds of techniques such as one parameter, poly parameter and ASM etc are being used for the prediction of K_{cw}. But these methods have some drawbacks in them such as they are laborious, parameter intensive, gave no mechanistic insight and require a lot of experimentation. Furthermore, OP doesn’t cover all the intermolecular interactions offered by the chemicals when released into the environment.

So, there is a need for the development a model which has the same prediction power as ASM, but the model doesn’t have the drawbacks as ASM.

Chapter 3

MATERIALS AND METHODS

3.1 Data Acquisition

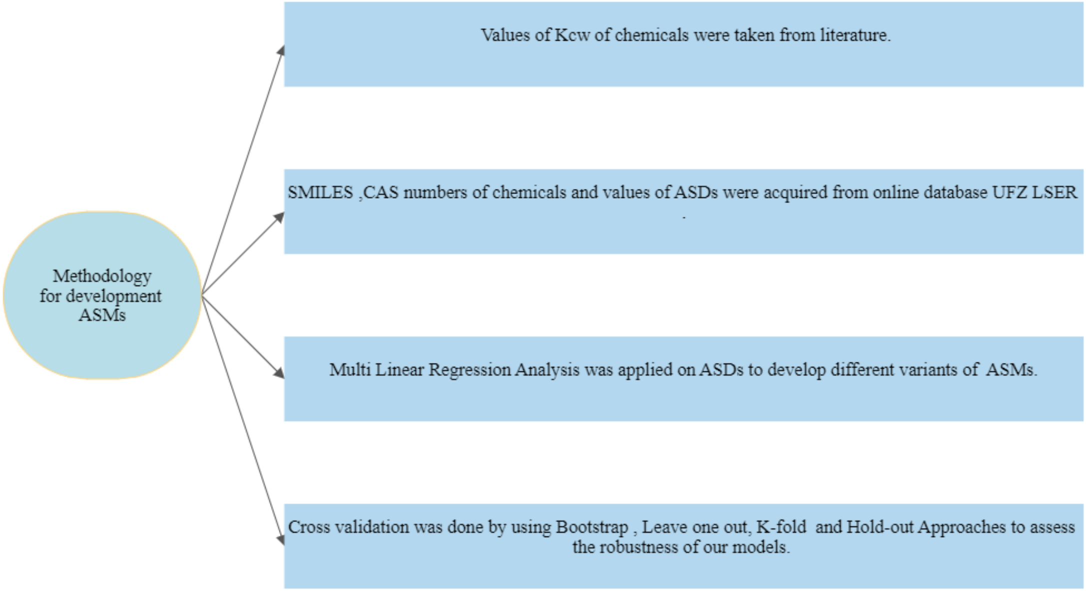

Methodology for development ASMs
Values of Kcw of chemicals were taken from literature.
SMILES ,CAS numbers of chemicals and values of ASDs were acquired from online database UFZ LSER .
Multi Linear Regression Analysis was applied on ASDs to develop different variants of ASMs.
Cross validation was done by using Bootstrap , Leave one out, K-fold and Hold-out Approaches to assess the robustness of our models.

Figure-3.1 Flow chart of the methodology for development solvation based models

Experimental values of cuticle-water partition (Kcw) for 109 chemicals were taken from the literature(Eddula et al., 2021c). The data covered diverse organic compounds like carbohyrates, alcohols, aromatic compounds, pesticides, alkanes, cycloalkanes, haloalkanes, olefins, ketones, esters, nitriles, and munition compounds (Qi et al., 2020b).Figure-1(bor plots)show us the diversity of compounds in our data. The range of values of Abraham Solvation Descriptors (ASDs),octanol/water partition coefficient($K_{o\text{-}w}$), and air/water partition coefficient($K_{a\text{-}w)}$ showing in below table1. The values for Abraham solvation descriptors(ASD), Simplified Molecular-Input Line-Entry System (SMILES) codes, and Chemical Abstracts Service (CAS) numbers of chemicals were taken from the freely available database UFZ - LSER Database. The estimated and experimental values of partition coefficients $\log K_{ow}$ and $\log K_{aw}$ were acquired from the open-source US-EPA software EPI Suite ^{TM} 4.1.(US EPA, 2016). Using the modules KOWWIN v1.68 for values of Kow and HenryWin v3.20 for values of Kaw. The experimental values of $\log K_{ow}$ were available for only 111chemicals out of 117 chemicals. We used the ASM model to calculate the $\log K_{ow}$ values for the remaining six compounds. We used the values of logK$_{ow}$ estimated from ASM in place of unavailable experimental values.

Table 3. Showing the range of Kow, Kaw, and ASDs values

Descriptor	Minimum value	Maximum value
E	-0.1	4.07
S	0	2.76
A	0	1.38
B	0	1.8
V	0.16	3.40
L	-1.741	13.3
logK$_{ow}$	-3.24	7.6
logK$_{aw}$	-14.97	1.97

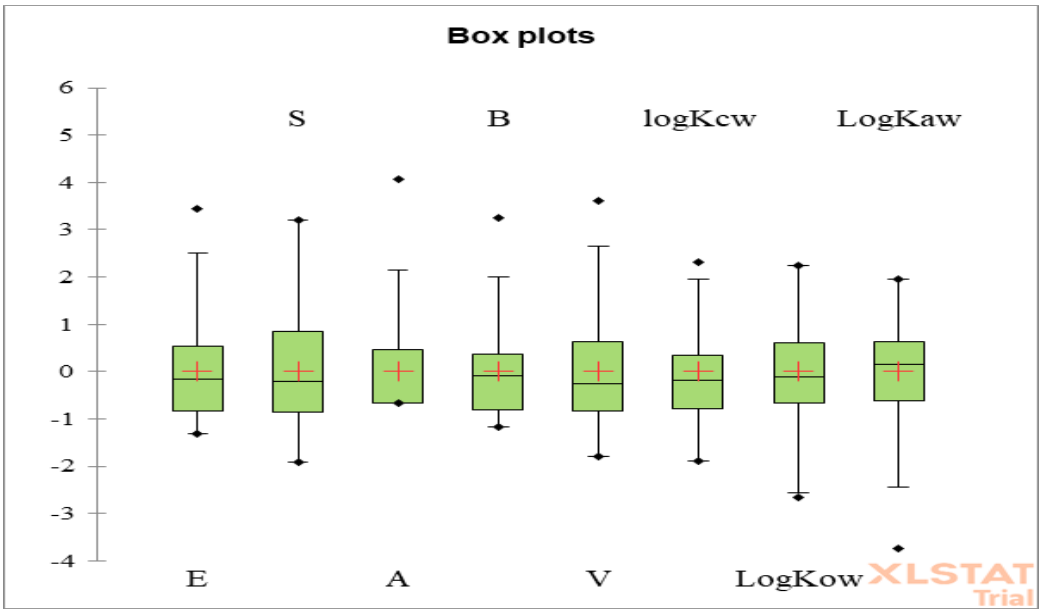


Figure-3.2 diversity of chemicals in the dataset

Box plots indicate to us the level of number and level of scores on a scale. These plots divide data into four equal sizes every one called by the name quartiles (25%data collected in each quartile) and numbering to these four quartiles starting from the bottom to top. Also used for visualization the rage in our data and other specification. **Figure-3.2** is indicating the diversity of chemicals in our dataset in which above the box's dots show the max values and below dots show the min values and also between boxes the red plus mark indicates the median values.

3.2 Test of Significance

Statistical significance tests are crucial in supporting the researcher in achieving this aim. Even when the gains are minor, a vigorous test helps the researcher to discover huge development. Then there's the matter of which statistical significance test Information-retrieval (IR) researchers should apply?. Because commonly these statistical significance tests are used, Student's t, bootstrap, randomization, Wilcoxon, and sign tests. Students' t, bootstrap, and randomization tests all agree to a great extent. Researchers that use one these three tests are most likely to be successful to reach similar conclusions about the statistical significance of their data. The Wilcoxon and sign tests contradict each other and the other tests. So then for no long time be utilized by Information-retrieval (IR) researchers for a variety of reasons that we describe. A test should create the statistic reported by the researcher(Smucker et al., 2007). Consequently, the t-test is only useful for determining the difference in means or comparing two groups' mean scores against an estimate of sample variability to see if they are statistically different(Rojewski et al., 2012; Smucker et al., 2007; Starkweather, 1988). T-tests can be used to calculate independent samples(with different contributors in each group) and dependent samples (with similar contributors in each group). types of T-test, One-sample T-test(If only one group is being compared to a reference value), two-sample T-test(If the two groups are from distinctive cultural categories,) and the paired t-test is a statistical method for comparing two groups of people (If the groups come from a single population)(Starkweather, 1988).s in this study we conducted the two-sample T-test. which we had three types of plant tissue isolated cuticle membrane(CM), polymer matrix membrane(MX), and whole plant biomass. So we want to know about that whether the accumulation of pollutants depends upon the plant tissue types or not we did a T-test. To check the variance of two groups i.e CM and MX, and to check whether we can combine the datasets of both the groups or not T-test was performed by extracting the common chemicals of both groups. T-test was performed using XLSTAT. The results of the t-test showed us that there was an insignificant difference between the mean of the two groups as well as the P >0.05, and we can combine the groups. below table-2 shows us the dataset for the T-Test.

Table -4. Dataset for T-Test

No	Chemicals Name	CM	Mx	No	Chemicals Name	CM	MX
1	4-Nitrophenol	1.97	2.03	21	2,4,5-Trichlorophenoxyacetic acid	3.11	3.2
2	Phenol	1.59	2	22	2,4-Dichlorophenoxyacetic acid	2.5	2.69
3	bis(2-Ethylhexyl) phthalate	7.48	7.66	23	4-Nitrophenol	1.79	1.76
4	Atrazine	2.19	2.2	24	di(2-ethylhexyl) phthalate	7.22	7.38
5	Perylene	6.55	6.58	25	Atrazine	2.15	2.17
6	1-Naphthylacetic acid	2.33	2.43	26	Perylene	6.45	6.59
7	Pentachlorophenol	4.66	4.72	27	1-Naphthaleneacetic acid	2.18	2.25
8	2-Nitrophenol	1.92	2.04	28	Pentachlorophenol	4.42	4.46
9	1-Naphthalenol	2.93	3.01	29	2,4,5-Trichlorophenoxyacetic acid	3.13	3.2
10	Naphthalene	3.37	3.39	30	2,4-Dichlorophenoxyacetic acid	2.47	2.48
11	2,4,5-Trichlorophenoxyacetic acid	3.21	3.26	31	4-Nitrophenol	1.89	1.91
12	2,4-Dichlorophenoxyacetic acid	2.76	2.89	32	Phenol	1.58	1.64
13	Hexachlorobenzene	5.8	5.82	33	di(2-ethylhexyl) phthalate	7.32	7.33
14	4-Nitrophenol	1.8	1.89	34	Atrazine	2.12	2.13
15	Phenol	1.51	1.69	35	Perylene	6.5	6.49
16	di(2-ethylhexyl) phthalate	7.28	7.58	36	Pentachlorophenol	4.57	4.7
17	Atrazine	2.16	2.15	37	2-Nitrophenol	1.83	1.99
18	Perylene	6.2	6.58	38	2,4,5-Trichlorophenoxyacetic acid	3.19	3.24
19	Pentachlorophenol	4.55	4.6	39	Acide 2,4-dichloro phenoxyacetique.	2.63	2.79
20	O-Nitrophenol	1.97	2.03	40			

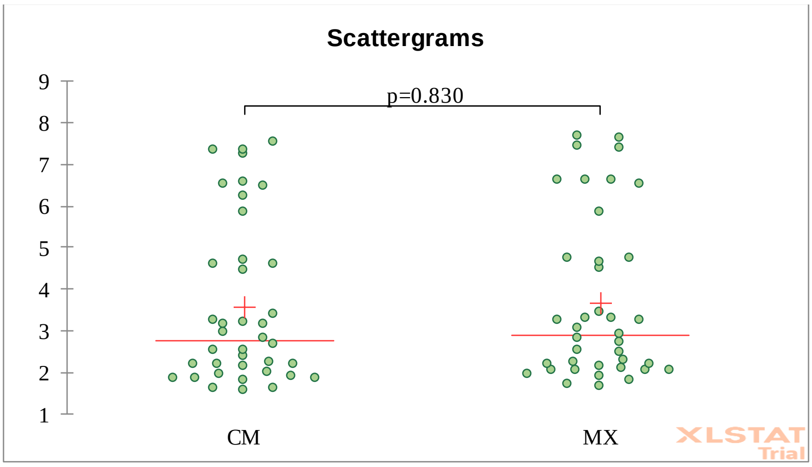


Figure -3.3 T-Test Resul

In this scattergram, dots indicate the number of observations. The plus mark indicates the median value and the red line indicate the mean of these observation in the data set.

Interpretation of the T-Test:

H_0: The difference between the means is equal to 0.

H_a: The difference between the means is different from 0.

The calculated p-value is greater than the significance level alpha=0.05, we should accept the null hypothesis H_0,which is the different between the means is equal to 0.

As the p-value is >0.05 difference between the groups is insignificant so we can combine the dataset. It was also proved from the literature that they combined the dataset for analysis.

3.3 Statistical Analysis

Statistical analyses including principal component analysis (PCA), multiple linear regression (MLR), Pearson correlation, and cross-validation tests were performed using (XLSTAT, 2020) and RStudio (version – 1.4.1106) (Core & Team, 2020). MLR was used to determine the optimum and significant number of descriptors based on statistical analysis. PCA was used to visualize the chemical space and to reduce the data redundancy. PCA was run on all the descriptors to quantify the variance in the data and to find out the contribution of each variable in the principal component. Pearson correlation analysis was done to determine the relationship between each variable. Cross-validation tests such as leave one out, K fold, External validation(Hold out), and the bootstrap method were used to assess the robustness of our models. External validation was done by splitting the data set into a training set and a test in 1:4 ratio.

Chapter4

RESULTS AND DISCUSSIONS

4.1 Development and validation of Abraham Solvation Models

After a lot of work, Abraham and co-workers were able to develop and novel set of solute descriptors that are also related to free energy and prefer to correlate equilibrium features(M. H. Abraham, 1992; M. H. Abraham et al., 1988). In chemical, biological, and environmental processes, the solvation parameter model is now generally recognized as a useful tool for computing quantitative structure-property relationships. The model connects a system's free-energy attribute to six descriptors derived from free-energy that characterise molecular properties. The solvation model could be used to physiochemically describe stationary phases. or to develop a quantitative structure-property relationship to aid in the prediction of additional system attributes for compounds that lack experimental values.(Mutelet, 2012).This model is generally known as the solvation parameter model which is based on the parameterization of the solution cavity model and for the transmission of neutral molecules between gas and condensed phases, as it's written(Poole et al., 2013b).

log SP = c + eE + sS + aA + bB + lL---(10)

log SP = c + eE + sS + aA + bB + vV--(11)

log SP = c + sS + aA + bB + vV + lL--(12)

log SP = c + sS + aA + bB + lL--(13)

In the above equations, SP is the quality of a solute,E and S are polarity/polarizability, A, and B are hydrogen bonding interactions, V, and L are cavity formation and small letters c, e, s, a, b, v, and l are specified coefficients for each two partitioning phases(Khawar & Nabi, 2021a; Poole et al., 2013b).

In the first step, for the verification purpose, we developed the different variants of Abraham Salvation Models(ASMs) for determining the cuticle-water partitioning coefficients(Kcw) properties (Figure- 5). The multilinear regression analysis was run on Abraham Salvation Descriptors (E, S, A, B, L and V) to evaluate the significance of each descriptor for every models. Then the ASMs with all possible combinations such as ESABV (W. R. Abraham et al., 2002), SABVL, ESABL etc were developed individually through MLR analysis. The ESABV model stood out as the best fit model having good statistics such as large coefficient of correlation (R^2) values and small root mean square error (RMSE) values. We developed the following equation(12) for Kcw as an outcome of development Asselin table-3 show us the dataset for Abraham solvation base models.

Table 5. Dataset of 109 chemicals for the development of ASMs

No	**Chemical_name**	**E**	**S**	**A**	**B**	**L**	**V**	**logKcw**
1	4-Nitrophenol	1.07	1.64	0.93	0.21	5.568	0.9493	1.876
2	Fenuron	1.05	1.59	0.41	0.9	6.812	1.3544	0.65
3	Phenol	0.805	0.89	0.6	0.3	3.766	0.7751	1.502
4	bis(2-Ethylhexyl) phthalate	0.64	1.25	0	1.02	12.7	3.4014	7.406
5	Monuron	1.14	1.5	0.47	0.78	7.18	1.4768	1.625
6	Chlortoluron	1.25	1.53	0.4	0.8	8.017	1.6177	2.16
7	Atrazine	1.22	1.29	0.17	1.01	7.783	1.6196	2.13
8	Perylene	3.26	1.76	0	0.42	12.053	1.9536	6.5
9	Cyanazine	1.41	1.31	0.26	1.15	8.373	1.7743	1.81
10	Diuron	1.28	1.6	0.57	0.7	8.06	1.5992	2.465
11	Isoproturon	1.2	1.79	0.46	0.93	8.742	1.7771	2.13
12	Chlorfenvinphos	1.21	1.52	0	1.27	0	2.3254	3.04
13	Permethrin	2.05	1.42	0	0.88	12.827	2.8186	5.51
14	Bitertanol	2.3	1.5	0	1.67	12.88	2.6736	3.896
15	Triadimenol	1.601	1.58	0.26	1.28	10.51	2.1882	3.298
16	Benzoic acid	0.73	0.9	0.59	0.4	4.657	0.9317	1.69
17	Phenanthrene	2.055	1.29	0	0.29	7.632	1.4544	4.68
18	1-Naphthaleneacetic acid	1.46	1.55	0.6	0.67	7.809	1.4416	2.29
19	Pentachlorophenol	1.22	0.91	0.66	0.06	6.805	1.3871	4.585
20	2-Nitrophenol	1.015	1.05	0.06	0.35	4.778	0.9493	1.935
21	1-Naphthalenol	1.52	1.1	0.66	0.34	6.264	1.1441	2.993
22	Naphthalene	1.34	0.92	0	0.2	5.161	1.0854	3.39
23	2,4,5-Trichlorophenoxyacetic acid	1.4	1.34	0.78	0.53	7.644	1.4985	3.1925
24	2,4-Dichlorophenoxyacetic acid	1.21	1.36	0.77	0.63	6.985	1.3761	2.6483
25	Phenylurea	1.11	1.33	0.79	0.79	6.332	1.0726	0.87
26	Paclobutrazole	1.534	1.39	0.21	1.46	10.455	2.2704	2.36
27	Hexachlorobenzene	1.49	0.75	0	0.09	6.986	1.4508	5.81
28	Naringenin	2.23	1.8	1.38	1.22	10.889	1.8888	2.836
29	Styrene	0.849	0.65	0	0.16	3.856	0.9552	2.89
30	Epichlorohydrin	0.395	1.11	0	0.27	2.82	0.6038	0.485
31	1,2-Dibromoethane	0.747	0.76	0.1	0.17	3.382	0.7404	1.855
32	1,2-Dichloroethane	0.42	0.64	0.1	0.11	2.573	0.6352	1.485

33	Acrylonitrile	0.297	0.83	0.03	0.3	1.995	0.5021	0.265
34	1-Nitropropane	0.242	0.95	0	0.31	2.894	0.7055	0.87
35	4-Methyl-2-pentanone	0.111	0.65	0	0.51	3.089	0.9697	0.885
36	Toluene	0.601	0.52	0	0.14	3.325	0.8573	2.55
37	Propyl acetate	0.092	0.6	0	0.45	2.819	0.8875	0.84
38	Pyridine	0.631	0.84	0	0.52	3.022	0.6753	0.41
39	1-Hexanol	0.21	0.42	0.37	0.48	3.61	1.0127	1.325
40	Butyl acetate	0.071	0.6	0	0.45	3.353	1.0284	1.395
N	**Chemical_name**	**E**	**S**	**A**	**B**	**L**	**V**	**logKcw**
41	1,4-Dioxane	0.329	0.75	0	0.64	2.892	0.681	-0.555
42	Limonene	0.501	0.31	0	0.23	4.688	1.323	2.675
43	Ethyl acetate	0.106	0.62	0	0.45	2.314	0.7466	0.485
44	2-Hexanol	0.187	0.36	0.33	0.56	3.33	1.0127	1.005
45	Ethanol	0.246	0.42	0.37	0.48	1.485	0.4491	-0.855
46	Methanol	0.278	0.44	0.43	0.47	0.97	0.3082	-1.087
47	Pyrene	2.808	1.71	0	0.28	8.833	1.5846	5.98
48	Benzo[ghi]perylene	4.073	1.9	0	0.45	13.447	2.0838	7.41
49	Fluoranthene	2.377	1.55	0	0.24	8.827	1.5846	5.89
50	Chrysene	3.027	1.73	0	0.36	10.334	1.8234	6.41
51	Benzo[a]pyrene	3.625	1.96	0	0.37	11.736	1.9536	7.01
52	Dibenzo[a,h]anthracene	4	2.04	0	0.44	12.96	2.1924	7.55
53	Benz[a]anthracene	2.992	1.7	0	0.35	10.291	1.8234	6.57
54	Acenaphthene	1.604	1.05	0	0.22	6.469	1.2586	4.27
55	2-Propanol	0.212	0.36	0.33	0.56	1.764	0.59	-0.61
56	Acetone	0.179	0.7	0.04	0.49	1.696	0.547	-0.175
57	Chloroform	0.43	0.49	0.15	0.02	2.48	0.6167	1.785
58	1-Butanol	0.224	0.42	0.37	0.48	2.601	0.7309	0.26
59	1-Pentanol	0.219	0.42	0.37	0.48	3.106	0.8718	0.796
60	Benzene	0.61	0.52	0	0.14	2.786	0.7164	2.06
61	Acetonitrile	0.237	0.9	0.07	0.32	1.739	0.4042	-0.315
62	Dichloromethane	0.39	0.57	0.1	0.05	1.818	0.4943	1.415
63	2-Methyl-2-propanol	0.18	0.3	0.31	0.6	1.963	0.7309	-0.38
64	Trichloronitromethane	0.461	0.82	0	0.1	3.208	0.7909	2.205
65	1,2-Dichloropropane	0.37	0.63	0	0.17	2.836	0.7761	1.86
66	2-Butanol	0.217	0.36	0.33	0.56	2.338	0.7309	-0.055
67	2-Butanone	0.166	0.7	0	0.51	2.287	0.6879	-0.1
68	o-Xylene	0.663	0.56	0	0.16	3.939	0.9982	2.885
69	Paclobutrazol	1.534	1.39	0.21	1.46	10.455	2.2704	2.545
70	4-Nitroanisole	0.89	1.33	0.04	0.38	5.345	1.0902	1.925
71	2,4,6-Trinitrotoluene	1.39	1.76	0.12	0.63	7.044	1.3799	2.05
72	2,4-Dinitrotoluene	1.16	1.9	0	0.52	6.752	1.2644	1.955
73	Hexahydro-1,3,5-trinitro-1,3,5-triazine	1.38	2.35	0.56	0.55	7.532	1.2447	2.17
74	Salicylic acid	1.38	2.35	0.56	0.55	7.532	1.2447	2.034
75	1,2-Dichlorobenzene	0.872	0.78	0	0.04	4.518	0.9612	3.1625
76	Anthracene	2.29	1.34	0	0.28	7.568	1.4544	5.2
77	Ethylbenzene	0.613	0.51	0	0.15	3.778	0.9982	2.82
78	1-Propanol	0.236	0.42	0.37	0.48	2.031	0.59	-0.31

79	3-Chloroprop-1-ene	0.327	0.56	0	0.05	2.109	0.6106	1.66
80	Chlorobenzene	0.718	0.65	0	0.07	3.657	0.8388	2.7
81	Cyclohexanone	0.403	0.86	0	0.56	3.792	0.8611	0.32
82	Tetrahydrofuran	0.289	0.52	0	0.48	2.636	0.6223	0.12
83	Cyclohexane	0.31	0.1	0	0	2.964	0.8454	3.13
84	Heptane	0	0	0	0	3.173	1.0949	4.47
85	Chlorotoluron	1.25	1.53	0.4	0.8	8.017	1.6177	2.09
86	Carbon tetrachloride	0.46	0.38	0	0	2.823	0.7391	2.49
87	2-Pentanol	0.2	0.36	0.33	0.56	2.84	0.8718	0.46
88	1,1,1-Trichloroethane	0.369	0.48	0	0.08	2.751	0.7576	2.44
89	1,1-Dichloroethene	0.362	0.34	0	0.05	2.11	0.5922	2.04
90	2-Methyl-2-butanol	0.194	0.3	0.31	0.63	2.722	0.8718	0.11
N	**Chemical_name**	**E**	**S**	**A**	**B**	**L**	**V**	**logKcw**
91	3-Methyl-3-pentanol	0.21	0.3	0.31	0.6	3.277	1.0127	0.61
92	2-Methyl-1,3-butadiene	0.313	0.23	0	0.1	2.101	0.7271	2.09
93	2-Methyl-1-propanol	0.217	0.39	0.37	0.48	2.413	0.7309	0.11
94	Trichloroethene	0.52	0.37	0.08	0.03	2.997	0.7146	2.56
95	Tetrachloroethene	0.64	0.44	0	0	3.584	0.837	3.05
96	Carbaryl	1.51	1.67	0.22	0.79	7.97	1.5414	2.17
97	Tributyl phosphate	-0.1	0.62	0	1.29	7.522	2.2388	2.54
98	Diethyl suberate	0.07	1.12	0	1.01	6.95	1.9482	1.96
99	2,4-Dichlorophenoxybutyric acid	1.2	1.3	0.55	0.64	7.882	1.6579	3.05
100	PCB 4	1.6	1.22	0	0.2	6.815	1.569	4.93
101	PCB 3	1.5	1.05	0	0.18	6.718	1.4466	4.83
102	PCB 138	2.18	1.74	0	0.11	9.772	2.0586	7.03
103	PCB 180	2.29	1.87	0	0.09	10.415	2.181	7.13
104	PCB 52	1.9	1.48	0	0.15	8.144	1.8138	5.93
105	PCB 101	2.04	1.61	0	0.13	8.868	1.9362	6.43
106	PCB 118	2.06	1.59	0	0.11	9.396	1.9362	6.73
107	PCB 28	1.76	1.33	0	0.15	7.904	1.6914	6.03
108	Tebuconazole	1.54	1.45	0.24	1.44	10.96	2.4113	3
109	Water	0	0.6	0.59	0.46	0.245	0.1673	-1.53

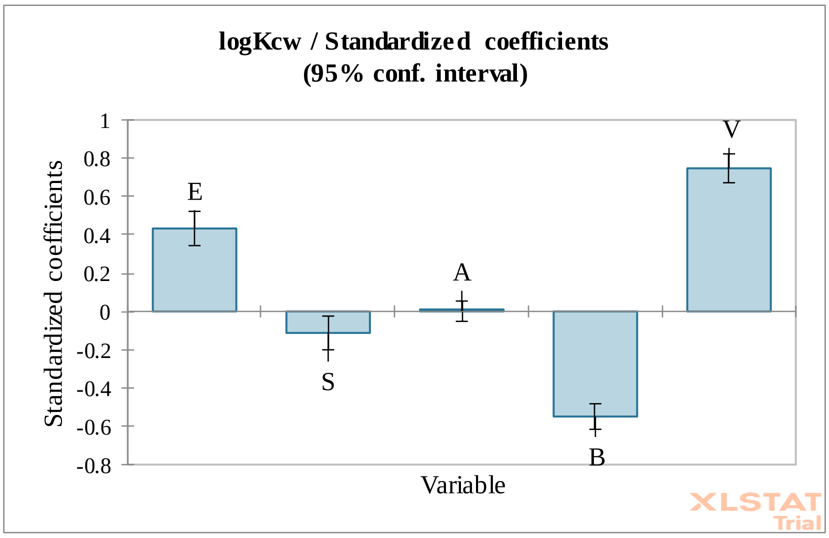


Figure-4.1 Standardized coefficient

Here this standardized coefficient chart shows us the contribution , influence, and strength of each independent variable to the dependent variable. The higher the value of the coefficient of the independent variable in this chart indicates the stronger influence and effect on the model. Here in this Abraham Sovation Model V and E descriptors are covering maximum variance and they have a positive influence on the model.

$$\boldsymbol{logK}_{\boldsymbol{cw}}\boldsymbol{=-0.1(\pm 0.075)+0.888(\pm 0.064)E-0.555(\pm 0.0.99)S-0.160(\pm 0.126)A-3.734(\pm 0.116)B}$$

$$\boldsymbol{+3.267(\pm 0.089)V}------------------------------------\boldsymbol{(14)}$$

N=109 R^2= 0.98 RMSE=0.3

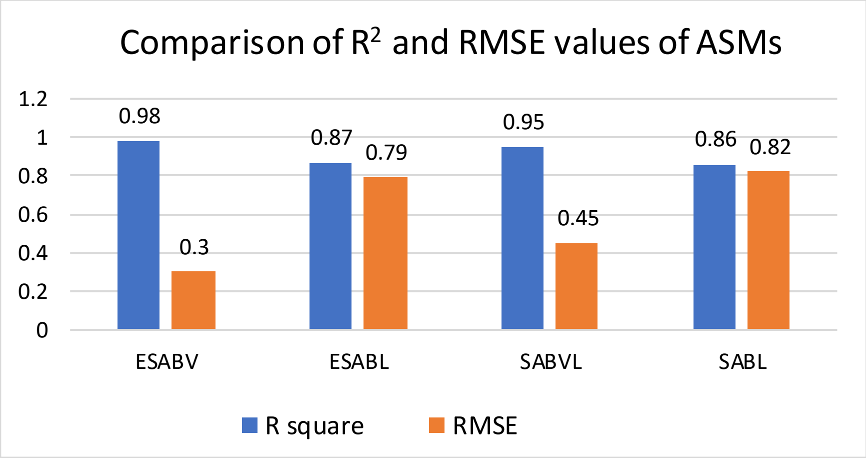


Figure- 4.2 comparisons of R^2 for different variants of ASMs.

Results of Multilinear regression (MLR) proved the ESABV model as the best fit model among all other variants based on values of Root Mean Square Error(RMSE=0.3) and R square R^2=0.98.

4.1.2 Scatter plot of ASDs

For roughly determination and explanation of the linear correlation between multiple variables, scatterplot matrices are one of the best ways. In the scatterplot matrix from top left to bottom right, the variables are written in a diagonal line. For instance, in the last rectangle of the first column in an independent scatter plot of logKcw and E, with logKcw as X-axis and E as the Y-axis. In the middle of the top row, the same plot is repeated just in the below part shown as a box plot and the above parts show us by number types correlation of these variables.In essence, the plots on the lower left and upper right sides of the scatterplot are mirror reflections of each other. The density distribution of each variable is shown along the diagonal direction.

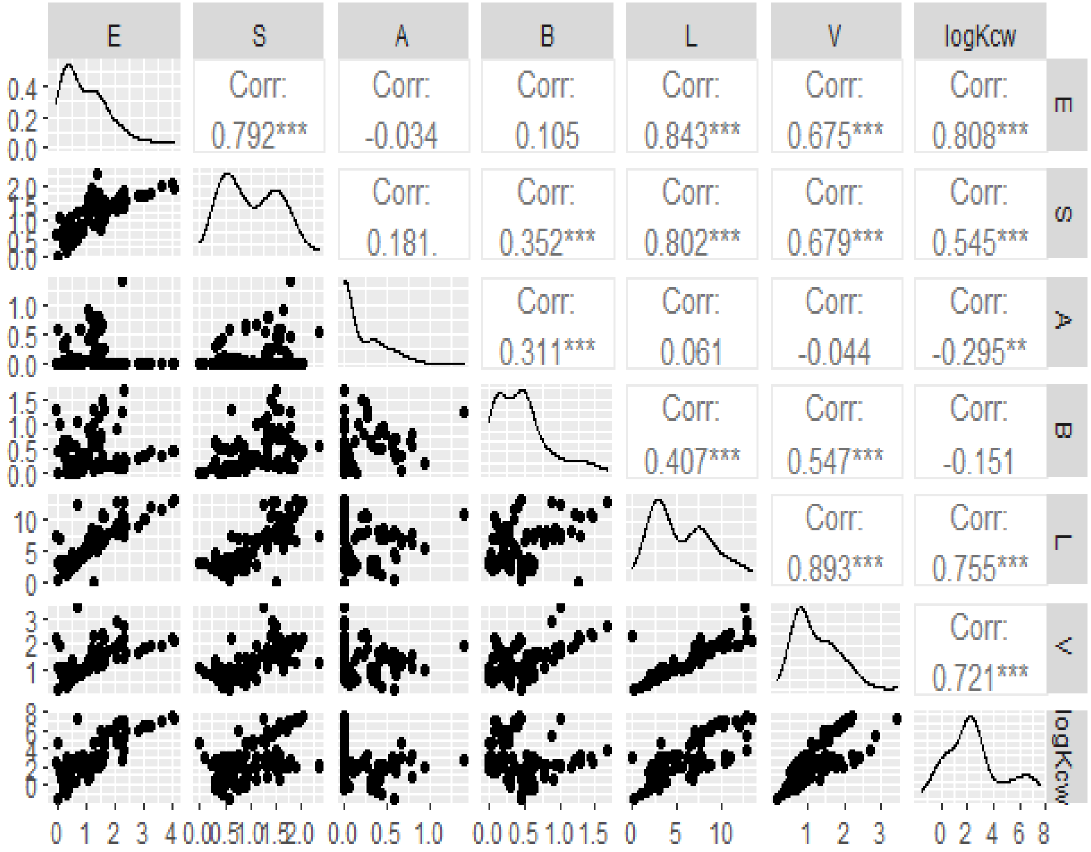


Figure-4.3 Scatter plot of ASDs

In this scatterplot, we can see that there is a good correlation between logKcw and E, S, V,L variables(positive correlation).There is probably less of a correlation between logKcw and A,B variables. More statistical analyses would be needed to confirm or deny this.

4.1.3 Cross-Validation of Abraham Solvation Models

The resampling technique which is used for the robustness and assessment of the model to avoid overfitting is called cross-validation(Berrar, 2018). Cross-validation was first proposed in the 1930s. Data is divided into two segments in cross-validation, one for model training and the other for model validation. The training and validation sets must cross over in successive rounds in standard cross-validation to validate each data point. Now it is widely acknowledged as a standard approach in the data mining and machine learning communities for performance estimation and model selection. Maine's goal of cross-validations uses a single algorithm, to estimate the trained model's performance from available data and assess the performance of two or more distinct algorithms to determine which one is the greatest fit for the data(Refaeilzadeh et al., 2020). The methods we used for our modle validation they are K-Fold cross validation , Leave-one out cross validation, Hold out and boot strap cross validation every one of them we will explain breafly in coming pages.

1-K-Fold Validation

At the start, data is divided into k folds or segments of equal size common folds are 5-10 folds and then there are k rounds of training and validation, with each iteration using a different fold of the data for validation and the remaining k -1 folds for learning(Refaeilzadeh et al., 2020).

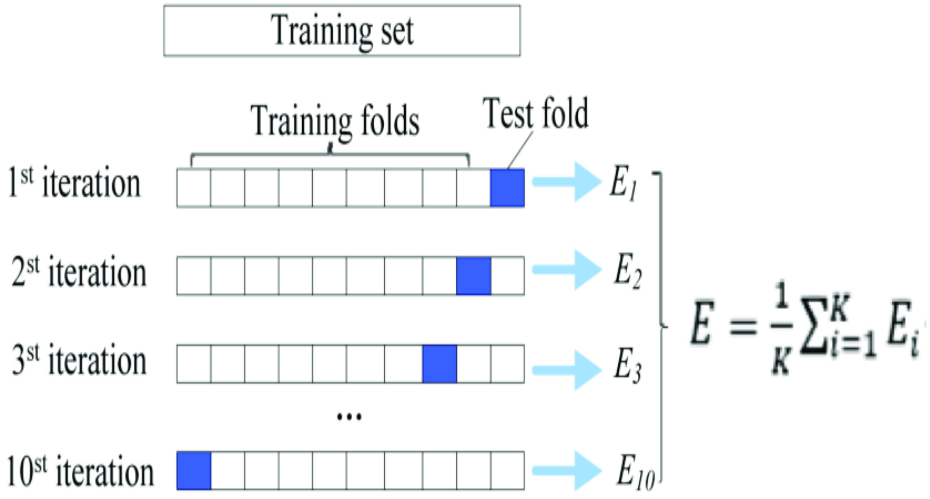


Figure-4.4 systematic diagram of K-fold validation(https://www.researchgate.net/)

2-Leave-One-Out Validation(LOOV)

It is a variant of k-fold cross-validation in which k is the number of samples in the data. In other words, except for a single observation, practically all of the data is utilized for training in each iteration, and the model is evaluated on that one observation. This approach is accurate and unbiased, but it has a huge variation, making estimations untrustworthy(Refaeilzadeh et al., 2020).

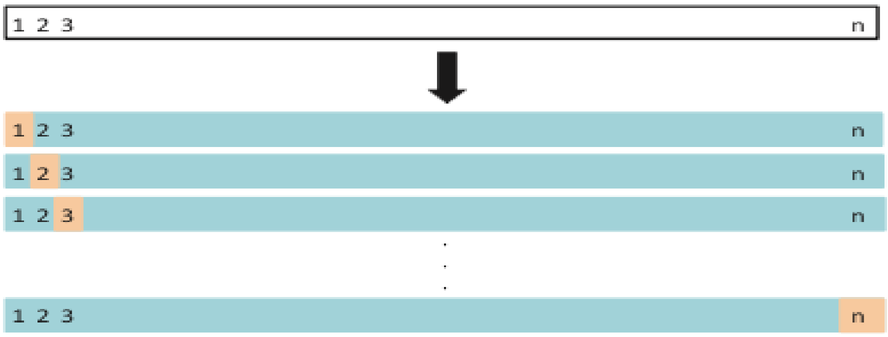


Figure-4.5 systematic diagram of LOOCV(https://www.researchgate.net/).

3- Hold-Out Validation

In this approach, the data set is bisection into two sub-samples training and testing data the validation commonly by the ratio of 1:4.This method avoids overlap between training and test samples. Generalization performance of this method is more accurate but the disadvantage is that result depends on split data for training and test. When data is few, in spite of that commonly employed, where only a few hundreds of data samples are available. (Refaeilzadeh et al., 2020).

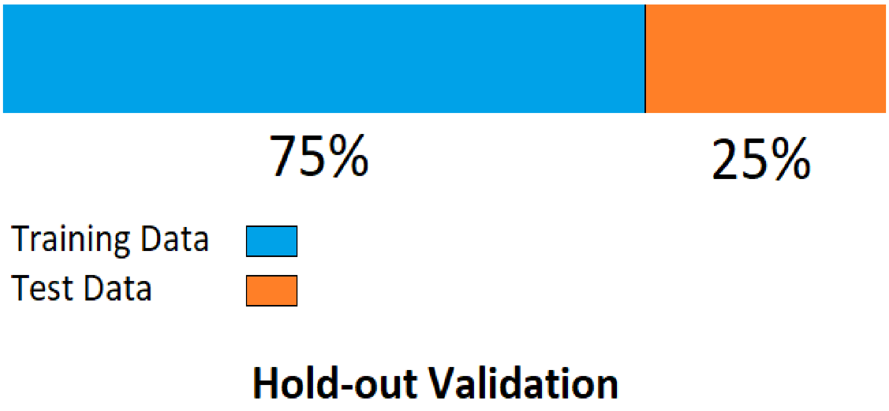


Figure-4.6 systematic diagram of Hold-out validation(https://medium.datadriveninvestor.com)

4- Bootstrap Validation

It is a randomized method. In the field of statistics has a long tradition(Koehn, 2004)

In this approach randomly select the observation from data set and iterate it,like here we take 100 observations for validation.

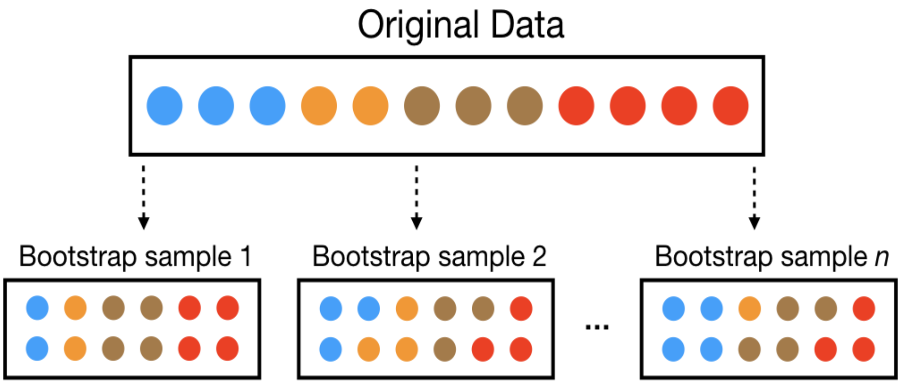


Figure-4.7 systematic diagram of Bootstrap cross-validation(https://bradleyboehmke.github.io/HOML/)

Table-6 Advantage and Disadvantage of different Cross-validation

Validation Approaches	**Advantages**	**Disadvantages**
Hold-out	Independent testing and training	Data for training and testing is few, and there is a lot of variation.
k-fold cross	Estimating performance with precision	Small performance estimation samples, overlapping training data, For contrast, there has been an increase in type I error, Exaggerated degree of freedom for comparison or understated performance variance
LOO	Estimation of performance unbiased	Extremely wide range
Bootstrap	It may be used for nonlinear regression and classification.	Sample sizes are tiny

The result of cross-validations was done on ASDs below.

Table-7 Cross-Validation of ASMs

Property	External Validation (Hold-out Approach)			
LogKcw	No of observations	RMSE	MAE	R^2
Train set (80%)	89	0.26	0.22	0.99
Test set (20%)	14	0.21	0.17	0.99

property	K-fold cross-validation			Leave one out approach			Bootstrap Approach		
LogKcw	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
	0.30	0.22	0.98	0.31	0.22	0.97	0.32	0.23	0.98

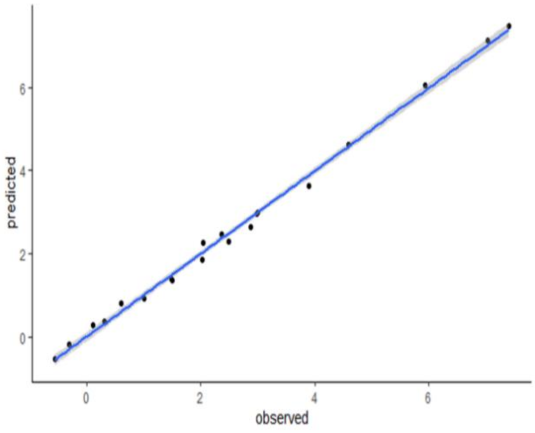


Figure-4.8 test sample observed and predicted values

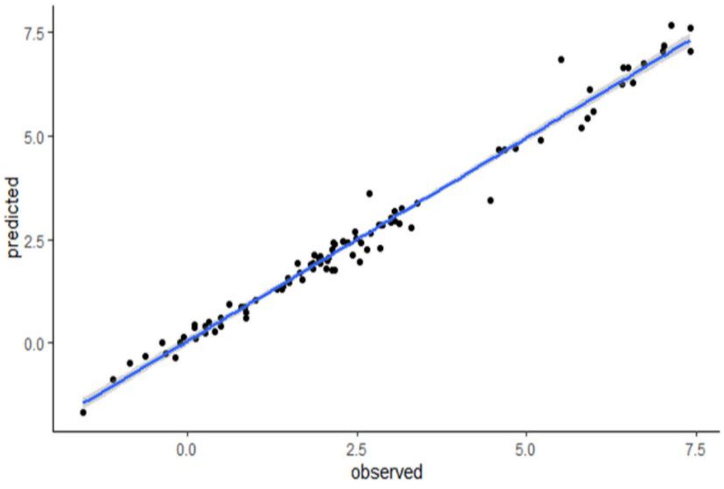


Figure-4.9train sample observed and predicted values

Figure-4.8 and **4.9** indicate the correlation of observed and predicted values of test and train samples in whin the blue line indicates the correlation line and the points indicate the observations in our data sets. The distance between the observation and correlation line is called residual.

4.2 Justification, Formulation, Validation Partition models (PMs) for Kcw estimation, and comparison of ASMs and the two-parameter base model

4.2.1 Justification of logKow and logKaw based two-parameter models

It was hypothesized that new two-parameter linear free energy relationship partitioning model PM using $\log K_{ow}$ and $\log K_{aw}$ would have comparable predictive ability to the ASMs. To explore our proposition, the information content present in the ASDs of the training sets of the ASMs was analyzed thoroughly. In the previous section, it was concluded that the minimum five dimensions are required to elucidate the variability in the Kcw data of the Abraham solvation models. To investigate this further, the PCA was run on individual data sets of Kcw. PCA was performed on Kcw data comprising E, S, A, B, and V descriptors. It was found that maximum information was captured in the first two dimensions. The table-5 shows chemicals used for conducting two-parameter base model.

Table-8 Dataset of 117 chemicals for creating a partitioning model(PM)

No	**Chemical_name**	**E**	**S**	**A**	**B**	**V**	**logKcw**	**LogKow**	**LogKaw**
1	4-Nitrophenol	1.07	1.72	0.82	0.26	0.9493	1.876	1.91	-7.03582
2	Fenuron	1.05	1.31	0.37	0.96	1.3544	0.65	0.98	-7.39299
3	Phenol	0.81	0.89	0.6	0.3	0.7751	1.502	1.46	-4.63125
4	bis(2-Ethylhexyl) phthalate	0.64	1.25	0	1.02	3.4014	7.406	7.6	-3.30833
5	Monuron	1.14	1.5	0.47	0.78	1.4768	1.625	1.94	-7.52348
6	Chlortoluron	1.11	1.5	0.47	0.81	1.6177	2.16	2.41	-7.48039
7	Atrazine	1.22	1.29	0.17	1.01	1.6196	2.13	2.61	-6.7299
8	Perylene	3.26	1.76	0	0.4	1.9536	6.5	6.25	-4.47173
9	Cyanazine	1.41	2	0.22	1.14	1.7743	1.81	2.22	-10.1107
10	Diuron	1.28	1.6	0.57	0.7	1.5992	2.465	2.68	-7.65348
11	Isoproturon	1.2	1.79	0.46	0.93	1.7771	2.13	2.87	-7.10375
12	Chlorfenvinphos	1.21	1.56	0	0.99	2.3254	3.04	3.81	-5.66672
13	Permethrin	2.05	1.42	0	0.88	2.8186	5.51	6.5	-4.92082
14	Bitertanol	2.3	1.5	0	1.67	2.6736	3.896	4.16	-10.168
15	Triadimenol	1.6	1.58	0.26	1.28	2.1882	3.298	3.08	-9.18155
16	Benzoic acid	0.73	0.9	0.59	0.4	0.9317	1.69	1.87	-5.34679
17	Phenanthrene	2.06	1.29	0	0.26	1.4544	4.68	4.46	-2.67009
18	1-Naphthaleneacetic acid	1.46	1.55	0.6	0.67	1.4416	2.29	2.24	-6.74473
19	Pentachlorophenol	1.22	0.91	0.66	0.06	1.3871	4.585	5.12	-5.2833
20	2-Nitrophenol	1.02	1.05	0.05	0.37	0.9493	1.935	1.79	-3.53573
21	1-Naphthalenol	1.52	1.05	0.61	0.37	1.1441	2.993	2.85	-5.64222
22	Naphthalene	1.34	0.92	0	0.2	1.0854	3.39	3.3	-1.65923
23	2,4,5-Trichlorophenoxyacetic acid	1.4	1.34	0.78	0.53	1.4985	3.1925	3.31	-6.54579
24	2,4-Dichlorophenoxyacetic acid	1.21	1.36	0.77	0.63	1.3761	2.6483	2.81	-6.41595
25	Phenylurea	1.11	1.33	0.79	0.79	1.0726	0.87	0.83	-8.07702

No	Chemical_name	E	S	A	B	V	logKcw	LogKow	LogKaw
26	Paclobutrazole	1.53	1.39	0.21	1.46	2.2704	2.36	3.2	-7.78915
27	Hexachlorobenzene	1.49	0.99	0	0	1.4508	5.81	5.73	-1.42985
28	Naringenin	2.23	1.8	1.38	1.22	1.8888	2.836	2.52	-14.9133
29	Styrene	0.85	0.65	0	0.16	0.9552	2.89	2.95	-0.9393
30	Epichlorohydrin	0.4	1.11	0	0.27	0.6038	0.485	0.45	-2.63047
31	1,2-Dibromoethane	0.75	0.76	0.1	0.17	0.7404	1.855	1.96	-1.26627
32	1,2-Dichloroethane	0.42	0.64	0.1	0.11	0.6352	1.485	1.48	-0.29743
33	Acrylonitrile	0.3	0.83	0.03	0.3	0.5021	0.265	0.25	-2.24033
34	1-Nitropropane	0.24	0.95	0	0.31	0.7055	0.87	0.87	-2.50981
35	4-Methyl-2-pentanone	0.11	0.65	0	0.51	0.9697	0.885	1.31	-2.31575
36	Toluene	0.6	0.52	0	0.14	0.8573	2.55	2.73	-0.60569
37	Propyl acetate	0.09	0.6	0	0.45	0.8875	0.84	1.24	-1.89025
38	Pyridine	0.63	0.84	0	0.52	0.6753	0.41	0.65	-3.53202
39	1-Hexanol	0.21	0.42	0.37	0.48	1.0127	1.325	2.03	-3.1347
40	Butyl acetate	0.07	0.6	0	0.45	1.0284	1.395	1.78	-1.76743
41	1,4-Dioxane	0.33	0.75	0	0.64	0.681	-0.555	-0.27	-3.60862
42	Limonene	0.495	0.295	0	0.22	1.323	2.675	4.38	1.199572
43	Ethyl acetate	0.11	0.62	0	0.45	0.7466	0.485	0.73	-2.01286
44	2-Hexanol	0.19	0.36	0.33	0.56	1.0127	1.005	1.76	-3.1347
45	Ethanol	0.25	0.42	0.37	0.48	0.4491	-0.855	-0.31	-3.62663
46	Methanol	0.28	0.44	0.43	0.47	0.3082	-1.087	-0.77	-3.74978
47	Pyrene	2.81	1.71	0	0.28	1.5846	5.98	4.88	-3.46113
48	Benzo[ghi]perylene	4.07	1.9	0	0.45	2.0838	7.41	6.763	-5.26294
49	Fluoranthene	2.38	1.55	0	0.24	1.5846	5.89	5.16	-3.46113
50	Chrysene	3.03	1.73	0	0.33	1.8234	6.41	5.81	-3.68037
51	Benzo[a]pyrene	3.63	1.98	0	0.44	1.9536	7.01	5.81	-3.68037
52	Dibenzo[a,h]anthracene	4	2.04	0	0.44	2.1924	7.55	6.54	-4.6909
53	Benz[a]anthracene	2.99	1.7	0	0.35	1.8234	6.57	5.76	-3.68037
54	Acenaphthene	1.6	1.05	0	0.22	1.2586	4.27	3.92	-1.92996
55	2-Propanol	0.21	0.36	0.33	0.56	0.59	-0.61	0.05	-3.50399
56	Acetone	0.18	0.7	0.04	0.49	0.547	-0.175	-0.24	-2.68473
57	Chloroform	0.43	0.49	0.15	0.02	0.6167	1.785	1.97	-0.87236
58	1-Butanol	0.22	0.42	0.37	0.48	0.7309	0.26	0.88	-3.38065
59	1-Pentanol	0.22	0.42	0.37	0.48	0.8718	0.796	0.25	-3.50399
60	Benzene	0.61	0.52	0	0.14	0.7164	2.06	2.13	-0.64862
61	Acetonitrile	0.24	0.9	0.04	0.33	0.4042	-0.315	-0.34	-2.89449
62	Dichloromethane	0.39	0.57	0.1	0.05	0.4943	1.415	1.25	-0.41927
63	2-Methyl-2-propanol	0.18	0.3	0.31	0.6	0.7309	-0.38	0.35	-3.38065
64	Trichloronitromethane	0.16	0.82	0	0.1	0.7909	2.205	2.09	-4.11539
65	1,2-Dichloropropane	0.37	0.63	0	0.17	0.7761	1.86	1.98	-0.17339
66	2-Butanol	0.22	0.36	0.33	0.56	0.7309	-0.055	0.61	-3.38065
67	2-Butanone	0.17	0.7	0	0.51	0.6879	-0.1	0.29	-2.56199
68	o-Xylene	0.66	0.56	0	0.16	0.9982	2.885	3.16	-0.56331
69	Paclobutrazol	1.53	1.39	0.21	1.46	2.2704	2.545	3.2	-7.78915
70	4-Nitroanisole	0.98	1.49	0	0.37	1.0902	1.925	2.03	-4.27984
71	2,4,6-Trinitrotoluene	1.43	1.84	0	0.63	1.3799	2.05	1.6	-7.81792
72	2,4-Dinitrotoluene	1.15	1.58	0	0.49	1.2057	1.955	1.98	-5.4136

73	Hexahydro-1,3,5-trinitro-1,3,5-triazine	1.63	1.44	0	0.97	1.2447	2.17	0.87	-5.57949
74	Salicylic acid	0.9	0.85	0.73	0.37	0.9904	2.034	2.26	-6.22792
75	1,2-Dichlorobenzene	0.87	0.78	0	0.04	0.9612	3.1625	3.43	-0.90892
76	Anthracene	2.29	1.34	0	0.28	1.4544	5.2	4.45	-2.67009
No	**Chemical_name**	**E**	**S**	**A**	**B**	**V**	**logKcw**	**LogKow**	**LogKaw**
77	Ethylbenzene	0.61	0.51	0	0.15	0.9982	2.82	3.15	-0.48313
78	1-Propanol	0.24	0.42	0.37	0.48	0.59	-0.31	0.25	-3.50399
79	3-Chloroprop-1-ene	0.33	0.56	0	0.05	0.6107	1.66	1.839	0.152543
80	Chlorobenzene	0.72	0.65	0	0.07	0.8388	2.7	2.84	-0.77924
81	Cyclohexanone	0.4	0.86	0	0.56	0.8611	0.32	0.81	-2.67179
82	Tetrahydrofuran	0.29	0.52	0	0.48	0.6223	0.12	0.46	-2.45438
83	Cyclohexane	0.31	0.1	0	0	0.8454	3.13	3.44	1.026329
84	Heptane	0	0	0	0	1.0949	4.47	4.66	1.975815
85	Chlorotoluron	1.11	1.5	0.47	0.81	1.6177	2.09	2.41	-7.48039
86	Carbon tetrachloride	0.46	0.38	0	0	0.7391	2.49	2.83	0.024622
87	2-Pentanol	0.2	0.36	0.33	0.56	0.8718	0.46	1.19	-3.25636
88	1,1,1-Trichloroethane	0.37	0.41	0	0.09	0.7576	2.44	2.49	-0.74978
89	1,1-Dichloroethene	0.36	0.34	0	0.05	0.5922	2.04	2.13	0.123579
90	2-Methyl-2-butanol	0.19	0.3	0.31	0.6	0.8718	0.11	0.89	-3.25636
91	3-Methyl-3-pentanol	0.21	0.3	0.31	0.6	1.0127	0.61	1.687	-3.1347
92	2-Methyl-1,3-butadiene	0.31	0.23	0	0.1	0.7271	2.09	2.42	0.706149
93	2-Methyl-1-propanol	0.22	0.39	0.37	0.48	0.7309	0.11	0.76	-3.38065
94	Trichloroethene	0.52	0.37	0.08	0.03	0.7146	2.56	2.42	-0.01848
95	Tetrachloroethene	0.64	0.44	0	0	0.837	3.05	3.4	-0.16273
96	Octagen	1.77	2.76	0	1.29	1.6596	1.58	0.16	-7.44219
97	Carbaryl	1.51	1.67	0.22	0.79	1.5414	2.17	2.36	-6.88328
98	Tributyl phosphate	-0.1	0.71	0	1.26	2.2388	2.54	4	-3.87642
99	Diethyl suberate	0.07	1.12	0	1.01	1.9482	1.96	2.883	-3.89877
100	2,4-Dichlorophenoxybutyric acid	1.2	1.3	0.55	0.64	1.6579	3.05	2.81	-6.41595
101	PCB 4	1.6	1.22	0	0.2	1.569	4.93	5	-2.02419
102	PCB 3	1.5	1.05	0	0.18	1.4466	4.83	4.61	-1.89307
103	PCB 138	2.18	1.74	0	0.11	2.0586	7.03	7.44	-2.54452
104	PCB 180	2.29	1.87	0	0.09	2.181	7.13	7.41	-2.6752
105	PCB 52	1.9	1.48	0	0.15	1.8138	5.93	6.09	-2.2833
106	PCB 101	2.04	1.61	0	0.13	1.9362	6.43	6.8	-2.41454
107	PCB 118	2.06	1.59	0	0.11	1.9362	6.73	7.12	-2.41454
108	PCB 28	1.76	1.33	0	0.15	1.6914	6.03	5.62	-2.1549
109	Tebuconazole	1.54	1.45	0.24	1.44	2.4113	3	3.7	-7.66588
110	Water	0	0.45	0.82	0.35	0.1673	-1.53	-1.38	-6.45542
111	Xylose	1.11	1.4	1.05	1.55	0.998	0.5	-3.02	-11.0792
112	Urea	0.5	1.49	0.83	0.84	0.4648	-0.7	-2.11	-7.81792
113	Glucose	1.34	1.7	1.14	1.8	1.1976	0.75	-3.24	-12.3925
114	Octachlorostyrene	1.8	1.15	0	0	1.9344	5.02	7.26	-2.01848
115	1,2,3-Trichlorobenzene	1.03	0.86	0	0	1.0836	2.75	4.05	-1.03977
116	1,2,3,4-Tetrachlorobenzene	1.18	0.92	0	0	1.206	3.47	4.6	-1.1707
117	Pentachlorobenzene	1.33	0.92	0.06	0	1.3284	4.37	5.17	-1.30103

4.2.2 Formulation of two-parameter models

For the formulation of two-parameter models, we conducted the below analysis.

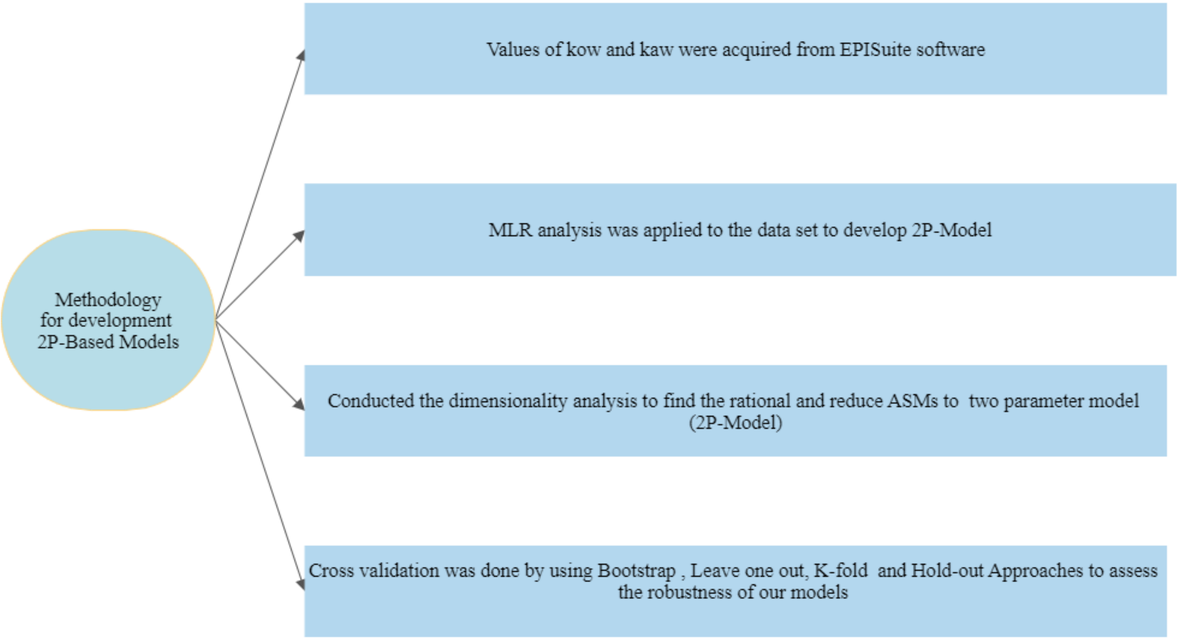

Methodology for development 2P-Based Models
Values of kow and kaw were acquired from EPISuite software
MLR analysis was applied to the data set to develop 2P-Model
Conducted the dimensionality analysis to find the rational and reduce ASMs to two parameter model (2P-Model)
Cross validation was done by using Bootstrap , Leave one out, K-fold and Hold-out Approaches to assess the robustness of our models

Figure-4.10 Flow chart of the methodology for development 2p based models

1- Multi-linear Regression analysis

Machine learning is wide interval utilized in a different of sectors to address complex issues that are difficult to solve using traditional computer methods. Linear regression is one of the most fundamental and extensively used machine learning approaches which is an approach to performing predictive analysis that is based on mathematics. In 1894, Sir Francis Galton for the first time, he proposed the concept of linear regression (Maulud & Abdulazeez, 2020). Linear regression is a mathematical test that evaluates and quantifies the relationship between the variables under consideration. Linear regression allows for projections of continuous/real or mathematical variables(Abdulqader et al., 2020). Linear regression is a modeling technique and popular mathematical research tool that allows you to assess and estimate anticipated effects versus many input variables. Also it develop linear relation between dependent and independent variables(Maulud & Abdulazeez, 2020). Multi-Linear-Regression MLR is a statistical approach that uses several illustrative factors to foresee the result of a response variable. The aim of (MLR) is to represent the linear relationship that will be evaluated between the independent variable x and the dependent variable y(Copy et al., 2019). MLR's fundamental model is as follows(Najat & Abdulazeez, 2018):

$$y = \beta 0 + \beta 1x1 + \cdots \beta mxm + \varepsilon \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (15)$$

First we conduct one- parameter linear free energy relationship model for the prediction of Kcw by octanol/water partition coefficient(Kow) which show the hydrophobicity and hydrophlicity of chemicals but it was not perfect model for highly polar and hydrophobic compounds.So we take Kaw(air/water patition coefficient)basically it show the volatility and hydrogen bond of the chemicals.below equation show one-parameter model.

$LogKcw = 0.94(\pm 0.02)Kow$…………………………………………………...(16)

R^2=0.94 n=117

For development of two-parameter models we did multiple linear regression analysis on Kcw as dependent variable and Kow and Kaw independent variables. As a result of MLR below equation developed.

$$Kcw = -0.052(\pm 0.016)Kaw + 0.902(\pm 0.022)Kow \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (17)$$

Or $Kcw = -0.052 * Kaw + 0.902 * Kow$

RMSE=0.71 R^2=0.95 N=117

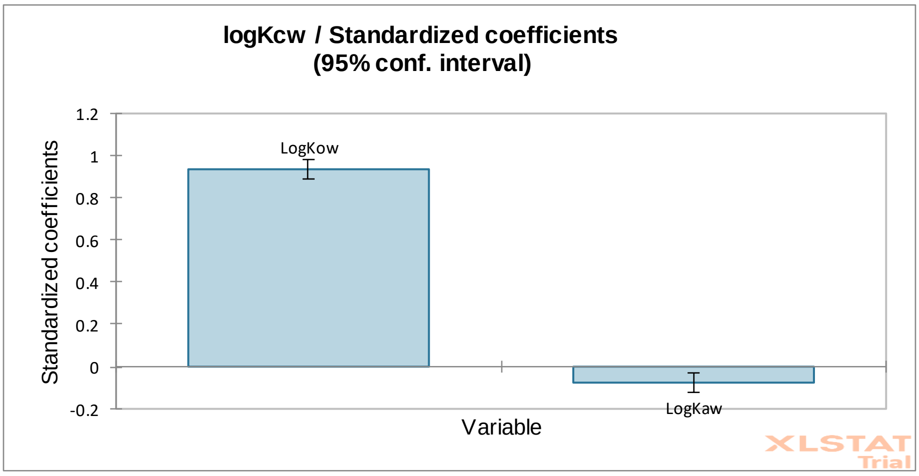


Figure-4.11 Standardized Coefficients

Herein **Figure-4.11** this standardized coefficient chart shows us the contribution , influence, and strength of each independent variable to the dependent variable. The higher the value of the coefficient of the independent variable in this chart indicates the stronger influence and effect on the model. Here in this 2P-Model Kow descriptor is covering maximum variance and it has a positive influence on the model.

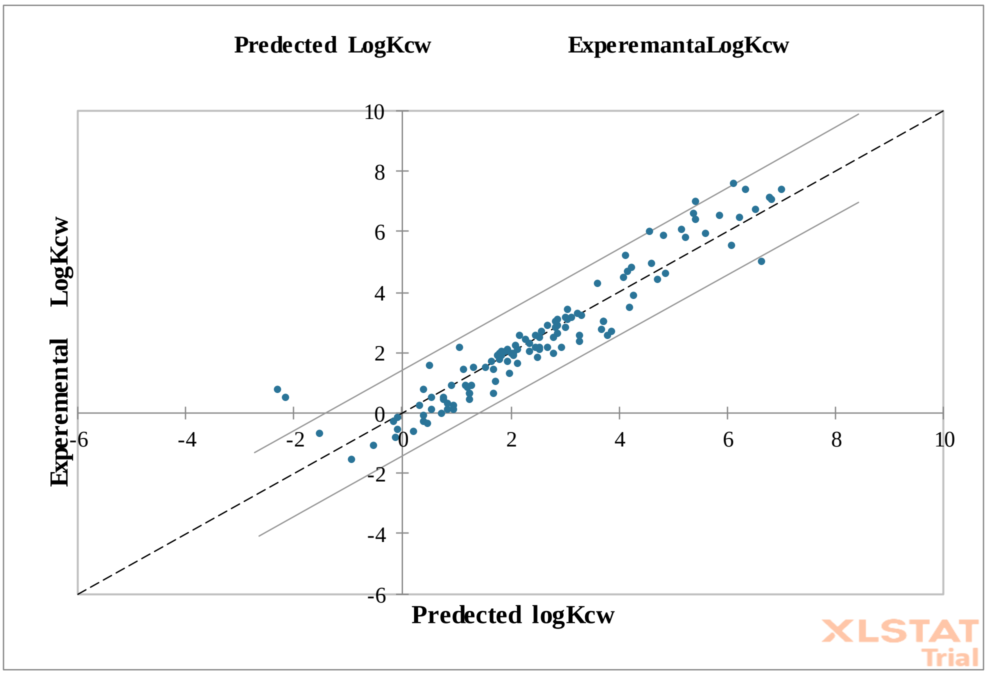


Figure-4.12 predicted and experimental Kcw

Figure-4.12 Show us the linear regression plot of the two-parameter model between experimental and predicted values of Kcw. Upper and lower green lines bound 95% confidence interval around.

2-Importance and interpretation of the dummy Variables

Dummy variables or indicator variables are arbitrary variables that were used to incorporate the datasets of all the equations in a single equation to get a common intercept and the same slope. Incorporating categorical variables into regression analysis is possible with the use of indicator variables. Incorporating categorical variables into regression analysis is possible with the use of indicator variables. The most common strategy for fitting this relation is to minimize the sum of squared errors between the observed values and the value that would fit under the hypothesized relationship(Bower, 2018). Indicator variables, also known as impulse dummies, and combinations of them are used to eliminate residuals that would otherwise be outliers in estimated time-series relationships. Any variations in the coefficients of deterministic variables, or combinations of distributions, as well as data measurement or recording mistakes, can cause outliers. Hendry (1999) adopted the notion of establishing an index to replace the original dummies in his analysis of US Food Expenditure(Singh & Balange, 2017). Here in our dataset indicator variables are plant tissue types Cuticular membrane(CM), Polymer Matrix Membrane(MX), and whole plant biomass. The purpose of doing so was to increase the chemical space and to have one equation for determining the Kcw for all kind of plant tissues discussed in this study. The indicator variables decide the mode of action of chemicals to determine the Kcw for different plant tissues types. For example, if we want to check the potency of ethanol for different plant tissues, we will use the corresponding dummy variable=1 reserved for that plant tissue along with its coefficient, while at the same time other dummy variables=0 by using dummy variable we developed below equation.

$$Kcw = (0.953)Kow - (0.028)Kaw - (0.057)CM - (0.139)MX - (0.107)Whole \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots.(18)$$

RMSE=0.57 R^2=0.97 N=227

3-Scatter plot of two parameter model

For roughly determination and explanation of the linear correlation between multiple variables, scatterplot matrices are one of the best ways. In the scatterplot matrix from top left to bottom right, the variables are written in a diagonal line. For instance, in the second rectangle of the first column in an independent scatter plot of logKcw and logKow, with logKcw as Y-axis and logKow as X-axis. In the middle of the top row, the same plot is repeated just in the below part shown as a box plot and the above parts show us by number types correlation of these variables.In essence, the plots on the lower left and upper right sides of the scatterplot are mirror reflections of each other. The density distribution of each variable is shown along the diagonal direction.

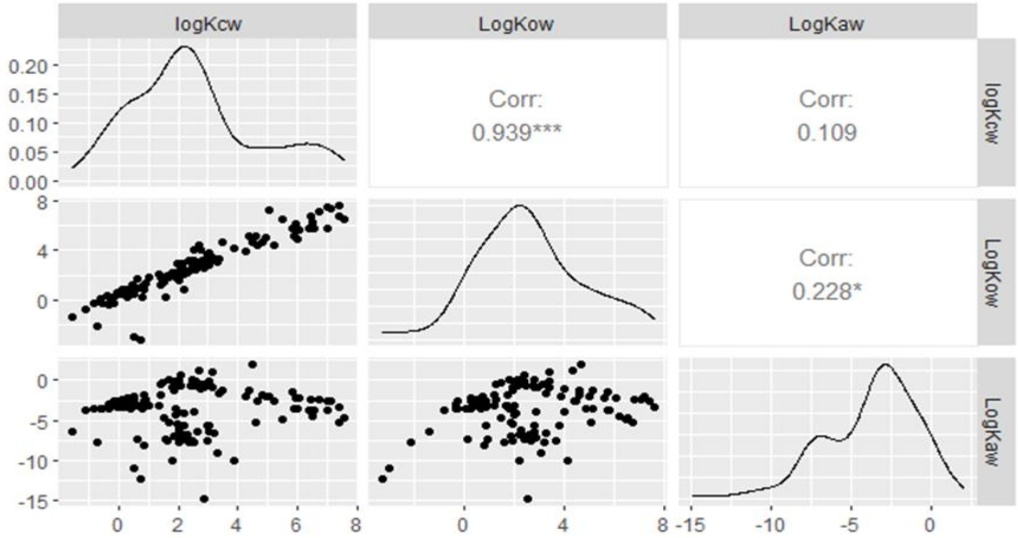
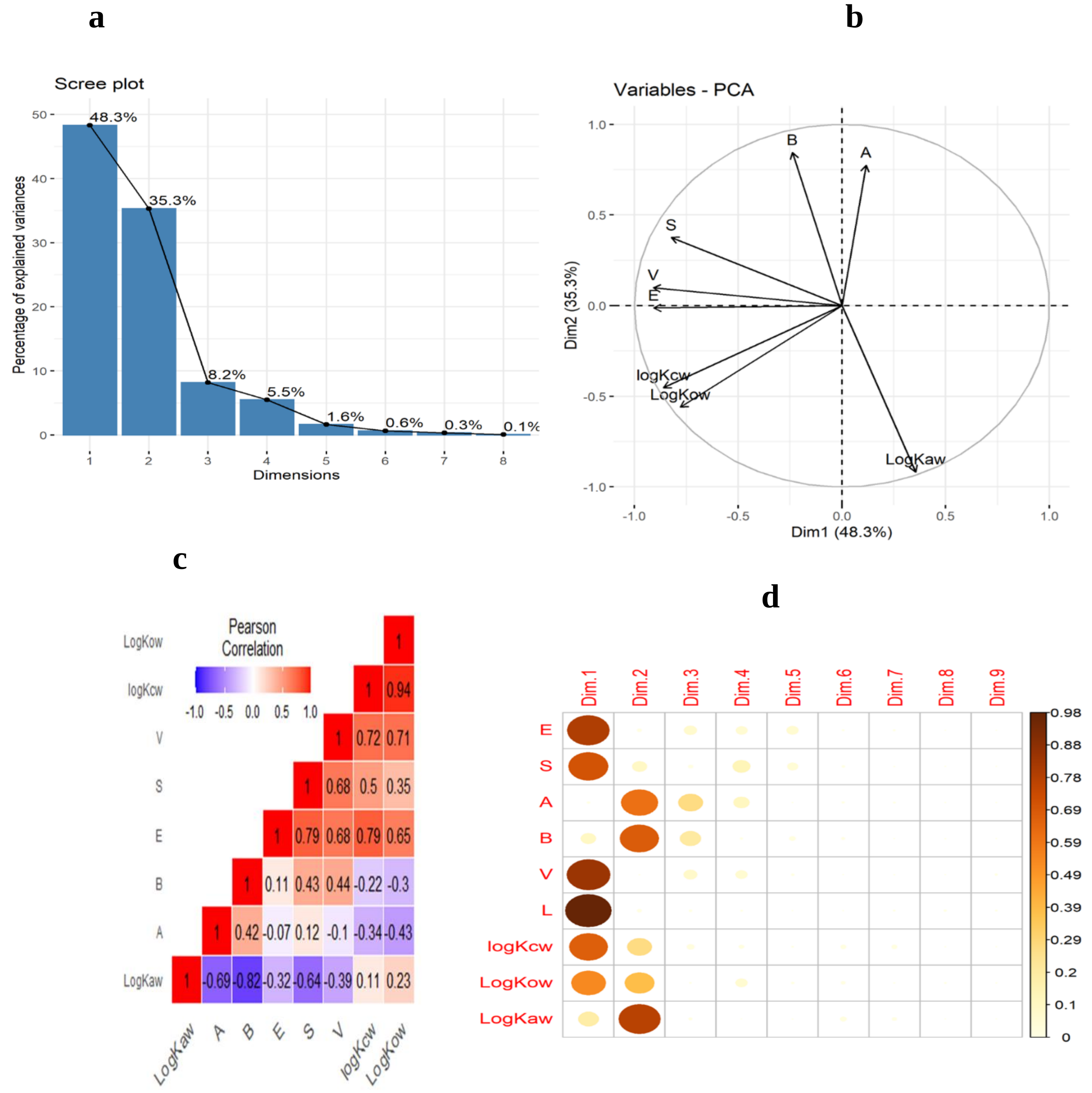


Figure-4.13 Scatterplot of two parameter model

4- Principal Component Analysis(PCA)

There are a variety of Dimensionality Reduction (DR) techniques like Fourier analysis, or wavelet decomposition, but Principal Component Analysis is the most often utilized. PCA is mostly a descriptive approach that does not attempt to foresee future data. In feature space, that will be curves or surfaces rather than directions. The purpose is to keep as much 'variance' as feasible while reducing data dimensionality. While it is widely used and constantly reinvented, it is primarily a statistical method applied in a range of sectors, with statisticians playing a key role in its development.PCA's task is to identify the best position for reducing information variance and vector dimensional characteristics. The PCA is an unsupervised learning method for reducing data dimensionality(Hassan et al., 2018). Karl Pearson devised PCA, a dimensionality reduction method, in 1901(Drennan, 2009). PCA has the following advantages: (I) it may be used to eliminate feature duplication in a data collection. (ii) Important information is collected about the high contrast that provides the best resolution. (iii)It allows for improved data display. (iv) It decreases complexity and boosts computing speed. (v)It's a robust tool that can be used to analyze datasets with (vi) multi-collinearity, (vii) values that are missing, (viii) data that is categorical, and (ix) inaccurate measurements. The objective is to summary key information from the data and show. it as a set of summary index known as principle components. One of the most usage of principal component analysis is speed up machine learning algorithms and reduction of dimentionality(Gajjar et al., 2018; Ning & You, 2018).For standard data, a PCA is frequently referred to as the PCA correlation matrix. The eigenvectors of this matrix describe linear combinations of the uncorrelated maximum-variance standard variables(Salih Hasan & Abdulazeez, 2021). The number of variables employed in the analysis is exactly proportional to the number of variables in the correlation matrix, therefore every correlation matrix PCA is the proportion of total variance divided by the number of variables in the PCA. PCAs are the best choice for datasets with a variety of scale changes for each variable and are invariant to linear changes in measurement units(Fujiwara et al., 2022).

The first dimension was formed by the combination of the following ASDs. E, S, V, and L with the minor contributions of descriptor B. The second dimension was majorly formed by descriptor A, B and the minor contributions from S descriptor (Fig -b).This analysis supported our hypothesis that the development of two-parameter models is possible without the loss of much information to estimate Kcw values for all data sets considered in this study but lead to the next important question: what can be the two new parameters that would be able to explain the same information as coded in the first two dimensions of PCA. The selection criterion for new variables was based on these considerations: the parameters should (i) be accessible (ii) have a simple chemical interpretation (iii) have a larger database and can be estimated computationally or determined experimentally by simple and inexpensive methods (iv) be able to incorporate all the physical interactions (intermolecular forces) as ASDs and the free energy changes during the transfer of the solute molecule from a gas phase to condensed phase (as in rate-limiting step). (v) be able to explain the mechanism of the process to a considerable extent (physically as well as thermodynamically). The partition coefficient K_{aw} (air to water) and K_{ow} (octanol to water) qualified for the selection criteria. To prove the suitability of K_{ow} and K_{aw} by analyzing the information variability in the first two dimensions, PCA was run on the Kcw data, comprising E, S, A, B, V, $\log K_{ow}$, $\log K_{aw}$ descriptors. It was found that the first two dimensions of each data set contain maximum information by incorporating $\log K_{aw}$, $\log K_{ow}$, and logKcw almost entirely in the first two dimensions along with contributions of ASDs.



	LogKaw	A	B	E	S	V	logKcw	LogKow
LogKow								1
logKcw							1	0.94
V						1	0.72	0.71
S					1	0.68	0.5	0.35
E				1	0.79	0.68	0.79	0.65
B			1	0.11	0.43	0.44	-0.22	-0.3
A		1	0.42	-0.07	0.12	-0.1	-0.34	-0.43
LogKaw	1	-0.69	-0.82	-0.32	-0.64	-0.39	0.11	0.23

The upper panels represent the results of PCA ran on a 117×5 matrix, [E, S, A, B &V] of ASM descriptors used to estimate Kcw values. (a) shows the eigenvalues scree plot which illustrates how much variability is contained by each major component. (b) The correlation circle shows the quality of representation and relationship among the variables in the first two dimensions. The angle between the arrows shows the correlation between the ASDs. The quality of representation of a parameter is proportional to the length of arrow lines. (c) depicts the correlation matrix's correlogram, as determined using Pearson correlation analysis. The red color is the representation of positive correlation while the blue color represents the negative correlation between each pair of variables. The values of r (correlation coefficient) show the magnitude of correlation. (d) shows the distribution of quality of representation into 8 dimensions obtained by PCA of 117×8 matrix [E, S, A, B, V, logKcw, logKow, and logKaw. The size of the circle and color intensity is proportional to the quality of representation of a parameter.

4.2.3 Cross Validation of two- parameter base model

The resampling technique which is used for the robustness and assessment of the model to avoid overfitting is called cross-validation(Berrar, 2018). Below methods, we used for cross-validation of the two-parameter model.

Table 9. cross-validation of the two-parameter model.

property	External Validation (Hold-out Approach)			
LogKcw	No of observations	RMSE	MAE	R^2
Train set (80%)	96	0.73	0.54	0.88
Test set (20%)	21	0.57	0.42	0.93

property	K-fold cross-validation			Leave one out approach			Bootstrap Approach		
LogKcw	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
	0.70	0.54	0.89	0.74	0.54	0.88	0.73	0.54	0.89

4.2.4 comparison of ASMs and the two-parameter base model

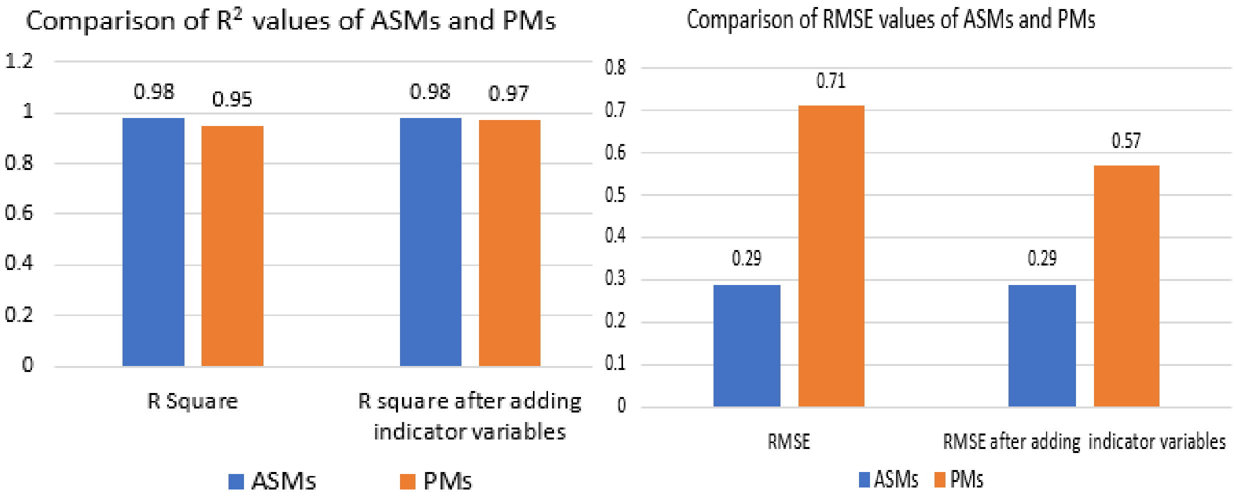


Figure-4.14 comparison of ASMs and the two-parameter base model

The lift chart shows the comparison of the R^2 value of ASMs and two-parameter model (2P-M) in which R^2 of ASMs is 0.98 and R^2 of 2P-M is 0.95 by adding indicator variables R^2 value of 2P-M is increased to 0.974. The right chart shows the comparison of RMSE of ASMs and 2P-M. RMSE of ASM is 0.29 and 2P-M RMSE is 0.71.by adding indicator variables the value of RMSE of 2P-M is decreased to 0.57.

4.2.3 Mechanistic Understanding of Kow and Kaw

1-Octanol-water partition coefficient (Kow)

For aqueous system octanol–water partition coefficient (Kow) is mostly applied partition coefficient and it also has its major role in air and other media like soil and others. In air system we have many organic compounds i.e. carbon-carbon or carbon-hydrogen bond molecules and as we know like solvents dissolve like solutes so, the solubility can be differentiated as these under normal conditions of pressure and temperature can easily get dissolved into organic solvents. In such case we term it as hydrophilic or lipophilic as organic compounds can easily 11 dissolve into water or lipids and if they do not readily dissolve into organic solvent under normal conditions it is termed as hydrophobic or lipophobic. Kow is defined as ratio of concentration of a substance in octanol to the concentration of substance in aqueous phase at equilibrium. In general octanol is a surrogate for solvents that are lipophilic and this is because it has affinity for water as well as organic compounds which is termed as amphibilic. It is also an indicator which is of importance in environmental partitioning (Vallero, 2014).

2-Air Water Partition Coefficient (Kaw)

In environment the transport of organic compounds is affected by the transfer between the atmosphere and aqueous systems (Schwarzenbach et al., 2004). Henry's law constant KH or Kaw is the air-water partition ratio for neutral compounds that are present in pure water at dilute solution concentrations. But as we know we don't have pure water but aqueous solutions which contain many chemicals, for this "air-water distribution ratio" is used, which is determined by approximating Henry's Law constant. Kaw is a unit less defined as the ratio of substance 12 abundance in air phase to that of aqueous or water phase at equilibrium. The compound's transfer depends upon the value of Kaw; it will get into air phase from water or aqueous phase if Kaw value is large (Ji et al., 2008).

Chapter 5

CONCLUSION AND RECOMMENDATION

5.1Conclusion

Models developed in this study are theoretically-rigorous and statistically robust respecting the principle of parsimony. The dimensionality analysis provides the statistical support for the justification in reducing 5p-ASM to 2p-LFER as it can be observed that it covers most of the information and supports the fact that only few information is lost. 2p-LFER models for plan cuticle to water partitioning R2 of 0.98 and 0.95were successfully developed in this study. Taken all the above together, the advantage of our model is that the values of logKcw can be estimated for the chemicals for which ASDs values are not available. The PMs based on log Kow and logKaw can be used to predict the Kcw values for different plant tissues types either by Eq. 17 in which the selection of indicator variables will allow the to estimate the Kcw values for required tissue type or the user can also use the equation devised for the general plant without specific plant tissue show in Eq.16 and the data set for this model analysis are mention in the table-8. Conducted two-parameter model (2P-M) has scientific and theoretical justification with the best predictive efficiency and based on Kow and Kaw is as good as parameter-intensive ASMs result showed in (a, b, c, and d) figures. Also, the dataset of descriptors in ASMs has limited just to 8000 chemicals but for the new model, the range of descriptors are about 60000chemicals. This proposed model is more easily accessible than the complicated multi-parameter model and 2P-M can give mechanistic insight into the Kcw process. Furthermore, is a good alternative to plant experimentation and testing to determine Kcw for the plant. But the conducted model cannot predict ionic species or complex mixtures and it can predict only within a specific range.

5.2 Recommendations

Extra descriptors can be used to include ionic species and the predicted values of PM can be integrated into the Estimation Program Interface (EPI) SuiteTM .

REFERENCES

Abdulqader, D. M., Mohsin Abdulazeez, A., & Zeebaree, D. Q. (2020). *Machine Learning Supervised Algorithms of Gene Selection: A Review*. *62*(03).

Abraham, M. H. (1992). *HMp7. 096*(5).

Abraham, M. H., Chadha, H. S., Martins, F., Mitchell, R. C., Bradbury, M. W., & Gratton, J. A. (1999). Hydrogen bonding part 46: A review of the correlation and prediction of transport properties by an LFER method: Physicochemical properties, brain penetration and skin permeability. *Pesticide Science*, *55*(1), 78–88. https://doi.org/10.1002/(SICI)1096-9063(199901)55:1<78::AID-PS853>3.0.CO;2-7

Abraham, M. H., Grellier, P. L., Hamerton, I., McGill, R. A., Prior, D. v., & Whiting, G. S. (1988). Solvation of gaseous non-electrolytes. *Faraday Discussions of the Chemical Society*, *85*(i), 107–115. https://doi.org/10.1039/DC9888500107

Abraham, W. R., Strömpl, C., Bennasar, A., Vancanneyt, M., Snauwaert, C., Swings, J., Smit, J., & Moore, E. R. B. (2002). Phylogeny of Maricaulis Abraham et al. 1999 and proposal of Maricaulis virginensis sp. nov., M. parjimensis sp. nov., M. washingtonensis sp. nov. and M. salignorans sp.nov. *International Journal of Systematic and Evolutionary Microbiology*, *52*(6), 2191–2201. https://doi.org/10.1099/ijs.0.02248-0

Berrar, D. (2018). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3, pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bower, K. (2018). On The Use of Indicator Variables in Regression Analysis. *Minitab.Com*, *October*, 2–5. http://www.minitab.com/en-US/uploadedFiles/Shared_Resources/Documents/Articles/indicator_variables_in_regression_analysis.pdf

Bräuer, P., Neinhuis, C., & Voigt, D. (2017). Attachment of honeybees and greenbottle flies to petal surfaces. *Arthropod-Plant Interactions*, *11*(2), 171–192. https://doi.org/10.1007/s11829-016-9478-0

Card, M. L., Gomez-Alvarez, V., Lee, W. H., Lynch, D. G., Orentas, N. S., Lee, M. T., Wong, E. M., & Boethling, R. S. (2017). History of EPI Suite™ and future perspectives on chemical property estimation in US Toxic Substances Control Act new chemical risk assessments. *Environmental Science. Processes & Impacts*, *19*(3), 203–212. https://doi.org/10.1039/c7em00064b

Copy, I. B., Network, C. N., & Hevc, C. (2019). *MULTIPLE LINEAR REGRESSION FOR HIGH EFFICIENCY VIDEO INTRA CODING Zhaobin Zhang University of Missouri Kansas City University of Science and Technology of China Tencent America*. 1832–1836.

Drennan, R. D. (2009). Principal Components Analysis. *Interdisciplinary Contributions to Archaeology*, 299–307. https://doi.org/10.1007/978-1-4419-0413-3_24

Eddula, S., Xu, A., Jiang, C., Huang, J., Tirumala, P., Liu, G., Acree, W. E., & Abraham, M. H. (2021a). Abraham solvation parameter model: updated correlations for describing solute partitioning into plant cuticles from water and from air. *Physics and Chemistry of Liquids*, *59*(5), 716–732. https://doi.org/10.1080/00319104.2020.1808659

Eddula, S., Xu, A., Jiang, C., Huang, J., Tirumala, P., Liu, G., Acree, W. E., & Abraham, M. H. (2021b). Abraham solvation parameter model: updated correlations for describing solute partitioning into plant cuticles from water and from air. *Physics and Chemistry of Liquids*, *59*(5), 716–732. https://doi.org/10.1080/00319104.2020.1808659

Eddula, S., Xu, A., Jiang, C., Huang, J., Tirumala, P., Liu, G., Acree, W. E., & Abraham, M. H. (2021c). Abraham solvation parameter model: updated correlations for describing solute partitioning into plant cuticles from water and from air. *Physics and Chemistry of Liquids*, *59*(5), 716–732. https://doi.org/10.1080/00319104.2020.1808659

Fernández, V., Bahamonde, H. A., Peguero-Pina, J. J., Gil-Pelegrín, E., Sancho-Knapik, D., Gil, L., Goldbach, H. E., & Eichert, T. (2017a). Physico-chemical properties of plant cuticles and their functional and ecological significance. *Journal of Experimental Botany*, *68*(19), 5293–5306. https://doi.org/10.1093/jxb/erx302

Fernández, V., Bahamonde, H. A., Peguero-Pina, J. J., Gil-Pelegrín, E., Sancho-Knapik, D., Gil, L., Goldbach, H. E., & Eichert, T. (2017b). Physico-chemical properties of plant cuticles and their functional and ecological significance. *Journal of Experimental Botany*, *68*(19), 5293–5306. https://doi.org/10.1093/jxb/erx302

Fujiwara, T., Kwon, O., & Ma, K. (2022). *Supporting Analysis of Dimensionality Reduction Results with Contrastive Learning*. *26*(1), 45–55.

Gajjar, S., Kulahci, M., & Palazoglu, A. (2018). Real-time fault detection and diagnosis using sparse principal component analysis. *Journal of Process Control*, *67*, 112–128. https://doi.org/10.1016/j.jprocont.2017.03.005

Goss, K. U., & Schwarzenbach, R. P. (2003). Rules of thumb for assessing equilibrium partitioning of organic compounds: Successes and pitfalls. *Journal of Chemical Education*, *80*(4), 450–455. https://doi.org/10.1021/ed080p450

Gupta, S., & Mallick, S. (2018). Modelling the water–plant cuticular polymer matrix membrane partitioning of diverse chemicals in multiple plant species using the support vector machine-based QSAR approach. *SAR and QSAR in Environmental Research*, *29*(3), 171–186. https://doi.org/10.1080/1062936X.2017.1419985

Hassan, O. M. S., Abdulazeez, A. M., & T?ryak?, V. M. (2018). Gait-Based Human Gender Classification Using Lifting 5/3 Wavelet and Principal Component Analysis. *ICOASE 2018 - International Conference on Advanced Science and Engineering*, 173–178. https://doi.org/10.1109/ICOASE.2018.8548909

Khawar, M. I., & Nabi, D. (2021a). Relook on the Linear Free Energy Relationships Describing the Partitioning Behavior of Diverse Chemicals for Polyethylene Water Passive Samplers. *ACS Omega*, *6*(8), 5221–5232. https://doi.org/10.1021/acsomega.0c05179

Khawar, M. I., & Nabi, D. (2021b). Relook on the Linear Free Energy Relationships Describing the Partitioning Behavior of Diverse Chemicals for Polyethylene Water Passive Samplers. *ACS Omega*, *6*(8), 5221–5232. https://doi.org/10.1021/acsomega.0c05179

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A Meeting of SIGDAT, a Special Interest Group of the ACL Held in Conjunction with ACL 2004*, 388–395.

Matschi, S., Vasquez, M. F., Bourgault, R., Steinbach, P., Chamness, J., Kaczmar, N., Gore, M. A., Molina, I., & Smith, L. G. (2020). Structure-function analysis of the maize bulliform cell cuticle and its potential role in dehydration and leaf rolling. *Plant Direct*, *4*(10), 1–21. https://doi.org/10.1002/pld3.282

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/jastt1457

Mutelet, F. (2012). The Use of Solvation Models in Gas Chromatography. *Chromatography - The Most Versatile Method of Chemical Analysis*. https://doi.org/10.5772/48381

Najat, N., & Abdulazeez, A. M. (2018). Gene clustering with partition around mediods algorithm based on weighted and normalized mahalanobis distance. *ICIIBMS 2017 - 2nd International Conference on Intelligent Informatics and Biomedical Sciences*, *2018-Janua*, 140–145. https://doi.org/10.1109/ICIIBMS.2017.8279707

Ning, C., & You, F. (2018). Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods. *Computers and Chemical Engineering*, *112*, 190–210. https://doi.org/10.1016/j.compchemeng.2018.02.007

Platts, J. A., & Abraham, M. H. (2000). Partition of volatile organic compounds from air and from water into plant cuticular matrix: An LFER analysis. *Environmental Science and Technology*, *34*(2), 318–323. https://doi.org/10.1021/es9906195

Poole, C. F., Ariyasena, T. C., & Lenca, N. (2013a). Estimation of the environmental properties of compounds from chromatographic measurements and the solvation parameter model. In *Journal of Chromatography A* (Vol. 1317, pp. 85–104). https://doi.org/10.1016/j.chroma.2013.05.045

Poole, C. F., Ariyasena, T. C., & Lenca, N. (2013b). Estimation of the environmental properties of compounds from chromatographic measurements and the solvation parameter model. In *Journal of Chromatography A* (Vol. 1317, pp. 85–104). https://doi.org/10.1016/j.chroma.2013.05.045

Qi, X., Li, X., Yao, H., Huang, Y., Cai, X., Chen, J., & Zhu, H. (2020a). Predicting plant cuticle-water partition coefficients for organic pollutants using pp-LFER model. *Science of the Total Environment*, *725*. https://doi.org/10.1016/j.scitotenv.2020.138455

Qi, X., Li, X., Yao, H., Huang, Y., Cai, X., Chen, J., & Zhu, H. (2020b). Predicting plant cuticle-water partition coefficients for organic pollutants using pp-LFER model. *Science of the Total Environment*, *725*. https://doi.org/10.1016/j.scitotenv.2020.138455

Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., & Scientist, C. D. (2020). Encyclopedia of Database Systems. *Encyclopedia of Database Systems*. https://doi.org/10.1007/978-1-4899-7993-3

Rojewski, J. W., Lee, I. H., & Gemici, S. (2012). Use of *t*-test and ANOVA in Career-Technical Education Research. *Career and Technical Education Research*, *37*(3), 263–275. https://doi.org/10.5328/cter37.3.263

Salih Hasan, B. M., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, *02*(01), 20–30. https://doi.org/10.30880/jscdm.2021.02.01.003

Singh, R., & Balange, A. (2017). PDFlib PLOP : PDF Linearization , Optimization , Protection Page inserted by evaluation version American Diet. *Journal of Food Processing and Preservation*, *January 2004*, 1919–1924.

Skrzydeł, J., Borowska-wykręt, D., & Kwiatkowska, D. (2021). Structure, assembly and function of cuticle from mechanical perspective with special focus on perianth. *International Journal of Molecular Sciences*, *22*(8). https://doi.org/10.3390/ijms22084160

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *International Conference on Information and Knowledge Management, Proceedings*, 623–632. https://doi.org/10.1145/1321440.1321528

Starkweather, Dr. J. (1988). Homogeneity of Variances Research and Statistical Support consultant This month we touch on a fundamental issue in statistical evaluation that often gets overlooked. Testing assumptions for parametric analysis. *Psychological Bulletin, 104(3), 396 – 404.*, *103*(3), 396–404.

US EPA. (2016). *Estimation Programs Interface SuiteTM for Microsoft Windows, v 4.00 (KowWIN, ver. 1.68)*.

Yeats, T. H., & Rose, J. K. C. (2013a). The formation and function of plant cuticles. In *Plant Physiology* (Vol. 163, Issue 1, pp. 5–20). American Society of Plant Biologists. https://doi.org/10.1104/pp.113.222737

Yeats, T. H., & Rose, J. K. C. (2013b). The formation and function of plant cuticles. *Plant Physiology*, *163*(1), 5–20. https://doi.org/10.1104/pp.113.222737

Zou, Y., Yin, H., Tan, Q., Chen, Y., Lv, G., & Hou, X. (2011). Polycyclic aromatic hydrocarbons (PAHs) pollution recorded in annual rings of gingko (Gingko biloba L.): Regression analysis and comparison to other pollutants. *Microchemical Journal*, *98*(2), 303–306. https://doi.org/10.1016/j.microc.2011.02.015