PREDICTION OF MIXTURE TOXICITY USING COMPUTATIONAL TOXICOLOGY METHODS

Towards Integrated Model for Environmental Risk Assessment

von
Jongwoon Kim
geboren in Suwon, Republik Korea

Angenommene Dissertation zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
Fachbereich 7: Natur- und Umweltwissenschaften
Universität Koblenz-Landau

Berichterstatter:
Prof. Dr. Gabriele E. Schaumann, Landau, Erste Berichterstatterin
PD Dr. Rolf Altenburger, Leipzig, Zweiter Berichterstatter

Tag der Disputation: 02.07.2013

DECLARATION

I herewith declare that I autonomously carried out the PhD dissertation entitled *"Prediction of Mixture Toxicity using Computational Toxicology Methods: Towards Integrated Model for Environmental Risk Assessment"*. All used assistances are declared and parts of involved contributors and other authors are clearly indicated. This dissertation has never been submitted elsewhere for an exam, as dissertation or for evaluation in a similar context; neither to any department of this university nor to any other scientific institution.

This dissertation is based on the following publications, which are referred to in the text by their roman numerals:

I. Kim, J., Kim, S., Schaumann, G.E., 2012. Reliable predictive computational toxicology methods for mixture toxicity: Toward the development of innovative integrated models for environmental risk assessment. *Reviews in Environmental Science and Bio/Technology*, DOI: 10.1007/s11157-012-9286-7. Published online (27 June 2012).

II. Kim, J., Kim, S., Schaumann, G.E., 2013. A case study and a computational simulation of the European Union draft technical guidance documents for chemical safety assessment of mixtures: Limitations and a tentative alternative. *Journal of Occupational & Environmental Hygiene*, 10:181-193.

III. Kim, J., Kim, S., Schaumann, G.E., 2012. Development of a partial least squares-based integrated addition model for predicting mixture toxicity. *Human and Ecological Risk Assessment: An International Journal*. DOI:10.1080/10807039.2012.754312. Accepted author version posted online (7 December 2012).

IV. Kim, J., Kim, S., Schaumann, G.E. Development of QSAR-based two-stage prediction model for estimating mixture toxicity. *SAR and QSAR in Environmental Research*, DOI:10.1080/1062936X.2013.815654. Accepted manuscript (Accepted on 29 April 2013).

Place, Date Signature

This dissertation is dedicated to my wife, our lovely two daughters, and my mother for their love, endless support, and encouragement

``Only love has no limits. In contrast,
our predictions can fail,
our communication can fail, and
our knowledge can fail.
For our knowledge is patchwork, and
our predictive power is limited.
But when perfection comes,
patchwork will disappear."

1 Corinthians 13:8 – 1

ACKNOWLEDGEMENTS

I would first like to thank God who gave me the grace and privilege to pursue this study and successfully complete it in spite of many challenges I confronted. During the journey for my PhD study, I could learn how to define the challenges and face them with confidence.

I would like to express my sincere appreciation to my advisor, Prof. Dr. Gabriele E. Schaumann, who guided, encouraged, and continuously supported me through the course of this study. I would like to extend my sincere thanks to my co-advisor, PD Dr. Rolf Altenburger, who encouraged and helped me think my arguments again one step back from them.

I am deeply grateful to my team leader, Dr. Sanghun Kim at Korea Institute of Science and Technology Europe, who heartfully supported me to do my study independently. I would like to express my thanks to my colleagues at KIST Europe Institute, Prof. Dr. H.S. Lee, Y.C. Joo, Dr. J.W. Hwang, Dr. S.W. Kim, Dr. H.P. Jeon, Dr. S.Y. Baik, Dr. B.H. Ryu, M.J. Kang, O. Weiss, Y.J. Jung, T.H. Kim, J.W. Yoo, E.J. Oh, D.Y. Chu, and S.J. Seo, with including former colleagues, Dr. K.C. Kim, Dr. Y.K. Park, Dr. J.M. Lee, Dr. J.S. Ra, Prof. J.M. Cha, and Dr. S. Aburous for their encouragements. I also thank Dr. D. Diehl, Dr. K.M. Sepúlveda, and many further colleagues at Soil and Environmental Chemistry Group for their kind comments and supporting during my study. I especially thanks to S. Recktenwald, and L.P. Loi who helped me do laboratory tests and programming work, respectively. I am also very thankful for all my many friends and neighbors.

Finally, I would like to dedicate this dissertation to my wife, Eunju, and our lovely daughters, Rina, and Lahee. The sacrifice you made throughout my years here are simply ineffable. I deeply thank my mother for her tender, loving care and compassion. I also want to thank my brother and his family for their kindness and warmth.

God bless all of you.

TABLE OF CONTENTS

ABSTRACT

Studies on the toxicity of chemical mixtures find that components at levels below no-observed-effect concentrations (NOECs) may cause toxicity resulting from the combined effects of mixed chemicals. However, chemical risk assessment frequently focuses on individual chemical substances, although most living organisms are substantially exposed to chemical mixtures rather than single substances. The concepts of additive toxicity, concentration addition (CA), and independent action (IA) models are often applied to predict the mixture toxicity of similarly and dissimilarly acting chemicals, respectively. However, living organisms and the environment may be exposed to both types of chemicals at the same time and location. In addition, experimental acquisition of toxicity data for every conceivable mixture is unfeasible since the number of chemical combinations is extremely large. Therefore, an integrated model to predict mixture toxicity on the basis of single mixture components having various modes of toxic action (MoAs) needs to be developed. The objectives of the present study were to analyze the challenges in predicting mixture toxicity in the environment, and to develop integrated models that overcome the limitations of the existing prediction models for estimating the toxicity of non-interactive mixtures through computational models. For these goals, four sub-topics were generated in this study. Firstly, applicable domains and limitations of existing integrated models were analyzed and grouped into three kinds of categories in this study. There are current approaches used to assess mixture toxicity; however, there is a need for a new research concept to overcome challenges associated with such approaches, which recent studies have addressed. These approaches are discussed with particular emphasis on those studies involved in computational approaches to predict the toxicity of chemical mixtures based on the toxicological data of individual chemicals. Secondly, through a case study and a computational simulation, it was found that the Key Critical Component (KCC) and Composite Reciprocal (CR) methods (as described in the European Union (EU) draft technical guidance notes for calculating the Predicted No Effect Concentration (PNEC) and Derived No Effect Level (DNEL) of mixtures) could derive significantly different results.

As the third and fourth sub-topics of this study, the following two integrated addition models were developed and successfully applied to overcome the inherent limitations of the CA and IA models, which could be theoretically used for either similarly or dissimilarly acting chemicals: i) a Partial Least Squares-Based Integrated Addition Model (PLS-IAM), and, ii) a Quantitative Structure-Activity Relationship-Based Two-Stage Prediction (QSAR-TSP) model. In this study, it was shown that the PLS-IAM might be useful to estimate mixture toxicity when the toxicity data of similar mixtures having the same compositions were available. In the case of the QSAR-TSP model, it showed the potential to overcome the critical limitation of the conventional TSP model, which requires knowledge of the MoAs for all chemicals. Therefore, this study presented good potential for the advanced integrated models (*e.g.*, PLS-IAM and QSAR-TSP), while considering various non-interactive constituents that have different MoAs in order to increase the reliance of conventional models and simplify the procedure for risk assessment of mixtures.

ZUSAMMENFASSUNG

In Studien zur Toxizität von Chemikaliengemischen wurde festgestellt, dass Gemische aus Komponenten in Konzentrationen ohne erkennbare Wirkung als Einzelstoff (NOECs) als Resultat der gemeinsamen Wirkung der Substanzen Toxizität verursachen können. Die Risikobewertung von Chemikalien konzentriert sich jedoch häufig auf einzelne chemische Substanzen, obwohl die meisten lebenden Organismen im Wesentlichen chemischen Gemischen anstatt einzelnen Substanzen ausgesetzt sind. Die Konzepte der additiven Toxizität, Konzentrationsadditivität (CA) und der unabhängigen Wirkung (IA), werden häufig angewendet, um die Mischungstoxizität von Gemischen ähnlich wirkender und unähnlich wirkender Chemikalien vorherzusagen. Allerdings können lebende Organismen, ebenso wie die Umwelt, beiden Chemikalienarten zur gleichen Zeit und am gleichen Ort ausgesetzt sein. Darüber hinaus wäre es nahezu unmöglich, auf experimentellem Wege Toxizitätsdaten für jede denkbare Mischung zu gewinnen, da die Anzahl der Möglichkeiten beinahe unendlich groß ist. Aus diesem Grund muss ein integriertes Modell zur Vorhersage der Mischungstoxizität, welches auf einzelnen Mischungskomponenten mit verschiedenen Arten toxischer Wirkung (MoAs) basiert, entwickelt werden. Die Ziele der vorliegenden Studie sind, die Problematik der Vorhersage der Mischungstoxizität in der Umwelt zu analysieren und integrierte Modelle zu entwickeln, die die Beschränkungen der vorhandenen Vorhersagemodelle zur Abschätzung der Toxizität nicht-interaktiver Mischungen mittels computergestützter Modelle überwinden. Für diese Zielsetzung wurden in dieser Studie vier Unterthemen bearbeitet. Als Erstes wurden Anwendungsbereiche und Beschränkungen bereits bestehender Modelle analysiert und in die drei Kategorien dieser Studie eingruppiert. Aktuelle Ansätze zur Einschätzung der Mischungstoxizität und die Notwendigkeit eines neuen Forschungskonzepts zur Überwindung bestehender Einschränkungen, die aus neueren Studien hervorgehen, wurden diskutiert. Insbesondere diejenigen, die computergestützte Ansätze einbeziehen um die Toxizität chemischer Gemische, basierend auf den toxikologischen Daten einzelner Chemikalien, vorherzusagen. Als Zweites wurde anhand einer

Fallstudie und mittels computergestützter Simulation festgestellt, dass die Key Critical Component (KCC) und die Composite Reciprocal (CR) methods, die im Entwurf des Technischen Leitfadens der Europäischen Union (EU) zu Berechnung der Predicted No Effect Concentration (PNEC) und des Derived No Effect Level (DNEL) von Gemischen beschrieben wurden, signifikant abweichende Ergebnisse hervorbringen. Als dritter und vierter Schritt dieser Studie wurden die zwei folgenden integrierten Nebenmodelle entwickelt und erfolgreich angewandt, um die dem CA und IA Modell innewohnenden Beschränkungen zu überwinden, welche theoretisch sowohl für Chemikalien mit ähnlichen, als auch mit abweichenden Reaktionen existieren: 1) Partial Least Squares-based Integrated Addition Model (PLS-IAM) und 2) Quantitative Structure-Activity Relationship-based Two-Stage Prediction (QSAR-TSP) Modell. In dieser Studie wurde gezeigt, dass das PLS-IAM angewandt werden könnte, wenn die toxikologischen Daten ähnlicher Gemische mit gleicher Zusammensetzung zur Verfügung stehen. Das QSAR-TSP Modell zeigt eine Möglichkeit zur Überwindung der kritischen Einschränkungen des herkömmlichen TSP Modells auf, bei der Kenntnisse der MoAs aller Chemikalien erforderlich sind. Diese Studie zeigt das hohe Potential der erweiterten integrierten Modelle, z.B. PLS-IAM und QSAR-TSP, die durch Berücksichtigung verschiedener nicht-interaktiver Komponenten mit unterschiedlichen MoA Gruppen, die Verlässlichkeit konventioneller Modelle erhöhen und das Verfahren der Risikobewertung von Gemischen aus wissenschaftlicher Sicht vereinfachen.

CHAPTER I

General Introduction

GENERAL INTRODUCTION

1. Study background

Is it necessary to study the prediction of mixture toxicity?

Studies on the toxicity of mixed chemicals find that components at levels below no-observed-effect concentrations (NOECs) may cause toxicity resulting from combined effects among substances (Kortenkamp and Altenburger, 1999; Rajapakse *et al.*, 2002; Walter *et al.*, 2002; Altenburger *et al.*, 2003; Vighi *et al.*, 2003; Lydy *et al.*, 2004; Breitholtz *et al.*, 2008). However, there is still a lack of knowledge as to the underlying mechanism for such interactions (Xu and Nirmalakhandan, 1998).

From a regulatory perspective, control levels are improving and the scope of global chemical regulations for protecting human health and the environment is being strengthened and extended. In the case of the European Union (EU), where regulations are aimed at securing human health and protecting the environment, legislation is broadly divided into two forms: 1) substance- and product-based legislations such as the Registration, Evaluation, Authorization, and Restriction of Chemicals regulation (REACH); the Placing of Plant Protection Products regulation (PPP); the Classification, Labeling and Packaging regulation (CLP); and, 2) the process- and media-based legislations such as the Integrated Pollution and Prevention Control Directive (IPPC) and the Water Framework Directive (WFD). However, current risk assessments even under such strict regulations place less focus on chemical mixtures as compared to single substances (Altenburger *et al.*, 2003; European Commission, 2003; Eggen *et al.*, 2004; Altenburger and Greco, 2008; Martin *et al.*, 2009; Syberg *et al.*, 2009). Two different methods, comprising of the Key Critical Component (KCC), and Composite Reciprocal (CR) are mentioned in the EU draft technical guidance notes (European Chemical Industry Council, 2005; European Chemical Agency, 2008a, b). The KCC method assumes that only one key component should be considered as equal to the whole mixture in terms of danger for developing risk management measures (European Chemical Industry Council, 2005). However, combined effects among mixture components are ignored under such a framework. By contrast, the CR method considers a multi-component mixture as an individual chemical unit by calculating a composite Predicted No Effect Concentration (PNEC) and a Derived No Effect Level (DNEL) for the mixture based on the PNECs and DNELs of single substances derived from available testing results for the environment and human health, respectively (European Chemical Industry Council, 2005; European Chemical Agency, 2008a, b). The CR method, using a fractional PNEC or DNEL summation with these values estimated from the lowest chronic toxicity data (*e.g.* NOEC) of the minimal toxicity datasets, is strictly not the same as the conventional concentration addition model, which uses identical effective concentration endpoints (*e.g.* EC_{50}). However, the 'additive toxicity' concept as employed in the concentration addition model, and as similarly assumed by the CR method, additionally assumes that the PNEC and DNEL of a mixture can be described as the sum of the PNECs and DNELs of components, respectively (European Chemical Industry Council, 2005). The two above-mentioned methods employ different concepts for estimating mixture toxicity, and basic assumptions of the KCC and CR methods are mutually contradictory (for detailed information, see the methodology in Chapter III). However, the EU draft technical guidance notes has not yet presented apparent criteria for the practical application of each method (*i.e.*, which approach performs best according to the characteristics of a mixture) (European Chemical Industry Council, 2005; European Chemical Agency, 2008a, b).

From the industrial and commercial perspective, over 100,000 chemical substances were placed on the market in the past few decades, and approximately 200 to 300 new chemicals have been tested in Europe every year (Hartung and Rovida, 2009). The number of test groups that can be created with n substances, at only one concentration level for each substance, is '2^n-1' for every possible combination and '$n(n$-1)/2' for binary combinations. For example, 20 substances can create 190 binary combinations and more than a million possible other combinations (*e.g.,* ternary, quaternary and so on) (Cassee *et al.*, 1998; Lydy *et al.*, 2004). Toxicological tests based on animal data for filling data gaps on the toxicity of every mixture may present a large economic burden to the chemical industry.

Some researchers insist that toxicity tests for mixtures are indispensable in validating untested assumptions and simplifications (Borgert, 2004). In practice, however, conducting toxicity tests on all conceivable combinations of chemical substances is unfeasible due to the very large number of possible combinations, as well as the changeable status of chemical combinations in the environment at any time (Cassee *et al.*, 1998; US ATSDR, 2004; Lydy *et al.*, 2004). In addition, toxicological tests using animals are expensive, time-consuming, and raise ethnical issues. Therefore, there is an essential need for appropriate mixture prediction models using knowledge on chemicals in order to facilitate practical chemical risk assessment that satisfies the scientific, regulatory, and industrial perspectives.

How well can we predict mixture toxicity using knowledge of mixture components?

In practice, developing reliable methods for estimating mixture toxicity based on single substances is one of the main challenges in ecotoxicology (Faust and Scholze, 2004). Conventionally two predictive models, including the concentration addition (CA, also referred to as dose addition) and the independent action (IA, also referred to as response addition) models, have been used frequently to estimate the additive toxicity of chemical mixtures with dose-response data of each component (*e.g.,* component-based approaches). The CA (Loewe and Muischnek, 1926) and IA (Bliss, 1939) models are based basically on contrary assumptions: every mixture component has either similar or dissimilar modes of toxic action (MoAs) (Faust *et al.*, 2003). The CA model calculates toxicity in the mixture by summation of the concentrations of each mixture component after modifying the differences in potencies (Loewe and Muischnek, 1926; Finney, 1942; Feron and Groten, 2002).

The IA model predicts mixture toxicity by summation of the responses (*e.g.,* toxicity effects) of each component in a mixture based on the probability theory. The IA model does not consider the contribution of constituents existing at no-effect concentrations into the overall mixture toxicity, in contrast to the CA model (Bliss, 1939; Finney, 1942; Cassee *et al.*, 1998; Feron and Groten, 2002). The overall toxicity calculated by the CA model, especially for low mixture concentrations, can be largely different from that predicted by the IA model (Drescher and Boedeker, 1995). Cedergreen *et al.* (2008) conducted a study that tested the accuracy of the CA and IA models on binary mixtures with various MoAs (*e.g.,* 158 toxicity datasets for 98 different mixtures comprised mainly of pesticides and pharmaceuticals tested on one or more of seven test organisms). The results showed that approximately 20% of the mixtures were properly predicted by the IA model and 10% were correctly estimated by the CA model. Both models could predict the results of another 20% of the testing datasets. Approximately half of the datasets could not be correctly addressed by either of the two models (Cedergreen *et al.*, 2008).

It has been argued that the CA model should be used as a default model from a regulatory point of view for determining aquatic toxicity of mixtures since it is usually more conservative and less data-demanding than the IA model (Arrhenius *et al.*, 2004; Backhaus *et al.*, 2004; Junghans *et al.*, 2006; Cedergreen *et al.*, 2008; Syberg *et al.*, 2009). The number of input parameters used in the calculation process of respective CA and IA models is same, but the type of each parameter used in these models is different: the effective concentration parameter (*e.g.,* EC_{50}) is used in the CA model, and the effect parameter (*e.g.,* effect-%) is used in the IA model. The effect concentration value calculated by the CA model is normally used to describe mixture toxicity in risk assessment rather than the effect estimate of the IA model. The difference between the parameter types mostly makes the IA model more data-demanding. For example, under the concept of CA, the EC_{50} of a mixture can be simply calculated from the EC_{50} of every mixture component. By contrast, according to the number of mixture constituents, the IA model may require full dose-response curves explaining the accurate toxicity responses elicited by the different concentrations of every individual component in order to estimate the EC_{50} of the mixture. If a mixture is based on the equitoxic concentration ratio of 10 components, the $EC_{6.7}$ of each component is needed to estimate the EC_{50} of the mixture of the whole. Nevertheless, common major drawbacks of the CA and IA models can be highlighted by the following background assumptions.

Firstly, in the reality of risk assessment, living organisms and the environment may be exposed to both similarly and dissimilarly acting chemicals simultaneously. However, both CA and IA models do not consider mixed similarly and dissimilarly acting chemical groups to simplify model development (Loewe and Muischnek, 1926; Bliss, 1939; Plackett and Hewlett, 1952; Mwense *et al.*, 2004). Secondly, the use of CA and IA models can be strictly limited unless accurate MoAs of all mixture constituents are readily available (Borgert *et al.*, 2004; Lambert and Lipscomb, 2007). Knowledge of such MoAs remains lacking (European Commission, 2009). Lastly, both models assume that no interactions (*e.g.*, synergism, antagonism, and potentiation) occur among mixture components (Plackett and Hewlett, 1952; Altenburger *et al.*, 2003). Therefore, from a scientific point of view, this leads to a need for developing integrated addition models (IAM) and combined CA and IA concepts, at least for calculating additive toxicity of non-interactive mixtures regardless of whether mixture components produce similar, dissimilar, or both similar and dissimilar MoAs (Mwense *et al.*, 2004).

As an IAM, the Two-Stage Prediction (TSP) model was developed to calculate the toxicity of non-interacting mixtures with different MoA groups (Altenburger *et al.*, 2002; Junghans *et al.*, 2004; Altenburger *et al.*, 2004; 2005). The TSP model executes the CA and IA calculations step by step as follows: (1) mixture constituents are classified into groups in accordance to their MoAs in the first stage so that the CA model is applied to estimate the effective concentrations of each group having similar MoAs; (2) in the second stage, the overall toxicity effect caused from the different groups is predicted by the IA model. From case studies, there is better prediction accuracy with the TSP model for estimating toxicities of mixtures of pesticides, nitrobenzenes, industrial organic compounds, or wastewater treatment plant effluents as compared to the CA and IA models (Junghans *et al.*, 2004; Altenburger *et al.*, 2005; Ra *et al.*, 2006; Wang *et al.*, 2009).

Qin *et al.* (2011) recently developed 'an integrated concentration addition with independent action based on a multiple linear regression (ICIM)' model by applying a multi-linear regression method that merges the CA and IA models for estimating toxicities resulting from 19 mixtures of pesticides and metals. An outstanding advantage of the ICIM model is that the information on MoAs of each component is not required to determine mixture toxicity; rather, only one set of dose-response data for a given mixture and its components is required. From the aspect of model performance, the ICIM may increase the prediction accuracy for estimating the toxicity of target mixtures by using dose-response data of similar mixtures. With respect to data requirement, however, it can be also highlighted that such dose-response data of mixtures are not required by the CA and IA models.

The ICIM model fundamentally uses a standard multi-linear regression (MLR) method based on ordinary least squares (OLS) regression for determining regression coefficients. However, in the case of a linear relationship between any pair of predictor variables (*i.e.*, multicollinearity problems causing high correlations between independent variables in multiple regression), prediction results through the OLS regression cannot be strictly guaranteed to work statistically well despite its ability to calculate good prediction values (Hastie *et al.*, 2001; Adler, 2009). Since the results of the CA and IA models were used as independent variables in the ICIM model (Qin *et al.*, 2011), a question may arise: what is the correlation between the CA and IA models? Related to this issue, Drescher and Boedeker (1995) demonstrated that such a relationship depends on the distribution functions (*e.g.*, Logistic, Weibull, Probit, etc.) for describing dose-response curves, the corresponding slope parameters, and the mixture concentrations administered.

Additionally, the ICIM model is restricted if dose-response curves of a target mixture and its components are not readily available. In the case of the conventional TSP model, its application is restricted to predict mixture toxicity if there is no accurate information of the MoAs for all mixture constituents. These restrictions lead us to the following research question:

How can the limitations of the existing IAM models be overcome?

2. Objectives

The objective of this study was to develop integrated computational models capable of estimating the toxicity of non-interactive mixtures, which overcome the limitations of existing prediction models. These integrated models aim at increasing the accuracy of the conventional models, as well as minimizing the burden of data generation required for model calculations. Therefore, in order to achieve this goal, the following was hypothesized and tested through this study:

i) Hypothesis I: Current approaches, the KCC, and CR methods described in the EU draft technical guidance notes (European Chemical Industry Council, 2005; European Chemical Agency, 2008a, b), for deriving PNECs and DNELs of mixtures, can result in significantly different results due to their contrary concepts. If there is difference between the results of the two methods, these results cannot be validated without testing for a whole mixture;

ii) Hypothesis II: Considering the applicability domain of prediction models for mixture toxicity, the integrated addition concept is more comprehensive than the conventional CA and IA models for estimating the toxicity of non-interactive mixtures consisting of different MoA groups. An advantage of the partial least squares (PLS, also referred to as projection to the latent squares) algorithm is that it offers a valid statistical model in the case of a high degree of multicollinearity between variables. Therefore, the PLS method for MLR can contribute to solving the multicollinearity problem, which can occur in the existing ICIM model when predicting the toxicity of mixtures using the toxicity data of similar mixtures; and,

iii) Hypothesis III: In the absence of MoA knowledge, chemicals can be grouped by their structural similarity due to the relationship between structures and biological activities (*i.e.*, this derives from the Quantitative Structure-Activity Relationship (QSAR) approach that assumes that the function of a substance follows a structural form). Therefore, QSAR techniques used for clustering chemicals can play a role in surmounting the significant disadvantage of the conventional TSP model that strictly requires the knowledge of MoAs of each mixture component.

This study is divided further into four sub-topics as follows:

i) Topic 1: *'Reliable predictive computational toxicology methods for mixture toxicity: Toward the development of innovative integrated models for environmental risk assessment'*, as described in Chapter II, aims to critically describe and summarize recent studies on the prediction models of mixture toxicity in an environmental risk assessment based on the toxicity of single chemicals. The present paper also focuses on integrated models that can be used to predict the toxicity of complex mixtures containing different MoA groups. On the basis of the current review, future challenges and a new research concept to improve the prediction model of mixture toxicity are described in this study. To our knowledge, this represents the first documentation of state-of-the-art computational approaches applied in the development of integrated models using quantum QSARs and machine learning algorithm (MLA).

ii) Topic 2: *'A case study and a computational simulation of the European Union draft technical guidance documents for chemical safety assessment of mixtures: Limitations and a tentative alternative'*, as addressed in Chapter III for 'the hypothesis I', evaluates existing methods, namely the KCC and CR methods, which are described in the EU draft technical guidance (European Chemical Industry Council, 2005; European Chemical Agency, 2008a, b) in order to determine the PNECs and DNELs of mixtures. A case study on coating products, which have different compounds, and a computational simulation were undertaken while considering influencing factors with a focus on the causes of the discrepancy in estimations between the two methods. In addition, this study discussed how the two methods should be considered for regulatory purposes in terms of three aspects: concept, implementation, and performance. Furthermore, as a tentative alternative method, a tiered approach combining 'Enhanced KCC (e-KCC)' and 'CR' methods is proposed and discussed in this study.

iii) Topic 3: *'Development of a partial least squares-based integrated addition model for predicting mixture toxicity'*, as elaborated in Chapter IV for 'the hypothesis II', aims at developing and evaluating a partial least square-based integrated addition model (PLS-IAM) not only to overcome the multicollinearity problem – which can occur between two independent variables (*e.g.*, CA and IA variables) – but also to combine them into an integrated addition model by using the latent variable. According to the original best-fit approach (Scholze *et al.*, 2001), different dose-response curve (DRC) functions were applied to each experimental data, and best-fit functions of each toxicant were employed in the PLS-IAM. The PLS-IAM was validated by four validation datasets. Each dataset consisted of training data for developing a prediction model and test data for validating the developed model. Dataset 1 was experimentally developed in this study for the mixture toxicity of ten pesticides (*e.g.*, five herbicides, four fungicides, and one insecticide) on *Vibrio fischeri*. The other three datasets (Datasets 2, 3, and 4) were derived from previously published studies (Faust *et al.*, 2003; Junghans *et al.*, 2003; Qin *et al.*, 2011) and were additionally used for further validation of the PLS-IAM. Those three datasets were then divided into three types of data: 1) Type 1, representing similarly acting components [Dataset 2: eight chloroacetanilide compounds on *Scenedesmus vacuolatus* (Junghans *et al.*, 2003)]; 2) Type 2, representing dissimilarly acting components [Dataset 3: 16 organics on *Scenedesmus vacuolatus* (Faust *et al.*, 2003)]; and, Type 3, representing a mixture with similarly and dissimilarly acting components [Dataset 4: five herbicides and four metals on *Vibrio qinghaiensis* (Qin *et al.*, 2011)].

iv) Topic 4: *'Development of QSAR-based two-stage prediction model for estimating mixture toxicity'*, as illustrated in Chapter V for 'the hypothesis III', finally aims at developing and evaluating a QSAR-Based TSP (QSAR-TSP) model as an IAM for non-interacting mixtures using the clustering methods that are based on the structural similarity between chemical substances in order to advance the conventional TSP model. In addition, the relatively important molecular descriptors for the chemical clustering were provided by applying the Random Forest (RF) analysis (Breiman, 2001; Shi and Horvath, 2006). Based on the best-fit approach (Scholze *et al.*, 2001), different DRC models were used in every experimental data, and then best-fit functions of each toxicant were employed in the QSAR-TSP model. The QSAR-TSP model was validated by two validation datasets: Dataset 1 was experimentally developed in this study for the mixture toxicity of ten pesticides (five herbicides, four fungicides, and one insecticide) on *Vibrio fischeri* (formerly *Photobacterium phosphoreum*). The following dataset for a complex mixture with similarly and dissimilarly acting components was also used for validation of the QSAR-TSP model: a mixture of 23 pesticides on *Scenedesmus vacuolatus* strain 211-15 (Dataset 2) (Junghans *et al.*, 2006).

Finally, major findings in this study are synthesized in Chapter VI, and final conclusions and further studies needed for validating and advancing the integrated addition models developed through this study are presented.

REFERENCES

Adler, J., 2009. R in a nutshell. O'Reilly Media, Inc., Sebastopol, CA, USA.

Altenburger, R., Greco, R.W., 2008. Extrapolation concepts for dealing with multiple contamination in environmental risk assessment. Integr. Environ. Assess. Manag. 5, 62-68.

Altenburger, R., Nendza, M., Schüürmann, G., 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ. Toxicol. Chem. 22, 1900-1915.

Altenburger, R., Schmitt, H., Schüürmann, G., 2002. Algal toxicity of nitrobenzenes – Combined effect analysis as a pharmacological probe for similar mode of action. In: Predicting toxic effects of contaminants in ecosystems using single species investigations. Habil. Diss., University of Bremen, Bremen., Germany.

Altenburger, R., Schmitt, H., Schüürmann, G., 2005. Algal toxicity of nitrobenzenes: Combined effect analysis as a pharmacological probe for similar modes of interaction. Environ. Toxicol. Chem. 24, 324-333.

Altenburger, R., Walter, H., Grote, M., 2004. What contributes to the combined effect of a complex mixture? Environ. Sci. Technol. 38, 6353-6362.

Arrhenius, A., Grönvall, F., Scholze, M., Backhaus, T., Blanck, H., 2004. Predictability of the mixture toxicity of 12 similarly acting congeneric inhibitors of photosystem II in marine periphyton and epipsammon communities. Aquat. Toxicol. 68, 351-367.

Backhaus, T., Arrhenius, A., Blanck, H., 2004. Toxicity of a mixture of dissimilarly acting substances to natural algal communities: Predictive power and limitations of independent action and concentration addition. Environ. Sci. Technol. 38, 6363-6370.

Bliss, C.I., 1939. The toxicity of poisons applied jointly. Ann Appl Biol 26, 586-615.

Borgert, C.J., 2004. Chemical mixtures: An unsolvable riddle? Hum. Ecol. Risk Assess. 10, 619-629.

Borgert, C.J., Quill, T.F., McCarty, L.S., Mason, A.M., 2004. Can mode of action predict mixture toxicity for risk assessment? Toxicol. Appl. Pharmacol. 201, 85-96.

Breitholtz, M., Nyholm, J.R., Karlsson, J., Andersson, P.L., 2008. Are individual NOEC levels safe for mixtures? A study on mixture toxicity of brominated flame-retardants in the copepod Nitocra spinipes. Chemosphere 72, 1242-1249.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5-32.

Cassee, F.R., Groten, J.P., van Bladeren, P.J., Feron, V.J., 1998. Toxicological evaluation and risk assessment of chemical mixtures. Crit. Rev. Toxicol. 28, 73-101.

Cedergreen, N., Cristensen, A.M., Kamper, A., Kudsk, P., Mathiassen, S.K., Streibig, J.C., Sørensen, H., 2008. A rivew of independent action compared to concentration addition as reference models for mixtures of compounds with different molecular target sites. Environ. Toxicol. Chem. 27, 1621-1632.

Drescher, K., Boedeker, W., 1995. Assessment of the combined effects of substances: The relationship between concentration addition and independent action. Biometrics 51, 716-730.

Eggen, R.I., Behra, R., Burkhardt-Holm, P., Escher, B.I., Schweigert, N., 2004. Challenges in ecotoxicology. Environ. Sci. Technol. 38, 58A-64A.

European Chemical Agency, 2008a. Guidance on information requirements and chemical safety assessment chapter R.8: Characterisation of dose (concentration)-response for human health. Guidance for the implementation of REACH.

European Chemical Agency, 2008b. Guidance on information requirements and chemical safety assessment chapter R.10: Characterisation of dose (concentration)-response for environment. Guidance for the implementation of REACH.

European Chemical Industry Council, 2005. Considerations on Safety Data Sheets and Chemical Safety Assessments of Preparations – Final report. Service Contract Nos. 22551-2004-12 F1SC ISP BE and 22552-2004-12 F1SC ISP BE.

European Commission, 2003. Technical guidance document in support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No. 1488/94 on Risk Assessment for Existing Substances. Luxembourg. ISBN:93-827-8011-2.

European Commission, 2009. State of the art report on mixture toxicity - Final report. The school of Pharmacy, University of London, London, UK.

Faust, M., Altenburger, R., Backhaus, T., Blanck, H., Boedeker, W., Gramatica, P., Hamer, V., Scholze, M., Vighi, M., Grimme, L.H., 2003. Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action. Aquat. Toxicol. 63, 43-63.

Faust, M., Scholze, M., 2004. Competing concepts for the prediction of mixture toxicity: Do the difference matter for regulatory purposes? Final Report of the European R&D Project

BEAM (EVK1-CT1999-00012), Bremen, Germany.

Feron, V.J., Groten, J.P., 2002. Toxicological evaluation of chemical mixtures Food Chem. Toxicol. 40, 825-839.

Finney, D.F., 1942. The analysis of toxicity tests on mixtures of poisons Ann. Appl. Biol. 29.

Hartung, T., Rovida, C., 2009. Chemical regulators have overreached.. Nature 460, 1080-1081.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning: Data mining, inference, and prediction. Springer, New York, NY, USA.

Junghans, M., Backhaus, T., Faust, M., Meyer, W., Scholze, M., Grimme, L.H., 2004. Predicting the joint algal toxicity of chemical mixtures using a mechanism based two stage prediction (TSP). In: Studies on combination effects of environmentally relevant toxicants. Ph.D Diss., University of Bremen, Germany.

Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2003. Predictability of combined effects of eight chloroacetanilide herbicides on algal reproduction. Pest Manag. Sci. 59, 1101-1110.

Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2006. Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. Aquat. Toxicol. 76, 93-110.

Kortenkamp, A., Altenburger, R., 1999. Approaches to assessing combination effects of oestrogenic environmental pollutants. Sci. Total Environ. 233, 131-140.

Lambert, J.C., Lipscomb, J.C., 2007. Mode of action as a determining factor in additivity models for chemical mixture risk assessment. Regul. Toxicol. Pharmacol. 49, 183-194.

Loewe, S., Muischnek, H., 1926. Über Kombinationswirkungen I. Mitteilung: Hilfsmittel der Fragestellung. . Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol. 114, 313-326.

Lydy, M., Belden, J., Wheelock, C., Hammock, B., Denton, D., 2004. Challenges in regulating pesticide mixtures. Ecol. Soc. 9, 1.

Martin, H.L., Svendsen, C., Lister, L.J., Gomez-Eyles, J.L., Spurgeon, D.J., 2009. Measurement and modelling of the toxicity of binary mixtures in the *nematode Caenorhabditis elegans* - a test of independent action. Environ. Toxicol. Chem. 28, 97-104.

Mwense, M., Wang, X.Z., Buontempo, F.V., Horan, N., Young, A., Osborn, D., 2004. Prediction of noninteractive mixture toxicity of organic compounds based on a Fuzzy set method. J. Chem. Inf. Comput. Sci. 44, 1763-1773.

Plackett, R.L., Hewlett, P.S., 1952. Quantal responses to mixtures of poisons. J. Royal Stat. Soc. B 14, 141-163.

Qin, L.-T., Liu, S.-S., Zhang, J., Xiao, Q.-F., 2011. A novel model integrated concentration addition with independent action for the prediction of toxicity of multi-component mixture. Toxicology 280, 164-172.

Ra, J.S., Lee, B.C., Chang, N.I., Kim, S.D., 2006. Estimating the combined toxicity by two-step prediction model on the complicated chemical mixtures from wastewater treatment plant effluents. Environ. Toxicol. Chem. 25,2107-2113.

Rajapakse, N., Silva, E., Kortenkamp, A., 2002. Combining xenoestrogens at levels below individual no-observed-effect concentrations dramatically enhances steroid hormone action. Environ. Health Perspect. 110, 917-921.

Scholze, M., Boedeker, W., Faust, M., Backhaus, T., Altenburger, R., Grimme, L.H., 2001. A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. Environ. Toxicol. Chem. 20, 448-457.

Shi, T., Horvath, S., 2006. Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. 15, 118-138.

Syberg, K., Jensen, T.S., Cedergreen, N., Rank, J., 2009. On the use of mixture toxicity assessment in REACH and the Water Framework Directive: A review. Hum. Ecol. Risk Assess. 15, 1257 - 1272.

US ATSDR, 2004. Guidance manual for the assessment of joint toxic action of chemical mixtures. U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry (ATSDR). Atlanta, GA, USA.

Vighi, M., Altenburger, R., Arrhenius, Backhaus, T., Bödeker, W., Blanck, H., Consolaro, F., Faust, M., Finizio, A., Froehner, K., Gramatica, P., Grimme, L.H., Grönvall, F., Hamer, V., Scholze, M., Walter, H., 2003. Water quality objectives for mixtures of toxic chemicals: problems and perspectives. Ecotoxicol. Environ. Saf. 54, 139-150.

Walter, H., Consolaro, F., Gramatica, P., Scholze, M., Altenburger, R., 2002. Mixture toxicity of priority pollutants at no observed effect concentrations (NOECs). Ecotoxicology 11, 299-310.

Wang, Z., Chen, J., Huang, L., Wang, Y., Cai, X., Qiao, X., Dong, Y., 2009. Integrated fuzzy concentration addition-independent action (IFCA-IA) model outperforms two-stage prediction (TSP) for predicting mixture toxicity. Chemosphere 74, 735-740.

Xu, S., Nirmalakhandan, N., 1998. Use of QSAR models in predicting joint effects in multi-component mixtures of organic chemicals. Water Res. 32, 2391-2399.

CHAPTER II

Reliable Predictive Computational Toxicology Methods for Mixture Toxicity: Toward the Development of Innovative Integrated Models for Environmental Risk Assessment

Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, 2012. *Reviews in Environmental Science and Bio/Technology*, DOI:10.1007/s11157-012-9286-7. Published online (27 June 2012).

Note by the author

This chapter is based on the following journal publication. Due to the copyright issue, the text of the chapter was replaced by the reference information and its web link to the published paper. Thus, the interested reader is kindly asked to read the published paper via the following reference:

Kim, J., Kim, S., Schaumann, G.E., 2012. Reliable predictive computational toxicology methods for mixture toxicity: Toward the development of innovative integrated models for environmental risk assessment. *Reviews in Environmental Science and Bio/Technology*, DOI: 10.1007/s11157-012-9286-7. Published online (27 June 2012).

Click here to see the paper on the publisher's website

CHAPTER III

A Case Study and a Computational Simulation of the European Union Draft Technical Guidance Documents for Chemical Safety Assessment of Mixtures: Limitations and a Tentative Alternative

Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann., 2013. *Journal of Occupational & Environmental Hygiene*, 10:181-193.

Note by the author

This chapter is based on the following journal publication. Due to the copyright issue, the text of the chapter was replaced by the reference information and its web link to the published paper. Thus, the interested reader is kindly asked to read the published paper via the following reference:

Kim, J., Kim, S., Schaumann, G.E., 2013. A case study and a computational simulation of the European Union draft technical guidance documents for chemical safety assessment of mixtures: Limitations and a tentative alternative. *Journal of Occupational & Environmental Hygiene*, 10:181-193.

Click here to see the paper on the publisher's website

CHAPTER IV

Development of a Partial Least Squares-Based Integrated Addition Model for Predicting Mixture Toxicity

Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, 2012. *Human and Ecological Risk Assessment: An International Journal*, DOI:10.1080/10807039.2012.754312. Accepted author version posted online (7 December 2012).

Note by the author

This chapter is based on the following journal publication. Due to the copyright issue, the text of the chapter was replaced by the reference information and its web link to the published paper. Thus, the interested reader is kindly asked to read the published paper via the following reference:

Kim, J., Kim, S., Schaumann, G.E., 2012. Development of a partial least squares-based integrated addition model for predicting mixture toxicity. *Human and Ecological Risk Assessment: An International Journal*. DOI:10.1080/10807039.2012.754312. Accepted author version posted online (7 December 2012).

Click here to see the paper on the publisher's website

CHAPTER V

Development of a QSAR-Based Two-Stage Prediction Model for Estimating Mixture Toxicity

Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann. *SAR and QSAR in Environmental Research*, DOI:10.1080/1062936X.2013.815654. Accepted manuscript (Accepted on 29 April 2013).

Development of QSAR-based two-stage prediction model for estimating mixture toxicity

Jongwoon Kim [a,b*], Sanghun Kim [b], and Gabriele E. Schaumann [a]

[a] *Institute of Environmental Sciences, University of Koblenz–Landau, Fortstrasse 7, D-76829 Landau, Germany*

[b] *KIST Europe, Korea Institute of Science and Technology, Campus E 7.1, D-66123 Saarbruecken, Germany*

* Corresponding author, J. Kim, contact information:

E-mail: jwkim@kist-europe.de

Phone: +49(681) 9382-322

Fax: +49(681) 9382-319

The 2nd author, S. Kim, contact information:

E-mail: shkim@kist-europe.de

Phone: +49(681) 9382-334

Fax: +49(681) 9382-319

The 3rd author, G. E. Schaumann, contact information:

E-mail: schaumann@uni-landau.de

Phone: +49(6341) 280-31571

Fax: +49(6341) 280-31576

Abstract

Conventionally, concentration addition (CA) and independent action (IA) models based on additive toxicity are often used to estimate the mixture toxicity of similarly- and dissimilarly-acting chemicals, respectively. Two-stage prediction (TSP) model has been developed as an integrated addition model that can perform the CA and IA calculations stage by stage. But, the use of the conventional TSP model is limited if the modes of toxic action (MoAs) for every mixture component is not readily known. The objective of this study was to develop and evaluate a quantitative structure-activity relationship-based TSP (QSAR-TSP) model for estimating mixture toxicity in the absence of knowledge on the MoAs of the constituents. For this purpose, different clustering methods of mixture constituents using computerised analysis based on the structural similarity between chemicals were applied as a part of the predictions of mixture toxicity. The relative importance of molecular descriptors was additionally determined by Random Forest analysis. This study highlights the prediction power of the QSAR-TSP model and its potential to overcome the limitations of the conventional TSP model, and how clustering methods of mixture components that employ chemical structural information to categorize might be applied to predict mixture toxicity effectively.

Keywords: mixture toxicity; QSAR; two-stage prediction; integrated addition model; concentration addition; independent action

1. Introduction

The concentration addition (CA) [1] and independent action (IA) [2] models are mainly used to predict the additive toxicity of mixture components, and were basically established on opposite assumptions: a mixture consists of components having either similar or dissimilar modes of toxic action (MoAs), respectively. Equations (1) and (2) define the model concepts of CA and IA, respectively [3]:

$$ECx_{mix} = (\sum_{i=1}^{n} \frac{p_i}{ECx_i})^{-1} \tag{1}$$

$$E(C_{mix}) = 1 - \prod_{i=1}^{n} [1 - E(C_i)] \tag{2}$$

where C_i is the concentration of the ith substance in a mixture with n components (i = 1...n); C_{mix} is the total concentration of substances in the mixture; ECx_i is the concentration of the ith substance that causes the effect x when applied individually; ECx_{mix} is the total concentration of substances in a mixture that causes the total effect x; $E(C_i)$ is the individual effect of the ith substance if present in the concentration C; $E(C_{\mathrm{mix}})$ is the total effect of the mixture with the total concentration C_{mix} of the mixture components; x is the definite value for the effect E; and, p_i is the relative proportion of the ith substance expressed as a fraction of the total concentration of substances in the mixture ($p_i = C_i/C_{\mathrm{mix}}$).

The prediction capability of both models can be strictly limited if accurate MoAs for all constituents are not readily available, even if the models' assumptions are reasonable [4,5]. To overcome this commonly encountered limitation of the CA and IA models, Qin *et al.* [6] developed 'an integrated concentration addition with independent action based on a multiple linear regression (ICIM)' model by applying a multi-linear regression method to combine the CA and IA models. The results of the respective CA and IA models were employed as independent variables in the ICIM model, which showed good prediction accuracy for estimating toxicities derived from 19 mixtures consisting of 5 pesticides and 4 metals [6]. An outstanding advantage of the ICIM model was that MoA information for each component was not required to determine mixture toxicity; rather, only one set of dose-response data for a given similar mixture and its components was required. However, the statistical robustness of prediction results obtained through the ICIM model based on ordinary least squares are not guaranteed in cases of high correlation between the results of the CA and IA models when used as independent variables (*i.e.*, the multicollinearity problem (for detailed information of the multicollinearity problem, refer to Hastie *et al.* [7] and Adler [8])) [9]. The Partial Least Square-Based Integrated Addition Model (PLS-IAM) was developed by Kim *et al.* [9] to maintain the advantages of the ICIM model while overcoming the multicollinearity problem that plagues it. The researchers showed that, compared with the existing CA, IA, and ICIM models, the PLS-IAM showed excellent predictive performance for toxicities of mixtures consisting of pesticides, organic compounds, or metals. However, it also bears mentioning that the application of the PLS-IAM is limited in cases where there is no dose-response data for a mixture having a similar composition to the target mixture.

Avoiding this weakness, the Two-Stage Prediction (TSP) model [10-13], an integrated addition model (IAM) that requires no toxicity data on a similar mixture, was developed to calculate the toxicity of non-interacting mixtures with different MoA groups. The TSP model executes the CA and IA calculations step by step: in the first stage, mixture constituents are classified into groups in accordance with their MoAs so that the CA model can be applied to estimate the effect concentrations of each of these groups; in the second stage, the overall toxicity effect caused by the different groups is predicted using the IA model. Case studies on the application of the TSP model have demonstrated better predictive performances for estimating toxicities of mixtures of pesticides, nitrobenzenes, industrial organic compounds, or wastewater treatment plant effluents than were achieved by the CA and IA models [11,13-15]. In the case of the conventional TSP model, however, its application is restricted if there is no accurate information about the MoAs of all mixture constituents.

This restriction leads to the following research question: How can the critical limitations of the IAMs (*e.g.*, the PLS-IAM and TSP models) be overcome? To find a possible answer to this question, we suggest that the main drawback of the TSP model (*i.e.*, that it requires MoA data for all mixture constituents) might be solved by clustering chemicals based on their structural similarities, instead of by MoAs, if robust relationships between structures and biological activities exist. In practice, quantitative structure activity relationship (QSAR) approaches, which are widely used in chemistry, pharmacology, toxicology, and other related fields, assume that the function of a substance follows its structural form: *i.e.*, a chemical's characteristics and biological responses to it are closely related to its molecular structure [16-18]. For example, QSARs for ecotoxicity describe mathematical relationships between molecular structure descriptors (*e.g.*, K_{ow}, LUMO, and HOMO) and ecotoxicological endpoints (*e.g.*, EC_{50} and LC_{50}) [19]. Recently, some studies on the QSAR analysis directly dealing with mixtures are being conducted by using mixture descriptors: *e.g.*, descriptors based on partition coefficients for mixtures, integral (whole-molecule) additive descriptors (*e.g.*, weighted sum of descriptors of components), integral non-additive descriptors of mixtures (*e.g.*, components are considered in a different approaches from the additive concept), and fragment non-additive descriptors (*e.g.*, structural parts of components are considered in same descriptor) [20] (for detailed information of the mixture descriptor type, refer to Muratov *et al.* [20]). Based on this understanding, we hypothesised the following:

- Considering the applicability domain of prediction models for mixture toxicity, the integrated addition concept is more comprehensive than conventional CA and IA models for estimating the toxicity of non-interacting mixtures consisting of different MoA groups;
- In the absence of knowledge regarding MoAs, chemicals can be grouped by structural similarity due to the robust relationship between the structures and biological activities of chemicals (*i.e.*, through QSAR approaches that assume that the function of a substance follows its structural form);
- Therefore, QSAR techniques used for clustering chemicals can play a role in surmounting the primary disadvantage of the conventional TSP model: its strict requirements that the MoAs of each mixture component must be known.

The objectives of this study were to develop and evaluate a QSAR-based TSP (QSAR-TSP) model as an IAM for non-interacting mixtures using clustering methods that classify based on the structural similarity between chemical substances in order to improve and advance the conventional TSP model. In addition, the relative important molecular descriptors for the chemical clustering were provided by applying Random Forest (RF) Analysis [21,22]. The QSAR-TSP model was validated by two datasets: Dataset 1, which was our previously published study [9] for the mixture toxicity of ten pesticides (five herbicides, four fungicides, and one insecticide) on *Vibrio fischeri* (formerly *Photobacterium phosphoreum*); and, Dataset 2, which was previously published by Junghans *et al.* [23] for a realistic pesticide mixture consisting of 23 pesticides on *Scenedesmus vacuolatus* strain 211-15.

2. Materials and Methods

2.1 Datasets

2.1.1 Dataset 1

Dataset 1, our previously published study [9], includes ten pesticides (*i.e.*, five herbicides, four fungicides, and one insecticide) widely used in European agricultural areas. These were selected as mixture components due to their different MoAs. Toxicity of the tested compounds was evaluated using the bioluminescent bacteria *Vibrio fischeri* in a short-term bioluminescence assay. However, not all MoAs in Dataset 1 originated from the test organism, *Vibrio fischeri* (as shown in Table 1). Physical properties, MoAs, and parameters for regression models of dose-response curves on *Vibrio fischeri* for ten pesticide chemicals are listed in Table 1.

The mixture toxicity test was conducted in a fixed ratio design. The fixed ratio design is often used to maximize the distribution of effect concentration range, to minimize the experiments needed to be performed, and to be suitable for a mixture with multiple components and multiple levels [15,24-30]. Table 1 presents the toxicity data on two mixtures with different mixture ratios (*i.e*., two equitoxic mixtures), which were examined based on the relative toxicity of each individual mixture component. The first equitoxic mixture (Mixture 1: EC_{50} mixture) mixed at 50% of the effective concentration (EC_{50}) of each component, and the second equitoxic mixture (Mixture 2: EC_{10} mixture) based on the 10% effective concentrations (EC_{10}) of components were employed as the model validation data. A "best-fit" approach [31] was used to select the best DRC models for each component and the mixtures to which they belonged. The best-fit regression models are shown in Table 1.

2.1.2 Dataset 2

Dataset 2, a 23-component mixture (Mixture 3) derived in a published study [23] reflecting a realistic exposure scenario in field run-off water, was also tested to provide additional validation for the proposed QSAR-TSP model. Table 3 shows the molecular weight, MoAs, and parameters of regression models for DRCs on *Scenedesmus vacuolatus* strain 211-15 for the realistic pesticide mixture. A total of eight MoAs, including an unknown MoA (on Carbofuran) found in Dataset 2, originated from the target organism, *Scenedesmus vacuolatus*. Detailed information regarding the organism and the testing conditions can be found in the original paper [23].

2.2 Development of the QSAR-TSP Model

The QSAR-TSP model requires no information on the MoAs of all mixture components; however, it does require the components' DRCs and chemical structures. Figure 1 shows the scheme of the QSAR-TSP model for estimating the toxicity of non-interacting mixtures. The QSAR-TSP modelling is basically divided into three sub-modules as follows:

(1) Module 1 (DRC modelling): For mixtures containing *n* constituents, the DRCs of all constituents were derived by applying sigmoidal regression functions selected to best describe the DRCs. These regression functions were used in the CA and IA calculation steps involved in the QSAR-TSP model for estimating the mixture toxicity.

(2) Module 2 (descriptor-based chemical clustering): The performance of the following four descriptor-based clustering methods were evaluated in the context of the QSAR-TSP model in this study: i) the *k*-means clustering via PCA; ii) the PAM clustering via PCA; iii) the k-means clustering via RF; and, iv) the PAM clustering via RF. For calculating molecular structural descriptors of all components, their molecular structures were modelled with CS ChemBio3D Ultra Ver. 12.0 (Cambridge, UK) in this study. The geometrical optimization of the chemical structures was conducted on the basis of the Parameterized Model number 6 (PM6) algorithm [32] within the MOPAC interface [33], a semi-empirical quantum chemistry program. The PM6 is a semi-empirical method developed from experimental and *ab initio* data (*i.e.*, modelled data) from over 9,000 chemicals that is used to perform quantum calculations of molecular electronic structures [32]. The software DRAGON Ver. 6.0 (Talete s.r.l, Italy) was employed to calculate the molecular structural descriptors. The principal component analysis (PCA) method [8,34] was used to reduce the number of molecular descriptors and thus the classification performance. The PCA technique is a mathematical procedure that transforms a large number of input variables into a set of fewer uncorrelated variables called principal components (PCs), which explains the total data while minimizing information loss [35]. However, it is frequently a difficult task to interpret what the respective PCs after compression by PCA mean due to the transformation of the original data [34,36]. Thus, the RF clustering method [37], which uses two specific importance measures, mean decrease accuracy (MDA) and mean decrease Gini (MDG) index [38-40], was additionally applied to find relative important descriptors. The RF method is an ensemble classifier consisting of many decision trees [41]. The MDA and MDG index can be used as general indicators of variable relevance, and their scores provide a relative ranking of the variables [40,42] (for detailed information on the calculation methods of the two importance measures, see Breiman and Cutler [39]). The corresponding Euclidean distance [43,44], based on principal component scores, and RF distance [22], based on the ranks of all descriptors, were computed to quantify the degree of structural similarity between each pair of mixture components. For the RF analysis, the RF decision tree algorithm [22] was used (the number of forests = 50; the number of trees = 1,000). In the present study, the similarities characterised by the respective distance values were applied to two cluster analysis methods, *k*-means and partitioning around medoids (PAM) algorithms, which are widely used as clustering techniques [8]. The aim of cluster analysis was to partition the observed data into several groups (*i.e.*, clusters) so that the similarities between data allocated to the same cluster tend to be

larger than between data across different clusters [7]. The *k*-mean clustering method [45] is an algorithm for determining clusters and cluster centres in a set of unlabeled data by optimising distances between objects and the centroids of clusters. The *k*-means procedure interactively moves the centroids to minimise the total cluster variance (the "*k*" in *k*-means refers to the number of cluster centres) [7]. The *k*-means algorithm calculates the means of objects in respective clusters to be the centroid of the clusters, whereas the PAM algorithm selects representative objects (also referred to as medoids), minimising a sum of dissimilarities for each cluster to create the cluster centres [46]. The clustering methods enabled mixture components with similar structures to be assigned to common clusters.

(3) Module 3 (mixture toxicity prediction): The toxicity of a given mixture was estimated by performing the CA and IA calculations step by step. In the first step, the total effective concentration of a given mixture of components in each cluster was determined by applying the CA model shown in Equation (1). The mixture toxicity from different clusters was calculated using the IA model shown in Equation (2). This TSP can be defined in Equation (3):

$$
\begin{aligned}
E(C_{mix,mix}) &= 1-\{(1-E(C_{CL1}))(1-E(C_{CL2}))...(1-E(C_{CL\,n}))\\
&= 1-\prod_{i=1}^{n}(1-E(C_{mix,CLi}))
\end{aligned}
\qquad (3)
$$

where $C_{mix,CLi}$ is the total concentration of the i_{th} cluster (C_{CLi}) having similar chemical structures; $E(C_{mix,CLi})$ is the mixture effect at $C_{mix,CLi}$; and, $E(C_{mix,mix})$ is the combined effect from different clusters.

Data analysis, statistical calculations, and clustering procedures used in the process of developing the QSAR-TSP model were performed with R software ver. 2.12.1 [47], a programming language and environment for statistical computing and graphics.

2.3 Validation of the QSAR-TSP Model

In this study, the DRCs of given complex mixtures in Datasets 1 and 2 shown in Tables 1 and 2 were used for the validation of the QSAR-TSP model. The prediction accuracy of the QSAR-TSP model was validated with the coefficient of determination for the modelled data (R^2_{test}) as well as the residual sum of squares (RSS). The R^2_{test} can be defined in Equation (4):

$$R^2_{test} = 1 - (\frac{SSE}{SST}) \tag{4}$$

where R^2_{test} is the coefficient of determination, *SSE* is the sum of squares of residuals, and *SST* is the total sum of squares.

The silhouette validation method was used to valid the model's determination of optimal cluster sizes for the four different clustering algorithms tested—namely, the *k*-means via PCA, PAM via PCA, k-means via RF, and PAM via RF [46,48]. This technique computes the average silhouette width for each cluster and the overall average silhouette width for a total dataset by comparing the tightness and separation of silhouettes [48]. The average silhouette width value is a measure of average geometric distances between elements in a given cluster that can help describe to what extent individual elements belong to their own clusters; it is often used for evaluating cluster validity and verifying the best number of clusters for datasets.

2.4 Evaluation of the QSAR-TSP Model

The Akaike's Information Criteria (AIC) [49] are frequently used to evaluate the performance of predictive models [50]. The AIC explains the goodness of fit of predictive models and penalises high numbers of regression parameters to avoid over-fitting (for more detailed information, see Burnham and Anderson [50]). The AIC can be described in Equation (5):

$$AIC = n\,ln(\frac{RSS}{n}) + 2K \tag{5}$$

where *n* is the number of observations in the data, *RSS* is the residual sum of squares of the model, and *K* is the number of model parameters.

For a comparison of the prediction capability of the CA, IA, TSP, and QSAR-TSP models, both *n* and *K* are kept constant for each of the four model fits. Thus, in this study, the difference in the AIC scores calculated from each model fit depends on the residual sum of squares only. The predictive model with the highest R^2_{test} and smallest *RSS* was selected as the best-fitting model.

Through comparing the QSAR-TSP model with the other reference models, the advantages and disadvantages of the QSAR-TSP were debated in three aspects: model performance, data availability, and application coverage. First, with respect to model performance, the efficacy of estimating mixture toxicity with the models was assessed. Second, with respect to data availability, the type of input data needed to be employed in the models was considered. Finally, the aspect of application coverage was discussed in terms of what types of mixtures can be considered under the models theoretically.

3. Results and Discussion

3.1 Feature generation and molecular descriptor-based chemical clustering

The software DRAGON (Ver. 6.0) computed 4,870 molecular descriptors from each compound in Datasets 1 and 2. After excluding descriptors with all values equal (*i.e.*, constant values) among the total descriptors, 2,920 descriptors from Dataset 1 and 3,154 descriptors from Dataset 2 were finally selected for the PCA and RF analyses, respectively. The optimum number of principle components (PCs) extracted by the PCA technique was determined when the PCA found the smallest set of PCs maximising the variance of transformed variables which could account for the original datasets as much as possible. The PCA extracted 9 and 20 PCs, which explained 100% and 99.4% of variances of the original molecular descriptors for the substances in Datasets 1 and 2, respectively (Tables S1 and S2). The PCs were used for quantifying the intermolecular Euclidean distances as the molecular distance geometry between any pair of substances in Datasets 1 and 2 (Tables S3 and S4). In calculating the intermolecular RF distances, all the original descriptors were used to find the relative important descriptors in clustering components (Tables S5 and S6).

Table 4 shows clustered results from the *k*-means and PAM clustering algorithms based on the intermolecular Euclidean and RF distances computed by the PCA and RF methods applied to Datasets 1 and 2. As presented in Table 4, the average silhouette width scores were calculated to determine the optimal cluster size through the applications of the *k*-means and PAM clustering algorithms via the PCA and RF methods. For Dataset 1, all clustering methods produced similar results, showing the optimal number of clusters to be the two with the largest average silhouette width. Therefore, two clusters were determined as the optimal size for Dataset 1. By contrast, the PCA-based (the *k*-means via PCA and PAM via PCA) and RF-based (the *k*-means via RF and PAM via RF) clustering methods showed different results in the case of Dataset 2. The PCA-based methods showed that three clusters were the best number for Dataset 2, but the RF-based methods demonstrated that two clusters were the optimal size. Therefore, two and three clusters were respectively applied in the QSAR-TSP calculations to predict the toxicity of Mixture 3 in Dataset 2. A comparison of the best number of clusters as calculated from the *k*-means-based (*i.e.*, the *k*-means via PCA and *k*-means via RF) versus the PAM-based clustering methods arrived at no significant differences between Datasets 1 and 2.

Figure 2(a)-(c) illustrates how mixture components in Datasets 1 and 2 were grouped into each cluster through the four PCA- and RF-based methods. Figure 2(a) shows the mixture components clustered into two clusters for Dataset 1. Figures 2(b) and 2(c) present the components clustered into two and three clusters for Dataset 2, respectively. In this study, among the four clustering methods, the *k*-means via PCA method showed the best discretisation performance for clustering chemicals, yielding the highest average silhouette width scores of 0.40 and 0.44 for Datasets 1 and 2, respectively. The average silhouette width values estimated by the *k*-means- and PAM-based clustering methods were slightly different, as shown in Table 4. However, the clusters reached by those methods contained exactly the same chemicals, regardless of whether PCA and RF techniques were used.

Therefore, it was concluded that the PCA- and RF-based methods were capable of showing different optimal cluster sizes for same-mixture compositions, but no differences between the clustered results from the *k*-means- and PAM-based clustering methods for the datasets used in this study were found. It is also notable that the *k*-means-based clustering method was shown to elicit higher average silhouette scores, and thus it might be more useful than the PAM-based method for clustering chemicals in the PCA- and RF-based methods in terms of its discretisation performance. Nevertheless, further research is still required to find which of these constitutes the more useful clustering method.

3.2 Finding relative importance descriptors

Through the RF analysis, the relative importance descriptors for clustering chemicals based on the structural similarity calculated between mixture components in Datasets 1 and 2 were found. Figure 3 illustrates the 20 most important descriptors for clustering chemicals in Datasets 1 and 2 with the MDA and MDG Index. On the basis of 2,920 molecular descriptors calculated from the chemical structures of the compounds in Dataset 1, the two plots of MDA and MDG [Figures 3(a) and 3(b)] had a common important descriptor in the solid line rectangle, ‘SpMax_Dz.e.’, a two-dimensional (2-D) matrix-based descriptor—the Barysz matrix weighted by Sanderson electronegativity [51,52]. 2-D matrix-based descriptors are topological indices computed by applying a set of algebraic operators to different graph-theoretical matrices denoting a hydrogen-depleted molecular graph obtained excluding all the hydrogen atoms [51,52]. Barysz matrices are symmetric weighted distance matrices explaining the presence of both heteroatoms and multiple bonds in the molecule [51,52]. In addition, three relative importance descriptors in solid line and dashed line rectangles in Figures 3(a) and 3(b) could be categorised into a common block as 2-D matrix-based descriptors (Tables S7 and S8).

In the case of Dataset 2, from 3,154 descriptors, the RF analysis on the top 20 relative importance descriptors found 5 common descriptors in solid line rectangles in the MDA and MDG plots [Figures 3(c) and 3(d)]: ‘SpAbs_B(e)’, SpDiam_Dz(m)’, ‘SpMaxA_Dz(v)’, ‘VE3_Dz(p)’, and ‘Wi_Dz(v)’. Those five common descriptors were 2-D matrix-based descriptors with one of Barysz matrices or Burden matrices (Table S9 and S10). Burden matrices are augmented adjacency matrices (*e.g.*, vertex matrices), mainly encoding information on the vertex (*i.e.* atom) connectivity and the distance matrix, obtained from a hydrogen-depleted molecular graph [51,52]. Furthermore, ten relative importance descriptors marked in solid line and dashed line rectangles in Figures 3(c) and 3(d) could be categorised into four common sub-blocks of 2-D matrix-based descriptors as follows: Burden matrix weighted by Sanderson electronegativity, Barysz matrix weighted by mass, Barysz matrix weighted by van der Waals volume, and Barysz matrix weighted by polarisability [51,52] (Tables S9 and S10).

The descriptions of the top 20 descriptors found in Datasets 1 and 2 are presented in Tables S7, S8, S9, and S10 in the supplementary material. Details on the descriptors and sub-blocks are given in references [51,52]. This study showed that the common important descriptors derived from Datasets 1 and 2 were all involved in the 2-D matrix-based descriptor categories based on Barysz distance matrices: *i.e.*, those descriptors highly contributed to discriminate among the molecular structures of mixture components. This finding also provides a possibility that the important descriptors based on Barysz distance matrices may be available as priority candidates to develop QSAR models for the datasets employed in this study. However, additional studies are needed to investigate if any specific relationship exists between descriptors, based on Barysz matrices, and toxicological responses derived from test organisms used in this study.

3.3 Mixture toxicity prediction and validation

This section presents the predictive performance and validation of the QSAR-TSP model used for estimating the toxicity of the three mixtures from Datasets 1 and 2. According to the clustering results for these datasets, Equation (3) was employed to estimate the mixture toxicity. In the case of Mixtures 1 and 2 in Dataset 1, the components fenamidone, cyanazine, MCPA, furalaxyl, and thiabendazole, could not dissolve into water at a higher concentration level to elicit an 80% or more toxicity effect (refer to Kim *et al.* [9]) due to their solubility limits in water. Thus, the mixture toxicity estimated by the QSAR-TSP model was validated by experimentally-observed data ranging from 5% to 75% of the effect, and these results were compared with those of reference models: namely, the CA, IA, and conventional TSP models. In the case of Mixture 3 in Dataset 2, the QSAR-TSP model was validated by observed data ranging from 2% to 97% of the effect.

Figure 4(a) shows the comparison results of the four prediction models for Mixture 1, the EC_{50} ratio mixture, which consisted of ten components (five herbicides, four fungicides, and one insecticide). The best prediction capability was found in the results of the QSAR-TSP model (R^2_{test} = 0.947, RSS = 3.70E+02) for Mixture 1, with the CA model showing a weaker result (R^2_{test} = 0.749, RSS = 1.76E+03). Interestingly, the conventional TSP model did not estimate the mixture toxicity correctly (R^2_{test} = 0.158, RSS = 5.89E+03). This result implies that correct MoAs of a test organism might be unavailable for the TSP model, by corroborated the fact that none of the MoAs listed in Dataset 1 originated from the test organism, *Vibrio fischeri*. The IA model had a negative R^2_{test} value (RSS = 9.58E+03), which would be equivalent to having no explained variation at all [53]. The QSAR-TSP and CA models overestimated the toxicity of Mixture 1 in the high effect range (> 40%), yet underestimated it in the low effect range (< 40%). However, in the whole effect range, the deviation between observed and predicted values from the QSAR-TSP model was relatively small as compared to the CA model. Also, the modelled values of the QSAR-TSP model were located within the standard deviation (SD) range.

Figure 4(b) illustrates the comparison results for Mixture 2, which is the EC_{10} ratio mixture of the same components as in Mixture 1. For Mixture 2, the QSAR-TSP model had the highest prediction capability (R^2_{test} = 0.923, RSS = 5.39E+02), but the CA model also predicted the toxicity of Mixture 2 well (R^2_{test} = 0.876, RSS = 8.68E+02). The conventional TSP model, which again was based on incorrect MoAs that did not originate from the test organism, and the IA model did not correctly calculate the toxicity of Mixture 2. The TSP and IA models showed much lower R^2_{test}s (0.337 and 0.034, respectively) and higher RSSs (4.64E+03 and 6.76E+03, respectively) than the QSAR-TSP and CA models. The CA model underestimated the toxicity of Mixture 2 in the effect range of up to 30%, but overestimated it at 30% or more. For the QSAR-TSP model, the toxicity of Mixture 2 was underestimated in the overall effect range.

Figure 4(c) shows the comparison results for Mixture 3, a realistic pesticide mixture composed of 23 chemicals with 8 different MoAs originating from the test organism, *Scenedesmus vacuolatus*. Since the PCA- and RF-based clustering methods provided different results for the best number of clusters for Mixture 3 (Dataset 2 in Table 4), the QSAR-TSP model was applied to estimate the toxicity of Mixture 3 on the basis of both two and three clusters, as derived by the PCA- and RF-based methods, respectively. The best prediction performance was achieved by the CA model (R^2_{test} = 0.985, RSS = 2.42E+02). The QSAR-TSP model with two clusters (R^2_{test} = 0.973, RSS = 4.46E+02), QSAR-TSP with three clusters (R^2_{test} = 0.974, RSS = 4.32E+02), and conventional TSP (R^2_{test} = 0.979, RSS = 3.45E+02) models gave excellent predictions on the toxicity of Mixture 3 as well. All the QSAR-TSP models with two and three clusters showed very similar prediction results for Mixture 3. The TSP model was based on correct MoA information originating from the target organism (*Scenedesmus vacuolatus*) for Dataset 2; this was most likely responsible for the model's much better prediction result than for Dataset 1. Along the lines of this result for Mixture 3, some previous studies had argued that the TSP model, based on reliable MoAs, might have better predictions for estimating mixtures of pesticides, nitrobenzenes, industrial organic compounds, or wastewater treatment plant effluents [11,14,15,54]. For Mixture 3, the IA model achieved a good prediction for mixture toxicity (R^2_{test} = 0.874, RSS = 2.10E+03), quite dissimilar to its poor performance for Mixtures 1 and 2. The IA and conventional TSP models showed a tendency of the deviations between the predicted and observed data on Mixture 3 increasing gradually concomitantly with the development of effective concentrations in the effect range of 30% or more. The quantitative difference between the CA and IA predictions for Mixture 3 was relatively smaller than Mixtures 1 and 2. Junghans *et al.* [23] theoretically argued that the deviation of EC_{50} values between the CA and IA predictions could not exceed a factor of 2.5 in the test system based on specific scenarios concerning pesticide mixtures (*e.g.*, Mixture 3) they used (for the information of the scenarios, see Junghans *et al.* [23]). This was due to the fact that the mixture ratio (*i.e.*, the concentration ratio) could influence the deviation between the CA and IA calculations [23,55]. The possible deviation between the two models could be maximised in proportion to the number of mixture components at the specific situation in which all components were strictly dissimilarly- and independently-acting chemicals [23,55]. In the case of Mixture 3 used as a realistic pesticide, however, it was a non-equitoxic mixture, and widely composed of both similarly- and dissimilarly-acting chemicals (refer to Table 3). Table 5 summarizes the RSSs and R^2_{test}s from the QSAR-TSP, TSP, CA, and IA models for the three mixtures in the validations of Datasets 1 and 2.

3.4 Evaluation of the QSAR-TSP model

This section addresses the advantages and disadvantages of the QSAR-TSP model by comparing the PLS-IAM with the other models used in this study from three perspectives: model performance, data availability, and application coverage. First, from the perspective of model performance on predicting the mixture toxicity of the three mixtures in this study, it was evaluated that the QSAR-TSP model, overall, showed excellent prediction power for all datasets (Table 5). The CA model also presented high prediction performance for Mixtures 2 and 3, but these mixture types, which included different MoAs, were essentially contrary to the model assumption. In the case of the conventional TSP model, it was shown that incorrect information on MoAs, which ideally should originate from reference data of the target organism for each mixture component, did not perform well for estimating the toxicity of the mixture in this study. When it comes to the performance of the clustering algorithms in the QSAR-TSP model applied in this study, the *k*-means-based methods showed higher average silhouettes than the PAM-based methods did. Among the methods, the *k*-means via PCA method presented not only the quickest computation, but also the largest overall average silhouette width, the size of which indicates how well the number of clusters was selected (Table 4). However, the PCA-based methods have a common critical disadvantage: they hardly describe which real molecular descriptors are important for clustering results due to the distortion of original data arising from their transformation into new features during the data compression process. Considering the capacity for a model's clustering interpretability, the *k*-means via RF method could also be preferred because RF analysis, advantageous in its handling of a large number of variables simultaneously, provides information on the importance of descriptors [56]. However, the *k*-means via the RF method handling a large set of data has a disadvantage in that it requires much more calculation time for the clustering mixture components than the k-means via PCA method does. Therefore, if one only considers the results from the clustering methods used in this study, either the *k*-means via PCA or the *k*-means via RF method might be selected as the optimal technique for the QSAR-TSP model—it depends on the needs of the risk assessors using it.

Second, from the perspective of data availability, the QSAR-TSP model has a notable characteristic advantage in that it does not require MoA information tailored to the target organism unlike the conventional TSP model. Borgert *et al.* [4] and Fent [57] highlighted that even predictions on MoAs may not be practical for most compounds due to uncertainties in MoA values. Our study highlights the QSAR-TSP model's high potential to minimize the required information and resources for predicting the toxicity of complex mixtures because it only needs one set of data on DRCs of single substances on a commonly employed test organism in order to overcome the critical limitation of the conventional TSP model.

Finally, from the perspective of model application coverage, the evidence produced by this study suggests the QSAR-TSP model, an IAM that assumes similarly- and dissimilarly-acting chemicals are involved in a mixture simultaneously, can be applied to mixtures containing both types of substances. Although more validations of the QSAR-TSP model still need to be conducted to evaluate its practical availability in mixture risk assessments, its application coverage seems more extended than those of the conventional CA and IA models theoretically. However, all current prediction models, including QSAR-TSP, essentially ignore interactions (*e.g.*, synergism, antagonism, and potentiation) that can be caused by combined effects among two or more mixture components, and as such are limited to non-interacting mixtures only.

4. Conclusions and Outlook

For the three pesticide mixtures used as model validation data in this study, the QSAR-TSP model based on the structural information of each compound, which functioned as an IAM combining the CA and IA concepts, successfully estimated mixture toxicity in the absence of knowledge of the MoAs of mixture components. Therefore, the QSAR-TSP model's success reflects its potential to overcome two critical limitations: the requirement for complete knowledge of the MoAs for all chemicals in the mixture set by the conventional TSP model, and the theoretical limits on either similarly- or dissimilarly-acting chemicals put in place by the CA and IA models. In addition, the relative important descriptors in calculations of structural information for clustering chemicals in the three target mixtures were found best by the RF analysis in this study. Further studies on the validation of the QSAR-TSP model should to be conducted with toxicity data based on different types of mixtures and test organisms

Acknowledgements

The authors are grateful to PD Dr. Rolf Altenburger for his constructive comments. This study is funded by the Korean Ministry of Knowledge Economy and the Korea Institute of Science and Technology.

References

[1] S. Loewe and H. Muischnek, *Über Kombinationswirkungen I. Mitteilung: Hilfsmittel der Fragestellung*, Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol. 114 (1926), pp. 313-326.

[2] C.I. Bliss, *The toxicity of poisons applied jointly*, Ann. Appl. Biol. 26 (1939), pp. 586-615.

[3] M. Faust, R. Altenburger, T. Backhaus, H. Blanck, W. Boedeker, P. Gramatica, V. Hamer, M. Scholze, M. Vighi, and L.H. Grimme, *Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action*, Aquat. Toxicol. 63 (2003), pp. 43-63.

[4] C.J. Borgert, T.F. Quill, L.S. McCarty, and A.M. Mason, *Can mode of action predict mixture toxicity for risk assessment?*, Toxicol. Appl. Pharmacol. 201 (2004), pp. 85-96.

[5] J.C. Lambert and J.C. Lipscomb, *Mode of action as a determining factor in additivity models for chemical mixture risk assessment*, Regul. Toxicol. Pharmacol. 49 (2007), pp. 183-194.

[6] L.-T. Qin, S.-S. Liu, J. Zhang, and Q.-F Xiao, *A novel model integrated concentration addition with independent action for the prediction of toxicity of multi-component mixture*, Toxicology 280 (2011), pp. 164-172.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer, New York, NY, 2001.

[8] J. Adler, *R in a nutshell*, O'Reilly Media, Inc., Sebastopol, CA, 2009.

[9] J. Kim, S. Kim, and G.E. Schaumann, *Development of a partial least squares-based integrated addition model for predicting mixture toxicity*, Hum. Ecol. Risk Assess. (2012), DOI:10.1080/10807039.2012.754312. Available at http://dx.doi.org/10.1080/10807039.2012.754312.

[10] R. Altenburger, *Predicting toxic effects of contaminants in ecosystems using single species investigations*, Habil. diss., University of Bremen, 2002.

[11] M. Junghans, *Studies on combination effects of environmentally relevant toxicants*, Ph.D. diss., University of Bremen, 2004.

[12] R. Altenburger, H. Walter, and M. Grote, *What contributes to the combined effect of a complex mixture?*, Environ. Sci. Technol. 38 (2004), pp. 6353-6362.

[13] R. Altenburger, H. Schmitt, and G. Schüürmann, *Algal toxicity of nitrobenzenes: Combined effect analysis as a pharmacological probe for similar modes of interaction*, Environ. Toxicol. Chem. 24 (2005), pp. 324-333.

[14] J.S. Ra, B.C. Lee, N.I. Chang, and S.D. Kim, *Estimating the combined toxicity by two-step prediction model on the complicated chemical mixtures from wastewater treatment plant effluents*, Environ. Toxicol. Chem. 25 (2006), pp. 2107-2113.

[15] Z. Wang, J. Chen, L. Huang, Y. Wang, X. Cai, X. Qiao, and Y. Dong, *Integrated fuzzy concentration addition-independent action (IFCA-IA) model outperforms two-stage prediction (TSP) for predicting mixture toxicity*, Chemosphere 74 (2009), pp. 735-740.

[16] D.W. Roberts, *QSAR issues in aquatic toxicity of surfactants*, Sci. Total Environ. 109 (1991), pp. 557-568.

[17] S. Xu and N. Nirmalakhandan, *Use of QSAR models in predicting joint effects in multi-component mixtures of organic chemicals*, Water Res. 32 (1998), pp. 2391-2399.

[18] S.W. Morrall, M.J. Rosen, Y.P. Zhu, D.J. Versteeg, and S.D. Dyer, *Physicochemical descriptors for development of aquatic toxicity QSARs for surfactants*, F. Chen and G. Schüürman, eds., SETAC Press, Pensacola, 1999, pp 299-313.

[19] G.M. Boeije, M.L. Cano, S.J. Marshall, S.E. Belanger, R. Van Compernolle, P.B. Dorn, H. Gümbel, R. Toy, and T. Wind, *Ecotoxicity quantitative structure-activity relationships for alcohol ethoxylate mixtures based on substance-specific toxicity predictions*, Ecotoxicol. Environ. Saf. 64 (2006), pp. 75-84.

[20] E.N. Muratov, E.V. Varlamova, A.G. Artemenko, P.G. Polishchuk, and V.E. Kuz'min, *Existing and Developing Approaches for Qsar Analysis of Mixtures*, Mol. Inf. 31 (2012), pp. 202-21.

[21] L. Breiman, *Random Forests*, Mach. Learn. 45 (2001), pp. 5-32.

[22] T. Shi, S. Horvath, *Unsupervised learning with random forest predictors*, J. Comput. Graph. Stat. 15 (2006), pp. 118-138.

[23] M. Junghans, T. Backhaus, M. Faust, M. Scholze, and L.H. Grimme, *Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures*, Aquat. Toxicol. 76 (2006), pp. 93-110.

[24] H. Könemann, *Fish toxicity tests with mixtures of more than two chemicals: A proposal for a quantitative approach and experimental results*, Toxicology 19 (1981), pp. 229-238.

[25] R. Altenburger, T. Backhaus, W. Boedeker, M. Faust, M. Scholze, and L.H. Grimme, *Predictability of the toxicity of multiple chemical mixtures to Vibrio fischeri: Mixtures composed of similarly acting chemicals*, Environ. Toxicol. Chem. 19 (2000), pp. 2341-2347.

[26] T. Backhaus, R. Altenburger, W. Boedeker, M. Faust, M. Scholze, and L.H. Grimme, *Predictability of the toxicity of a multiple mixture of dissimilarly acting chemicals to Vibrio fischeri*, Environ. Toxicol. Chem. 19 (2000), pp. 2348-2356.

[27] K.-T. Fang, D.K.J. Lin, P. Winker, and Y. Zhang, *Uniform Design: Theory and Application*, Technometrics 42 (2000), pp. 237 - 248.

[28] J. Payne, N. Rajapakse, M. Wilkns, and A. Kortenkamp, *Prediction and assessment of the effects of mixtures of four xenoestrogens*, Environ. Health Perspect. 108 (2000), pp. 983-987.

[29] Y.-Z. Liang, K.-T. Fang, and Q.-S. Xu, *Uniform design and its applications in chemistry and chemical engineering*, Chemometr. Intell. Lab. 58 (2001), pp. 43-57.

[30] Y.H. Zhang, S.S Liu, H.L. Liu, and Z.Z. Liu, *Evaluation of the combined toxicity of 15 pesticides by uniform design*, Pest Manag. Sci. 66 (2010), pp. 879-887.

[31] M. Scholze, W. Boedeker, M. Faust, T. Backhaus, and R. Altenburger, L.H. Grimme, *A general best-fit method for concentration-response curves and the estimation of low-effect concentrations*, Environ. Toxicol. Chem. 20 (2001), pp. 448-457.

[32] J.J.P. Stewart, *Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements*, J. Mol. Model. 13 (2007), pp. 1173-1213.

[33] *Molecular Orbital Package (MOPAC) 2009*. Stewart Computational Chemistry, Colorado Springs, CO, USA, 2008; sofware available at http://openmopac.net/.

[34] I.T. Jolliffe, *Principal component analysis*, 2nd ed., Springer-Verlag, New York, NY, 2002

[35] C.R. Rao, *The use and interpretation of principal component analysis in applied research*, Sankhya Ser. A 26 (1964), pp. 329-358.

[36] H. Zou, T. Hastie, and R. Tibshirani, *Sparse Principal Component Analysis*, J. Comput. Graph. Stat. 15 (2006), pp. 265-286.

[37] T. Shi, S. Horvath, *Unsupervised learning with random forest predictors*, Tech. Rep., Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, 2005. Available at http://www.genetics.ucla.edu/labs/horvath/publications/RFclusteringShiHorvath.pdf.

[38] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and regression trees*. Chapman and Hall, New York, NY, 1984.

[39] L. Breiman and A. Cutler, *Manual - Setting Up, Using, And Understanding Random Forests V4.0*, Tech. Rep., Department of Statistics, University of California Berkeley, Berkely, CA, 2003. Available at http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf.

[40] S. Kuhn, B. Egert, S. Neumann, C. Steinbeck, *Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction*, BMC Bioinformatics 9 (2008), p. 400.

[41] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling*, J. Chem. Inf. Comp. Sci. 43 (2003), pp. 1947-1958.

[42] B. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. Hamprecht, *A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data*. BMC Bioinformatics 10 (2009), p. 213.

[43] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, and L. Kaufman, *Chemometrics: a textbook, in series: Data handling in science and technology*, Vol. 2, Elsevier Science Publishers B.V., Amsterdam, 1988.

[44] B.D. Gute, S.C. Basak, D. Mills, and D.M. Hawkins, *Tailored similarity spaces for the prediction of physicochemical properties*, Internet Electron. Mol. Des. 1 (2002), 374-387.

[45] J.B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the 5th of Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkely, CA, 1967.

[46] L. Kaufman and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons., Hoboken, NJ, 1990.

[47] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing., Vienna, 2010; software available at http://www.R-project.org/.

[48] P.J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, J. Comput. Appl. Math. 20 (1987), pp. 53-65.

[49] H. Akaike, *Information theory as an extension of the maximum likelihood principle*, 2nd International Symposium on Information Theory, Akademiai Kiado, Bupdapest, 1973.

[50] K.P. Burnham and D.R. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd ed., Springer, New York, NY, 2010.

[51] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*. *Volume 1: Alphabetical Listing; Volume 2: Appendices, References*, Wiley-VCH, Weinheim, 2009.

[52] *DRAGON software for molecular descriptor calculation 6.0*. Talete s.r.l., Milano, Italy, 2010; software available at http://www.talete.mi.it.

[53] M. Mittlböck, *Calculating adjusted R^2 measures for Poisson regression models*, Comput. Methods Programs Biomed. 68 (2002), pp. 205-214.

[54] R. Altenburger, H. Schmitt, G. Schüürmann, *Algal toxicity of nitrobenzenes: Combined effect analysis as a pharmacological probe for similar modes of interaction*. Environ. Toxicol. Chem. 24 (2005), 324-333.

[55] M. Faust, Kombinationseffekte von Schadstoffen auf aquatische Organismen: Prüfung der Vorhersagbarkeit am Beispiel einzelliger Grünalgen, GCA-Verlag, Herdecke, Germany, 1999.

[56] Y. Kim, H. Kim, *Application of random forests to association studies using mitochondrial single nucleotide,* Genomics Inform. 5 (2007), 168-173.

[57] K. Fent, *Ecotoxicological problems associated with contaminated sites*. Toxicol. Lett. 140-141 (2003), 353-365.

Table 1. Dataset 1: Physical properties, MoAs, and parameters for regression models of dose-response curves on *Vibrio fischeri* for ten pesticide chemicals in different MoA groups and their mixtures [9]

	CAS RN[1]	*MW*[2]	*Use*	*RM*[3]	*MoA*	*Model parameters*			
						α[4]	*β*[5]	*γ*[6]	*r^2*
Single component									
Alachlor	15972-60-8	269.77	Herbicide	S	VLCFA[7]	0.8694	82.6671	438.3162	0.985
Napropamide	15299-99-7	271.35	Herbicide	L2	VLCFA	1.4374	-1.3685	379.549	0.986
Cyanazine	21725-46-2	240.69	Herbicide	C	PSII[8]	0.9284	0.0021	0.9084	0.977
Isoproturon	34123-59-6	206.28	Herbicide	C	PSII	1.0014	0.0022	0.5547	0.991
Thiabendazole	148-79-8	201.25	Fungicide	L2	MCD[9]	1.0296	-1.622	427.0463	0.964
Thiophanate-methyl	23564-05-8	342.39	Fungicide	C	MCD	0.9377	0.0925	1.4674	0.995
Fenamidone	161326-34-7	311.40	Fungicide	H	Res[10]	2.11E+05	0.7027	5.89E+09	0.981
Furalaxyl	57646-30-7	301.34	Fungicide	S	NAS[11]	0.7639	71.5347	337.6356	0.979
MCPA[12]	94-74-6	200.62	Herbicide	C	AIAA[13]	3.4423	0.0004	1.4864	0.976
Chlorfenvinphos	470-90-6	359.57	Insecticide	L2	AChE[14]	2.0744	-0.6235	188.0293	0.971
Mixture									
EC_{50} ratio mixture[15]	-	-	-	C	-	1.0960	0.001	0.623	0.988
EC_{10} ratio mixture	-	-	-	L2	-	9.56E+05	-0.8743	6.75E+09	0.976

Notes: [1]Chemical Abstracts Services Registry Number; [2]Molecular weight (g/mol); [3]Regression model (refer to Table 2); [4]Height; [5]Slope; [6]Centre point; [7]Inhibition of very long chain fatty acid formation; [8]Inhibition of photosynthesis at photosystem II; [9]Mitosis and cell division; [10]Respiration; [11]Nucleic acids synthesis; [12]2-methyl-4-chlorophenoxyacetic acid; [13]Action-like indole acetic acid (synthetic auxins); [14]Acetylcholinesterase (AChE) inhibitors; [15]The ECx ratio mixture: an equitoxic mixture-based ratio at x% effective concentration of each component.

Table 2. Regression models used for describing the dose-response curve for chemical substances and mixtures in this study

Regression model	Function
Logit (L1)	$E(c) = \frac{1}{1 + exp(-\alpha - \beta log_{10}(c))}$
Probit (P)	$E(c) = \frac{1}{2\pi} \int_{-\infty}^{\alpha + \beta log_{10}(c)} exp(\frac{-u^2}{2}) du = \Phi(\alpha + \beta log_{10}(c))$
Weibull (W)	$E(c) = 1 - exp(-exp(\alpha + \beta log_{10}(c)))$
Generalized Logit (GL)	$E(c) = \frac{1}{[1 + exp(-\alpha - \beta log_{10}(c))]^{\gamma}}$
Box-Cox-Weibull (BCW)	$E(c) = 1 - exp(-exp(\alpha + \beta(\frac{c^{\gamma} - 1}{\gamma})))$
Sigmoid (S)	$E(c) = \frac{\alpha}{1 + exp(-\frac{c - \gamma}{\beta})}$
Logistic (L2)	$E(c) = \frac{\alpha}{1 + (\frac{c}{\gamma})^{\beta}}$
Hill (H)	$E(c) = \frac{\alpha c^{\beta}}{\gamma^{\beta} + c^{\beta}}$
Chapman (C)	$E(c) = \alpha(1 - exp(-\beta c))^{\gamma}$

Notes: E(c): the fractional effect elicited at concentration c; α, β, and γ: parameters of regression models (corresponding statistical estimates); Φ(y): the cumulative normal (Gaussian) distribution, meaning that the probability of a standard normal random variable is less than y.

Table 3. Dataset 2: Physical properties, MoAs, and parameters of regression models for DRCs on *Scenedesmus vacuolatus* strain 211-15 for the realistic pesticide mixture involving 23 components [23]

	CAS RN	*MW(g/mol)*	*Use*	*RM*[1]	*MoA*	*Regression Coefficients*		
						α	*β*	*γ*
Single component								
2,4-D	94-75-7	221.04	herbicide	GL	Narcotic	-37.540	11.106	0.1392
Aclonifen	74070-46-5	264.67	herbicide	BCW	Porphyrin	2.402	0.408	-0.3400
Alachlor	15972-60-8	269.77	herbicide	W	VLCFA[3]	4.060	5.193	
Atrazin	1912-24-9	215.69	herbicide	GL	PSII[4]	6.765	17.391	0.1118
Bromoxynil	1689-84-5	276.91	herbicide	L1	PSII	-19.600	9.267	
Carbofuran	1563-66-2	221.26	insecticide	W	Unknown	-4.564	1.978	
Chloridazon	1698-60-8	221.65	herbicide	W	PSII	-2.375	2.777	
Cycloxydim	101205-02-1	325.47	herbicide	W	Narcotic	-5.232	1.990	
Ethofumesate	26225-79-6	286.35	herbicide	W	VLCFA	-2.126	1.108	
Ioxynil	1689-83-4	370.92	herbicide	W	PSII	-3.785	2.229	
Isofenphos	25311-71-1	345.39	insecticide	GL	Narcotic	-3.373	2.186	0.4219
Isoproturon	34123-59-6	206.39	herbicide	BCW	PSII	1.246	1.073	-0.0235
Isoxaflutol	141112-29-0	359.32	herbicide	W	Plastoquinone	-5.313	2.529	
Lenacil	2164-08-1	234.3	herbicide	GL	PSII	14.991	14.338	0.1845
Linuron	330-55-2	249.1	herbicide	W	PSII	1.769	2.020	
MCPA[2]	94-74-6	200.62	herbicide	P	Narcotic	-4.501	1.551	
Metamitron	41394-05-2	202.22	herbicide	W	PSII	-0.995	1.912	
Metolachlor	51218-45-2	283.8	herbicide	BCW	VLCFA	0.239	3.156	0.4930
Pendimethalin	40487-42-1	281.31	herbicide	W	Microtubule	5.752	2.957	
Terbuthylazine	5915-41-3	229.71	herbicide	W	PSII	4.165	3.908	
Thifensulfuron-methyl	79277-27-3	387.38	herbicide	L1	ALS[5]	-2.093	1.837	
Triasulfuron	82097-50-5	401.82	herbicide	W	ALS	0.093	1.684	
Tribenuron-methyl	101200-48-0	395.39	herbicide	W	ALS	0.670	1.735	
Mixture								
Mixture of 23 substances	-	-	-	BCW	-	1.090	1.896	0.3659

Notes: [1]Regression model (refer to Table 2); [2]2-methyl-4-chlorophenoxyacetic acid; [3]Inhibition of very long chain fatty acid formation; [4]Inhibition of the D1 protein in the photosystem II; [5]Inhibition of acetolactate synthase.

Table 4. Determination of optimal cluster size using average silhouette width for datasets from different clustering algorithms based on PCA or RF

No. of clusters	*Average Silhouette width*			
	k-means via PCA[1]	*PAM*[2] *via PCA*	*k-means via RF*[3]	*PAM via RF*
Dataset 1				
2	0.40	0.28	0.15	0.09
3	0.25	0.22	0.09	0.06
4	0.22	0.16	0.09	0.04
5	0.23	0.16	0.07	0.03
6	0.14	0.13	0.06	0.03
7	0.10	0.1	0.04	0.02
8	0.11	0.06	0.03	0.02
9	0.05	0.03	0.01	0.01
Dataset 2				
2	0.38	0.36	0.24	0.20
3	0.44	0.43	0.17	0.12
4	0.28	0.28	0.14	0.09
5	0.26	0.30	0.16	0.11
6	0.30	0.29	0.15	0.11
7	0.26	0.28	0.17	0.11
8	0.25	0.26	0.15	0.12
9	0.28	0.26	0.15	0.11

Notes: [1]Principal component analysis; [2]Partitioning around medoids; [3]Random forest.

Table 5. Summary of the AIC and R^2_{test} of the QSAR-TSP, TSP, CA, and IA models calculated from validations of Datasets 1 and 2

Model	*Mixture components in datasets*	*RSS*[1]	R^2_{test}[2]
Dataset 1			
QSAR-TSP	Mixture 1: the EC_{50} ratio mixture	3.70E+02	0.925
TSP	(10 pesticides in different MoA groups)	5.89E+03	0.158
CA		1.76E+03	0.749
IA		9.58E+03	n.v.
QSAR-TSP	Mixture 2: the EC_{10} ratio mixture	5.39E+02	0.923
TSP	(10 pesticides in different MoA groups)	4.64E+03	0.337
CA		8.68E+02	0.876
IA		6.76E+03	0.034
Dataset 2			
QSAR-TSP (with 2 clusters)	Mixture 3: 23 pesticides in different MoA	4.32E+02	0.973
QSAR-TSP (with 3 clusters)	groups	4.46E+02	0.974
TSP		3.45E+02	0.979
CA		2.42E+02	0.985
IA		2.10E+03	0.874

Notes: [1]The residual sum of squares; [2]The coefficient of determination for the modelled data.

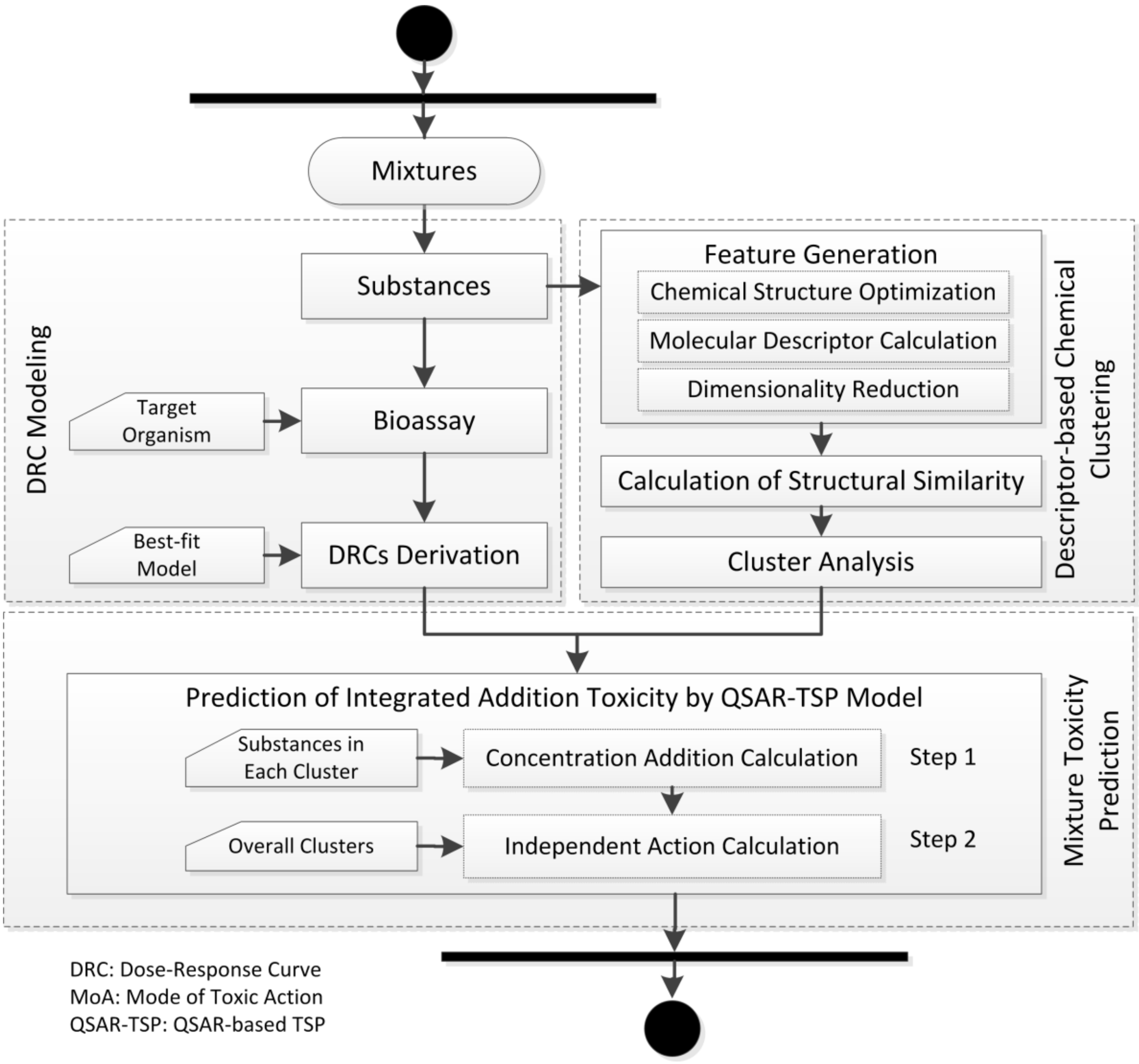
Mixtures
DRC Modeling
Substances
Target Organism
Bioassay
Best-fit Model
DRCs Derivation
Descriptor-based Chemical Clustering
Feature Generation
Chemical Structure Optimization
Molecular Descriptor Calculation
Dimensionality Reduction
Calculation of Structural Similarity
Cluster Analysis
Mixture Toxicity Prediction
Prediction of Integrated Addition Toxicity by QSAR-TSP Model
Substances in Each Cluster
Concentration Addition Calculation
Step 1
Overall Clusters
Independent Action Calculation
Step 2
DRC: Dose-Response Curve
MoA: Mode of Toxic Action
QSAR-TSP: QSAR-based TSP

Figure 1. The scheme of the QSAR-TSP model.

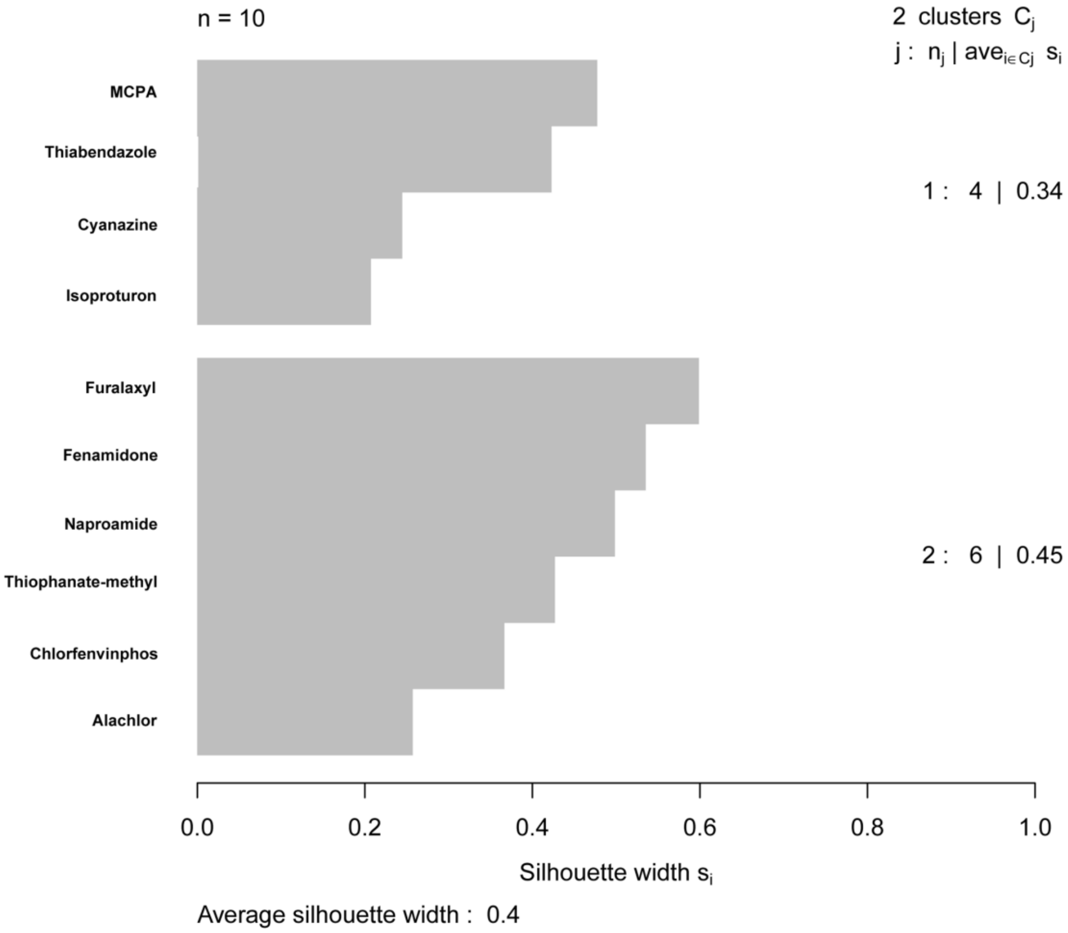

Figure 2(a). Clustered results and maximum average silhouette widths of mixture components in Dataset 1 through the k-means and PAM clustering methods based on the PCA and RF techniques (*i.e.*, the *k*-means via PCA, PAM via PCA, the *k*-means via RF, and PAM via PCA methods).

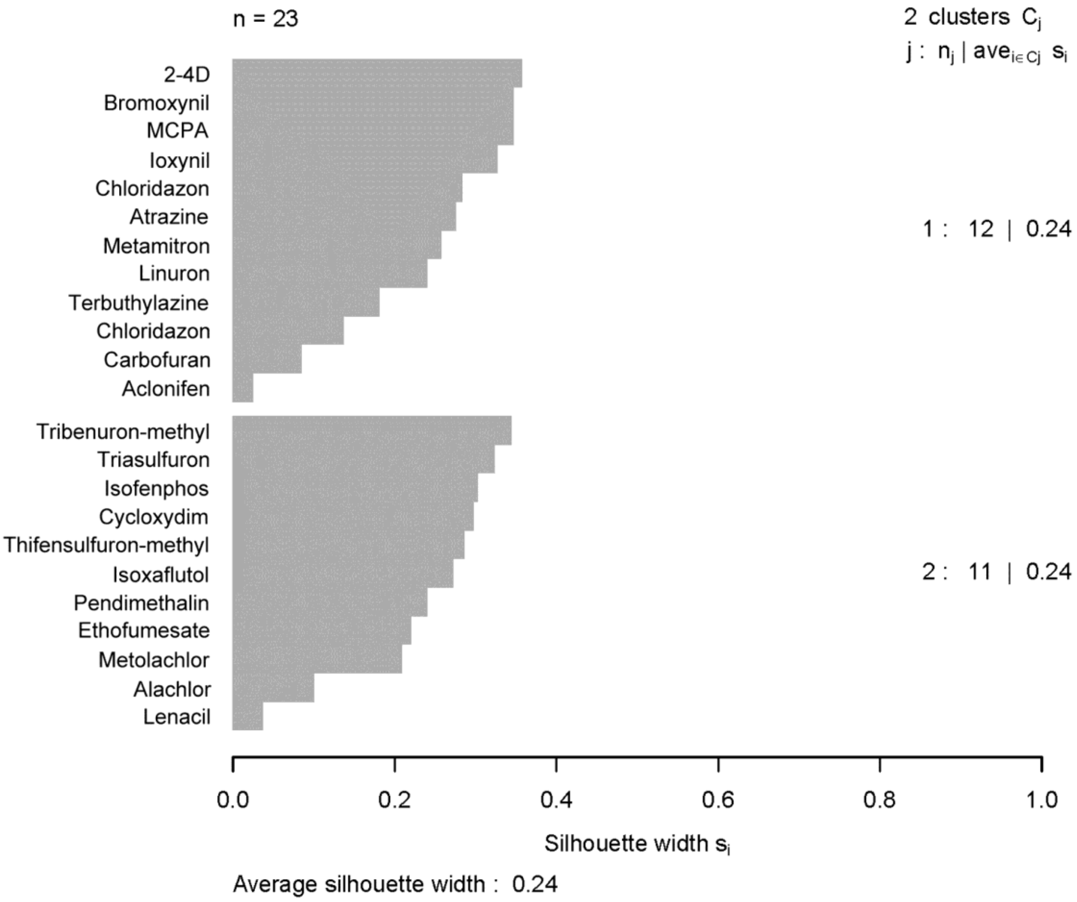

Figure 2(b). Clustered results and maximum average silhouette widths of mixture components in Dataset 2 through the k-means and PAM clustering methods based on the PCA technique (*i.e.*, the k-means via PCA and PAM via PCA methods) using principle components derived from the original molecular descriptors.

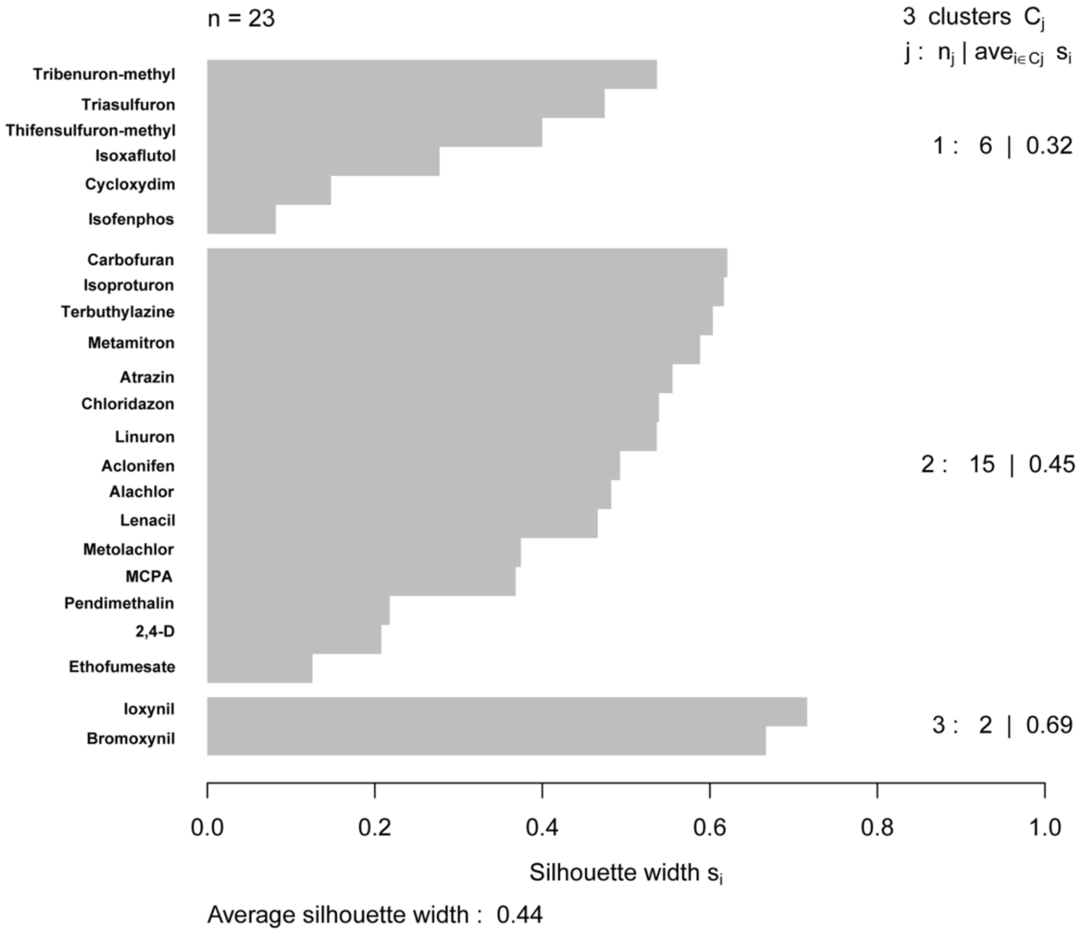

Figure 2(c). Clustered results and maximum average silhouette widths of mixture components in Dataset 2 through the k-means and PAM clustering methods based on the RF technique (*i.e.*, the *k*-means via RF and PAM via RF methods) using all the original molecular descriptors.

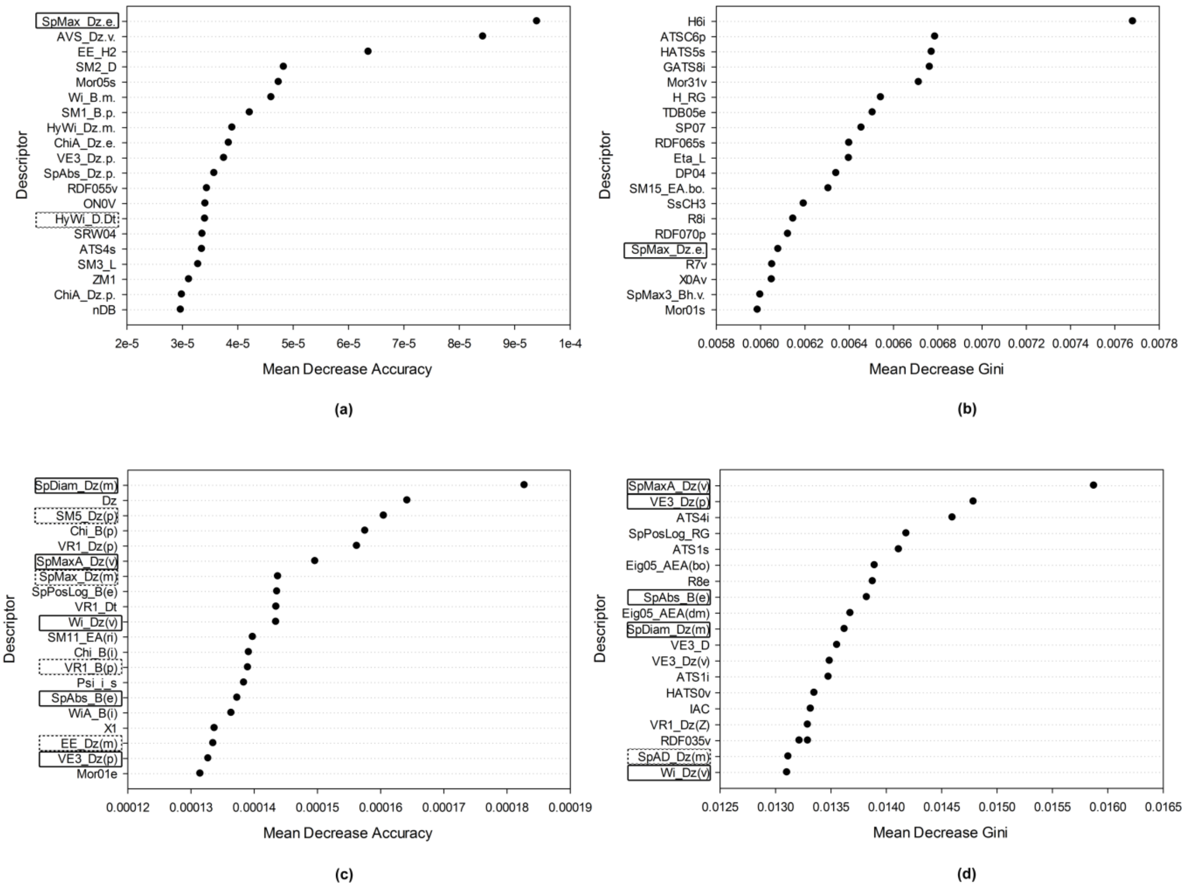

Figure 3. Molecular descriptor importance plots with top 20 descriptors for clustering chemicals in Dataset 1[(a) and (b)] and Dataset 2[(c) and (d)] through RF analysis with the MDA and MDG Index.

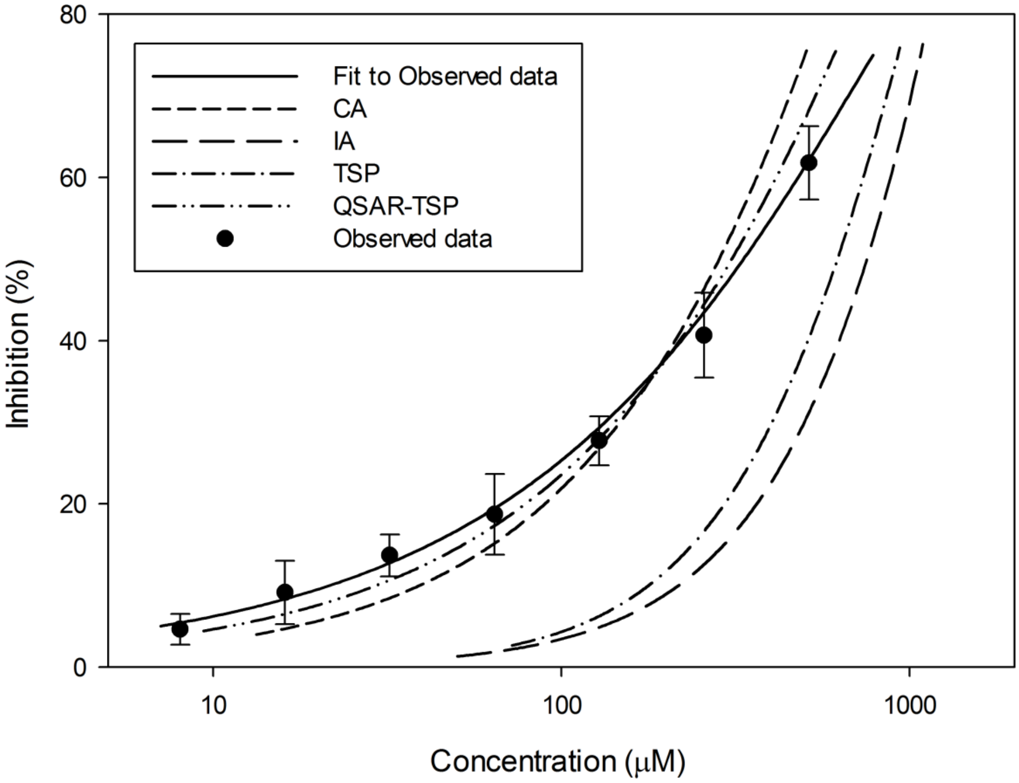

Figure 4(a). Comparison of the CA, IA, TSP, and QSAR-TSP predictions against observed toxicity for Mixture 1 (the EC_{50} ratio mixture), an equitoxic mixture-based ratio at 50% effective concentrations of each component in Dataset 1 (the data points are geometric means ± SD of experimentally-observed data [9].).

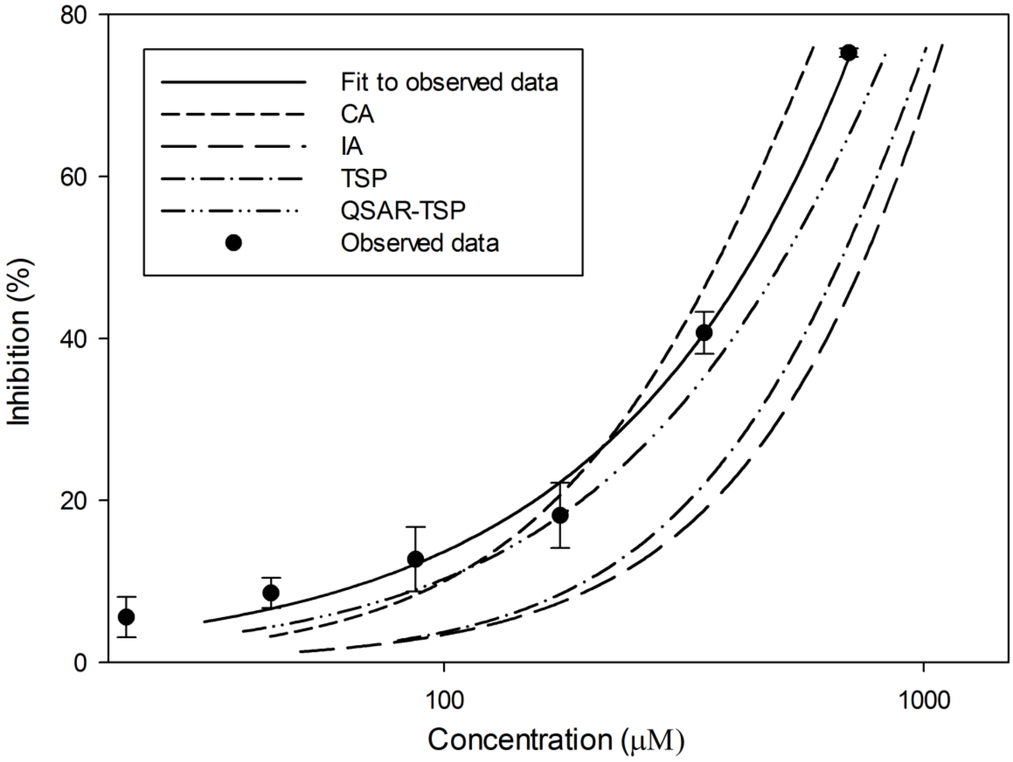

Figure 4(b). Comparison of the CA, IA, TSP, and QSAR-TSP predictions against observed toxicity for Mixture 2 (the EC_{10} ratio mixture), an equitoxic mixture-based ratio at 10% effective concentrations of each component in Dataset 1 (the data points are geometric means ± SD of experimentally-observed data [9].).

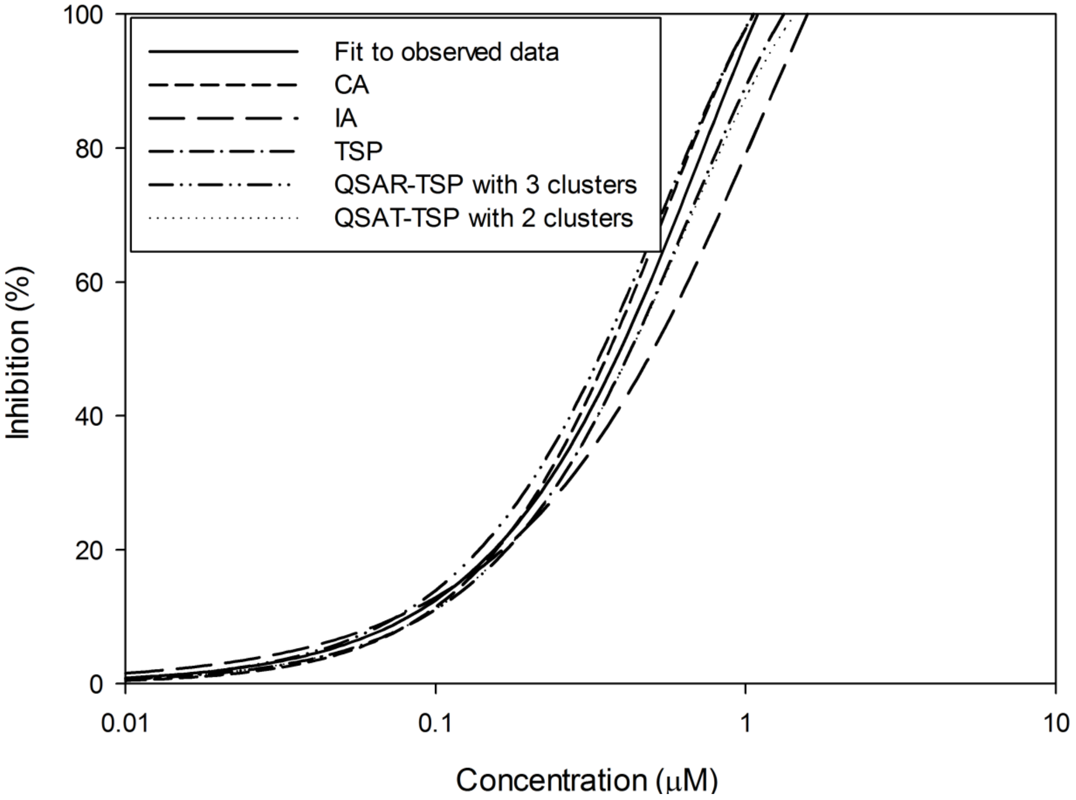

Figure 4(c). Comparison of the CA, IA, TSP, and QSAR-TSP predictions against observed toxicity for Mixture 3, a realistic pesticide mixture in Dataset 2 (the fitted regression line for observed data is plotted by the regression function in Table 3 [22]).

Supplementary material

Development of QSAR-based two-stage prediction model for estimating mixture toxicity

Jongwoon Kim, Sanghun Kim, and Gabriele E. Schaumann

Table S1. Principal component scores for the 10 substances in Dataset 1

Compounds	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Thiophanate-methyl	-3.28E+01	2.20E+01	3.45E+01	-1.58E+01	1.53E+01	-3.90E+00	4.43E+00	-5.66E+00	-1.98E+00
Alachlor	-5.46E+00	4.27E+00	-2.73E+01	1.47E+01	6.55E+00	7.05E-01	1.39E+01	-1.80E+01	-4.57E+00
Chlorfenvinphos	-1.57E+01	1.75E+01	-2.36E+01	-3.01E+01	-1.84E+01	-1.01E+01	-6.14E+00	1.15E+00	1.86E+00
Cyanazine	2.18E+01	1.20E+01	1.44E+01	5.17E+00	-2.35E+01	2.07E+01	1.01E+01	3.42E+00	-4.07E+00
Fenamidone	-3.55E+01	-2.73E+01	9.29E-01	-3.19E+00	5.23E-01	1.72E+01	-1.90E+01	-7.09E+00	2.14E-01
Furalaxyl	-3.45E+01	-1.12E+01	-6.20E+00	6.04E+00	5.43E+00	1.68E+00	1.38E+01	1.28E+01	1.62E+01
Isoproturon	2.12E+01	1.62E+01	9.23E+00	2.99E+01	-5.27E+00	-1.12E+01	-1.49E+01	-3.75E+00	1.01E+01
MCPA	5.49E+01	1.11E+01	-1.35E+01	-9.20E+00	2.32E+01	9.81E+00	-6.58E+00	8.66E+00	-1.35E+00
Naproamide	-2.32E+01	-8.91E+00	-1.26E+00	1.63E+01	2.83E-02	-1.19E+01	-2.27E+00	1.33E+01	-1.75E+01
Thiabendazole	4.92E+01	-3.57E+01	1.28E+01	-1.38E+01	-3.95E+00	-1.30E+01	6.74E+00	-4.86E+00	1.03E+00

Notes) PC: Principal Component.

Table S2. Principal component scores for the 23 substances in Dataset 2

Compounds	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13	PC 14	PC 15	PC 16	PC 17	PC 18	PC 19	PC 20
Tribenuron-methyl	-6.05E+01	-2.43E+01	6.21E-02	1.14E+01	-7.32E-01	-7.92E+00	3.77E+00	3.53E-01	4.80E+00	1.37E+00	-6.00E+00	1.02E+00	-5.54E-01	2.99E+00	-8.01E+00	4.02E+00	-1.17E+01	1.25E+01	-1.80E+00	-5.61E-01
2,4-D	4.33E+01	-5.38E+00	1.89E+01	1.05E+01	-1.20E+01	2.24E+01	-2.56E+00	-7.29E+00	-3.14E+00	7.12E+00	3.11E-01	-3.99E+00	-1.40E+00	4.93E+00	-1.52E+00	-7.45E-01	8.19E-01	3.37E+00	1.13E-01	2.41E+00
Aclonifen	2.40E+00	-7.13E+00	1.77E+01	-3.07E+00	9.24E+00	-2.70E+00	1.06E+01	9.39E+00	-9.19E+00	-1.44E-01	-1.54E+00	-5.24E+00	-9.12E+00	3.78E-02	5.23E+00	-1.83E+01	-5.31E+00	-6.46E-01	-2.71E+00	-2.82E-01
Alachlor	-4.59E+00	1.90E+01	-5.50E+00	-5.25E+00	-2.64E+01	-4.66E+00	8.33E+00	6.81E-01	8.37E+00	-6.28E+00	5.42E+00	-1.07E+00	-3.40E+00	-8.48E+00	-1.80E+00	-2.76E+00	2.12E+00	-4.11E-01	-2.38E+00	2.55E+00
Atrazin	2.60E+01	2.07E+01	3.03E+00	1.70E+01	7.04E+00	-4.74E+00	-1.81E+01	2.76E+00	7.06E+00	4.18E+00	9.67E+00	-9.62E-01	-1.64E+00	-1.73E+00	2.04E+00	-4.87E-01	-5.15E-01	9.79E-01	-2.16E+00	-1.18E+01
Bromoxynil	6.72E+01	-2.91E+01	-2.13E+01	-2.25E+00	2.45E+00	-4.02E+00	3.49E-01	3.49E-01	3.54E+00	1.56E+00	-1.26E+01	-1.71E+01	8.09E+00	-1.31E+01	4.03E+00	1.28E+00	4.67E-01	1.28E+00	3.12E-01	-5.57E-01
Carbofuran	8.34E+00	1.01E+01	3.87E+00	-7.29E+00	1.34E+00	-9.31E+00	-1.38E+00	-1.81E+01	1.06E-01	-3.88E+00	1.41E+00	-2.52E+00	3.22E+00	1.73E+00	-2.75E+00	1.17E+00	-1.39E+01	-1.16E+01	5.04E+00	7.00E-04
Chloridazon	2.34E+01	-8.93E-01	1.85E+01	-5.49E+00	1.13E+01	-4.01E+00	1.32E+01	5.28E+00	-5.07E+00	3.86E+00	7.85E+00	1.69E+00	-7.33E+00	-5.72E+00	2.25E+00	7.66E+00	2.28E+00	4.54E+00	1.29E+01	-1.25E-01
Cycloxydim	-4.33E+01	2.39E+01	-2.97E+01	6.92E-01	1.85E+01	3.12E+01	8.70E+00	6.71E-01	6.39E+00	6.42E+00	-4.17E+00	2.99E+00	-4.64E+00	-4.05E+00	3.77E+00	6.88E-01	-3.23E+00	-2.27E+00	5.55E-01	9.81E-02
Ethofumesate	-2.03E+01	3.38E+00	-1.16E+00	-8.85E+00	6.86E+00	-2.40E+00	-4.45E+00	-1.86E+01	-8.05E+00	-1.75E+01	-9.86E+00	-7.44E-01	-8.23E+00	6.58E+00	1.08E+01	3.13E+00	5.79E+00	4.57E+00	-1.75E+00	-2.10E+00
Ioxynil	6.77E+01	-3.70E+01	-3.53E+01	-4.49E+00	-1.97E-01	-1.79E+00	9.32E-01	2.26E+00	-9.91E-01	-3.17E+00	1.09E+01	1.29E+01	-6.66E+00	1.04E+01	-2.22E+00	-1.34E+00	-4.95E-01	-9.24E-01	-6.66E-01	3.15E-01
Isofenphos	-3.52E+01	1.32E+01	-2.07E+01	6.60E+00	-1.19E+01	-6.65E+00	-8.17E+00	2.28E+00	-3.14E+01	1.11E+01	-8.03E-02	-1.24E+00	1.95E+00	-2.54E+00	-2.77E+00	1.33E+00	1.20E+00	-1.08E+00	2.26E-01	1.15E-01
Isoproturon	1.20E+01	1.79E+01	-5.42E-01	2.81E+00	8.46E+00	-2.15E+00	-3.08E+00	8.69E+00	7.98E-01	-1.19E+01	-1.61E+01	9.48E+00	6.53E+00	-2.60E-01	-1.43E+01	-6.35E+00	5.07E+00	-2.48E-01	5.37E+00	-1.22E+00
Isoxaflutol	-4.53E+01	-2.85E+01	1.26E+01	-3.59E+01	-8.92E-01	1.32E+01	-2.47E+01	5.96E+00	8.45E-01	-1.73E+00	5.17E+00	1.07E+00	9.63E-01	-7.18E+00	-3.08E+00	2.41E-01	1.44E-01	-8.32E-01	-3.25E-01	-2.57E-01
Lenacil	-2.54E+00	1.88E+01	-2.72E+00	-2.02E+01	1.38E+01	-6.05E+00	9.63E+00	-1.20E+01	2.59E+00	9.30E+00	1.10E+01	-1.43E+00	1.66E+01	5.66E+00	-2.44E+00	-5.23E+00	4.52E+00	6.64E+00	-2.40E+00	9.56E-01
Linuron	2.23E+01	2.71E-01	1.47E+01	6.88E+00	-9.84E+00	6.40E+00	3.95E+00	5.62E+00	-4.41E+00	-4.98E+00	-1.98E+00	1.69E+01	1.65E+01	-3.82E+00	1.37E+01	2.69E+00	-4.58E+00	3.12E-01	-2.60E+00	3.11E-01
MCPA	3.82E+01	1.70E+00	1.41E+01	7.64E+00	-1.16E+01	1.69E+01	-1.18E+00	-8.34E+00	1.70E-01	2.38E+00	-4.06E+00	-3.58E+00	-2.27E+00	5.90E+00	-8.70E+00	5.86E-01	2.07E+00	-8.90E-01	-6.91E-01	-2.20E-01
Metamitron	2.03E+01	4.85E+00	1.55E+01	-7.02E+00	1.52E+01	-8.21E+00	1.13E+01	5.88E+00	-2.23E+00	3.42E+00	-3.68E+00	1.84E+00	-5.57E+00	-2.65E+00	-6.70E+00	1.11E+01	2.19E+00	-6.04E+00	-1.13E+01	7.15E-01
Metolachlor	-1.27E+01	2.00E+01	-7.31E+00	-4.31E+00	-2.40E+01	-3.08E+00	8.53E+00	2.90E-01	5.49E+00	-7.05E+00	5.35E+00	-1.16E+00	-6.13E+00	-6.72E+00	-6.62E-01	4.18E-01	2.13E-01	2.64E+00	9.73E-03	-1.54E+00
Pendimethalin	-1.83E+01	9.00E+00	-8.52E-01	-1.25E+01	-1.35E+01	-6.54E+00	-6.04E-01	1.70E+01	8.92E+00	7.98E+00	-8.70E+00	-6.54E+00	1.58E+00	1.97E+01	7.28E+00	3.34E+00	1.17E+00	-3.34E+00	2.75E+00	-5.78E-01
Terbuthylazine	1.55E+01	2.24E+01	-8.78E-01	1.50E+01	1.29E+01	-9.20E+00	-2.28E+01	4.15E+00	6.06E+00	-2.27E+00	4.00E+00	-1.77E+00	-3.58E+00	-4.88E-01	3.58E+00	2.25E-01	-1.04E+00	2.46E+00	-4.71E-01	1.09E+01
Thifensulfuron-methyl	-4.71E+01	-2.89E+01	5.78E+00	1.41E+01	-3.25E+00	-1.23E+01	-5.12E-01	-1.27E+01	1.04E+01	1.48E+01	-5.51E+00	1.03E+01	-3.09E+00	-5.13E+00	3.50E+00	-4.34E+00	7.75E+00	-5.42E+00	1.55E+00	6.53E-01
Triasulfuron	-5.71E+01	-2.42E+01	1.32E+00	2.41E+01	7.22E+00	5.69E+00	8.21E+00	5.40E+00	-1.04E+00	-1.45E+01	1.32E+01	-1.09E+01	8.25E+00	3.85E+00	-1.15E+00	1.75E+00	4.98E+00	-5.53E+00	4.53E-01	2.01E-01

Notes) PC: Principal Component.

Table S3. Euclidean distances between all pairs of substances in Dataset 1

Compounds	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	7.84E+01	7.24E+01	7.84E+01	7.03E+01	6.42E+01	8.17E+01	1.04E+02	6.56E+01	1.05E+02
S2	7.84E+01	0	6.20E+01	6.64E+01	6.75E+01	5.46E+01	6.28E+01	7.72E+01	5.27E+01	8.70E+01
S3	7.24E+01	6.20E+01	0	7.35E+01	7.11E+01	6.50E+01	7.99E+01	8.81E+01	6.54E+01	9.52E+01
S4	7.84E+01	6.64E+01	7.35E+01	0	8.15E+01	7.65E+01	5.37E+01	6.85E+01	6.98E+01	7.08E+01
S5	7.03E+01	6.75E+01	7.11E+01	8.15E+01	0	4.89E+01	8.50E+01	1.04E+02	5.22E+01	9.55E+01
S6	6.42E+01	5.46E+01	6.50E+01	7.65E+01	4.89E+01	0	7.80E+01	9.95E+01	4.32E+01	9.62E+01
S7	8.17E+01	6.28E+01	7.99E+01	5.37E+01	8.50E+01	7.80E+01	0	6.94E+01	6.43E+01	7.72E+01
S8	1.04E+02	7.72E+01	8.81E+01	6.85E+01	1.04E+02	9.95E+01	6.94E+01	0	9.28E+01	6.75E+01
S9	6.56E+01	5.27E+01	6.54E+01	6.98E+01	5.22E+01	4.32E+01	6.43E+01	9.28E+01	0	8.85E+01
S10	1.05E+02	8.70E+01	9.52E+01	7.08E+01	9.55E+01	9.62E+01	7.72E+01	6.75E+01	8.85E+01	0

Notes) S1: Thiophanate-methyl; S2: Alachlor; S3: Chlorfenvinphos; S4: Cyanazine; S5: Fenamidone; S6: Furalaxyl; S7: Isoproturon; S8: MCPA; S9: Naproamide; S10: Thiabendazole.

Table S4. Euclidean distances between all pairs of substances in Dataset 2

Compounds	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23
S1	0	1.15E+02	7.90E+01	8.24E+01	1.04E+02	1.35E+02	8.63E+01	9.74E+01	7.80E+01	6.81E+01	1.38E+02	6.81E+01	9.17E+01	6.74E+01	8.77E+01	9.70E+01	1.09E+02	9.53E+01	7.60E+01	7.07E+01	9.75E+01	4.17E+01	4.48E+01
S2	1.15E+02	0	6.33E+01	7.36E+01	5.65E+01	6.90E+01	6.29E+01	5.32E+01	1.11E+02	8.38E+01	7.86E+01	9.99E+01	6.65E+01	1.09E+02	7.93E+01	4.75E+01	1.66E+01	5.94E+01	7.84E+01	8.36E+01	6.67E+01	1.05E+02	1.13E+02
S3	7.90E+01	6.33E+01	0.	5.92E+01	6.16E+01	8.82E+01	5.02E+01	4.16E+01	8.61E+01	5.58E+01	9.62E+01	7.43E+01	5.37E+01	7.90E+01	5.80E+01	5.28E+01	5.96E+01	4.32E+01	6.00E+01	5.78E+01	6.11E+01	7.45E+01	7.97E+01
S4	8.24E+01	7.36E+01	5.92E+01	0	6.11E+01	9.76E+01	4.77E+01	6.27E+01	7.64E+01	5.75E+01	1.04E+02	6.19E+01	5.35E+01	8.60E+01	5.47E+01	5.79E+01	6.37E+01	6.09E+01	1.21E+01	4.80E+01	6.01E+01	7.92E+01	8.64E+01
S5	1.04E+02	5.65E+01	6.16E+01	6.11E+01	0	8.15E+01	5.05E+01	5.36E+01	9.32E+01	7.22E+01	8.99E+01	8.21E+01	4.83E+01	1.06E+02	6.30E+01	5.37E+01	5.01E+01	5.25E+01	6.37E+01	7.06E+01	2.85E+01	9.65E+01	1.03E+02
S6	1.35E+02	6.90E+01	8.82E+01	9.76E+01	8.15E+01	0	8.47E+01	7.70E+01	1.32E+02	1.05E+02	5.10E+01	1.21E+02	8.57E+01	1.29E+02	9.81E+01	7.89E+01	6.90E+01	7.90E+01	1.03E+02	1.06E+02	8.77E+01	1.25E+02	1.35E+02
S7	8.63E+01	6.29E+01	5.02E+01	4.77E+01	5.05E+01	8.47E+01	0	4.92E+01	8.37E+01	4.90E+01	9.28E+01	7.03E+01	4.72E+01	8.58E+01	4.43E+01	5.24E+01	5.29E+01	4.59E+01	4.94E+01	5.74E+01	4.96E+01	8.04E+01	9.01E+01
S8	9.74E+01	5.32E+01	4.16E+01	6.27E+01	5.36E+01	7.70E+01	4.92E+01	0	9.66E+01	6.70E+01	8.51E+01	8.66E+01	5.38E+01	9.36E+01	5.73E+01	4.74E+01	5.02E+01	3.22E+01	6.47E+01	6.86E+01	5.77E+01	9.01E+01	9.75E+01
S9	7.80E+01	1.11E+02	8.61E+01	7.64E+01	9.32E+01	1.32E+02	8.37E+01	9.66E+01	0	6.93E+01	1.35E+02	6.73E+01	8.08E+01	8.95E+01	7.39E+01	9.67E+01	1.04E+02	9.26E+01	6.94E+01	7.36E+01	8.65E+01	8.47E+01	7.83E+01
S10	6.81E+01	8.38E+01	5.58E+01	5.75E+01	7.22E+01	1.05E+02	4.90E+01	6.70E+01	6.93E+01	0	1.11E+02	6.04E+01	5.88E+01	7.05E+01	5.60E+01	6.89E+01	7.58E+01	6.39E+01	5.25E+01	5.59E+01	6.43E+01	6.71E+01	7.23E+01
S11	1.38E+02	7.86E+01	9.62E+01	1.04E+02	8.99E+01	5.10E+01	9.28E+01	8.51E+01	1.35E+02	1.11E+02	0	1.24E+02	9.41E+01	1.33E+02	1.04E+02	8.65E+01	7.86E+01	8.86E+01	1.08E+02	1.12E+02	9.60E+01	1.29E+02	1.38E+02
S12	6.81E+01	9.99E+01	7.43E+01	6.19E+01	8.21E+01	1.21E+02	7.03E+01	8.66E+01	6.73E+01	6.04E+01	1.24E+02	0	7.33E+01	8.26E+01	7.07E+01	8.31E+01	9.31E+01	8.38E+01	5.58E+01	6.07E+01	7.59E+01	7.18E+01	7.37E+01
S13	9.17E+01	6.65E+01	5.37E+01	5.35E+01	4.83E+01	8.57E+01	4.72E+01	5.38E+01	8.08E+01	5.88E+01	9.41E+01	7.33E+01	0	9.40E+01	5.55E+01	5.09E+01	5.49E+01	4.67E+01	5.55E+01	6.00E+01	4.47E+01	8.92E+01	9.38E+01
S14	6.74E+01	1.09E+02	7.90E+01	8.60E+01	1.06E+02	1.29E+02	8.58E+01	9.36E+01	8.95E+01	7.05E+01	1.33E+02	8.26E+01	9.40E+01	0	8.55E+01	9.54E+01	1.05E+02	9.34E+01	8.29E+01	7.33E+01	1.00E+02	6.98E+01	7.55E+01
S15	8.77E+01	7.93E+01	5.80E+01	5.47E+01	6.30E+01	9.81E+01	4.43E+01	5.73E+01	7.39E+01	5.60E+01	1.04E+02	7.07E+01	5.55E+01	8.55E+01	0	6.59E+01	7.14E+01	5.40E+01	5.50E+01	5.73E+01	6.13E+01	8.46E+01	9.08E+01
S16	9.70E+01	4.75E+01	5.28E+01	5.79E+01	5.37E+01	7.89E+01	5.24E+01	4.74E+01	9.67E+01	6.89E+01	8.65E+01	8.31E+01	5.09E+01	9.54E+01	6.59E+01	0	4.57E+01	5.06E+01	6.18E+01	6.72E+01	5.90E+01	8.92E+01	9.63E+01
S17	1.09E+02	1.66E+01	5.96E+01	6.37E+01	5.01E+01	6.90E+01	5.29E+01	5.02E+01	1.04E+02	7.58E+01	7.86E+01	9.31E+01	5.49E+01	1.05E+02	7.14E+01	4.57E+01	0	5.13E+01	6.87E+01	7.56E+01	5.99E+01	1.00E+02	1.09E+02
S18	9.53E+01	5.94E+01	4.32E+01	6.09E+01	5.25E+01	7.90E+01	4.59E+01	3.22E+01	9.26E+01	6.39E+01	8.86E+01	8.38E+01	4.67E+01	9.34E+01	5.40E+01	5.06E+01	5.13E+01	0	6.39E+01	6.44E+01	5.44E+01	8.93E+01	9.76E+01
S19	7.60E+01	7.84E+01	6.00E+01	1.21E+01	6.37E+01	1.03E+02	4.94E+01	6.47E+01	6.94E+01	5.25E+01	1.08E+02	5.58E+01	5.55E+01	8.29E+01	5.50E+01	6.18E+01	6.87E+01	6.39E+01	0	4.57E+01	6.21E+01	7.61E+01	8.10E+01
S20	7.07E+01	8.36E+01	5.78E+01	4.80E+01	7.06E+01	1.06E+02	5.74E+01	6.86E+01	7.36E+01	5.59E+01	1.12E+02	6.07E+01	6.00E+01	7.33E+01	5.73E+01	6.72E+01	7.56E+01	6.44E+01	4.57E+01	0	6.62E+01	7.17E+01	7.90E+01
S21	9.75E+01	6.67E+01	6.11E+01	6.01E+01	2.85E+01	8.77E+01	4.96E+01	5.77E+01	8.65E+01	6.43E+01	9.60E+01	7.59E+01	4.47E+01	1.00E+02	6.13E+01	5.90E+01	5.99E+01	5.44E+01	6.21E+01	6.62E+01	0	9.21E+01	9.75E+01
S22	4.17E+01	1.05E+02	7.45E+01	7.92E+01	9.65E+01	1.25E+02	8.04E+01	9.01E+01	8.47E+01	6.71E+01	1.29E+02	7.18E+01	8.92E+01	6.98E+01	8.46E+01	8.92E+01	1.00E+02	8.93E+01	7.61E+01	7.17E+01	9.21E+01	0	5.61E+01
S23	4.48E+01	1.13E+02	7.97E+01	8.64E+01	1.03E+02	1.35E+02	9.01E+01	9.75E+01	7.83E+01	7.23E+01	1.38E+02	7.37E+01	9.38E+01	7.55E+01	9.08E+01	9.63E+01	1.09E+02	9.76E+01	8.10E+01	7.90E+01	9.75E+01	5.61E+01	0

Notes) S1: Tribenuron-methyl; S2: 2,4-D; S3: Aclonifen; S4: Alachlor; S5: Atrazin; S6: Bromoxynil; S7: Carbofuran; S8: Chloridazon; S9: Cycloxydim; S10: Ethofumesate; S11: Ioxynil; S12: Isofenphos; S13: Isoproturon; S14: Isoxaflutol; S15: Lenacil; S16: Linuron; S17: MCPA: S18: Metamitron; S19: Metolachlor; S20: Pendimethalin; S21: Terbuthylazine; S22: Thifensulfuron-methyl; S23: Triasulfuron.

Table S5. Random Forest distances between all pairs of substances in Dataset 1

Compounds	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	9.21E-01	8.77E-01	9.09E-01	8.65E-01	8.55E-01	9.20E-01	9.48E-01	8.72E-01	9.47E-01
S2	9.21E-01	0	8.68E-01	8.81E-01	8.94E-01	8.62E-01	8.81E-01	8.97E-01	8.57E-01	9.15E-01
S3	8.77E-01	8.68E-01	0	8.87E-01	8.77E-01	8.66E-01	9.18E-01	9.12E-01	8.77E-01	9.28E-01
S4	9.09E-01	8.81E-01	8.87E-01	0	9.33E-01	9.28E-01	8.18E-01	8.38E-01	9.11E-01	8.47E-01
S5	8.65E-01	8.94E-01	8.77E-01	9.33E-01	0	7.98E-01	9.43E-01	9.59E-01	8.31E-01	9.38E-01
S6	8.55E-01	8.62E-01	8.66E-01	9.28E-01	7.98E-01	0	9.34E-01	9.56E-01	8.09E-01	9.45E-01
S7	9.20E-01	8.81E-01	9.18E-01	8.18E-01	9.43E-01	9.34E-01	0	8.49E-01	8.94E-01	8.70E-01
S8	9.48E-01	8.97E-01	9.12E-01	8.38E-01	9.59E-01	9.56E-01	8.49E-01	0	9.51E-01	8.12E-01
S9	8.72E-01	8.57E-01	8.77E-01	9.11E-01	8.31E-01	8.09E-01	8.94E-01	9.51E-01	0	9.29E-01
S10	9.47E-01	9.15E-01	9.28E-01	8.47E-01	9.38E-01	9.45E-01	8.70E-01	8.12E-01	9.29E-01	0

Notes) S1: Thiophanate-methyl; S2: Alachlor; S3: Chlorfenvinphos; S4: Cyanazine; S5: Fenamidone; S6: Furalaxyl; S7: Isoproturon; S8: MCPA; S9: Naproamide; S10: Thiabendazole.

Table S6. Random Forest distances between all pairs of substances in Dataset 2

Compounds	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23
S1	0	9.75E-01	8.82E-01	8.77E-01	9.65E-01	9.75E-01	9.16E-01	9.51E-01	7.53E-01	7.88E-01	9.77E-01	7.37E-01	9.33E-01	7.15E-01	8.92E-01	9.44E-01	9.72E-01	9.46E-01	8.34E-01	7.92E-01	9.42E-01	6.45E-01	5.96E-01
S2	9.75E-01	0	8.71E-01	9.06E-01	7.48E-01	6.59E-01	8.54E-01	7.50E-01	9.75E-01	9.48E-01	6.90E-01	9.70E-01	8.53E-01	9.58E-01	9.14E-01	7.37E-01	5.28E-01	7.91E-01	9.34E-01	9.50E-01	8.35E-01	9.57E-01	9.67E-01
S3	8.82E-01	8.71E-01	0	8.93E-01	8.91E-01	8.85E-01	8.66E-01	8.06E-01	8.99E-01	8.54E-01	8.92E-01	8.94E-01	8.89E-01	8.58E-01	8.70E-01	8.49E-01	8.78E-01	8.18E-01	8.94E-01	8.71E-01	8.86E-01	8.68E-01	8.73E-01
S4	8.77E-01	9.06E-01	8.93E-01	0	8.85E-01	9.14E-01	8.38E-01	8.98E-01	8.46E-01	8.54E-01	9.16E-01	8.34E-01	8.73E-01	8.88E-01	8.43E-01	8.78E-01	8.90E-01	8.94E-01	6.90E-01	8.14E-01	8.79E-01	8.86E-01	8.81E-01
S5	9.65E-01	7.48E-01	8.91E-01	8.85E-01	0	7.85E-01	8.33E-01	8.01E-01	9.45E-01	9.23E-01	7.97E-01	9.40E-01	8.09E-01	9.58E-01	8.82E-01	8.05E-01	7.40E-01	7.96E-01	9.07E-01	9.25E-01	6.99E-01	9.51E-01	9.57E-01
S6	9.75E-01	6.59E-01	8.85E-01	9.14E-01	7.85E-01	0	8.62E-01	7.64E-01	9.80E-01	9.57E-01	4.61E-01	9.72E-01	8.62E-01	9.66E-01	9.13E-01	7.91E-01	6.89E-01	7.93E-01	9.40E-01	9.57E-01	8.46E-01	9.60E-01	9.74E-01
S7	9.16E-01	8.54E-01	8.66E-01	8.38E-01	8.33E-01	8.62E-01	0	8.33E-01	9.08E-01	8.52E-01	8.69E-01	9.02E-01	8.45E-01	9.07E-01	8.19E-01	8.42E-01	8.33E-01	8.24E-01	8.65E-01	8.85E-01	8.35E-01	9.05E-01	9.16E-01
S8	9.51E-01	7.50E-01	8.06E-01	8.98E-01	8.01E-01	7.64E-01	8.33E-01	0	9.55E-01	9.21E-01	7.75E-01	9.52E-01	8.57E-01	9.33E-01	8.69E-01	7.73E-01	7.59E-01	6.61E-01	9.14E-01	9.25E-01	8.38E-01	9.36E-01	9.44E-01
S9	7.53E-01	9.75E-01	8.99E-01	8.46E-01	9.45E-01	9.80E-01	9.08E-01	9.55E-01	0	7.90E-01	9.78E-01	7.19E-01	9.06E-01	8.03E-01	8.43E-01	9.50E-01	9.68E-01	9.42E-01	7.89E-01	8.00E-01	9.15E-01	8.07E-01	7.52E-01
S10	7.88E-01	9.48E-01	8.54E-01	8.54E-01	9.23E-01	9.57E-01	8.52E-01	9.21E-01	7.90E-01	0	9.57E-01	7.89E-01	8.87E-01	7.79E-01	8.42E-01	9.20E-01	9.38E-01	9.09E-01	8.22E-01	8.06E-01	8.94E-01	8.05E-01	7.95E-01
S11	9.77E-01	6.90E-01	8.92E-01	9.16E-01	7.97E-01	4.61E-01	8.69E-01	7.75E-01	9.78E-01	9.57E-01	0	9.70E-01	8.69E-01	9.65E-01	9.15E-01	8.03E-01	7.10E-01	8.07E-01	9.41E-01	9.57E-01	8.57E-01	9.61E-01	9.73E-01
S12	7.37E-01	9.70E-01	8.94E-01	8.34E-01	9.40E-01	9.72E-01	9.02E-01	9.52E-01	7.19E-01	7.89E-01	9.70E-01	0	9.09E-01	8.00E-01	8.63E-01	9.38E-01	9.64E-01	9.44E-01	7.80E-01	7.81E-01	9.08E-01	7.78E-01	7.52E-01
S13	9.33E-01	8.53E-01	8.89E-01	8.73E-01	8.09E-01	8.62E-01	8.45E-01	8.57E-01	9.06E-01	8.87E-01	8.69E-01	9.09E-01	0	9.34E-01	8.64E-01	8.36E-01	8.39E-01	8.20E-01	8.89E-01	9.01E-01	7.78E-01	9.37E-01	9.34E-01
S14	7.15E-01	9.58E-01	8.58E-01	8.88E-01	9.58E-01	9.66E-01	9.07E-01	9.33E-01	8.03E-01	7.79E-01	9.65E-01	8.00E-01	9.34E-01	0	8.81E-01	9.32E-01	9.58E-01	9.34E-01	8.54E-01	8.01E-01	9.41E-01	7.25E-01	7.33E-01
S15	8.92E-01	9.14E-01	8.70E-01	8.43E-01	8.82E-01	9.13E-01	8.19E-01	8.69E-01	8.43E-01	8.42E-01	9.15E-01	8.63E-01	8.64E-01	8.81E-01	0	8.92E-01	9.04E-01	8.50E-01	8.43E-01	8.40E-01	8.72E-01	8.98E-01	8.95E-01
S16	9.44E-01	7.37E-01	8.49E-01	8.78E-01	8.05E-01	7.91E-01	8.42E-01	7.73E-01	9.50E-01	9.20E-01	8.03E-01	9.38E-01	8.36E-01	9.32E-01	8.92E-01	0	7.49E-01	8.06E-01	9.01E-01	9.20E-01	8.39E-01	9.26E-01	9.37E-01
S17	9.72E-01	5.28E-01	8.78E-01	8.90E-01	7.40E-01	6.89E-01	8.33E-01	7.59E-01	9.68E-01	9.38E-01	7.10E-01	9.64E-01	8.39E-01	9.58E-01	9.04E-01	7.49E-01	0	7.82E-01	9.21E-01	9.42E-01	8.24E-01	9.54E-01	9.64E-01
S18	9.46E-01	7.91E-01	8.18E-01	8.94E-01	7.96E-01	7.93E-01	8.24E-01	6.61E-01	9.42E-01	9.09E-01	8.07E-01	9.44E-01	8.20E-01	9.34E-01	8.50E-01	8.06E-01	7.82E-01	0	9.14E-01	9.13E-01	8.22E-01	9.36E-01	9.46E-01
S19	8.34E-01	9.34E-01	8.94E-01	6.90E-01	9.07E-01	9.40E-01	8.65E-01	9.14E-01	7.89E-01	8.22E-01	9.41E-01	7.80E-01	8.89E-01	8.54E-01	8.43E-01	9.01E-01	9.21E-01	9.14E-01	0	7.77E-01	8.93E-01	8.47E-01	8.42E-01
S20	7.92E-01	9.50E-01	8.71E-01	8.14E-01	9.25E-01	9.57E-01	8.85E-01	9.25E-01	8.00E-01	8.06E-01	9.57E-01	7.81E-01	9.01E-01	8.01E-01	8.40E-01	9.20E-01	9.42E-01	9.13E-01	7.77E-01	0	9.06E-01	8.11E-01	8.07E-01
S21	9.42E-01	8.35E-01	8.86E-01	8.79E-01	6.99E-01	8.46E-01	8.35E-01	8.38E-01	9.15E-01	8.94E-01	8.57E-01	9.08E-01	7.78E-01	9.41E-01	8.72E-01	8.39E-01	8.24E-01	8.22E-01	8.93E-01	9.06E-01	0	9.36E-01	9.37E-01
S22	6.45E-01	9.57E-01	8.68E-01	8.86E-01	9.51E-01	9.60E-01	9.05E-01	9.36E-01	8.07E-01	8.05E-01	9.61E-01	7.78E-01	9.37E-01	7.25E-01	8.98E-01	9.26E-01	9.54E-01	9.36E-01	8.47E-01	8.11E-01	9.36E-01	0	6.62E-01
S23	5.96E-01	9.67E-01	8.73E-01	8.81E-01	9.57E-01	9.74E-01	9.16E-01	9.44E-01	7.52E-01	7.95E-01	9.73E-01	7.52E-01	9.34E-01	7.33E-01	8.95E-01	9.37E-01	9.64E-01	9.46E-01	8.42E-01	8.07E-01	9.37E-01	6.62E-01	0

Notes) S1: Tribenuron-methyl; S2: 2,4-D; S3: Aclonifen; S4: Alachlor; S5: Atrazin; S6: Bromoxynil; S7: Carbofuran; S8: Chloridazon; S9: Cycloxydim; S10: Ethofumesate; S11: Ioxynil; S12: Isofenphos; S13: Isoproturon; S14: Isoxaflutol; S15: Lenacil; S16: Linuron; S17: MCPA: S18: Metamitron; S19: Metolachlor; S20: Pendimethalin; S21: Terbuthylazine; S22: Thifensulfuron-methyl; S23: Triasulfuron.

Table S7. Descriptions on the 20 most important descriptors descending by the Mean Decrease Accuracy in clustering chemicals in Dataset 1

Rank	Name	Description	Block	Sub-block
1	SpMax_Dz(e)	leading eigenvalue from Barysz matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Barysz matrix weighted by Sanderson electronegativity (Dz(e))
2	AVS_Dz(v)	average vertex sum from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))
3	EE_H2	Estrada-like index (log function) from reciprocal squared distance matrix	2D matrix-based descriptors	Reciprocal squared distance matrix (H2)
4	SM2_D	spectral moment of order 2 from topological distance matrix	2D matrix-based descriptors	Topological distance matrix (D)
5	Mor05s	signal 05 / weighted by I-state	3D-MoRSE descriptors	Weighted by I-state
6	Wi_B(m)	Wiener-like index from Burden matrix weighted by mass	2D matrix-based descriptors	Burden matrix weighted by mass (B(m))
7	SM1_B(p)	spectral moment of order 1 from Burden matrix weighted by polarizability	2D matrix-based descriptors	Burden matrix weighted by polarizability (B(p))
8	HyWi_Dz(m)	hyper-Wiener-like index (log function) from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
9	ChiA_Dz(e)	average Randic-like index from Barysz matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Barysz matrix weighted by Sanderson electronegativity (Dz(e))
10	VE3_Dz(p)	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
11	SpAbs_Dz(p)	graph energy from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
12	RDF055v	Radial Distribution Function - 055 / weighted by van der Waals volume	RDF descriptors	Weighted by van der Waals volume
13	ON0V	overall modified Zagreb index of order 0 by valence vertex degrees	Topological indices	Vertex degree-based indices
14	HyWi_D/Dt	hyper-Wiener-like index (log function) from distance/detour matrix	2D matrix-based descriptors	Distance / detour matrix (D/Dt)
15	SRW04	self-returning walk count of order 4	Walk and path counts	Self-returning walk counts
16	ATS4s	Broto-Moreau autocorrelation of lag 4 (log function) weighted by I-state	2D autocorrelations	Broto-Moreau autocorrelations
17	SM3_L	spectral moment of order 3 from Laplace matrix	2D matrix-based descriptors	Laplace matrix (L)
18	ZM1	first Zagreb index	Topological indices	Vertex degree-based indices
19	ChiA_Dz(p)	average Randic-like index from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
20	nDB	number of double bonds	Constitutional indices	Basic descriptors

Table S8. Descriptions on the 20 most important descriptors descending by the Mean Decrease Gini Index in clustering the chemicals in Dataset 1

Rank	Name	Description	Block	Sub-block
1	H6i	H autocorrelation of lag 6 / weighted by ionization potential	GETAWAY descriptors	H-indices
2	ATSC6p	Centred Broto-Moreau autocorrelation of lag 6 weighted by polarizability	2D autocorrelations	Centred Broto-Moreau autocorrelations
3	HATS5s	leverage-weighted autocorrelation of lag 5 / weighted by I-state	GETAWAY descriptors	H-indices
4	GATS8i	Geary autocorrelation of lag 8 weighted by ionization potential	2D autocorrelations	Geary autocorrelations
5	Mor31v	signal 31 / weighted by van der Waals volume	3D-MoRSE descriptors	Weighted by van der Waals volume
6	H_RG	Harary-like index from reciprocal squared geometrical matrix	3D matrix-based descriptors	Reciprocal squared geometrical distance matrix (RG)
7	TDB05e	3D Topological distance based descriptors - lag 5 weighted by Sanderson electronegativity	3D autocorrelations	TDB autocorrelations
8	SP07	shape profile no. 7	Randic molecular profiles	Shape profiles
9	RDF065s	Radial Distribution Function - 065 / weighted by I-state	RDF descriptors	Weighted by I-state
10	Eta_L	eta local composite index	ETA indices	Basic descriptors
11	DP04	molecular profile no. 4	Randic molecular profiles	Molecular profiles
12	SM15_EA(bo)	spectral moment of order 15 from edge adjacency mat. weighted by bond order	Edge adjacency indices	Spectral moments
13	SsCH3	Sum of sCH3 E-states	Atom-type E-state indices	E-State sums
14	R8i	R autocorrelation of lag 8 / weighted by ionization potential	GETAWAY descriptors	R-indices
15	RDF070p	Radial Distribution Function - 070 / weighted by polarizability	RDF descriptors	Weighted by polarizability
16	SpMax_Dz(e)	leading eigenvalue from Barysz matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Barysz matrix weighted by Sanderson electronegativity (Dz(e))
17	R7v	R autocorrelation of lag 7 / weighted by van der Waals volume	GETAWAY descriptors	R-indices
18	X0Av	average valence connectivity index of order 0	Connectivity indices	Kier-Hall molecular connectivity indices
19	SpMax3_Bh(v)	largest eigenvalue n. 3 of Burden matrix weighted by van der Waals volume	Burden eigenvalues	Largest eigenvalues
20	Mor01s	signal 01 / weighted by I-state	3D-MoRSE descriptors	Weighted by I-state

Table S9. Descriptions on the 20 most important descriptors descending by the Mean Decrease Accuracy in clustering the chemicals in Dataset 2

Rank	Name	Description	Block	Sub-block
1	SpDiam_Dz(m)	spectral diameter from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
2	Dz	Pogliani index	Topological indices	Vertex degree-based indices
3	SM5_Dz(p)	spectral moment of order 5 from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
4	Chi_B(p)	Randic-like index from Burden matrix weighted by polarizability	2D matrix-based descriptors	Burden matrix weighted by polarizability (B(p))
5	VR1_Dz(p)	Randic-like eigenvector-based index from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
6	SpMaxA_Dz(v)	normalized leading eigenvalue from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))
7	SpMax_Dz(m)	leading eigenvalue from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
8	SpPosLog_B(e)	logarithmic spectral positive sum from Burden matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Burden matrix weighted by Sanderson electronegativity (B(e))
9	VR1_Dt	Randic-like eigenvector-based index from detour matrix	2D matrix-based descriptors	Detour matrix (Dt)
10	Wi_Dz(v)	Wiener-like index from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))
11	SM11_EA(ri)	spectral moment of order 11 from edge adjacency mat. weighted by resonance integral	Edge adjacency indices	Spectral moments
12	Chi_B(i)	Randic-like index from Burden matrix weighted by ionization potential	2D matrix-based descriptors	Burden matrix weighted by ionization potential (B(i))
13	VR1_B(p)	Randic-like eigenvector-based index from Burden matrix weighted by polarizability	2D matrix-based descriptors	Burden matrix weighted by polarizability (B(p))
14	Psi_i_s	intrinsic state pseudoconnectivity index - type S	Topological indices	E-state indices
15	SpAbs_B(e)	graph energy from Burden matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Burden matrix weighted by Sanderson electronegativity (B(e))
16	WiA_B(i)	average Wiener-like index from Burden matrix weighted by ionization potential	2D matrix-based descriptors	Burden matrix weighted by ionization potential (B(i))
17	X1	connectivity index of order 1 (Randic connectivity index)	Connectivity indices	Kier-Hall molecular connectivity indices
18	EE_Dz(m)	Estrada-like index (log function) from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
19	VE3_Dz(p)	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
20	Mor01e	signal 01 / weighted by Sanderson electronegativity	3D-MoRSE descriptors	Weighted by Sanderson electronegativity

Table S10. Descriptions on the 20 most important descriptors descending by the Mean Decrease Gini Index in clustering the chemicals in Dataset 2

Rank	Name	Description	Block	Sub-block
1	SpMaxA_Dz(v)	normalized leading eigenvalue from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))
2	VE3_Dz(p)	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability	2D matrix-based descriptors	Barysz matrix weighted by polarizability (Dz(p))
3	ATS4i	Broto-Moreau autocorrelation of lag 4 (log function) weighted by ionization potential	2D autocorrelations	Broto-Moreau autocorrelations
4	SpPosLog_RG	logarithmic spectral positive sum from reciprocal squared geometrical matrix	3D matrix-based descriptors	Reciprocal squared geometrical distance matrix (RG)
5	ATS1s	Broto-Moreau autocorrelation of lag 1 (log function) weighted by I-state	2D autocorrelations	Broto-Moreau autocorrelations
6	Eig05_AEA(bo)	eigenvalue n. 5 from augmented edge adjacency mat. weighted by bond order	Edge adjacency indices	Eigenvalues
7	R8e	R autocorrelation of lag 8 / weighted by Sanderson electronegativity	GETAWAY descriptors	R-indices
8	SpAbs_B(e)	graph energy from Burden matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	Burden matrix weighted by Sanderson electronegativity (B(e))
9	Eig05_AEA(dm)	eigenvalue n. 5 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices	Eigenvalues
10	SpDiam_Dz(m)	spectral diameter from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
11	VE3_D	logarithmic coefficient sum of the last eigenvector from topological distance matrix	2D matrix-based descriptors	Topological distance matrix (D)
12	VE3_Dz(v)	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))
13	ATS1i	Broto-Moreau autocorrelation of lag 1 (log function) weighted by ionization potential	2D autocorrelations	Broto-Moreau autocorrelations
14	HATS0v	leverage-weighted autocorrelation of lag 0 / weighted by van der Waals volume	GETAWAY descriptors	H-indices
15	IAC	total information index on atomic composition	Information indices	Basic descriptors
16	RDF035v	Radial Distribution Function - 035 / weighted by van der Waals volume	RDF descriptors	Weighted by van der Waals volume
17	VR1_Dz(Z)	Randic-like eigenvector-based index from Barysz matrix weighted by atomic number	2D matrix-based descriptors	Barysz matrix weighted by atomic number (Dz(Z))
18	RDF035v	Radial Distribution Function - 035 / weighted by van der Waals volume	RDF descriptors	Weighted by van der Waals volume
19	SpAD_Dz(m)	spectral absolute deviation from Barysz matrix weighted by mass	2D matrix-based descriptors	Barysz matrix weighted by mass (Dz(m))
20	Wi_Dz(v)	Wiener-like index from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors	Barysz matrix weighted by van der Waals volume (Dz(v))

CHAPTER VI

Synthesis and General Conclusions

SYNTHESIS AND GENERAL CONCLUSIONS

This chapter synthesizes the major results observed from the four sub-topics (Chapters II to V) conducted in this study and derives a final conclusion based on the respective results. Future outlook for further studies is also presented in this chapter.

1. Conclusions

The results derived in this study lead us to the following conclusions:

i) Table 1 shows a brief summary of studies related to the major integrated models published from 1997 to 2010 for predicting toxicity of chemical mixtures in the environment. A conceptual relationship network of the integrated models is illustrated in Figure 1. The conceptual relationship network presents how different model concepts and algorithms are theoretically related to each other to develop integrated models for predicting mixture toxicity. Nine of seventeen integrated models surveyed in this study belonged to QSAR models developed for the single-compound- or mixture toxicity to ultimately estimate the toxicity of target mixtures, but the QSAR models had no conceptual relationships to the CA and IA models. For instance, Altenburger et al. (2005) applied a QSAR model developed for nitrobenzenes to estimate the toxicity of their mixtures, and then their predicted toxicity values were used to calculate their mixture toxicity by using a combined CA and IA model. Whereas, Zhang et al. (2007) used QSAR models developed for directly assessing the toxicity of polar and non-polar narcotic mixtures by using non-empirical descriptors. The CA and IA models, however, were basically employed in the IAM, IAI, and MLA models. As combining the CA and IA models, the existing integrated models mostly presented good prediction results for estimating the toxicity of complex mixtures containing different MoA groups; however, they were more data-demanding (for dose-response curves, and MoAs) than the CA and IA models. Among those integrated

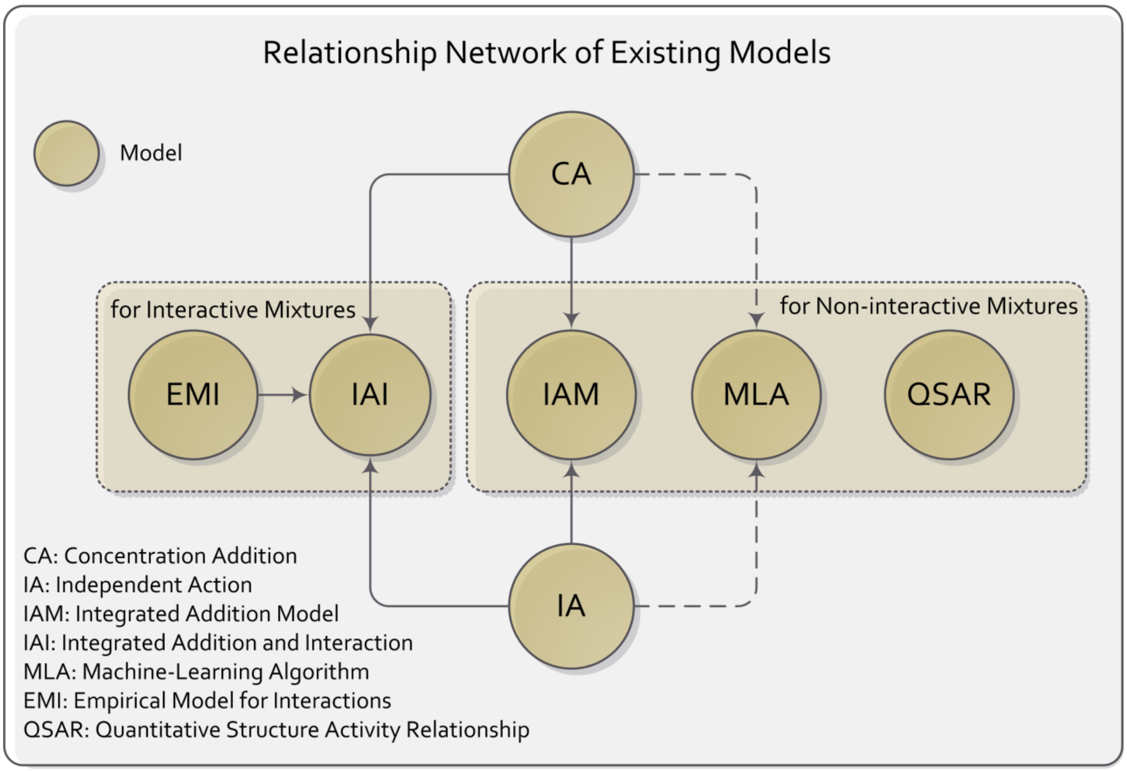
Relationship Network of Existing Models
Model
CA
for Interactive Mixtures
for Non-interactive Mixtures
EMI
IAI
IAM
MLA
QSAR
IA
CA: Concentration Addition
IA: Independent Action
IAM: Integrated Addition Model
IAI: Integrated Addition and Interaction
MLA: Machine-Learning Algorithm
EMI: Empirical Model for Interactions
QSAR: Quantitative Structure Activity Relationship

Figure 1. A conceptual relationship network of the existing models in Table 1 surveyed in this study.

models, the IAI model was a highly data-demanding model since the IAI model additionally used an empirical model to determine experimental coefficients for describing interactions among mixture components. This fact becomes a critical barrier for applying such models to predict mixture toxicity in practice. Therefore, not only to increase the accuracy of prediction models, but also to minimize the burden of data generation for model calculations, the advanced models need to be developed continuously;

Table 1. A brief summary on studies related to the integrated models (published from 1997 to 2010) for predicting toxicity of chemical mixtures in the environment (modified from Table 1 in Chapter II)

Model type	Nº of related studies	Remarks
IAM[1)]	3	· integrated CA[5)] and IA[6)] models
IAI[2)]	1	· integrated CA and IA models using empirical constants determined by experimental test for toxicological interactions between chemicals
QSAR[3)]	9	· 7 studies on empirical QSAR models using partition coefficients; and 2 non-empirical models using quantum descriptors as predictors
MLA[4)]	4	· 3 studies uses Fuzzy theory; and 1 uses ANN[7)] algorithm
Total	17	-

Notes. 1) integrated addition model; 2) integrated addition and interaction model; 3) quantitative structure-activity relationship-based model; 4) machine learning algorithm-based model; 5) concentration addition; 6) independent action; and, 7) artificial neural network.

ii) The three hypotheses described in Chapter I were tested to achieve the objective of this study for developing integrated prediction models which overcome the limitations of existing integrated models for estimating the toxicity of non-interactive mixtures. The study described in Chapter III for 'the hypothesis I' was the first to investigate and yield supportive evidence based on a case study and a computational simulation for evaluating major factors influencing the KCC and CR methods used in determining the PNEC and DNEL of mixtures. This observation necessarily leads us to conclude that the number of mixture components with similarly weighted PNECs and DNELs in the same exposure pathway first requires checking suitability before the application of the KCC or CR methods. From a risk assessment point of view, we firstly suggest that the CR method becomes a general default method for the sake of regulatory purposes based on 'the precautionary principal' if a choice between the two methods is given. The reason for this belief was clearly illustrated and discussed by the results of the case study and computational simulation in this study. The CR method appears more conservative than the KCC method because the KCC method basically ignores additive toxicity, which is a combined effect among components. In addition to the conservatism of the CR method, this method may give manufacturers or formulators, who function as risk assessors, the possibility to conduct a preliminary assessment on what components in a mixture need to be screened or substituted with compounds of less (or no) concern in their development process in order to produce safer mixture products. As a tentative alternative to applying either the KCC or CR method, we also propose a tiered approach that integrates the e-KCC and CR methods for satisfying the precautionary principle as well as maintaining the advantages of the original KCC and CR methods simultaneously. The case study and simulation showed that the e-KCC method might be used to maintain the advantage of the original KCC method and reduce concern about the non-additive toxicity concept of the

KCC method. The PNEC and DNEL values calculated by the e-KCC method were less than those produced from the CR method. Therefore, the CR method can be considered as the second tier only when the risk characterization ratio (*e.g.*, exposure levels to DNELs or PNECs) derived from the e-KCC method exceeds 1. Nevertheless, the KCC and CR methods ultimately require updating or substitution by more scientific concepts and methodologies for better risk assessment of mixtures;

iii) The PLS-IAM developed in Chapter IV for 'the hypothesis II' combined the CA model with the IA model based on the partial least squares regression technique, in order to overcome the critical limitation of the ICIM model, *i.e.*, the multicollinearity problem. Through the four test datasets, this study showed that the PLS-IAM overall outperformed the other reference models, including the CA, IA, and ICIM models. Therefore, it was shown that the PLS-IAM might be useful when the toxicity data of similar mixtures having the same compositions are available. Nevertheless, further studies need to be conducted to determine the following: 1) how the difference in DRC shapes between training and test datasets influences the prediction accuracy of the PLS-IAM; and, 2) how reliably the PLS-IAM predicts the high effect concentrations (>50%) of non-interactive mixtures when the training dataset composed of substances in the very low effect concentration (<5%) range is used;

iv) Through the study described in Chapter V for 'the hypothesis III', the QSAR-TSP model based on the structural information of each compound successfully developed and estimated mixture toxicity in the absence of knowledge on MoAs of mixture components. This advantage of the QSAR-TSP model reflects potential to overcome the critical limitation of not only the conventional TSP model, which requires knowledge on the MoAs of all chemicals, but also that of the CA and IA models, which

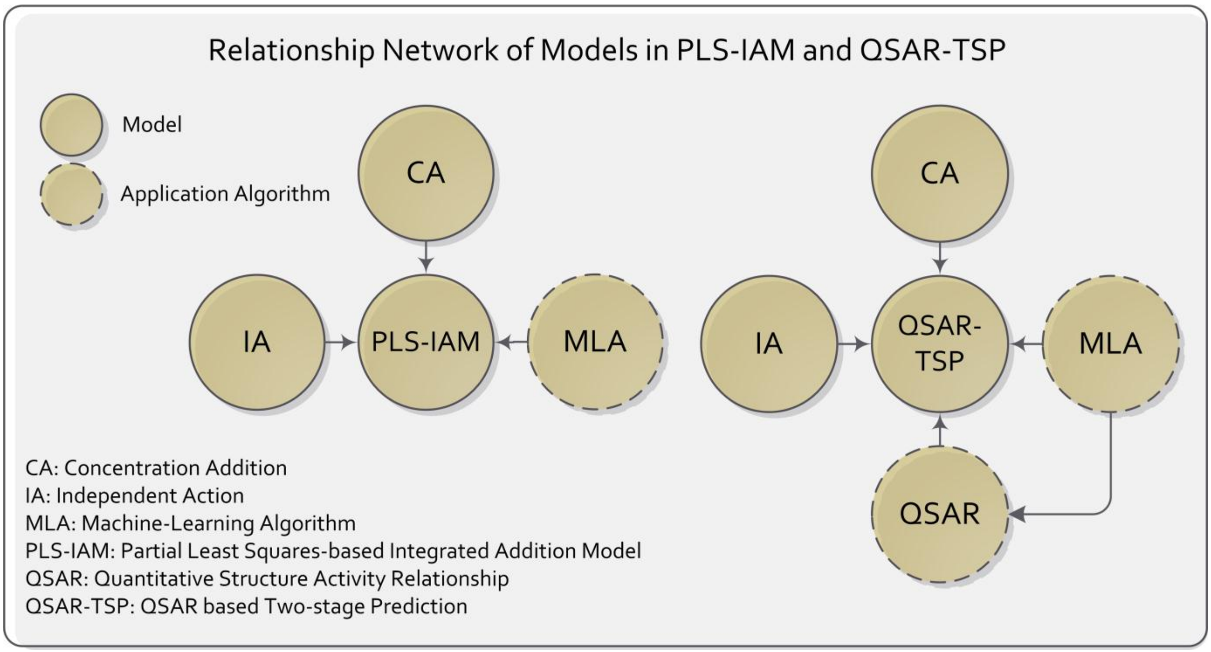
Relationship Network of Models in PLS-IAM and QSAR-TSP
Model
Application Algorithm
CA
IA
PLS-IAM
MLA
CA
IA
QSAR-TSP
MLA
QSAR
CA: Concentration Addition
IA: Independent Action
MLA: Machine-Learning Algorithm
PLS-IAM: Partial Least Squares-based Integrated Addition Model
QSAR: Quantitative Structure Activity Relationship
QSAR-TSP: QSAR based Two-stage Prediction

Figure 2. A conceptual relationship network of models in the PLS-IAM and QSAR-TSP developed in this study.

can be theoretically limit to either similarly or dissimilarly acting chemicals. In addition, the relatively important descriptors used in calculations of structural information for clustering chemicals in the three target mixtures were found by the RF analysis in this study. Further studies for the validation of the QSAR-TSP model need to be conducted with toxicity data based on different types of mixtures and test organisms;

Consequently, when comparing with the existing models shown in Figure1 and Table 1, the PLS-IAM and QSAR-TSP models successfully employed the MLA and QSAR techniques to integrate the CA and IA models as well as minimizing the burden of data generation. Figure 2 illustrates a conceptual relationship network of models and algorithms used in the PLS-IAM and QSAR-TSP models. This study presents good potential for these integrated models, which consider various non-interactive constituents having different MoA groups, and can be used to increase the reliance of conventional models. Figure 3 shows these models also simplify the conventional procedure of

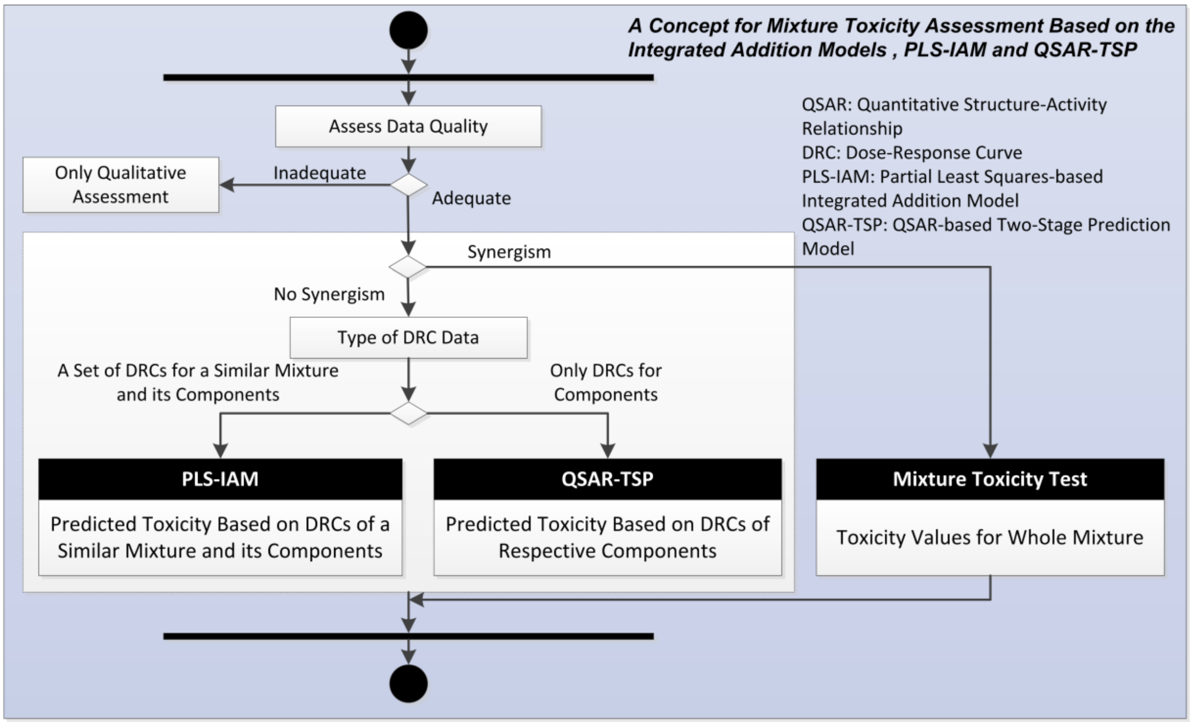
A Concept for Mixture Toxicity Assessment Based on the Integrated Addition Models , PLS-IAM and QSAR-TSP
Assess Data Quality
Only Qualitative Assessment
Inadequate
Adequate
QSAR: Quantitative Structure-Activity Relationship
DRC: Dose-Response Curve
PLS-IAM: Partial Least Squares-based Integrated Addition Model
QSAR-TSP: QSAR-based Two-Stage Prediction Model
Synergism
No Synergism
Type of DRC Data
A Set of DRCs for a Similar Mixture and its Components
Only DRCs for Components
PLS-IAM
Predicted Toxicity Based on DRCs of a Similar Mixture and its Components
QSAR-TSP
Predicted Toxicity Based on DRCs of Respective Components
Mixture Toxicity Test
Toxicity Values for Whole Mixture

Figure 3. A concept for mixture toxicity assessment based on the integrated addition models, the PLS-IAM and QSAR-TSP.

mixture risk assessment, as described in Figure 1 in Chapter II, from the scientific perspective. For non-interactive mixtures, the PLS-IAM might be useful when the toxicity data of similar mixtures having the same compositions are available. In case of no available data on the toxicity of similar mixtures and MoAs of every component, the QSAR-TSP can be considered for estimating mixture toxicity with only DRCs of the components.

2. Outlook: A blueprint for 'Smart Assessment Tools for Mixture Toxicity: the Integrated Model of Synergism-Screening and Addition Toxicity'

Although the PLS-IAM and QSAR-TSP models developed as the IAMs showed excellent results for predicting the toxicity of different mixtures used in this study, further studies with

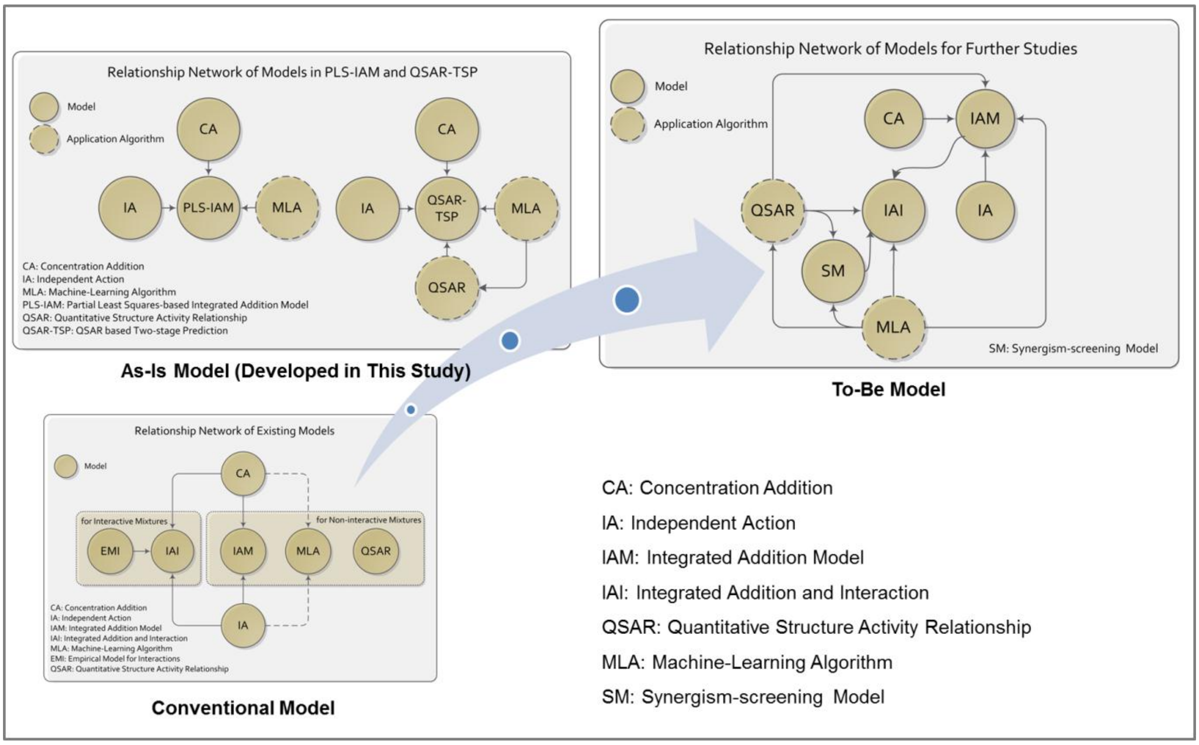
Relationship Network of Models in PLS-IAM and QSAR-TSP
Model
Application Algorithm
CA
IA
PLS-IAM
MLA
CA
IA
QSAR-TSP
MLA
QSAR
CA: Concentration Addition
IA: Independent Action
MLA: Machine-Learning Algorithm
PLS-IAM: Partial Least Squares-based Integrated Addition Model
QSAR: Quantitative Structure Activity Relationship
QSAR-TSP: QSAR based Two-stage Prediction
As-Is Model (Developed in This Study)
Relationship Network of Models for Further Studies
Model
Application Algorithm
CA
IAM
QSAR
IAI
IA
SM
MLA
SM: Synergism-screening Model
To-Be Model
Relationship Network of Existing Models
Model
CA
for Interactive Mixtures
for Non-interactive Mixtures
EMI
IAI
IAM
MLA
QSAR
IA
CA: Concentration Addition
IA: Independent Action
IAM: Integrated Addition Model
IAI: Integrated Addition and Interaction
MLA: Machine-Learning Algorithm
EMI: Empirical Model for Interactions
QSAR: Quantitative Structure Activity Relationship
Conventional Model
CA: Concentration Addition
IA: Independent Action
IAM: Integrated Addition Model
IAI: Integrated Addition and Interaction
QSAR: Quantitative Structure Activity Relationship
MLA: Machine-Learning Algorithm
SM: Synergism-screening Model

Figure 4. Conceptual relationship networks of the conventional and newly developed models (‘As-Is’ models) in this study. A future, 'To-Be' model for the integrated addition and interaction model using various computational approaches is also presented.

various types of mixtures and test organisms are needed to verify and validate applicability. In addition, the application of the PLS-IAM and QSAR-TSP models are limited to non-interactive mixture components. Therefore, it is necessary to conduct further studies for developing a comprehensive integrated model for estimating the additive toxicity as well as the synergistic effects that may occur among chemicals in the long term (as shown in Figure 4).

The generation of an 'ultimate model' to predict additive toxicity and synergistic effects still seems to be fleeting. This is due to the fact that knowledge on the biological mechanisms of mixture toxicity on diverse living organisms lacks, and also the difficulty in empirically assessing and finding synergism among extremely large numbers of chemicals exists in practice. Since the quantitative prediction of synergism on the basis of toxicity among components seems to be considerably difficult to attain in the near future, we carefully

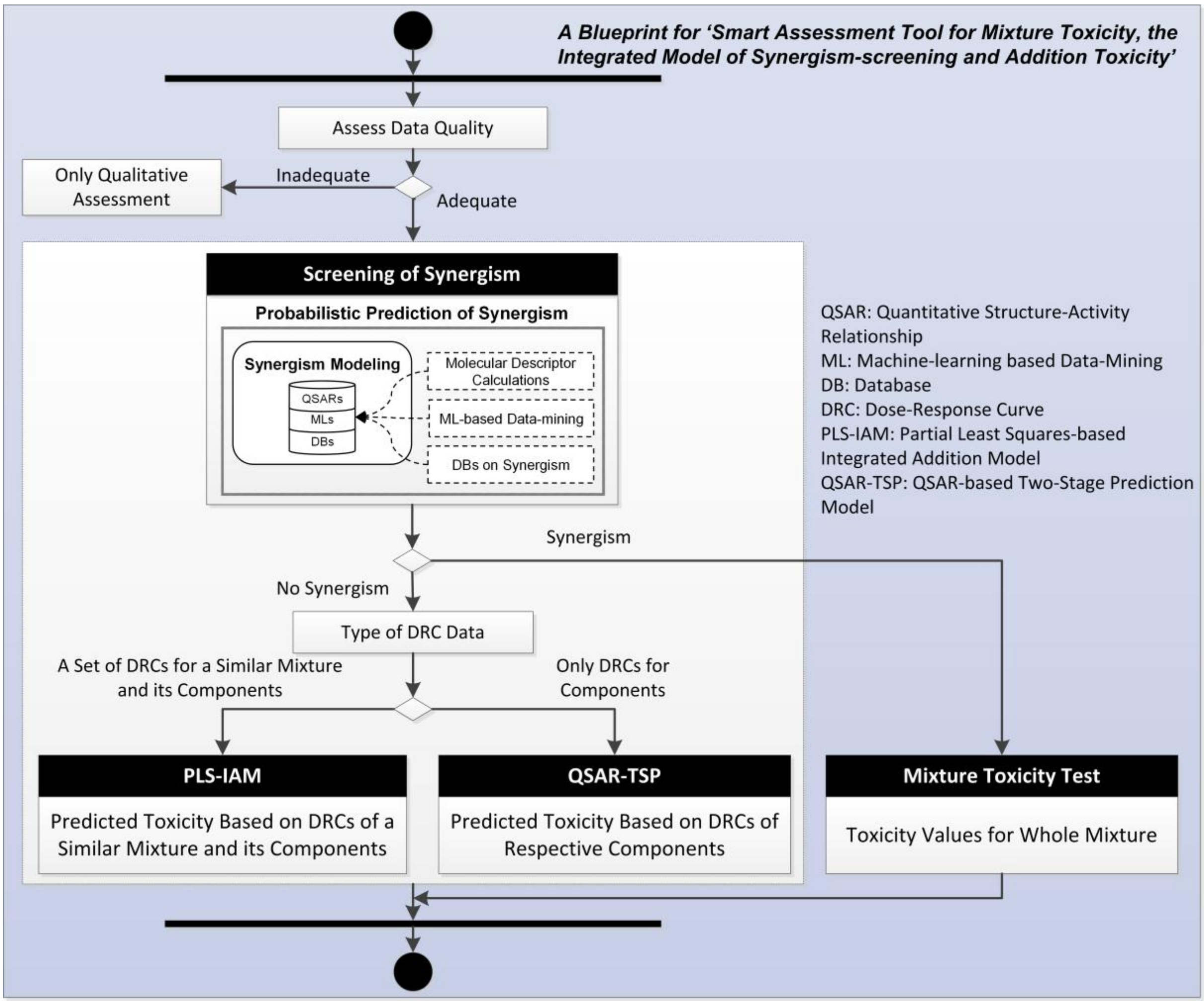
A Blueprint for 'Smart Assessment Tool for Mixture Toxicity, the Integrated Model of Synergism-screening and Addition Toxicity'
Assess Data Quality
Only Qualitative Assessment
Inadequate
Adequate
Screening of Synergism
Probabilistic Prediction of Synergism
Synergism Modeling
QSARs
MLs
DBs
Molecular Descriptor Calculations
ML-based Data-mining
DBs on Synergism
QSAR: Quantitative Structure-Activity Relationship
ML: Machine-learning based Data-Mining
DB: Database
DRC: Dose-Response Curve
PLS-IAM: Partial Least Squares-based Integrated Addition Model
QSAR-TSP: QSAR-based Two-Stage Prediction Model
Synergism
No Synergism
Type of DRC Data
A Set of DRCs for a Similar Mixture and its Components
Only DRCs for Components
PLS-IAM
Predicted Toxicity Based on DRCs of a Similar Mixture and its Components
QSAR-TSP
Predicted Toxicity Based on DRCs of Respective Components
Mixture Toxicity Test
Toxicity Values for Whole Mixture

Figure 5. A blueprint for 'Smart Assessment Tools for Mixture Toxicity: the Integrated Model of Synergism-Screening and Addition Toxicity (IMSAT) model'.

sketched a blueprint for a 'Smart Assessment Tool for Mixture Toxicity: the Integrated Model of Synergism-Screening and Addition Toxicity (IMSAT) model' as an alternative concept that can screen the synergism qualitatively as shown in Figure 5.

Figure 5 was derived from Figures 1 and 3 in Chapter II, which show a general mixture risk assessment concept and the concept of the IMSAT model, respectively, by applying the assessment concept of mixture toxicity based on the PLS-IAM and QSAR-TSP models illustrated in Figure 3 of this chapter. With the hypothesis that a large number of datasets detailing synergism will become available in the future, and that there are hidden relationships between the predictor (descriptor) and response (synergism), the various data-mining techniques can be considered to search for relationships and patterns. These ideas may also form the basis of further studies for a synergism-screening module of the integrated model as outlined in Figure 5.

In order to determine how much synergism data are available at the present time, a literature survey of synergism was performed by collecting journal articles and reviews published from 1997 to 2010 in the fields of toxicology, environmental science, and engineering as available on the ISI Web of Science database in April 2011. From a total of 304 journal articles, the synergism combination list for binary mixtures of pesticides could be compiled and summarized as follows:

- The number of total synergism combinations was 185 (including 98 combinations of pesticide synergists);
- The number of total non-synergism combinations was 106; and,
- The largest number of synergism combinations across the taxonomic groups surveyed was 88 on Insecta.

Table 2 shows a brief summary of the synergism combination list. Unfortunately, the current number of synergism combinations for each taxonomic group does not seem (yet) to be enough for consideration of use in data-mining techniques. This is because the data-mining techniques are generally useful when the number of sample data is larger than that of the variables and when the number of positive and negative datasets is almost balanced.

Nevertheless, it is still expected that the potential of data-mining techniques can be tested if data on synergism are sufficiently available for the techniques in the future, or if any algorithm can be used or developed to accommodate the current situation. Understanding all the mechanisms in mixture toxicity of environmental pollutants is virtually unfeasible, thus, new concepts should be utilized to develop more advanced predictive tools for mixture toxicity.

Table 2. Brief summary of the synergism combination list in binary mixtures of pesticides surveyed in this study

Taxonomic group	Combination	Test organism
Synergism combinations:	***185***	
Algae	11	*Raphidocelis subcapitata; Dunaliella tertiolecta; Chlamydomonas einhardtii; Scenedesmus vacuolatus*
Amphibia	3	*Xenopus laevis; Larbal amphibians(Rana pipiens; Bufo americanus*
Bacteria	11	*Vibrio fischeri; Vibrio-qinghaiensis sp.-Q67; activated sludge microorganisms; Bacilus thuringiensis*
Crustacea	30	*Daphnia magna Straus; Schizopera knabeni; Hyalella azteca; Daphnia magna; Tigriopus brevicornis; Homarus americanus; Ceriodaphnia dubia*
Osteichthyes	17	*Pimephales promelas; Oreochromis niloticus; Tilapia Nilotica fish; Oreochromis mossambicus; Oncorhynchus mykiss; acific Salmon; Gambusia yucatana; Channa punctatus; Carassius auratus*
Fungi	2	*Fusarium oxysporum*
Insecta	88	*Chironosmus tentans; Aedes aegypti; Culex quinquefasciatus; Culex pipiens pallens Coq; Plutella xylostella; Culex quinquefasciatus; Oligonychus pratensis; Sesamia nonagrioides; Boophilus microplus; Grain weevil; Apis mellifera; Diglyphus begini*
Mammalia	21	*Rat; Mouse; Partridge; Coturniz quail*
Mollusca	2	*Crassostrea gigas; Lymnaea acuminata*
Non-synergism combinations:	**106**	
Algae	54	*Chrorella fusca; Scenedesmus vacuolatus; Pseudokirchneriella subcapitata; Pseudokirchneriella subcapitata;*
Monocots	17	*Lemna minor*
Crustacea	5	*Ceriodaphnia dubia; Daphnia magna; Neomysis mercedis; Oncorhynchus mykiss; Lepomis macrochirus; Fundulus heteroclitus*
Osteichthyes	24	*Salmo clarki; Oncorhynchus mykiss; Lepomis macrochirus*
Insecta	6	*Chironomus tentans*

REFERENCES

Altenburger, R., Schmitt, H., Schüürmann, G., 2005. Algal toxicity of nitrobenzenes: Combined effect analysis as a pharmacological probe for similar modes of interaction. Environ. Toxicol. Chem. 24, 324-333.

Zhang, L., Zhou, P.J., Yang, F., Wang, Z.-d., 2007. Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives. Chemosphere 67: 396-401.

LIST OF ABBREVIATIONS

α, β, γ, δ	regression coefficients, or empirical constants
A^{MH}, B^{MH}	joint effects of hydrogen bond in a mixture (similar to Lewis acidity), which are quantified by different partition coefficients of a mixture in various organic phase/water systems
AChE	acetylcholinesterase
$Adj.\ R^2_{test}$	adjusted coefficient of determination for modeled data
AF	assessment factor
AIC	Akaike's Information Criterion
ASW	average silhouette width
b_0	constant
b_1, b_2	regression coefficient
CA	concentration addition
Ca	modifies the effective concentration of chemical *i*
CAS RN	chemical abstracts service registry number
C_{CLi}	total concentration of the ith cluster having similar chemical structures
CEFIC	European chemical industry council
C_i	concentration of the *i*th substance
CLP	classification, labeling and packaging regulation
C_{mix}	concentration of a mixture
CR	composite reciprocal
CSA	chemical safety assessment
DNEL	derived no-effect level
$DNEL_i$	DNEL of the i^{th} substance
$DNEL_{mixture}$	DNEL of a mixture
DPD	dangerous preparation directive
DRC	dose-response curve
E	effect
$E(C_i)$	individual effect of the ith substance if present in the concentration C
$E(C_{mix})$	total effect of the mixture with the total concentration
$E(C_{mix}, C_{Li})$	mixture effect at total concentration of the i_{th} cluster
$E(C_{mix,mix})$	combined effect from different clusters
$E(ECx_{mix})$	overall effect caused by the total effect concentration $ECx_{mix,}$ of a mixture
E_{HOMO}	energy of the highest occupied molecular orbital
E_{LUMO}	energy of the lowest unoccupied molecular orbital
EC_{50i}	concentration of the *i*th chemical that causes 50% of the maximum response (effect)
EC_{50M}	concentration of a mixture that causes 50% of the maximum response (effect)
ECx	concentration that causes the effect *x*
ECx_i	concentration of the *i*th substance that causes the effect *x*

ECx_{mix}	total concentration of substances in a mixture that causes the total effect x%
$ECx_{mix,exp}$	experimental concentration of a mixture eliciting x% toxicity effect
EMI	empirical model for interactions
ES	exposure scenario
EU	European Union
f_i	weight fraction; or function used to describe the DRC of the ith component.
$GAP_{h\text{-}lM}$	difference of E_{HOMO} and E_{LUMO}
$GAPV_{mM}$	absolute value of the difference of a binary mixture's molar volume
GHS	globally harmonized system
HOMO	highest occupied molecular orbital
HPLC	high performance liquid chromatography
HRAC	herbicide resistance action committee
IA	independent action
IAI	integrated addition and interaction
IAM	integrated addition model
IC_{50mix}	50% of the inhibition concentration of the mixture
ICIM	integrated concentration addition with independent action based on a MLR model
IPPC	integrated pollution and prevention control directive
$k_{a,i}$	a function describing the extent to which chemical a presents in the mixture as concentration Ca modifies the effective concentration of chemical i
K_{bw}	t-butyl ether-water partition coefficient
K_{chw}	chloroform-water partition coefficient
K_{cw}	cyclohexane-water partition coefficient
K_{MD}	the C18-EmporeTM disk/water partition coefficient for a mixture
K_{MOW}	octanol-water partition coefficient of a mixture
K_{OW}	octanol-water partition coefficient
K_{SDi}	partition coefficient of the single chemical i
K_{tw}	carbon tetrachloride–water partition coefficient
KCC	key critical component
$lgEnr_M$	logarithm of the nuclear repulsion energy
LUMO	lowest occupied molecular orbital
$LUMO_{mix}$	LUMO of the mixture
μ	dipole moment
MDA	mean decrease accuracy
MDG	mean decrease Gini index
MLA	machine-learning algorithm
MLR	multi-linear regression
MoA	mode of toxic action
MW	molecular weight
n	total number of single chemicals in a mixture
NaNs	not a numbers
NOAEL	no observed adverse effect level
NOEC	no observed effect concentration

OECD	organization for economic cooperation and development
OLS	ordinary least squares
p'	average power of the individual chemicals within a chemical group
p_i	relative proportion of the *i*th substance expressed as a fraction of the total concentration of substances in the mixture ($p_i = C_i / C_{\mathrm{mix}}$)
P_{mix}	n-octanol/water partition coefficient of the mixture calculated by the summed partitioning of single substances based on the independence assumption
PAM	partitioning around medoids
PBO	P450 inhibitor piperonylbutoxide
PEC	predicted effect concentration
$PNEC_i$	PNEC of the i^{th} substance
$\mathrm{PNEC_{mixture}}$	PNEC of a mixture
$PNEC_{W.i}$	concentration weighted PNEC of the i^{th} substance
PLS	partial least squares
PLS-IAM	PLS-based IAM
PNEC	predicted no-effect concentration
PPP	placing of plant protection Products regulation
q_M^-	largest negative atomic charge on an atom
Q^0_{water}	initial amount of chemical *i*
QSAR	quantitative structure-activity relationship
QSAR-TSP	QSAR-based TSP
R_{mix}	combined toxicity of chemical groups
R-phrase	risk-phrase
R^2_{test}	coefficient of determination for modeled data
REACH	regisgration, evaluation, authorisation, and restriction of chemical
RET	risk-based emission threshold
RF	random forest
RMM	risk management measure
RM	regression model
RSS	residual sum of squares
SDS	safety data sheet
SIDS	OECD screening information dataset
TSP	two-stage prediction
US EPA	environmental protection agency of the United States of America
V	volume of the hydrophobic phase
VCI	German chemical industry association
VLCFA	very-long-chain fatty acid
W	volume of the solution
WF	weight fraction
WFD	water framework directive
x	definite value (concentration) for the effect *E*

Curriculum Vitae

Name:	Jongwoon Kim
Date of Birth:	24.11.1977
Place of Birth:	Suwon, Republic of Korea (South Korea)
Nationality:	Republic of Korea
Address:	KIST Europe, Campus E7.1, D-66123 Saarbrücken, Germany
Email:	jwkim@kist-europe.de / with.jwkim@gmail.com

EDUCATION

2/2005 - 3/2003

MASTER OF SCIENCE IN ENVIRONMENTAL SCIENCE AND ENGINEERING (M.SC.)

Hankuk University of Foreign Studies

Graduate School, Dept. of Environmental Science and Engineering, Seoul, Korea

Thesis Title: *Study of Factors Affecting on the Remediation of Diesel Contaminated Soil by Microwave-enhanced SVE*

- Researcher in Hazardous Materials Laboratory (2003 to 2004)
- Teaching Assistant for Waste Water Treatment Unit Process (2004)

2/2003 - 3/1996

BACHELOR OF SCIENCE IN ENVIRONMENTAL SCIENCE (B.SC. 1ST MAJOR), AND

BACHELOR OF ARTS IN ENGLISH (B.A. 2ND MAJOR)

Hankuk University of Foreign Studies

College of Natural Science, Dept. of Environmental Science, Yongin, Korea

Thesis Title: *Removal of Semi-volatile Organic Compounds from Diesel Contaminated Soil by Thermally Enhanced SVE System Using Microwave Heating*

- Research Assistant in Hazardous Materials Laboratory (2001 to 2002)
- Research Assistant in Air Pollution Control Laboratory (1997)

PROFESSIONAL EXPERIENCE

Present - 2007

KIST EUROPE, Saarbrücken, Germany

Korea Institute of Science and Technology

Chemical Risk Management Laboratory

Position: Research Scientist / Ph.D. candidate

2007	National Institute of Environmental Research, Incheon, Korea
- 2006	*Ministry of Environment*
	Department of Chemical Registration and Evaluation
	Position: Expert Advisor
2006	DMEC Co, Ltd., Seoul, Korea
- 2005	*Innovation Business Team in the Head Office*
	Position: Senior Researcher
2005	Han River Environment Research, Yangpyung, Korea
- 2004	*National Institute of Environmental Research*
	Water Environment Monitoring and Modeling Laboratory
	Position: Researcher
2004	Institute of Environmental Science, Yongin, Korea
- 2000	*Hankuk University of Foreign Studies*
	Hazardous Materials Laboratory
	Position: Researcher / Teaching Assistant (2004)
2000	Military Service, Jechun, Korea
- 1998	*Republic of Korea Army*
	Position: Sergeant
1998	Institute of Environmental Science, Yongin, Korea
- 1997	*Hankuk University of Foreign Studies*
	Air Pollution Control Engineering Laboratory
	Position: Research Assistant

PUBLICATIONS

Peer Reviewed Publications:

- "Development of QSAR-based Two-Stage Prediction Model for Estimating Mixture Toxicity", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *SAR and QSAR in Environmental Research,* 2013. DOI:10.1080/1062936X.2013.815654. (Accepted on 29 April 2013).
- "A Case Study and a Computational Simulation of the European Union Draft Technical Guidance Documents for Chemical Safety Assessment of Mixtures: Limitations and a Tentative Alternative", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *Journal of Occupational and Environmental Hygiene*, 2013, 10:181-193.

- “Reliable Predictive Computational Toxicology Methods for Mixture Toxicity: Toward the Development of Innovative Integrated Models for Environmental Risk Assessment”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *Reviews in Environmental Science and Bio/Technology*, 2012, DOI:10.1007/s11157-012-9286-7. Published online (27 June 2012).
- “Development of a Partial Least Squares-Based Integrated Addition Model for Predicting Mixture Toxicity”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *Human and Ecological Risk Assessment: An International Journal*, 2012, DOI:10.1080/10807039.2012.754312. Accepted author version posted online (7 December 2012).
- “Technical Review on Methodology of Generating Exposure Scenario in eSDS of EU REACH”, Eun Kyung Choe, Jongwoon Kim, Sang Hun Kim, Sung Won Byun, *The Korean Society of Clean Technology, Vol. 7, No.4,* 2011 (Korean).
- “Remediation of the Diesel Contaminated Soils Using Thermally Enhanced Soil Vapor Extraction Process with Microwave Heating”, Jongwoon Kim, Kapsong Park, *Korean Society of Soil and Groundwater Environment*, Vol. 9, No.1, pp.39-46, 2004 (Korean).

NON-PEER REVIEWED PUBLICATIONS:

- “Comparative Study of Risk Assessment Approaches Based on Different Methods for Deriving PNEC and DNEL of Chemical mixtures”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *EKC 2009 Proceedings of EU-Korea Conference on Science and Technology*, Springer Proceedings in Physics, Vol. 135, 2010.
- “Research for REACH Compliances of Korean Chemical Industry”, Changro Yun, Sanghun Kim, Jongwoon Kim *et al.*, *Official Report of the Ministry of Environment, Korea*, 2008 (Korean).
- “Construction and Operation for Stream flow and Water Quality Monitoring Network of the Han River Basin”, Dongil Chung, Seuk Chun, Jihyoung Park, Jongwoon Kim, Namhee Kim, *et al.*, *Official Report of The Han River Basin Management Committee, the Ministry of Environment,* 12/2004 (Korean).

ORAL PRESENTATIONS:

- “A QSAR-based Two-Stage Prediction for Estimating Mixture Toxicity”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *32nd SETAC North America Annual Meeting,* Boston, United States of America, 2011.
- “Case Study: Risk Assessment of Chemical Mixtures using different approaches based on PNEC and DNEL”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *The 1st SETAC YES Meeting,* Landau, Germany, 2009.
- “Approach to Ecological Risk Assessment based on Non-Testing Methods under REACH”, Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *The VeKNI Annual Meeting 2007, Verein Koreanischer Naturwissenschaftler und Ingenieure in der BRD e.V.,* Essen, Germany, 2007.

- "Strategic REACH Pre-registration for Company within EU", *Joint Workshop for REACH Implementation, Ministry of Environment and Ministry of Knowledge Economy*, Frankfurt, Germany, 2008
- "Case Study and FAQs for REACH", Jongwoon Kim, *Korea-EU Joint Workshop, KIST*, Seoul, Korea, 2007
- "REACH Compliances and Strategy for EU Chemical Company", Jongwoon Kim, *REACH Training Program, Korea Intl. Trade Association*, Einthoven, The Netherlands, 2007
- "REACH Compliances and Strategy for EU Chemical Company", Jongwoon Kim, *REACH Workshop, KIST Europe*, Frankfurt, Germany, 2007
- "European Chemical Agency and Only Representative", Jongwoon Kim, *REACH Training Program, Samsung Co. Ltd.*, Seoul, Korea, 2007
- "Understanding the REACH with FAQs", Jongwoon Kim, *The 2-day REACH In-depth Seminar for the Chemical Industry* , Seoul, Korea, 2007
- "Introduction to EU REACH Regulation", Jongwoon Kim, *The 5-day Lecturing Tour Seminar of the Ministry of Environment*, Seoul, Korea, 2007.

POSTER PRESENTATIONS:

- "Challenges in Predicting Mixture Toxicity using Computational Toxicology Methods: Toward Integrated Environmental Hazard Assessment", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *13th International Congress of Toxicology 2013*, Seoul, Korea, 2013.
- "Characterization and In-vitro Toxicity Test of Surface Modified Metallic Nanoparticles", Younjung Jung, Seungyun Baik, Jongwoon Kim, Hyunpyo Jeon and Sanghun Kim, *EuroNanoForum*, Dublin, Ireland, 2013.
- "Comparative Ecotoxicological Assessment of Nanomaterials by In vitro Screening Tools", Younjung Jung, Seungyun Baik, Jongwoon Kim, Hyunpyo Jeon and Sanghun Kim, *13th International Congress of Toxicology 2013*, Seoul, Korea, 2013.
- "A Partial Least Squares Based Integrated Addition Model for Estimating Mixture Toxicity", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *6th SETAC World Congress 2012, 22nd SETAC Europe Annual Meeting*, Berlin, Germany, 2012.
- "Status quo and Challenges in the EU Scheme of Environmental Risk Assessment for Nanomaterials", Jongwoon Kim *et al.*, *6th SETAC World Congress 2012, 22nd SETAC Europe Annual Meeting*, Berlin, Germany, 2012.
- Study on a Tiered Approach based on Occupational Exposure Models under EU REACH: ECETOC TRA & Stoffenmanager, Jongwoon Kim, Jaehong Jang, Jongmoon Cha, Sanghun Kim, Eun Kyung Choe, *Korean Society of Environmental Engineers 2012 Conference*, Changwon, Korea, 2012.
- "Study on the Application of ECETOC TRA Tool", Jongmoon Cha, Jongwoon Kim, Sanghun Kim, Eun Kyung Choe, *The Korean Society of Environmental Health and Toxicology 2012 Conference*, Seoul, Korea, 2012.

- "Multi-linear Regression Methods to Develop an Integrated Addition Model for Predicting Mixture Toxicity", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *32nd SETAC North America Annual Meeting,* Boston, United States of America, 2011.
- "Predictability of the Toxicity of a Mixture of Fungicides and Herbicides Using Two-Stage Prediction Model", Jongwoon Kim, Svenja Recktenwald, Sanghun Kim, Gabriele E. Schaumann, Rolf Altenburger, *3rd SETAC Special Symposium,* Brussels, Belgium, 2011.
- "An Application of Different Multi-linear Regression Methods to Develop an Integrated Addition Model for Predicting Mixture Toxicity", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *EU-Korea Conference on Science and Technology 2011,* Paris, France, 2011.
- "Development of Rational Exposure Assessment Strategy under K-REACH for Risk Assessment for Workers by ECETOC TRA", Bohyun Ryu, Jongwoon Kim, Sanghun Kim, *EU-Korea Conference on Science and Technology 2011,* Paris, France, 2011.
- "Prediction of Mixture Toxicity Based on the Categorization of Mixture Constituents Using Computerised Analysis of Chemical Similarity", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *20th SETAC Europe Meeting,* Seville, Spain, 2010.
- "Use of Higher Tiered Exposure Assessment for Worker in Risk Assessment under REACH, Jongmun Cha, Jongwoon Kim, Sanghun Kim, Eunkyung Choe, *EU-Korea Conference on Science and Technology 2010,* Vienne, Austria, 2010.
- "Study on the Validation of ECETOC TRA Tool - An Improved Exposure Estimation Model for Risk Assessment", Yawei Zhang, Jongwoon Kim, Samer Aburous, Sanghun Kim, Eunkyung Choe, *SETAC Asia/Pacific,* Guangzhou, China, 2010.
- "Use of Exposure Assessment Model in Risk Assessment under REACH", Jongmun Cha, Jongwoon Kim, Sanghun Kim, *95th Korean Society for Geosystem Engineering Conference,* Busan, Korea, 2010.
- "Comparative Study of Risk Assessment Approaches Based on Different Methods for Deriving PNEC and DNEL of Chemical Mixtures", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *EU-Korea Conference on Science and Technology 2009,* Reading, United Kingdom, 2009.
- "Case Study on Environmental Risk Assessment of Chemical Mixture", Jongwoon Kim, Sanghun Kim, Gabriele E. Schaumann, *19th SETAC Annual Meeting,* Göteborg, Sweden, 2009.
- "Estimation of Water Quantity per Sub-watersheds for the Han River Basin Management", Jihyoung Park, Saeuk Chun, Jongwoon Kim, et al., *Korean Society of Limnology,* Daejeon, Korea, 2005.
- "Removal of Diesel Hydrocarbons by Microwave-Enhanced Soil Vapor Extraction: Focused on Loss and Kinetic Constants", Jongwoon Kim, Kapsong Park, *Korean Society of Soil and Groundwater Environment*, Incheon, Korea, 2004.

PROJECTS

SCIENCE PROJECTS:

- *“*Development and Application of Sonication on the Removal of Cyanobacteria and its Toxicity in Early Stage”, *Korea Institute of Science and Technology*, 2013 ~ 2016.
- “Establishment of Infrastructure for Nanomaterial Risk Assessment”, *Korea Institute of Science and Technology*, 2012 ~ 2013.
- “Development of Substitution Technology for Substances of Very High Concern: Development of Technology for Exposure Scenario of Chemical Product”, *Korean Ministry of Knowledge Economy*, 2008 ~ 2013.
- “National Agenda: A Study on Chemical Safety Assessment (CSA) under the EU REACH Regulation”, *Korea Institute of Science and Technology*, 2008 ~ 2011.
- “Integrated Approach for Ecological Risk Assessment Based on a Computer-assisted Prediction Model”, A cooperation research project between *University of Koblenz-Landau,* and *Korea Institute of Science and Technology (KIST) Europe*, 2007 ~ 2012.
- “Construction and Operation for Stream Flow and Water Quality Monitoring Network of the Han River Basin”, *Korean Ministry of Environment*, 2004 ~ 2005.

POLICY PROJECTS:

- “Technology Information Survey Report: High-sensitivity Analysis Techniques for Monitoring Chemical Substances in the Environment”, *Korea Institute of Science and Technology Information*, 2012 ~ 2013 (as Principal Investigator (PI) Position).
- “National Environmental Technology Information Report”, *Korea Environmental Industry and Technology Institute*, 2007 ~ present (as Project Manager (PM) Position since 2010).
- “Plan for the Development of Underlying Laws and Ordinances for the Registration and Evaluation of Chemical Substances (Korean REACH-like chemical regulation)”, *Korean Ministry of Environment*, 2012 ~ 2013.
- “Development of Technical Guidance for Implementing the Korean REACH”, *Korean Ministry of Environment*, 2012 ~ 2013.
- “Infrastructure for Pre-compliance with Global Environmental Regulations”, *Korean Ministry of Knowledge Economy*, 2012.
- “Research for REACH Compliances of Korean Chemical Industry”, *Korean Ministry of Environment*, 2008.
- “The State of the Art of the EU REACH Only Representatives”, *Korean Ministry of Environment*, 2008.
- “Development of the EU REACH Navigation Tool for Korean Industry”, *Korean Ministry of Environment*, 2007 ~ 2008.
- “Korea-EU Science & Technology Cooperation Program in Europe”, *Korean Ministry of Science and Technology*, 2006 ~ 2007.
- “Roadmap Development for REACH”, *Korean Ministry of Environment*, 2006 ~ 2007.

INDUSTRY PROJECTS:

- “Development of Compliance Guidance on the EU Cosmetic Regulation”, *Foundation of Korea Cosmetic Industry Institute*, 2012.
- “Strategy for the EU REACH Compliance”, *Samsung Fine Chemicals*, 2008 ~ 2009.
- “REACH Implementation as an Only Representative of Korean Chemical Companies”, *Korean Chemical Industries: Samsung, LG, S-Oil, etc.*, 2008 - Present.

HONORS / AWARDS

- Teaching Assistant Scholarship, Hankuk University of Foreign Studies, Seoul, Korea, Mar. 2004.
- Distinction Scholarship, Hankuk University of Foreign Studies, Seoul, Korea, Sep. 2003.
- Extramural Scholarship, Korea Research Fund, Seoul, Sep. 2002.
- Excellency Scholarship, Hankuk University of Foreign Studies, Yongin, Korea, Mar. 2002.
- Welfare Scholarship, Hankuk University of Foreign Studies, Yongin, Korea, Sep. 2001.
- Second Distinction Scholarship, Hankuk University of Foreign Studies, Yongin, Korea, Sep. 1996.
- Distinction Scholarship II (Department Top Entrance Scholarship) in College of Natural Science, Hankuk University of Foreign Studies, Yongin, Korea, Feb. 1996.

QUALIFICATIONS / ACTIVITIES

LICENSES:

- National Industrial Engineer Water Pollution Environmental, *Republic of Korea*, 2002.
- National Industrial Engineer Office Automation, *Republic of Korea*, 2006.

.