Analysis of the environmental parameters for risk assessment of pesticides by machine learning approach

筑波大学審査学位論文（博士）

２０２１

小林　由幸

筑波大学大学院
ビジネス科学研究科　企業科学専攻

Analysis of the environmental parameters for risk assessment of pesticides by machine learning approach

Doctoral thesis

September 2021

Yoshiyuki Kobayashi

Doctoral Program in Systems Management
Graduate School of Business Sciences
University of Tsukuba

Abstract

Pesticides are artificially synthesized chemical compounds with high biological activity that are widely used in agriculture for controlling pests and insects. Some pharmaceuticals and pesticides have common active ingredients. However, unlike pharmaceuticals, pesticides are intentionally released into the environment for crop control. As a result, they may remain in agricultural products and then be unknowingly ingested via food over a long period. Conducting safety tests for the registration of pesticides is usually expensive and takes up to several years. Depending on the obtained test results, a pesticide may or may not be approved for registration or its registration may be cancelled altogether.

Thus, it is necessary to predict the main safety parameters at the early stage of pesticide development. In previous studies, some prediction models for such parameters have been developed. However, calculated properties were used as explanatory variables for these models because it was difficult to determine them experimentally. In this study, we have developed and modified a prediction model for soil adsorption coefficient (K_{oc}) and bioconcentration factor (BCF), which are two of the most important parameters used in risk and hazard assessments to determine the registrability of pesticides in the European Union (EU). K_{oc} represents the distribution ratio of chemicals between the soil/sediment phase and the aqueous phase. BCF is a widely used hazard assessment criterium and represents the ratio of the chemical concentration in fish to its concentration in water. By using the experimental and calculated values of physicochemical properties that are rarely used as explanatory variables, we have managed to establish more accurate prediction models than the

conventional ones and express the relationship between selected parameters on one side and K_{oc} and BCF on the other side. By applying these models, a preliminary environmental risk assessment can be performed without conducting time-consuming experiments. Consequently, the proposed models may significantly contribute to the development of new chemical compounds, including pesticides. This work represents a new effective approach to evaluating the K_{oc} and BCF parameters as well as the applicability of the data mining method.

In Chapter 3, we developed prediction models for K_{oc} values. In previous studies, molecular and topological descriptors were mainly used as explanatory variables because their magnitudes could be easily calculated from chemical structures. In contrast, physicochemical properties are closely related to the K_{oc} value, but they have not been analyzed in sufficient detail. For this reason, we collected experimental data for different physicochemical properties listed in pesticide evaluation reports. In addition, we calculated the corresponding molecular descriptors using Cheminformatics software. By utilizing these physicochemical properties and molecular descriptors, a new prediction model based on the gradient boosting decision tree (GBDT) algorithm was developed. The obtained results revealed that the proposed high-performance model was more accurate than the previously reported models.

In Chapter 4, we propose prediction models for K_{oc} values based only on calculated parameters. In the previous chapter, physicochemical properties were obtained from a peer-reviewed report of the European Food Safety Agency and Chemistry Dashboard of the U.S. Environmental Protection Agency (EPA). However, collecting experimental data for a wide range of compounds requires a considerable amount of time and effort. Hence, the objective of this chapter was to develop an accurate predictive model by obtaining physicochemical properties using a relatively easy approach

and freely available software. In addition, we employed a larger dataset of K_{oc} values than that utilized in the previous chapter. The resultant model demonstrated much better prediction ability than those of the previously developed models. Although the model based on the experimental values of physicochemical properties exhibited a good fit, high prediction accuracy, and robustness, the approach proposed in this chapter is a good substitute for actual values if the latter are difficult to obtain.

In Chapter 5, we establish prediction models for the estimation of logarithmic BCF values. Similar to the previous chapter, we calculated physicochemical properties by OPERA software because the quantitative structure-property relationships model for K_{oc} developed in the previous chapter exhibited superior performance. Hence, the new model based on the GBDT algorithm and properties calculated by OPERA is more accurate than the existing BCF prediction model. The proposed method is applicable for the development of prediction models for various parameters used in environmental risk assessment.

In Chapter 6, we discuss the results of this work and their significance. In Chapter 7, we summarize the conclusions of this work and outline the directions of future research studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and purpose of this research

Pesticides are artificially synthesized chemical compounds with high biological activity that are widely used in agriculture for controlling pests and insects. Antibiotics such as streptomycin and oxytetracycline also contain active ingredients commonly present in pesticides and pharmaceuticals [Thiele-Bruhn, 2003]. However, unlike pharmaceuticals, pesticides are intentionally released into the environment for crop control, as shown in Figure 1.1. As a result, they may remain in agricultural products and then unknowingly be ingested as food over a long period. Furthermore, the incorrect use of pesticides may cause environmental pollution and produce a negative impact on the entire ecosystem [Damalas and Eleftherohorinos, 2011]. Therefore, to achieve a proper balance between risks and benefits, it is necessary to adopt special measures restricting or prohibiting the pesticide application when the risk is too high.

There are four main safety requirements related to pesticide utilization. First, it is necessary

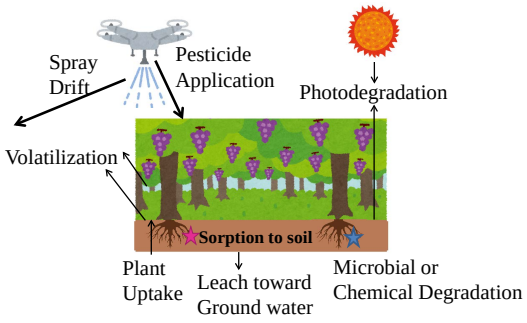


Figure 1.1: Environmental effects of pesticide applications

to ensure the safety of agricultural chemical users by assessing the potential for acute poisoning in a case of accidental human exposure to pesticides during application. Second, it is imperative to maintain the safety of crops by examining whether their growth or yield/quality are affected by pesticide spraying. Third, the safety of consumers must be ensured by evaluating the effects of both the short-term and long-term pesticide exposures on human health due to the continuous consumption of pesticide-treated crops. Finally, an environmental safety assessment is required.

During the registration of pesticides with proper authorities, the manufacturer must submit an application form including the results of a safety evaluation to an EU inspection country. A draft evaluation report, which is called a draft assessment report (DAR) (`https://www.efsa.europa.eu/en/publications`), is issued after a certain period. Subsequently, the European Food Safety Agency (EFSA) publishes the DAR and invites the public to comment on this document. After that, EFSA issues the final evaluation report, and the European Commission votes on pesticide registration. The evaluation reports issued after July 29, 2005 are published on the Web, and each report includes the results, considerations, and evaluations of various safety tests used in the application. Meanwhile, OPERA, a software package for calculating the physicochemical properties of chemical compounds, has been released by the U.S. Environmental Protection agency (EPA)

(https://github.com/kmansouri/OPERA). The U.S. EPA is an administrative agency of the U.S. federal government created to protect the public health and preserve the natural environment. Its main objectives include controlling air pollution, water pollution, and soil pollution.

Thus, by analyzing the DAR and physicochemical properties calculated by OPERA, it is possible to develop a predictive model for environmental risk / hazard assessment parameters. Note that these parameters can be predicted before the initiation of long-term studies and expensive trials for pesticide registration, which affect R&D decisions.

In this work, we have developed prediction models for risk assessment parameters based on machine learning by using the physicochemical properties provided in the evaluation reports for various pesticides registered in Europe and calculated by OPERA. In previous studies, linear relationships between some explanatory variables and the objective variable have been established. However, because these parameters depend on various factors, it is difficult to adequately describe them by a simple linear model [Lombardo et al., 2010]. Therefore, machine learning was used in the present study to develop a highly accurate prediction model utilizing various explanatory variables.

In particular, we proposed prediction models for the soil adsorption coefficient (K_{oc}) and bio-concentration factor (BCF), which represented the key parameters of environmental risk / hazard assessments [Chi et al., 2018].

1.2 Structure of this research study

The structure of this paper can be described as follows. In Chapter 2, previous research works on the prediction models of safety test parameters used in the environmental risk / hazard assessment of pesticides are summarized. Such studies mainly focused on evaluating quantitative structure-activity / property relationships via computational methods.

In Chapter 3, we outline a prediction model for K_{oc}, which is one of the main parameters of the environmental risk assessment procedure. K_{oc} is closely related to the physicochemical properties of compounds; however, their values can be obtained only by conducting scientific experiments. Therefore, a linear prediction model using the value of molecular descriptors calculated from a structural formula has been developed. In this chapter, we discuss the pesticide evaluation reports available on the EFSA website and experimentally determined physicochemical properties, K_{oc}. Note that EFSA is the regulatory authority in the European Union, providing scientific information on food safety through the risk assessment conducted by experts. We also calculated molecular descriptors using Cheminformatics software. The results of the developed prediction model were compared with the parameters reported in selected papers. Owing to the use of both the molecular descriptors calculated from structural formulas and experimentally determined physicochemical properties from the literature and open databases, the prediction accuracy of the developed model was considerably higher than those of the previously reported models. In addition, a prediction model based on the obtained experimental data was developed in Chapter 3. Note that collecting experimental data for a wide range of compounds requires a considerable amount of time and efforts.

The objective of Chapter 4 was to establish a predictive model with improved accuracy by analyz-

ing the physicochemical properties of compounds using a simplified approach and freely available software. For this purpose, the U.S. EPA has developed the OPEn structure-activity Relationship App (OPERA) software [Mansouri et al., 2018] that utilizes the data obtained for a wide range of chemical compounds listed in the corresponding database (PHYSPROP)[Howard and Meylan, 2000]. This software can predict various physicochemical properties and environmental effects of pollutants with high reliability. By applying a prediction model based on machine learning for K_{oc} using the physicochemical properties and environmental parameters calculated by OPERA instead of the experimental data, a highly accurate model may be rapidly developed. Hence, in addition to the good fit, high prediction accuracy, and robustness demonstrated in the previous chapter, the theoretical approach utilized in this chapter can serve as a substitute for experimental values if the latter are difficult to obtain.

In Chapter 5, we combine the data and algorithms proposed in the previous chapters to construct a prediction model for the bioconcentration factor (BCF), which is one of the multiple indicators used for risk and hazard assessments during pesticide registration.

In Chapter 6, we discuss the results of this work and their significance. In Chapter 7, we draw conclusions from the obtained data and outline the main directions of future research works.

Chapter 2

Related works

2.1 Environmental risk/hazard assessment of pesticides

The main difference between pesticides and pharmaceuticals is that the former require an environmental risk assessment. In particular, it is necessary to evaluate environmental parts containing pesticides such as soil and rivers, their effects on individual plants, the human influence on water, and pesticide degradability. For this reason, authorities in each country have established pesticide registration systems based on laws, regulations, and various guidelines, and mandate conducting multiple safety tests.

The pesticide registration system currently existing in Europe and the United States serves as an international standard. In addition, the European pesticide registration system is one of the strictest systems in the entire world. After the EU establishment in 1993, the 91/414/EEC Directive was issued, and the European pesticide registration system was significantly changed [Directive, 1991]. In particular, the required number of data points was increased, and the evaluation period was

extended. Furthermore, the evaluation criteria specified in this directive were applied not only to new pesticides, but also to the pesticides previously registered in European countries.

During evaluation, a comprehensive risk assessment was performed for each active ingredient of pesticides. Furthermore, the European Commission advocated the "precautionary principle" in 2000. In the "EU Thematic Strategy for Pesticides" published in July 2006 [EuropeanComission, 2006], the main goal was to minimize risks to the human health and environment caused by the use of pesticides. After three years of discussion, "Regulation (EC) No. 1107/2009" was adopted [EuropeanComission, 2009]. Owing to the recent increase in food safety and security requirements, the new European Food Safety Agency (EFSA) was established in 2002 as an independent entity from the European Commission to assess the safety of agricultural chemicals in the EU. After that, the data requirements related to pesticide registration have been reviewed regularly, and the number of data points requested at the time of application increased in January 2014.

Because the persons conducting safety tests must possess advanced skills in the fields of toxicology and environmental chemistry, the number of research institutions that can perform these tests is limited. Consequently, the costs of such tests remain relatively high. In addition, some long-term tests take more than two years to complete. Traditionally, the development of pesticides requires approximately 10 years and costs a significant amount of money. Furthermore, as mentioned above, due to the increase in the number of safety tests and their complexity, the pesticide development period and related costs also increased. According to the survey conducted by CropLife International, which is an international pesticide industrial organization, the cost of developing pesticides was approximately $286 million per product in 2015, nearly double that in 1995 [Nishimoto, 2019, McDougall, 2016]. Moreover, the pesticide development period has been

extended to an average of 11 years, while the probability of a candidate chemical compound to be registered as a pesticide is one in tens of thousands. Pesticide re-evaluation is performed every 10 years after the initial registration, requiring the submission of additional data in accordance with the latest standards [EuropeanComission, 2014]. As estimated by the European Crop Control Association (ECPA), the cost of reassessment per agent is approximately 6 million euro. At the start of the re-evaluation process, 954 pesticides were registered in the EU. However, an over half of these pesticides were subsequently deleted due to the tightening of regulations and increase in the related costs [Yokota, 2014].

In particular, the majority of pesticides were removed because of the strict environmental risk/hazard assessment and non-approval of the registration account. When a pesticide is applied to soil, it remains near the soil surface and is decomposed by soil microorganisms. However, its residual behavior in soil significantly depends on the pesticide and soil types. In Europe, there are many areas where groundwater is mainly used as drinking water. Therefore, strict environmental risk assessments are performed in Europe assuming that the active ingredients and their metabolites flow through the soil into rivers and groundwater [EFSA, 2018, EPRS, 2018, Schäfer et al., 2019]. In particular, PECgw, a very strict standard for groundwater pollution, specifies a limits of 0.1 ppb or less for active ingredients and 0.75 ppb or less for metabolites [EuropeanComission, 2003].

Note that a global risk assessment procedure was initially established in the EU; however, in the developed countries, the concept of cut-off criteria based on a hazard type was adopted for the first time based on the precautionary principle. Hazard assessment is a procedure that is restricted to evaluating potential toxicity and hazards of compounds [Klopffer, 1994, Henschel et al., 1997]. Unlike risk assessment, these criteria do not take into account the amounts of pesticides exposed

to living organisms and the environment. Only the compound composition determines whether its approval is possible. Various cut-off hazard-based criteria prohibiting active ingredients that are carcinogenic, genotoxic, reproductive, or pose significant environmental risks are listed in Table 2.1. Furthermore, "Persistent Organic Pollutants (POPs)", "Persistent, Bioaccumulative, and Toxic (PBT) Substances", and "Highly Persistent / Highly Bioaccumulative substances (vPvB)" have been established. Thus, due to the rigorous risk and hazard assessment of pesticide application data in the EU, a pesticide registration may be rejected during a new application or the re-evaluation process.

Because advanced skills are required to experimentally determine risk assessment parameters, and very strict guidelines such as those developed by the Organization for Economic Cooperation and Development (OECD), Directorate General for Health and Food Safety (DG SANTE), and Office of Chemical Safety and Pollution Prevention（OCSSP) must be followed during this procedure, a limited number of research institutions can conduct safety tests. Moreover, the experimental determination of such parameters for individual compounds is an expensive process that may take several months or even a half year to complete. Since the utilized compound classification and safety criteria are regulated internationally (including the European countries, the United States, and Japan), pesticide registration may be limited or prohibited altogether if these criteria are not satisfied [EuropeanComission, 2009, Markell, 2010]. Therefore, the establishment of an accurate prediction model for estimating the risk and hazard assessment parameters of safety tests at the early stage of pesticide development should facilitate their practical implementation.

Table 2.1: Perticide cut-off criteria [Moermond et al., 2012, Matthies et al., 2016]

Environmental		
POP	Persistence	DT50 (Water) >2 months or
		DT50 (Soil) >6 months or
		DT50 (Sediment) >6 months
	Bioaccumulation	BCF >5,000 or logPow >5
	Long Range Transport	The criteria utilized here include the measured distance from the source, monitoring data and environmental properties (e.g. DT50 (AIR) >2 days
PBT	Persistence	DT50 (Marine Water) >60 days or
		DT50 (Fresh/Estuarine Water) >40 days or
		DT50 (Marine Sediment) >180 days or
		DT50 (Fresh/Estuarine Sediment) >120 days or
		DT50 (soil) >180 days
	Bioaccumulation	BCF >2000
	Ecotoxicology and Toxicity	NOEC (Marine and Fresh Water Organisms) <0.01 mg/l or
		Classified as Categories 1A or 1B for mutagenic and carcinogenic (for reproduction, see also Category 2) or
		Classified STOT RE 1 or STOT RE 2.
vPvB	Persistence	DT50 (Marine/Fresh/Estuarine Water) >60 days or
		DT50 (Marine/Fresh/Estuarine Sediment) >180 days or
		DT50 (soil) >180 days
	Very bioaccumulative	BCF >5000
Endocrine Disrupter	ED potential for environmental organisms	
Ground water pollution	PECgw >0.1 ppb (Parent), >0.75 ppb (Metabolite)	
Toxicology		
Carcinogenicity Mutagenicity Reproductive toxicity		When classified into the 1A and 1B toxicity categories Cat. 1A: Known toxicants Cat. 1B: Presumed human toxicants Cat 2: Suspected human toxicants

2.2 Quantitative Structure-Activity Relationship

Quantitative structure-activity relationship (QSAR) is a method for predicting the toxicity of chemical compounds based on statistical parameters [Hansch, 1993]. In QSAR, the numerical value of each element such as chemical structure and chemical properties is called a descriptor. Typical descriptors include fingerprints indicating the presence or absence of a specific chemical structure and measured / estimated values of physicochemical properties (such as molecular dipole moment, charge, and energy) of a chemical compound.

In the 1960s, Hansch and Fujita reported the existence of a correlation between structure and activity, which was subsequently called the Hansch-Fujita method [Hansch and Fujita, 1964]. It utilized linear and non-linear multiple regression analysis procedures as well as the physicochemical and thermodynamic parameters derived from a compound structure. Kowalski applied a K-nearest neighbor method (k-NN) based on pattern recognition to anticancer drugs and successfully classified them into active and inactive groups [Kowalski and Bender, 1974]. Various QSAR toxicity models have been also developed in the past couple of decades [Kubinyi, 1997, Tropsha, 2010, Halder et al., 2018].

2.2.1 QSARs established for pharmaceuticals

QSARs have been mainly used in the pharmaceutical field for predicting the safety, efficacy, and toxicity of drugs as well as the interaction between the target and a compound.

Cos et al. determined a QSAR for flavonoids used as the inhibitors of xanthine oxidase and scavengers of superoxide radicals produced by enzyme xanthine oxidase. The developed model was

able to identify gout treatment compounds more effectively than the widely used drug allopurinol, which acted as a xanthine oxidase inhibitor [Cos et al., 1998]. Rice-Evans et al. published a review on the relationship between the antioxidant activity as a free radical scavenger and chemical structure describing the biological properties of flavonoids and phenolic acids[Rice-Evans et al., 1996]. Zhao et al. evaluated the human intestinal absorption of 241 drugs by their bioavailability. They found that 1) the urinary excretion ratio of drug-related substances after oral administration, 2) cumulative urinary excretion ratio of drug-related substances after oral / intravenous administration, and 3) Abraham descriptor could accurately predict the human intestinal absorption [Zhao et al., 2001].

Similarly, QSARs are widely used in various areas of the pharmaceutical field, increasing the efficiency of a new drug development procedure.

2.2.2 QSARs established for pesticides

QSAR efficacy and safety models were also developed for pesticides [Coats, 1990, Sparks et al., 2001, Hamadache et al., 2016, Braeuning et al., 2018]. However, unlike pharmaceuticals, pesticides require an environmental impact assessment. For this reason, QSAR environmental toxicity models were established as well. For example, Toropova et al. developed a toxicity prediction model for Daphnia and rainbow trout [Toropov and Benfenati, 2006, Toropov et al., 2017]. Tremolada et al. reported a model predicting the acute toxicity of pesticides for fish and daphnia. The obtained results revealed that the correlation between fish (rainbow trout) and daphnia toxicity was statically significant [Tremolada et al., 2004]. Devillers developed a QSAR model based on neural network that took into account the fish weight, exposure time, temperature, pH value, and water hardness to predict the acute toxicity of pesticides for bluegill [Devillers, 2001]. Furthermore, Bradbury

compiled a review paper focusing on a QSAR quantic toxicity model [Bradbury, 1995]. Hence, multiple QSAR environmental impact assessment models have been developed in the past. Moreover, Benfenati et al. published a book on pesticide regulatory purposes, which summarized the QSAR model outline, utilized algorithms, and validation methods [Benfenati, 2011].

2.2.3 QSAR BCF prediction models

Some linear BCF QSAR models were developed by Devillers et al. [Devillers et al., 1996], Papa et al. [Papa et al., 2007], and Garg and Smith [Garg and Smith, 2014]. Gissi et al. attempted to predict the BCF values of 851 compounds, as reported in the Alternative Non-Testing Methods Assessed for REACH Substances BCF dataset [Gissi et al., 2015]. The most widely used BCF prediction models, i.e., CAESAR and Meylan, were utilized to develop a more reliable integrated approach [Meylan et al., 1999, Lombardo et al., 2010]. Pramanik and Roy reported two BCF prediction models that included multiple linear regression algorithms and a partial least squares analysis procedure [Pramanik and Roy, 2014]. These models were based on a training set that included 324 compounds. They were applied to verify the performance of a testing set containing 198 compounds. Additionally, many other QSAR BCF prediction models have been developed for various compounds over the last 20 years [Gramatica and Papa, 2005, Pavan et al., 2008, Nolte and Ragas, 2017].

In general, the process of bioaccumulation is strongly influenced by the physicochemical properties of compounds such as molecular size, fat- and water-solubility, and biological characteristics of organisms including type and size [Veith et al., 1979, Connell, 1988, Garg et al., 2014]. However, such properties are typically determined experimentally. As a result, many QSAR models were established by calculating molecular descriptors such as PaDEL-Descriptor [Yap, 2011]

Table 2.2: Previous research studies on BCF prediction models

Previous research	Data				Algorithm	Remarks
	Compounds	Properties	m	Nt/Np		
[Meylan et al., 1999]	Diverse	BCF	2	694	MLR	R2 pred is close to 0.7 Using paid software for statistical analysis Not performing a train-test split
[Papa et al., 2007]	Diverse	BCF	4	290/315	GA-VSS	Using paid software for descriptor calculation
[Lombardo et al., 2010]	Diverse	BCF	1	327/81	LR	Some paid software is used Only one variable was used R2 is less than 0.6
[Pramanik and Roy, 2014]	Diverse	BCF	4	324/198	GFA-MLR	Using paid software for model development
[Garg and Smith, 2014]	Highly hydrophobic organic chemicals	BCF	3	24/5	MLR	Limited number of data sets
[Gissi et al., 2015]	Diverse	BCF	-	851	-	Using existing software

and DRAGON [Mauri et al., 2006] and using them as explanatory variables [Zhao et al., 2008, Pramanik and Roy, 2014, Toropova et al., 2020] determined by Cheminformatics software.

The most frequently employed models using physicochemical properties include the linear and nonlinear models based on the n-octanol/water partition coefficient ($\log P_{\mathrm{ow}}$) [Connell and Hawker, 1988, Bintein et al., 1993, ECHA, 2017]. The bioconcentration of organic compounds in fish mainly depends on their hydrophobicity [Devillers et al., 1996], and $\log P_{\mathrm{ow}}$ is closely related to BCF. Furthermore, $\log P_{\mathrm{ow}}$ is one of the essential physicochemical properties required for registering a chemical substance that can be determined by a simple experimental method [Klein et al., 1988]. Thus, $\log P_{\mathrm{ow}}$ was used in many QSAR models. Note that there are few reports of QSAR models utilizing other physicochemical properties [Isnard and Lambert, 1988, Pavan et al., 2008]. A summary of the previously developed QSAR BCF predition models is provided in Table 2.2.

2.3 Quantitative Structure-Property Relationship

Quantitative structure-property relationships (QSPRs) have been used as another method for predicting the physicochemical properties of chemical compounds based on statistical parame-

ters. Various QSPR models were established to improve the development efficiency of compounds, including pharmaceuticals. For instance, such models were employed to predict the octanol/water partition coefficients (log P_{ow}) [Haeberlein and Brinck, 1997, Zeng et al., 2012], water solubilities [Katritzky et al., 1998, Freire et al., 2010], vapor pressures [Katritzky et al., 1998, Gharagheizi et al., 2012], and melting points [Katritzky et al., 2002, Liang et al., 2013] of different substances.

The soil adsorption coefficient (K_{oc}) represents the distribution ratio of chemicals between the soil/sediment and aqueous phases. It is a critical parameter used in environmental risk assessments such as PECgw. A chemical with high K_{oc} tends to be strongly adsorbed by soil. Over the past 20–30 years, many QSPR models for predicting the K_{oc} values of various compounds have been developed [Gawlik et al., 1997, Gramatica, 2010, Nolte and Ragas, 2017]. They included a model for predicting the adsorption of drugs by the soil surface using an artificial neural network (ANN) [Berthod et al., 2017]. In general, these models demonstrated the importance of taking into account hydrophobicity, charge, and molecular shape when studying sorbate-sorbent interactions. Khan et al. developed a model for calculating K_{oc} values using multiple linear regression (MLR) and applied it to 344 environmental pollutants [Kahn et al., 2005]. The obtained results revealed that hydrophobicity, molecular size, molecular shape, and charge distribution strongly influenced the interactions between different pollutants. Goudarzi et al. developed a model for predicting pesticide properties using MLR, ANN, and the descriptors calculated by DRAGON software [Goudarzi et al., 2009]. Shao et al. proposed a support vector machine (SVM) model using various molecular descriptors, including the Moriguchi octanol-water partition coefficient (MLogP), atomic bond index, and three-dimensional (3D) structure of molecules, which were also calculated by DRAGON software

Table 2.3: Previous research studies on K_{oc} prediction models

Previous research	Data				Algorithm	Remarks
	Compounds	Properties	m	Nt/Np		
[Gramatica et al., 2000]	Pesticide	K_{oc}	6	143/20	MLR	R2 pred is less than 0.7
[Huuskonen, 2003]	Pesticide	K_{oc}	12	143/20	MRL	Many explanatory variables (12)
[Duchowicz et al., 2007]	Pesticide	K_{oc}	6	143/20	MLR	Using paid software for descriptor calculation
[dos Reis et al., 2014]	Pesticide	K_{oc}	4	143/20	MLR	Using paid software for model development
[Shao et al., 2014]	Diverse	K_{oc}	4	643/321	LS-SVM	Using paid software for descriptor calculation Using limited access software that is available to some people for model development
[Olguin et al., 2017]	Diverse	K_{oc}	1	643/321	LR	Using limited access software that is available to some people for model development Low performance compared other models due to the use of a single linear regression

[Shao et al., 2014].

Although most QSPR predicting models for K_{oc} values utilized small datasets for compounds with specific properties, Shao et al. employed an extensive dataset containing 964 compounds for their QSPR model. In all these studies, molecular and topology descriptors were mainly employed as explanatory variables. The advantage of using these descriptors is that once the structure of a chemical compound is identified, they can be easily calculated by Cheminformatics software. In previous works, many models were developed using molecular descriptors calculated by the PaDEL-Descriptor [Yap, 2011] or DRAGON [Mauri et al., 2006] software as explanatory variables. A summary of the QSPR models constructed for K_{oc} is provided in Table 2.3.

2.4 Outline of this research study

In this research study, we propose a new QSAR / QSPR model with higher accuracy and readability than those of previous developed models using the actual experimental data for safety studies and calculated physicochemical properties as explanatory variables. For this purpose, we obtained both the experimental and calculated physicochemical properties from the pesticide evaluation report

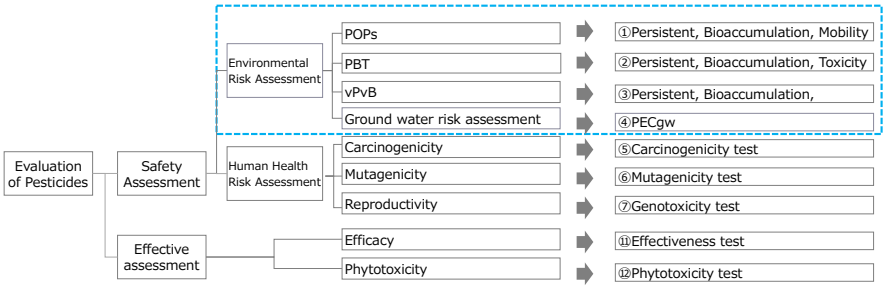


Figure 2.1: Conceptual diagram of pesticide evaluation and scope of this study

published by EFSA and utilized the Cheminformatics software developed by U.S. EPA.

Because pesticides are often deregistered after conducting an environmental risk assessment such as PECgw, we have developed a QSPR model for predicting K_{oc} values during the PECgw evaluation. In addition, we established a QSAR model for predicting BCF values, representing an evaluation standard for all hazardous compounds such as POP, PBT, and vPvB. The related conceptual diagram is shown in Figure 2.1.

Many previous studies conducted in the past used molecular descriptors calculated by PaDEL-Descriptor or DRAGON as explanatory variables; however, very few works utilized the experimentally obtained physicochemical properties. For instance, such physicochemical parameters as water solubility, octanol/water partition coefficient (log P_{ow}), and vapor pressure strongly affect the soil adsorption process. Furthermore, log P_{ow} is closely related to the K_{oc} prediction procedure [dos Reis et al., 2013]. However, because these values must be obtained experimentally, there are few reports on QSPR models using physicochemical properties. The prediction model described in the previous section uses the data calculated by the above-mentioned software, and it is not easy to explain the relationship between the explanatory variable and the objective variable. Other studies

utilized expensive software and paid databases. To select and optimize a large number of candidate compounds with a minimum investment at the initial development stage, the cost of a conducted study should be as small as possible.

Thus, in this work, physicochemical properties were objectively selected using free software and open data sources and utilized them for predicting safety parameters. In previous research studies, neural network-based and SVM models were used; however, it was difficult to investigate the prediction process involving objective variables in these models. It was also difficult to evaluate the importance and contribution of each explanatory variable to the objective variable. Therefore, we employed a simple algorithm with high readability and gradient boosting as a prediction method.

To evaluate the robustness of the proposed model, the latter was validated in accordance with the principles of the OECD [Gramatica, 2007]. By calculating the international validation index utilized in many studies,

To increase the readability of algorithms, the European General Data Protection Regulation (GDPR) also requires a brief explanation of the processing logic and conducting periodic reviews of the accuracy and validity of the decision making process [EPRS, 2020].

Accordingly, an algorithm clarifying the relationship between the explanatory variables and prediction results is desirable, and this paper proposes a prediction model for environmental risk assessment parameters using the open data processing software and physicochemical properties and verifies its performance. Finally, we summarized the issues encountered in previous research studies and the present work in Table 2.4.

Table 2.4: Research tasks from previous studies and responses in this study

Previous research	Contribution	Remarks	Topics addressed in this study
Huuskonen (2003)	Their model is based on the atom-type electrotopological state indices calculated only from chemical structures.	Many explanatory variables (12)	Using selected properties by machine learning algorithms
Duchowicz et al. (2007)	Their model is based on the molecular descriptors calculated only from chemical structures.	Using paid software for descriptor calculation	Using experimental data from the EU evaluation report Using calculated descriptors by the freely available software
Papa et al. (2007)	They used a large dataset which included 640 compounds and a wide range of log BCF values.	Using paid software for descriptor calculation	
Lombardo et al. (2010)	They provided freely available GUI software.	Some paid software was used for descriptor calculation	
		R^2_{PRED} is less than 0.6	
dos Reis et al. (2014)	They developed simple models using few explanatory variables.	Using paid software for model development	Using a free programming language for the prediction model development
Pramanik et al. (2014)	Their model is based on molecular descriptors calculated only from chemical structures.	Using paid software for model development	
Meylan et al. (1999)	They have used large dataset which include 694 compounds.	R^2_{PRED} is close to 0.7	
		Using paid software for statistical analysis	
		Not performing a train-test split	
Giss et al. (2015)	They compared the performances of the existing software packages	Using the existing software	
Shao et al. (2014)	They developed both linear and non-linear prediction models.	Using paid software for descriptor calculation	Using experimental data from the EU evaluation report Using descriptors calculated by freely available software Using a free programming language for the prediction model development
		Using limited access software that is available to some people for model development	
Olguin et al. (2017)	They adopted various external validation parameters.	Using limited access software that is available to some people for model development	
		Low performance compared to those of other models due to the use of a single linear regression	
Garg and Smith (2014)	They developed simple linear models.	Limited number of data sets	Using a large data set

Chapter 3

Prediction of soil adsorption coefficient using experimental physicochemical properties and molecular descriptors

3.1 Introduction

The soil adsorption coefficient (K_{oc}) of pesticides plays an important role in risk assessment calculations, based on which, the pesticides can be allowed for application [FOCUS, 2000]. K_{oc} represents the distribution ratio of chemicals between the soil/sediment phase and the aqueous phase. Higher K_{oc} values result in strong adsorption into soils [Jury, 1986]. Advanced experimental skills are required to measure the K_{oc}; thus, limited research organizations can perform the study. Additionally, the cost to obtain the K_{oc} value of a pesticide is high and the experimental period is approximately a year. The pesticides submitted to the authorities need to undergo a strict risk assessment. Despite conducted safety assessment studies with the high cost and long test duration, the submitted pesticides can be rejected in the risk assessment depending on K_{oc} [EFSA, 2010]. Therefore, it is necessary to develop an accurate prediction model for K_{oc} values to efficiently develop pesticides.

In this chapter, we propose a QSPR model, which is a rapid and inexpensive method to predict

the physicochemical properties theoretically [Hansch and Leo, 1995]. QSPR focuses on the quantitative relationships between chemical structures and physicochemical properties to predict various parameters from the structure of chemical substances. QSPR provides a numerical representation of each element of the chemical structure. The chemical property is termed as the descriptor. Common descriptors include fingerprints expressing the presence or absence of a specific chemical moiety. These descriptors help estimate values of molecular descriptors (i.e., molecular dipole moment, charge, and energy) of chemical substances [Karelson et al., 1996].

Various QSPR models have been developed to predict physicochemical properties that are important for pharmaceutical development including octanol/water partition coefficient (log P_{ow}) [Padmanabhan et al., 2006] and water solubility [Chen et al., 2002]. Numerous QSPR models to predict the soil adsorption coefficient have also been developed for various compounds over the last 20-30 years [Gawlik et al., 1997, Gramatica, 2010, Nolte and Ragas, 2017]. Kahn et.al. developed a model to calculate the K_{oc} value using multiple linear regression (MLR) for 344 environmental pollutants [Kahn et al., 2005]. They suggested that hydrophobicity, compound size, shape, and charge distribution were related to the interaction between environmental pollutants. Additionally, Jiao developed a model to calculate the K_{oc} value via principal component analysis and back propagation for polychlorinated biphenyls [Jiao, 2012]. Another model using log P_{ow}, atomic bond index, and molecular three-dimensional structures as descriptors has also been developed to calculate the K_{oc} value [Shao et al., 2014]. A partial least squares regression model using the two-dimensional chemical structure as a descriptor was developed for herbicides [Freitas et al., 2014]. In these studies, molecular descriptors and topological descriptors are mainly used as explanatory variables. The advantage of using molecular descriptors and topological descriptors is that they can

be easily calculated using the structure.

In contrast, physicochemical properties including water solubility, octanol/water partition coefficient (log P_{ow}), and vapor pressure of compounds are closely related to the soil adsorption process. However, the properties have not been analyzed in detail in previous studies. Laboratory experiments are required to obtain the experimental data. Thus, few experimental data have been used to develop QSPR models to date. Specifically, previous studies indicated that log P_{ow} is closely related to the prediction of K_{oc}. A model using predicted values of log P was developed by dos Reis et al. [dos Reis et al., 2013].

The objective of our study was to create a fast and inexpensive estimation method of soil adsorption; thus, we compared the performances of K_{oc} prediction models using both molecular descriptors and physicochemical properties with existing models. We gathered physicochemical properties by using data mining technique and referring to chemical database. Gramatica et al., Huuskonen, Duchowicz et al., and dos Reis et al. used a dataset that included 163 types of pesticides and developed QSPR models using molecular descriptors [Gramatica et al., 2000, Huuskonen, 2003, Duchowicz et al., 2007, dos Reis et al., 2014]. In this chapter, we used the same data set and developed QSPR models for K_{oc} to analyze the effect of physicochemical properties and molecular descriptors. The data set is one of the most famous data set. The reason why the data set is used previous studies that it includes many chemical classes. In addition, the range of the log K_{oc} values are relatively wide. Thus, it is suitable for building QSPR models by using the data set.

3.2 Material and methods

3.2.1 Data set

To collect the experimental data of physicochemical properties, we used a previous study on pesticide risk assessment in EU. For the application for the registration of pesticides in the EU, the evaluation reports issued since July 2005 are published on the internet by EU authorities. The results, discussion, and evaluation of the toxicity test, environmental fate test, and physicochemical properties test used for the application are described in each evaluation report. Regarding pesticides which were not listed in the EU evaluation reports, we gathered the values of physicochemical properties from US EPA's CompTox Chemicals Dashboard which was originally titled the Chemistry Dashboard [Williams et al., 2017]. If there are more than one experimental data for a compound, we used representative experimental data (i.e. submitted to the authorities or first displayed on the database) which is described in the EFSA peer review report or CompTox Chemicals Dashboard of a compound. Using the data, the QSPR model for the prediction of log K_{oc} based on a machine learning technique was developed, and the performance of the model was evaluated.

To compare the model developed in the study with models in previous studies, we used the K_{oc} experimental data (log K_{oc}) of 163 pesticides used from previous studies [Gramatica et al., 2000, Huuskonen, 2003, Duchowicz et al., 2007, dos Reis et al., 2014]. In these studies, 143 pesticides were used as the training set and 20 pesticides were used as the test set. The test set is famous as the test set of prediction of K_{oc} [Gramatica et al., 2000] because it includes large variety of chemical properties of pesticides. As we would like to compare our models with previous models, we used same dataset. The chemical compound datasets belong to several pesticide classes: six acetanilides,

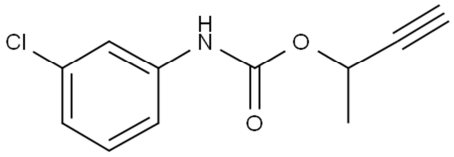
29 carbamates, eight dinitroanilines, eight organochlorides, 28 organophosphates, 44 phenylureas, 13 triazines, and seven di- and triazoles. The log K_{oc} values ranged from 0.42 to 5.31 for the training sets and from 0.56 to 4.50 for the test sets.

3.2.2 Software and program

We used Python 3.7 as a programming language [Oliphant, 2007]. We used python modules Numpy [Walt et al., 2011], Scipy [Jones et al., 2001] and matplotlib[Hunter, 2007] for calculations and visualization. To implement linear regression and SVM models, we used scikit-learn which is a machine learning package in Python [Pedregosa et al., 2011]. We optimized SVM model by grid search implemented in scikit-learn.

We calculated the molecular descriptors by employing the following procedure. First, simplified molecular input line entry system (SMILES) of the compound in the dataset was collected from PubChem [Kim et al., 2016] and ChemSpider [Pence and Williams, 2010]. Specifically, SMILES corresponds to a character string of the chemical structure of a molecule with alphanumeric characters of encoding standard for electronic communication called American Standard Code for Information Interchange (ASCII). It is typically used in chemical molecular software. The example of SMILES was shown in Figure 3.1. Subsequently, we normalized the molecular structure of the obtained compound by MMFF 94 Force Field (mmff 94) [Halgren, 1996] using Open Babel [O'Boyle et al., 2011] to calculate the molecular descriptors.

Based on the above molecular structure, a total of 1,826 molecular descriptors of different types were calculated using freely available software mordred [Moriwaki et al., 2018]. We can easily install mordred and use it on the command line interface, web application, and Python package.



Compound name: Chlorbufam
SMILES: CC(C#C)OC(=O)NC1=CC(=CC=C1)Cl

Figure 3.1: Chemical structural formula and the simplified molecular input line entry system SMILES = simplified molecular input line entry system

Mordred is at least twice as fast as PaDEL-Descriptor [Yap, 2011] which is a major software for performance benchmark tests.

The molecular descriptors calculated by mordred describe the structural diversity of the compound. Types of descriptors include 1) constitutional descriptors, 2) topological descriptors, 3) 2D matrix-based descriptors, 4) 3D matrix-based descriptors, and 5) WHIM descriptors. Table 3.1 lists the representative descriptors. Details of the descriptors and calculation procedures are described in the molecular descriptor handbook [Todeschini and Consonni, 2008].

3.2.3 Obtaining physicochemical properties of experimental data

The European Food Safety Authority (EFSA) publishes a peer review report of pesticide risk assessment after reviewing a dossier submitted by a pesticide manufacturer. The peer review report includes various experimental data and endpoints for the pesticide such as chemical properties, toxicological, ecotoxicological, and environmental fate. The published reports are opened to public via website of EFSA (https://efsa.onlinelibrary.wiley.com/journal/18314732). We obtained the

experimental data of physicochemical properties of each pesticide from a previously published peer reviewed report. With respect to the pesticides that are not registered in the EU and do not have a corresponding EFSA Peer review report, experimental data were obtained from U.S. EPA's Chemistry Dashboard [Williams et al., 2017]. Seven experimental data on the physicochemical properties of pesticides were extracted. Table 3.2 shows the physicochemical properties and their meaning.

3.2.4 Model development and validation

We developed a prediction model based on the gradient boosting decision tree (GBDT) algorithm [Roe et al., 2005] as a non-linear model to compare the performance with models in previous studies. Gradient boosting was proposed by Friedman [Friedman, 2001] and consists of a gradient descent method and a boosting method. The boosting algorithm is a part of ensemble learning and is used to construct an entire learner by integrating a plurality of weak learners. Decision trees are often used as weak learners. Using a decision tree as a weak learner in boosting has advantages such as being strong against outliers in data and being able to handle discrete variables, and missing values. Therefore, it is one of the most used algorithms in various data analysis contests such as Kaggle and KDD Cup [Chen and Guestrin, 2016]. In addition, tree models including gradient boosting trees are generally excellent in readability. In this chapter, it is possible to read which physicochemical properties or molecular descriptors contribute to the model. We used XGBoost [Chen and Guestrin, 2016], which is a machine learning package for the GBDT algorithm of Python and has a highly scalable end-to-end tree boosting system. XGBoost employs an algorithm that determines the direction of tree branching in advance for sparse data such as many missing values.

Table 3.1: Representative descriptors calculated via Mordred

Descriptor type	Number	Representative descriptor
Acidic group count	1	nAcid
ALOGP	3	ALogP, ALogp2, AMR
Atom count	14	nAtom, nHeavyAtom, nH, nB, nC, nN, nO, nS, nP, nF
Bond count	10	nBonds, nBonds2, nBondsS, nBondsS2, nBondsS3
Atom type electrotopological state	489	nHBd, nwHBd, SHBd, minHBd, LipoaffinityIndex, MAXDN
Molecular linear free energy relation	6	MLFER_A, MLFER_BH, MLFER_BO, MLFER_S, MLFER_E
Rule of five	1	LipinskiFailures
Topological	3	topoRadius, topoDiameter, topoShape
Topological distance matrix	11	SpMax_D, SpDiam_D, SpAD_D, SpMAD_D, EE_D, VE1_D
Van der Waals volume	1	VABC
3D autocorrelation	80	TDB1u, TDB2u, TDB3u, TDB4u, TDB5u, TDB6u, TDB7u
Charged partial surface area	29	PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3
RDF	210	RDF10u, RDF15u, RDF20u, RDF25u, RDF30u, RDF35u,
WHIM	91	L1u P1u, E1u, Tu, Au, Du, L1m, P1m, E1m, Km, Dm

Table 3.2: Representative descriptors calculated via Mordred

Experimental data	Description
log S	Limit amount at which a certain solute dissolve in a certain amount of water
log P	Dimensionless number of the hydrophobicity and migration of chemical substances
log H	Constant representing the solubility in the liquid of the components in the gas
log VP	Gas phase pressure of the substance in phase equilibrium between solid and liquid
Flash Point (FP)	Lowest temperature at which the material can volatilize to make a flammable mixture with air
Melting Point (MP)	Temperature at which solid melts to liquid
Surface tension (ST)	Characteristics trying to make the surface as small as possible

The model search is accelerated by parallel distributed processing.

We performed parameter tuning to find the best parameters for the prediction model using GBDT algorithm. We conducted a grid search for the values of max depth, min child weight, n estimators, and reg alpha. Although there are other search methods such as optuna [Akiba et al., 2019] which is a parameter optimization framework, we adopted a grid search. It is a conventional and quick approach to find the best parameter because of the quick algorithm of GBDT.

A total of 1,826 descriptors were calculated by mordred. It took 14 seconds to calculate all descriptors for 163 pesticides. Most calculated molecular descriptors were not significantly important in the calculation of log K_{oc} values; thus, we selected the important descriptors to predict the ability of the model by feature importance from GBDT. GBDT has three indicators—weight, gain, and cover of feature importance. Weight is an indicator of how many times a feature is used to split the data across all trees to only observe the existing number. There is no information on how close the branch is to prediction or how much of the branch is used for the input. Gain is an indicator of how much the evaluation criteria can be improved. The tree-based model finds the variable and the threshold value that maximizes this at each branch. Cover is the sum of second order gradient of training data classified to the leaf. Square loss simply corresponds to the number of instances in that branch. We adopted gain which is the default parameter as an indicator of feature selection.

We also develop linear regression models based on the MLR and support vector machine (SVM) using two molecular descriptors and five physicochemical properties selected by feature selection. We compared these two prediction models with prediction model using GBDT algorithm.

The constructed model was evaluated based on the OECD principles for model validation [Gramatica, 2007] and previous studies [dos Reis et al., 2014]. The fitting performance, robust-

ness, and prediction ability of the model were evaluated using the coefficient of determination (R^2), leave-one-out cross validation of correlation coefficient (Q^2_{LOO}), the residual sum of squares (RSS), and the standard error of calibration (SEC). To evaluate the QSPR models, some recent studies have suggested that R^2 values should be greater than 0.7 and that values of RSS and SEC should be close to 0. Other detailed definitions and calculations of the parameters are given in references [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012]. We also evaluated the Applicability Domain (AD). The AD is defined as the response and chemical structure space in which the QSAR model makes predictions with given reliability. We conducted a standardization approach using software called “AD using standardization approach” [Roy et al., 2015]. In the approach, it was mentioned that 99.7% of the population will remain within the range mean $\pm$ 3 standard deviation (SD) in keeping with ideal data distribution. We also characterized applicability domain by Euclidean distance-based method using a software called “Euclidean-Distance 1.0” [Ambure et al., 2015]. The Euclidean distance is the ordinary distance between two points in the Euclidean space. In the Euclidean-Distance 1.0 software, the Euclidean distance scores and the mean distance scores are calculated followed by the normalization within the interval of zero to one.

3.2.5 Comparison of the developed models with EPI Suite and models in the previous studies

EPI suite (The Estimations Programs Interface for Windows) was also used for comparison with the model constructed in this study. EPI suite was jointly developed by US EPA and Syracuse. It is a model that predicts and calculates the physicochemical properties of each substance based

on the chemical structure [Card et al., 2017].This software has been used in various fields for risk assessment, such as for the examination of new chemical substances under the Toxic Substance Control Act (TSCA) when products containing new chemical substances by the US EPA are imported to the United States.

The EPI suite incorporates a model that predicts various parameters such as bioconcentration and biodegradability. In addition, KOCWIN, which is a subset of EPI suite, is incorporated as a model to predict log K_{oc}. KOCWIN uses an estimation method using the molecular connectivity index (MCI), the group contribution coefficient, and an estimation method based on log K_{ow}. Prediction parameters were calculated and compared for each compound in the training set and test set using both the MCI estimation method and the log K_{ow} estimation method.

We compared the prediction abilities of the QSPR model and models developed in previous four studies using a molecular descriptor and experimental data of physicochemical properties [Gramatica et al., 2000, Huuskonen, 2003, Duchowicz et al., 2007, dos Reis et al., 2014].

3.3 Result and Discussion

3.3.1 Result of the developed models

We optimized the standard parameters for GBDT via a grid search using the data set. We obtained the best values of R^2, SEC, and standard error of prediction (SEP) for both the training and test data sets using maximum depth = 2, minimum child weight = 4, n estimators = 100, and reg alpha = 0.5.

We developed a prediction model via GBDT using experimental values of physicochemical properties and molecular descriptors for pesticides in the data set. We used 143 pesticides as a

training set and 20 pesticides as the test set, in accordance with previous studies. We compared 3 models using only physicochemical properties, only molecular descriptors, and both of these. We obtained the best prediction ability by using both physicochemical properties and molecular descriptors. The statistical parameters of each model are shown in Table 3.3.

We obtained the best prediction ability by using both physicochemical properties and molecular descriptors.

For the physicochemical properties, we performed feature importance selection by GBDT, and 5 types of values were selected. The selected physicochemical properties were log S, log P_{ow}, flash point, log H, and surface tension, which are listed in Table 3.2. The most important physicochemical property was log S, which is a logarithm of water solubility. Log P_{ow}, which is a dimensionless number of the hydrophobicity and migration of chemical substances, was also important for the prediction of log K_{oc}. In previous studies on the prediction of log K_{oc}, water solubility and log P_{ow} were used as a part of the parameters. It was also indicated that increases in the value of log P_{ow} increase the ability of soil adsorption [Sabljić et al., 1995, Gao et al., 1996]. With respect to the molecular descriptors, we also performed feature importance selection from the 1826 descriptors calculated by mordred. Three molecular descriptors were selected for model development. The selected molecular descriptors are presented in Table 3.4.

The most important descriptor was FilterItLogS, which denotes the calculated value of log S by Filter - it™. The next important descriptor was ATSC2dv, which is a centered Moreau - Broto autocorrelation of log function of topological distance (lag 2) weighted by valence electrons. The autocorrelation of a topological structure (ATS) descriptors describe how properties are distributed along the topological structure. The selected physicochemical properties and molec-

Table 3.3: Comparison of statistical parameters between the prediction models by GBDT algorithm using physicochemical properties and molecular descriptors

No.	Model	No. of variables	N_t/N_p	R^2	R^2_{PRED}	Q^2_{LOO}	SEC	SEP
1	physicochemical properties	1360	143/20	0.911	0.722	0.913	0.269	0.49
2	Molecular descriptors	11	143/20	0.887	0.682	0.873	0.303	0.524
3	(1) + (2)	7 / (1360 + 11)	143/20	0.935	0.775	0.931	0.231	0.421

Q^2_{LOO} = leave one out cross validation of correlation coefficient; SEC = standard error of calibration; SEP = standard error of prediction; N_p = number of test set; N_t = number of training set

Table 3.4: Molecular descriptors selected by GBDT

Molecular descriptor	Description
FilterItLogS	Log S calculated by Filter-it™
ATSC2dv	Centered moreau-broto autocorrelation of lag 2 weighted by valence electrons
AATS4v	Geary autocorrelation of lag 3 weighted by polarizability

Filter-it™ : A software that eliminates unwanted properties of molecules of chemicals.

ular descriptors are consistent with the results of previous studies. The sorption of nonionic organic compounds in soil is related to a mechanism that makes hydrophobicity the driving force [Wen et al., 2012, dos Reis et al., 2013, dos Reis et al., 2014]. Thus, the physicochemical properties and molecular descriptors related to the hydrophobicity of the pesticides are relevant for the adsorption process. This is an explanation that log S, log P_{ow}, surface tension, and FilterItLogS have high correlation with log K_{oc}. The flash point is the lowest temperature at which vapor is released at a concentration sufficient to form a flammable mixture. Thus, the flash point is dependent on the boiling point and vapor pressure of the liquid [Fujii and Hermann, 1982]. Previous studies have shown a correlation between volatilization fluxes with the log of the ratio of vapor pressure and K_{oc} [Woodrow et al., 1997, Alvarez-Benedi et al., 1999]. This correlation is evidence that a binding mechanism exists, including solute exchange between the adsorbed, dissolved, and vapor phases. Besides, the importance of ATSC2dv and AATS4v as topological descriptors was also shown. These are topological descriptors regarding topological distance weighted by valence electrons and polarizability. The probability of H - bonding with soil and water is based on the number of electronegative atoms in the molecule [Gramatica et al., 2000]. The results indicated that the K_{oc} values of pesticides were mainly affected by molecular lipo - hydrophobic properties and topological properties of molecules. To evaluate the accuracy of the models, the coefficient of determination (R^2), SEC, SEP, average relative error of prediction, and concordance correlation coefficient were calculated. Each expression is detailed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_{obs} - y_{\mathrm{pred}}\right)^2}{\sum_{i=1}^{n} \left(y_{\mathrm{obs}} - y_{\mathrm{means}}\right)^2} \tag{3.1}$$

$$SEC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{n - p - 1}} \tag{3.2}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{n}} \tag{3.3}$$

In Equations 3.1 to 3.3, y_{pred} denotes the predicted value of log K_{oc} and y_{obs} denotes the experimental data of log K_{oc}. In addition, y_{means} denotes the average value of log K_{oc}, and n denotes the number of samples. The predicted values calculated via the developed model are shown in Table 3.5. In addition, plots of predicted values and measured values of log K_{oc} are shown in Figure 3.2.

An MLR model was developed using the same data set, and explanatory variables explained in the section 3.3.1 were used. The regression equation of the prediction model is given in Equation 3.4.

$$\begin{aligned} logKoc \;=\; & -0.1875logS \;+\; 0.1927logP \;-\; 0.0004FP \;+\; 0.0355logH \;+\; 0.0084ST \\ & -\; 0.1089FilterItLogS \;-\; 0.0010ATSC2dv \;+\; 0.0052AATS4v \;+\; 0.7867 \end{aligned} \tag{3.4}$$

The R^2 values for the training set and test set of experimental data were relatively low (0.807 and 0.637, respectively), and Q^2_{LOO} was 0.767. Statistical parameters for the developed MLR model are shown in Table 3.6.

We also developed SVM models using the same data set. In the present study, we compared linear and nonlinear kernels. Polykernel, radial basis function kernel, and sigmoid kernel were used

Table 3.5: Test set with experimental and calculated log K_{oc} values

No.	Compound name	Exp.log K_{oc}	GBDT	MLR	SVM
1	Aldicarb sulfoxide	0.56	1.49	2.36	1.41
2	Anilazine	3.00	3.00	3.26	3.4
3	Asulam	2.48	1.69	1.35	1.4
4	Chlorbufam	2.21	2.27	2.46	2.48
5	Cyromazine Terbuthylazine	2.30	1.66	2.85	1.48
6	Demeton-S-methyl	1.49	1.59	3.92	1.55
7	Dichlorvos	1.67	1.50	2.59	1.89
8	EPN	3.12	3.58	1.31	3.9
9	Fenobucarb	1.71	2.17	3.45	2.37
10	Iprobenfos	2.40	2.56	1.37	2.44
11	Leptophos	4.50	4.12	2.43	4.23
12	Methidathion	1.53	2.44	1.83	2.53
13	Neburon	3.40	3.04	1.45	2.95
14	Piperophos	3.44	3.12	1.74	2.88
15	Pirimicarb	1.90	1.92	3.82	1.73
16	Pirimiphos methyl	3.00	2.68	2.24	2.99
17	Sulprofos	4.08	4.18	2.42	4.01
18	Terbuthylazine	2.32	2.55	4.42	2.72
19	Thiodicarb	2.54	2.36	2.39	2.64
20	Xylicarb	1.71	1.99	2.94	2.05

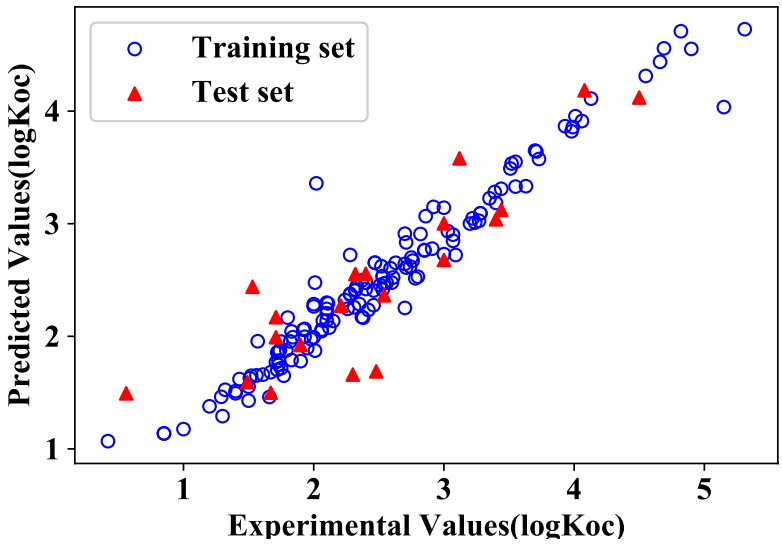


Figure 3.2: Plots of the experimental data and predicted values of log soil adsorption coefficient by gradient boosting decision tree

Table 3.6: Comparison of statistical parameters between the prediction models developed by the GBDT, the MLR and the SVM

Model	R^2	R^2_{PRED}	Q^2_{LOO}	SEC	SEP
GBDT	**0.935**	**0.775**	**0.931**	**0.231**	**0.421**
MLR	0.822	0.742	0.812	0.397	0.539
SVM	0.817	0.720	0.812	0.401	0.528

as nonlinear kernels. Three parameters, C (regularization parameter), gamma (the relative weight of the regression error), and sigma (kernel parameters), were optimized via a grid search. With respect to the result of the grid search, SVM using the linear kernel (C = 10) exhibited the optimal score. The R^2 values for the training set and test set of experimental data were relatively low (0.817 and 0.720, respectively), and Q^2_{LOO} was 0.812. Statistical parameters for the developed SVM model are shown in Table 3.6.

To confirm the performance of the 3 developed models, we compared the calculated log K_{oc} and statistical parameters in Tables 3.5 and 3.6. Although some predicted values by the MLR and SVM models showed the best accuracy, over half of the values predicted by the GBDT algorithm exhibited the best accuracy. In addition, considering the statistical parameters shown in Table 3.6, we obtained the best value for all statistical parameters. Thus, we assumed that the best of the three prediction models is based on the GBDT algorithm.

3.3.2 Applicability domain

We evaluated the applicability domain (AD) by a software which is called “AD Using Standardization Approach”. We checked that the test set pesticides were not outside the AD and that the training set compounds were not outliers. The results of the calculation of outliers by AD using standardization approach showed that there are no outliers in the training set, with the normal

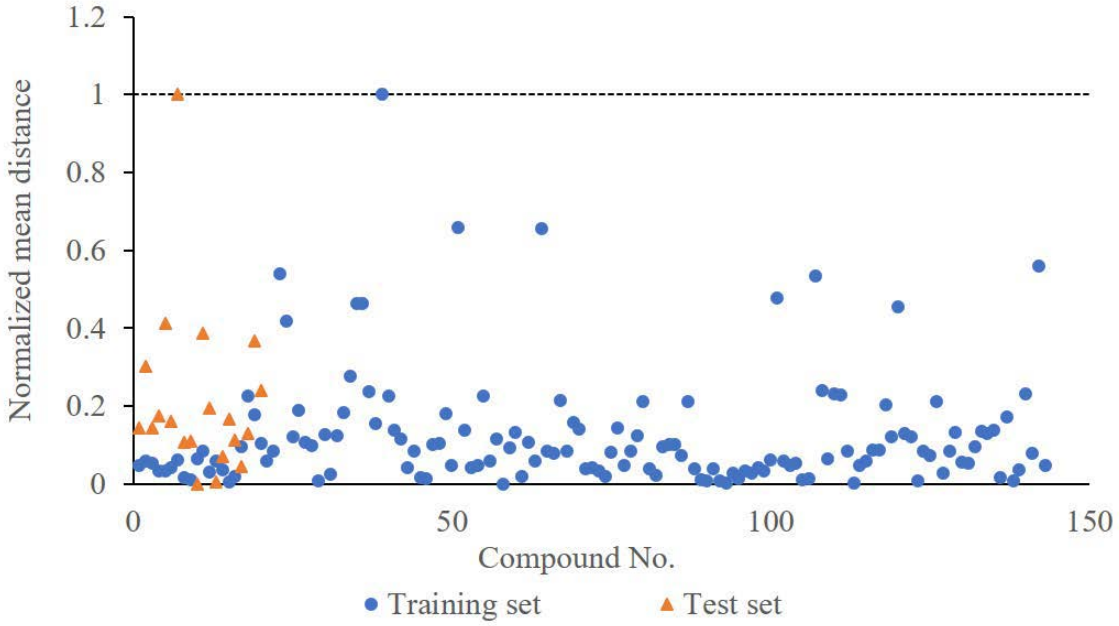


Figure 3.3: Plot of applicability domains characterized by the Euclidean distance 1.0

distribution pattern of approximately 99.7% of the population remaining with the range mean of ± 3 SD. Also, there is no test compound outside the AD. In addition, the Euclidean graph generated by Euclidean - Distance 1.0 is shown in Figure 3.3

According to the normalized mean distances in the graph, only one training compound (No.39) and one test compound (No.7) located outside the AD. Therefore, the QSAR model developed in the present study can make predictions with given reliability.

3.3.3 Comparison of the developed models with OPERA and models in the previous studies

We compared our developed model with the calculation values of log K_{oc} from EPI Suite. EPI is software that predicts and calculates the physical properties of each chemical substance based on

Table 3.7: Comparison of statistical parameters between GBDT based prediction model and EPI suite

Model	R^2	R^2_{PRED}	SEC	SEP
GBDT	**0.935**	**0.775**	**0.231**	**0.421**
EPI Suite (MCI)	0.749	0.661	0.458	0.565
EPI Suite (log Kow)	0.734	0.675	0.492	0.585

Table 3.8: Overall summary of statistical parameters for all QSPR models

Model	No. of variables	R^2	R^2_{PRED}	Q^2_{LOO}	SEC
Gramatica et al., 2000	6	0.843	0.67	0.824	0.35
Huuskonen, 2003	12	0.82	**0.79**	0.79	0.37
Duchowicz et al., 2007	6	0.9	0.71	0.89	0.29
dos Reis et al., 2014	4	0.852	0.743	0.84	0.343
EPI Suite (MCI)	-	0.749	0.661	-	0.458
EPI Suite (log Kow)	-	0.734	0.675	-	0.492
MLR	7	0.822	0.742	0.812	0.397
SVM	7	0.817	0.72	0.812	0.401
GBDT	7	**0.935**	0.775	**0.931**	**0.231**

its structure. We calculated log K_{oc} using KOCWIN by the method of MCI and log K_{ow}. Table 3.7 compares the prediction abilities of GBDT based prediction models and EPI Suite. Although EPI Suite has a larger chemical space compared to our developed models, our model developed by GBDT showed better prediction ability than KOCWIN from EPI Suite with respect to both MCI and log K_{ow}. Table 3.8 compares the prediction abilities of QSPR models of GBDT using a molecular descriptor and physicochemical properties with the models used in previous studies.

For our developed GBDT based prediction model, the prediction accuracy, fitness, and robustness were higher than those obtained in previous studies except for the model by Huuskonen in terms of R^2. Our model showed a lower value of SEC than the model by the author. In addition, the author used 12 explanatory variables, and we used 7 explanatory variables. The results indicated that the experimental data of the physicochemical properties and molecular descriptors are

important for log K_{oc} estimation. Therefore, the present study showed that log K_{oc} can be estimated from the structure of the compound and preliminary physicochemical properties without expensive laboratory studies on log K_{oc}. Based on this finding, we were able to reduce the cost for log K_{oc} estimation as required for pesticide development and shortened the development period. In addition, we can use the results for preliminary risk assessment. Specifically, we can stop development of pesticides with weak prospects halfway based on the QSPR models. This is because the decision of the project for development is currently determined via experts' heuristics. The results provide objective justification for the aforementioned types of decisions. An overall summary of the models is shown in Table 3.8.

3.4 Conclusion

In this chapter, we have developed prediction models for K_{oc} values. We proposed GBDT based prediction model used the experimental data of physicochemical properties by gathering evaluation report of pesticide, and molecular descriptors calculated by cheminformatics software. As a result, the following results were obtained.

- By using both the molecular descriptors calculated from the structural formulas and experimental physicochemical properties from the literature and open databases, the model prediction accuracy was significantly improved.

- By utilizing the GBDT algorithm, the prediction accuracy of the proposed model was further increased.

- The prediction models based on the GBDT algorithm demonstrated the best prediction abilities among the different machine learning models.
- The proposed models were developed using the open data sources and free software.

The results of this chapter showed that it is possible to perform preliminary environmental risk assessment at a low cost and without time-consuming studies. However, it takes a considerable amount of time and effort to collect data from huge number of documents such as literature. Therefore, in the next chapter, we take steps to gathering physicochemical properties by using latest cheminformatics software.

Chapter 4

Prediction of soil adsorption coefficient using calculated physicochemical properties and molecular descriptors

4.1 Introduction

Physicochemical properties such as water solubility, octanol/water partition coefficient (log P), and vapor pressure of compounds are considered to be closely related to the soil adsorption process. Specifically, previous studies have shown that log P is closely associated with K_{oc} prediction [dos Reis et al., 2013]. However, as experiments are required to obtain these values, there are few reports on QSPR models using physicochemical properties.

In the previous chapter, we developed a machine learning-based QSPR model with five physicochemical properties and three molecular descriptors for 163 pesticides. We reported that a high-performance model in terms of accuracy was established, as opposed to a model using only molecular descriptors. Physicochemical properties were collected from EFSA peer review report of pesticides and the U.S. EPA's) Chemistry Dashboard. However, collecting experimental data for a wide range of compounds requires a considerable amount of time and effort.

In the work presented herein, the goal was to develop an accurate predictive model by gathering physicochemical properties using a relatively easy approach and freely available software. Also, we used largest dataset of K_{oc} to improve accuracy and versatility compared to the previous chapter. The U.S. EPA has developed the OPEn structure-activity Relationship App (OPERA) software to predict physicochemical properties [Mansouri et al., 2018]. OPERA has been developed using data of a wide range of chemical substances listed in the physicochemical properties database (PHYSPROP). It can predict various physicochemical properties and endpoints of environmental fate, offering highly reliable model performance. By developing a prediction model by machine learning for K_{oc} using physicochemical properties and environmental fate endpoints calculated by OPERA instead of experimental data, a highly accurate model may be rapidly developed.

In this chapter, we used a dataset containing 964 different, many chemicals obtained in a previous study [Shao et al., 2014]. We also changed the collection method of physicochemical properties by using OPERA software. We developed machine learning models using molecular descriptors, physicochemical properties, and environmental fate endpoints as calculated by OPERA. Although using experimental data is ideal for the development of QSPR models, our proposed procedure is considered an acceptable substitute for experimental data if the latter is difficult to obtain. In addition, as we used the biggest K_{oc} dataset available, it was possible to apply the developed models to a diverse range of chemical compounds. After developing the model, we confirmed its performance and the contribution of molecular descriptors and physicochemical properties.

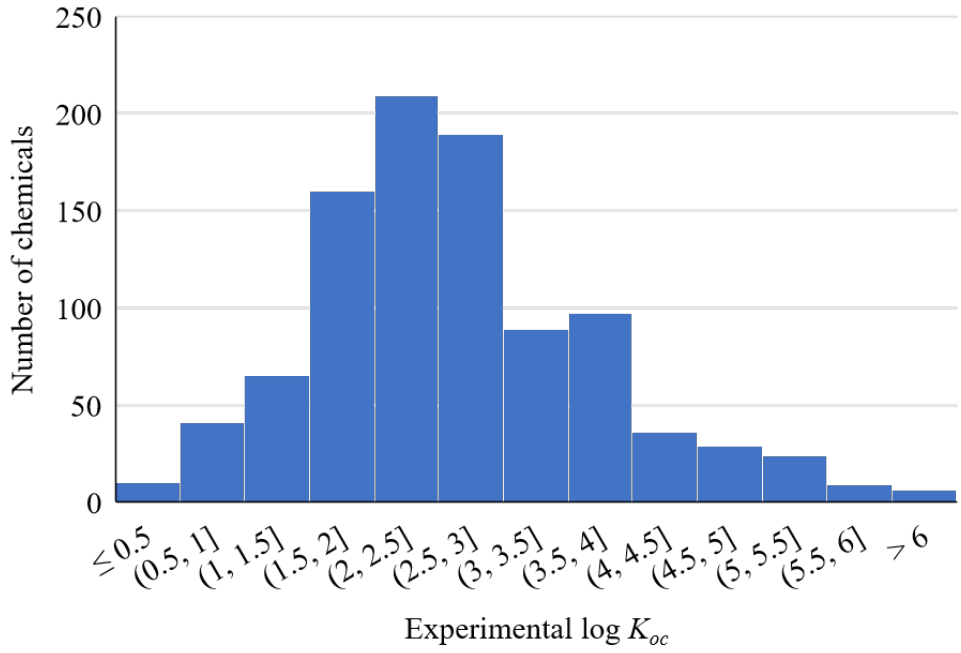


Figure 4.1: Histogram of distribution of experimental log K_{oc} in the dataset

4.2 Material and methods

4.2.1 Dataset

A dataset of 964 nonionic chemicals was obtained from previous studies [Shao et al., 2014, Olguin et al., 2017]. In this study, the dataset was divided into training and test sets using a Y-ranking method [Bhhatarai and Gramatica, 2011], resulting in 644 and 320 chemicals, respectively. We used the same training and test sets to compare each model. The dataset is ideal for QSPR modeling because it is the largest dataset of K_{oc} available, with many varieties of chemicals included. The log K_{oc} values ranged from -0.386 to 6.469 for the training sets and from -0.630 to 6.100 for the test sets. The distribution of K_{oc} values and chemical groups of the dataset are shown in Figure 4.1 and Table 4.1.

Table 4.1: Chemical groups of the target chemicals in the dataset

Chemical group	Number of chemicals
Alcohols	53
Alkanes	26
Alkenes and Alkynes	28
Amides	26
Amines	30
Anilines	44
Aromatic heterocycles	35
Benzene derivatives	85
Benzenes and Alkylbenzenes	36
Biphenyls	42
Carbonyl compounds	35
Esters	45
Ethers	20
Halogenated benzenes	31
Halogenated alkanes	64
Halogenated alkenes	13
Heterocycles	10
Nitriles	16
Nitroalkanes	6
Nitrobenzenes	22
Organic acids	67
Organophosphorus compounds	20
Organosulfur compounds	18
Other Compounds	32
Phenols	66
Phenyl ureas	24
Polyaromatic heterocycles	14
Polyciclyc aromatic hydrocarbons	49
Triazines	7
Total	964

Table 4.2: The parameters calculated by OPERA.

Descriptor type	Number	Representative descriptor
Acidic group count	1	nAcid
ALOGP	3	ALogP, ALogp2, AMR
Atom count	14	nAtom, nHeavyAtom, nH, nB, nC, nN, nO, nS, nP, nF
Bond count	10	nBonds, nBonds2, nBondsS, nBondsS2, nBondsS3
Atom type electrotopological state	489	nHBd, nwHBd, SHBd, minHBd, LipoaffinityIndex, MAXDN
Molecular linear free energy relation	6	MLFER_A, MLFER_BH, MLFER_BO, MLFER_S, MLFER_E
Rule of five	1	LipinskiFailures
Topological	3	topoRadius, topoDiameter, topoShape
Topological distance matrix	11	SpMax_D, SpDiam_D, SpAD_D, SpMAD_D, EE_D, VE1_D
Van der Waals volume	1	VABC
3D autocorrelation	80	TDB1u, TDB2u, TDB3u, TDB4u, TDB5u, TDB6u, TDB7u
Charged partial surface area	29	PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3
RDF	210	RDF10u, RDF15u, RDF20u, RDF25u, RDF30u, RDF35u,
WHIM	91	L1u P1u, E1u, Tu, Au, Du, L1m, P1m, E1m, Km, Dm

4.2.2 Software and program

Python 3.7 was the programming language used and the Python modules, Matplotlib, Numpy, and Scipy were used for calculation and visualization. We also used scikit-learn, a machine learning package in Python.

Mordred was used to calculate molecular descriptors. A total of 1826 molecular descriptors were calculated by Mordred. The descriptors and calculation procedures are provided in the molecular descriptor handbook [Todeschini and Consonni, 2008].

The calculated physicochemical properties and environmental fate endpoints were generated by OPERA. OPERA is available in Matlab, C, and C++ languages and is based on PaDEL-descriptors. OPERA was developed using the publicly available PHYSPROP database. A total of 46 properties can be generated by OPERA and representative properties calculated by OPERA are shown in Table 4.2. OPERA demonstrated good predictive performance for determining physicochemical properties: the R^2 test values ranged from 0.71 to 0.96 (average: 0.82).

4.2.3 Model development and validation

In this chapter, the GBDT algorithm was also adopted as a non-linear model and an ensemble algorithm for the prediction models. A boosting algorithm such as XGboost [Chen and Guestrin, 2016] or CatBoost [Prokhorenkova et al., 2017] is one of the most popular additions in data analysis competitions. We confirmed the explanatory variables that contributed to the developed model as tree models, generally have excellent readability. Among the boosting algorithms, LightGBM, provided by Microsoft, has attracted great attention [Ke et al., 2017]. LightGBM has significantly improved performance compared to other GBDT algorithm in terms of computational speed, memory consumption, and communication costs for parallel learning. In cheminformatics, there is little research available on the use of LightGBM [Zhang et al., 2019, Su et al., 2021]. The use of this algorithm ensures accurate, rapid parameter tuning as well as efficient prediction.

The SVM and MLR models were developed using the same explanatory variables used in the GBDT based prediction model. The performance of these models was compared to the performance of the GBDT based prediction model.

Model validation was conducted for the developed models in accordance with the principles of the Organization for Economic Co-operation and Development (OECD) [Gramatica, 2007]. The fitting performance, prediction ability, and model robustness were evaluated by the coefficient of determination (R^2), the coefficient of multiple determinations of 10-fold cross-validation (R^2_{10fold}), leave-one-out cross-validation of correlation coefficient (Q^2_{LOO}), concordance correlation coefficient in the internal validation (CCC), and the root mean square error (RMSE). To eliminate the possibility that the relationship between the explanatory variable and the objective variable is

accidental, a Y-scrambling test was performed and R^2_{Yscr}, Q^2_{Yscr}, and RMSE AVY_{Yscr} were calculated. The possibility by chance is ruled out if the values of R^2 and Q^2_{LOO} are greater than R^2_{Yscr} and Q^2_{Yscr}, respectively, and the value of RMSE is less than RMSE AVY_{Yscr}. A detailed definition of the statistical parameters is provided in the literature [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012]. We compared these statistical parameters with those of previous studies [Shao et al., 2014, Olguin et al., 2017].

In addition, to evaluate the external predictivity of our model, we performed validation according to the previous literature [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012, Roy et al., 2015]. We calculated the values of variance explained in external prediction (Q^2_{F1} and Q^2_{F2}), root mean square error in external prediction (RMSEext), the modified coefficient of determination of the external validation (rm^2), and the value of the concordance correlation coefficient (CCC_{ext}). The equations and criteria are shown in Table 4.3.

4.2.4 Applicability Domain

The definition of the applicability domain (AD) is the response and chemical structure space in which the QSPR model shows reliable predictions. The AD of the dataset was evaluated using three different methods: a standardization approach, the Euclidean distance-based method, and the one-class support vector machine (OCSVM) method. The first method uses the AD using the standardization approach software developed by Roy et al. [Roy et al., 2015]. In this approach, 99.7% of the population remains within the range of mean $\pm$ three standard deviations (SD) to maintain ideal data distribution.

The second method, the Euclidean distance-based method, is commonly used in distance mea-

Table 4.3: Statistical metrics and criteria for external validation

Statistical metrics	Definition	Equations	Criteria
Q^2_{F1}	External predictive ability	$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{TR})^2}$	$Q^2_{F1} > 0.70$
Q^2_{F2}	External predictive ability	$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{EXT})^2}$	$Q^2_{F2} > 0.70$
$RMSE_{ext}$	Root Mean Square Error in external prediction	$RMSE_{EXT} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n_{EXT}}}$	-
CCC_{ext}	Concordance correlation coefficient	$CCC = \frac{2\sum_{i=1}^{n_{EXT}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT}(\bar{y} - \bar{\hat{y}})^2}$ $\bar{\hat{y}}$: the average of all $\hat{y}_i$	$CCC_{ext} > 0.85$
$\overline{r_m^2}$	Average of r_m^2 and r'^2_m	$\overline{r_m^2} = \frac{r_m^2 + r'^2_m}{2}$ $r_m^2 = R^2\left(1 - \sqrt{R^2 - R_0^2}\right)$ $r'^2_m = R^2\left(1 - \sqrt{R^2 - R_0'^2}\right)$	$\overline{r_m^2} > 0.65$
Δr_m^2	Difference between r_m^2 and r'^2_m	$\Delta r_m^2 = \lvert r_m^2 - r'^2_m \rvert$	$\Delta r_m^2 < 0.2$

sures using the "Euclidean-Distance 1.0" software [Ambure et al., 2015]. The Euclidean distance is an ordinary distance between two points in Euclidean space [Golmohammadi et al., 2012]. The software can calculate the Euclidean distance and mean distance scores, followed by normalization within intervals of zero to one. The calculations of Euclidean distance are shown in the following equations:

$$dij = \|Xi - Xj\| = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2} \tag{4.1}$$

$$\overline{d_\iota} = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \tag{4.2}$$

$$\text{MeanDistance}_{\text{(Normalized)}} = \frac{(\text{ MeanDistance } - \text{ Min\_MeanDistance })}{(\text{ Max\_MeanDistance } - \text{ Min\_MeanDistance })} \tag{4.3}$$

where d_{ij} is the distance score between two compounds, and $\overline{d_i}$ is the mean distance.

The third method, OCSVM, is a method that applies the SVM to the area estimation problem [Kaneko and Funatsu, 2015]. The data density can be estimated continuously, and it is applied to outlier detection, outlier sample detection, and setting of the model application range.

4.2.5 Comparison of the developed models with OPERA and the models in the previous study

The log K_{oc} calculated by OPERA was also compared with our developed model. OPERA incorporates a model that directly predicts log K_{oc} values. The data from the OPERA model are

derived from PHYSPROP, a collection of a wide variety of sources built by the Syracuse Research Corporation (USA). Experimental protocols of different parts of the data may be traced back to the original referenced literature from the database. PHYSPROP's K_{oc} data were collected from Arnot and Gobas [Arnot and Gobas, 2006]. The original data collected from the PHYSPROP database underwent a series of processes to curate chemical structures and remove duplicates.

The prediction abilities of the developed QSPR model were compared with those of the models developed in previous studies. Two studies used the same dataset; Shao et al. developed the LS-SVM, genetic algorithm-multiple linear regression (GA-MLR), and local linear regression (LLR) models and [Shao et al., 2014]. Olguin et al. developed QSPR models based on the calculated log P_{ow} by linear regression [Olguin et al., 2017].

Our proposed model development procedure was compared with our previous chapter, which adopted the experimental data of physicochemical properties as explanatory variables. In the previous chapter, 163 pesticides were used as the dataset. Five physicochemical properties and three molecular descriptors were used as explanatory variables.

4.3 Results and discussion

4.3.1 Result of the developed models

Parameter tuning of the GBDT based prediction model was carried out by a grid search using the dataset. We obtained the best R^2 values and RMSE for the training and test datasets using a max depth = 15, a min child weight = 5, n estimators = 400, and gamma = 0.001.

For the development of the GBDT based prediction model, we used 324 chemicals as a training

Table 4.4: Statistical parameters of the GBDT based prediction models using physicochemical properties, environmental fate endpoints, and molecular descriptors

No.	Model	No. of variables	R_t^2	R_p^2	Q_{LOO}^2	R_{10fold}^2	$RMSE_t$	CCC
1	Physicochemical properties and environmental fate endpoints	5	0.95	0.835	0.95	0.677	0.248	0.912
2	Molecular descriptors	5	0.958	0.874	0.959	0.713	0.227	0.932
3	1 and 2	6	**0.990**	**0.904**	**0.995**	**0.777**	**0.110**	**0.944**

Nt/Np: the number of compounds in the training set/test set; R_t^2: coefficient of determination for training set; R_p^2: coefficient of determination for test set; Q_{LOO}^2: correlation coefficient of leave-one-out cross validation; R_{10fold}^2: coefficient of determination of 10-fold cross validation; RMSEt: root mean square error of training set; RMSEt: root mean square error of test set; CCC: the value of the concordance correlation coefficient

set and 198 chemicals as the test set following the process from a previous study. Although a model using molecular descriptors calculated by Mordred as explanatory variables demonstrated an excellent performance compared with the previous model (R^2 > 0.85), a model using physicochemical properties and environmental fate endpoints generated by OPERA and molecular descriptors generated by Mordred also demonstrated excellent performance (R^2 > 0.90). In addition, the GBDT based prediction model showed the best performance comparing to other models. The statistical parameters of each model are listed in Table 4.4.

The feature importance selection of explanatory variables was performed to select physicochemical properties. The selected physicochemical property was Log P pred, the prediction value of log P_{ow}, as calculated by OPERA. In terms of the molecular descriptors, we also conducted feature importance selection for the 1826 descriptors calculated by Mordred, and four molecular descriptors were selected. The most essential descriptor was Slog P_VSA2, representing different aspects of the van der Waals surface area contribution to the chemical's lipophilicity [Leszczynski and Puzyn, 2012]. The next imperative descriptor was S log P, a calculated log P_{ow}

based on the atomic contribution mode [Wildman and Crippen, 1999]. The SIC0 is structural information content (neighborhood symmetry of 0-order) [Magnuson et al., 1983] and GATS1Z is Geary's coefficient of log function for topological distance (lag 1) weighted by atomic number. The final imperative descriptor was AATSC1v, an average centered Broto-Moreau autocorrelation of lag 1 weighted by van der Waals volumes.

The selected physicochemical properties and molecular descriptors had the same tendencies as those observed in previous research. The mechanism of sorption in soil was related to the hydrophobicity of the chemicals as a driving force [Wen et al., 2012, dos Reis et al., 2014]. This mechanism demonstrated the underlying reason why Log P pred, Slog P_VSA2, and SLogP are highly related to log K_{oc}. Additionally, the importance of GATS1Z and AATSC1v as topological descriptors was also indicated. These topological descriptors were related to the topological distance weighted by the valence electrons and polarizability. he probability of H-bonding with water and soil is dependent on the number of electronegative atoms of the molecule [Gramatica, 2010].

We calculated the R^2, Q^2_{LOO}, R^2_{10fold}, CCC, and RMSE to evaluate the developed models. The expressions are shown in the following equations:

$$\boldsymbol{R}^2 = \mathbf{1} - \frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{\sum_{i=1}^{n} \left(y_{obs} - y_{\mathrm{means}}\right)^2} \tag{4.4}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{n}} \tag{4.5}$$

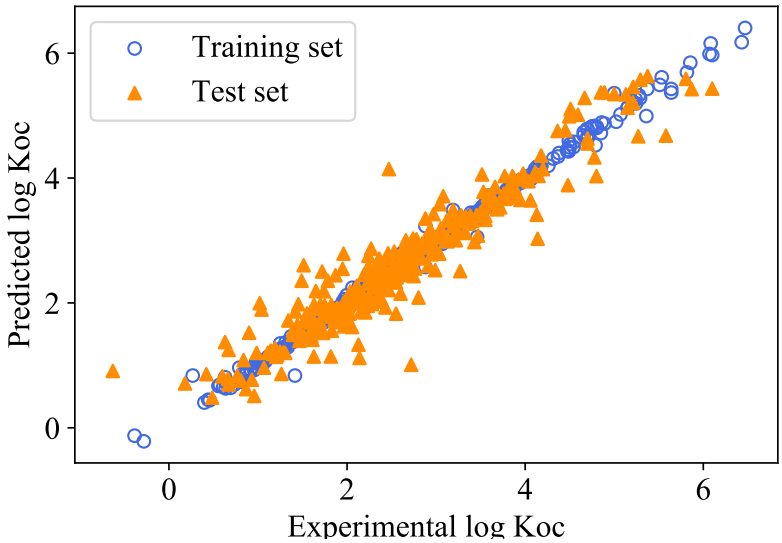


Figure 4.2: Plots of the experimental and predicted values of log K_{oc}

$$CCC = \frac{2\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)\left(y_i^{pred}-\bar{y}^{pred}\right)}{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)^2+\sum_{i=1}^{n}\left(y_i^{pred}-\bar{y}^{pred}\right)^2+n\left(\bar{y}^{obs}-\bar{y}^{pred}\right)^2} \tag{4.6}$$

where y_{obs}, and y_{obs} denote the experimental data and the predicted value of log K_{oc}, respectively; and ymeans and n denote the average value of log K_{oc} and the number of samples, respectively. The R^2 values for the training and test sets were 0.990 and 0.904, respectively, and the Q^2_{LOO} and R^2_{10fold} were 0.995 and 0.777, respectively. The plots of the predicted and measured values of log K_{oc} are shown in Figure 4.2.

The MLR and SVM models were developed using the same dataset and explanatory variables used for the GBDT based prediction model. In the MLR model, the regression equation of the prediction model was shown in the following equations::

$$logKoc = 0.4323LogP_{pred} + 0.0021SLogP\_VSA2 - 0.0031AATSC1v$$
$$+0.0774SLogP + 0564GATS1Z - 1.0836SIC0 + 1.7039 \quad (4.7)$$

The R^2 for the training set and test sets was relatively low (0.817 and 0.815, respectively), compared to the GBDT based prediction model. The Q^2_{LOO} and R^2_{10fold} were 0.818 and 0.656 shown in the Table 4.3.

We also compared four kernels for the SVM models: one linear kernel and three non-linear kernels, including a poly kernel, a radial basis function kernel, and a sigmoid kernel. Two parameters, C (regularization parameter) and gamma (the relative weight of the regression error), were optimized using a grid search. The SVM model with the RBF kernel (C = 100, gamma = 0.001) had the best prediction score, and the R^2 values for the training and test sets were also relatively low, 0.849 and 0.828, respectively. The Q^2_{LOO} and R^2_{10fold} were 0.847 and 0.692, respectively. The overall statistical parameters for the developed MLR and SVM models are shown in Table 4.5.

4.3.2 Applicability Domain

Evaluation of the AD of the dataset using the standardization approach was performed using a software called "AD using standardization approach." There were 23 compounds (Nos.2, 6, 8, 9, 10, 11, 13, 15, 20, 21, 51, 191, 252, 277, 292, 294, 421, 525, 526, 527, 609, 633, 634, 639) from the training set considered to be outliers. In addition, there were ten compounds (Nos.2, 6, 8, 9, 11, 13, 15, 20, 21, 51, 191, and 252) outside the AD. The results of the AD using the standardization

Table 4.5: Statistical parameters of the optimal and previous models.

Study	Algorithm	Nt	Np	No. of variables	R_t^2	R_p^2	Q_{LOO}^2	R_{10fold}^2	$RMSE_t$	$RMSE_p$
Shao et al. 2014	LS-SVM	643	321	4	0.904	0.846	0.840	-	0.344	0.431
	GA-MLR	644	320	4	0.817	0.808	0.813	-	0.490	0.475
	LLR	NA	NA	4	0.873	0.831	0.824	-	0.398	0.45
Olguin et al. 2017	LR	639	321	1 (ALOGPs)	0.85	0.809	0.849	-	0.428	0.48
	LR	639	321	1 (KOWWIN)	0.85	0.796	0.848	-	0.428	0.496
	LR	639	321	1 (XLOGP3)	0.85	0.79	0.848	-	0.428	0.504
OPERA	-	643	321	-	0.912	0.849	-	-	0.330	0.426
Current work	GBDT	643	321	6	0.99	0.904	0.995	0.777	0.110	0.335
	MRL	643	321	6	0.817	0.815	0.818	0.656	0.475	0.472
	SVM	643	321	6	0.849	0.828	0.847	0.692	0.432	0.455

Nt/Np: the number of compounds in the training set/test set; R_t^2: coefficient of determination for training set; R_p^2: coefficient of determination for test set; Q_{LOO}^2: correlation coefficient of leave-one-out cross validation; R_{10fold}^2: coefficient of determination of 10-fold cross validation; RMSEt: root mean square error of training set; RMSEt: root mean square error of test set

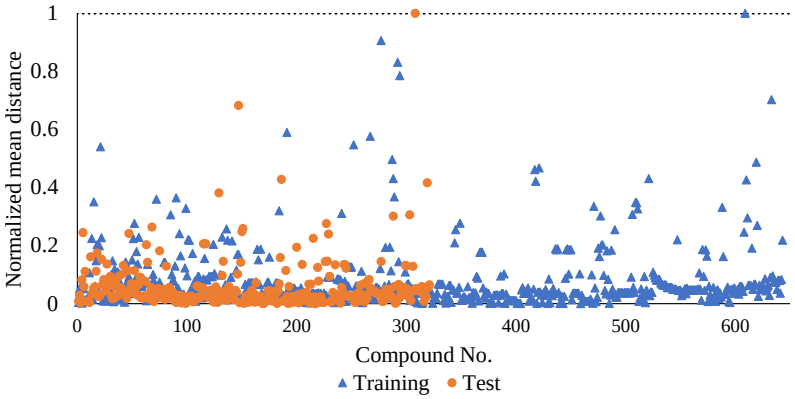
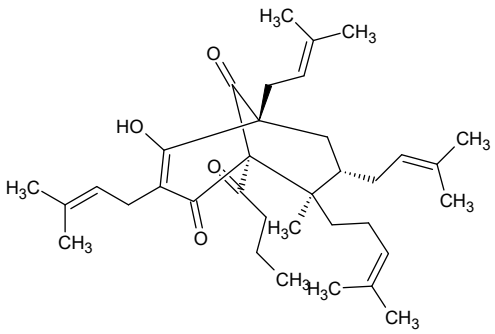


Figure 4.3: The plot of applicability domains by the Euclidean-Distance 1.0.

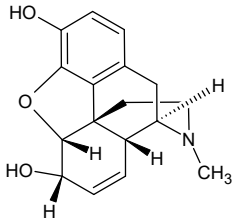
approach showed few outliers in the test set with a normal distribution pattern of approximately 99.7% of the population remaining within the range mean $\pm$ 3 standard deviation (SD).

Next, we implemented the Euclidean-Distance approach and generated a Euclidean graph using Euclidean-Distance 1.0, as shown in Figure 4.3. Only one compound in the training set (No.609; nicosulfuron) and one compound in the test set (No.308; morphine) were located outside the AD. Both compounds have a chiral center and cyclic structure in the molecule; the chemical structures of the two compounds are shown in the Figure 4.4.

Finally, we conducted OCSVM and generated a graph. Only seven compounds in the training set (Nos.1, 5, 126, 147, 151, 200, and 308) were located outside the AD. The graph is shown in the Figure 4.5. The three AD methods showed that the developed QSPR models were able to make reliable predictions.



609: nicosulfuron



308: morphine

Figure 4.4: The chemical structures of outliers determined by Euclidean-Distance 1.0

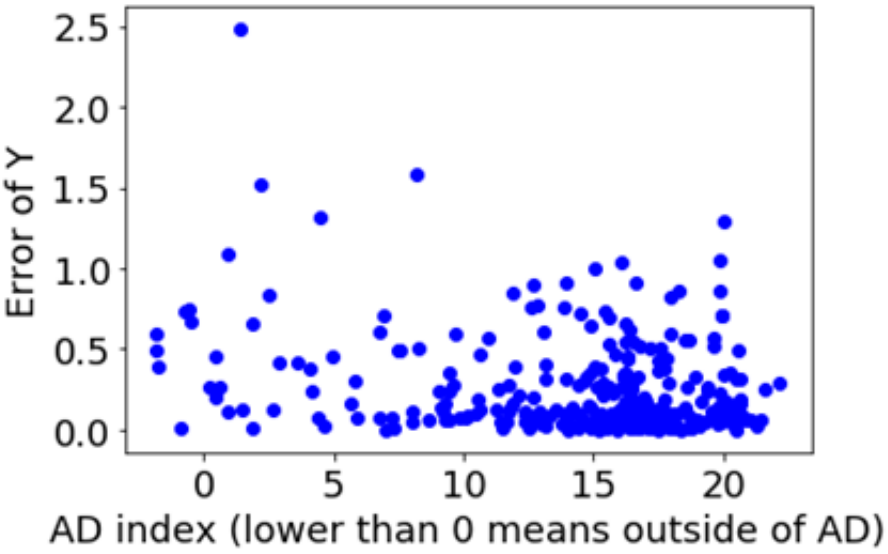


Figure 4.5: The plot of applicability domains by OCSVM

Table 4.6: Statistical parameters of external validation

Model	Q2F1	Q2F1	RMSEext	CCC_{ext}	$\overline{r_m^2}$	Δr_m^2
MLR	0.988	0.828	0.455	0.908	0.757	0.11
SVM	0.987	0.815	0.471	0.898	0.736	0.155
GBDT	0.993	0.903	0.341	0.95	0.862	0.038

4.3.3 Comparison of the developed models with OPERA and models in the previous studies

The statistical parameters of the three developed models are shown in Table 4.5. There exist a few chemicals whose prediction values from the MLR and SVM models demonstrated the best prediction ability. However, considering the statistical parameters of the models, the best value for all statistical parameters was demonstrated by the GBDT based prediction model. Thus, we concluded that the GBDT based prediction model was the superior model. External validation of the three models was also performed with the results shown in Table 4.6. The results presented include all values that were greater than the criteria shown in Table 4.3. These criteria were recommended in previous studies [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012, Roy et al., 2012].

We compared the prediction ability of our developed models with the calculated values of log K_{oc} from OPERA. The prediction abilities of the GBDT based prediction model and OPERA are listed in Table 4.5. The GBDT based prediction model demonstrated a better prediction ability than OPERA.

The prediction abilities of our developed QSPR models and models used in previous research [Shao et al., 2014, Oliphant, 2007] were compared and are shown in Table 4.5. Our GBDT based prediction model demonstrated the highest fitness, prediction accuracy, and robustness than the

Table 4.7: Features of QSPR models in current work and previous study

Study	Algorithm	Software for calculation of explanatory variables	Software for model development	Advantages of developed model compared with other models
Shao et al. 2014	LS-SVM	DRAGON 5.4 (Shareware)	Matlab/C toolbox (Shareware)	-Highest prediction ability -Only using freeware -No need for registration -Versatile programming language
Olguin et al. 2017	LR	ALOGPs algorithm (Freeware)	QSARINS (Freeware, Limited)	
Current work	GBDT	Mordred, OPERA (Freeware)	Python (Freeware)	

models reported in previous research. As a result, the physicochemical properties and molecular descriptors were found to be essential for the determination of K_{oc}. Therefore, this study shows that K_{oc} may be estimated from chemical structures without expensive and time-consuming laboratory studies. Based on these findings, the experimental cost for the determination of K_{oc} for chemical development may be reduced, along with the duration required for chemical development.

Furthermore, it is possible to use the developed model for preliminary environmental risk assessment during the early stages of chemical development. As such, we can make a determination as to whether the research and development of chemicals should be proceeded by QSPR models. This decision for the development project is currently determined by experts. Our models can provide objective justification for decisions associated with product development. The overall summary of the model performance is shown in Table 4.5. We have also summarized the features of all QSPR models developed in previous research and the current work in Table 4.7.

4.3.4 Comparison of the developed models with the previous chapter using experimental data

The model development procedure proposed in this work was compared to the procedure in the previous chapter using the experimental data of physicochemical properties. We developed a GBDT

Table 4.8: Statistical parameters of the various models including a model by experimental data of physicochemical properties for the 163 pesticides

Reference	Algorithm	Nt	Np	No. of variables	R_t^2	R_p^2	Q_{LOO}^2	RMSE
Gramatica et al. (2000)	GA-MLR	143	20	6	0.843	0.670	0.824	0.350
Huuskonen (2003)	MLR	143	20	12	0.820	0.790	0.790	0.370
Duchowicz et al. (2007)	MLR	143	20	6	0.900	0.710	0.890	0.290
Rinaldo dos Reis (2014)	MLR	143	20	4	0.852	0.743	0.840	0.343
Models in Chapter 3	GBDT	143	20	7	0.935	0.775	0.931	0.231
Current chapter	GBDT	143	20	6	0.897	0.729	0.910	0.290

based prediction model using properties calculated from OPERA and Mordred for 163 pesticides used in the previous chapter. The performance of our model and previous models are shown in Table 4.8. The model using experimental data of physicochemical properties demonstrated the highest fitness, prediction accuracy, and robustness. However, the models developed in this chapter showed better performance with some of the models in the previous chapter. Although the actual experimental data is considered to be the best choice for development for QSPR models, our proposed procedure is a good substitute for actual values if these are difficult to obtain.

4.4 Conclusion

In this chapter, we have improved prediction models for K_{oc} values by using only calculated values. We have also proposed GBDT based prediction model as in the previous chapter used physicochemical properties and molecular descriptors quantified by OPERA and Mordred for a large dataset. As a result, the following results were obtained.

- By using both the physicochemical properties and molecular descriptors calculated from structural formulas, the model prediction accuracy was considerably increased as compared

with that of the model described in the previous chapter.

- By using the GBDT algorithm, the prediction accuracy was further improved.
- The prediction models based on the GBDT algorithm exhibited the best prediction abilities among various machine learning models.
- The performance of the model developed in this chapter was much higher than those of the models developed in the previous chapter.

Although the model using the experimentally determined physicochemical properties demonstrated a good fit, high prediction accuracy, and robustness, the method proposed in this chapter can be used instead of the actual values if the latter are difficult to acquire. The results contribute to the establishment of a new chemical development process with quick and easy procedure.

Chapter 5

Prediction of fish bioconcentration factors using calculated physicochemical properties and molecular descriptors

5.1 Introduction

When applying for the registration of chemicals such as pesticides, risk and hazard assessments must be performed to evaluate the impact not only on humans and animal species but also on the environment [Danaei et al., 2005, Damalas and Eleftherohorinos, 2011, EFSA, 2013]. Both the application method and the quantity of the chemical substances to be applied are considered in the risk assessment [van der Oost et al., 2003, Hernando et al., 2006]. On the other hand, the hazard assessment is an index that is restricted to evaluating potential toxicity and hazards [Klopffer, 1994, Henschel et al., 1997].

BCFs are widely used criteria for hazard assessment [Arnot and Gobas, 2006] that represent the ratio of the concentration of the chemical in the fish to the concentration of the chemical substance in the water [Mackay and Fraser, 2000]. The higher the value of BCF, the more likely the chemical is to concentrate in the organism. The conceptual diagram of BCF was shown in the Figure 5.1.

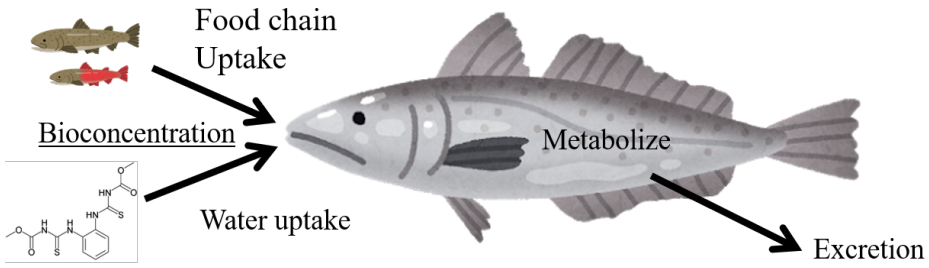


Figure 5.1: Conceptual diagram of Bioconcentration

Various bioconcentration properties and bioconcentration factors used as indicators have been considered due to differences in the exposure routes and evaluation status [Miyamoto et al., 1990, OECD, 2012]. The subjects of evaluation are aquatic, benthic, and soil organisms. Among these, the most widely evaluated species are fish. The establishment of safe daily consumption limits for fish and pesticide concentrations in water, using BCF, aids in the estimation of daily pesticide intake by the organism. Thus, it is crucial to investigate the BCF of pesticides.

Due to the high level of skill required to determine the BCF and the need to conform to strict guidelines [OECD, 2012], only a limited number of research institutions can conduct BCF tests. Moreover, an experiment to obtain the BCF of a single compound is a costly undertaking and requires a period of several months to half of a year to complete. In some cases, it may cost $125,000 [Weisbrod et al., 2007]. Furthermore, since the BCF classification and criteria are regulated internationally, including countries in Europe, the United States, and Japan, the registration of the pesticide may be limited or prohibited if the criteria are exceeded [Moss et al., 2000, EuropeanComission, 2009, Markell, 2010]. Therefore, the development of an accurate prediction model for estimating BCFs in the early stage of pesticide development is considered necessary to facilitate efficient research and development.

QSAR is a quick and inexpensive method of evaluating the toxicity of a compound without the need to perform actual experiments [Hansch and Leo, 1995]. Various QSAR models of toxicity have been developed in the past couple of decades [Kubinyi, 1997, Tropsha, 2010, Halder et al., 2018]. Some linear BCF QSAR models were developed by Devillers et al.[Devillers et al., 1996] , Papa et al. [Papa et al., 2007], and Garg et al. [Garg and Smith, 2014]. Gissi et al. attempted to predict the BCF of 851 compounds, as reported in the ANTARES (Alternative Non-Testing methods Assessed for REACH Substances) BCF dataset [Gissi et al., 2015]. The most widely used BCF prediction models, i.e., CAESAR and Meylan, were utilized to develop a more reliable integrated model for predictions. Pramanik and Roy developed two models for BCF prediction that include multiple linear regression algorithms and partial least squares analysis [Pramanik and Roy, 2014]. These models were based on a training set that included 324 compounds. The models were applied to the test set with 198 compounds to verify performance. Additionally, many other QSAR models to predict the BCFs of various compounds have also been developed over the last 20 years [Gramatica and Papa, 2005, Pavan et al., 2008, Nolte and Ragas, 2017].

It is believed that the process of bioaccumulation is impacted by physicochemical properties such as the molecular size, fat- and water-solubility of the compound, and the biological characteristics of the organisms such as species and size [Veith et al., 1979, Connell, 1988, Garg et al., 2014]. However, to obtain the physicochemical properties, experiments are basically needed. As a result, many QSAR models have been developed by calculating molecular descriptors using cheminformatics software such as PaDEL-Descriptor [Yap, 2011] and DRAGON [Mauri et al., 2006] and using molecular descriptors as explanatory variables [Zhao et al., 2008, Pramanik and Roy, 2014, Toropova et al., 2020].

The most frequently reported models using physicochemical properties are linear and non-linear models using n-octanol/water partition coefficient (log P_{ow}) [Connell and Hawker, 1988, Bintein et al., 1993, ECHA, 2017]. The bioconcentration of organic compounds in fish mainly depends on the hydrophobicity of the compounds [Devillers et al., 1996], and log P_{ow} is closely related to the BCF. Furthermore, log P_{ow} is one of the essential physicochemical properties required when registering a chemical substance and can be calculated using a simple experimental method [Klein et al., 1988]. Thus, log P_{ow} was used in many QSAR models. There are few reports of QSAR models using other physicochemical properties [Isnard and Lambert, 1988, Pavan et al., 2008]. In the chapter 3 and 4, we have achieved to develop good prediction models by using molecular descriptors and physicochemical properties. In order to confirm the generality and versatility of the proposed method in the previous chapters, we have applied the method to the model development of the BCF which is one of the important parameters of environmental risk/hazard assessment.

We have developed a machine learning based prediction model with experimental and calculated physicochemical properties and of pesticides. The physicochemical properties were collected from the EFSA peer review report and the U.S. EPA's Chemistry Dashboard. Despite the development of a high-performance model in terms of accuracy compared to that in the previous study, it is difficult and time-consuming to collect experimental data for diverse chemicals. In this study, we propose to develop an accurate prediction model by gathering physicochemical properties using a simple method employing a freely available software.

The U.S. EPA released OPERA, which is a software that predicts the physicochemical properties and environmental fate endpoints [Mansouri et al., 2018]. OPERA was developed using data collected from the PHYSPROP database and can predict various physicochemical properties and

environmental fate endpoints. It is a model constructed using a wide range of chemical substances, and its performance is good; the R^2 test ranges from 0.71 to 0.96 (average 0.82). Thus, it is believed that a highly accurate prediction model can be developed by machine learning using the properties calculated by OPERA in addition to the molecular descriptors. Furthermore, BCF results can be predicted before the start of expensive long-term experiments using the developed model, which will contribute to the decision making for chemical substance development.

In this chapter, we collected physicochemical properties, environmental fate and toxicology estimated value calculated by OPERA and molecular descriptors calculated by Mordred [Moriwaki et al., 2018]. Then, by incorporating the calculated values, we developed a QSAR model for the dataset of chemicals used in the previous research [Pramanik and Roy, 2014]. Moreover, we compared our developed models with previously reported models and the BCF value calculated by OPERA was used as a reference.

5.2 Material and methods

5.2.1 Dataset

We used the dataset of 522 chemicals from previous studies [Fernández et al., 2012, Pramanik and Roy, 2014]. The reason why we used the data set is that it includes a large number of chemicals. In addition, we used the dataset because we intended to check if the combination of molecular descriptor and physicochemical properties would perform better than previous models. In these studies, 60% (324) and 40% (198) of the chemicals were used as the training and test sets. We use same way of splitting in accordance with previous studies. The test set was suitable for model validation because

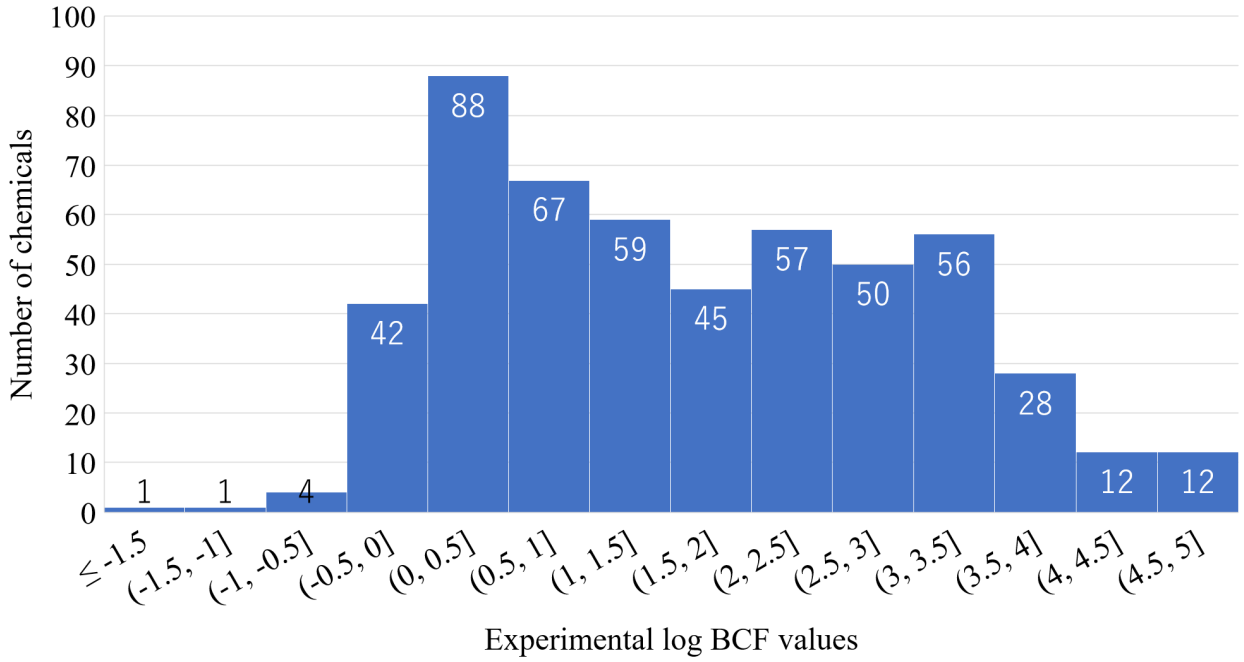


Figure 5.2: Histogram of distribution of experimental log BCF values in the dataset

it includes a wide range of chemicals. The chemical datasets were classified into various categories: aliphatic and aromatic hydrocarbons, alcohols, anilines, amides, amines, cyanides, esters, ethers, halogenated derivatives, heteroaromatics, nitriles, nitroaromatics, organochlorines, organophosphates, phenols, phosphate esters, sulfonic acids, and thiols. The industrial classifications of these chemicals were as follows: agrochemicals, industrial chemicals, pharmaceuticals, plant secondary metabolites, and pollutants. The logarithm of BCF (log BCF) was used for the QSAR models. The log BCF values ranged from –1.70 to 4.85 and -0.73 to 4.85 for the training and test sets, respectively. The distribution of the dataset is shown in Figure 5.2.

5.2.2 Software and program

Python 3.7 was used as the programming language. Python modules such as Matplotlib, NumPy, and SciPy were used for calculations and visualization. Scikit-learn was used for implementing linear regression and SVM models. We optimized the SVM model using a grid search method implemented in scikit-learn. XGBoost was used for the development of the gradient boosting decision tree model.

We have conducted the standardization for the molecular structure of the obtained compounds by MMFF 94 Force Field (mmff 94) using Open Babel [O'Boyle et al., 2011] before the calculating the molecular descriptors. A total of 1,826 molecular descriptors were calculated using Mordred [Moriwaki et al., 2018].

Table 5.1: Representative descriptors calculated by Mordred

Descriptor type	Number	Representative descriptor
Acidic group count	1	nAcid
ALOGP	3	ALogP, ALogp2, AMR
Atom count	14	nAtom, nHeavyAtom, nH, nB, nC, nN, nO, nS, nP, nF
Molecular linear free energy relation	6	MLFER_A, MLFER_BH, MLFER_BO, MLFER_S, MLFER_E
Rule of five	1	LipinskiFailures
Topological	3	topoRadius, topoDiameter, topoShape
Topological distance matrix	11	SpMax_D, SpDiam_D, SpAD_D, SpMAD_D, EE_D, VE1_D
3D autocorrelation	80	TDB1u, TDB2u, TDB3u, TDB4u, TDB5u, TDB6u, TDB7u
WHIM	91	L1u P1u, E1u, Tu, Au, Du, L1m, P1m, E1m, Km, Dm

We gathered predictive physicochemical properties and environmental fate estimated value for the dataset using OPERA, which is a standalone free and open-source software for predicting

physicochemical properties and environmental fate endpoints developed by the U.S. EPA. OPERA can predict the properties of chemicals based on PaDEL-descriptors, which were available in MATLAB, C, and C++ languages. We calculated 46 properties for the dataset using OPERA. The calculated properties are shown in Table 5.2.

Table 5.2: The parameters calculated by OPERA

Type	Description
MolWeight	Molecular weight
nbAtoms	Number of atoms
nbHeavyAtoms	Number of heavy atoms (i.e. not hydrogen)
nbC	Number of carbon atoms
nbO	Number of oxygen atoms
nbN	Number of nitrogen atoms
nbAromAtom	Number of aromatic atoms
nbRing	Number of rings
nbHeteroRing	Number of rings containing heteroatoms (N, O, P, S, or halogens)
Sp3Sp2HybRatio	Fraction of sp3 carbons to sp2 carbons
nbRotBd	Number of rotatable bonds, excluding terminal bonds
nbHBdAcc	Number of hydrogen bond acceptors (using CDK HBondAcceptorCount Descriptor algorithm)

ndHBdDon	Number of hydrogen bond donors (using CDK HBondDonorCount Descriptor algorithm)
nbLipinskiFailures	Number failures of the Lipinski's Rule Of 5
TopoPolSurfAir	Topological polar surface area
MolarRefract	Molar refractivity
CombDipolPolariz	Combined dipolarity/polarizability
BP	Boiling Point at 760 mm Hg
HL	Henry's Law constant (air/water partition coefficient) at 25 °C
KOA	The octanol/air partition coefficient.
LogP	Octanol-water partition coefficient
MP	Melting Point
VP	Vapor Pressure
WS	Water solubility at 25 °C
RT	HPLC retention time.
Pka	Logarithmic (acid) dissociation constant
logD	Octanol-water distribution coefficient
LogBCF	Fish bioconcentration factor
AOH	OH rate constant for the atmospheric, gas-phase reaction between photochemically produced hydroxyl radicals and organic chemicals

BioDeg	biodegradation half-life for compounds containing only carbon and hydrogen
RBioDeg	Ready biodegradability of organic chemicals
KM	The whole body primary biotransformation rate (half-life) constant for organic chemicals in fish.
KOC	soil adsorption coefficient of organic compounds.
CERAPP-Binding	Collaborative Estrogen Receptor Activity Prediction Project. Binding consensus
CERAPP-Agonist	Collaborative Estrogen Receptor Activity Prediction Project. Agonist consensus
CERAPP-Antagonist	Collaborative Estrogen Receptor Activity Prediction Project. Antagonist consensus
CoMPARA-Binding	Collaborative Modeling Project for Androgen Receptor. Binding consensus
CoMPARA-Agonist	Collaborative Modeling Project for Androgen Receptor. Agonist consensus
CoMPARA-Antagonist	Collaborative Modeling Project for Androgen Receptor. Antagonist consensus
CATMoS-VT	Collaborative Acute Toxicity Modeling Suite. very_toxic LD50 $\leq$ 50 mg/kg vs LD50 >50 mg/kg

CATMoS-NT	Collaborative Acute Toxicity Modeling Suite. Nontoxic. LD50 >2000 mg/kg vs LD50 ≤ 2000 mg/kg
CATMoS-EPA	Collaborative Acute Toxicity Modeling Suite. EPA_categories. 1 is LD50 ≤ 50 mg/kg; 2 is LD50 >50 to ≤ 500 mg/kg; 3 is LD50 >500 to LD50 ≤ 5000 mg/kg; 4 is LD50 >5000 mg/kg
CATMoS-GHS	Collaborative Acute Toxicity Modeling Suite. GHS_categories. 1 is LD50 ≤ 5 mg/kg; 2 is LD50 >5 to ≤ 50 mg/kg; 3 is LD50 >50 to LD50 ≤ 300 mg/kg; 4 is LD50 >300 to LD50 ≤ 2000 mg/kg; 5 is LD50 >2000 mg/kg
CATMoS-LD50	Collaborative Acute Toxicity Modeling Suite. LD50 point estimate model
FuB	Human plasma fraction unbound
Clint	Human hepatic intrinsic clearance
MolWeight	Molecular weight
nbAtoms	Number of atoms
nbHeavyAtoms	Number of heavy atoms (i.e. not hydrogen)
nbC	Number of carbon atoms

nbO	Number of oxygen atoms
nbN	Number of nitrogen atoms
nbAromAtom	Number of aromatic atoms
nbRing	Number of rings
nbHeteroRing	Number of rings containing heteroatoms (N, O, P, S, or halogens)
Sp3Sp2HybRatio	Fraction of sp3 carbons to sp2 carbons

5.2.3 Model development and validation

The GBDT based prediction model was developed as a non-linear model and an ensemble algorithm. This GBDT based prediction model was developed using physicochemical properties, environmental fate endpoints, and molecular descriptors for the chemicals in the dataset. Gradient boosting consisted of gradient descent and boosting methods developed by Friedman [Friedman, 2001]. The boosting algorithm was a part of ensemble learning and was used to integrate multiple weak learners, such as decision trees, to build the entire learner. The use of a decision tree as a weak learner in boosting had advantages such as good resistance to outliers in the data and the ability to withstand discrete variables and missing values. Accordingly, the boosting algorithm is one of the most popular algorithms in data analysis competitions, such as KDD Cup and Kaggle. Since the tree models are also generally excellent in readability, it was possible to determine the explanatory variables that contributed to the developed model. XGBoost, which is a machine-learning package

for the GBDT algorithm [Chen and Guestrin, 2016], was executed in Python. XGBoost had a very scalable end-to-end tree boosting system. XGBoost employed an algorithm that determined the direction of tree branching in advance for sparse data, i.e., when there are many missing values. The model search was accelerated by parallel distributed processing. XGBoost adopted an algorithm that determined the branch direction of a tree in advance for sparse data, including many missing values, and model retrieval was accelerated by parallel distributed processing.

Parameter tuning was performed by a grid search to determine the optimum parameters for the GBDT based prediction model. The grid search was performed in 5-fold cross-validation on the training set and evaluated by R^2 value. The values of maxdepth, min child weight, n estimators, and regalpha were optimized. Grid search is a conventional, reliable, and rapid approach that was used to determine the optimum parameter due to quick algorithms such as GBDT.

A total of 1,826 descriptors and 46 physicochemical properties and environmental fate endpoints were generated using Mordred and OPERA. The descriptors for all 522 chemicals were calculated within 20 second using Mordred. The physicochemical properties and environmental fate endpoints were calculated in 35 min using OPERA. Majority of the molecular descriptors and properties were not essential for the calculation of log BCF. Therefore, we selected the descriptors that were relevant to the predictability of the developed model using the feature importance implemented in the GBDT. The tree-based model was used to determine the variable and threshold value that maximized the model at each branch. We adopted gain as an indicator, which is the default and critical parameter for feature selection.

Linear regression models based on the multiple linear regression (MLR) and support vector machine (SVM) were developed using molecular descriptors, physicochemical properties, and

environmental fate endpoints selected by feature selection. We compared these models with the GBDT based prediction models.

The performance of the developed model was evaluated based on the OECD principles for model validation [Gramatica, 2007]. The prediction ability, fitting performance, and robustness of the model were evaluated using R^2, the leave-one-out cross-validation of correlation coefficient (Q^2_{LOO}), the coefficient of multiple determinations of 10-fold cross-validation (R^2_{10fold}), and the root-mean-square error (RMSE). Other detailed definitions and calculations of the parameters are provided in references [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012].

5.2.4 Applicability Domain

We evaluated the applicability domain (AD) of the dataset using same approach in the previous chapter. We adopted two approaches for evaluating AD. The first was a standardization approach using the AD using standardization approach program [Roy et al., 2015]. The second approach was the Euclidean distance-based method using Euclidean-Distance 1.0 software [Ambure et al., 2015].

$$dij = \|Xi - Xj\| = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2} \tag{5.1}$$

$$\overline{d_i} = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \tag{5.2}$$

$$\text{MeanDistance}_{\text{ (Normalized)}} = \frac{(\text{ MeanDistance } - \text{ Min\_MeanDistance })}{(\text{ Max\_MeanDistance } - \text{ Min\_MeanDistance })} \tag{5.3}$$

where, d_{ij} is the distance score between two compounds; and $\overline{d_i}$ is the mean distance.

5.2.5 Comparison of the developed models with OPERA and models in previous studies

The EPI (Estimations Programs Interface for Windows) suite was also used for comparison with the model developed in this study.

The predictive values of log BCF calculated by OPERA were also compared with our developed model. OPERA contained a model for the direct prediction of log BCF values. These data yielded by the OPERA were from the PHYSPROP, which is a collection of a wide variety of sources built by Syracuse Research Corporation. Experimental protocols for the different components of the data were traced to the originally referenced literature from the database. PHYSPROP's BCF data were collected from Arnot and Gobas [Arnot and Gobas, 2006]. The original data collected from the PHYSPROP database underwent a series of processes to curate the chemical structures and remove duplicates. We compared the predictive abilities of the QSAR model and models developed in the previous study by Pramanik [Pramanik and Roy, 2014]. They developed two models using genetic function approximation followed by multiple linear regression (GFA-MLR) [Rogers and Hopfinger, 1994] and subsequent partial least squares (PSL) regression models [Eriksson et al., 2013].

Table 5.3: Statistical parameters of the developed models using physicochemical properties, environmental fate estimated value, and molecular descriptors

No.	Model	No. of variables	R^2	R^2_{PRED}	Q^2_{LOO}	R^2_{10fold}	$RMSE_t$	$RMSE_p$
1	Physicochemical properties and e-fate estimated value	5	0.927	0.833	0.907	0.802	0.368	0.545
2	Molecular descriptors	7	0.91	0.831	0.896	0.763	0.409	0.549
3	1 + 2	5	0.923	0.863	0.917	0.815	0.378	0.494

5.3 Results and discussion

5.3.1 Result of the developed models

The standard parameters for the developed models were optimized via a grid search using the dataset. We obtained the optimum values of R^2 and RMSE for both the training and test datasets using maxdepth, min child weight, n estimators, and gamma at measures of 30, 15, 200, and 0.001, respectively. Several studies have suggested that R^2 values should be greater than 0.7 and RMSE values should be as low as possible [Chirico and Gramatica, 2011, Chirico and Gramatica, 2012].

We used 324 chemicals as a training set and 198 chemicals as the test set, as outlined in a previous study [Pramanik and Roy, 2014]. A model using molecular descriptors calculated by Mordred as an explanatory variable demonstrated excellent performance compared with the previous model (R^2 > 0.8). However, the models that used both values calculated by OPERA and Mordred showed the best parameters. The statistical parameters of the model are shown in Table 5.3.

We selected the descriptors using the feature importance. There are three indicators of feature importance: cover, gain, and weight.Cover is the sum of the quadratic gradients of the training data classified into leaves. The squared loss corresponds to the number of instances on that branch and gain is an indicator of the extent to which the evaluation criteria can be improved. Weight is a

measure of the number of times the feature was used to split the data across all trees to observe only the existing number. There is no information on the proximity of the branch to the prediction or the number of branches used for input. We adopted gain as an indicator.

After conducting the feature importance selection of the explanatory variable, six values were selected. The selected physicochemical properties and environmental fate endpoints were LogD74 and LogKM_pred, which are listed in Table 5.2. LogD74 is a predictive logarithmic distribution coefficient at pH 7.4. Log D represents a measure of lipophilicity at the physiologically relevant pH. Log P represents the concentration ratio at the neutral species, whereas log D is defined as the total concentration of all charge-state forms of the substance dissolved in the lipid (octanol) phase divided by the total concentration dissolved in water at a selected pH. Thus, log D is a more relevant parameter for describing the biological effects of chemicals because it considers ionization at the relevant pH. LogKM_pred is the logarithmic predictive fish biotransformation half-life.

The fish biotransformation half-life is closely related to the bioconcentration process [Papa et al., 2007]. Of the 1,826 descriptors calculated by Mordred, the three molecular descriptors that were selected for model development are presented. The most important descriptor was XLogP, which denotes the calculated value of log *P*. XlogP was calculated by decomposing the molecule into individual atoms and calculating the total contribution of each. A correction term was included in the calculation. The value of Log *P* published in PubChem is based on XLogP. The next important descriptor was TopoPSQ (NO), which is a topological polar surface area. Last, the fMF is a descriptor characterizing the complexity of a molecule. The fMF was described in a previous report [Yang et al., 2010] and is an approach used for characterizing molecular complexity based on the Murcko framework present in the molecule. The descriptor is the ratio of heavy atoms per the total number of atoms in

the molecule. By definition, acyclic molecules that have no frameworks will have a value of zero. Last, the FilterItLogS is the calculated value of the base 10 logarithm of the solubility (log *S*) via Filter-it™.

The selected physicochemical properties, environmental fate endpoints, and molecular descriptors are consistent with the results of previous studies. The bioconcentration of a chemical substance in fish is related to mechanisms and properties such as the molecular size, the fat- and water-solubility of the compound, and the biological characteristics of the organisms such as species and size [Veith et al., 1979, Connell, 1988, Garg and Smith, 2014]. The accuracy of the models was evaluated by calculating R^2 , Q^2_{LOO}, R^2_{10fold}, and RMSE. Each expression is detailed as follows:

$$\boldsymbol{R}^2 = \mathbf{1} - \frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{\sum_{i=1}^{n} \left(y_{obs} - y_{\text{means}}\right)^2} \tag{5.4}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{obs} - y_{pred}\right)^2}{n}} \tag{5.5}$$

where y_{pred} denotes the predicted value of log BCF and y_{obs} denotes the experimental data of log BCF Additionally, y_{means} denotes the average value of log BCF, and n denotes the number of samples. The R^2 values for the training and test sets were 0.923 and 0.863, respectively. Q^2_{LOO} and R^2_{10fold} were 0.917 and 0.815, respectively. The predicted values calculated via the developed model are shown in Table 5.4. Additionally, plots of predicted values and measured values of log BCF are shown in Figure 5.3.

An MLR model was developed using the same dataset, and explanatory variables were used in GBDT based prediction model. The regression equation of the prediction model is shown in the

Table 5.4: Statistical parameters of the GBDT, MLR, and SVM based prediction models

Model	R_t^2	R_p^2	Q_{LOO}^2	R_{10fold}^2	$RMSE_t$	$RMSE_p$
GBDT	0.923	0.863	0.917	0.815	0.378	0.494
MLR	0.727	0.757	0.739	0.704	0.712	0.658
SVM	0.783	0.805	0.808	0.75	0.635	0.589

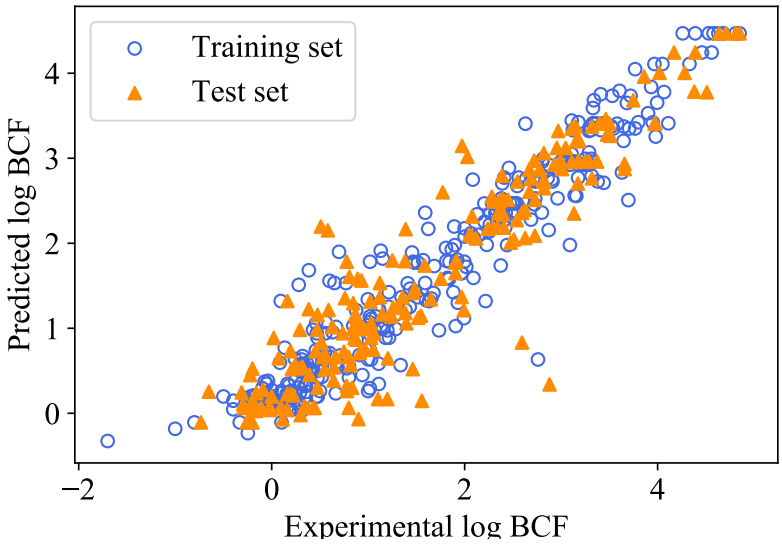


Figure 5.3: Plot of the experimental data and predicted values of log BCF by GBDT based prediction model

following equations:

$$logBCF = 0.1615LogD74 - 0.7043LogKMpred + 0.0327XLogP$$
$$-0.0061TopoPSA + 1.1044fMF + 0.0604FilterItLogS + 1.1595 \quad (5.6)$$

The R^2 values for the training and test sets were relatively low (0.727 and 0.757, respectively) compared to the GBDT based prediction model. Q_{LOO}^2 and R_{10fold}^2 were 0.739 and 0.704. Statistical parameters for the developed MLR model are shown in Table 5.4.

The SVM model was also developed using the same dataset. We compared linear and non-linear

kernels; poly kernel, radial basis function (RBF) kernel, and the sigmoid kernel were used as the non-linear kernels. Two parameters, C (regularization parameter) and gamma (the relative weight of the regression error), were optimized by the grid search. Based on the result of the grid search, SVM using the RBF kernel (C = 100, gamma = 0.001) exhibited the optimal score. The R^2 values for the training and test sets were relatively low (0.783 and 0.805, respectively) compared to the values obtained using the GBDT based prediction models. Q^2_{LOO} and R^2_{10fold} were 0.808 and 0.750, respectively. Statistical parameters for the developed SVM model are shown in Table 5.4.

5.3.2 Applicability Domain

We evaluated the Applicability Domain (AD) using two approaches: the standardization approach and the Euclidean-Distance approach. We confirmed that only four chemicals (Nos.29, 125, 196, and 278) in the training set compounds were outliers, and one chemical (No.90: 6,6'-Ureylene-Bis (1-Naphthol-3-Sulfonic Acid)) in the test set lied outside the applicability domain. The results by AD using the standardization approach showed that there are few outliers in the training set with the regular distribution pattern in which 99.7% of the population remained within the range of mean $\pm$ 3 SD. The Euclidean graph generated by Euclidean-Distance 1.0 was prepared and shown in Figure 5.4. The Euclidean distance is an ordinary distance between two points in the Euclidean space. Based on the normalized mean distances in the graph, only one training compound (No.125) and one test compound (No.90) were located outside the AD. The two AD methods showed an almost identical result, and the QSAR model developed in this study can make predictions with excellent reliability.

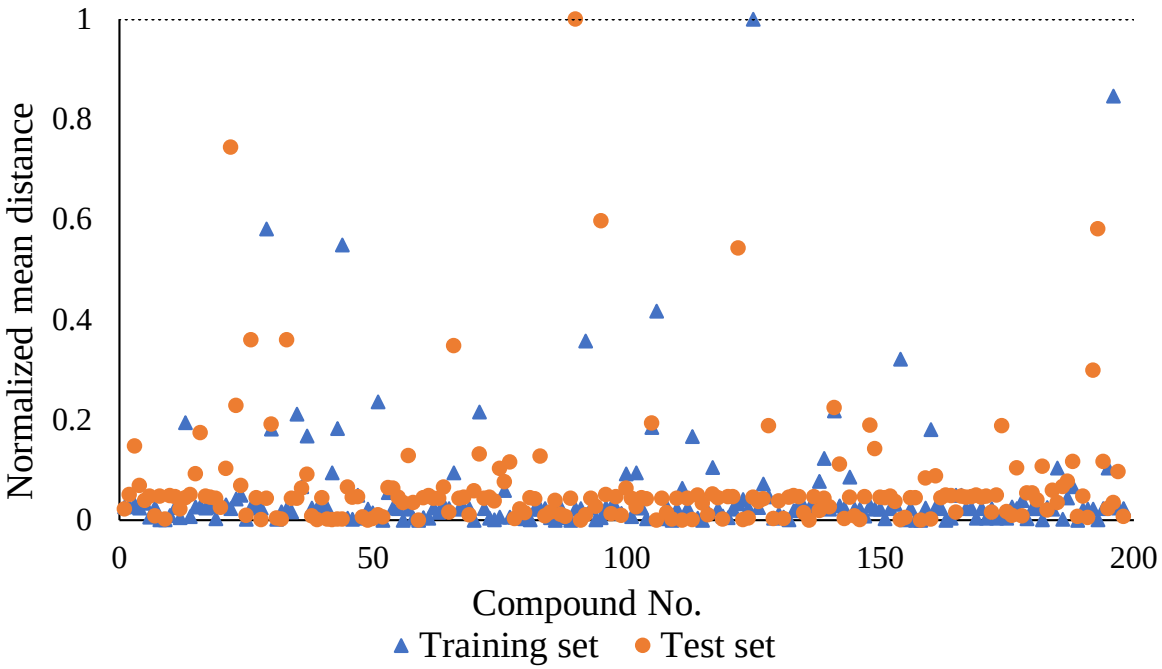


Figure 5.4: Plot of Euclidean-Distance applicability domain

5.3.3 Comparison of the developed models with OPERA and models in the previous studies

We compared the calculated log BCF to evaluate the performance of the three developed models. The statistical parameters are shown in Table 5.4. Some values that were predicted by the MLR and SVM models exhibited the greatest accuracy. However, considering the statistical parameters shown in Table 5.4, we obtained the best values for all statistical parameters by the GBDT based prediction model; hence, we concluded that GBDT algorithm was the best of the three models.

We compared the predictive ability of our developed model with the calculated values of log BCF using the EPI suite and OPERA. EPI suite is a model that was jointly developed by the U.S. EPA and Syracuse to predict and calculate the physicochemical properties of each substance based on its

chemical structure [Card et al., 2017]. This software has been used for risk assessment in various fields, such as in the examination of new chemical substances under the Toxic Substances Control Act (TSCA) when products containing new chemical substances are imported into the United States. The EPI suite incorporated a model that predicts various parameters, such as bioconcentration and biodegradability. BCFBAF™, formerly called BCFWIN™, is a subset of the EPI suite. This program estimates the logarithm of the fish BCF by two different methods. The first estimation method is the regression model based on log Pow and any relevant correction factors; this method is analogous to the WSKOWWIN™ method. The second method is the Arnot-Gobas method, which calculates the BCF from mechanistic first principles. The predictive abilities of the GBDT based prediction model, EPI suite, and OPERA are shown in Table 5.5. OPERA demonstrated good predictive ability compared to the EPI Suite. Although the EPI suite and OPERA possess a more extensive chemical space compared to our developed models, our GBDT based prediction model exhibited better predictive ability than the EPI Suite, and is compare with OPERA.

Table 5.5: Overall summary of statistical parameters for all QSPR models

Model	No. of variables	R_t^2	R_p^2	Q_{LOO}^2	R_{10fold}^2	$RMSE_t$	$RMSE_p$
Pramanik et al., 2014 (Model 1)	7	0.641	0.659	0.62	-	-	0.78
Pramanik et al., 2014 (Model 2)	4	0.614	0.696	0.597	-	-	0.738
CAESAR reported by Pramanik et al.	-	-	0.828	-	-	-	-
TEST reported by Pramanik et al.	-	-	0.83	-	-	-	-
OPERA	-	0.798	0.885	-	-	-	0.463
EPI Suite (BCFBAF™)	-	0.618	0.661	-	-	-	0.669
MLR	5	0.725	0.761	0.739	0.706	0.715	0.652
SVM	5	0.783	0.805	0.808	0.75	0.635	0.589
GBDT	5	0.923	0.863	0.917	0.815	0.378	0.494

Table 5.5 compares the predictive abilities of our developed QSAR models with the models used in previous studies reported by Pramanik [Pramanik and Roy, 2014]. Our developed GBDT based

prediction model showed higher predictive accuracy, fitness, and robustness than those reported in the previous study. The results indicated that the physicochemical properties, environmental fate endpoints, and molecular descriptors are important for log BCF estimation. Therefore, this study demonstrated that log BCF could be determined from the chemical structure without expensive laboratory studies on log BCF. On the basis of this finding, the experimental cost for BCF calculation as required for chemical development, may be significantly reduced and shorten the development period.

Additionally, we can apply the developed model for preliminary environmental risk assessment, whereby we can determine whether the development of the chemical should be pursued or aborted by the QSAR models based on the chemical structure. Though the decision of the project for development is currently determined via expert heuristics, these results provide objective justification for the use to guide product development. An overall summary of the models is shown in Table 5.5.

5.4 Conclusion

In this chapter, we have developed prediction models for the estimation of BCF. We have used calculated physicochemical properties by OPERA as in the previous chapter because the QSPR models for K_{oc} developed in the previous chapter showed high performance comparing to the previous models. As a result, the following results were obtained.

- By using both the physicochemical properties and molecular descriptors calculated from structural formulas, the model prediction accuracy was considerably improved.
- By using the GBDT algorithm, the model prediction accuracy was further increased.

- The prediction models using the GBDT algorithm demonstrated the best prediction abilities among various machine learning models.

The results suggest that by using molecular descriptors, physicochemical properties, and environmental fate endpoints as explanatory variables, we have developed a high-performance prediction model in terms of accuracy that is comparable to the existing model as well as QSPR models for K_{oc} in the previous chapter. Thus, we showed the generality and versatility of the procedure developed in the chapter 3 and 4.Our proposed procedure is applicable for development of prediction models for various parameters for environmental risk assessment. A preliminary environmental risk assessment can be performed without the need to perform time-consuming experiments. Consequently, the developed models can significantly contribute to the development process of new chemicals. This work is not only an effective proposal for the BCF evaluation method, but also evaluated the applicability of the data mining method.

Chapter 6

Discussion

In previous research studies, different predictive models for the toxicity and environmental effects of various chemicals were developed; however, no research works have focused on the risk and hazard assessment of pesticides using the data obtained from EU pesticide evaluation reports and physicochemical properties calculated by OPERA software. The results obtained in the present work revealed that it was possible to predict whether a particular pesticide would be registered or not by building an evaluation report database using publicly available pesticide application information and constructing an appropriate prediction model. In addition, the data contained in the officially issued evaluation reports have uniform standards and are highly reliable.

Therefore, it can be concluded that it is possible to predict environmental risk/hazard assessment parameters related to the registrability of pesticides using publicly available information and experimentally determined physicochemical properties and constructing an appropriate prediction model.

This study will serve as a model for predicting the outcomes of environmental tests and other

parameters involved in the registration of pesticides, facilitating the decision-making process regarding pesticide registration at an early stage. Compared to the traditional empirical and logic-based approaches utilized in drug discovery, our research is expected to strongly contribute to the rapid and efficient development of pesticides by reducing the timeframe required for issuing recommendations by R&D organizations and related costs.

Chapter 7

Conclusion and future works

Pesticides are artificially synthesized biologically active chemicals that are applied to agricultural crops to control pests and diseases. Because pesticides remain in agricultural products and people may ingest them unknowingly over a long period via food, it is necessary to ensure the safety of these products. For this reason, authorities in all countries across the globe have established pesticide registration systems based on laws and regulations, which require conducting numerous safety tests and risk and hazard assessments. Some of these tests take several years to complete and are very expensive. Depending on the obtained results, a pesticide may or may not be approved for registration or its registration may be cancelled.

Thus, it is necessary to be able to efficiently predict safety parameters at the early development stage of chemical compounds. In this study, we developed a prediction model for K_{oc} and BCF, which are two of the most important parameters used in the risk and hazard assessments conducted in Europe by analyzing the DAR and physicochemical properties calculated by OPERA software.

In Chapter 3, we developed prediction models for K_{oc} values based on the GBDT algorithm.

For this purpose, we collected experimentally determined physicochemical properties from pesticide evaluation reports. In addition, we calculated molecular descriptors by the Cheminformatics software. As a result, the following results were obtained.

- By using both the molecular descriptors calculated from the structural formulas and experimental physicochemical properties from the literature and open databases, the model prediction accuracy was significantly improved.
- By utilizing the GBDT algorithm, the prediction accuracy of the proposed model was further increased.
- The prediction models based on the GBDT algorithm demonstrated the best prediction abilities among the different machine learning models.
- The proposed models were developed using the open data sources and free software.

The results of this work revealed that it was possible to perform a preliminary environmental risk assessment at a relatively low cost and without conducting time-consuming experiments. However, retrieving data from a large number of documents takes a considerable amount of time. Therefore, in the next chapter, physicochemical properties were evaluated by using the latest version of Cheminformatics software.

In Chapter 4, we established prediction models for K_{oc} values utilizing only calculated parameters. We also proposed a prediction model based on the GBDT algorithm that utilized the physicochemical properties and molecular descriptors quantified by the OPERA and Mordred (for large datasets) software packages. After performing these steps, the following results were obtained.

- By using both the physicochemical properties and molecular descriptors calculated from structural formulas, the model prediction accuracy was considerably increased as compared with that of the model described in the previous chapter.
- By using the GBDT algorithm, the prediction accuracy was further improved.
- The prediction models based on the GBDT algorithm exhibited the best prediction abilities among various machine learning models.
- The performance of the model developed in this chapter was much higher than those of the models developed in the previous chapter.

Although the model using the experimentally determined physicochemical properties demonstrated a good fit, high prediction accuracy, and robustness, the method proposed in this chapter can be used instead of the actual values if the latter are difficult to acquire.

In Chapter 5, we developed prediction models for the estimation of BCF values. In particular, we calculated physicochemical properties by OPERA software because the QSPR models for K_{oc} developed in the previous chapter demonstrated higher performance than those of the previously constructed models. As a result, the following conclusions were drawn.

- By using both the physicochemical properties and molecular descriptors calculated from structural formulas, the model prediction accuracy was considerably improved.
- By using the GBDT algorithm, the model prediction accuracy was further increased.
- The prediction models using the GBDT algorithm demonstrated the best prediction abilities among various machine learning models.

The obtained results suggest that by using the GBDT algorithm for constructing a prediction model including molecular descriptors, physicochemical properties, and environmental parameters as explanatory variables, it is possible to establish a high-performance model that is comparable to the existing models as well as to the QSPR models for K_{oc} described in the previous chapter. Thus, the proposed approach is applicable for the development of prediction models for various parameters with a potential utilization in the environmental risk assessment.

Future research studies should focus on the following topics.

1. In Chapters 3 and 4, we developed prediction models for the K_{oc} values of pesticides; however, because a large pesticide database called "Pesticide Properties Database" [Lewis et al., 2016] is publicly available, prediction models for compounds with a wider range of physicochemical properties can be developed as well. Furthermore, because this dataset includes other parameters in addition to K_{oc} magnitudes, it can be applied to develop a more comprehensive prediction model for the risk assessment of pesticides.

2. In Chapter 5, we proposed a prediction model for BCF, which is the most representative factor describing the bioaccumulation of chemicals. Other bioconcentration parameters include the bioaccumulation factor, which is a bioconcentration factor of the surrounding environment (water, soil, and food); biomagnification factor, a bioconcentration factor characterizing the oral intake of food; and biota-substrate (soil/sediment) accumulation factor, which represents a bioconcentration factor of the soil and sediment of the habitat substrate [Crookes and Brooke, 2011, Burkhard et al., 2012].

   Thus, a comprehensive prediction model that includes these bioaccumulation factors should

be developed in the future.

3. While the prediction models for K_{oc} and BCF were established in Chapters 3 and 4 as part of the environmental risk/hazard assessment procedure, other important environmental parameters include the mobility and degradation of soils [Arias-Estévez et al., 2008]. In addition to environmental assessments, human health risk assessments are also routinely conducted for the safety evaluation of pesticides, and various toxicity tests such as carcinogenicity, mutagenicity, and genotoxicity tests are performed as well [Bolognesi, 2003, Tchounwou et al., 2004].

   The methodology of the QSAR/QSPR models proposed in this study is highly versatile and can be applied to other environmental impact assessments and human and animal toxicity evaluations. Therefore, a comprehensive prediction model must be established for various pesticide evaluation parameters to further improve the accuracy of predicting whether a particular pesticide should be registered or not.

4. In Chapters 3, 4, and 5, we summarized the physical properties of chemical compounds and developed prediction models for K_{oc} and BCF values. In a recent research work, the structures of chemical compounds were considered graph structures, which were successfully used for designing neural network-based applications [Altae-Tran et al., 2017, Gilmer et al., 2017]. In the future, the possibility of applying graph convolutional networks to the development of prediction models for chemical risk assessment parameters should be explored.

Acknowledgement

With great appreciation, I would like to sincerely thank my academic advisor Professor Kenichi Yoshida for taking the time to consult with me, who had no knowledge or skills in data science at the time of admission, and for taking the form of my doctoral dissertation. I also thank my associate thesis advisors - Professor Kazuhiko Tsuda and Professor Hua Xu for their time and advice.

I would like to express my sincere gratitude to Takumi Uchida in AI Consulting of Tsukuba, Corp. for his advice and assistance in my research.I would like to express my sincere gratitude to Dr. Kei Nakagawa, Dr. Naoki Kobayakawa, Takumi Uchida, Tomonori Manabe, Tomohisa Aoshima, Shigeki koda, Toshiaki Hirata, Wataru Suzuki, Takafumi Takeda, and Kenta Suzuki who has helped me and spent time with me in graduate school and Prof. Yoshida's laboratory. I also would like to thank my friend, former colleague in Astellas Pharma Inc. and college classmate in Graduate School of Business Sciences, University of Tsukuba, Masaya Fujita. We have had a relationship based on friendly and rivalry.

In addition, I wish to thank my colleagues in NTT DATA INSTITUTE OF MANAGEMENT CONSULTING, Inc. I would like to express my gratitude to my former colleagues in Nippon Soda Co. Ltd., Yoshiyuki Eguchi, Hirotaka Sugiyama, Mariko Hashi, Dr. Yoshikane Itoh, colleagues in

Regulatory affairs department, Department of Environmental Science and Toxicology in Odawara Research Center, Technology Laboratory in Takaoka Plant, NCAS Odawara Laboratory who allowed me to enroll in a doctoral course while I was working at the Nippon Soda and gave me the understanding to carry out this research in parallel with my work.

Finally, I would like to express my deep gratitude to my wife, Asami, my daughter, Ellie, my parents, Masaaki and Kaoru for all their patience, supports and encouragements throughout doctoral course at the University of Tsukuba.

References

[Akiba et al., 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

[Altae-Tran et al., 2017] Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4):283–293.

[Alvarez-Benedi et al., 1999] Alvarez-Benedi, J., Tabemero, M. T., Atienza, J., and Bolado, S. (1999). A coupled model representing volatilisation and sorption of soil incorporated herbicides. *Chemosphere*, 38(7):1583–1593.

[Ambure et al., 2015] Ambure, P., Aher, R. B., Gajewicz, A., Puzyn, T., and Roy, K. (2015). " nanobridges " software: Open access tools to perform qsar and nano-qsar modeling. *Chemometrics and Intelligent Laboratory Systems*, 147:1–13.

[Arias-Estévez et al., 2008] Arias-Estévez, M., López-Periago, E., Martínez-Carballo, E., Simal-Gándara, J., Mejuto, J.-C., and García-Río, L. (2008). The mobility and degradation of pesticides

in soils and the pollution of groundwater resources. *Agriculture, Ecosystems & Environment*, 123(4):247–260.

[Arnot and Gobas, 2006] Arnot, J. A. and Gobas, F. A. (2006). A review of bioconcentration factor (bcf) and bioaccumulation factor (baf) assessments for organic chemicals in aquatic organisms. *Environmental Reviews*, 14(4):257–297.

[Benfenati, 2011] Benfenati, E. (2011). *Quantitative structure-activity relationships (QSAR) for pesticide regulatory purposes*. Elsevier.

[Berthod et al., 2017] Berthod, L., Whitley, D. C., Roberts, G., Sharpe, A., Greenwood, R., and Mills, G. A. (2017). Quantitative structure-property relationships for predicting sorption of pharmaceuticals to sewage sludge during waste water treatment processes. *Sci Total Environ*, 579:1512–1520.

[Bhhatarai and Gramatica, 2011] Bhhatarai, B. and Gramatica, P. (2011). Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. *Environ Sci Technol*, 45(19):8120–8.

[Bintein et al., 1993] Bintein, S., Devillers, J., and Karcher, W. (1993). Nonlinear dependence of fish bioconcentration on n-octanol/water partition coefficient. *SAR QSAR Environ Res*, 1(1):29–39.

[Bolognesi, 2003] Bolognesi, C. (2003). Genotoxicity of pesticides: a review of human biomonitoring studies. *Mutation Research/Reviews in Mutation Research*, 543(3):251–272.

[Bradbury, 1995] Bradbury, S. P. (1995). Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicology Letters*, 79(1):229–237.

[Braeuning et al., 2018] Braeuning, C., Braeuning, A., Mielke, H., Holzwarth, A., and Peiser, M. (2018). Evaluation and improvement of qsar predictions of skin sensitization for pesticides. *SAR QSAR Environ Res*, 29(10):823–846.

[Burkhard et al., 2012] Burkhard, L. P., Cowan-Ellsberry, C., Embry, M. R., Hoke, R. A., and Kidd, K. A. (2012). Bioaccumulation data from laboratory and field studies: are they comparable? *Integrated environmental assessment and management*, 8(1):13–16.

[Card et al., 2017] Card, M. L., Gomez-Alvarez, V., Lee, W.-H., Lynch, D. G., Orentas, N. S., Lee, M. T., Wong, E. M., and 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.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, page 785–794, New York, NY, USA. Association for Computing Machinery.

[Chen et al., 2002] Chen, X., Cho, S. J., Li, Y., and Venkatesh, S. (2002). Prediction of aqueous solubility of organic compounds using a quantitative structure–property relationship. *Journal of Pharmaceutical Sciences*, 91(8):1838–1852.

[Chi et al., 2018] Chi, Y., Zhang, H., Huang, Q., Lin, Y., Ye, G., Zhu, H., and Dong, S. (2018). Environmental risk assessment of selected organic chemicals based on toc test and qsar estimation models. *Journal of Environmental Sciences*, 64:23–31.

[Chirico and Gramatica, 2011] Chirico, N. and Gramatica, P. (2011). Real external predictivity of qsar models: how to evaluate it? comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J Chem Inf Model*, 51(9):2320–35.

[Chirico and Gramatica, 2012] Chirico, N. and Gramatica, P. (2012). Real external predictivity of qsar models. part 2. new intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J Chem Inf Model*, 52(8):2044–58.

[Coats, 1990] Coats, J. R. (1990). Mechanisms of toxic action and structure-activity relationships for organochlorine and synthetic pyrethroid insecticides. *Environmental Health Perspectives*, 87:255–262.

[Connell, 1988] Connell, D. W. (1988). Bioaccumulation behavior of persistent organic chemicals with aquatic organisms. *Reviews of environmental contamination and toxicology*, pages 117–154.

[Connell and Hawker, 1988] Connell, D. W. and Hawker, D. W. (1988). Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. *Ecotoxicology and Environmental Safety*, 16(3):242–257.

[Cos et al., 1998] Cos, P., Ying, L., Calomme, M., Hu, J. P., Cimanga, K., Van Poel, B., Pieters, L., Vlietinck, A. J., and Berghe, D. V. (1998). Structure − activity relationship and classification

of flavonoids as inhibitors of xanthine oxidase and superoxide scavengers. *Journal of Natural Products*, 61(1):71–76.

[Crookes and Brooke, 2011] Crookes, M. and Brooke, D. (2011). Estimation of fish bioconcentration factor (bcf) from depuration data. *Science Report SCHO0811BUCE - E - E. Environment Agency, Bristol, UK*.

[Damalas and Eleftherohorinos, 2011] Damalas, C. A. and Eleftherohorinos, I. G. (2011). Pesticide exposure, safety issues, and risk assessment indicators. *Int J Environ Res Public Health*, 8(5):1402–19.

[Danaei et al., 2005] Danaei, G., Vander Hoorn, S., Lopez, A. D., Murray, C. J. L., and Ezzati, M. (2005). Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *The Lancet*, 366(9499):1784–1793.

[Devillers, 2001] Devillers, J. (2001). A general qsar model for predicting the acute toxicity of pesticides to lepomis macrochirus. *SAR and QSAR in Environmental Research*, 11(5-6):397–417.

[Devillers et al., 1996] Devillers, J., Bintein, S., and Domine, D. (1996). Comparison of bcf models based on log p. *Chemosphere*, 33(6):1047–1065.

[Directive, 1991] Directive, C. (1991). Council directive 91/414/eec of 15 july 1991 concerning the placing of plant protection products on the market. *Official Journal of the European Communities L*, 230:1–32.

[dos Reis et al., 2013] dos Reis, R. R., Sampaio, S. C., and de Melo, E. B. (2013). The effect of different log p algorithms on the modeling of the soil sorption coefficient of nonionic pesticides. *Water Res*, 47(15):5751–9.

[dos Reis et al., 2014] dos Reis, R. R., Sampaio, S. C., and de Melo, E. B. (2014). An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficients. *Water Res*, 53:191–9.

[Duchowicz et al., 2007] Duchowicz, P. R., González, M. P., Helguera, A. M., Natália Dias Soeiro Cordeiro, M., and Castro, E. A. (2007). Application of the replacement method as novel variable selection in qspr. 2. soil sorption coefficients. *Chemometrics and Intelligent Laboratory Systems*, 88(2):197–203.

[ECHA, 2017] ECHA (2017). Guidance on information requirements and chemical safety assessment, chapter r. 11: Pbt/vpvb assessment.

[EFSA, 2010] EFSA (2010). Conclusion on the peer review of the pesticide risk assessment of the active substance carbendazim. *EFSA Journal*, 8(5).

[EFSA, 2013] EFSA (2013). Scientific opinion on the hazard assessment of endocrine disruptors: scientific criteria for identification of endocrine disruptors and appropriateness of existing test methods for assessing effects mediated by these substances on human health and the environment. *EFSA Journal*, 11(3):3132.

[EFSA, 2018] EFSA (2018). Scientific risk assessment of pesticides in the european union (eu): Efsa contribution to on-going reflections by the ec. *EFSA Supporting Publications*, 15(1):1367E.

[EPRS, 2018] EPRS (2018). Regulation (ec) 1107/2009 on the placing of plant protection products on the market. *Brussels: European Parliament*, page PE 615.668.

[EPRS, 2020] EPRS (2020). The impact of the general data protection regulation (gdpr) on artificial intelligence. *Brussels: European Parliament*, page PE 641.530.

[Eriksson et al., 2013] Eriksson, L., Byrne, T., Johansson, E., Trygg, J., and Vikström, C. (2013). *Multi-and megavariate data analysis basic principles and applications*, volume 1. Umetrics Academy.

[EuropeanComission, 2003] EuropeanComission (2003). Guidance document on the assessment of the relevance of metabolites in groundwater of substances regulated under council directive 91/414/eec. *The European Commission, Health & Consumer Protection Directorate-General. Sanco/221/2000-rev*.

[EuropeanComission, 2006] EuropeanComission (2006). A thematic strategy on the sustainable use of pesticides. *COMMISSION OF THE EUROPEAN COMMUNITIES*, COM(2006) 372 final.

[EuropeanComission, 2009] EuropeanComission (2009). Regulation (ec) no 1107/2009 of the european parliament and of the council of 21 october 2009 concerning the placing of plant protection products on the market and repealing council directives 79/117/eec and 91/414/eec. *Off J Eur Comm L*, 309:1–50.

[EuropeanComission, 2014] EuropeanComission (2014). Guidance document on the renewal of approval of active substances to be assessed in compliance with regulation (eu) no 844/2012. sanco/2012/11251 - rev. 4, 12 december 2014.

[Fernández et al., 2012] Fernández, A., Lombardo, A., Rallo, R., Roncaglioni, A., Giralt, F., and Benfenati, E. (2012). Quantitative consensus of bioaccumulation models for integrated testing strategies. *Environment international*, 45:51–58.

[FOCUS, 2000] FOCUS (2000). Focus groundwater scenarios in the eu review of active substances. *FOCUS Reference Sanco/321/2000 rev.*, 2:202.

[Freire et al., 2010] Freire, M. G., Neves, C. M. S. S., Ventura, S. P. M., Pratas, M. J., Marrucho, I. M., Oliveira, J., Coutinho, J. A. P., and Fernandes, A. M. (2010). Solubility of non-aromatic ionic liquids in water and correlation using a qspr approach. *Fluid Phase Equilibria*, 294(1-2):234–240.

[Freitas et al., 2014] Freitas, M. R., Freitas, M. P., and Macedo, R. L. (2014). Aug-mia-qspr modeling of the soil sorption of carboxylic acid herbicides. *Bull Environ Contam Toxicol*, 93(4):489–92.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

[Fujii and Hermann, 1982] Fujii, A. and Hermann, E. R. (1982). Correlation between flash points and vapor pressures of organic compounds. *Journal of Safety Research*, 13(4):163–175.

[Gao et al., 1996] Gao, C., Govind, R., and Tabak, H. H. (1996). Predicting soil sorption coefficients of organic chemicals using a neural network model. *Environmental Toxicology and Chemistry*, 15(7):1089–1096.

[Garg et al., 2014] Garg, A., Garg, A., Tai, K., and Sreedeep, S. (2014). An integrated srm-multi-gene genetic programming approach for prediction of factor of safety of 3-d soil nailed slopes. *Engineering Applications of Artificial Intelligence*, 30:30–40.

[Garg and Smith, 2014] Garg, R. and Smith, C. J. (2014). Predicting the bioconcentration factor of highly hydrophobic organic chemicals. *Food Chem Toxicol*, 69:252–9.

[Gawlik et al., 1997] Gawlik, B. M., Sotiriou, N., Feicht, E. A., Schulte-Hostede, S., and Kettrup, A. (1997). Alternatives for the determination of the soil adsorption coefficient, koc, of non-ionicorganic compounds — a review. *Chemosphere*, 34(12):2525–2551.

[Gharagheizi et al., 2012] Gharagheizi, F., Eslamimanesh, A., Ilani-Kashkouli, P., Mohammadi, A. H., and Richon, D. (2012). Qspr molecular approach for representation/prediction of very large vapor pressure dataset. *Chemical Engineering Science*, 76:99–107.

[Gilmer et al., 2017] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR.

[Gissi et al., 2015] Gissi, A., Lombardo, A., Roncaglioni, A., Gadaleta, D., Mangiatordi, G. F., Nicolotti, O., and Benfenati, E. (2015). Evaluation and comparison of benchmark qsar models to predict a relevant reach endpoint: The bioconcentration factor (bcf). *Environ Res*, 137:398–409.

[Golmohammadi et al., 2012] Golmohammadi, H., Dashtbozorgi, Z., and Acree, W. E., J. (2012). Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci*, 47(2):421–9.

[Goudarzi et al., 2009] Goudarzi, N., Goodarzi, M., Araujo, M. C., and Galvao, R. K. (2009). Qspr modeling of soil sorption coefficients (k(oc)) of pesticides using spa-ann and spa-mlr. *J Agric Food Chem*, 57(15):7153–8.

[Gramatica, 2007] Gramatica, P. (2007). Principles of qsar models validation: internal and external. *QSAR & Combinatorial Science*, 26(5):694–701.

[Gramatica, 2010] Gramatica, P. (2010). Chemometric methods and theoretical molecular descriptors in predictive qsar modeling of the environmental behavior of organic pollutants. *Recent Advances in QSAR Studies*, pages 327–366.

[Gramatica et al., 2000] Gramatica, P., Corradi, M., and Consonni, V. (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere*, 41(5):763–777.

[Gramatica and Papa, 2005] Gramatica, P. and Papa, E. (2005). An update of the bcf qsar model based on theoretical molecular descriptors. *QSAR & Combinatorial Science*, 24(8):953–960.

[Haeberlein and Brinck, 1997] Haeberlein, M. and Brinck, T. (1997). Prediction of water–octanol partition coefficients usingtheoretical descriptors derived from the molecular surface area and theelectrostatic potential. *Journal of the Chemical Society, Perkin Transactions 2*, (2):289–294.

[Halder et al., 2018] Halder, A. K., Moura, A. S., and Cordeiro, M. N. D. (2018). Qsar modelling: a therapeutic patent review 2010-present. *Expert Opinion on Therapeutic Patents*, 28(6):467–476.

[Halgren, 1996] Halgren, T. A. (1996). Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17(5 - 6):490–519.

[Hamadache et al., 2016] Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L., and Si Moussa, C. (2016). A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *J Hazard Mater*, 303:28–40.

[Hansch, 1993] Hansch, C. (1993). Quantitative structure-activity relationships and the unnamed science. *Accounts of Chemical Research*, 26(4):147–153.

[Hansch and Fujita, 1964] Hansch, C. and Fujita, T. (1964). p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8):1616–1626.

[Hansch and Leo, 1995] Hansch, C. and Leo, A. (1995). Exploring qsar. fundamentals and applications in chemistry and biology. acs professional reference book. *American Chemical Society*, 1:557–1037.

[Henschel et al., 1997] Henschel, K. P., Wenzel, A., Diedrich, M., and Fliedner, A. (1997). Environmental hazard assessment of pharmaceuticals. *Regulatory Toxicology and Pharmacology*, 25(3):220–225.

[Hernando et al., 2006] Hernando, M. D., Mezcua, M., Fernandez-Alba, A. R., and Barcelo, D. (2006). Environmental risk assessment of pharmaceutical residues in wastewater effluents, surface waters and sediments. *Talanta*, 69(2):334–42.

[Howard and Meylan, 2000] Howard, P. and Meylan, W. (2000). Physprop database. *Syracuse Research Corp., Syracuse, NY*.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *IEEE Annals of the History of Computing*, 9(03):90–95.

[Huuskonen, 2003] Huuskonen, J. (2003). Prediction of soil sorption coefficient of organic pesticides from the atom-type electrotopological state indices. *Environmental Toxicology and Chemistry*, 22(4):816–820.

[Isnard and Lambert, 1988] Isnard, P. and Lambert, S. (1988). Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility. *Chemosphere*, 17(1):21–34.

[Jiao, 2012] Jiao, L. (2012). Qspr study on the soil-water partition coefficient of polychlorinated biphenyls by using artificial neural network. *Advanced Materials Research*, 455-456:930–934.

[Jones et al., 2001] Jones, E., Oliphant, T., and Peterson, P. (2001). Scipy: Open source scientific tools for python.

[Jury, 1986] Jury, W. (1986). Adsorption of organic chemicals onto soil. *Vadose Zone Modeling of Organic Pollutants. Lewis Publishers, Inc., Chelsea Michigan. 1986. p 177-189, 2 tab, 39 ref.*

[Kahn et al., 2005] Kahn, I., Fara, D., Karelson, M., Maran, U., and Andersson, P. L. (2005). Qspr treatment of the soil sorption coefficients of organic pollutants. *Journal of Chemical Information and Modeling*, 45(1):94–105.

[Kaneko and Funatsu, 2015] Kaneko, H. and Funatsu, K. (2015). Strategy of structure generation within applicability domains with one-class support vector machine. *Bulletin of the Chemical Society of Japan*, 88(7):981–988.

[Karelson et al., 1996] Karelson, M., Lobanov, V. S., and Katritzky, A. R. (1996). Quantum-chemical descriptors in qsar/qspr studies. *Chemical Reviews*, 96(3):1027–1044.

[Katritzky et al., 2002] Katritzky, A. R., Lomaka, A., Petrukhin, R., Jain, R., Karelson, M., Visser, A. E., and Rogers, R. D. (2002). Qspr correlation of the melting point for pyridinium bromides, potential ionic liquids. *Journal of Chemical Information and Computer Sciences*, 42(1):71–74.

[Katritzky et al., 1998] Katritzky, A. R., Wang, Y., Sild, S., Tamm, T., and Karelson, M. (1998). Qspr studies on vapor pressure, aqueous solubility, and the prediction of water − air partition coefficients. *Journal of Chemical Information and Computer Sciences*, 38(4):720–725.

[Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

[Kim et al., 2016] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. (2016). Pubchem substance and compound databases. *Nucleic Acids Res*, 44(D1):D1202–13.

[Klein et al., 1988] Klein, W., Kördel, W., Weiß, M., and Poremski, H. J. (1988). Updating of the oecd test guideline 107 "partition coefficient n-octanol/water": Oecd laboratory intercomparison test on the hplc method. *Chemosphere*, 17(2):361–386.

[Klopffer, 1994] Klopffer, W. (1994). Environmental hazard : Assessment of chemicals and products part i: General assessment principles. *Environ Sci Pollut Res Int*, 1(1):47–53.

[Kowalski and Bender, 1974] Kowalski, B. R. and Bender, C. F. (1974). The application of pattern recognition to screening prospective anticancer drugs. adenocarcinoma 755 biological activity test. *J Am Chem Soc*, 96(3):916–8.

[Kubinyi, 1997] Kubinyi, H. (1997). Qsar and 3d qsar in drug design part 1: methodology. *Drug discovery today*, 2(11):457–467.

[Leszczynski and Puzyn, 2012] Leszczynski, J. and Puzyn, T. (2012). Towards efficient designing of safe nanomaterials: Innovative merge of computational approaches and experimental techniques.

[Lewis et al., 2016] Lewis, K. A., Tzilivakis, J., Warner, D. J., and Green, A. (2016). An international database for pesticide risk assessments and management. *Human and Ecological Risk Assessment: An International Journal*, 22(4):1050–1064.

[Liang et al., 2013] Liang, G., Xu, J., and Liu, L. (2013). Qspr analysis for melting point of fatty acids using genetic algorithm based multiple linear regression (ga-mlr). *Fluid Phase Equilibria*, 353:15–21.

[Lombardo et al., 2010] Lombardo, A., Roncaglioni, A., Boriani, E., Milan, C., and Benfenati, E. (2010). Assessment and validation of the caesar predictive model for bioconcentration factor (bcf) in fish. *Chem Cent J*, 4 Suppl 1:S1.

[Mackay and Fraser, 2000] Mackay, D. and Fraser, A. (2000). Bioaccumulation of persistent organic chemicals: mechanisms and models. *Environmental Pollution*, 110(3):375–391.

[Magnuson et al., 1983] Magnuson, V., Harriss, D., and Basak, S. (1983). *Studies in physical and theoretical chemistry*, pages 178–191. Elsevier Amsterdam.

[Mansouri et al., 2018] Mansouri, K., Grulke, C., Judson, R., and Williams, A. (2018). Opera: A free and open source qsar tool for predicting physicochemical properties and environmental fate endpoints. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 255. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA.

[Markell, 2010] Markell, D. (2010). An overview of tsca, its history and key underlying assumptions, and its place in environmental regulation. *Wash. UJL & Pol'y*, 32:333.

[Matthies et al., 2016] Matthies, M., Solomon, K., Vighi, M., Gilman, A., and Tarazona, J. V. (2016). The origin and evolution of assessment criteria for persistent, bioaccumulative and toxic (pbt) chemicals and persistent organic pollutants (pops). *Environmental Science: Processes & Impacts*, 18(9):1114–1128.

[Mauri et al., 2006] Mauri, A., Consonni, V., Pavan, M., and Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, 56(2):237–248.

[McDougall, 2016] McDougall, P. (2016). R&d expenditure in 2014 and expectations for 2019. *Phillips McDougall*.

[Meylan et al., 1999] Meylan, W. M., Howard, P. H., Boethling, R. S., Aronson, D., Printup, H., and Gouchie, S. (1999). Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environmental Toxicology and Chemistry*, 18(4):664–672.

[Miyamoto et al., 1990] Miyamoto, J., Mikami, N., and Takimoto, Y. (1990). The fate of pesticides in aquatic ecosystems. *Progress in pesticide biochemistry and toxicology*, 7:123–147.

[Moermond et al., 2012] Moermond, C. T., Janssen, M. P., de Knecht, J. A., Montforts, M. H., Peijnenburg, W. J., Zweers, P. G., and Sijm, D. T. (2012). Pbt assessment using the revised annex xiii of reach: a comparison with other regulatory frameworks. *Integr Environ Assess Manag*, 8(2):359–71.

[Moriwaki et al., 2018] Moriwaki, H., Tian, Y. S., Kawashita, N., and Takagi, T. (2018). Mordred: a molecular descriptor calculator. *J Cheminform*, 10(1):4.

[Moss et al., 2000] Moss, K. T., Boethling, R. S., Nabholz, J. V., and Auer, C. M. (2000). *US Environmental Protection Agency new chemicals program PBT chemical category: Screening and risk management of new PBT chemical substances*. ACS Publications.

[Nishimoto, 2019] Nishimoto, R. (2019). Global trends in the crop protection industry. *Journal of pesticide science*, pages D19–101.

[Nolte and Ragas, 2017] Nolte, T. M. and Ragas, A. M. J. (2017). A review of quantitative structure–property relationships for the fate of ionizable organic chemicals in water matrices and identification of knowledge gaps. *Environmental Science: Processes & Impacts*, 19(3):221–246.

[O'Boyle et al., 2011] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33.

[OECD, 2012] OECD (2012). *Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure*. OECD Publishing.

[Olguin et al., 2017] Olguin, C. J. M., Sampaio, S. C., and Dos Reis, R. R. (2017). Statistical equivalence of prediction models of the soil sorption coefficient obtained using different log p algorithms. *Chemosphere*, 184:498–504.

[Oliphant, 2007] Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20.

[Padmanabhan et al., 2006] Padmanabhan, J., Parthasarathi, R., Subramanian, V., and Chattaraj, P. K. (2006). Qspr models for polychlorinated biphenyls: n-octanol/water partition coefficient. *Bioorg Med Chem*, 14(4):1021–8.

[Papa et al., 2007] Papa, E., Dearden, J. C., and Gramatica, P. (2007). Linear qsar regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. *Chemosphere*, 67(2):351–8.

[Pavan et al., 2008] Pavan, M., Netzeva, T., and Worth, A. (2008). Review of literature-based quantitative structure–activity relationship models for bioconcentration. *QSAR & Combinatorial Science*, 27(1):21–31.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[Pence and Williams, 2010] Pence, H. E. and Williams, A. (2010). Chemspider: An online chemical information resource. *Journal of Chemical Education*, 87(11):1123–1124.

[Pramanik and Roy, 2014] Pramanik, S. and Roy, K. (2014). Modeling bioconcentration factor (bcf) using mechanistically interpretable descriptors computed from open source tool "padel-descriptor". *Environ Sci Pollut Res Int*, 21(4):2955–65.

[Prokhorenkova et al., 2017] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2017). Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*.

[Rice-Evans et al., 1996] Rice-Evans, C. A., Miller, N. J., and Paganga, G. (1996). Structure-antioxidant activity relationships of flavonoids and phenolic acids. *Free Radical Biology and Medicine*, 20(7):933–956.

[Roe et al., 2005] Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584.

[Rogers and Hopfinger, 1994] Rogers, D. and Hopfinger, A. J. (1994). Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences*, 34(4):854–866.

[Roy et al., 2015] Roy, K., Kar, S., and Ambure, P. (2015). On a simple approach for determining applicability domain of qsar models. *Chemometrics and Intelligent Laboratory Systems*, 145:22–29.

[Roy et al., 2012] Roy, K., Mitra, I., Kar, S., Ojha, P. K., Das, R. N., and Kabir, H. (2012). Comparative studies on some metrics for external validation of qspr models. *J Chem Inf Model*, 52(2):396–408.

[Sabljić et al., 1995] Sabljić, A., Güsten, H., Verhaar, H., and Hermens, J. (1995). Qsar modelling of soil sorption. improvements and systematics of log koc vs. log kow correlations. *Chemosphere*, 31(11-12):4489–4514.

[Schäfer et al., 2019] Schäfer, R. B., Liess, M., Altenburger, R., Filser, J., Hollert, H., Roß-Nickoll, M., Schäffer, A., and Scheringer, M. (2019). Future pesticide risk assessment: narrowing the gap between intention and reality. *Environmental Sciences Europe*, 31(1):21.

[Shao et al., 2014] Shao, Y., Liu, J., Wang, M., Shi, L., Yao, X., and Gramatica, P. (2014). Integrated qspr models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods. *Atmospheric Environment*, 88:212–218.

[Sparks et al., 2001] Sparks, T. C., Crouse, G. D., and Durst, G. (2001). Natural products as insecticides: the biology, biochemistry and quantitative structure-activity relationships of spinosyns and spinosoids. *Pest Manag Sci*, 57(10):896–905.

[Su et al., 2021] Su, R., Wu, H., Liu, X., and Wei, L. (2021). Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies. *Brief Bioinform*, 22(1):428–437.

[Tchounwou et al., 2004] Tchounwou, P. B., Centeno, J. A., and Patlolla, A. K. (2004). Arsenic toxicity, mutagenesis, and carcinogenesis – a health risk assessment and management approach. *Molecular and Cellular Biochemistry*, 255(1):47–55.

[Thiele-Bruhn, 2003] Thiele-Bruhn, S. (2003). Pharmaceutical antibiotic compounds,in soils - a review (vol 166, pg 145, 2003). *Journal of Plant Nutrition and Soil Science*, 166(4):546–546.

[Todeschini and Consonni, 2008] Todeschini, R. and Consonni, V. (2008). *Handbook of molecular descriptors*, volume 11. John Wiley & Sons.

[Toropov and Benfenati, 2006] Toropov, A. A. and Benfenati, E. (2006). Qsar models for daphnia toxicity of pesticides based on combinations of topological parameters of molecular structures. *Bioorg Med Chem*, 14(8):2779–88.

[Toropov et al., 2017] Toropov, A. A., Toropova, A. P., Marzo, M., Dorne, J. L., Georgiadis, N., and Benfenati, E. (2017). Qsar models for predicting acute toxicity of pesticides in rainbow trout using the coral software and efsa's openfoodtox database. *Environ Toxicol Pharmacol*, 53:158–163.

[Toropova et al., 2020] Toropova, A. P., Duchowicz, P. R., Saavedra, L. M., Castro, E. A., and Toropov, A. A. (2020). The use of the index of ideality of correlation to build up models for bioconcentration factor. *Mol Inform*, 39(7):e1900070.

[Tremolada et al., 2004] Tremolada, P., Finizio, A., Villa, S., Gaggi, C., and Vighi, M. (2004). Quantitative inter-specific chemical activity relationships of pesticides in the aquatic environment. *Aquat Toxicol*, 67(1):87–103.

[Tropsha, 2010] Tropsha, A. (2010). Best practices for qsar model development, validation, and exploitation. *Mol Inform*, 29(6-7):476–88.

[van der Oost et al., 2003] van der Oost, R., Beyer, J., and Vermeulen, N. P. E. (2003). Fish bioaccumulation and biomarkers in environmental risk assessment: a review. *Environmental Toxicology and Pharmacology*, 13(2):57–149.

[Veith et al., 1979] Veith, G. D., DeFoe, D. L., and Bergstedt, B. V. (1979). Measuring and estimating the bioconcentration factor of chemicals in fish. *Journal of the Fisheries Research Board of Canada*, 36(9):1040–1048.

[Walt et al., 2011] Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

[Weisbrod et al., 2007] Weisbrod, A. V., Burkhard, L. P., Arnot, J., Mekenyan, O., Howard, P. H., Russom, C., Boethling, R., Sakuratani, Y., Traas, T., Bridges, T., Lutz, C., Bonnell, M., Woodburn, K., and Parkerton, T. (2007). Workgroup report: review of fish bioaccumulation databases used to identify persistent, bioaccumulative, toxic substances. *Environ Health Perspect*, 115(2):255–61.

[Wen et al., 2012] Wen, Y., Su, L. M., Qin, W. C., Fu, L., He, J., and Zhao, Y. H. (2012). Linear and non-linear relationships between soil sorption and hydrophobicity: model, validation and influencing factors. *Chemosphere*, 86(6):634–40.

[Wildman and Crippen, 1999] Wildman, S. A. and Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873.

[Williams et al., 2017] Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., and Richard, A. M. (2017). The comptox chemistry dashboard: a community data resource for environmental chemistry. *J Cheminform*, 9(1):61.

[Woodrow et al., 1997] Woodrow, J. E., Seiber, J. N., and Baker, L. W. (1997). Correlation techniques for estimating pesticide volatilization flux and downwind concentrations. *Environmental Science & Technology*, 31(2):523–529.

[Yang et al., 2010] Yang, Y., Chen, H., Nilsson, I., Muresan, S., and Engkvist, O. (2010). Investigation of the relationship between topology and selectivity for druglike molecules. *Journal of Medicinal Chemistry*, 53(21):7709–7714.

[Yap, 2011] Yap, C. W. (2011). Padel-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 32(7):1466–74.

[Yokota, 2014] Yokota, T. (2014). Registration of plant protection products in eu : Regulation (ec) no.1107/2009. *Plant Protection*, 68(3):117–121.

[Zeng et al., 2012] Zeng, X. L., Wang, H. J., and Wang, Y. (2012). Qspr models of n-octanol/water partition coefficients and aqueous solubility of halogenated methyl-phenyl ethers by dft method. *Chemosphere*, 86(6):619–25.

[Zhang et al., 2019] Zhang, J., Mucs, D., Norinder, U., and Svensson, F. (2019). Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity-application to the tox21 and mutagenicity data sets. *J Chem Inf Model*, 59(10):4150–4158.

[Zhao et al., 2008] Zhao, C., Boriani, E., Chana, A., Roncaglioni, A., and Benfenati, E. (2008). A new hybrid system of qsar models for predicting bioconcentration factors (bcf). *Chemosphere*, 73(11):1701–7.

[Zhao et al., 2001] Zhao, Y. H., Le, J., Abraham, M. H., Hersey, A., Eddershaw, P. J., Luscombe, C. N., Boutina, D., Beck, G., Sherborne, B., Cooper, I., and Platts, J. A. (2001). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (qsar) with the abraham descriptors. *Journal of Pharmaceutical Sciences*, 90(6):749–784.

Achievement

Chapter 3 Yoshiyuki Kobayashi, Takumi Uchida and Kenichi Yoshida, "Prediction of soil adsorption coefficient of pesticides via experimental values", Society of Environmental Toxicology and Chemistry North America 40th Annual Meeting, 2019

Yoshiyuki Kobayashi, Takumi Uchida and Kenichi Yoshida, "Prediction of soil adsorption coefficient in pesticides using physicochemical properties and molecular descriptors by machine learning models", Environmental Toxicology and Chemistry 39 (7), 1451-1459, 2020

Chapter 4 Yoshiyuki Kobayashi and Kenichi Yoshida, "Quantitative structure–property relationships for the calculation of the soil adsorption coefficient using machine learning algorithms with calculated chemical properties from open-source software", Environmental Research 196, 110363 (2021)

Chapter 5 Yoshiyuki Kobayashi and Kenichi Yoshida, "Development of QSAR models for prediction of fish bioconcentration factors using physicochemical properties and molecular descriptors with machine learning algorithms", Ecological Informatics 63, 101285 (2021)