Quantitative Structure-Property Relationships Modeling of Rate Constants of Selected Micropollutants in Drinking Water Treatment Using Ozonation and UV/H_2O_2

by

Xiaohui Jin

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2012

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Concern over the occurrence of micropollutants in drinking water and their health effects is increasing. Therefore, there is a growing interest in understanding micropollutant removal during drinking water treatment. Ozonation and advanced oxidation processes (AOPs) have been found to be effective in the degradation of many micropollutants. Ozonation involves reactions with both molecular ozone (direct pathway) and hydroxyl radicals (indirect pathway), while hydroxyl radicals are the main oxidants in advanced oxidation processes. Reaction rate constants of micropollutants with molecular ozone (k_{O3}) and hydroxyl radicals (k_{OH}) are indicators of their reactivity and are therefore useful in assessing their removal efficiency in ozonation and AOPs. However, to date, only a limited number of rate constants are available for micropollutants, especially emerging micropollutants such as endocrine disrupting chemicals (EDCs) and pharmaceuticals. Quantitative structure-property relationships (QSPR) are therefore desirable for predicting rate constants of numerous untested micropollutants without experimentation. The overall objective of this thesis was to develop predictive QSPR models which correlate the rate constants of a wide range of structural diverse micropollutants to their structural characteristics.

To ensure the wide applicability of the QSPR models, the training set compound selection is critical and a group of heterogeneous compounds which are structurally representative of many others is preferred. A systematic compound selection approach which involves principal component analysis (PCA) and D-optimal onion design was applied for the first time in water treatment research. As a result, 22 micropollutants with diverse structures were selected as representatives from a large pool of micropollutants of interest (182 compounds). In addition, 12 molecular descriptors were identified which link relevant structural features to the removal mechanisms of oxidation processes.

The k_{O3} and k_{OH} values of the 22 selected micropollutants were then determined experimentally in bench-scale reactors at neutral pH using high performance liquid chromatography equipped with a photodiode array detector (HPLC-PDA). Three methods, competition kinetics, compound monitoring, and ozone monitoring were used for k_{O3} measurement, and competition kinetics was used for k_{OH} measurement. As expected, k_{O3} values span a wide range from 10^{-2} to 10^{7} M^{-1} s^{-1} because of the selective nature of molecular ozone. The general trends of micropollutant reactivity with ozone can be explained by the micropollutant structures and the electrophilic nature of ozone reactions. The k_{OH} values range from 10^{8} to 10^{10} M^{-1} s^{-1} because hydroxyl radicals are relatively non-selective in their

reactions. For the majority of these micropollutants k_{O3} and k_{OH} values were not reported prior to this study. Thus they provide valuable information for modeling and designing of ozonation and AOP treatment.

QSPR models for k_{O3} and k_{OH} prediction were then developed with special attention to model validation, applicability domain and mechanistic interpretation. With the experimentally determined rate constants, QSPR models were developed for predicting k_{O3} values using the selected 22 micropollutants as the training set and the 12 identified descriptors as model variables. As a result, two QSPR models were developed using piecewise linear regression (PLR) both showing an excellent goodness-of-fit. Model 1 was governed by average molecular weight and number of phenolic functional groups, and Model 2 was dominated by two principal components extracted from the descriptor matrix. The models were then validated using an external validation set collected from the literature, showing good predictive power of both models. Prior to applying these models to unknown micropollutants they need to be classified as high-reactive ($\log k_{O3} > 2\ M^{-1}\ s^{-1}$) or low-reactive ($\log k_{O3} \leq 2\ M^{-1}\ s^{-1}$), so that the appropriate submodel of the PLR can be applied. A classification function using linear discriminant analysis (LDA) was therefore developed which worked very well for both training and validation sets. With the help of additional compounds collected from the literature, and DRAGON molecular descriptors, a QSPR model for k_{OH} prediction in the aqueous phase was developed using multiple linear regression. As a result, 7 DRAGON descriptors were found to be significant in modeling k_{OH}, which related k_{OH} of micropollutants to their electronegativity, polarizability, presence of double bonds and H-bond acceptors. The model fitted the training set very well and showed great predictive power as assessed by the external validation set. In addition, the model is applicable to a wide range of micropollutants. The model's applicability domain was defined using a leverage approach.

The main contributions of this thesis lie in the successful development of QSPR models for k_{O3} and k_{OH} value prediction which, for the first time, can be used for a wide range of structurally diverse micropollutants. In addition, all QSPR models were externally validated to verify their predictive power, and the applicability domains were defined so that the applicability of the models to new compounds can be determined.

Finally, the applicability of the model to natural water was explored by combining the QSPR models with the established R_{ct} concept which predicts micropollutant removals during ozone treatment of natural water but requires kinetic data as input. Results show that the kinetic data from

the QSPR model predictions worked well in the R_{ct} model providing reliable estimations for most of the selected micropollutants. This approach can therefore be used in water treatment for initial assessment and estimation of ozonation efficiency.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisors Dr. Peter Huck and Dr. Sigrid Peldszus, for their knowledge and guidance of my research. Over the past five years, Dr. Huck has been providing me strong continuous support both scientifically and personally. I also thank Dr. Sigrid Peldszus for her support, patience, encouragement, and speedy feedback to my numerous questions. Her expertise, feedback and guidance were crucial to the completion of my thesis. This dissertation would have been impossible without their great efforts.

I would also like to express my great thanks to Dr. William B. Anderson, Dr. Michele I. Van Dyke, and Ms. Dana Herriman for their valuable advice and support through my study. Special thanks to Dr. Douglas I. Sparkes for his important suggestions to my modeling work. My sincere thanks are extended to the other members of my advisory committee for their many useful comments during this research. They are: Dr. Susan Andrews, Dr. Thomas A. Duever, and Dr. William A. Anderson. I would also like to express my gratitude to Dr. Urs von Gunten for agreeing to be my external examiner and for his thorough review of this thesis.

I am very thankful for the exceptional technical support from Mark Sobon and Terry Ridgway in the laboratory to this research project. I appreciated their patience and strong problem solving abilities. Many thanks to Cynthia Hallé, Zirui Yu, Feisal Rahman, Ahmed El-Hadidy, Fei Chen, Mohamed Hamouda, Shoeleh Shams, Avid Banihashemi, and all the NSERC Chair graduate students for their help, the laughs and the memories.

Finally, I would like to express my most sincere gratitude to my wife Jinghua Li for her unconditional encouragement, inspiration and support to my research. Her love and support is the foundation for me to go through all the difficulties and challenges encountered in this research. She is going to give me the best gift of my life, our first child, in the middle of May this year. I would also like to express my gratitude to my parents for their selfless support.

I acknowledge the Ontario Ministry of Research and Innovations (MRI), and the Natural Sciences and Engineering Research Council of Canada (NSERC) and our Industrial Research Chair partners (www.civil.uwaterloo.ca/watertreatment) for financial support of this research.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANN	Artificial Neural Networks
AOPs	Advanced Oxidation Processes
CASRN	Chemical Abstract Service Registry Number
EDCs	Endocrine Disrupting Chemicals
HPLC	High Performance Liquid Chromatography
LDA	Linear Discriminant Analysis
MLR	Multiple Linear Regression
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PDA	Photodiode Array Detector
PLR	Piecewise Linear Regression
PLS	Partial Least Squares Regression
PPCPs	Pharmaceuticals and Personal Care Products
QSAR	Quantitative Structure-Activity Relationships
QSPR	Quantitative Structure-Property Relationships
SMILES	Simplified Molecular Input Line Entry System
SMD	Statistical Molecular Design
VIF	Variance Inflation Factor
VIP	Variable Importance in the Projection

Chapter 1
Introduction

1.1 Micropollutants and Water Treatment

There has been growing concern about the occurrence of micropollutants in the aquatic environment in recent years. The term micropollutants (or microcontaminants) is used since the concentrations of these contaminants in the aquatic environment are in most cases in the range of ng/L up to μg/L. Pharmaceuticals and personal care products (PPCPs), endocrine disrupting chemicals (EDCs) and pesticides are groups of micropollutants which have been detected in surface water (Heberer 2002; Kolpin *et al*., 2002) and even in finished drinking water (Snyder *et al*., 2007; Benotti *et al*., 2009; Huerta-Fontela *et al*., 2011; Loos *et al*., 2007).

The occurrence of micropollutants such as PPCPs and EDCs can pose a serious problem to the safety of drinking water (Fent *et al*., 2006). For example, exposures to EDCs may disturb hormonal regulation and the normal endocrine system, and affect hormonal balance and reproduction in humans and wildlife (Colborn *et al*., 1993). For most of the PPCPs, their potential impact on the environment and public health is highly unknown (Kümmerer 2001; Stackelberg *et al*., 2004). The primary concern of PPCPs is the potential chronic health effects associated with long term exposure in trace concentration (Snyder 2008). Furthermore, the large number of micropollutants that are present in surface water as a complex mixture can produce combined effects (Cleuvers 2004; Thorpe *et al*., 2003). Based on the precautionary principle, these micropollutants should be removed or at least minimized in drinking water.

Therefore, the efficiency of drinking water treatment processes for the removal of micropollutants from drinking water has been of concern to water utilities and environmental agencies. These micropollutants create unique challenges to water treatment because of the number of compounds detected and the diversity and complexity of their physicochemical properties. Not surprisingly, a wide range of such compounds are not readily removed by conventional water treatment processes such as coagulation/flocculation/sedimentation, dual-media gravity filtration, and chlorination (Westerhoff *et al*., 2005; Stackelberg *et al*., 2004, Bundy *et al*., 2007). However, recent studies have shown that ozonation and advanced oxidation processes (AOPs), adsorption on granular and powdered activated carbon, reverse osmosis and nanofiltration were effective technologies for removing micropollutants from drinking water (von Gunten 2003; Ternes *et al*., 2002, 2003; Snyder *et al*., 2007).

Due to their high oxidation potential, ozonation and AOPs have been widely used in water and wastewater treatment for the oxidation of a wide range of organic compounds. Ozone reacts with organic contaminants through two pathways, direct reaction with molecular ozone, and indirect reaction with hydroxyl radicals produced by ozone decomposition. Molecular ozone reactions are selective to organic molecules having double bonds, activated aromatic systems, and deprotonated amines (von Gunten 2003). In contrast, the hydroxyl radical is a relatively non-selective, highly reactive oxidant. For ozone-resisting compounds, AOPs can be applied for their degradation. AOPs combine chemical agents and auxiliary energy sources to accelerate the generation of hydroxyl radicals (Ikehata *et al*., 2006). Examples of AOPs include O_3/H_2O_2, O_3/UV, UV/H_2O_2, Fenton (Fe^{2+}/H_2O_2), and γ-radiolysis.

To investigate the removal efficiency of various organic micropollutants during ozonation and AOPs in natural waters, it is necessary to obtain kinetic data (i.e., the reaction rate constants of micropollutants with ozone and hydroxyl radicals). Rate constants are needed to predict the extent to which contaminants are eliminated from water. Kinetic data are providing therefore important information for designing and optimizing treatment processes. If rate constants are low unsatisfactory removals may be achieved and additional treatment steps or a different treatment technology may be required. In addition, some models have been developed to describe the removal efficiency of contaminants in natural water matrices incorporating rate constants, e.g., R_{ct} model for ozonation (Elovitz and von Gunten 1999) and $R_{OH,UV}$ model for UV/H_2O_2 AOP (Rosenfeldt and Linden 2007). Although kinetic data are available for a large number of chemicals for their reactions with ozone and hydroxyl radicals (Hoigné and Bader 1983a; Buxton *et al.*, 1988), due to the complexity of the analytical methods and the high cost of determining rate constants experimentally, there is still a data gap especially for emerging micropollutants.

1.2 QSPR Approach

It is impractical to determine the rate constants of all the micropollutants of interest with ozone and hydroxyl radicals. Therefore, it is highly desirable to develop a reliable model to predict the rate constants of numerous micropollutants. Quantitative Structure-Property Relationships (QSPR) has been widely used as a modeling tool to develop relationships between the properties (e.g., p*K*a) of micropollutants and their structural characteristics. Therefore the properties of unstudied contaminants can be predicted without experimentation (Dunn *et al*., 1989; Eriksson and Johansson 1996; Eriksson *et al*., 2003). The QSPR approach has been widely applied in pharmaceutical and

environmental chemistry, and in environmental toxicology. Most recently, an increasing number of papers utilizing QSPR applications in water treatment have been published (Metivier-Pignon *et al*., 2007; Yangali-Quintanilla *et al*., 2010). QSPR models can relate the compound physico-chemical characteristics to their properties (e.g. removal, adsorption, transport, rejection, etc.) in water treatment processes, providing improved knowledge on different removal mechanisms and interactions between organic compounds and physical/chemical treatment processes.

To date, only a small number of studies have been published focusing on predicting the reaction rate constants of micropollutants with ozone (Gurol and Nekoulnalnl 1984; Benitez *et al*., 2007; Hu *et al*., 2000) and hydroxyl radicals (Kusic *et al*., 2009). However, the existing models are based on groups of structural similar compounds, and only applicable to a small range of chemicals. In addition, the existing models were built without proper external validation; therefore, the predictive power of the models is in question. A reliable QSPR model which can be applied to various, structural diverse chemicals and predict the rate constants with reasonable error is currently not available.

1.3 Objectives and Overview of the Thesis

The overall objective of this thesis is therefore to develop reliable QSPR models which correlate ozone and hydroxyl radical reaction rate constants of a wide range of structural diverse micropollutants to their structural characteristics. These models are then used to predict the rate constants of untested micropollutants without experimentation. Furthermore, the goal is to predict the percent removal of micropollutants in natural waters by ozonation and AOPs by combining the predicted rate constants with the existing R_{ct} and $R_{OH,UV}$ models.

Several key elements are essential in QSPR model development which will determine the predictive power of the model. Key elements are the selection of training set compounds, the selection of the molecular descriptors, statistical methods for model development, and model validation (Eriksson *et al*., 2003). To reach the overall goal, this research was divided into several phases with the following sub-objectives:

- Phase one: Selection of structural representative compounds. The objective of this phase is to select a small number of representative compounds from a large pool of structurally diverse micropollutants so that they will cover the entire chemical collection systematically in a well-balanced manner. The selected compounds will serve as a training set for QSPR model development.

- Phase two: Determination of rate constants. The objective is to determine the ozone and hydroxyl radical rate constants of the selected micropollutants (from phase one) in bench-scale experiments.
- Phase three: QSPR model development and validation. The objective is to establish QSPR models for rate constant prediction by using the previously experimentally determined rate constants as the training set (from phase two), and to validate the models using existing data collected from the literature.
- Phase four: Model application in natural water. The objective is to explore the application of the developed QSPR models (from phase three) by using the predicted rate constants together with R_{ct} and/or $R_{OH,UV}$ models to predict micropollutant removals in natural water.

1.4 Thesis Structure

The thesis consists of eight Chapters that were written in journal article format. Four out of five key Chapters (Chapters 3–6) are based on papers that have been published or are ready for submission to peer-reviewed journals. Chapter 2 will be edited further and will be submitted after the completion of the thesis. The structure of the thesis is shown in Figure 1.1.

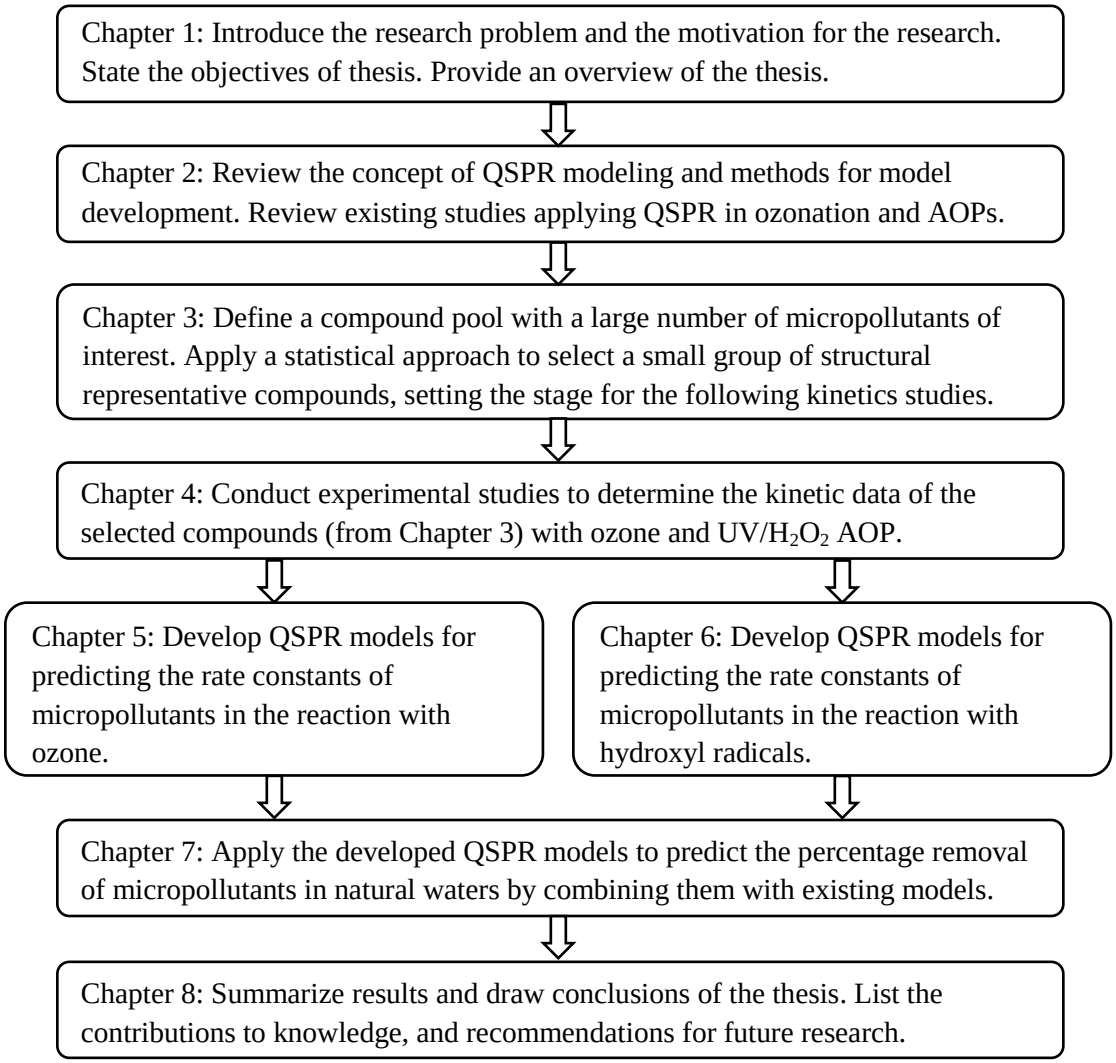


Figure 1.1 Thesis structure

Chapter 2

Quantitative Structure-Property Relationships (QSPR) Applications in Modeling the Properties of Micropollutants in Ozonation and Advanced Oxidation Processes (AOPs): A Review

This Chapter is based on a review paper which will be submitted to a journal for publication.

This Chapter is a literature review of QSPR applications in water treatment studies, specifically to predict the rate constant of micropollutants in ozonation and advanced oxidation processes. This Chapter mainly consists of two parts: (1) the key elements of QSPR model development are reviewed, including selection of training set, selection of molecular descriptors, statistical methods for modeling, model evaluation, model validation, and applicability domain; (2) the QSPR studies in modeling the rate constant of micropollutants in ozonation and advanced oxidation processes are reviewed. In addition, knowledge gaps and research needs are identified.

Outline: QSPR models have shown great predictive power for modeling environmental processes, drug design, and predicting the physico-chemical and biological properties of compounds. However, applications of the QSPR approach in water treatment are rare but increasing. QSPR can be used as a predictive tool to assess the removal of numerous contaminants during water treatment processes, especially those emerging contaminants where the experimental data are currently unavailable. In this review, first of all, the general scheme of a QSPR model is introduced and the main components of QSPR modeling are identified, namely the selection of training set, selection of molecular descriptors, statistical methods for modeling, model evaluation and validation, and applicability domain. Following, the commonly used statistical methods of each main component are reviewed. The existing QSPR studies in modeling the rate constant of micropollutants in ozonation and advanced oxidation processes are then reviewed. Finally, discussions on knowledge gaps and research needs are presented.

Keywords: EDCs, PPCPs, ozone, hydroxyl radical, rate constant

2.1 Introduction

Concerns about the occurrence of organic micropollutants in source waters for drinking water supply are increasing (Fent *et al.,* 2006; Heberer 2002). Diverse groups of these micropollutants including pharmaceuticals and personal care products (PPCPs) (e.g., antibiotics, anticonvulsants, contrast media agents, and sunscreen agents), endocrine disrupting chemicals (EDCs) (e.g. natural and synthetic estrogens), insecticides, herbicide, and many others have been detected at very low concentrations (ng/L – µg/L). These micropollutants may enter the aquatic environment via agricultural and urban runoff, landfill leachates, municipal sewage, industrial effluent, waste disposal, etc., and will eventually reach drinking water supplies. Hence, some have also been detected in finished drinking water (Benotti *et al.*, 2009; Snyder 2008; Huerta-Fontela *et al.*, 2011). Thus, there is a growing interest in understanding the removal efficiency of micropollutants during drinking water treatment processes.

Recent studies have shown that advanced technologies such as ozonation, advanced oxidation processes (AOPs), adsorption on activated carbon, reverse osmosis, and nanofiltration were effective in removing most micropollutants from drinking water (Westerhoff *et al.,* 2005; Snyder *et al.,* 2007). To assess the removal efficiency of micropollutants during these technologies, it is convenient and cost-effective to develop quantitative structure-property relationships (QSPR) models and apply them to micropollutants for which experimental studies have not been performed (Eriksson *et al.,* 2003; Eriksson and Johansson 1996). QSPRs have been widely used in the pharmaceutical industry for drug design, toxicity prediction, and regulatory decisions (e.g., US EPA uses QSPR predicted values for some regulatory purposes (Cronin *et al.,* 2003)). QSPR can also be used as a modeling tool to correlate the physico-chemical characteristics of micropollutants and their properties (e.g. reaction rate constants, removal, adsorption, rejection, etc.) in water treatment processes, thus providing improved knowledge on removal mechanisms for organic compounds in treatment processes.

QSPR applications in drinking water treatment studies are increasing. A number of studies have been published focusing on predicting the reaction rate constants of organic compounds in oxidation processes such as ozonation and advanced oxidation processes (Jiang *et al.*, 2010; Kusic *et al.*, 2009), the equilibrium adsorption constants on activated carbon (Metivier-Pignon *et al.,* 2007), and the rejection during membrane filtration (Kusic *et al.*, 2009; Metivier-Pignon *et al.,* 2007; Yangali-Quintanilla *et al.,* 2010). However, a comprehensive review on QSPR applications in water treatment

is not available. The scope of the following review is limited to oxidation processes in water treatment such as ozonation and AOPs as this is relevant to this thesis.

Oxidation processes such as ozonation and AOPs are effective technologies in degrading micropollutants from drinking water (von Gunten 2003; Ternes *et al.*, 2002; 2003; Westerhoff *et al.*, 2005; Snyder *et al.*, 2007). However, during ozonation and AOPs micropollutants are not completely mineralized. Instead, micropollutants are transformed into a multitude of degradation by-products, and the toxicity of most by-products is unknown. During ozonation, oxidation occurs via molecular ozone and hydroxyl radicals, which are produced through ozone decomposition in natural waters. Processes which involve the formation of highly reactive hydroxyl radicals are generally referred to as AOPs (e.g., O_3/H_2O_2, UV/H_2O_2). Oxidation efficiencies of micropollutants are characterized by chemical reaction kinetics where the reactivity of compounds during ozonation and AOPs can be measured by their reaction rate constants with molecular ozone and with hydroxyl radicals. Generally, rate constants are experimentally determined in pure water under laboratory conditions. In natural waters, it is impossible to assess the removal efficiency using rate constants alone. Models such as R_{ct} model for ozone (Elovitz and von Gunten 1999) and $R_{OH,UV}$ model for UV/H_2O_2 (Rosenfeldt and Linden 2007) have been developed to describe the removal efficiency of contaminants in natural waters incorporating standard reaction rate constants. These reaction rate constants are available for many organic contaminants (von Gunten 2003; Buxton *et al.*, 1988; NIST 2002) but they are still limited for emerging micropollutants. Therefore, QSPR models have been developed to correlate the compound structure to its rate constant, and then used to predict the reaction rate constants of new compounds (e.g., Kusic *et al.*, 2009).

The objective of this review is to provide an in-depth overview of QSPR applications for predicting rate constants of micropollutants during ozonation and AOPs. Elements of QSPR model development are presented first, and followed by a critical review of QSPR applications in ozonation and AOPs. Discussions on research needs and suggestions for QSPR applications in water treatment are also provided.

2.2 QSPR Model Development

2.2.1 General Approach of QSPR

QSPR models are mathematical relationships between the physico-chemical characteristics of compounds and their properties. The fundamental assumption of QSPR is that properties of a

compound can be related to its structural and physico-chemical features. A general QSPR modeling scheme is shown in Figure 2.1, QSPR model development starts with a compound pool which includes all the compounds of interest. A group of structural similar compounds (e.g., aromatic compounds) is usually defined as the compound pool. However, the applicability of those models is therefore limited to compounds with similar characteristics. It is still a challenge to include a large number of structural diverse compounds in a single QSPR model, as different mechanisms can be dominant for different sub-groups of compounds in the compound pool. Next, the compound pool is split into the training set and validation set. The molecular descriptors are then selected and calculated by software or searched in databases. Molecular descriptors with clear physico-chemical meanings which may be able to explain the underlying mechanisms of the property to be predicted are preferred. The properties to be predicted are usually determined by laboratory analysis for the training and validation set compounds. The QSPR models can then be developed using various modeling techniques which find mathematical relationships between the molecular descriptors and the properties to be predicted. Once validated, the model can be used for prediction of untested compounds. Without the proper validation, the predictive power of the developed QSPR model remains unknown. In addition, the applicability domain of the QSPR model should be defined. Predictions for compounds outside of the domain cannot be used without great caution.

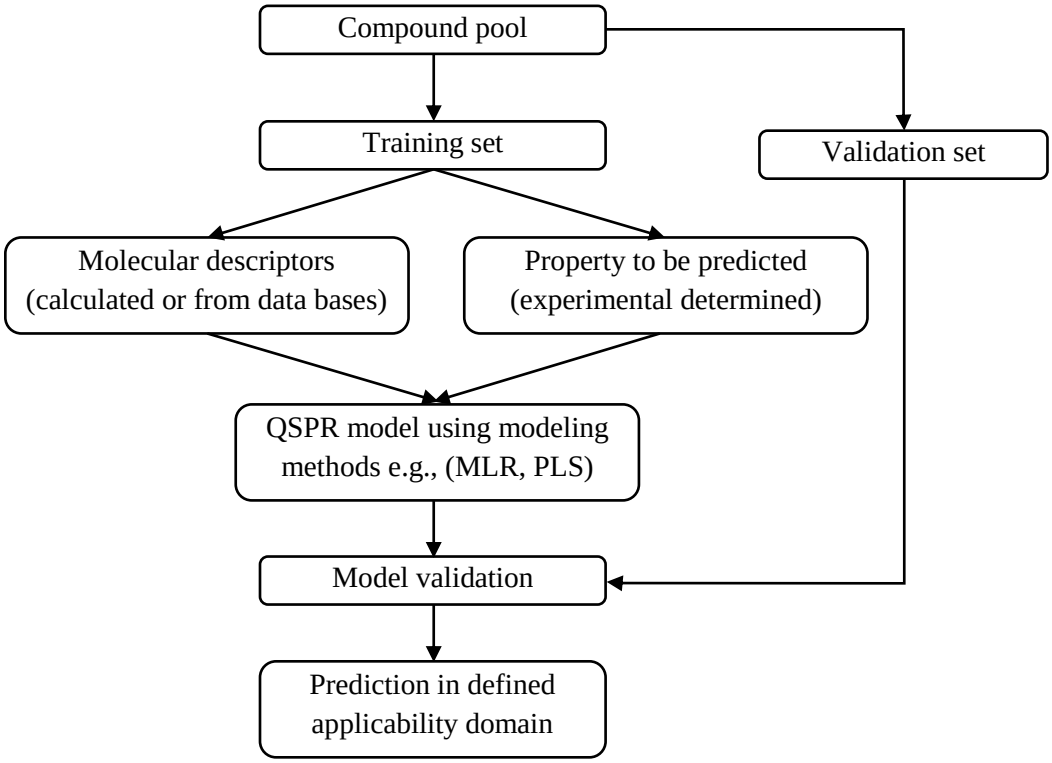


Figure 2.1 The general processes of QSPR model development

Several key elements are essential in the process of QSPR model development, including selection of training set, selection of molecular descriptors, statistical methods for model development, model validation, and determination of the applicability domain. The quality and predictive power of QSPR models depend on the proper applications of all the elements. As shown in Table 2.1, various statistical methods have been used for each element of QSPR modeling, and these methods are discussed briefly in the following sections.

Table 2.1 Statistical methods reviewed for each element of QSPR modeling

QSPR Modeling Element	Methods Reviewed
Training set selection	Random selection, sorted property sampling, *k*-means clustering, statistical molecular design
Descriptor selection	Subjective selection based on mechanistic knowledge, statistical criteria of correlations, statistical test and diagnostics, forward/backward/stepwise algorithm, variable importance in the projection, genetic algorithm.
Modeling techniques	Multiple linear regression, partial least square regression, principal

	component regression, artificial neural networks
Model validation	internal cross-validation, external validation
Applicability domain	Range based method, distance based method

2.2.2 Selection of Training Set

The training set is the compound set which is used to develop QSPR models, and the validation set (i.e., test set) is the external compound set used to test the predictive power of the developed models. QSPR models are built on the common features of the training set compounds, and the models use these features to predict the property of unknown, new compounds. Therefore, a new compound which has very little in common with the training set compounds will unlikely be predicted very well (Guha and Jurs 2005). The representativeness of the training set has a direct impact on the predictive accuracy and confidence for unknown compounds. The selection of a suitable training set is therefore an important step in QSPR analysis since the resulting model depends on the data quality of the training set and the applicability domain of the model is defined by the size and the diversity of the training set (Leonard and Roy 2006). QSPR model will likely fail to predict chemicals which are outside their applicability domain even if the model fits the training set perfectly.

There are several possible approaches for the selection of training set (and validation set). One approach is random selection where the available compound pool is randomly divided into training set and validation set (Oberg 2005). Another approach is sorted property sampling which is based on ranking of property to be predicted. In this method compounds are sorted according to the magnitude of the property to be predicted, and for example every other compound is selected for the training set, and the remaining compounds will be used as the validation set (Leonard and Roy 2006). These methods are simple and straight forward. The weakness of these methods is that they cannot guarantee that the selected training and validation set compounds represent the entire descriptor space of the original dataset (Golbraikh and Tropsha 2002a).

As an alternative, certain statistical techniques can be applied for compound selection to ensure the representativeness of the training set compounds. Cluster analysis is a group of statistical methods which assign compounds to clusters so that similar compounds are grouped together. Among many clustering methods, the *k*-means clustering method is commonly used in QSPR studies for training set selection (Burden *et al.*, 2000). This method classifies all compounds into *k* sub-groups (clusters) so to minimize the within-cluster sum of squares. Compounds within each cluster are then split into

training set and validation set. In this method, all chemical classes will be well represented in training and validation set. *K*-means clustering shows better result than random selection in many ways and has been recommended as a reliable method (Leonard and Roy 2006). However, it can be difficult to determine the number of clusters and an inappropriate choice of *k* may yield poor results. In addition, the performance of the method depends on the initial partition, and it is difficult to compare the quality of clusters produced by different initial partitions.

Statistical molecular design is one of the most commonly used method for training set selection. This approach differs from the above methods that only a small number of representative compounds will be selected from the pool. This approach is particular useful when the data of property to be predicted are not available yet, which is often the case when dealing with emerging contaminants. In this method representative training set compounds are selected by experimental design methods such as fractional factorial design (Wold *et al.*, 2004), D-optimal design (De Aguiar *et al.*, 1995), D-optimal onion design (Olsson *et al.*, 2004). The initial data matrix containing all compounds in the pool and their molecular descriptors are first analyzed by principal component analysis (PCA). As a result, a few informative latent variables, principal components (PCs) are derived to explain the main variation of the original data matrix. The obtained PCs are limited in number and mathematically independent, therefore ideal for experimental design. Representative training set compounds are then selected by applying experimental design methods (e.g., Knekta *et al.*, 2004; Papa *et al.*, 2007). This method results in a small number of informative and representative compounds, in which all major structural and chemical characteristics are well represented in a well-balanced manner (Eriksson *et al.*, 2003; Eriksson *et al.*, 2006).

2.2.3 Selection of Molecular Descriptors

Molecular descriptors are numerical values that characterize the properties of molecules such as physico-chemical properties, and structural features. Molecular descriptors can be determined experimentally or they can be calculated by software. Experimentally determined physico-chemical descriptors have historically been widely used; however, their availability is restricted because their measurement is time-consuming and expensive, and they are usually not available for many emerging contaminants. Therefore, the application of calculated descriptors is readily increasing with the help of modern computational techniques and chemistry software packages. Studies on the comparison between physico-chemical and calculated molecular descriptors have shown that they contain similar

information and calculated descriptors are suitable to use for developing models (e.g. Andersson *et al.*, 2000). A good collection and review of molecular descriptors can be found in Todeschini and Consonni (2000). Nowadays, computer programs can calculate over a thousand descriptors which cover a wide variety of descriptor classes. Several most commonly used software packages are DRAGON (Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy), HyperChem (Hypercube, Inc.), ChemOffice (ChembridgeSoft), etc.

The selection of descriptor variables from the many available molecular descriptors is a crucial step in QSPR model development (Andersen and Bro 2010). It is always desirable to build an adequate model with as few variables as possible, and those descriptors with clear physico-chemical meanings are preferred. Excluding redundant, irrelevant variables may not only improve the statistical properties of the model especially its predictive ability, but also make the model interpretation easier. In addition, including highly correlated descriptors violates the underlying assumptions of some modeling techniques (e.g. multiple linear regression). In such a case, the model will be ill-conditioned and the calculated regression coefficients will be unstable and uninterpretable, for example, coefficients with the wrong sign may be found or the coefficients are much larger than expected (Eriksson *et al.*, 2003).

Researchers can either start with as many descriptors as possible or they can start with a smaller set of preselected descriptors considered to be important based on available mechanistic knowledge. Either way, when developing a model a small set of relevant descriptors will be selected by statistical techniques from the initial descriptor sets while unrelevant descriptors will be eliminated. Prior to the variable selection, constant descriptors should be removed.

When an initial, statistically valid model is developed, some model parameters or diagnostic test can be applied to test the significance of the regression coefficients or loadings of the descriptors. It is possible to improve the model by removing descriptors with relatively low loadings or low standard regression coefficients. For example, in multiple linear regression, non-significant variables can be identified by using the student *t*-test or the associated *p*-value. Some modeling techniques are combined with a variable selection feature, such as forward, backward or stepwise algorithms. For example in forward multiple linear regression, the modeling process starts without any descriptors in the model, then the descriptors are tested one by one, and individual variables are added to the model if they are statistically significant. The procedure terminates when no variable meets the inclusion criterion, or when the available improvement falls below some critical value (Andre *et al.*, 2003).

Variable importance in the projection (*VIP*) is a measure of how much a variable contributes to both the dependent variable (i.e., property to be predicted) and independent variables (i.e., descriptors) and can be used for descriptor selection in projection method such as principal component analysis (PCA) and partial least squares regression (PLS) (Wold *et al.,* 2001).

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} W_{jf}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}} \qquad \textbf{(2.1)}$$

Where W_{jf} is the weight value for descriptor j in component (latent variable) f, SSY_f is the sum of squares of explained variance for the component f and J the number of descriptors, SSY_{total} is the total sum of squares of the dependent variables, and F is the total number of components. A variable with a *VIP* value smaller than one indicates a non-important variable. However, it is not as simple as removing all variables below one since useful information may be excluded. It is therefore recommended to remove a few variables with the lowest *VIP* values and check if the model is improved or not. This approach is repeated until no further improvements can be achieved (Andre *et al.,* 2003).

A genetic algorithm (GA) is an optimization algorithm which utilizes the concepts of the Darwinian evolution of species in the biological world (Leardi *et al.,* 1992). One application of GA in QSPR modeling is to find optimal subsets of descriptors that can be used to build predictive models. GA is a very effective tool with many advantages compared to other variable selection methods (Xu and Zhang 2001). The general approach of GA is to create different subsets of descriptors and evaluate their performance. The first step of the GA is to initialize the first generation of descriptor subsets and corresponding models. A number of descriptor subsets of similar size are randomly generated and each descriptor subset is then used to build a model (e.g. by multiple linear regression). The models are then ranked based on the fitness of individual compounds. Best models are selected as "parents" for reproduction of the next generation. A next generation is formed by different kinds of combination of randomly selected subset descriptors such as reproduction, mutation and crossover. The least-fit models are replaced by newly generated models. Therefore the average fitness of the next generation models has increased because only the best from the previous generation are selected for reproduction. The whole process is repeated until no more improvement is found or a fixed number of generations (e.g., 1000) are reached. At the end, the top ranked descriptor subsets can be used as the optimal subset of descriptors. The combination of GA and multiple linear regression

analysis has been used for prediction of rate constants for hydroxyl radical degradation of aromatic pollutants in a water matrix (Kusic *et al.,* 2009).

2.2.4 Statistical Methods for Model Development

The most common statistical methods used to develop QSPR models are linear methods such as multiple linear regression (MLR), principal component regression (PCR), and partial least squares regression (PLS). MLR is usually preferred because it produces apparently easily interpretable models. However, MLR cannot analyze data with correlated descriptors, and is unable to handle multiple responses in the same model (Box and Draper 1987). In contrast, multivariate techniques such as partial least squares regression (PLS) can handle collinear variables and can model several responses simultaneously (Eriksson and Johansson 1996). Artificial neural networks (ANN) method is one of the most used non-linear modeling tools. It can model complex relationships between inputs and outputs or it finds patterns in data sets (Yangali-Quintanilla *et al.,* 2009).

2.2.4.1 Multiple Linear Regression (MLR)

MLR is among the most widely used modeling methods in QSPR studies. MLR models a dependent variable (property to be predicted), *y*, as a linear combination of *p* independent variables (molecular descriptors) by determining the regression coefficients to each molecular descriptor. The coefficients are chosen to minimize the squares of the errors between the predicted and the observed property.

$$y = b + b_1 x_1 + b_2 x_2 ... + b_P x_P + e \qquad \textbf{(2.2)}$$

Where b_1, b_2, … b_p are regression coefficients and b is the constant, x_1, x_2, … x_p are molecular descriptors, y is the property to be predicted, and e represents the residuals. The equation can be written in the matrix form:

$$Y = XB + E \qquad \textbf{(2.3)}$$

Where Y is the matrix of property, X is the matrix of molecular descriptors, B is the matrix of regression coefficients, and E represents the matrix of residuals.

MLR assumes that the relationship between variables is linear and that the predictor variables are mathematically independent (orthogonal). In practice, the linear assumption can virtually never be

confirmed. Researches may have to consider either transforming the variables or applying non-linear models if necessary. If the data set shows multicollinearity among the molecular descriptors, the model will be ill-conditioned and the calculated regression coefficients will be unstable and uninterpretable (Eriksson *et al.,* 2003). Tolerance or variance inflation factor (*VIF*) can be used to detect the presence of multicollinearity in a model (Roy and Roy 2009).

$$tolerance = 1 - R_j^2 \quad \textbf{(2.4)}$$

$$VIF = \frac{1}{tolerance} = \frac{1}{1 - R_j^2} \quad \textbf{(2.5)}$$

Where R_j^2 is the coefficient of determination of a regression of descriptor *j* on all the other descriptors. If the tolerance value is less than a preset cut-off value (e.g., 0.1) or the *VIF* is higher than a cut-off value (e.g., 10), a multicollinearity problem exist in the descriptor set. MLR is satisfactory applied in QSPR studies if the main problem of the multicollinearity among variables is solved. Another limitation is that MLR requires a higher number of input data than the number of predictor variables, which refers to a large compounds-to-descriptors ratio in QSPR modeling. It has been recommended that the ratio should be at least 5 (Topliss and Edwards 1979).

2.2.4.2 Principle Component Regression (PCR)

Another regression-based method is PCR. In PCR, a principal component analysis (PCA) is first conducted to evaluate the original data matrix (descriptors) and a few principal components (PCs) are extracted. The PCs are orthogonal (mathematical independent) and able to explain most of the main variation in the original data matrix. Eriksson *et al.,* (2006) identified PCA as the most suitable technique for variables reduction and generation of orthogonal latent variables. A reduced set of variables such as the generated PCs is much easier to analyze and interpret. As a common procedure to avoid the influence of the unit of variables, data for PCA are usually pre-processed by means of mean-centering and scaling to unit variance. As a result, the mean values of all the variables for each observation are equal to zero (Eriksson *et al.,* 2006). PCA then decomposes the *X*-matrix (descriptors) into the product of two matrices, the score matrix *T* and the loading matrix *P'*, plus a residuals matrix *E*. The product of score matrix and loading matrix *TP'* is used to model the initial data matrix *X* (Wold *et al.,* 1987). The number of principal components (i.e., the number of columns in score matrix and the number of rows in loading matrix), is determined by cross-validation.

$$X = TP' + E \quad \textbf{(2.6)}$$

In the next step, the first few significant principal components are used as predictor variables in a MLR with the dependent variables. Because the principal components are mathematically independent, multicollinearity among original variables is no longer a problem as it is in MLR.

$$y = b_1 PC_1 + b_2 PC_2 + ... + b_q PC_q + b \quad \textbf{(2.7)}$$

Where y is the property to be predicted, b_1, b_2, … b_q are regression coefficients and b is the constant, PC_1, PC_2, … PC_q are principal components extracted by PCA, and q is the number of significant principal component. The regression model can be found by using the usual MLR algorithm, and same statistical parameters as used in MLR can be applied to assess the quality of models.

2.2.4.3 Partial Least Squares Regression (PLS)

PLS is a recently developed regression method, which can be viewed as a generalization of multiple linear regression (Eriksson *et al.*, 2006). PLS is a projection method which finds new variables (latent variables) which are linear combinations of the original variables and orthogonal, and also well correlated to the dependent variable(s). The dependent variable(s) can be a single property (e.g. rate constant) or multiple responses (e.g. toxicity determined in several testing systems). PLS is similar to PCA, but PCA works with an X matrix while PLS works with two matrices, X and Y, respectively. In other words, the difference between PCR and PLS is that PLS finds the latent variables and the regression coefficients at the same time. PLS projects the matrix X into a lower-dimensional hyper-plane, and several latent variables are introduced to describe the positions of the projected data. Latent variables are used to correlate the values of Y and X. If the response matrix Y contains multiple responses, Y will be summarized by another projected lower-dimensional hyper-plane. The number of significant dimensions in PLS is determined using cross-validation. Models will be set up between latent variables of X and Y.

$$Y = TB + E \quad \textbf{(2.8)}$$

Where T is the matrix containing the scores of the extracted latent variables, ($T = XF$, where F is the matrix of the loadings of the original variables in the principal component scores, and X is the matrix of mean-centered molecular descriptors) B is the matrix of the PLS regression coefficients, and

E is the unexplained variance in *Y*. The PLS regression model can be presented in terms of the original molecular descriptors.

$$Y = XQ + E \quad \textbf{(2.9)}$$

Where $Q = FB$ is the matrix of regression coefficient for original descriptors. In addition, the original mean-centered descriptors can also be expressed by the latent variables.

$$X = TP + K \quad \textbf{(2.10)}$$

Where *P* is the loading matrix, and *K* is the unexplained part of *X*. A detailed tutorial of PLS method with some good examples can be found in the literature (Wold *et al.*, 2001). Unlike MLR, PLS works well when data are strongly correlated since the extracted latent variables are orthogonal and limited in number.

2.2.4.4 Artificial Neural Networks (ANN)

As one of the most commonly used non-linear method for QSPR modeling, ANN is a prediction method inspired by the biological nervous systems. ANN contains at least three layers: input, hidden and output layers (Roy and Roy 2009). The multi-layer perceptron (MLP) is one of the most commonly used models of neural networks. Basically a MLP consists of a network of several neurons assembled in layers, and the neurons of a specific layer are generally all connected to the neurons of the following layer. Each neuron is able to linearly combine its input values to an output by means of a certain transfer function. Each input of the neuron has an adaptive weight specifying the importance of the input. The weights are adjusted during a supervised training phase, and the differences between estimated and expected results (output) are calculated. The optimized weights are obtained by minimization of the error between estimated and expected values during the training phase. MLP neural networks contain a single input layer, which are formed by the molecular descriptors, one or more hidden layers (usually only one hidden layer to avoid over-fitting), which process the descriptors into internal representations and an output layer utilizing the internal representation to produce the final prediction. The advantage of ANN is that this method is adaptive and can learn from the data without any human interference. The drawback is the lack of model transparency. ANN is labeled as a “black box” approach to modeling the relationships between structure and property,

because it provides little information on the relative influence of the descriptors in the predictive process, making it difficult to understand the underlying mechanisms (Yang *et al.*, 2005).

2.2.5 Model Evaluation

A few parameters such as the squared multiple correlation coefficient (R^2), adjusted R^2, and variance ratio (F) are used to judge the statistical qualities of the equations. R^2 is defined as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad \textbf{(2.11)}$$

Where y_i and $\hat{y}_i$ are observed and calculated property values, respectively, while $\bar{y}$ is the mean of the observed property. R^2 is a measure of how much of the variation in the data set is explained by the regression model. R^2 ranges from 0 to 1, the closer the value of R^2 to 1, the better the variations among the observed data are explaind by the regression model. The value of R^2 depends on the number of compounds (n) and number of descriptors (p), therefore another statistical parameter can be used, called adjusted R^2 (R^2_{adj}).

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \quad \textbf{(2.12)}$$

Adjusted R^2 shows similar information as R^2 but adjusted by the number of compounds and number of descriptors. Because of the inflation of R^2 with the number of independent variables, adjusted R^2 is a more appropriate and meaningful parameter to compare models with different numbers of independent variables.

The dispersion of the observed dependent variable about the regression line (surface) can be assessed by the value of the standard error of estimate s. Larger value of s means worse statistical fit of the model.

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}} \quad \textbf{(2.13)}$$

The statistical significance of a regression equation can be assessed by means of the Fisher (F) statistic. A regression model is considered to be statistically significant if the F value is greater than a

tabulated value for the chosen level of significance (typically 95% level) and the corresponding degrees of freedom.

$$F = \frac{\sum (y_i - \bar{y})^2 / p}{\sum (y_i - \hat{y}_i)^2 / (n - p - 1)} \quad \textbf{(2.14)}$$

2.2.6 QSPR Model Validation

In order to assess the model and test the predictive power, QSPR models must be validated, preferably by an external data set (validation set) which was not included in model calibration (Eriksson *et al.*, 2006). Internal validation and external validation are two commonly applied methods for model validation.

When the training set is sufficiently large, the internal validation process also known as cross-validation, is applicable. It is used to assess the predictivity in addition to the robustness of the model (stability of QSPR model parameters). Cross-validation is performed by using leave-one-out or leave-many-out procedure. Leave-one-out cross-validation tends to overestimate the predictive capacity and leave-many-out is preferred (Golbraikh and Tropsha 2002b). The process can be describe as follows: Leave one or more compounds out of the training set and use the remaining compounds to develop the model; use the model to predict the left-out compound(s); calculate the predictive residual error sum of squares (*PRESS*) value; repeat the processes above until all compounds have been left out once and only once; calculated the overall *PRESS* and total sum of squares (*SST*), providing a cross-validated Q^2.

$$Q^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum (y_i - \hat{y}_{i/i})^2}{\sum (y_i - \bar{y})^2} \quad \textbf{(2.15)}$$

Where y_i is observed dependent variable, $\hat{y}_{i/i}$ is calculated dependent variable from a model developed without that data point, $\bar{y}$ is the mean value of training set compound. Many authors (e.g., Tropsha *et al.*, 2003) consider Q^2 greater than 0.5 as an indicator of the robust model robustness with sufficient predictive ability.

However, when the number of compounds in the training set is not sufficient enough, especially when the training set is selected by experimental design, then the internal validation is unlikely to

provide a reliable measure of the model's predictive power. In addition, a recent study indicated that even high values of Q^2 from an internal validation may not be a suitable indicator (Golbraikh and Tropsha 2002b). External validation is therefore the only way to establish reliable QSPR models (Golbraikh *et al.,* 2003; Cronin and Schultz 2003; Hawkins *et al.,* 2003). In external validation, the predicted and observed values of a sufficiently large external validation set of compounds that were not used in the model development are compared. The predictive R^2 (R^2_{pred}) which measures the truly predictive capacity of a model for new compounds (validation set), can be calculated using the following equation:

$$R^2_{pred} = 1 - \frac{\sum \left(y_{validation} - \hat{y}_{validation}\right)^2}{\sum \left(y_{validation} - \bar{y}_{training}\right)^2} \quad \textbf{(2.16)}$$

Where: $\hat{y}_{validation}$ and $y_{validation}$ are predicted and observed values of validation set compounds; $\bar{y}_{training}$ is the mean value of training set compounds.

2.2.7 Applicability Domain

QSPR models are developed using a limited number of training set compounds with limited structural characteristics. It is unlikely these the models can be applied to every chemical. The applicability domain defines the scope of a QSPR model in which it is appropriate to make predictions for new compounds. Predictions should be made within this applicability domain by interpolation and not by extrapolation. QSPR model will likely fail to predict compounds outside the applicability domain. The multivariate space occupied by the training set compounds is the basis for defining the applicability domain. The simplest method is range based; for example, use the ranges of the descriptors which define an *n*-dimensional hyper-rectangle. However, such approach may cover lots of empty space if data are not uniformly distributed (Jaworska *et al.,* 2005). Distance based methods are commonly used (Eriksson *et al.,* 2003; Tropsha *et al.,* 2003). For example, leverage is used to measure the distance of a compound to the centroid of the model. A validation compound with a high leverage (i.e., structurally distant from the training compounds) will likely not be predicted reliably, as a result of substantial extrapolation of the model. A leverage value of 3 is often taken as a critical value which represents 3 standard deviations from the mean (Eriksson *et al.,* 2003). To visualize the

applicability domain of a QSPR model, Williams plot (standardized cross-validated residuals vs. leverage values) can be used (Kusic *et al.*, 2009; Wang *et al.*, 2009).

2.3 QSPR Models in Ozonation and AOPs

The QSPR approach has been applied in water treatment to predict the kinetics of compounds in their reaction with molecular ozone and hydroxyl radicals and only rarely to assess the removal efficiency of these processes for organic contaminants in natural waters. Most of the studies focused on predicting second-order rate constants (e.g., Kusic *et al.*, 2009). However, it is worth mentioning that Lei and Snyder (2007) applied the QSPR technique to predict percentage removals of organic contaminants by ozone and free chlorine in natural water. Their QSPR model for ozonation provides a useful prescreening tool to preliminary evaluate removal of organic contaminants. However, the degradation of organic compounds depends not only on kinetics but also on the water matrix, pH, flow rate, ozone dose, etc., this removal model is therefore case-specific, and it is not possible to apply it elsewhere. Only a small number of QSPR studies have been published, but various modeling approaches and techniques have been applied such as the linear free energy relationship approach (Haag and Yao 1992), the group contribution method (Minakata *et al.*, 2009), as well as linear and non-linear regression methods. In addition, a number of molecular descriptors have been identified as suitable for descripting the physico-chemical characteristics of compounds relevant for compound reactivity during ozonation and AOPs. In the following sections, the existing QSPR studies on reactivity of compound in ozonation and AOPs are reviewed. First of all, the concept of the rate constant is introduced, followed by a discussion of relevant molecular descriptors, and the modeling techniques used in oxidation studies are reviewed at the end of this section.

2.3.1 Properties to be Predicted: Rate Constant

Before applying the QSPR approach, the endpoint of the modeling needs to be specified. The reaction rate constant represents a suitable endpoint in the process of correlating the reactivity of a compound to its structure and has been used as such. The reactivity of compounds varies with their structures; therefore pH can play a role for dissociating compounds.

For a non-dissociating compound, the reaction rate constant of compound *P* with oxidant ozone and hydroxyl radicals are determined by the equations below, respectively. Where $k_{O3\text{-}P}$ is the second-order rate constant for the reaction with ozone, and $k_{OH\text{-}P}$ is the second-order rate constant for the reaction with hydroxyl radicals.

$$\ln\left(\frac{[P]_t}{[P]_0}\right) = -k_{O3-P}\int [O_3]dt \qquad \textbf{(2.17)}$$

$$\ln\left(\frac{[P]_t}{[P]_0}\right) = -k_{OH-P}\int [OH]dt \qquad \textbf{(2.18)}$$

For dissociating compound, the overall rate constant is pH dependent because the neutral and ionic species of the compound can have different reaction rates with the oxidants. At a certain pH, the rate constant is shown by the equation below (assume one ionic species), where $k_{app,p}$ is the apparent rate constant at certain pH; k_1, k_2 are the specific rate constants for neutral and ionic species, respectively; α_1, α_2 are the ionization fraction for neutral and ionic species, respectively (α_1 and α_2 can be calculated from the dissociation constant p*K*a and the specific pH of the solution).

$$\ln\left(\frac{[P]_t}{[P]_0}\right) = -k_{app,P}\int [O_3]dt = -(k_1\alpha_1 + k_2\alpha_2)\int [O_3]dt \qquad \textbf{(2.19)}$$

$$\ln\left(\frac{[P]_t}{[P]_0}\right) = -k_{app,P}\int [OH]dt = -(k_1\alpha_1 + k_2\alpha_2)\int [OH]dt \qquad \textbf{(2.20)}$$

QSPR models have been developed to correlate the absolute rate constant (Gurol and Nekouinaini 1984; Hoigné and Bader 1983b) and apparent rate constant around pH 7 (Jiang *et al.*, 2010; Hu *et al.*, 2000) with the structural descriptors of organic contaminants.

2.3.2 Molecular Descriptors Suitable for QSPR Modeling

To develop a significant correlation between the property to be predicted and chemical structure, it is crucial to employ appropriate descriptors, whether they are theoretical, empirical or experimental. Descriptors reflecting simple molecular properties that can provide insight into the property to be predicted are preferred. It is not the objective of this review to give a complete overview of all

possible descriptors, only those descriptors that have been widely used and are suitable for QSPR studies on oxidation processes are reviewed here.

Functional Groups or Substructures. Ozone attack is selective to compounds with double bonds, activated aromatic systems, and deprotonated amines (von Gunten 2003), therefore molecular descriptors describing the presence of these functional groups could potentially be used in QSPR studies, for example, number of double bonds, number of deprotonated amines.

Aromatic systems can be activated by electron donor substituents (e.g. -OH, $-OCH_3$, $-NH_2$), while they are being deactivated by electron withdrawing groups (e.g. -Cl, $-NO_2$, etc.). Thus the presence of a phenolic group (Ar-OH), methoxybenzene ($Ar-OCH_3$), or aminobenzene ($Ar-NH_2$) could be used as molecular descriptors for aromatic compounds. Another type of descriptor for aromatic systems is the Hammett constant. It is a measure of the electron withdrawing or donating abilities of the substituents on benzene (positive values for electron-withdrawing substituents and negative values for electron-donating substituents). Hammett constants have become the most common descriptors in predicting the effect of substituent on reactivity of aromatic system (Canonica and Tratnyek 2003). Extensive reviews of Hammett constants and related substituent properties are available in the literature (Brown and Okamoto 1958; Hansch *et al.*, 1991; Hansch and Leo 1995; Hansch and Gao 1997).

In addition, in a recent kinetics study on antibacterial compounds (Dodd *et al.*, 2006), some substructures reactive to ozone attack have been identified and the corresponding reaction rate constants were reported. These reactive substructures may potentially be used as indicator variables for modeling purpose.

Quantum Chemical Descriptors. Quantum chemistry provides a description of the electronic properties of molecules and their interactions. Quantum chemical descriptors can describe many aspects of molecular electronic properties such as atomic charge, molecular orbital energy, electron density, polarizability, and dipole moment, etc. An extensive review of quantum chemistry descriptors in QSPR studies has been published (Karelson *et al.*, 1996). Quantum chemical descriptors can be calculated by various software using semi-empirical methods.

Recent QSPR studies have employed quantum chemical descriptors alone or in combination with conventional descriptors (Kusic *et al.*, 2009). Quantum chemical descriptors such as energy of the highest occupied molecular orbital (HOMO), and energy of the lowest unoccupied molecular orbital

(LUMO) are becoming increasingly favored as molecular descriptors as they are related to the energy of oxidation (and reduction) reactions (Canonica and Tratnyek 2003). HOMO energy can also be described as a measure of the tendency that a molecule will be attacked by electrophiles, and LUMO energy as a measure of the tendency that a molecule will be attacked by nucleophiles. HOMO energy is related to the ionization potential, and LUMO energy is related to the electron affinity. The difference in energy between the HOMO and LUMO, i.e. the HOMO-LUMO gap is a measure of the stability of molecule. A large HOMO-LUMO gap suggests low reactivity in chemical reactions and high stability (Gramatica *et al.*, 2003). HOMO has been found important in modeling the hydroxyl radicals rate constant of aromatic pollutants in water (Kusic *et al.*, 2009) and in modeling the ozone rate constant of a few pesticides in water (Hu *et al.*, 2000).

Empirical descriptors. In addition, the unsaturation index which related to the presence of reactive functional groups has been used for QSPR modeling of the tropospheric degradation of ozone (Gramatica *et al.*, 2003). Polarizability, molecular weight and average molecular weight were also found important in predicting the hydroxyl radical rate constants of aromatic compounds in water (Kusic *et al.*, 2009).

2.3.3 QSPR Modeling Techniques Used in Oxidation Studies

Several modeling techniques have been applied to build QSPR models for oxidation processes, including linear methods such as Hammett-type linear free energy relationships, group contribution method, multiple linear regression, partial least squares regression and nonlinear artificial neural networks.

2.3.3.1 Hammett-type Linear Free Energy Relationships

The Hammett equation, originally developed by Hammett in 1937 (Hammett 1937), describes a linear free energy relationship for reaction rate constants of substituted benzene derivatives and benzoic acids. Hammett-type relationships where descriptor variables are substituent constants of various types have become the most common type of QSPR in predicting the effect of substituent on reactivity (Canonica and Tratnyek 2003). The general form of the Hammett equation may be written as

$$\log k_S = \log k_0 + \rho \cdot \sigma \qquad \textbf{(2.21)}$$

where k_S and k_0 are the rate constants for the substituted and the unsubstituted reference compounds (e.g. benzene), σ, the substituent constant, is a measure of the electron withdrawing or donating abilities of the substituent (positive values for electron-withdrawing substituents and negative for electron-donating substituents); ρ, the slope, is a measure of the sensitivity of the reaction rate to substituent effects. The larger the slope, the more sensitive the reaction to the electronic effects of the substituents; If $\rho = 0$, the reaction is insensitive to electronic effects. Electrophilic reactions such as ozonation and hydroxyl radical reactions are indicated by a negative Hammett slope, which indicates that the reaction is favored by electron-donating groups ($\sigma > 0$) and disfavored by electron-withdrawing groups ($\sigma < 0$). This equation describes a linear correlation between the logarithm of the reaction rate constants for substituted compounds and the constant for the corresponding substituents.

Hammett-type relationships have been proven to be useful in predicting the reaction rate constants of substituted aromatics with various oxidants, such as ozone (Gurol and Nekouinaini 1984; Hoigné and Bader 1983a; Benitez *et al.*, 2007) and hydroxyl radicals (Hansch and Gao 1997; Haag and Yao 1992; Peres *et al.*, 2010; Einschlag *et al.*, 2003; Zimbron and Reardon 2005) (as shown in Table 2.2). About half of the models show a good fit (i.e., high R^2 value) whereas the fit for the other half is relatively poor (i.e., low R^2 value). It is also worth to mention that the models with good fit using structural similar compounds and models with poor fit using structural diverse compounds. For example, although statistically significant, a relative poor model with low r^2 value was developed by Haag and Yao (1992) because of the structural diversity of the compound set or the high variability of the rate constants. Good QSPR models obtained for a set of structure similar compounds, for example, only one substituent varied at a fixed position (Hansch and Gao 1997). As shown in Table 2.2, all the equations indicate negative slope, the reaction constant ρ, which are expected for both ozonation and hydroxyl radical reactions where the reactions are favored by increased electron density at the aromatic ring. The magnitude of the slope is a measure of how susceptible a reaction is to the electronic characteristics of the substituent. The small absolute values of ρ for hydroxyl radical reactions (0.14 – 0.60) reflect the low sensitivity (selectivity) of hydroxyl radical reactions towards substituents at the aromatic ring; whereas the large absolute values for ozone (2.81 – 8.0) indicate that the reaction of ozone with aromatic compounds is highly selective.

Table 2.2 Hammett-type relationship for hydroxyl radical and ozone reactions in the aqueous phase

Correlation	R^2	n	Compound types	Reference
Reactions with hydroxyl radicals				
$\log(k_{OH}) = 9.829 - 0.318\Sigma\sigma$	0.595	25	X_n-C_6H_{6-n}	Haag and Yao 1992
$\log(k_{OH}) = 8.58 - 0.21\sigma^+$	0.88	12	X-C_6H_5	Hansch and Gao 1997
$\log(k_{OH}) = 8.70 - 0.27\sigma^+$	0.98	9	X-C_6H_5-COOH	Hansch and Gao 1997
$\log(k_{OH}) = 9.96 - 0.60\sigma$	0.986*	8	X_n-C_6H_{5-n}-NO_2	Einschlag *et al.*, 2003
$\log(k_{OH}) = 10.0 - 0.15\Sigma\sigma$	0.54	24	Cl_n-C_6H_{6-n}	Zimbron and Reardon 2005
$\log(k_{OH}) = 10.0 - 0.14\Sigma\sigma^+$	0.50	24	Cl_n-C_6H_{6-n}	Zimbron and Reardon 2005
$\log(k_{OH}) = 9.5 - 0.28\Sigma\sigma$	0.62*	10	X_n-C_6H_{4-n}-OH-COOH	Peres *et al.*, 2010
Reactions with ozone				
$\log(k_{O3}) = a - 8.0\sigma$	NA	9	$(CH_3)_n$-C_6H_{5-n}-OH	Gurol and Nekoulnalnl 1983
$\log(k_{O3}) = a - 3.1\sigma^+$	NA	7	X_n-C_6H_{6-n}	Hoigné and Bader 1983a
$\log(k_{O3}) = a - 2.81\sigma$	0.988	3	X_n- C_6H_{5-n}-$C_3H_7ON_2$	Benitez *et al.*, 2007
$\log(k_{O3}) = 8.9 - 2.4\Sigma\sigma^+$	0.96	13	X_n-C_6H_{5-n}-OH (anionic species)	Suarez *et al.*, 2007
$\log(k_{O3}) = 3.4 - 3.4\Sigma\sigma^+$	0.94	7	X_n-C_6H_{5-n}-OH (neutral species)	Suarez *et al.*, 2007

*R^2 was calculated from r values, a is constant which is not specified in the reference, σ is the Hammett's constant (Hansch *et al.*, 1991), and σ^+ is the modified Hammett's constant (Brown and Okamoto 1958).

The advantage of Hammett-type relationship is that σ values are additive for multiple substituents. Therefore, it is possible to correlate a variety of substituted aromatics by calculating substituent effects from a limited set of σ values. However, Hammett-type relationships are only applicable to substituted aromatics with known substituent constants; they cannot be applied to other non-aromatic compounds. Only a limited number of sigma constants is currently available, which makes it impossible to explore new compounds with more complex substituents. Overall, it seems impossible to apply Hammett-type relationship to determine the reactivity of structural diverse compounds beyond simple substituted aromatic compounds. In addition, none of the published studies validated the Hammett-type relationships with an external data set, nor did they determine an

applicability domain. This limits the applicability of these relationships to new compounds as the predictive ability of these models is unknown.

2.3.3.2 Group Contribution Method

The Group contribution method was originally developed to predict hydroxyl radical rate constants in the gaseous phase (Atkinson 1987; Atkinson 1988). In this group/fragment contribution methodology, the total estimated rate constant is the summation of all applicable reaction pathways, such as H-atom abstraction from aliphatic bonds, hydroxyl radical addition to olefinic and acetylenic bonds, and aromatic rings, hydroxyl radical reaction with nitrogen, sulfur and phosphorus atom-containing substructure, etc. A table of substituents (or groups) and the corresponding factors (coefficients) were given for each pathway (Kwok and Atkinson 1995; Atkinson 2000). The group contribution method has been proven to be robust and is widely used for predicting gaseous phase reaction rates. For example, it has been used as an estimation method in U.S. EPA software AOPWIN which can predict the atmospheric hydroxyl radical and ozone rate constants.

However, before applying the group contribution methods to aqueous phase rate constants, the differences in the reaction mechanisms between gaseous phase and aqueous phase need to be considered For example, the hydrogen bond and polarity of water molecules will play a role in the aqueous phase Monod and coworkers modified and applied Atkinson's group contribution methodology to predict the hydroxyl radical constants of aliphatic organic compounds in the aqueous phase (Monod *et al.*, 2005; Monod and Doussin 2008). Monod and coworkers focused on the oxygenated aliphatic compounds in which the H-atom abstraction is the dominant mechanism. Similarly to Atkinson's method, this method calculated the overall rate constant of a compound as a summation of the partial rate constants of each reactive site (elementary reaction). There were group rate constants which represented the reaction mechanisms and substituent factors which took into account the field and resonance effects of the neighboring groups (α-position). In addition, the next-nearest neighboring (β-position) effects were also considered by introducing *G* parameters. With the additional *G* parameters, a better agreement between calculated and experimental data was obtained (Monod and Doussin 2008). In this study, a group of 72 aliphatic compounds (including alkanes, alcohols, organic acids, bases and polyfunctional compounds containing at least two of these functions) which relevant to atmospheric chemistry, and 7 function groups were investigated. As a result, 60% of the estimated values were found within the range of 80% of the experimental values. The correlation (R^2) between the estimated and experimental logk_{OH} was 0.89. Compared to other

estimation methods (the correlations between the aqueous phase reactivity and the bond dissociation energy, the correlations between gas- and aqueous-phase reactivity, and the neural network) of the rate constant (k_{OH}) in aqueous phase, the author claimed that the group contribution method is the most easy to used method and gave the best performance (Monad and Doussin 2008). However, the model was not externally validated, and this method is only applicable to compounds where H-atom abstraction mechanism is dominant.

In a recent work by Minakata *et al.,* (2009), other mechanisms were also discussed such as OH addition to alkenes and aromatic compounds, OH interaction with nitrogen-, sulfur-, or phosphorus-atom-containing compounds. Therefore, the aqueous phase hydroxyl radical constants can be applied to compounds with a wide range of functional groups. The resulting group contribution model included 66 group rate constants and 80 group contribution factors, which characterize the effect of the chemical structure groups and the neighboring functional groups. In this study, 310 compounds with literature-reported k_{OH} values were used as training set and another 124 compounds were used to validate the model. As a result, the estimated values of 83% (257 compounds) of the training set compounds and 62% (77 compounds) of the validation set compounds were within 0.5-2 times of the experimental values. In addition, Minakata *et al.,* (2009) also applied this method to predict 11 emerging aromatic compounds and compared the predicted values with experimental ones, and found that the difference were in an acceptable range.

Overall the group contribution method is only applicable to a small group of compounds with specific structure, i.e., aliphatic organic compounds. The group contribution method is reasonably reliable when applied to compounds similar to those used as training set. However, extrapolation to chemical structures significantly different from those in the experimental database may result in significant estimation error.

2.3.3.3 Regression Methods

Multiple Linear Regression. MLR has been applied to model the rate constants of micropollutants in the reaction with ozone (Hu *et al.,* 2000) and hydroxyl radicals (Kusic *et al.,* 2009; Wang *et al.,* 2009) in the aqueous phase.

Hu *et al.,* (2000) determined the apparent rate constants (at pH 7.5) of 24 pesticides (including 4 phenolic-, 8 organonitrogen-, 8 phenoxyalkylacitic-, and 4 heterocyclic N-pesticides) in the reaction with ozone, and found a good correlation (R^2 = 0.84) between logarithm rate constants of all the

pesticides studied and their HOMO energy (one parameter QSPR model). Even better correlations with HOMO energy ($R^2 > 0.9$) can be found in separate groups except phenoxyalkylacitic pesticides. Phenoxyalkylacitic pesticides can be modeled accurately by a two-parameter QSPR model using absolute electronegativity and HOMO energy ($R^2 = 0.97$). However, without external validation, the predictive power of the model remains unknown.

A well designed QSPR study was developed for the aqueous-phase hydroxyl radical reaction rate constants of 55 phenols, alkanes and alcohols using stepwise MLR with a $R^2 = 0.905$ (Wang *et al.*, 2009). The model was internally validated ($Q^2 = 0.806$) by the leave-many-out technique and externally validated ($Q^2 = 0.922$) using an external validation set, and the applicability domain was also analyzed by a Williams plot. Four out of fifteen quantum chemical descriptors were found to be the governing descriptors using stepwise MLR; they were the HOMO energy, average net atomic charges on hydrogen atoms, molecular surface area, and dipole moment. The model obtained was applicable to phenols, alkanes and alcohols but not applicable to complex structures. A poor model ($R^2 = 0.365$) was developed when various classes of chemicals were included in the previous model (Wang *et al.*, 2009). Another satisfactory study for predicting the aqueous-phase hydroxyl radical rate constants was conducted by Kusic *et al.*, 2009. In that study, the QSPR models were developed with 78 aromatic compounds using MLR combined with a variable selection genetic algorithm (GA). The combination of GA-MLR approach was used to find the best few descriptors from a large number of original descriptors. As a result, the logarithm of rate constants was correlated to HOMO energy and several other descriptors relating to molecular polarizability.

Partial Least Squares Regression. In a recent paper (Jiang *et al.*, 2010), the rate constants of ozone with 39 aromatic compounds were determined, and a QSPR model ($R^2 = 0.791$, standard deviation = 0.126) was developed using PLS regression. Quantum chemical descriptors LUMO energy, the most positive partial charge on a hydrogen atom (qH^+), and thermodynamic descriptor Connolly molecular area were found important. The QSPR model showed that the main contribution to degradation was the Connolly molecular area. However, no external validation was applied, and the experimental data reported are questionable as the ozonation of aromatic compounds was found to be zero-order. In addition, the chemical structure was linked to the pseudo-rate constants which were ozone exposure dependent, but according to the information provided in the reference ozone concentration was not measured.

Artificial Neural Networks. An artificial neural network, the multi-layer perceptron (MLP), was used to relate the functional groups of the molecule and the rate constant of the molecule reacting with hydroxyl radicals in the aqueous phase (Dutot *et al.*, 2003). A large group of 209 compounds from the review of Buxton *et al.,* (1988) were used in this study, and the initial compound set was divided into three different groups for model development and evaluation of the prediction capability: the training, test and validation set. The training and test set were used to optimize the parameters of the neural regression, and the validation set was used to assess the performance of model prediction. The molecular descriptors used in this study were 17 functional groups. It was found that the standard error of prediction was 0.24 for $\log k_{OH}$ with a data range from 8.22 – 10.37 ($\log k_{OH}$). In a predicted vs. experimental data analysis using the validation set (not included in the model training), the R^2 was found to be significant ($R^2 = 0.81$) showing a good predictive power. The author claimed that ANN performs better than linear regression methods because of the parsimony of ANN.

2.4 Knowledge Gaps and Research Needs

Important elements of QSPR development were reviewed, including selection of the training set, selection of descriptors, various modeling techniques, model evaluation and validation, and applicability domain. Existing QSPR studies developing models for oxidation process in water treatment namely ozonation and AOPs were also reviewed. A number of studies are available for modeling reaction rates with hydroxyl radicals whereas there are much fewer for reactions with molecular ozone. Based on this review knowledge gap and research needs were identified as follows:

- Most studies used a group of structurally relative homogeneous compounds as the training set. The similarity of the compounds generally ensures fairly high predictive power of the developed QSPR models. However, the applicability of these QSPR models is limited to a small range of compounds which are structural similar to the training set compounds. For example, the model developed by Kusic *et al.,* (2009) is only applicable to aromatic compounds. It still remains a challenge to develop QSPR models which are widely applicable to many structural diverse compounds, mainly because various mechanisms may be involved in a large and diverse compound pool.
- The existing QSPR studies predicting the rate constant of compounds mostly focus on conventional contaminants with higher concentrations in aqueous phase compared to many of the emerging micropollutants. The interests in these micropollutants especially EDCs, and

PPCPs are increasing. Therefore predictive models applicable to these newer micropollutants are needed.

- Successful application of the QSPR approach in model development for water treatment processes involves several key elements as discussed above. Extensive statistical knowledge is needed and misuse or missing one or more key elements can lead to incorrect or poor models. In most of the existing studies, external validation and definition of applicability domain are not included. Therefore it is not possible to know the predictive power of the model and whether or not the predictions are applicable and valid.
- Molecular descriptor sets should ideally be related to the mechanisms of the various treatment technologies for which models are to be developed. Further research is needed to identify sets of descriptors which can well describe the structural characteristics related to oxidation, as well as other treatment processes such as adsorption and membrane filtration. Descriptors with clear meanings which are easy to explain are preferred.
- The pH-dependence of the rate constants for dissociating-compounds has not often been modeled. Ideally, separate models should be developed for predicting rate constants of neutral species and ionic species. Then, the overall rate constant can be calculated for any specified pH. However, models for ionic compounds seem not be available in the literature, and suitable descriptors to describe the charged status of ionic compounds have not been identified. To make it more complex, some compounds may involve two or more ionic species (more than one p*K*a). For drinking water studies, the pH range of water is typically around 6-8. A model which can estimate the overall rate constant of dissociating-compounds is therefore needed.

Experimental data can never be replaced completely with QSPR predictions; however, obtaining experimental kinetic data for the large number of existing micropollutants would take years with high costs. Therefore, QSPR can be used as a valuable modeling tool for the initial evaluation of the effectiveness of a water treatment process for micropollutants of interest. Results obtained from valid QSPR models can be used; to estimate the removal of micropollutants in drinking water treatment processes; to guide the selection of treatment process for target compounds and guide the design of a testing strategy; to improve the evaluation and understanding of existing data; and to provide mechanistic explanations of physico-chemical treatment processes.

Chapter 3

Selection of Representative Emerging Micropollutants for Drinking Water Treatment Studies: A Systematic Approach

This Chapter is based on a paper of the same title has been published in Science of the Total Environment in January 2012 (Jin and Sigrid, 2012, 414, 653-663).

This article focuses on several tasks: (1) identify suitable molecular descriptors which link the structural characteristics of micropollutants to mechanisms of water treatment processes; (2) develop a systematic approach to select structural representative micropollutants for water treatment studies; (3) select a set of representative compound for oxidation processes studies. The kinetics of selected micropollutants in ozonation and advanced oxidation process UV/H_2O_2 will be studied in detail and rate constant will be determined (Chapter 4). Furthermore, QSPR models will be built for predicting the rate constant of micropollutants in the reactions with ozone (Chapter 5) and hydroxyl radicals (Chapter 6).

Outline: Micropollutants remain of concern in drinking water, and there is a broad interest in the ability of different treatment processes to remove these compounds. To gain a better understanding of treatment effectiveness for structurally diverse compounds and to be cost effective, it is necessary to select a small set of representative micropollutants for experimental studies. Unlike other approaches to-date, in this research micropollutants were systematically selected based solely on their physico-chemical and structural properties that are important in individual water treatment processes. This was accomplished by linking underlying principles of treatment processes such as coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration to compound characteristics and corresponding molecular descriptors. A systematic statistical approach not commonly used in water treatment was then applied to a compound pool of 182 micropollutants (identified from the literature) and their relevant calculated molecular descriptors. Principal component analysis (PCA) was used to summarize the information residing in this large dataset. D-optimal onion design was then applied to the PCA results to select structurally representative compounds that could be used in experimental treatment studies. To demonstrate the applicability and flexibility of this selection approach, two sets of 22 representative micropollutants are presented. Compounds in the first set are representative when studying a range of water treatment processes (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration), whereas

the second set shows representative compounds for ozonation and advanced oxidation studies. Overall, selected micropollutants in both lists are structurally diverse, have wide-ranging physico-chemical properties and cover a large spectrum of applications. The systematic compound selection approach presented here can also be adjusted to fit individual research needs with respect to type of micropollutants, treatment processes and number of compounds selected.

Keywords: molecular descriptors relevant in water treatment processes, principal component analysis, D-optimal onion design, representative and structurally diverse compounds, customizing selection approach

3.1 Introduction

Emerging micropollutants such as endocrine disrupting chemicals (EDCs) and pharmaceuticals and personal care products (PPCPs) are of environmental and public concern (Daughton and Ternes 1999). They enter the aquatic environment continuously through wastewater discharge, agricultural runoff, and municipal landfill leachates. Not surprisingly they have therefore been detected in surface water, groundwater (Kolpin *et al.*, 2002), and also in finished drinking water at very low concentrations (Snyder 2008; Benotti *et al.*, 2009; Huerta-Fontela *et al.*, 2011; Loos *et al.*, 2007). It is expected that larger numbers of micropollutants in higher concentrations may be detected in the future as water reuse becomes more prevalent. Of primary health concern is the potential chronic health effects associated with long term exposure to multiple micropollutants at trace concentrations (Snyder 2008). Hence, there is a growing interest in understanding the removal of these micropollutants during drinking water treatment. Conventional processes have been shown to be largely ineffective for emerging micropollutants (e.g. Stackelberg *et al.*, 2004; Westerhoff *et al.*, 2005), while various advanced treatment technologies such as ozonation and advanced oxidation processes (AOPs), activated carbon adsorption, reverse osmosis and nanofiltration are effective (Westerhoff *et al.*, 2005; Snyder *et al.*, 2007).

To assess the removal of micropollutants during drinking water treatment processes, it is usually necessary from a practical perspective to select a relatively small set of micropollutants for experimental study. Selection criteria commonly used are: occurrence (Pereira *et al.*, 2007), production volume and extent of usage (Wu *et al.*, 2007), suggested monitoring lists from authorities (Shemer *et al.*, 2006) and known health effects (Rosenfeldt and Linden, 2004). Some studies even consider several criteria simultaneously, e.g. Snyder *et al.*, (2007) considered occurrence, structural diversity and analytical methods together. Such approaches provide improved knowledge on water treatment processes for the selected micropollutants. However, these findings may not necessarily be applicable to other micropollutants which then need to be studied in turn. Although thorough, this 'one at a time' approach is not cost-effective. Further, it is simply not possible to assess all of the numerous contaminants currently in use or detected in water. To gain an overall understanding of the removal of numerous, structurally diverse micropollutants with an efficient level of experimental study, it is highly desirable to select a relatively small set of representative compounds from a large pool of micropollutants of interest. However, a well-defined compound selection method for water treatment studies is currently not available.

The challenge is to select micropollutants which are representative of others with respect to their behavior (e.g. removal, reaction rate constant, adsorbability, rejection, etc.) during water treatment, which applies physical, chemical and biological processes to remove contaminants. Therefore, the physico-chemical and structural properties of contaminants play an important role in their removals. Structurally similar compounds are expected to behave similarly in water treatment processes, based on the underlying theory of quantitative structure-property relationships (QSPR) (e.g. Lei and Snyder 2007; Kusic *et al.,* 2009; Magnuson and Speth 2005; Yangali-Quintanilla *et al.,* 2010). Thus, it is not desirable to select a group of relatively homogenous contaminants for water treatment studies as their behavior is expected to be similar. On the contrary, a structurally diverse set of contaminants which covers a wide spectrum of physico-chemical properties is more desirable since these will be representative of a larger pool of contaminants. A statistical process which combines principal component analysis (PCA) and experimental design has been developed to select structural representative compounds for QSPR studies (Eriksson and Johansson 1996; Knekta *et al.,* 2004; Papa *et al.,* 2007; Jin *et al.,* 2009). In those studies, a large number of available molecular descriptors were used for selecting structurally diverse and representative compounds. However, not all of these numerous descriptors and therefore properties are important in water treatment processes and hence, differences in these descriptors may not necessarily lead to differences in compound behavior in treatment processes. It follows that a statistical approach needs to be developed that is suitable for water treatment studies. This can be accomplished by linking structural characteristics of micropollutants to the removal mechanisms of treatment processes by identifying relevant properties/descriptors, which simultaneously reduces the number of descriptors employed in the selection process.

The objective of this study was therefore to develop a procedure to systematically select representative micropollutants for water treatment studies that reflect the structural diversity of micropollutants, while taking into account the underlying removal mechanisms of the treatment processes. The approach was based on both an improved statistical approach (Eriksson and Johansson 1996) and the knowledge that certain physico-chemical properties and structural functionalities of micropollutants are relevant to certain water treatment processes. As a result a large pool of micropollutants was characterized, and from this representative micropollutants for future investigations were identified. Conducting experimental treatment studies on these selected micropollutants will improve the overall understanding of contaminant removal in water treatment processes in a cost-effective manner.

3.2 Approach and Background

3.2.1 Overall Approach

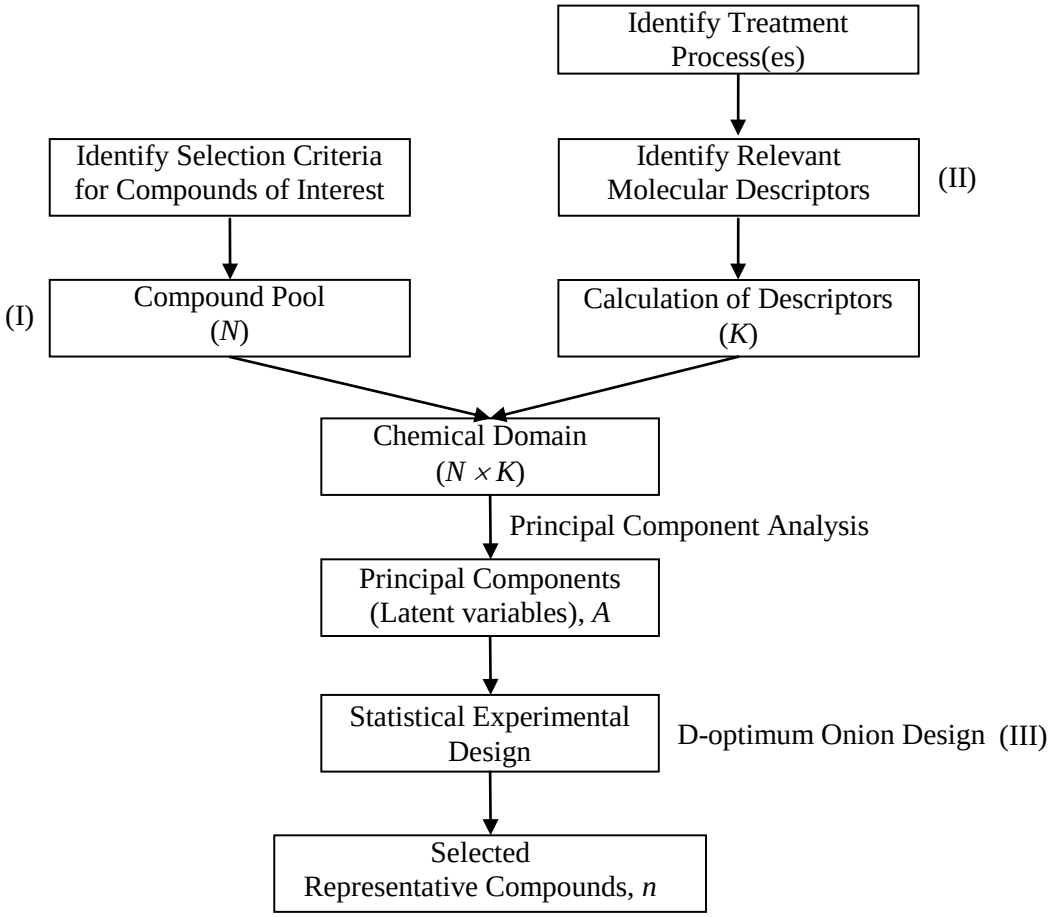


Figure 3.1 Approach to the selection of structurally representative compounds; *N*: number of compounds in the pool, *K*: number of molecular descriptors, *A*: number of principal components derived, *n*: number of selected compounds from the initial pool *N*.

The statistical approach which has to-date mostly been used in other fields (Knekta *et al.*, 2004; Papa *et al.*, 2007) is summarized in Figure 3.1. First, an initial compound pool which included a large number of heterogeneous micropollutants of environmental concern was defined (see section 3.2.2). Then, the original selection approach was modified by identifying a set of molecular descriptors, which explained the underlying removal mechanisms of various water treatment processes (details in the results in section 3.3.1). These descriptors were calculated for each compound (see section 3.2.3). As a result, a multivariate dataset, the chemical domain, consisting of 182 compounds and their

molecular descriptors was created. PCA was applied to characterize the information contained in this very large dataset and principal components (PCs) were extracted (see section 3.2.4). These PCs were limited in number and mathematically independent from each other, which was ideal for statistical experimental design. A small set of representative compounds was then selected by applying a D-optimal onion design (Eriksson *et al.*, 2004) instead of D-optimal design which was employed in the original selection approach (Eriksson and Johansson 1996), to the PCs (see section 3.2.5). D-optimal onion design is more flexible and provides a more controlled coverage of the chemical domain compared to D-optimal design.

Overall, this systematic approach (Figure 3.1) ensured that only a few, representative yet structurally diverse compounds were selected which were evenly distributed over the entire chemical domain. To illustrate the applicability and flexibility of this approach, two sets of compounds were identified. The first set (Compound set 1) serves for experimental screening/treatment studies investigating a wide range of water treatment techniques (Treatment set 1: coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration). The second set (Compound set 2) is representative for oxidation studies investigating ozonation and advanced oxidation processes (Treatment set 2).

3.2.2 Compound Pool

The compound pool can be defined in such a way that it fits the researcher's special interest and needs. For example, Papa *et al.,* (2007) defined a pool with 92 antibiotics. On the other hand, Knekta *et al.,* (2004) defined their pool using a potential hazardous contaminant list including 397 heterogeneous compounds since they aimed for a large range of structural diverse compounds and had no interest in a special compound group. The compound pool defines the boundary and applicability of the selection, i.e. the selected compounds are representative for compounds within the pool.

In this study, the intention was to select micropollutants for water treatment studies. Therefore micropollutants were included in the pool based mainly on two criteria: Occurrence in water and wastewater (1997 - 2008), and availability of kinetics or removal studies investigating water and wastewater treatment processes. Toxicity was not used as a criterion as available toxicity data are limited for many micropollutants, especially for emerging contaminants such as personal care products and potential endocrine disrupting chemicals, their impact is largely unknown (e.g., Kümmerer 2001; Stackelberg *et al.*, 2004**)**. Furthermore, available toxicity data have mostly been

established at concentrations much higher than those reported in the aquatic environment. However, the potential health and environmental impact of micropollutants was partly considered in this paper because some of the occurrence studies used in establishing the compound pool were based on contaminants with known or potential health effects (e.g., Kolpin *et al.,* 2002; Loraine and Pettigrove 2006). Overall 182 micropollutants were identified from the literature for the inclusion in the compound pool. Altogether 25 micropollutants (e.g., atrazine, trifluralin, dicamba) in this compound pool are regulated in Canada and/or the U.S. (Appendix A. Table A.1). Furthermore, 11 micropollutants are included in the Third Contaminant Candidate List (CCL3, US EPA) meaning that these compounds may require regulation due to potential health effects (Appendix A. Table A.1). The size of the contaminant pool was relative small compared to the fact that numerous contaminants are produced and consumed, and may subsequently be discharged into receiving waters. However, micropollutants in this pool were very heterogeneous in structure and included a number of chemical classes (e.g. phenols, PAHs, alkanes, halogenated aromatic compounds, organophosphorus compounds, etc.) thus covering a wide spectrum of physico-chemical properties. Most of the commonly found micropollutants are likely to fall within this spectrum. As shown in Table 3.1, 182 micropollutants in the pool covered a wide range of properties, e.g. their molecular weights (MW) ranged from 94.12 (phenol, no.115) to 777.12 g/mol (iomeprol, no.111), estimated logK_{ow} values ranged from -2.52 (iomeprol, no.111) to 7.07 (di(2-ethylhexyl)phthalate, no.165) and the number of isolated double bonds (nDB) ranged from 0 (e.g. hexachlorobenzene, no.118) to 7 (sultamicillin, no.6). A few compounds in this compound pool are similar in structure (e.g. α-HCH/β-HCH, *p,p'*-DDD/*p,p'*-DDT) but it is highly unlikely that the selection approach used here would select both compounds from a given compound pair.

Table 3.1 Diversity in properties of compounds pool (n = 182) compared to selected compound sets.

	Compound pool				Set 1[a]	Set 2[b]
	max.	min.	median	range*	% range[#]	% range[#]
logK_{ow}	7.07	-2.52	2.57	5.49	99	116
logS (mol/L)	-0.18	-8.31	-3.71	3.91	88	124
logD	8.03	-7.55	1.6	6.55	95	124
MW	777.12	94.12	278.38	226.64	89	94
AMW	23.73	5.51	7.81	6.09	136	111
nDB	7	0	1	3	130	100
nAB	24	0	6	12	118	100
nArOH	3	0	0	1	100	100
nN	2	0	0	1	100	100
Ui	4.64	0	3.17	2.41	95	88
HOMO (eV)	-7.019	-11.730	-9.125	1.507	105	99
LUMO (eV)	1.003	-2.099	-0.318	1.507	125	99
GAP (eV)	11.841	6.504	8.887	1.917	142	123
P	56.53	9.74	29.27	23.71	86	114
PSA (Å^2)	217.78	0	52.60	132.23	108	84
DM (debye)	10.27	0	2.93	5.83	104	83
L (Å)	23.4781	7.7760	13.4231	6.6224	83	88
W (Å)	14.8367	6.3506	9.1308	3.6992	87	157
RLW	2.8258	1.0154	1.4212	0.6579	76	93
HA	4.92	0	0.92	2.00	106	99
HD	2.97	0	0.46	1.41	101	153
Df($\times10^{-6}$ cm^2/s)	9.66	3.08	4.59	6.41	97	97

For abbreviations of properties see Table 3.2. *Range between 10 percentile and 90 percentile. [#]Percent represents the property range of the selected compounds set (10-90 percentile, n = 22) divided by the property range of the pool (10-90 percentile, n = 182). [a] Selected compounds for a range of water treatment processes (coagulation/flocculation, oxidation, activated carbon adsorption, membrane filtration); [b] Selected compounds for oxidation processes.

3.2.3 Calculation of Molecular Descriptors

To keep data processing manageable and to simplify interpretation of results, the number of descriptors employed in the selection process was minimized while retaining the important information. Note that large numbers of molecular descriptors can be generated though using software packages (e.g. over one thousand descriptors can be calculated by the DRAGON software).

However, 22 molecular descriptors (physico-chemical properties) which are relevant to treatment processes, were identified (section 3.3.1) in this study.

In some instances, physico-chemical properties can either be determined experimentally or by software calculated descriptors. Although experimentally determined data may be preferable, they are not available for many emerging micropollutants and the experimental processes to obtain these data are time-consuming and expensive. In addition, experimental data may be available from several sources, and in some cases these may differ. On the other hand, software packages can calculate numerous descriptors quickly, and they can be applied to large numbers of compounds including untested compounds. However, researchers have to exercise caution in that unreliable data may be provided if the target compounds are outside the applicability range of the software. In this study, calculated descriptors were favored because of the limited availability of experimental data. Andersson *et al*. (2000) found that calculated descriptors representing physico-chemical properties correlated well with the experimental data.

To calculate the identified descriptors, first of all, the structure of each compound was obtained from the online database ChemIDplus Advanced, United States National Library of Medicine (http://chem.sis.nlm.nih.gov/chemidplus/). For dissociable compounds, Marvin predictive tool (ChemAxon, http://www.chemaxon.com/marvin/sketch/index.jsp) was used to predict p*K*a values and to determine the dominant species between pH 5.5 and 8.5. If the dominant species was charged then the corresponding neutral structure was modified accordingly by ChemDraw (ChemOffice 2006, ChembridgeSoft). Software E-DRAGON (VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org, 2005), Molecular Modeling Pro Plus (MMP+ 6.25, ChemSW, Inc.), Marvin (ChemAxon, Inc.), and HyperChem (HyperChem 7.5, Hypercube, Inc.) were used for calculations (Appendix A. Table A.2). The input for these programs was either the molecular structure or the Simplified Molecular Input Line Entry System (SMILES) code both obtained from the ChemIdplus Advanced database. Then the calculated molecular weight was compared with the value reported in the ChemIdplus Advanced database. Only those compounds passing this validation were included. Quantum-chemical descriptors (HOMO and LUMO) were estimated using software package HyperChem (HyperChem 7.5, Hypercube, Inc.). The initial 3D structure of each compound was built by HyperChem. The molecular mechanics (MM+) force field was applied to optimize the conformation of each compound. The conformations with minimal energy were found using the semi-empirical AM1 method and the Polak-Ribiere algorithm, with a convergence limit of 0.01 kcal/mol

and a maximum number of calculation cycles set at 10 000. The overall resulting multivariate data matrix (compound × descriptors) is shown in Appendix A. Table A.3. Molecular descriptors are explained in detail elsewhere (Todeschini and Consonni 2000).

3.2.4 Principal Component Analysis (PCA)

The 182 micropollutants in the pool and their 22 calculated molecular descriptors formed a large, multivariate dataset (e.g. 182 × 22 data matrix). Statistical experimental design methods provide a methodology to choose representative compounds from this pool (Eriksson *et al.*, 2004), but they are only applicable to datasets with few variables which are independent from each other. With a multitude of compounds and descriptors in the chemical domain, it is likely that many of them will be correlated resulting in multicollinearity and multivariance of descriptors. To achieve dimensionality reduction and generate latent orthogonal (i.e. mathematically independent) variables, several techniques are available such as PCA, Factor Analysis, and Linear Discriminant Analysis. Eriksson *et al.* (2006) identified PCA as the most suitable technique for this purpose, and it has since been applied successfully in a number of studies (e.g. Knekta *et al.*, 2004; Papa *et al.*, 2007; Harju *et al.*, 2002). PCA reduces an original large data set to a small set of latent variables (PCs) that still contains most of the variation (i.e. information) of the original data set. A reduced set is much easier to analyze and interpret; more importantly, the PCs are mathematically independent, which is ideal for experimental design. Therefore, the combination of PCA and experimental design provides a powerful tool for representative compound selection (Eriksson *et al.*, 2006). In this study, numerical ranges of individual descriptors differed considerably and prior to PCA, all descriptors were mean-centered and scaled to unit variance. Some descriptors were log transformed to reduce skewness while obtaining an approximately normal distribution. The number of PCs was determined by cross-validation. Results are visually displayed by score plots (i.e. projecting compounds in relation to PC values, Figure 3.2a and Figure 3.3a) and loading plots (i.e. projecting descriptors in relation to PC values, Figure 3.2b and Figure 3.3b). Software SIMCA-P (Version 11.5, Umetrics, Sweden) was used for PCA.

3.2.5 D-optimal Onion Design

Statistical experimental design methods are tools for choosing representative and diverse compounds (Eriksson *et al.*, 2004). Classical design methods such as fractional factorial and central composite designs (Box and Draper 1987) are commonly used when factors (e.g., time and temperature) are

independent and can be continuously varied in a regular design region (e.g. selecting experimental conditions for treatment processes). However, molecular descriptors and the resulting PCs are inherent to a particular compound. They are therefore not controllable and as a result they form an irregular design region (property space). D-optimal design is more favorable in such situations because the selected compounds are distributed in such a way that they span a maximum volume in the property space. D-optimal onion design is an extension of D-optimal design, which splits the property space into several layers with a separate D-optimal design in each layer to achieve efficient and controlled coverage of the entire space (Eriksson *et al.*, 2004). In addition, it is very flexible because the number of layers and the regression model targeted within each layer can be altered, thus controlling the degree of coverage while balancing it with the number of compounds selected. This makes D-optimal onion design superior to other techniques such as the fractional factorial design, grid and cell based design, and space-filling design (Eriksson *et al.*, 2006) and it was therefore applied in this study. In this study, software MODDE (Version 8.0, Umetrics, Sweden) was used for D-optimal onion design.

3.3 Results and Discussion

3.3.1 Identification of Molecular Descriptors for Individual Treatment Processes

The most important modification to the original compound selection approach (Eriksson and Johansson 1996) was the targeted selection of molecular descriptors. Removal efficiencies in drinking water treatment depend to a large extent on the properties of the micropollutants as has been shown for many processes. It is therefore important to identify molecular descriptors which represent the properties important in individual treatment processes and include them in the compound selection process (Figure 3.1). Factors such as water quality and operational parameters are important when assessing removals for specific water sources but were not considered here as they are not pertinent for compound selection. Properties relevant in individual treatment processes can differ substantially (see below) and the discussion below provides justification for the inclusion of certain descriptors for individual treatment processes.

Only those descriptors with clear physical meanings which link back to properties identified as being relevant for a particular treatment process in the literature were included in this study. Descriptors such as molecular connectivity indices and topological descriptors are unlikely to provide interpretable insights into removal mechanisms and were therefore not included. In addition, the

selected descriptors should be applicable to the wide range of structural diverse compounds included in this study. Descriptors should not be restricted to a specific compound group. Detailed reasoning for the inclusion or rejection of properties/descriptors into Table 3.2 is provided below for individual water treatment processes.

Table 3.2 Selected molecular descriptors for water treatment processes.

Treatment Processes	Removal Mechanisms	Molecular Descriptors
Coagulation/flocculation[1]	Hydrophobic Interactions	logK_{ow}, water solubility (logS)
Ozonation and AOPs[1,2]	Functional groups	number of conjugated double bonds (nDB), number of isolated double bonds (nAB), number of primary and secondary amines (nN), number of phenolic group (nArOH)
	Energy of reaction	HOMO*, LUMO*, HOMO-LUMO gap (GAP), polarizability (P)
	Empirical QSPR models	molecular weight (MW) and average molecular weight (AMW), unsaturation index (Ui), diffusivity (Df)
Adsorption processes[1]	Van der Waals interactions	MW, molecular length (L) and width (W), length-width ratio (RLW)
	Hydrophobic interaction	logK_{ow}, logS, logD
	Electrostatic interaction	Polarizability
	Hydrogen bonding	Hydrogen bond acceptor (HA) and donor (HD)
	Mass transfer	Diffusivity (Df)
Membrane filtration[1]	Size exclusion	MW, molecular length (L) and width (W), length-width ratio (RLW)
	Electrostatic repulsion	Polarizability, polar surface area (PSA), dipole moment (DM)
	Adsorption	logK_{ow}, logS, logD

*HOMO: highest occupied molecular orbital energy, LUMO: lowest unoccupied molecular orbital energy. [1] Treatment set 1, includes coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration. [2] Treatment set 2 includes oxidation processes (ozonation and AOPs).

3.3.1.1 Coagulation/Flocculation

Coagulation/flocculation is typically used for suspended solids removal in conventional drinking water treatment plants preceding rapid filtration. Coagulation/flocculation alone has been shown to be ineffective in removing especially polar micropollutants (Ternes *et al.*, 2002; Adams *et al.*, 2002). However, removal may occur if the compounds partition onto particulates or onto precipitated flocs through hydrophobic interactions. Compounds with a high hydrophobicity as indicated by a high

octanol-water partitioning coefficient (logK_{ow}) value tend to partition onto the solid phase. Westerhoff *et al*., (2005) confirmed that logK_{ow} could be a good indicator. In addition, water solubility is inversely related to the hydrophobic property of a chemical. Thus, logK_{ow} and water solubility were included.

3.3.1.2 Oxidation Processes

Ozonation and advanced oxidation processes (AOPs) are effective in degrading micropollutants in drinking water (e.g., Huber *et al.*, 2003; Crosina *et al.*, 2006). Ozone is also applied for other treatment objectives such as disinfection or taste and odor control. During ozonation, oxidation occurs by molecular ozone (O_3) and hydroxyl radicals (·OH), which are produced through ozone decomposition in natural waters. Processes which involve the formation of highly reactive ·OH are generally referred to as AOPs (e.g. O_3/H_2O_2, UV/H_2O_2).

Oxidation efficiencies of pollutants are characterized by chemical reaction kinetics. The second-order rate constants for oxidation of micropollutants by O_3 cover a range of more than 9 orders of magnitude. O_3 selectively attacks organic compounds with a high electron density that is with functional groups such as double bonds, activated aromatic systems, and deprotonated amines (von Gunten, 2003). Carbamazepine, for example, containing a double bond showed a high reactivity ($k_{O3} = 3 \times 10^5$ $M^{-1}s^{-1}$) with ozone (Huber *et al.*, 2003). Activation of aromatic systems by electron donor groups (e.g., -OH) leads to increased rate constants, while electron withdrawing groups (e.g. -Cl, -NO_2, etc.) cause slower rate constants. Thus, benzene itself is relatively unreactive ($k_{O3} = 2$ $M^{-1}s^{-1}$) (Hoigné and Bader 1983a) while compounds with phenolic structures are highly reactive, e.g., 17α-ethinylestradiol ($k_{O3} = 7 \times 10^9$ $M^{-1}s^{-1}$) (Huber *et al.*, 2003). The amino group is only reactive in its deprotonated, neutral form and almost non-reactive in its protonated form (Benner and Ternes, 2009). Hence, the apparent rate constant is pH dependent. Molecular descriptors identified for the oxidation by molecular ozone were therefore: the number of isolated/conjugated double bonds, the number of primary and secondary amines, and the number of phenolic groups. In addition, the unsaturation index was included which related to the presence of reactive functional groups, as has been reported for QSPR modeling of the tropospheric degradation of ozone (Gramatica *et al.*, 2003).

In contrast, ·OH reacts non-selectively with micropollutants which is reflected by near diffusion-controlled second-order rate constants (10^8 - 10^{10} $M^{-1}s^{-1}$). Thus the diffusivity of compounds was included. Polarizability, molecular weight and average molecular weight were included as they were

found important and validated in predicting the hydroxyl radical rate constants of aromatic compounds in water (Kusic *et al.,* 2009). Quantum-chemical descriptors such as the highest occupied molecular orbital energy (HOMO), the lowest unoccupied molecular orbital energy (LUMO), and the HOMO-LUMO energy gap (GAP) were also included since they are widely used in predicting the reactivity of compounds in ozonation and hydroxyl radical reactions (Gramatica *et al.,* 2003). These parameters are directly related to the energy of the reaction. HOMO can be approximated with the negative ionization potential and LUMO with the negative electron affinity. A large GAP implies great kinetic stability and low chemical reactivity.

3.3.1.3 Activated Carbon Adsorption

Activated carbon adsorption is frequently used to remove organic micropollutants from contaminated water sources by granular activated carbon adsorbers or by powdered activated carbon addition. Factors affecting overall treatment efficiency are adsorption capacity, kinetics and competition/preloading of natural organic matter. However, equilibrium adsorption capacity of an adsorbent often serves as a starting point for process design, and hence, various models have been developed to predict these. An example is linear solvation energy relationships (LSERs) (Luehrs *et al.,* 1996), in which molar volume, polarizability, hydrogen-bonding acceptor and donor were identified as important parameters. In addition, QSPR models have been proposed correlating adsorption (Brasquet *et al.,* 1997; Magnuson and Speth, 2005) to the chemical structure of the adsorbate. Molecular connectivity indices which do not have a clear physical meaning were often used in such QSPR models. But since they do not aid in the phenomenological understanding of the adsorption mechanisms, they were not included here. Generally, adsorption depends on the properties of the activated carbon sorbent and the properties of the adsorbate. Van der Waals forces, hydrophobic interactions, electrostatic interactions, and hydrogen bonding are driving mechanisms. Van der Waals forces are related to molecular size which can be approximated by molecular weight. The size of the adsorbate also determines its access to the internal pores and therefore the available surface area for adsorption. In this case molecular size can be described by molecular weight but also molecular length, width and length-width ratio. LogK_{ow} was chosen to represent hydrophobic interactions (Magnuson and Speth, 2005) as hydrophobic compounds tend to leave the bulk solution more easily to attach to the adsorptions sites. However, Yu *et al.,* (2008) reported that at ng/L concentrations the compound with the highest logK_{ow} had the lowest adsorptivity on virgin carbon. Once the carbon was preloaded the adsorptivities followed the expected pattern (Yu *et al.,* 2009a).

Electrostatic interactions are best described by polarizability as this encodes information about the charge distribution in molecules (Magnuson and Speth 2005). Hydrogen bonding between water and the adsorbate decreases the adsorbate's affinity for the activated carbon, thus hydrogen bond acceptor/donor parameters are also important (Crittenden *et al.,* 1999). Mass transfer of the adsorbate to the adsorption site is another important factor to be considered. At ng/L concentrations this process is dominated by film diffusion through the boundary layer (Yu *et al.,* 2009b). Compound diffusivity is an important term when determining film diffusion coefficients and is therefore included.

3.3.1.4 Membrane Filtration

Membrane filtration such as reverse osmosis (RO) and tight nanofiltration (NF) can achieve high removals of organic compounds (e.g. Makdissy *et al.,* 2007). Rejection is a complex process driven by steric, hydrophobic and charge interactions which are influenced by membrane properties and solute structure (Bellona *et al.,* 2004). Main transport mechanisms are convection and diffusion (Kim *et al.,* 2007). Especially small, non-polar compounds have been reported to be able to diffuse across even dense membranes (Comerton *et al.,* 2008). Higher $\log K_{ow}$ values favor this behavior through initial adsorption onto the membrane surface and $\log K_{ow}$ was therefore included (Yangali-Quintanilla *et al.,* 2010; Bellona *et al.,* 2004; Comerton *et al.,* 2008). However, steric exclusion in which solutes larger than the membrane molecular weight cut-off (MWCO) are rejected remains the dominant mechanism for many compound-membrane combinations (Yangali-Quintanilla *et al.,* 2010; Van der Bruggen *et al.,* 1999). Molecular weight is often used but does not provide any information on the geometry of a molecule. Molecular size descriptors such as molecular length and width, and length-width ratio have been shown to be better predictors (Yangali-Quintanilla *et al.,* 2010; Van der Bruggen *et al.,* 1999) with clear physical meanings and were therefore included. Negatively charged compounds are electrostatic repelled by the negatively charged membrane surface, and at least partial rejection can be achieved even at molecular weights below the MWCO of the membrane (Van der Bruggen *et al.,* 1999). Polarizability which describes the electronic aspects of the whole molecule, polar surface area, and dipole moment were identified as indirect descriptors for electrostatic repulsion.

3.3.2 The pH Effect

The original statistical approach employed in QSPR studies (Eriksson and Johansson, 1996) almost exclusively focuses on neutral molecules. However, this study considered the effect of pH on the

molecular structure and hence, on the numerical value of the molecular descriptors. Depending on their p*K*a values charged species can be dominant at natural water pH values. As a result, not only compound properties, but also their behavior during treatment can change substantially affecting removals. LogK_{ow} values are specific to neutral species and do not consider any dissociation, whereas the distribution coefficient (logD) values are indicative of the pH dependent hydrophobicity of an ionizable compound. Polarizability and polar surface area are also affected by pH for ionizable compounds. In addition, molecular length and width are changed when compounds adjust to new structures by optimizing their geometry after losing or adding a proton. Therefore, the dominant species of all compounds at pH values typical for water treatment (pH 5.5 - 8.5) were determined first, prior to calculating their descriptors.

3.3.3 Compounds Selection Using PCA and D-optimal Onion Design

3.3.3.1 Compounds Selection for Treatment Set 1

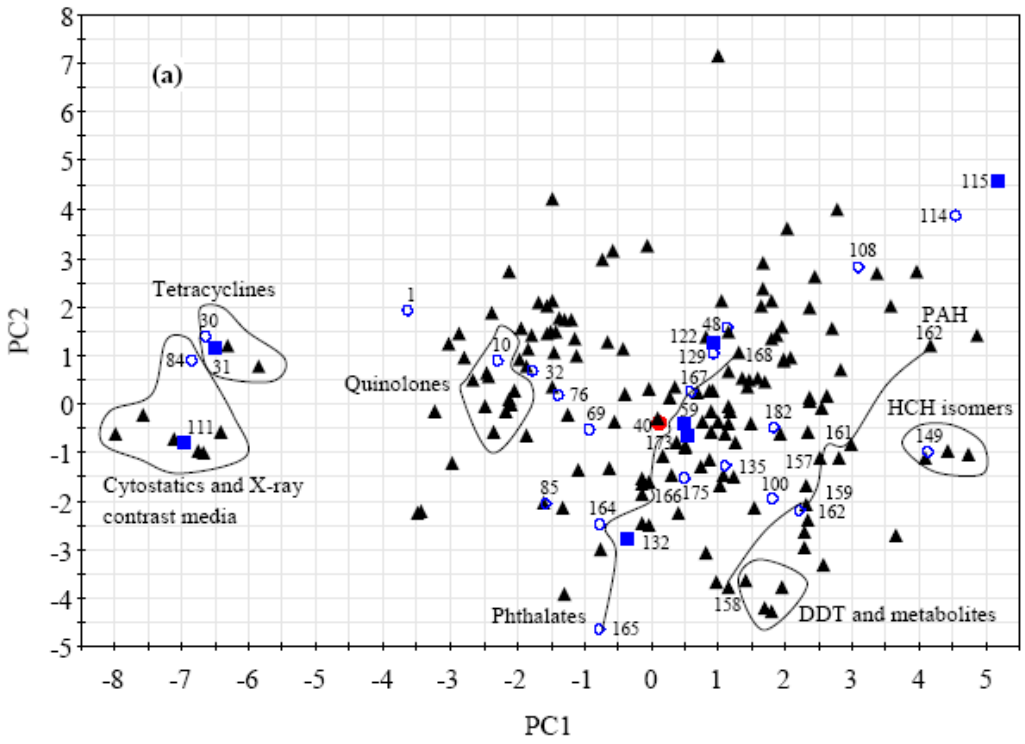
Treatment set 1 included coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration. PCA was applied to the multivariate dataset containing all 182 compounds and all 22 descriptors from Table 3.2 (i.e. descriptors for all treatment processes). This resulted in a five-dimensional model which explained cumulatively 78.2% (PC1-5: 31.2%, 16.0%, 14.2%, 9.9% and 6.9%) of the systematic variation in this dataset.

The score plot (Figure 3.2a) provides a summary of the chemical and structural variation among the 182 micropollutants. Compounds with similar properties are located close to each other forming clusters, e.g. HCH isomers (α-HCH, β-HCH, γ-HCH, and δ-HCH, no.147-150), DDT and its metabolites (*p,p'*-DDD, *o,p'*-DDT, *p,p'*-DDT, and *p,p'*-DDE, no.153-156), tetracycline antibiotics (chlortetracycline, doxycycline, oxytetracycline, and tetracycline, no.28-31), cytostatica and X-ray contrast media (iomeprol, and iopamidol, no.111-112), and quinolone antibiotics (ciprofloxacin, enoxacin, enrofloxacin, levofloxacin, lomefloxacin, norfloxacin, and ofloxacin, no.9-15). Compounds within these groups are expected to behave similarly during treatment studies and hence, one compound from each of these groupings should be well representative of the others in their group. The molecular descriptors are presented in the loading plot (Figure 3.2b). It is correlated to the score plot and indicates how descriptors are related to each other. By examining the loading values of each PC, it is possible to understand the contributions of the original descriptors to each PC. PC1 is strongly positively related to diffusivity, and inversely related to polar surface area, polarizability and

molecular weight. Thus, for example, cytostatica and X-ray contrast media which are largest in molecular weight and polar surface area are located at the far left of PC1. PC2 is strongly influenced by water solubility, and negatively related to logK_{ow}, logD and molecular weight. Therefore hydrophobic compounds such as DDT and its metabolites with their high logK_{ow} and logD values are located on the negative side of PC2. It is interesting that phthalates (butylbenzyl phthalate, di(2-ethylhexyl) phthalate, di-n-butyl phthalate, diethyl phthalate, and dimethyl phthalate, no.164 - 168) do not form a cluster although they have a common base structure. Instead they spread out in an almost vertical line starting from the central bottom. The significant properties that make them spread are logK_{ow} (as well as related logD and water solubility) and molecular weight. Among these phthalates, logK_{ow} values varied from 7.07 for di(2-ethylhexyl)phthalate (no.165) located at the bottom to 1.96 for dimethylphthalate (no.168) located in the center of the plot. A similar trend was also observed for polycyclic aromatic hydrocarbons (PAHs). The driving forces that make them spread vertically are again logK_{ow} and molecular weight. Benzo[a]pyrene (no.158) with its high molecular weight (252.32 g/mol) and low logK_{ow} (6.39) is located at the bottom, while naphthalene (no.160) with its smaller molecular weight (128.18 g/mol) and much lower logK_{ow} (3.33) is at the top. Being located at the center of the loading plot, variables such as HOMO, length-width ratio, and number of phenolic groups contributed minimally to PC1 and PC2. PC3 and PC4 capture significantly less information than PC1 and PC2, but they were still important for characterizing the overall chemical domain. PC3 versus PC4 score and loading plots are shown in Appendix A. Figure A.1.

To select representative micropollutants the D-optimal onion design was applied to the score plots of PC1-4. PC5 was excluded to limit overall complexity. The chemical domain was divided into 3 layers and a linear model was used in each layer. As a result, 22 compounds were selected (Table 3.3, Figure 3.2a) which are evenly distributed throughout the chemical space including one central point. D-optimal design allows for replacing of compounds with similar PC score values and structures. The original compound selection was in some instances modified considering information on commercial availability, analytical methods or existing studies. Regulated compounds were preferably selected wherever possible. Some compounds were replaced due to limited commercial availability and two others were included as they have been studied in detail (Table 3.3). The selected compounds (Table 3.1) displayed a wide spectrum of structural and physico-chemical properties (e.g. molecular weights ranged from 94.12 (phenol, no.115) to 777.12 g/mol (iomeprol, no.111), estimated logK_{ow} values ranged from -2.52 (iomeprol, no.111) to 7.07 (di(2-ethylhexyl)phthalate, no.165) and in spite of their small number (n = 22), they provide good coverage of the original range of properties.

Selected compounds included antibiotics, prescription and nonprescription drugs, pesticides, industrially relevant compounds and hormones. Their structures are available in Appendix A. Figure A.3.



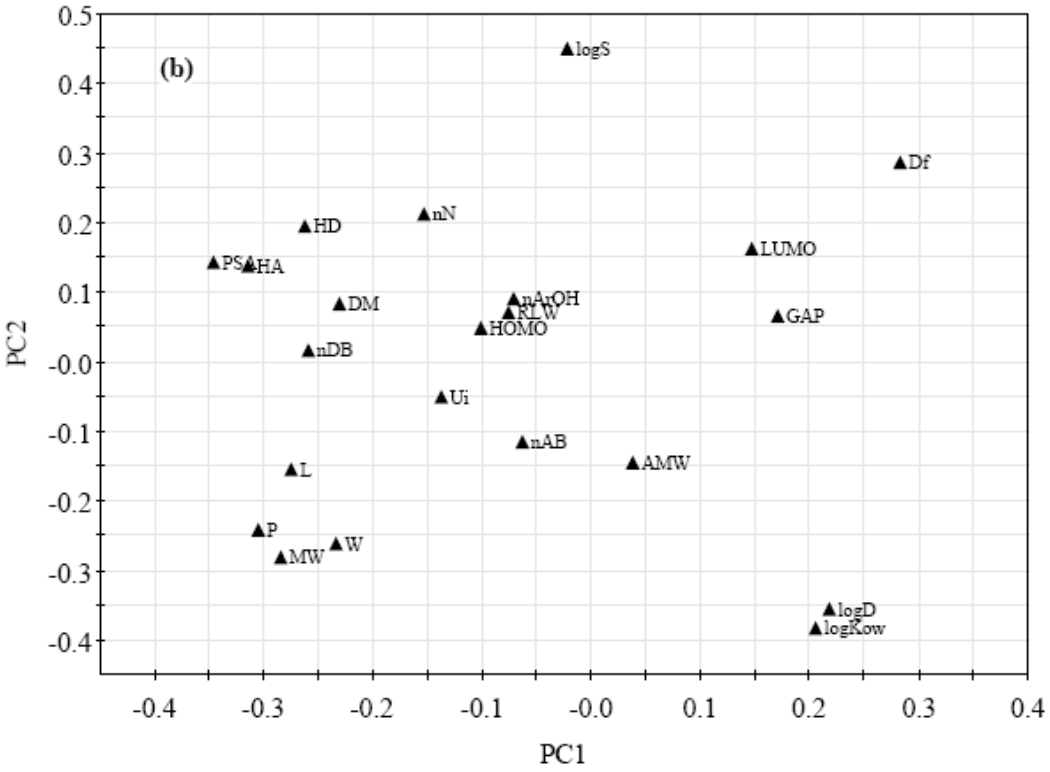
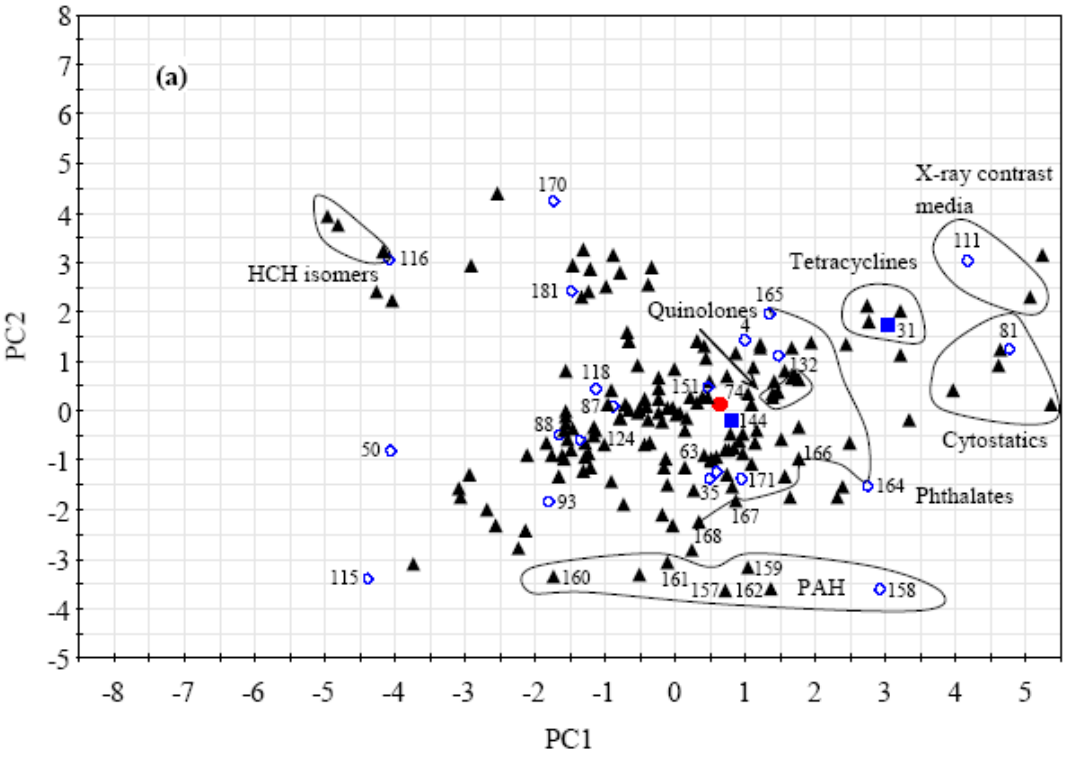


Figure 3.2 PCA analysis of chemical domain (182 compounds × 22 descriptors) covering all treatment processes listed in Table 3.2 (a) Score plot of PC1 and PC2 (showing micropollutant positions in relation to PC1 and PC2). D-optimal onion design applied (3 layers) for compound selection of treatment set 1. Black triangles: compounds not selected, blue circles: selected compounds, red dot: center compound, blue boxes: compounds selected to replace similar compounds; (b) Loading plot of PC1 and PC2 (showing the contributions of each descriptor to PC1 and PC2). For abbreviations see Table 3.2.


(a)
PC2
PC1
HCH isomers
Quinolones
Tetracyclines
X-ray contrast media
Cytostatics
Phthalates
PAH
116
170
181
165
4
31
111
81
132
118
151
74
87
144
88
124
63
166
50
171
164
35
93
167
168
115
160
161
159
157
162
158

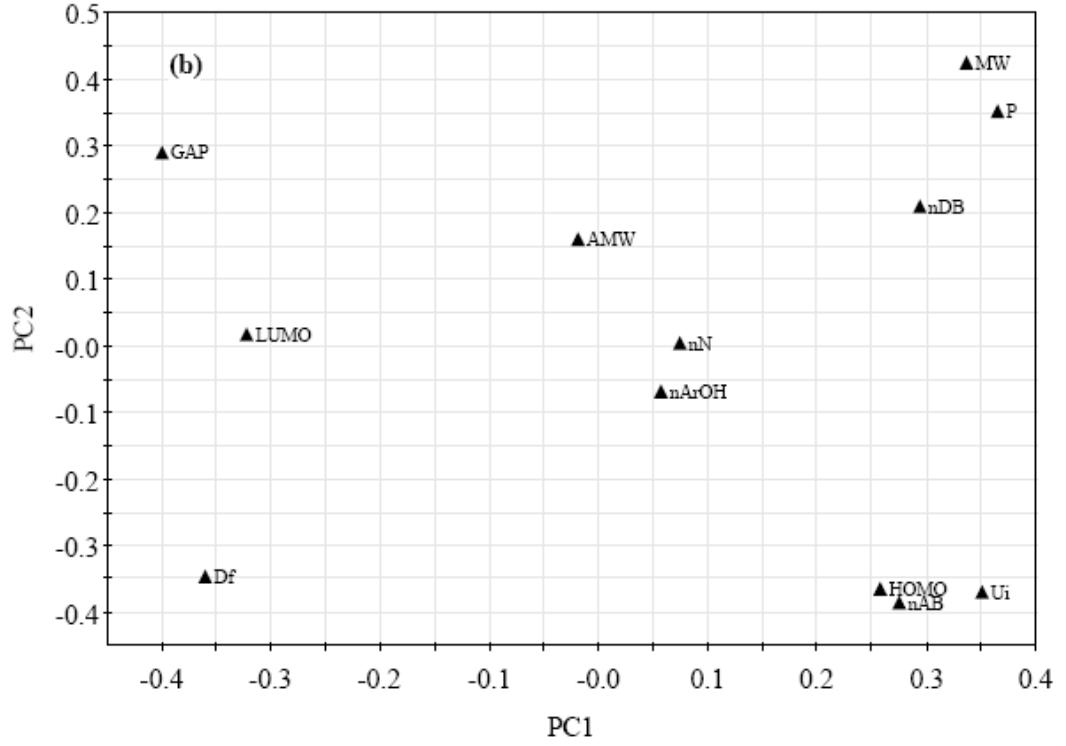

(b)
PC2
PC1
MW
P
GAP
nDB
AMW
LUMO
nN
nArOH
Df
HOMO
nAB
Ui

Figure 3.3 PCA analysis of chemical domain (182 compounds × 12 descriptors) covering oxidation processes listed in Table 3.2. (a) Score plot of PC1 and PC2 (showing micropollutant positions in relation to PC1 and PC2). D-optimal design applied (3 layers) for compound selection of treatment set 2. Black triangles: compounds not selected, blue circles: selected compounds, red dot: center compound, blue boxes: compounds selected to replace similar compounds; (b) Loading plot of PC1 and PC2 (showing the contributions of each descriptor to PC1 and PC2). For abbreviations see Table 3.2.

3.3.3.2 Compounds Selection for Treatment Set 2

Treatment set 2 includes oxidation processes i.e. ozonation and AOPs. PCA was applied to the multivariate dataset using only the 12 descriptors identified for oxidation processes (Table 3.2). For example logK_{ow}, logD and water solubility are not included. The resulting three-dimensional model explained 69.6% (PC1-3: 31.7%, 23.2%, and 14.7%) of the variation in the data. The loading plot (Figure 3.3b) shows HOMO, number of isolated double bonds and unsaturation index clustered together with high PC values whereas low PC values were observed for these descriptors identified for treatment set 1 (Figure 3.2b). This is fitting since unsaturation index and isolated double bonds are specific for the reactivity with molecular ozone and HOMO is related to the reactivity of a compound. In addition GAP, which describes compound reactivity, has a much higher PC value in the oxidation loading plot (Figure 3.3b) compared to that of set 1 (Figure 3.2b). For most of the other descriptors, positions on the loading plots are inversed but of similar magnitude. For example, diffusivity, polarizability and molecular weight influence PC1 and PC2 substantially in both plots. Polarizability and molecular weight are also located closely together in both cases.

In the score plot (Figure 3.3a), structurally similar compounds were again grouped together and diverse compounds were far apart. Compared to the score plot of treatment set 1 (Figure 3.2a) there were some similarities but also differences. Generally, compound positions were inversed due to the inverse behavior of many descriptors in the loading plots. The score plot of treatment set 1 (Figure 3.2a) has several clusters at high PC values which are mirrored in the oxidation (treatment set 2) score plot (Figure 3.3a) at lower PC values, e.g. HCH isomers are located in the upper left corner instead of the lower right. Tetracyclines, cytostatica and X-ray contrast media behaved in the same manner. They are all heavily influenced by molecular weight and polarizability which are of similar weight in both loading plots but in inverse directions. Interesting trends were found for phthalates and PAHs. Unlike in Figure 3.2a, phthalates are scattered in the central part of the score plot, mainly because the

previous dominant descriptor logK_{ow} was excluded. Their positions were now strongly influenced by molecular weight and polarizability, both important for PC1. PAHs were found to be distributed horizontally at the bottom of the score plot. Similarly, because of the exclusion of logK_{ow}, molecular weight and number of aromatic double bonds (nAB) became the influential descriptors of PC1 for the PAHs determining the positions of individual PAHs. Benzo[a]pyrene (no.158) with the highest molecular weight (252.32 g/mol) and high nAB (24) is at the far right, while naphthalene (no.160), the PAH with the lowest values for molecular weight (128.18 g/mol) and nAB (11), is at the far left. Score and loading plots for PC2 versus PC3 are shown in Appendix A. Figure A.2.

D-optimal onion design was then applied to the PCs and 22 representative compounds were identified (Table 3.3, and Appendix A. Figure A.4). Again, selected compounds span a wide range of properties and applications (Table 3.1) covering the range of properties of the original micropollutant pool well.

The selected micropollutants are all heterogeneous in structure to each other as indicated by the Tanimoto coefficients. These coefficients are commonly used as a tool to determine the similarity (or dissimilarity) of a compound pair. If a Tanimoto coefficient is higher than 0.85 then compounds are considered similar (Matter 1997). In this study, the Tanimoto coefficients of all the compound pairs of the selected compound sets 1 and 2 were calculated using a free web service tool ChemMine (http://chemmine.ucr.edu) in which the algorithm is based on the maximum common substructure (Cao *et al.*, 2008). As shown in the Appendix A. Tables A.4 and A.5, all compounds in the two sets are dissimilar to each other (Tanimoto coefficient < 0.85).

Table 3.3 Representative micropollutants selected by D-optimal onion design

Treatment set 1 (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration)			**Treatment set 2** (oxidation processes)		
No. in Chemical Domain	Compound Name	Layer in D-optimal Onion Design	No. in Chemical Domain	Compound Name	Layer in D-optimal Onion Design
10	Enoxacin	1	4	Methicillin	1
48	Primidone	1	35	Triclosan	1
69	Carazolol	1	87	Gemfibrozil	1
85	Bezafibrate	1	88	Clofibric acid	1
100	Celestolide	1	151	Dicofol	1
122	Atrazine* (no.129, Prometone [a])	1	171	Equilenin	1
167	Diethylphthalate	1	93	Butylated hydroxyanisole	2
1	Amoxicillin	2	124	Dicamba*	2
32	Carbadox	2	132	Trifluralin*	2
108	Hydrocinnamic acid	2	164	Butylbenzylphthalate	2
132	Trifluralin* (no.164, Butylbenzyl phthalate [a])	2	165	Di(2-ethylhexyl)phthalate	2
135	α-Endosulfan	2	181	Testosterone	2
173	17α-Ethinylestradiol [c] (no.175, Mestranol [b])	2	31	Tetracycline (no.81, Epirubicin [b])	3
31	Tetracycline (no. 30, Oxytetracycline [b])	3	50	Metformin	3
76	Azathioprine	3	63	Fenoterol	3
111	Iomeprol (no.84, Methotrexate [b])	3	111	Iomeprol	3
115	Phenol [c] (no.114, 4-Methyl phenol)	3	115	Phenol	3
149	γ-HCH	3	116	Tris(chloroethyl)phosphate	3
162	Pyrene	3	118	Hexachlorobenzene	3
165	Di(2-ethylhexyl)phthalate	3	158	Benzo[a]pyrene*	3
182	Androsterone	3	170	Tri(2-butoxyethyl)phosphate	3
59	Ketoprofen (no.40, Hydrocodone [b])	center	144	Methoxychlor* (No.74, Sotalol [a, b])	center

Compounds in brackets were replaced with similar compounds on the left. * Regulated compounds in Guidelines for Canadian Drinking Waterloo Quality (2008). [a] Compounds were replaced with regulated compounds; [b] Compounds were replaced due to limited commercial availability; and [c] compounds were included as they have been studied in detail.

3.4 Relevance of Representative Micropollutants Lists and Applicability of Selection Approach

One of the key features of the selection process is that compound properties important for the water treatment processes under investigation were incorporated in the selection process. Hence, the selected representative compound set 1 (Table 3.3) can be used in experimental drinking water treatment studies screening all treatment processes listed in Table 3.1. These experimental results are particularly suited to assess treatability. However, if different micropollutants than the ones included in the original compound pool are of interest then the selection can be modified as described later in this section.

The representative compounds set 2 (Table 3.3) is tailored to ozone and AOPs studies as only descriptors identified as relevant for these processes have been used in the selection process. Compound set 2 may serve as a training set to develop QSPR models correlating compound structure to rate constants for reactions with O_3 and ·OH. Once validated, these models can be used to predict other untested compounds as long as the untested compounds reside within the chemical domain. These models would be valuable for screening compound behavior and assessing suitability of oxidation processes.

It should be pointed out though that both lists are not suitable for water quality surveys since a survey aims to provide information on occurrence and not treatability.

Analytical methods are available for all of the selected compounds in set 1 and set 2 since one of the selection criteria for inclusion in the original compound pool was that they are either detected in water at trace concentration or studied in depth (e.g. Kolpin *et al.*, 2002; Snyder 2008; Westerhoff *et al.,* 2005; Benotti *et al.,* 2009b). Realistically, more than one method will have to be employed for either list due to the diversity of the compounds. As with any study targeting these trace contaminants complex sample preparation methods and advanced instrumentation such as GC-MS and HPLC-MS will be required. However, the time and effort spent on these experimental studies are well worth it since the selected compounds are well representative of many others and an overall understanding of their behavior during treatment can be obtained. Thanks to the flexibility of the overall selection approach, in particular the D-optimal onion design, difficult to analyze compounds may be excluded or replaced as has been described in the latter part of section 3.3.1.

The selection approach presented here is flexible and it can be tailored to fit individual needs. Individual steps of the overall approach (indicated by I, II or III in Figure 3.1) can be customized. First of all, if other classes of micropollutants are targeted than the ones in this paper then the compound pool can be tailored by only including the compounds of interest (I in Figure 3.1). Secondly, individual treatment processes or a combination of treatment processes (e.g. adsorption-membrane filtration) can be selected to fit ones needs as long as the relevant descriptors are identified (II in Figure 3.1). As the knowledge of treatment processes grows the list of molecular descriptors can be modified. Third, the number of compounds selected for experimentation can either be reduced or increased by adjusting the number of layers in the D-optimal onion design (III in Figure 3.1).

It should be noted that the lists of selected compounds presented here (i.e. set 1 and set 2) are examples to show the applicability and flexibility of the selection approach. To recommend representative compounds to the water industry for specific purposes or for particular research projects, many factors need to be considered and the compound selection may have to be customized as described above to fulfill the needs of a particular project.

3.5 Summary and Conclusions

Micropollutants such as EDCs and PPCPs may pose a risk to drinking water consumers. To assess the effectiveness of water treatment processes, it is necessary to select a small group of representative micropollutants for experimental treatment studies. Unlike others to-date, this study proposes a systematic selection approach which identifies representative micropollutants solely based on their physico-chemical and structural properties relevant in individual water treatment processes. Results are summarized as follows.

- Physico-chemical properties (i.e. molecular descriptors) of micropollutants determine to a large extent their removal from drinking water. A set of 22 molecular descriptors which are relevant to the removal mechanisms of individual treatment processes (i.e. coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration) was identified. Only descriptors with clear physical meanings were included.
- A systematic statistical approach combining principal component analysis and experimental design was modified and applied to a pool of heterogeneous micropollutants and their molecular descriptors. Principal component analysis summarized the variation in this original multivariate dataset and extracted latent variables, the principal components. D-optimal onion

design was applied to these principal components to select structural representative compounds.

- To demonstrate the applicability of the selection approach, it was applied to a pool of 182 micropollutants and two sets of 22 representative micropollutants were selected. The first set is suitable for experimentally studying a range of water treatment processes (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration) whereas the second set can be used for studying oxidation processes. The small number of selected micropollutants (22 out of 182) provided very good coverage over the entire property space and thus represented the original micropollutant pool well.
- Maximum information on treatability of compounds with very diverse structures can be obtained with a minimum amount of experimental study when using the selected compounds, therefore making treatment studies more cost effective.
- The selection approach presented here is flexible and can be customized to fit individual needs by for example reducing the number of compounds, applying it to other processes such as adsorption and/or membrane filtration, or studying other classes of micropollutants by re-defining the compound pool.

Chapter 4
Kinetics of Selected Micropollutants in Ozonation and Advanced Oxidation Processes (UV/H_2O_2)

This Chapter is based on a paper of the same title was submitted to Water Research in April 2012.

This article focuses on the determination of the reaction rate constants of selected micropollutants in their reaction with ozone and hydroxyl radicals. Micropollutants included in this study were selected as representative compounds from a large initial compound pool (Chapter 3). The experimental data obtained in this study were later used for model development in Chapters 5 and 6.

Outline: Second-order reaction rate constants of micropollutants with ozone (k_{O3}) and hydroxyl radicals (k_{OH}) are essential for evaluating their removal efficiencies from water during ozonation and advanced oxidation processes. But kinetic data are lacking for many of the newer micropollutants. Twenty-four micropollutants with very diverse structures and applications including endocrine disrupting chemicals, pharmaceuticals, and personal care products were selected, and their k_{O3} and k_{OH} values were determined using batch-scale reactors. Three different methods were used to determine k_{O3} values whereas competition kinetics method was applied for measuring k_{OH} values. Reactions with ozone were highly selective as indicated by k_{O3} values ranging from 10^{-2} to 10^{7} $M^{-1} s^{-1}$. The general trend of ozone reactivity can be explained by micropollutant structures in conjunction with the electrophilic nature of ozone reactions. All of the studied compounds are highly reactive with hydroxyl radicals as shown by their high k_{OH} values (10^{8} to 10^{10} $M^{-1} s^{-1}$) even though they were structurally very diverse. For compounds with a low reactivity towards ozone, hydroxyl radicals based treatment such as O_3/H_2O_2 or UV/H_2O_2 is a viable alternative. This study contributed to filling the data gap pertaining kinetic data of organic micropollutants while confirming results reported in the literature where available.

Keywords: ozone, hydroxyl radicals, rate constants k_{O3} and k_{OH}, water treatment

4.1 Introduction

Studies have documented a great variety of micropollutants including endocrine disrupting chemicals (EDCs) and pharmaceuticals and personal care products (PPCPs) in surface water and groundwater (Ternes 1998; Kolpin *et al*., 2002). Concerns about their effects on the environment and human health are increasing. Many of these micropollutants cannot be completely removed by drinking water treatment processes, and they can therefore be detected in finished drinking water (Benotti *et al*., 2009).

Although insufficient removals of many micropollutants from drinking water by conventional treatment processes have been observed, advanced technologies have shown great abilities to degrade/remove many of these micropollutants. In particular, ozonation and hydroxyl radicals based advanced oxidation processes (AOPs) are effective means for degrading micropollutants during drinking water treatment (Ikehata *et al.,* 2006). However, the major disadvantage of toxic by-products and incomplete mineralization has to be considered (von Gunten 2003). During ozonation the oxidation occurs through direct reactions with molecular ozone and indirect reactions with hydroxyl radicals which are produced by ozone decomposition. While molecular ozone selectively attacks organic compounds with high electron density functional groups such as double bonds, activated aromatic systems, and deprotonated amines, hydroxyl radicals react non-selectively with organic contaminants (von Gunten 2003). During AOPs micropollutants are degraded mainly by hydroxyl radicals which can be generated by various combinations of reactants such as UV/H_2O_2, O_3/H_2O_2, Fenton/photo-Fenton, UV/TiO_2, etc.

In order to evaluate the potential for removing micropollutants by ozonation or AOPs, kinetic data are needed to predict to what extent micropollutants will be degraded after a specified duration of treatment (e.g., ozone dose). Based on this initial assessment treatment processes can be optimized in pilot studies, or if current settings fail to reach a satisfactory removal alternative treatment processes may be considered. In addition, models have been developed to describe the removal efficiency of micropollutants in natural waters incorporating reaction rate constants, e.g., R_{ct} model for ozonation (Elovitz and von Gunten 1999) and $R_{OH,UV}$ model for UV/H_2O_2 AOP (Rosenfeldt and Linden 2007). Although kinetic data are available for a large number of chemicals for the reactions with ozone and hydroxyl radicals (Hoigné and Bader 1983a; Buxton *et al*., 1988), due to the complexity of analytical methods and the high cost of experimentation, there is still a data gap especially for emerging micropollutants.

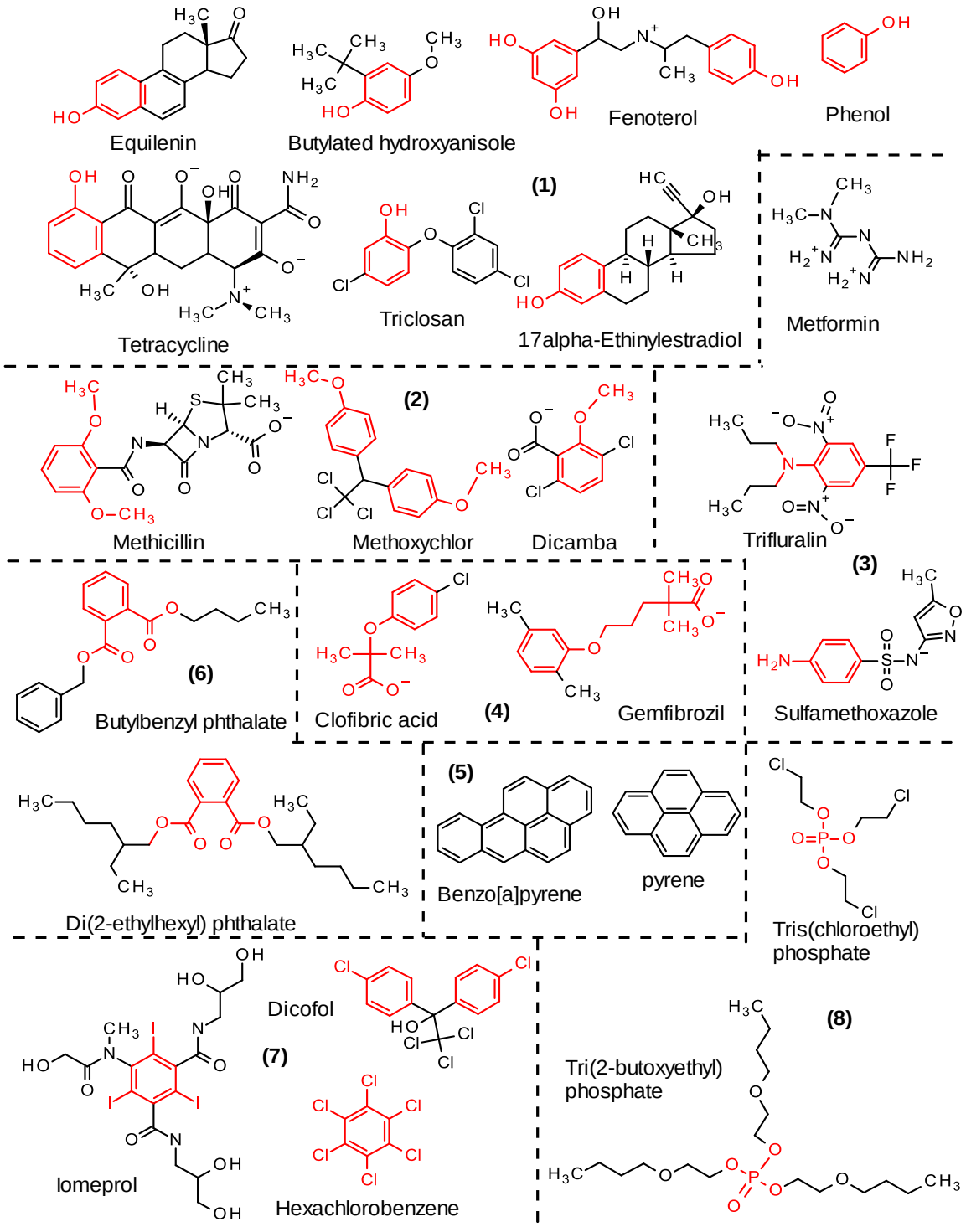


Figure 4.1. Structure of the selected 24 micropollutants at neutral pH. Micropollutants are divided into 8 groups based on their chemical structures: (1) phenolic compounds; (2) anisole derivatives; (3)

aniline and amine derivatives; (4) phenoxyalkanoic acid derivatives; (5) polycyclic aromatic hydrocarbons; (6) phthalates; (7) halo-substituted aromatics; (8) organophosphorus compounds. The common structural features for compounds in the same group were highlighted in red color.

The objective of this study was therefore to determine second-order reaction rate constants of twenty-four selected micropollutants (as shown in Figure 4.1) for the reaction with ozone (k_{O3}) and with hydroxyl radicals (k_{OH}), and therefore to assess the potential of ozonation and AOPs to degrade micropollutants with very diverse structures. Twenty-two of the micropollutants were selected using a statistical approach (principal component analysis followed by D-optimal onion design) from a compound pool containing 182 structurally diverse compounds. The selection process is based on linking the structural characteristics of micropollutants to the removal mechanisms of oxidation process such as ozonation and AOP. The 22 selected compounds were considered as structural representatives and better understandings can be gained by studying them in detail (Jin and Peldszus, 2012). In addition, two micropollutants (pyrene and sulfamethoxazole) were also studied for reference purposes. The determination of the rate constants was carried out in bench-scale experiments in pure aqueous solutions, and where available results were compared with literature data. The reactivity of the studied micropollutants was then linked to their structural characteristics.

4.2 Materials and Methods

4.2.1 Standards and Reagents

Standard chemicals were purchased from Sigma-Aldrich (Oakville, ON), and Toronto Research Chemicals (North York, ON) of the highest purity available (>97%). Stock solutions of these compounds were made by dissolving the individual compounds into ultrapure water (Millipore). Compounds with extreme low water solubility (e.g., hexachlorobenzene) were dissolved into methanol or acetonitrile first, and then a very small volume of the stock solutions was added into ultrapure water to reach the desired concentrations for the experiment. Indigo blue and *para*-chlorobenzoic acid (*p*CBA) were of the highest grade commercially available (99%). All other chemicals and solvents used were reagent grade and used without further purification.

4.2.2 Analytical Methods

Aqueous ozone concentrations were determined by UV absorbance at 258nm (ε = 3000 $M^{-1}cm^{-1}$) when there was no interference present in the reaction solution, otherwise the standard indigo colorimetric method (Eaton *et al.,* 2005) was used. All micropollutants were analyzed by a high-performance liquid chromatography (HPLC) system (Waters 600E system controller, Waters 717 plus autosampler, Waters 996 photodiode array detector, and Empower 2 chromatography data software). The column used was a Zorbax SB-C18 column (3.5 μm, 4.6 × 150 mm) (Agilent). Eluents consisted of 10 mM phosphoric acid buffer and methanol or acetonitrile. Varied eluent ratios were used depending on the compounds analyzed (as shown in Appendix B Table B.1). The injected sample volumes ranged from 20 to 50 μL depending on concentrations analyzed and quantification limits.

4.2.3 Determination of Rate Constant for the Micropollutants with Ozone

Ozone gas was generated using a water-cooled corona discharge generator (Ozotec Type 'S', model 2, Hankin Atlas ozone system, Canada) from oxygen feed gas (oxygen > 99%). Aqueous ozone stock solutions were prepared by continuously bubbling gas phase ozone produced by the ozone generator into ultrapure water (for a minimum of 1 h) chilled in an ice-water bath. The concentrations of the ozone stock solutions ranged from 15-20 mg/L. The pH was adjusted to a value of 7 ($\pm$ 0.1) in all experiments by adding orthophosphate buffers.

Three different methods were used in this study to determine the k_{O3}, namely the competition kinetics method, the compound monitoring method, and the ozone monitoring method. The appropriate experimental method was selected based on the expect reactivity of the target compound towards ozone. The competition kinetics method is convenient to use because ozone decay does not need to be monitored. However, it is only applicable to fast reacting compounds (k_{O3} > 1000 $M^{-1}s^{-1}$). Phenol is commonly used as the reference compound (Deborde *et al*., 2005), and pH control is important since the k_{O3} of phenol varies greatly with pH ($k_{neutral,O3}$ = 1.3×10^{3} $M^{-1}s^{-1}$ and $k_{anion,O3}$ = 1.3×10^{9} $M^{-1}s^{-1}$, p*K*a = 9.9). The compound monitoring method is used for slow reacting compounds (k_{O3} < 1000 $M^{-1}s^{-1}$) in which the decrease of the compound in presence of ozone (at least 10-fold excess) is monitored together with the ozone decay (Hoigné and Bader 1983b). This method is not applicable to fast reacting compound because the compound would be exhausted in a very short period of time before consecutive samples could be taken. Both methods require an analytical method to determine the remaining concentration of the compounds involved and the reference compound in

the case of the competition method. This typically involves the use of analytical instrumentation such as HPLC or GC-MS. If the instrumentation or a suitable analytical method is not available, an alternative indirect way to determine the rate constant of slow reacting compounds is the ozone monitoring method. Instead of monitoring the concentration of a compound, this method monitors the concentration of ozone as a function of time in the presence of the target compound (in excess) (Yao and Haag 1991).

Competition kinetics (for fast reacting compounds, $k_{O3} > 1000\ M^{-1}s^{-1}$). The target compound and a reference compound (phenol) with a similar reaction rate constant to the target compound were added to a series of 25 mL flask at the same initial concentrations (1-10 μM). This solution was buffered at pH 7 (± 0.1) using phosphate buffer, and contained *tert*-butyl alcohol (10 mM) as a hydroxyl radicals scavenger. The experiments were carried out at 20-22°C. Seven under-stoichiometric concentrations of the ozone stock solution were injected into individual solutions of the reaction mixture. The solutions in the serum vials were vigorously stirred to guarantee the even distribution of ozone during ozone injection. One minute after each injection, a 1 mL sample was withdrawn and samples were analyzed by HPLC. The experiment was repeated two to three times for each compound. The Equation 4.1 was used to calculate the rate constant.

$$\ln\left(\frac{[P]}{[P]_0}\right) = \ln\left(\frac{[R]}{[R]_0}\right)\frac{k_{O3,P}}{k_{O3,R}} \qquad \textbf{(4.1)}$$

Where $k_{O3,R}$ and $k_{O3,P}$ are rate constant of the reference compound (R) and target compound (P), respectively. $[R]_0$ and $[P]_0$ represent the concentrations of the reference and target compound before adding the ozone solution, respectively. $[R]$ and $[P]$ represent the remaining concentration of the reference and target compound after the ozone reaction, respectively. By plotting $\ln([R]/[R]_0)$ versus $\ln([P]/[P]_0)$, the ratio of $k_{O3,P}$ and $k_{O3,R}$ which was represented by the slope of the straight line can be determined. Then the rate constant of the target compound can be calculated since the rate constant of the reference compound is already known.

Phenol was selected as the reference compound. The apparent second-order rate constant of phenol at pH 7 was calculated ($1.8\times10^6\ M^{-1}s^{-1}$ at pH = 7) based on the literature data using Equation 4.2.

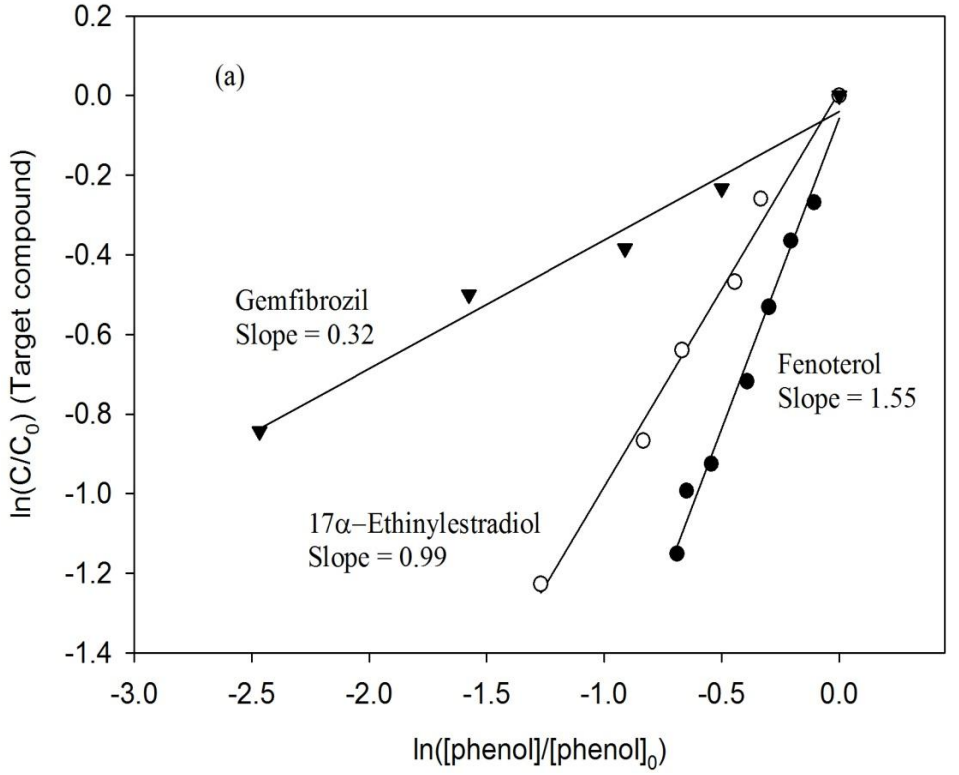
$$k_{appPhenol} = k_{O3,phenol}\frac{10^{-pH}}{10^{-pKa}+10^{-pH}} + k_{O3,phenolate}\frac{10^{-pKa}}{10^{-pKa}+10^{-pH}} \qquad \textbf{(4.2)}$$

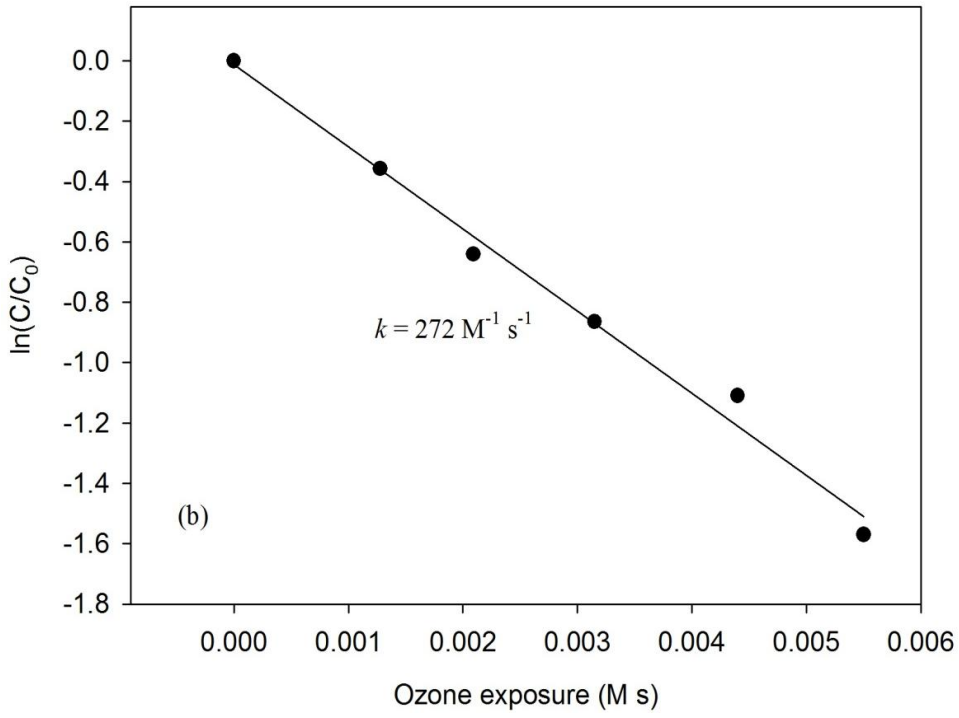
With $k_{O3,phenol}$ = 1.3 × 10^3 $M^{-1}s^{-1}$, $k_{O3,phenolate}$ = 1.4 × 10^9 $M^{-1}s^{-1}$, and p*K*a = 9.9 (Hoigné and Bader 1983b).

Compound monitoring method (for slow reacting compounds, k_{O3} < 1000 $M^{-1}s^{-1}$). The second-order rate constants were determined by following the decrease of the compound in the presence of ozone in at least a 10-fold excess. The experiments were carried out in 500 mL glass bottles at 20-22 °C and pH 7 (± 0.1), and containing *tert*-butyl alcohol (10 mM) as a hydroxyl radicals scavenger. An aliquot of the ozone stock solution was injected with a syringe to start the reaction. The initial ozone concentrations ranged from 15-300 μM depending on the expected reactivity of the compound towards ozone. Several 1 mL samples (at least 6 samples) were withdrawn with a dispenser system into HPLC vials, which contained fresh sodium sulfite solution (25 mM) to quench residual ozone. The total reaction time varied from a few minutes to several hours. Ozone decomposition had to be taken into account for experiments which were run for more than a few minutes. In these cases, the ozone exposure was determined by withdrawing additional samples for ozone analysis. The concentrations of the micropollutant were then plotted versus the ozone exposure. The rate constant was determined by the slope of the plot according to Equation 4.3.

$$\ln\left(\frac{[P]}{[P]_0}\right) = -k_{O3,P}\int [O_3]dt \qquad \textbf{(4.3)}$$

Ozone monitoring method (Yao and Haag 1991). This method involved monitoring the concentration of ozone as a function of time in the presence of at least a 5-fold excess of the organic compound. For compounds with negligible UV absorbance at 258 nm and water solubility higher than 200 μM, the reactions were initiated by injecting ozone stock solution into a 10-cm cuvette containing the compound, pH buffer (pH 7±0.1) and hydroxyl radicals scavenger (*tert*-butyl alcohol, 10 mM). The ozone concentration was monitored by a spectrophotometer at the single wavelength of 258 nm. The concentration of ozone as a function of time was measured by the indigo method for compounds where the UV absorbance of the compound interfered with the spectrophotometric determination of ozone at 258nm.


(a)
Gemfibrozil
Slope = 0.32
17α−Ethinylestradiol
Slope = 0.99
Fenoterol
Slope = 1.55
ln(C/C0) (Target compound)
ln([phenol]/[phenol]0)
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0


(b)
k = 272 M-1 s-1
ln(C/C0)
Ozone exposure (M s)
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
-1.8
0.000
0.001
0.002
0.003
0.004
0.005
0.006

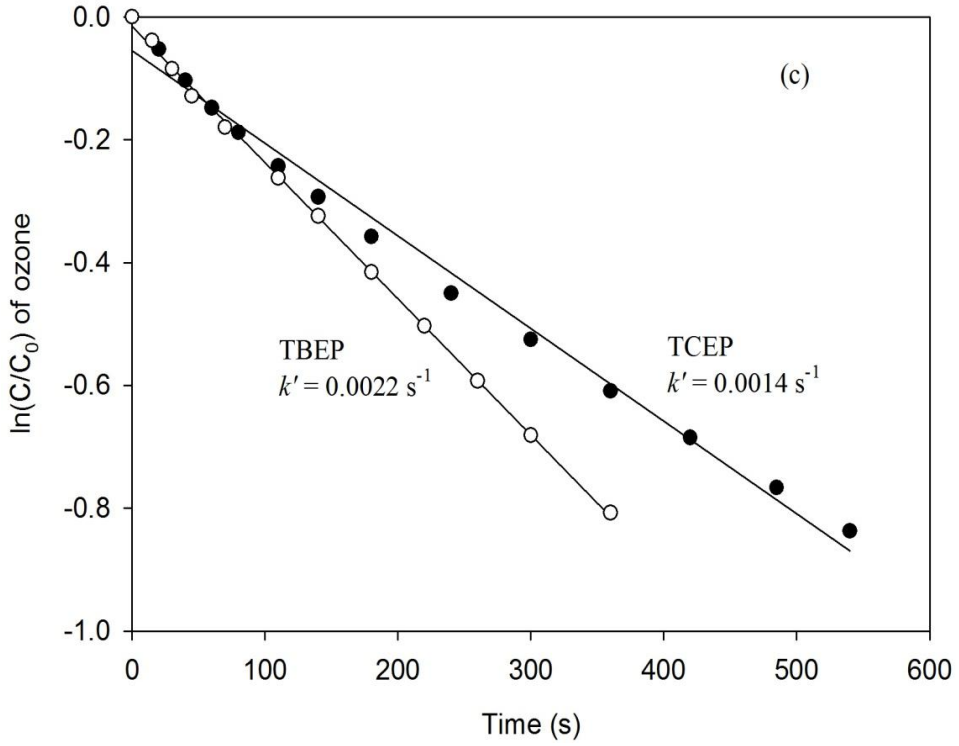


Figure 4.2 Determination of rate constant with ozone using three different methods. (a) competition kinetics method: phenol was the reference compound; (b) compound monitoring method: methoxychlor; (c) ozone monitoring method: the pseudo-first order rate of ozone decay (*k'*) was measured at pH = 7, TBEP represents tris(2-butoxyethyl) phosphate, TCEP represents tris(2-chloroethyl) phosphate.

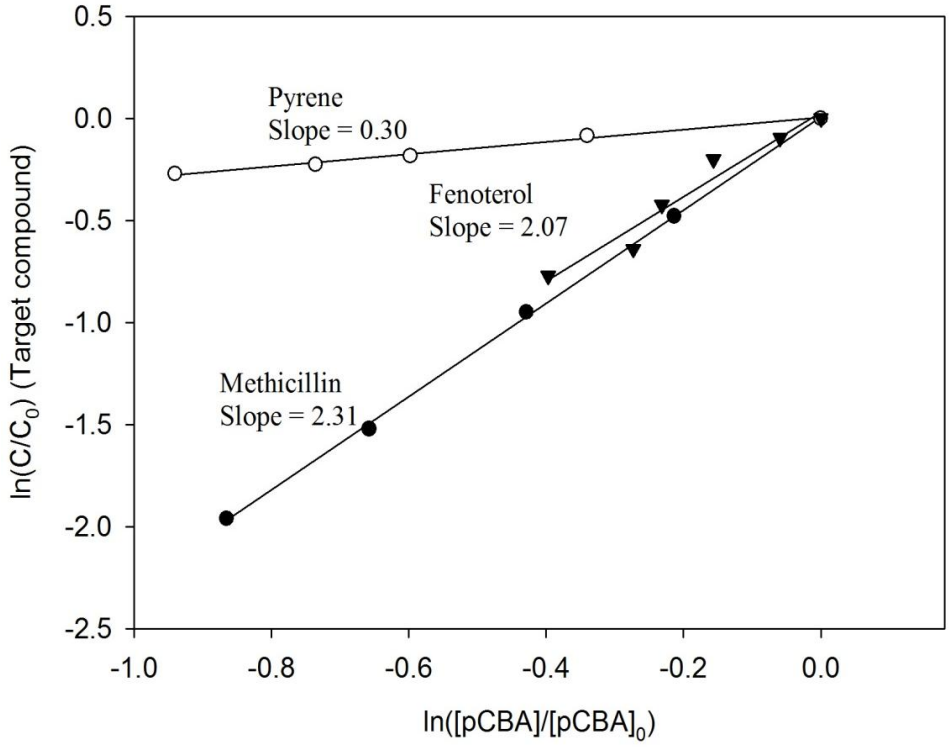


Figure 4.3 Determination of rate constant with hydroxyl radicals using competition kinetics method.

4.2.4 Determination of Rate Constant for the Micropollutants with Hydroxyl Radical

The competition kinetics method was used in the present study to determine the k_{OH}. Hydroxyl radicals were generated by UV/H_2O_2. It is preferable to apply the UV/H_2O_2 method to compounds which have low susceptibility to UV photolysis. An alternative to generate hydroxyl radicals is pulse radiolysis which decomposes water molecules without affecting the compounds being studied. It was found that the hydroxyl radical rate constants determined by UV/H_2O_2 and pulse radiolysis were in good agreement (Elovitz *et al*., 2008).

Experiments with UV/H_2O_2 were performed under a collimated beam apparatus (Calgon Carbon Corp, Pittsburgh, PA) equipped with a 1 kW medium pressure (MP) mercury lamp (ozone-free, Hanovia #6806A441, Union, NJ) which emits a broadband spectrum from 200 to 600 nm. UV fluence (mJ/cm^2) was calculated as the average irradiance multiplied by the exposure time. The average UV irradiance in the test water was determined from the incident irradiance, UV absorbance, and sample depth using a spreadsheet program developed by Bolton and Linden (2003). A UV radiometer (IL

1700, SED 240 detector, International Light, Peabody, MA) was used to measure incident irradiance (mW/cm^2) at the surface of the test water. The radiometer was calibrated at 2 nm intervals in the range of 200 to 400 nm. The UV absorbance (200 – 300 nm) of the test water was measured in a UV-visible spectrophotometer (HP 8453, Agilent Technologies, Santa Clara, CA). The exposure time (seconds) was determined by dividing the desired UV fluence (up to 100 mJ/cm^2) by the average UV irradiance. For the MP UV source, the fluence was calculated as the total UV output in the 200 – 300 nm range.

All experiments were performed in ultrapure purified water at room temperature (20-22°C) and the pH was buffered to 7 (± 0.1) using 5 mM phosphate buffer. The competition kinetics method was used to determine the second-order rate constants for the reaction with hydroxyl radicals. The probe compound *para*-chlorobenzoic acid (*p*CBA) was used as the reference compound with a known rate constant of $k_{OH} = 5 \times 10^9\ M^{-1}s^{-1}$ (Buxton *et al*., 1988). The p*K*a of *p*CBA is 3.98 (Park *et al.*, 2004) and at pH 7 *p*CBA is in its protonated form. The exposures were performed under the collimated beam apparatus, with spiked solutions (100 ml) containing equal concentrations (1 μM) of target compound and reference compound, which were placed in Pyrex crystallizing dishes (7.6 cm diameter, Fisher Scientific, Ottawa, ON) containing a small stir bar to provide constant mixing. Before and after exposing the spiked samples to UV irradiation, a calibrated radiometer was placed under the UV source at the same height as the water level in the Pyrex dish, to take incident irradiance measurements. Hydroxyl radicals were generated by photolysis of H_2O_2. 10 mg/L of hydrogen peroxide were added. The hydrogen peroxide residual were determined using the I_3^- method (Klassen *et al*. 1994). After exposure, catalase (4 g/L) was used to quench the residual H_2O_2 to stop the reaction.

Samples were repeatedly irradiated ($n \geq 2$) for constant time intervals. Samples were withdrawn at preset irradiation intervals (at least 5) and then analyzed by HPLC. The second order rate constants were calculated based on the Equation 4.4.

$$\ln\left(\frac{[P]}{[P]_0}\right) = \ln\left(\frac{[R]}{[R]_0}\right)\frac{k_{OH,P}}{k_{OH,R}} \qquad \textbf{(4.4)}$$

where $k_{OH,R}$ and $k_{OH,P}$ are hydroxyl radical rate constants for the reference (*R*) and target compound (*P*), respectively. Control experiments without H_2O_2 addition were performed to determine if the compounds undergo direct photolysis. If the compounds undergo significant direct photolysis,

the contribution from direct photolysis has to be considered in the calculations using the formula introduced by Shemer *et al.*, (2006) and Elovitz *et al.*, (2008).

$$k_{OH,P} = \frac{k_P^{'} - \frac{E_{avg(H2O2)}^{'}}{E_{avg(w/oH2O2)}^{'}} k_{d,P}^{'}}{k_R^{'} - \frac{E_{avg(H2O2)}^{'}}{E_{avg(w/oH2O2)}^{'}} k_{d,R}^{'}} \times k_{OH,R} \qquad \textbf{(4.5)}$$

Where k'_P, k'_R are the overall observed time-based pseudo-first-order rate constant for the degradation of target and reference compound using UV/H_2O_2, respectively. The terms $k'_{d,P}$, $k'_{d,R}$ are the pseudo-first-order direct photolysis rate constant using UV alone for the target and reference compound, respectively. The ($E'_{avg(H2O2)}/E'_{avg(w/o\ H2O2)}$) term in Equation 4.5 is a ratio of the average irradiance with H_2O_2 ($E'_{avg(H2O2)}$) and the average irradiance without H_2O_2 ($E'_{avg(w/o\ H2O2)}$).

4.3 Results and Discussion

Measured k_{O3} and k_{OH} values for the selected 24 micropollutants are shown in Table 4.1 . For most of these compounds kinetic data were not available and hence, the data presented here aid in filling a data gap pertaining to micropollutants relevant to the water industry. For some compounds rate constants have recently become available in the literature and results reported here confirmed the published data.

Examples of experimental results for k_{O3} and k_{OH} determinations are shown in Figure 4.2 and Figure 4.3, respectively. Initially, several compounds with known rate constants were tested to validate the experimental methods for rate constant determination. For example, the validity of the compound monitoring method was tested by determining k_{O3} of phenol, and the result ($1.1\times10^3\ M^{-1}s^{-1}$ at pH 2) was in good agreement with the reported value of $1.3\times10^3\ M^{-1}s^{-1}$ at pH 2 (Hoigné and Bader 1983a). The ozone monitoring method was tested by measuring k_{O3} of trichloroethylene ($13\pm0.3\ M^{-1}s^{-1}$), which was comparable to the reported k_{O3} values for trichloroethylene of $15\pm2\ M^{-1}s^{-1}$ by Yao and Haag (1991) and of $17\pm4\ M^{-1}s^{-1}$ by Hoigné and Bader (1983a). The competition kinetics method was tested by measuring k_{O3} of 17α-ethinylestradiol ($1.8\times10^6\ M^{-1}s^{-1}$ at pH 7) and k_{OH} of phenol ($6.1\times10^9\ M^{-1}s^{-1}$), and again results were consistent with the literature as shown in Table 4.1 (Buxton *et al*., 1988; Huber *et al*., 2003).

The 24 micropollutants studied here cover a wide range of applications and usage, including various pharmaceutical classes (e.g., antibiotics, lipid regulators and X-ray contrast media), disinfectants, pesticides and herbicides, fire retardants, natural and synthetic hormones, phthalates, etc. Up to now, kinetics for the reactions with ozone and hydroxyl radicals have not been investigated for many of these selected micropollutants. Furthermore, these compounds were selected from 182 micropollutants by a systematic statistical approach (Jin and Peldszus 2011) and they are therefore structurally representatives of many other similar micropollutants. Structural characteristics are very diverse (Figure 4.1) ranging from very large halogenated compounds (i.e. iomeprol MW = 777 g/mol) to small, nitrogen containing compounds (i.e. metformin MW = 129 g/mol). Hence, it is expected that their rate constants, especially k_{O3}, will cover a fairly large range because of the selective nature of ozone. This was confirmed by the measured ozone rate constants, k_{O3}, which ranged from <0.01 (hexachlorobenzene) to 1×10^{7} $M^{-1}s^{-1}$ (equilenin), and standard deviations for repeated measurements ($n \geq 2$) were all within 20% of the mean, except for benzo(a)pyrene (37%). Hydroxyl radical rate constants, k_{OH} ranged from 10^{8}~10^{10} $M^{-1}s^{-1}$, and repeat measurements ($n \geq 2$) were very consistent. This fairly narrow range of hydroxyl radical rate constants was expected since the hydroxyl radicals are very reactive, relatively non-selective oxidant.

Rate constants determined in this study were compared with those from the literature where available, and most of the measured rate constants were very close and thus confirming previously reported values (Table 4.1). For some cases, similar results were obtained albeit different experimental methods were employed. For example, similar k_{O3} results were obtained for pyrene and benzo(a)pyrene, even though competition kinetics was used in this study and the compound monitoring method was used in the reference (Butkovic *et al.*, 1983). In addition, similar results were found in the determination of k_{OH} for tetracycline, even though UV/H_2O_2 was used to generate hydroxyl radicals in this study, but the γ-radiolysis technique was applied by Dodd *et al*. (2006). In the case of sulfamethoxazole, different results were obtained when different reference compounds were used for competition kinetics. The k_{O3} of sulfamethoxazole obtained in this study was determined with phenol as a reference compound and it was close to the value reported by Huber *et al.,* (2003) who also used phenol as a reference. However, Dodd *et al.* (2006) using cinnamic acid as reference reported a k_{O3} value which was one magnitude lower.

For dissociating compounds rate constants, in particular k_{O3}, are pH dependent (Hoigné and Bader 1983b). Instead of determining the specific rate constant of each species involved, the apparent rate

constants at pH 7 were determined in this study as these are more relevant in drinking water practice. Over half of the studied compounds (13 out of 24 compounds) are dissociating compounds. At pH 7 either the protonated form or deprotonated form will be prevalent for almost all of these compounds (except triclosan and fenoterol) because their p*K*a values differ from pH 7 by at least 2 units (Table 4.1). The pH values in natural water often ranges from 6-8 and it follows that for most of these compounds rate constants determined at pH 7 are sufficient for preliminary assessment or modeling of compound reactivity under water treatment conditions.

In the following paragraphs, the relationship between the reactivity of micropollutants as described by their experimentally determined rate constants and their structural features will be discussed. For ease of discussion, the 24 micropollutants are divided into 8 sub-groups based on their structure, for example as phenolic compounds, anisole derivatives, aniline derivatives, etc. (as shown in Figure 4.1).

For aromatic systems, electron-donating substituents (such as –OH, $–NH_2$, $–OCH_3$, $–CH_3$) activate the benzene ring towards electrophilic attack such as molecular ozone and hydroxyl radicals, and electron-withdrawing substituents (such as $–NO_2$, –CN, –X) deactivate the benzene ring (von Gunten 2003). Therefore, in general, phenolic compounds (-OH), anisole derivatives ($-OCH_3$), and aniline derivatives ($-NH_2$) are expected to be of higher reactivity whereas halo-substituted compounds are expected to be of low reactivity. It will become apparent that reaction rates are influenced by a combination of structural features where some will have a stronger influence than others.

Phenolic Compounds. Phenols are aromatic systems activated by an electron-donating substituent (–OH). Phenol is a dissociating compound with an acid dissociation constant of p*K*a = 9.9, and the ionic form ($k_{O3,phenolate} = 1.4 \times 10^9$ $M^{-1}s^{-1}$) is much more reactive than the neutral form ($k_{O3,phenol} = 1.3 \times 10^3$ $M^{-1}s^{-1}$) (Hoigné and Bader 1983b). As a consequence phenol is much more reactive towards ozone at high pH values. As shown in Figure 4.4, all compounds with phenolic moiety show high reactivity towards ozone at pH 7, including equilenin, butylated hydroxyanisole, fenoterol, tetracycline, triclosan, phenol, and 17α-ethinylestradiol with k_{O3} values ranging from 1.8×10^6 $M^{-1}s^{-1}$ to 1.0×10^7 $M^{-1}s^{-1}$. The p*K*a values of these compounds ranged from 8.1 (triclosan) to 10.6 (butylated hydroxyanisole) and their rate constants are therefore expected to increase by several orders of magnitude when the phenolic group dissociates at higher pH values. However, with the exception of triclosan this increased reactivity with ozone at high pH values will likely not be relevant for drinking water treatment scenarios as pH values would have to be increased well beyond what is typically

encountered in natural waters. The k_{O3} values of tetracycline and 17α-ethinylestradiol determined in this study were very close to the reported values. But approximately one order difference in magnitude was observed in k_{O3} values for triclosan which may be due to the different methods employed. In this study the k_{O3} value of triclosan was measured at pH 7 by the competition kinetics method, while Suarez *et al*. (2007) monitored the loss of triclosan in presence of excess ozone using a continuous-flow quenched-reaction monitoring system. With this methodology they determined elementary rate constant of all possible species (Suarez *et al.*, 2007) which were used to calculate the k_{O3} value at pH 7 given in Table 4.1.

Equilenin shows the highest reactivity toward ozone and hydroxyl radicals as demonstrated by its high k_{O3} and k_{OH} values. Both equilenin and 17α-ethinylestradiol are estrogenic steroid hormone compounds with similar structures and p*K*a values. But equilenin is likely more reactive because of the fused pair of benzene rings (i.e. naphthalene moiety) which makes it more reactive towards an electrophilic attack than a single benzene ring. The k_{OH} of 17α-ethinylestradiol determined in this study is about half of the reported value by Huber *et al*. (2003) although the determination methods were basically the same. However, a factor of two is often considered a reasonable variation in terms of the absolute values of second-order rate constant of organic compounds in the reaction with OH radials (Elovitz *et al*., 2008). Haag and Yao (1992) also considered a factor of two as an acceptable range for the purpose of estimating rate constants during treatment processes.

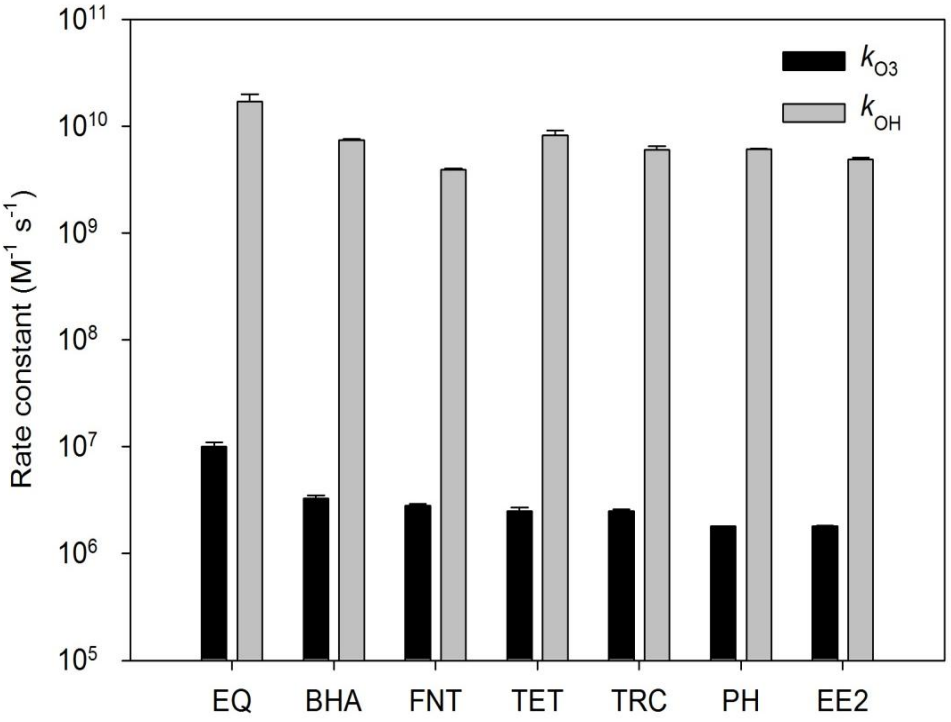


Figure 4.4 Experimentally determined k_{O3} and k_{OH} values of the selected phenolic compounds at pH 7: EQ is equilenin, BHA is butylated hydroxyanisole, FNT is Fenoterol, TET is tetracycline, TRC is Triclosan, PH is phenol, and EE2 is 17α-ethinylestradiol. The k_{O3} of phenol was from literature (Hoigné and Bader 1983b).

Anisole Derivatives. The methoxy group (-OCH_3) has electron donating properties and substitutions on a benzene ring with methoxy groups, especially on the *ortho* and *para* positions, increase therefore the electron density which favors electrophilic attack by ozone. As a result, the reactivity of anisole towards ozone (290 $M^{-1}s^{-1}$) is over one hundred times higher than the reactivity of benzene (2 $M^{-1}s^{-1}$) (Hoigné and Bader 1983a). Anisole derivatives such as methicillin, methoxychlor, and dicamba were investigated in the present study (Figure 4.5). Methicillin is a *β*-lactam antibiotic of the penicillin class. Methicillin shows the highest reactivity towards ozone among the three compounds, probably because of the absence of chlorine substitution which is a strong electron-withdrawing group. In contrast, dicamba is very resistant to ozone attack due to two chlorine substitutions on the benzene ring which decreases the electron density. The reactivity of methoxychlor falls in between methicillin and dicamba, probably because the chlorine groups are not directly substituted on the benzene ring, which softens the electron withdrawing effect of chlorine. It

should be noted that k_{O3} for the compounds with a phenolic substituent (Figure 4.4) were all higher than those measured for the anisole derivatives which is consistent with stronger electron donating properties of the –OH group. All of the anisole derivative compounds studied showed high reactivity toward hydroxyl radicals and thus can be efficiently degraded by hydroxyl radicals dominated treatment processes (Figure 4.5).

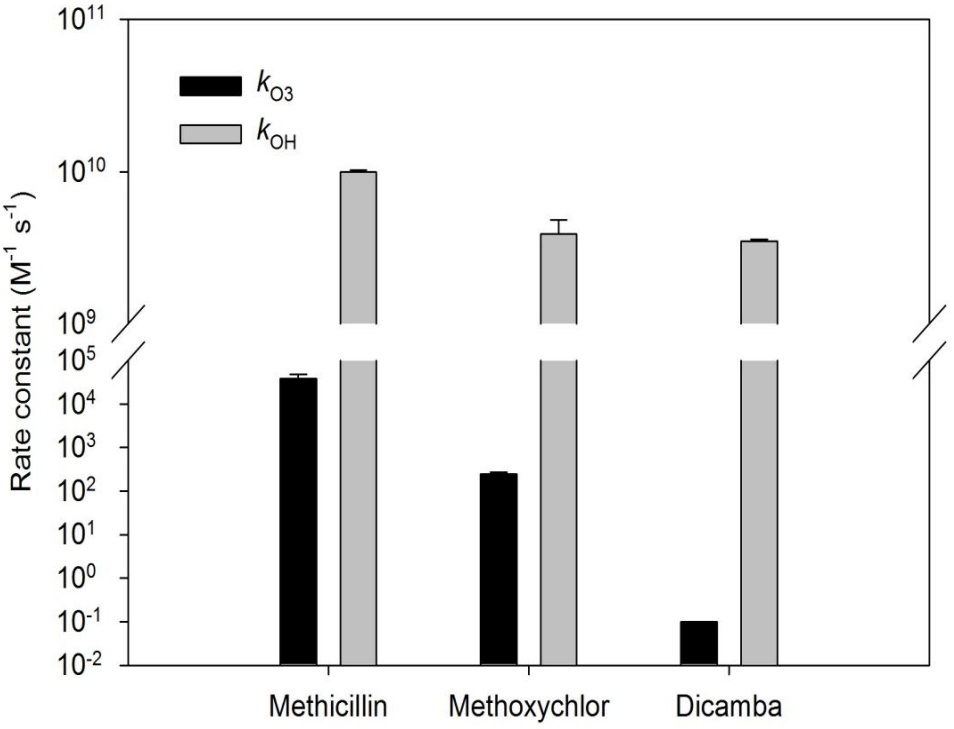


Figure 4.5 Experimentally determined k_{O3} and k_{OH} values of the selected anisole derivatives at pH 7.

Aniline and Amine Derivatives. Aniline (aminobenzene, Ar-NH_2) also represents an activated aromatic system, which can explain the high reactively observed for sulfamethoxazole (~10^6 $M^{-1}s^{-1}$). Sulfonamide antibiotics such as sulfamethoxazole, sulfapyridine, and sulfisoxazole have a common core chemical structure (*p*-aminobenzene sulfonamide) (Ikehata *et al*., 2006) in which amino substitution plays a key role. Dodd *et al*. (2006) identified several substructures (moieties) of molecules which are key sites for the ozone attack, and he reported k_{O3} values of those substructures. The k_{O3} value of the aminobenzene sulfonamide substructure was given as 4.7×10^4 $M^{-1}s^{-1}$. While the amino groups substituted to a pyrimidine structure instead of benzene ring, the moiety is also reactive (1.3×10^6 $M^{-1}s^{-1}$ by Dodd *et al*., 2006). Another aniline derivative studied here was trifluralin, a very widely used herbicide. Trifluralin shows a moderate reactivity toward ozone (1.9×10^3 $M^{-1}s^{-1}$ in this

study), probably because of the highly electron-withdrawing substitution groups $–NO_2$ and $–CF_3$. Metformin (1,1-dimethylbiguanide) is commonly used an oral antihyperglycaemic drugs, i.e. in the control of diabetes. Metformin with its amine substituents shows very low reactivity toward ozone (1.2 $M^{-1}s^{-1}$ at pH 7). It is known that the amino group is only reactive in its deprotonated, neutral form and almost non-reactive in its protonated form (Munoz and von Sonntag 2000). The deprotonated secondary and tertiary amines are reactive with ozone with a rate constant of around 10^6 $M^{-1}s^{-1}$, while primary amines react more slowly (Munoz and von Sonntag 2000). As shown in Figure 4.1, metformin is protonated at neutral pH which explains the low reactivity toward ozone.

Phenoxyalkanoic Acid Derivatives. The fibrate lipid regulators clofibric acid and gemfibrozil are phenoxyalkanoic acid derivatives, which are used as pharmaceuticals to accelerate the clearance of lipoproteins. These lipid regulator compounds have been detected in the aquatic environment (Kolpin *et al*., 2002). As indicated by its k_{O3} of 4.9×10^5 $M^{-1}s^{-1}$ at pH 7, gemfibrozil is more reactive toward ozone than clofibric acid, ($k_{O3} = 5.0\times10^3$ $M^{-1}s^{-1}$ at pH 7, in Table 4.1). Clofibric acid is less reactive probably because of the presence of chlorine on the aromatic ring. In addition, the electron-donating methyl substitution ($-CH_3$) increases the electron density of gemfibrozil, making it more susceptible to reactions with molecular ozone than clofibric acid. This is supported by a study on ozonation of wastewater where clofibric acid was shown to be relatively resistant to ozone treatment (Huber *et al*., 2005). It follows that advanced oxidation is more suitable for the degradation of such compounds, as their reactivity with hydroxyl radicals is very high as is apparent form their high k_{OH} values (Table 4.1).

Polycyclic Aromatic Hydrocarbons (PAHs). Two PAHs were measured in the present study, namely pyrene (four rings, 3.6×10^4 $M^{-1}s^{-1}$) and benzo(a)pyrene (five rings, 7.5×10^3 $M^{-1}s^{-1}$) and these data were in good agreement with a previous study (Butkovic *et al*., 1983). Note that for PAHs pH is expected to have a negligible influence on k_{O3} and this was confirmed by measuring k_{O3} at pH values ranging from 1 to 7 (Butkovic *et al*., 1983). Hence, values form different studies should be comparable even when differing pH values were used. When evaluating k_{O3} values from this study together with k_{O3} values determined by others it becomes apparent that an increase in rate constants is observed with an increasing number of rings i.e. benzene (1 ring, 2 $M^{-1}s^{-1}$ at pH 2 (Hoigné and Bader, 1983a)), naphthalene (2 rings, 3×10^3 $M^{-1}s^{-1}$ at pH 2 (Hoigné and Bader, 1983a)), phenanthrene (3 rings, 1.57×10^4 $M^{-1}s^{-1}$ at pH 2 (Butkovic *et al*., 1983)) and pyrene (4 rings, 3.6×10^4 $M^{-1}s^{-1}$; this study). A drop was observed though for benzo(a)pyrene with 5 rings, (7.5×10^3 $M^{-1}s^{-1}$; this study).

These increased k_{O3} values are due to the fused aromatic systems which have a substantially higher electron density than benzene alone. This in turn favors electrophilic attack by molecular ozone and hence, the rate constants of the fused aromatic systems are more than one thousand times faster than that of benzene.

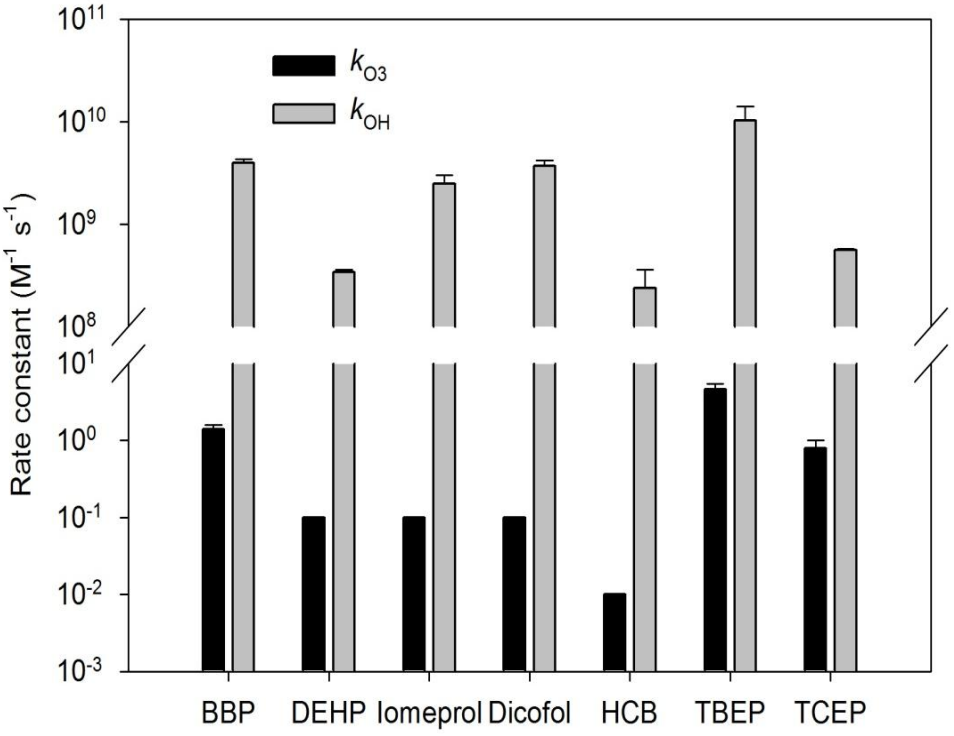


Figure 4.6 Experimentally determined k_{O3} and k_{OH} values of the selected micropollutants including phthalates (BBP and DEHP), halogen-substituted aromatics (iomeprol, dicofol, and HCB), and organophosphorus compounds (TBEP and TCEP) at pH 7. BBP is butylbenzyl phthalate, DEHP is di(2-ethylhexyl) phthalate, TBEP is tris(2-butoxyethyl) phosphate, and TCEP is tris(2-chloroethyl) phosphate. The k_{OH} of TBEP and TCEP were from the literature (Watts and Linden 2009).

Phthalates. Two phthalate compounds butylbenzyl phthalate (BBP) and di(2-ethylhexyl) phthalate (DEHP) were measured in the present study (Figure 4.6). Molecular ozone reacts with aromatic compounds by electrophilic aromatic substitution. Low k_{O3} values of phthalate are expected since the strong electron-withdrawing substitutions decrease the electron density of the aromatic rings and therefore lower their reactivity. Some k_{O3} values of other phthalates are also available, such as dimethyl phthalate (DMP) and diethyl phthalate (DEP) which both show very low reactivity toward

ozone (Yao and Haag 1991). But despite their low ozone reactivity, phthalates measured in this study i.e. BBP and DEHP are quite reactive toward hydroxyl radicals, with k_{OH} values of 4.0×10^9 $M^{-1}s^{-1}$ and 3.4×10^8 $M^{-1}s^{-1}$, respectively. As a comparison, DMP and DEP show similar reactivity with hydroxyl radicals as their rate constants are reported as 4.0×10^9 $M^{-1}s^{-1}$ (Haag and Yao 1992). Therefore, phthalates cannot be effectively removed by ozonation, but can be efficiently degraded during advanced oxidation processes since high k_{OH} values were observed (Figure 4.6). This figure includes three groups of micropollutants with low reactivity towards ozone.

Halogen-substituted Aromatics. The chlorine substituted compounds dicofol and hexachlorobenzene (HCB) have been used as herbicides/insecticides, and both show very low reactivity towards molecular ozone as indicated by small k_{O3} values. This low reactivity can be explained by the electron-withdrawing Cl groups on the benzene ring, which decrease the electron density of the ring and hence decrease their susceptibility to an electrophilic attack by molecular ozone. With six Cl substitutions on the benzene ring, HCB shows no reactivity toward ozone under the experimental conditions employed i.e. a detectable decrease of the HCB concentration could not be observed even after long exposure times to excess ozone. The k_{O3} of HCB is therefore reported as < 0.01 $M^{-1}s^{-1}$. Similarly, dicofol showed no reactivity although it has fewer Cl substitutions on the aromatic ring compared to HCB. The degradation of these compounds can be improved by applying advanced oxidation processes as their k_{OH} are in the range of 10^8~10^9 $M^{-1}s^{-1}$ found in the present study. Roche and Prados (1995) also found HCBs resistant to ozone but surprisingly more than half of HCB remained after treatment with O_3/H_2O_2 in their study. For dicofol and dicamba, however, addition of H_2O_2 into ozone enhanced the oxidation and nearly complete conversion was observed (Ikehata and El-Din 2005).

X-Ray contrast media compounds such as triiodinated benzene derivatives iomeprol and iopamidol are used to improve the visibility of internal body structure by X-ray imaging technologies. These compounds are highly hydrophilic and persistent in the aquatic environment, and have been detected in surface water and raw drinking water (Ikehata *et al*., 2006). Iomeprol is resistant to ozonation because of the triiodinated substitution structure. As shown in Table 4.1, the k_{O3} of iomeprol is very small (< 0.1 $M^{-1}s^{-1}$). A similar value was also reported by Huber *et al*. (2003). Iopamidol has almost the same chemical structure, and hence, shows a similar rate constant with a k_{O3} < 0.8 $M^{-1}s^{-1}$ (Huber *et al*., 2005).

Organophosphorus Compounds. Certain organophosphorus compounds are employed as flame retardants and plasticizer in a large variety of consumer products, and a number of studies became available over the last decade. The chlorinated ester tris(2-chloroethyl) phosphate (TCEP) is a flame retardant plasticizer, whereas the non-chlorinated ester tris(2-butoxyethyl) phosphate (TBEP) is widely used as a plasticizer in rubber and plastics. Ozone is very effective in eliminating many micropollutants, but fails to remove ozone refractory compounds ($k_{O3} < 10\ M^{-1}s^{-1}$) such as organic phosphates. In this study the k_{O3} of TCEP and TBEP were measured by the ozone monitoring method. As saturated aliphatic compounds, the ozone rate constants of TCEP and TBEP were found to be very low (0.8 and 4.6 $M^{-1}s^{-1}$, respectively). TBEP is somewhat more reactive than TCEP due to the non-chlorinated alkyl chain structure. Similar low rate constants ($< 2\ M^{-1}s^{-1}$) were also reported for other organic phosphate compound such as tri-n-butyl phosphate (TnBP) and tris(2-chloroisopropyl) phosphate (TCPP) (Pablo Pocostales *et al.*, 2010). Therefore, ozonation is not an effective treatment process for organic phosphate degradation. However, advanced oxidation processes show good potential to effectively remove such compounds by, for example, UV/H_2O_2 (Watts and Linden, 2009). The k_{OH} of TBEP was determined to be $1.03\times10^{10}\ M^{-1}s^{-1}$ and TCEP was $5.60\times10^{8}\ M^{-1}s^{-1}$ (Watts and Linden 2009). A significant reduction in the TBEP concentration was observed in model surface water treated with UV/H_2O_2 at neutral pH after up to 1000 mJ/cm^2 UV exposure, but little to no TCEP degradation was observed. A significant increase of initial H_2O_2 is required to reach a substantial TCEP degradation (Watts and Linden 2009).

Implications of Oxidation Kinetics during Water Treatment. The rate constants of twenty-four micropollutants encompassing diverse chemical structures were determined for the reaction with ozone and hydroxyl radicals. Oxidation by molecular ozone is selective for compounds with activated aromatics moiety such as phenolic compounds, anisole derivatives, aniline derivatives and PAHs. More than half of these micropollutants showed high-reactivity toward ozone with rate constants over 100 $M^{-1}s^{-1}$, which represents a half-life time of less than 5 minutes for a 1 mg/L ozone exposure. These micropollutants are expected to be largely degraded in water by ozonation. In the contrast, ozone will be ineffective for compounds with smaller k_{O3} values. Alternatively, advanced oxidation can effectively remove all the micropollutants studied because of the relatively non-selective nature of the hydroxyl radicals ($k_{OH} = 10^8 \sim 10^{10}\ M^{-1}\ s^{-1}$).

4.4 Conclusions

Second-order rate constants for the reaction of twenty-four structural diverse micropollutants with ozone and hydroxyl radicals were measured at pH 7 and 20-22 ℃. Competition kinetics, compound monitoring, and ozone monitoring methods were used for k_{O3} measurement; competition kinetics was used for k_{OH} measurement; the degradation of micropollutants was monitored by a HPLC with PDA detector. In view of the results we may conclude the following:

- The k_{O3} values determined in this study ranged from 10^{-2} to 10^{7} M^{-1} s^{-1}. For the majority of the compounds these values were not known yet thus providing valuable, basic information for modeling and design of ozonation and AOPs treatment.
- The general trend of micropollutant reactivity with ozone can be explained by the micropollutant structures and the electrophilic nature of ozone reactions. In general, compounds with activated aromatic rings including phenolic compounds, anisole derivatives, and aniline derivatives show high reactivity (~10^{4} to 10^{7} M^{-1} s^{-1}) toward ozone except dicamba and methoxychlor. Phenoxyalkanoic acid derivatives and polycyclic aromatic hydrocarbons show moderate reactivity (~10^{3} to 10^{5} M^{-1} s^{-1}). Compounds with deactivated aromatic rings such as phthalate and halo-substituted compounds show moderate to very low reactivity (~10^{-2} to 1.4 M^{-1} s^{-1}) toward ozone. Saturated aliphatic compounds such as organophosphorus compounds have a very low reactivity (< 10 M^{-1} s^{-1}) towards ozone as well.
- The k_{OH} values determined in this study ranged from 10^{8} to 10^{10} M^{-1} s^{-1} indicating that all selected micropollutants are highly reactive toward hydroxyl radicals, since hydroxyl radicals are relatively non-selective.
- For compounds with low reactivity toward ozone, ozonation treatment could be insufficient for removing them from drinking water, therefore hydroxyl radicals based treatment techniques such as O_3/H_2O_2 or UV/H_2O_2 are recommended.

Table 4.1 The k_{O3} and k_{OH} determined for 24 selected micropollutants at pH 7 and room temperature (20-22 ℃).

Compounds	p*K*a	k_{O3} ($M^{-1}s^{-1}$)			k_{OH} ($\times10^9$ $M^{-1}s^{-1}$)	
		Method	Measured	Reference	Measured	Reference
Equilenin	9.8 [a]	CK	$1.0(\pm0.1)\times10^7$		17±3	
Butylated hydroxyanisole	10.6 [a]	CK	$3.3(\pm0.2)\times10^6$		7.4±0.2	
Fenoterol	8.6 [a]	CK	$2.8(\pm0.1)\times10^6$		3.9±0.1	
Tetracycline	3.3, 7.7, 9.7 [b]	CK	$2.5(\pm0.2)\times10^6$	1.9×10^6 (Dodd *et al*., 2006)	8.2±0.9	7.7±1.2 (Dodd *et al*., 2006) 6.3±0.1 (Jeong *et al*., 2010)
Triclosan	8.1 [c]	CK	$2.5(\pm0.1)\times10^6$	3.8×10^7 (Suarez *et al*., 2007)	6.0±0.5	5.4±0.3 (Latch *et al*., 2005)
Phenol	9.9 [c]		ND	1.8×10^6 (Hoigné and Bader, 1983b)	6.1±0.1	6.6 (Buxton *et al*., 1988)
17α-Ethinylestradiol	10.4 [c]	CK	$1.8(\pm0.02)\times10^6$	1.6×10^6 (Deborde *et al*., 2005)	4.9±0.2	9.8 (Huber *et al*., 2003)
Methoxychlor	NA	CM	250(±24)	270±80 (Yao and Haag, 1991)	3.9±0.9	
Dicamba	2.0 [c]	CM	<0.1		3.5±0.1	
Methicillin	2.8 [c]	CK	$3.9(\pm0.9)\times10^4$		10±0.2	
Metformin	10.3 [a]	OM	1.2(±0.2)		ND	
Sulfamethoxazole	5.7 [c]	CK	$2.0(\pm0.1)\times10^6$	5.5×10^5 (Dodd *et al*., 2006) 2.5×10^6 (Huber *et al*., 2003)	ND	8.5±0.3 (Mezyk *et al*., 2007)
Trifluralin	NA	CK	$1.9(\pm0.3)\times10^3$		1.3±0.1	
Gemfibrozil	4.4 [a]	CK	$4.9(\pm0.9)\times10^5$		7.1±0.3	10.0±0.60 (Razavi *et al*., 2009)
Clofibric acid	3.4 [a]	CK	$5.0(\pm1.0)\times10^3$		5.2±0.4	6.98±0.12 (Razavi *et al*., 2009)
Benzo(a)pyrene	NA	CK	$7.5(\pm2.8)\times10^3$	6.2×10^3 (Butkovic *et al*., 1983)	0.94±0.4	
pyrene	NA	CK	$3.6(\pm0.4)\times10^4$	$3.9(\pm0.5)\times10^4$ (Butkovic *et al*., 1983)	1.4±0.2	
Butylbenzyl phthalate	NA	CM	1.4(±0.2)		4.0±0.3	
Di(2-ethylhexyl) phthalate	NA	CM	<0.1		0.34±0.02	

Iomeprol	NA	CM	<0.1	<0.8 (Huber *et al*., 2003)	2.5±0.5	2.03±0.13 (Cooper *et al*., 2010)
Dicofol	NA	CM	<0.1		3.7±0.5	
Hexachlorobenzene	NA	CM	<0.01		0.24±0.12	
Tris(2-butoxyethyl) phosphate	NA	OM	4.6(±0.9)		ND	10.3±3.8 (Watts and Linden, 2009)
Tris(2-chloroethyl) phosphate	NA	OM	0.8(±0.2)		ND	0.56±0.02 (Watts and Linden, 2009)

[a] estimated by Marvin software; [b] from Dodd *et al.*, 2003; [c] obtained by searching from the ChemIDplus online database. CK: competition kinetics. CM: compound monitoring method. OM: ozone monitoring method. NA: not applicable. ND: not determined.

Chapter 5
Modeling Ozone Reaction Rate Constants of Micropollutants Using Quantitative Structure–Property Relationships

This Chapter is based on a paper of the same title was submitted to Environmental Science and Technology in April 2012.

This article focuses on developing quantitative structure-property relationship models for predicting the rate constants of diverse micropollutants in the reaction with molecular ozone. The training set compounds were selected as representatives from a large compound pool (Chapter 3), and their rate constants were determined in experimental analysis (Chapter 4). In addition, a set of micropollutants collected from literature were used as validation set for model validation.

Outline: Quantitative structure-property relationship (QSPR) models were developed to predict the second-order rate constants of micropollutants with ozone (k_{O3}) from their structural characteristics. The models were developed using 12 molecular descriptors for 22 pre-selected structural diverse micropollutants, and then validated with an external data set. Piecewise linear regression (PLR) with a pre-defined breakpoint ($\log k_{O3}$ = 2.00 $M^{-1} s^{-1}$) provided the best results since reactions of low-reactive (sub-model $\log k_{O3}(<2)$) and high-reactive micropollutants (sub-model $\log k_{O3}(\geq 2)$) are governed differently. A classification function was developed using linear discriminant analysis (LDA) classifying micropollutants into high-reactive or low-reactive compounds before predicting $\log k_{O3}$ through the appropriate PLR sub-model. Overall, the PLR-LDA approach was able to predict the ozone rate constants of structural diverse micropollutants with a high certainty as indicated by R^2_{pred} = 0.858 for Model 1 (governed by log*AMW* and *nArOH*) and by R^2_{pred} = 0.865 for Model 2 (governed by t_2 and t_3). The applicability of the models to new micropollutants can be determined by Williams plots as has been demonstrated for the validation set. The predicted $\log k_{O3}$ values are indicative for the reactivity of micropollutant in ozonation. They can also be used when experimentally determined $\log k_{O3}$ values are not available for estimating micropollutant degradation by ozonation in natural waters with existing ozonation models.

Keywords: water treatment, piecewise linear regression, classification, linear discriminant analysis, model validation.

5.1 Introduction

Organic micropollutants such as endocrine disrupting chemicals (EDCs) and pharmaceuticals and personal care products (PPCPs) have been detected in surface water (Kolpin *et al.*, 2002) and even in finished drinking water (Benotti *et al.*, 2009; Huerta-Fontela *et al.*, 2011). Hence, there has been growing interest in determining removal efficiencies of drinking water treatment processes for these micropollutants, with recent studies showing that ozonation can be very effective in their degradation (von Gunten 2003; Westerhoff *et al.*, 2005).

Substantial removals can be achieved during ozonation, as oxidation of micropollutants occurs via molecular ozone (k_{O3}) and hydroxyl radicals (k_{OH}) produced by ozone decomposition (von Gunten 2003). Ozonation efficiency in natural waters can be predicted through the R_{ct} model which incorporates k_{O3} and k_{OH} (Elovitz and von Gunten 1999). Knowledge of these rate constants is therefore essential for assessing micropollutants reactivity and for R_{ct} model predictions. This paper focuses on k_{O3} whereas a companion paper deals with k_{OH} (Chapter 6). Although k_{O3} data are available for numerous micropollutants (von Gunten 2003; Hoigné and Bader 1983a; 1983b), there is a gap for emerging micropollutants such as EDCs and PPCPs. As it is impractical to experimentally determine k_{O3} for all micropollutants, predictive models for k_{O3} are of interest. One approach is to develop quantitative structure-property relationship (QSPR) models where the property to be predicted (i.e. k_{O3}) is correlated to chemical characteristics (i.e., molecular descriptors) using a group of experimentally studied micropollutants. The established model can then be applied to predict k_{O3} for untested compounds without experimentation (Eriksson *et al.*, 2003).

To date, only few QSPR studies have been published focusing on predicting k_{O3} of organic compounds in the aqueous phase (Benitez *et al.*, 2007; Hu *et al.*, 2000). Existing models are typically based on groups of structural similar compounds and often lack external validation, leading to poorly defined predictive power and limited applicability.

The objective of this study was therefore to develop a reliable QSPR model linking structural features to k_{O3} which is predictive for a wide range of structurally diverse micropollutants. This builds upon previous work (Jin and Peldszus 2012) using molecular descriptors selected based on current mechanistic knowledge, and a set of structurally diverse compounds selected by means of principal component analysis (PCA) and D-optimal onion design. For these compounds k_{O3} values were experimentally determined at neutral pH (Chapter 4). In this paper, these experimentally studied compounds were used as training set for developing QSPR models, and an external compound set

from the literature was used for model validation. The developed models were able to reliably predict k_{O3} for structurally diverse micropollutants and therefore may be used to screen organic micropollutants for their treatability with molecular ozone.

5.2 QSPR Model Development

The success of any QSPR model depends on the accuracy of the input data, the selection of the molecular descriptors, the statistical techniques used to develop the model, and its validation (Tropsha *et al.,* 2003).

5.2.1 Data Sets

The training set consisted of 22 micropollutants which were pre-selected from a large compound pool by a statistical approach that incorporated chemical characteristics relevant in ozone removal mechanisms (Jin and Peldszus 2012). Their k_{O3} values were then determined experimentally at pH 7 and room temperature (Chapter 4). Model validation used a second data set containing 33 micropollutants selected from the literature. The reaction rate constants for both sets (Appendix C Table C.1) are expressed in logarithm form ($\log k_{O3}$), and cover similar ranges (training set: -2.00 to 7.00; validation set: -1.40 to 6.43). Both data sets were comprised of compounds with a wide range of structural properties and very diverse applications (e.g. pharmaceutically active compounds, hormones, EDCs, pesticides, flame retardants, etc. (Appendix C Figures C.1 and C.2).

5.2.2 Molecular Descriptors and Data Preparation

The 12 molecular descriptors for all compounds in both data sets are given in Appendix C Table C.1. To achieve an approximate normal distribution, molecular weight (*MW*), average molecular weight (*AMW*) and diffusivity (*Df*) were log transformed. The discrete variables number of double bonds (*nDB*), number of aromatic bonds (*nAB*), number of phenolic groups (*nArOH*), and number of primary and secondary amines (*nN*) were converted into categorical variables with two levels (dichotomous) by coding which assigned values “1” and “0” to represent their presence and absence, respectively.

5.2.3 Statistical Analysis

QSPR models were developed using the training set, and then validated externally using the validation set. Four methods were tested to develop quantitative models: (1) step-wise multiple linear regression (MLR), (2) principal component regression (PCR), (3) partial least squares (PLS) regression, and (4) piecewise linear regression together with linear discriminant analysis (PLR-LDA). These statistical analyses used the software SIMCA-P (Umetrics), IBM SPSS (IBM Corp.), and STATISTICA (StatSoft. Inc.).

Briefly (details provided in Appendix C), MLR can work well if the molecular descriptors are independent from each other. PCR and PLS are projection based methods in which latent variables are extracted and then used as independent variables, reducing the impact of multicollinearity among the original variables. PLR works well with data showing piecewise linear features such as switching slopes at breakpoints. Local sub-models are determined using linear regression after meaningful breakpoints distinguishing between the local trends have been defined. Before applying PLR models for predictions it is necessary to determine the group membership of any new compound and forward stepwise LDA (Worth and Cronin 2003) was used to generate a classification function. The tolerance parameter (proportion of variance that is unique to the respective variable) was set as the default value (0.01) for minimum acceptable tolerance. Wilk's λ and the Mahalanobis distance were used to test the quality of the discriminant functions derived.

5.2.4 Model Validation and Accuracy

Statistical parameters such as correlation coefficient (R^2), and adjusted R^2 (R^2_{adj}) (details see Appendix C) indicate how well the model fits the training set but they are not a measure of the models predictive capability. Hence, internal and external validations were applied to test the predictive power (Eriksson *et al.*, 2006) of the developed QSPR models. The internal validation approach using a leave-one-out procedure provides a cross-validated Q^2.

$$Q^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum (y_i - \hat{y}_{i/i})^2}{\sum (y_i - \bar{y})^2} \qquad \textbf{(5.1)}$$

Where *PRESS:* predictive residual error sum of squares; *SST:* total sum of squares; y_i: observed dependent variable; $\hat{y}_{i/i}$: calculated dependent variable from a model developed without that data

point; $\bar{y}$: mean value of training set compound. The external validation provides a predictive R^2 (R^2_{Pred}), which is a measure of the predictive capacity for any new compounds.

$$R^2_{pred} = 1 - \frac{\sum \left(y_{validation} - \hat{y}_{validation}\right)^2}{\sum \left(y_{validation} - \bar{y}_{training}\right)^2} \quad \textbf{(5.2)}$$

Where: $\hat{y}_{validation}$ and $y_{validation}$ are predicted and observed values of validation set compounds; $\bar{y}_{training}$ is the mean value of training set compounds.

The root mean squared of errors for the training set (*RMSE*) and root mean squared of errors for the validation set (*RMSEP*) which summarize the overall error of the model were also calculated as indicators of the accuracy of the proposed models.

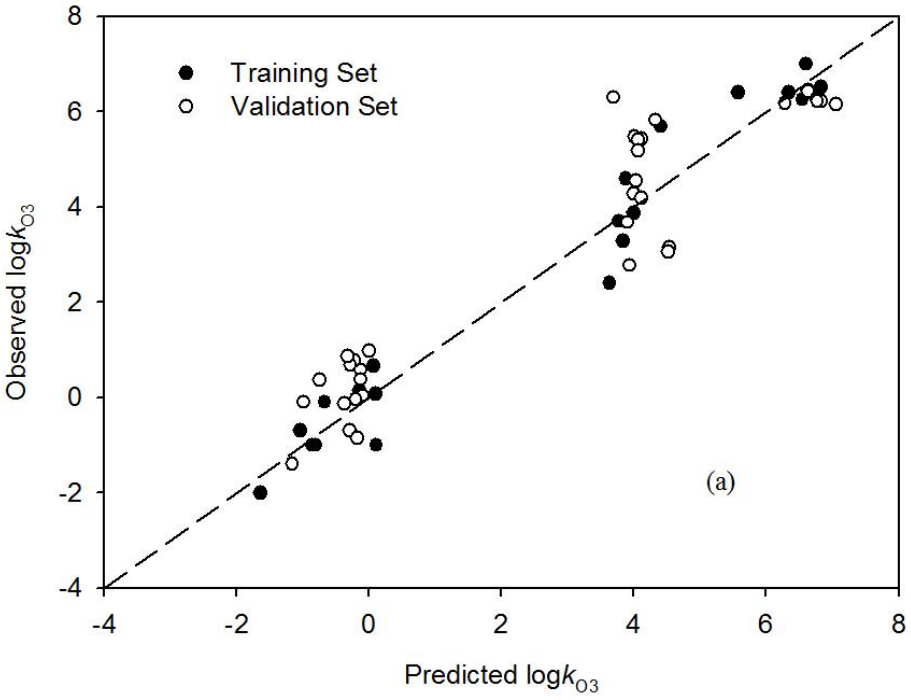
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}} \quad \textbf{(5.3)}$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2}{n_{ext}}} \quad \textbf{(5.4)}$$

Where n is the number of training set compounds and n_{ext} is the number of validation set compounds.

5.3 Results and Discussion

5.3.1 Preliminary Analysis by Stepwise MLR, PLS and PCR



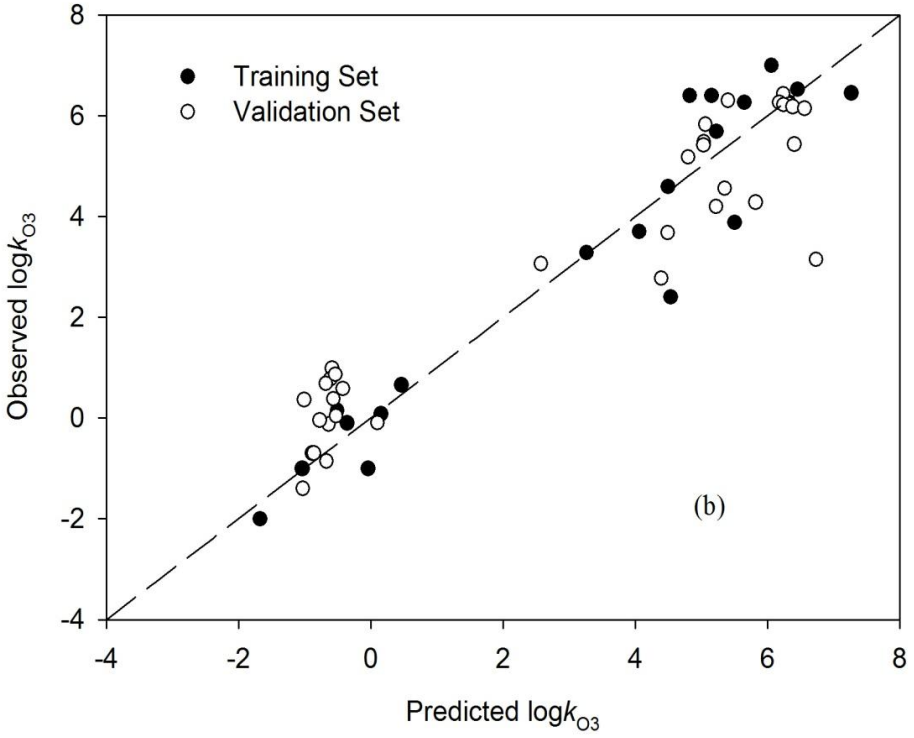


Figure 5.1 Plot of predicted logk_{O3} vs. observed logk_{O3}. Comparison of the results obtained by (a) PLR using molecular descriptors log*AMW* and *nArOH*; (b) PLR using principal components t_2 and t_3.

Several initial QSPR models (MLS, PLS, and PLR) were developed with a training set of 22 experimentally studied compounds and 12 descriptors. These were then validated with a separate validation set of 33 compounds selected from the literature. Initially, forward stepwise MLR was used to establish a preliminary QSPR model and to identify statistically significant molecular descriptors from the original group. The following MLR model was obtained.

$$\log k_{O3} = 6.863 - 5.818 \log AMW + 4.711 nArOH \qquad (\mathbf{5.5})$$

$n_{training}$ = 22, R^2 = 0.681, R^2_{adj} = 0.647, Q^2 = 0.590, $F(2,19)$ = 20.27 ($p < 0.0001$), *RMSE* = 1.763

$n_{validation}$ = 33, R^2_{pred} = 0.338, *RMSEP* = 2.170

Only two variables, average molecular weight in logarithm scale (log*AMW*) and number of phenolic groups (*nArOH*) were found to be significant. Phenolic groups are reactive with ozone and the positive coefficient indicates that rate constants increase with increasing *nArOH*. The negative coefficient of log*AMW* points to a decrease in rate constants with increasing log*AMW*. This decrease in reactivity may be linked to electron-withdrawing halogens which are present in higher proportions

in compounds with increased *logAMW*. Also, *nArOH* and log*AMW* are not highly correlated as the correlation coefficient is only -0.28 (Appendix C Table C.2). Note that number of aromatic bonds (*nAB*) would have been the next variable, but *nAB* (t-test, $p = 0.092$) was not significant at the 0.05 significance level. This initial model does not fit the training set data very well as shown by the relative low R^2 and large error ($R^2_{adj} = 0.647$, $RMSE = 1.763$). Further, when comparing the predicted to the measured rate constants (Appendix C Figure C.3) many compounds are either over- or underestimated. Two strategies were employed to test model validity. First, leave-one-out cross validation ($Q^2 = 0.618$) indicated a relatively robust model since values greater than 0.5 are considered a criterion for robustness (Fan *et al.*, 2001), although some disagree (Golbraikh and Tropsha 2002b). Second, external validation was performed using the validation set. The low predictive R^2 and large error ($R^2_{pred} = 0.338$, $RMSEP = 2.170$) indicate very poor predictive ability.

Next, an alternate model was developed by PLS regression starting with all 12 molecular descriptors. Descriptors with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value. The following equation was obtained with only 3 molecular descriptors remaining.

$$\log k_{O3} = 6.658 - 7.574 \log AMW + 2.390 nAB + 4.010 nArOH \qquad (\mathbf{5.6})$$

$n_{training} = 22$, $R^2 = 0.728$, $R^2_{adj} = 0.683$, $Q^2 = 0.662$, $F(3,18) = 16.10$ ($p < 0.0001$), $RMSE = 1.626$

$n_{validation} = 33$, $R^2_{pred} = 0.306$, $RMSEP = 2.222$

In this case, R^2_{adj} and its internal validation Q^2 were both slightly better than for the MLR model. Compared to MLR, the variable *nAB* is included in the equation with a positive coefficient meaning that increased presence of aromatic double bonds will increase reactivity. But again the predictive ability ($R^2_{pred} = 0.306$, $RMSEP = 2.222$) with the validation set was poor and the observed and predicted values of $\log k_{O3}$ showed severe over- and under-prediction (Appendix C Figure C.4).

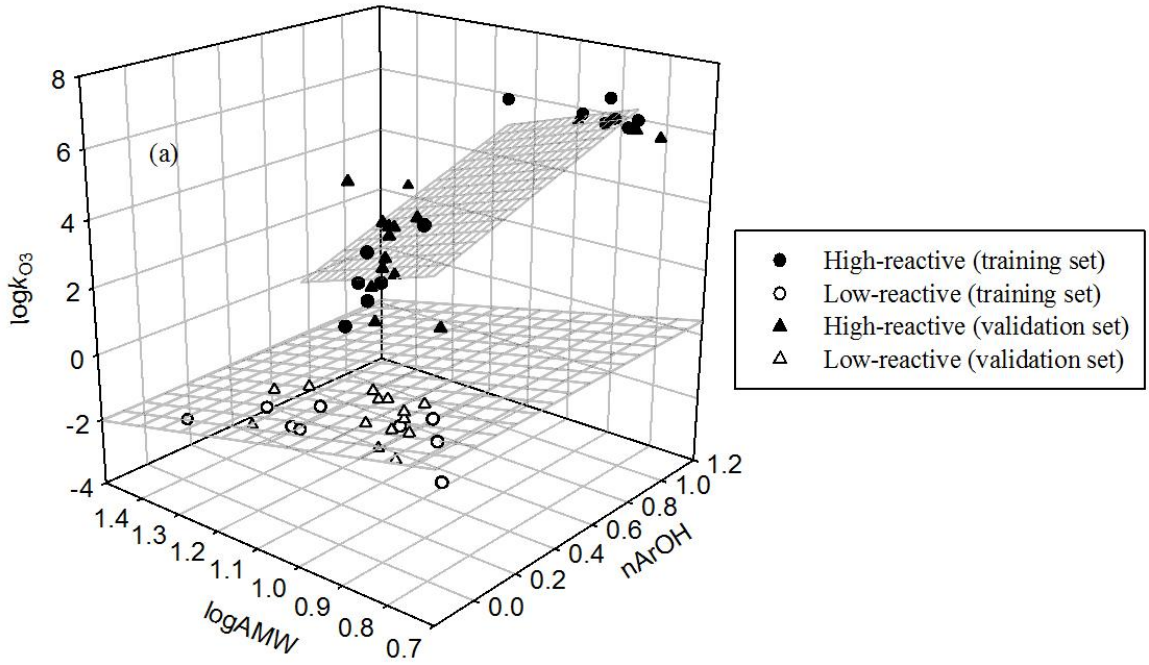
Since the performance of these models proved to be insufficient, a PCR model was considered. Before regression, the *X*-matrix (22 training set compounds × 12 molecular descriptors) was analyzed by PCA, resulting in the extraction of three significant principal components (PCs), explaining 75.9% (35.6% by PC1, 25.8 by PC2, and 14.5% by PC3) of the variance. Loading and score plots are provided in Appendix C Figures C.5 and C.6. Next, with the extracted PCs as independent variables and using stepwise MLR the following was obtained.

$$\log k_{O3} = 2.906 + 1.033t_2 - 0.878t_3 \quad (\mathbf{5.7})$$

$n_{training}$ = 22, R^2 = 0.455, R^2_{adj} = 0.398, Q^2 = 0.298, $F(2,19)$ = 7.924 (p = 0.001), *RMSE* = 2.305

$n_{validation}$ = 33, R^2_{pred} = 0.100, *RMSEP* = 2.530

Where t_2, t_3 represent PC2 and PC3, respectively. Although PC1 could explain 35.6% of the variation in the *X*-matrix, it was found to be insignificant. This is not surprising since PC1 is dominated by variables such as unsaturation index (*Ui*) and *HOMO-LUMO* gap (*GaP*) (Appendix C Figure C.5) which were insignificant in the previous models. The important variables n*ArOH* and log*AMW* are mainly coded into PC2 and PC3, respectively. The resulting PCR model could only explain 45.5% variation with large prediction error (*RMSE* = 2.305) and low Q^2 (0.368), so it is not surprising that it failed to predict the validation set (R^2_{pred} = 0.100, *RMSEP* = 2.530). The predicted versus measured logk_{O3} values were plotted in Appendix C Figure C.7.



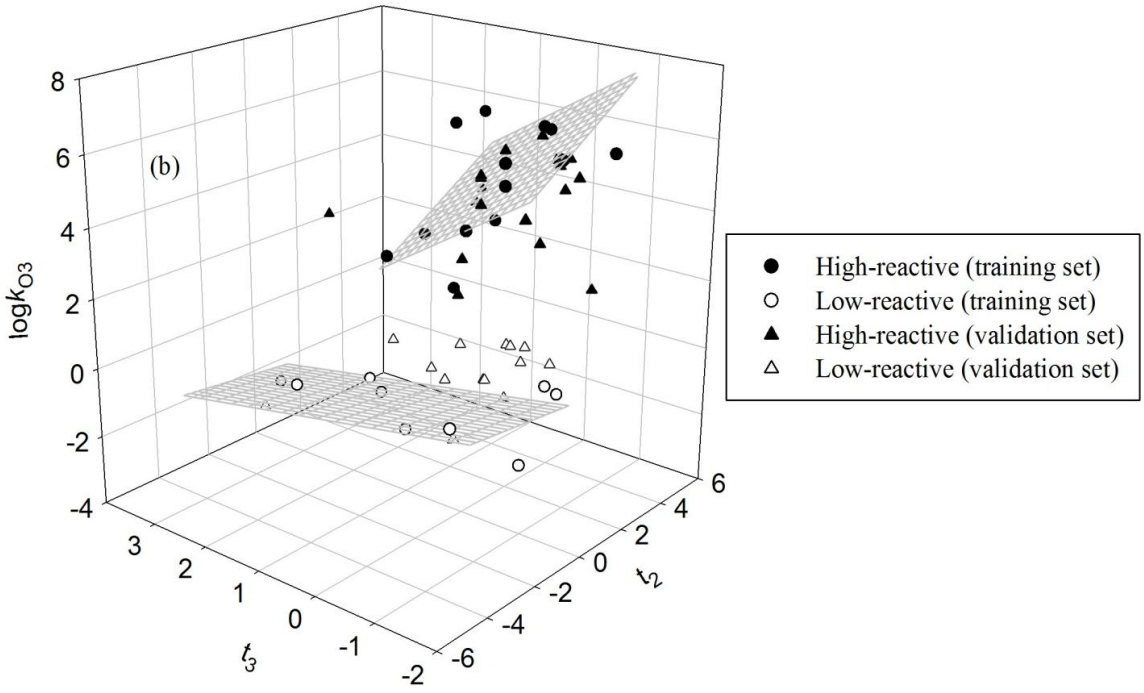


Figure 5.2 3D plot of QSPR models, (a) Model 1: PLR with *nArOH* and log*AMW*; (b) Model 2: PLR with t_2 and t_3.

5.3.2 Reassessment Using PLR-LDA Approach

5.3.2.1 Modeling of the Rate Constants Using PLR

As discussed, QSPR models developed by stepwise MLR, PLS and PCR showed poor to moderate fitting of the data and they failed to adequately predict the external validation set. This also confirms that external validation is essential in assessing the predictive power of QSPR models, and that internal validation alone is not sufficient to ensure the QSPR models have predictive power (Golbraikh and Tropsha 2002a). However, in plots of predicted versus observed values (Appendix C Figures C.3 and C.4) it can be noted that most of the compounds with low observed rate constants are overestimated and many compounds with high observed rate constants are underestimated. This indicates a breakpoint, differentiating compounds with high and low rate constants, and suggests that different mechanisms may be predominant. For example, the reactivity towards ozone may be governed by different molecular properties for low-reactive and high-reactive compounds; or by certain ozone-reactive functional groups which are present in high-reactive compounds and absent in low-reactive compounds. In such situations, it is possible to fit piecewise linear regression (PLR)

models which separate the data into several groups (two groups in this study), followed by fitting a linear sub-model to each.

The challenge of applying PLR is to find a meaningful breakpoint which splits the dataset into subsets. In complex cases, quasi-Newton algorithms can be used to search for a breakpoint or multiple breakpoints (Molina *et al.,* 2008). Here, a clear distinction between under- and overestimation is apparent at $\log k_{O3} = 2.00$ ($k_{O3} = 100$ $M^{-1}s^{-1}$) which in water treatment practice represents 5 minutes half-life time at 1 mg/L of ozone exposure (Appendix C Figures C.3 and C.4). Using a breakpoint at $\log k_{O3} = 2.00$ creates two subclasses: $\log k_{O3} \geq 2.00$ (i.e., $k_{O3} \geq 100$ $M^{-1}s^{-1}$) are classified as high-reactive compounds (labeled as "1"), and those with a $\log k_{O3} < 2.00$ (i.e., $k_{O3} < 100$ $M^{-1}s^{-1}$) are classified as low-reactive compounds (labeled as "-1").

Next, two PLR models were developed each using a different set of molecular descriptors. Model 1 includes log*AMW* and *nArOH*, the descriptors selected by the stepwise MLR algorithm. Model 2 uses the two significant principal components (t_2 and t_3) identified by PCR. Each model consists of two linear regression sub-models, one describing the rate constant of high-reactive compounds, designated by $\log k_{O3}(\geq 2)$, and the other describing those for low-reactive compounds, designated by $\log k_{O3}(< 2)$.

Model 1 was established using log*AMW* and *nArOH*:

$$\begin{aligned} \log k_{O3}(<2) &= 2.327 - 2.876 \log AMW \\ \log k_{O3}(\geq 2) &= 7.747 - 4.171 \log AMW + 2.382 nArOH \end{aligned} \qquad (\mathbf{5.8})$$

$n_{training} = 22$, $R^2 = 0.964$, $R^2_{adj} = 0.960$, $F(2,19) = 257.3$ ($p < 0.00001$), $RMSE = 0.589$

$n_{validation} = 33$, $R^2_{pred} = 0.858$, $RMSEP = 0.978$

Note that the *nArOH* parameter is missing in the first equation, simply because none of the slow reacting compounds in this group contain a phenolic group. The statistical parameters above correspond to the regression obtained by the combination of both sub-models, i.e., the statistical parameters were obtained by fitting the observed values to those predicted by both pieces of the model.

Model 2 was obtained using principal components:

$$\log k_{O3}(<2) = -0.534 - 0.097t_2 - 0.311t_3$$
$$\log k_{O3}(\geq 2) = 4.612 + 0.486t_2 - 1.158t_3 \quad \textbf{(5.9)}$$

$n_{training}$ = 22, R^2 = 0.929, R^2_{adj} = 0.922, $F(2,19)$ = 137.4 (p < 0.00001), *RMSE* = 0.830

$n_{validation}$ = 33, R^2_{pred} = 0.865, *RMSEP* = 1.057

Both models are very good in terms of fit (R^2_{adj} > 0.92) and external validation (R^2_{pred} > 0.84) i.e. excellent predictive ability. The observed logk_{O3} versus predicted logk_{O3} are shown in Figure 5.1. The PLR Model 1 with log*AMW* and *nArOH* fits the training set slightly better than Model 2 with principal components t_2 and t_3 in terms of the R^2_{adj} value, but Model 2 has a slightly higher predictive ability in terms of the R^2_{pred}. Figure 5.2 shows the 3-dimensional PLR models where the models are fit to 2 dimensional planes. It clearly shows the change in slope for low-reactive and high-reactive compounds, indicating that different mechanisms may be dominant for different groups of compounds.

Model 2 with principal components t_2 and t_3 was considered better than Model 1 with log*AMW* and *nArOH*. First, the primary objective of QSPR modeling is to predict the rate constants of new compounds, therefore models with better predictive power are preferred. Second, Model 2 combines partial contributions from several molecular descriptors thus considering multiple molecular properties. Last, Model 1 is strongly influenced by phenolic groups and may overemphasize their importance. High-reactive compounds cluster in two groups as a function of the number of phenolic groups in Model 1 (Figure 5.1a). In the absence of phenolic groups a near vertical trend was found for high-reactive compounds (i.e. at a predicted logk_{O3} value of approximate 4) when plotting observed vs. predicted logk_{O3} values, because of the small variation of log*AMW*. However, compounds without a phenolic group can still be very reactive, as for example, carbamazepine due to the presence of double bonds (logk_{O3} = 5.48 (Huber *et al.*, 2003)), where its rate constant was predicted with a larger error by Model 1 (predicted logk_{O3} = 4.01) than by Model 2 (predicted logk_{O3} = 5.04).

5.3.2.2 Classification Using LDA

To carry out the prediction using PLR models for new compounds, it is necessary to classify them into either high-reactive or low-reactive compounds as this determines which of the sub-models should be applied.

The group membership of a compound was determined by a Canonical discriminant function. If this function gives a value of Class ≥ 0 the compound was classified as high-reactive ($\log k_{O3} \geq 2.00$), and if Class < 0 it was classified as a low-reactive compound ($\log k_{O3} < 2.00$). A pool of molecular descriptors containing a multitude of descriptors calculated by DRAGON (Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy) was used to obtain the classification function by step-wise linear LDA. The Canonical discriminant function found is given below:

$$Class = 28.98 - 50.21 \times X1A - 1.92 \times ATS3m + 2.12 \times MATS6p + 4.60 \times GATS2v - 1.79 \times GATS3v \quad (\mathbf{5.10})$$

$n_{training}$ = 22, Canonical R = 0.91, Wilks' λ = 0.17, $F(5,16)$ = 15.6 ($p < 0.0001$), D^2 = 18.33.

Where *X*1*A*: average connectivity index chi-1; *ATS*3*m*: Broto-Moreau autocorrelation of a topological structure - lag 3/weighted by atomic masses; *MATS*6*p*: Moran autocorrelation - lag 6/weighted by atomic polarizabilities; *GATS*2*v*: Geary autocorrelation - lag2/ weighted by atomic van der Waals volumes; *GATS*3*v*: Geary autocorrelation - lag 3/weighted by atomic van der Waals volumes. Detailed explanations of these descriptors can be found elsewhere (Todeschini and Consonni 2000).

The classification function correctly grouped the training set into either high- or low-reactive compounds (Table 5.1), having a Wilks' λ of 0.17 indicating very good discrimination. The fairly large value of the squared Mahalanobis distance between the two group centroids (D^2 = 18.33) indicates significant separation. Further, the canonical correlation R of 0.91 (normal range: 0 to 1), measuring the association between the groups and the given discriminant function, indicates a high correlation. Subsequently the discriminant function was applied to the external validation set, correctly classifying 89.5% (17/19) of the high-reactive compounds and 85.7% (12/14) of the low-reactive compounds (Table 5.1).

Table 5.1 Classification Results

	Actual group membership		Predicted group membership			
		Total Count	-1 (low-reactive) Count	(%)	1 (high-reactive) Count	(%)
Training Set	-1 (low-reactive)	9	9	(100)	0	(0)
	1 (high-reactive)	13	0	(0)	13	(100)
Validation Set	-1 (low-reactive)	14	12	(85.7)	2 [a]	(14.3)
	1 (high-reactive)	19	2 [b]	(10.5)	17	(89.5)

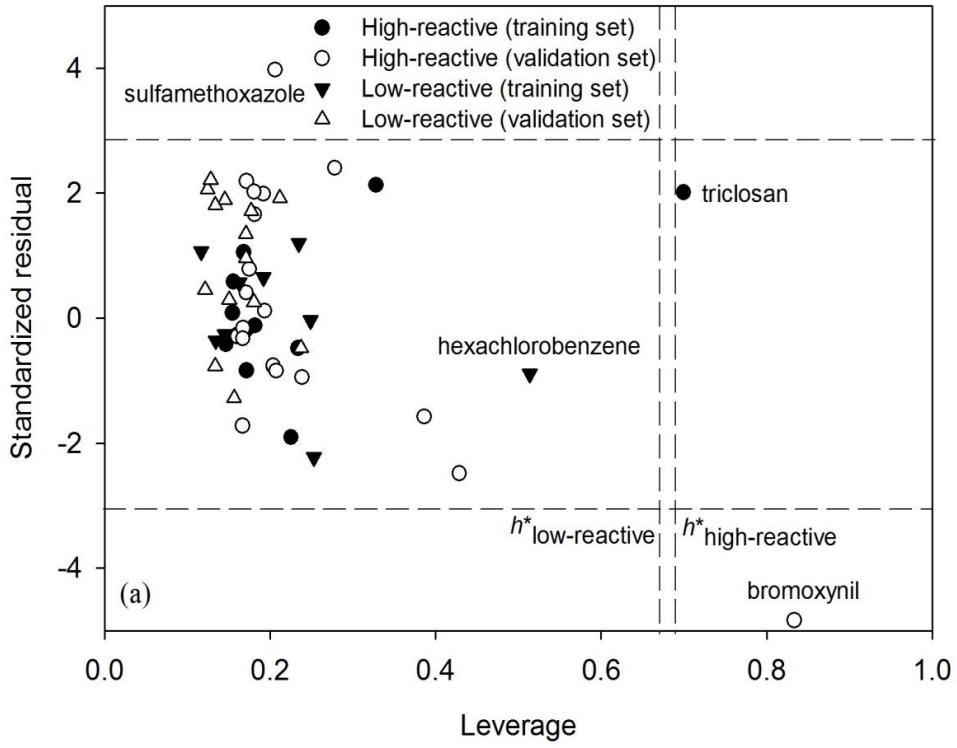
[a] wrong cases for -1 group (low-reactive) are diazepam and propachlor; [b] wrong cases for +1 group (high-reactive) are carbamazepine and trimethoprim. The calculation details can be found in Appendix C Table C. 3.

5.3.2.3 Applicability Domain

The applicability domain is the chemical space defined by the properties of the training set. Predictions for new compounds falling within this space are expected to be reliable since their properties are close to those used to establish the PLR model. Several methods are available for defining the applicability domain of QSPR models (Netzeva *et al.,* 2005). The most common method is to determine the leverage of each compound ($h_i = x_i^T(X^TX)^{-1}x_i$, where x_i is the descriptor vector of the considered compound, and X is the descriptor matrix, and then plot standardized residuals versus leverages for each compound, i.e. Williams plot. The applicability domain is established by a squared area within ± 3 standard deviations and a leverage threshold h^* ($h^* = 3p/n$, where p is the number of model variables plus one, and n the number of training set compounds). Thus, compounds with standardized residuals > 3 standard deviation units and h_i > leverage threshold h^* are considered as outliers. However, a high leverage training set compound with small residual is not necessary an outlier (Gramatica *et al.,* 2004).

As shown in the Williams plots (Figure 5.3), all training set compounds are inside the square area for Model 1 and Model 2, except for triclosan in Model 1. Its highest average molecular weight (*AMW* = 12.06) places it far from the centroid of the descriptor space. However, its residual is relatively small (2.02), thus it stabilizes the model and makes the model more precise. There are no outliers for the training set of the QSPR models. However, in the validation set 2 compounds (bromoxynil and sulfamethoxazole) were identified outside of the applicability domain for Model 1

(high-reactive group), and 3 compounds (bromoxynil, lincomycin, and metoprolol) were identified for Model 2 (high-reactive group). Bromoxynil is structurally anomalous because of its bromine substitution, and lincomycin is the only one without an aromatic structure in the high-reactive group (Appendix C Figure C.2). In Model 2, the leverage values of bromoxynil and lincomycin exceed *h** but standardized residuals remain acceptable. This indicates that even at high leverage values, predictions for these compounds fell within the model's expected uncertainty range. While this demonstrates the model's applicability to a wide range of structurally diverse compounds, predictions for compounds falling outside the applicability domain should be used with caution.



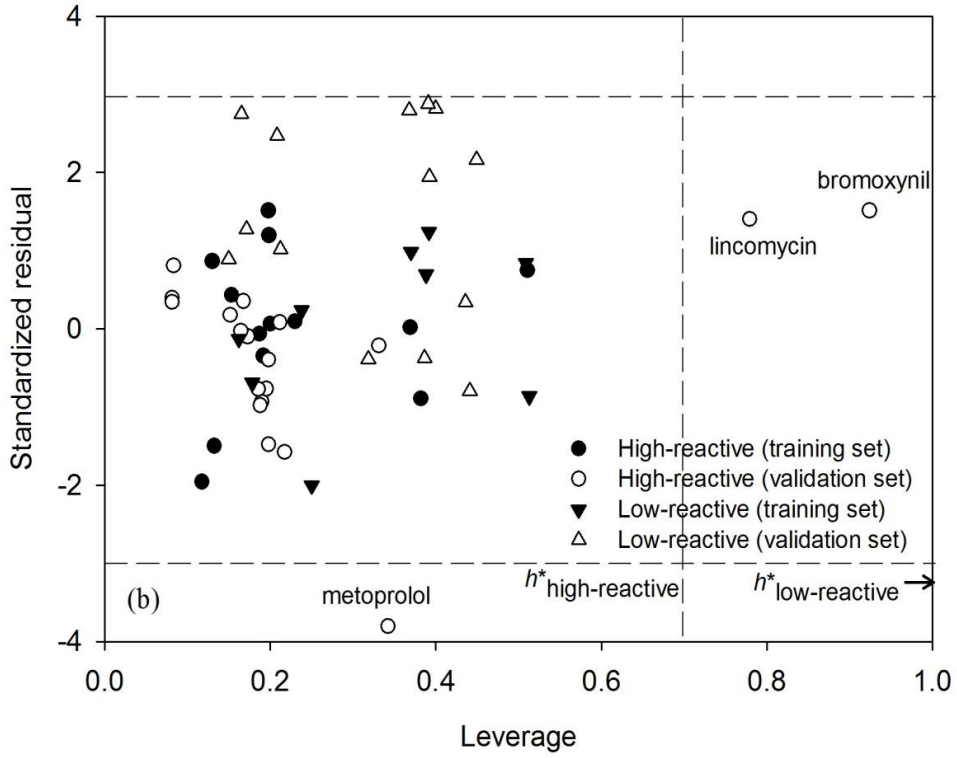


Figure 5.3 Williams plot showing the application domain of QSPR models, (a) Model 1; and (b) Model 2.

5.3.3 The PLR-LDA Models in Ozonation Practice

The PLR-LDA models are useful to water treatment engineers, researchers and regulators to predict the reactivity of micropollutants, and determine their remaining concentration after ozone exposure. Further, when pre-screening suitable water treatment technologies for degradation of a particular micropollutant, it may be sufficient to simply apply the classification function to assess whether ozonation is appropriate.

The application of the PLR-LDA QSPR models for predicting a new compound's ozone rate constant is described in Appendix C Figure C.8. When applying these (or any other) models, the user should be aware of the associated errors, which include the classification and model prediction errors. If the classification function fails to correctly identify a compound as high-reactive or low-reactive, then a large prediction error would be expected. This will only be the case for a small fraction of the compounds however, and with identification of chemical substructures favoured by ozone attack (e.g.,

activated aromatics system, double bond) classification error can be minimized. The error associated with model predictions, assessed by *RMSEP*, is close to one for both models. This value is acceptable considering the models were developed using compounds with very diverse structures and are therefore applicable to many different compounds.

One of the limitations of the developed models is the uncertainty for dissociating compounds with a dissociation constant (p*K*a) around 7. For dissociating compounds, k_{O3} is pH-dependent and the normal approach is to predict the specific k_{O3} for the neutral and ionic species separately, then calculate the apparent k_{O3} at the desired pH based on the p*K*a (Canonica and Tratnyek 2003). However, QSPR models successful for predicting k_{O3} of ionic species have not been reported to-date. In this study, an approximation was used to estimate the apparent k_{O3} at pH 7 which is relevant in drinking water treatment. The dominant species at pH 7 were used for calculating molecular descriptors, i.e. either the neutral or the ionic form, and the experimentally determined apparent k_{O3} values at pH 7 were used as dependent variables. This approach assumes that the contributions from non-dominating species are minimal, which holds true if the p*K*a differs by at least one unit from 7. As shown in Appendix C Table C.1, none of the micropollutants in the training set and only 5 out of 33 in the validation set have p*K*a values close to 7 (lincomycin, sulfamethoxazole, amoxicillin, trimethoprim, and enrofloxacin). But predicted values were close to reported values for these compounds (except sulfamethoxazole predicted by Model 1). Nevertheless, caution needs to be exercised in the interpretation of predicted k_{O3} values for compounds with a p*K*a close to 7.

Chapter 6
QSPR Modeling for the Hydroxyl Radical Reaction Rate Constants of Organic Micropollutants in Aqueous Phase

This Chapter is in paper format which will be revised accordingly and submitted to a peer reviewed journal.

This Chapter focuses on developing quantitative structure-property relationship models for predicting the rate constants of diverse micropollutants in their reaction with hydroxyl radicals. Initially, QSPR models were developed with the 22 training set compounds selected from a large compound pool and 12 molecular descriptors as described in Chapter 3. Their hydroxyl radical rate constants were determined by experimental analysis (Chapter 4). However, an unsatisfactory QSPR model was obtained (Appendix D). Therefore, the modeling approach was revised to a conventional QSPR approach using a large number of compounds (collected from the literature) and a large number of DRAGON descriptors from which the best subset descriptors were selected. A satisfactory empirical predictive QSPR model was developed and externally validated. Together with the QSPR model for the reaction with molecular ozone (Chapter 5), we were able to predict both oxidation pathways (direct oxidation with molecular ozone, and indirect oxidation with hydroxyl radicals), and assess the percent removal of various contaminants in natural water during ozonation (Chapter 7).

Outline: Quantitative structure-property relationship (QSPR) models which predict hydroxyl radical rate constants (k_{OH}) in the aqueous phase for a wide range of micropollutants especially EDCs and PPCPs are needed to assess the removal efficiencies of advanced oxidation processes. QSPR models for the prediction of k_{OH} were developed with special attention to model validation, applicability domain and mechanistic interpretation. In this study, 118 compounds including those experimentally determined by the author and literature data were collected and randomly divided into the training set (n = 89) and the validation set (n = 29). The QSPR model was calibrated using the training set and multiple linear regression (forward selection) was applied. Seven DRAGON descriptors were found to be important in predicting the k_{OH} values which related to the electronegativity, polarizability, and double bonds, etc. The model fits the training set very well as indicated by the high R^2 value (R^2_{adj} = 0.823) and the low prediction error *RMSE* (0.204). A high Q^2 (0.773) was obtained indicating good robustness and good internal predictivity. The model was then

externally validated with the validation set showing good predictive power (R^2_{pred} = 0.772). The applicability domain of this model was then assessed using the Williams plot and two outlier compounds were identified. The QSPR model was then further improved by removing these two outlier compounds form the original model. Overall, the developed QSPR model provides a valuable tool for assessing the removal efficiency of micropollutants by AOPs.

Keywords: molecular descriptors, reaction rate constant, external validation, applicability domain, outlier detection

6.1 Introduction

Micropollutants such as endocrine disrupting chemicals (EDCs) and pharmaceutical and personal care products (PPCPs) create unique challenges to water treatment because of the number of compounds detected and the diversity and complexity of their physico-chemical properties. The efficiency of drinking water treatment processes for the removal of micropollutants from drinking water has been of concern to water utilities and environmental agencies. Advanced oxidation processes (AOPs) such as O_3/H_2O_2, UV/H_2O_2, UV/TiO_2 produce a highly reactive oxidant, the hydroxyl radical, which reacts rapidly with most organic micropollutants and leads to their degradation (Huber *et al.*, 2003). To investigate the removal efficiency of various organic micropollutants during AOPs in natural waters, it is necessary to obtain the reaction rate constants of micropollutants for their reaction with hydroxyl radicals (k_{OH}). Rate constants are valuable when predicting the extent to which the original contaminants are eliminated from water, and they are therefore important for designing and optimizing treatment processes. Although kinetic data are available for a large number of chemicals for their reactions with hydroxyl radicals (Buxton *et al.*, 1988), there is still a data gap especially for emerging micropollutants such as EDCs and PPCPs.

Due to the complexity of the analytical methods and the high cost associated with the determination of reaction rate constants, it is highly desirable and cost-effective to develop a reliable model to predict the rate constants of numerous micropollutants. Quantitative structure-property relationships (QSPR) have been widely used as a modeling tool to develop relationships between the properties (e.g., p*K*a) of chemicals and their structural characteristics (Eriksson *et al.*, 2003). QSPR models can relate the physico-chemical characteristics of compounds to their properties relevant in water treatment processes (e.g., removal, adsorption, and rejection), providing improved knowledge on removal mechanisms and interactions between organic compounds and physical/chemical treatment processes.

To date, only a small number of QSPR studies have been published focusing on predicting the reaction rate constants of organic compounds with hydroxyl radicals in the aqueous phase, but the applicability of these models is limited. For example, a QSPR model was developed using linear regression to predict k_{OH} of aromatic compounds in the aqueous phase (Kusic *et al.*, 2009), but this model is not applicable to non-aromatic compounds. Neural networks were applied to correlate functional groups and the k_{OH} values of a great variety of organic compounds (Dutot *et al.*, 2003). However, compounds used as training set were mostly conventional, small micropollutants using data

from Buxton *et al.*, (1988), and only a few micropollutants especially EDCs, PPCPs, or pesticides were utilized in model development. The group contribution method has also been used to predict aqueous phase k_{OH} values for compounds with a wide range of functional groups (Monod and Doussin 2008; Minakata *et al.*, 2009). However, certain assumptions such as availability of data for all possible functional groups and additivity of rate constants limit the use of the group contribution method (Minakata *et al.*, 2009).

The objective of this study was therefore to develop a robust, validated QSPR model for predicting the aqueous phase k_{OH} of a wide range of micropollutants. A large number of micropollutants with diverse structures including many EDCs and PPCPs were collected for model development. The data set was then split into training and validation sets, and the training set was used to calibrate the model which was then externally validated using the validation set. In addition, the applicability domain of the model was defined by a leverage approach so that the applicability of the model to a new, unknown compound can be determined. This overall approach ensured that the developed models were applicable to micropollutants with diverse structures and a wide range of k_{OH}, and they will therefore be helpful in assessing the efficiency of AOPs technologies with respect to the degradation of micropollutants.

6.2 Materials and Methods

6.2.1 Data Set

A total of 118 micropollutants were used for developing the QSPR models in this study, in which k_{OH} values of 22 micropollutants were determined experimentally in a previous study using competition kinetics (Chapter 4), and the other 96 micropollutants were collected from the literature. Micropollutants included in this study were very heterogeneous in structure and included a number of chemical classes (e.g., phenols, polycyclic aromatic hydrocarbons, alkanes, halogenated aromatic compounds, organophosphorus compounds, etc.) thus covering a wide spectrum of physico-chemical properties. A list of the micropollutants included and their k_{OH} values are provided in Table 6.1. The k_{OH} values range from 5.4×10^{7} (M^{-1} s^{-1}) to 1.7×10^{10} (M^{-1} s^{-1}). The total compound set was divided into a training set and a validation set through property sampling as described in Leonard and Roy (2006). This was accomplished by ordering the micropollutants according to their descending k_{OH} values, taking every fourth compound from the set to be used as the validation set, and the remaining

compounds used as the training set. As a result, about 25% of the total data set was used for the validation set (n = 89 for the training set, and n = 29 for the validation set).

A large number of different molecular descriptors were calculated using DRAGON software, and these were then used as independent variables for modeling. The chemical name or registration number was used to search the SMILES code of the chemical structure from the ChemIDplus Advanced online database (United States National Library of Medicine). The SMILES code of the chemical structure was then used as input for the software DRAGON (Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy) to generate the molecular descriptors. As a result, 951 descriptors including constitutional descriptors, topological descriptors, connectivity indices, information indices, 2D autocorrelations, eigenvalue-based indices, 3D MoRSE descriptors, WHIM descriptors, molecular properties, functional group counts, and atom-centered fragments, etc. were calculated. A list of the DRAGON descriptors is available (Todeschini *et al.*, 2005). Most of these descriptors are reviewed in a textbook by Todeschini and Consonni (2000). The correctness of the SMILES code was then validated by comparison of the molecular weights reported in the databases with those calculated by the software. To minimize the redundant information, descriptors with constant values among micropollutants (n = 142, mostly functional group count descriptors) were removed, and descriptors found to be pairwise correlated by greater than 95% (n = 110) were excluded.

6.2.2 QSPR Modeling

Multiple linear regression (MLR) was used in this study to identify a linear relationship between k_{OH} and a set of molecular descriptors. MLR is among the most widely used modeling methods in QSPR studies, which models a dependent variable (property to be predicted) as a linear combination of independent variables (molecular descriptors) with regression coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \quad \textbf{(6.1)}$$

Where β_1, β_2, … β_p are regression coefficients and β_0 is a constant, x_{i1}, x_{i2}, … x_{ip} are molecular descriptors of the i'th compound, y_i is the property to be predicted, and ε_i represents the residuals.

As there are a large number of DRAGON descriptors, the forward selection method was used to screen the descriptors. The modeling process starts without any descriptors in the model; the descriptors are then tested one by one to find the descriptor that achieves the best fit, i.e., the largest

R^2 value when added to the model. This descriptor is then added to the model. The procedure continues to find the second descriptor to be added to the model in the same manner, and so on. This procedure terminates when no variable meets the inclusion criterion, or when the available improvement falls below some critical value (Andre *et al.*, 2003).

Data analysis and modeling were carried out using the software NCSS 2007 (NCSS, Kaysville, Utah, US). Before modeling, k_{OH} was transformed to its decadic logarithm ($\log k_{OH}$). The discrete molecular descriptors, such as the functional groups counts, and atom-centered fragments were converted to categorical variable with two categories ("0" represents absence and "1" represents presence). The other descriptors were used with no transformation.

6.2.3 Model Evaluation

The model was not only evaluated by the goodness-of-fit to the training set, but also verified with respect to its internal and external predictive performance (Tropsha *et al.*, 2003; Gramatica 2007). The model fit was assessed using the adjusted coefficient of determination (R^2_{adj}). The internal predictivity of QSPR models was assessed by the leave-one-out cross-validated correlation coefficient (Q^2). For the external predictivity, the external validation R^2_{pred} parameter was calculated.

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1} \qquad \textbf{(6.2)}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_{i/i} - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad \textbf{(6.3)}$$

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{n_{ext}}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n_{ext}}\left(y_i - \bar{y}_{tr}\right)^2} \qquad \textbf{(6.4)}$$

Where R^2 is the coefficient of determination, n is the number of training set compounds, and p is the number of descriptors involved in the model; Q^2 is calculated using the training set, where y_i is the measured $\log k_{OH}$ values, $\hat{y}_i$ is the predicted $\log k_{OH}$ values, $\bar{y}$ is the mean value of training set compounds, $\hat{y}_{i/i}$ is the predicted value of the response calculated excluding the i^{th} compound from

the model computation; R^2_{pred} is calculated with the validation set, but $\bar{y}_{tr}$ is the mean value of the training set compounds.

The root mean squared of errors for the training set (*RMSE*) and root mean squared of errors for the validation set (*RMSEP*), which summarize the overall error of the model, were also calculated as indicators of the accuracy of the proposed models (Gramatica and Papa 2005).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad \textbf{(6.5)}$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2}{n_{ext}}} \quad \textbf{(6.6)}$$

The variance inflation factor (*VIF*) is a measure of multicollinearity. A *VIF* of 10 or more indicates that multicollinearity is a problem in the data set (Roy and Roy 2009).

$$VIF = \frac{1}{1 - R_j^2} \quad \textbf{(6.7)}$$

Where R_j^2 is the unadjusted R^2 when the j^{th} variable is regressed against all the other variables in the model.

The applicability domain is the chemical space defined by the properties of the training set. Predictions for new compounds falling within this space are expected to be reliable since their properties are close to those used to establish the model. The applicability domain of the QSPR model is visualized by plotting the standardized residuals versus the leverage (Kusic *et al.*, 2009). Leverage indicates the compound's distance from the centroid of compound space. The leverage of a compound is defined as:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \quad \textbf{(6.8)}$$

Where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as:

$$h^* = 3p/n \quad \textbf{(6.9)}$$

Where *n* is the number of training set compounds, and *p* is the number of descriptors in the model plus one.

6.3 Results and Discussion

6.3.1 QSPR Modeling and Validation

A preliminary analysis was conducted with a data of 22 selected micropollutants (k_{OH} values were determined experimentally in Chapter 4) and 12 selected molecular descriptors (details are shown in Appendix D). It was found that the model developed with these selected micropollutants and descriptors failed to develop a satisfactory predictive QSPR model, and additional compounds and better molecular descriptors were needed to improve the QSPR model. Therefore, in this study, a conventional QSPR approach with a large number of compounds (118 micropollutants collected from literature including those experimental determined micropollutants shown in Chapter 4) and many new descriptors (951 DRAGON descriptors) was applied. The best subset of descriptors which can capture the structural features related to the hydroxyl radical reactions was selected by forward MLR. In this study, the prior knowledge in the descriptor selection was not considered therefore an empirical model was developed.

As a result of the MLR (forward selection), the following 7-variable model with the highest R^2 value (Model 1: Equation 6.10) was established.

$$\log k_{OH} = 17.215 - 7.564 \times Me + 0.160 \times nDB - 0.625 \times CH2RX + 0.310 \times nHAcc - 0.563 \times Vindex - 0.362 \times MATS2m - 0.427 \times Mor27p \quad \textbf{(6.10)}$$

$n_{\text{training}} = 89$, $R^2 = 0.837$, $R^2_{adj} = 0.823$, $Q^2 = 0.773$, $F(7, 81) = 59.435$ ($p < 0.0001$), $RMSE = 0.204$

$n_{\text{validation}} = 29$, $R^2_{pred} = 0.772$, $RMSEP = 0.329$

where *Me* is the mean atomic Sanderson electronegativity, *nDB* is the number of double bonds, *nCH*2*RX* is the number of CH2RX (primary alkyl halides) functional group, *nHAcc* is the number of acceptor atoms for H-bonds (N, O, F), *MATS*2*m* is the Moran autocorrelation of lag 2 weighted by mass, *Vindex* is the Balaban V index, and *Mor*27*p* is signal 27/weighted by polarizability. The detailed explanation of these descriptors can be found elsewhere (Todeschini and Consonni 2000). The mechanistic interpretation of these descriptors will be discussed further in section 6.3.3.

Table 6.1 Compounds used for the QSPR modeling.

No.	Compound	k_{OH} ($M^{-1}s^{-1}$)	*Me*	*nDB*	*nCH2RX*	*nHAcc*	*Vindex*	*MATS2m*	*Mor27p*	Reference
1	Bezafibrate	8.00×10^{9}	1.01	2	0	5	0.225	-0.048	-0.192	1
2	DEET	4.95×10^{9}	0.98	1	0	2	0.471	0.199	-0.192	2
3	Atenolol	7.05×10^{9}	1	1	0	5	0.328	-0.13	-0.056	3
4	Metoprolol*	8.39×10^{9}	0.99	0	0	4	0.322	-0.191	-0.038	3
5	Propranolol	1.07×10^{10}	0.99	0	0	3	0.277	-0.154	-0.152	3
6	Penicillin G	7.97×10^{9}	1.02	3	0	6	0.237	-0.127	-0.127	4
7	Penicillin V	8.76×10^{9}	1.02	3	0	7	0.222	-0.16	-0.172	4
8	Amoxicillin*	6.94×10^{9}	1.02	3	0	8	0.235	-0.139	0.087	4
9	Levofloxacin	7.60×10^{9}	1.02	3	0	8	0.242	-0.061	-0.099	5
10	Lomefloxacin*	8.04×10^{9}	1.03	3	0	8	0.278	-0.071	-0.12	5
11	Norfloxacin	6.61×10^{9}	1.02	3	0	7	0.274	-0.036	-0.195	5
12	Orbifloxacin	6.94×10^{9}	1.03	3	0	9	0.242	-0.101	0.003	5
13	Flumequine	8.26×10^{9}	1.03	3	0	5	0.328	0.015	-0.136	5
14	Marbofloxacin	9.03×10^{9}	1.03	3	0	9	0.242	-0.042	-0.251	5
15	Danofloxacin	6.15×10^{9}	1.02	3	0	7	0.203	-0.034	-0.108	5
16	Enrofloxacin	7.95×10^{9}	1.01	3	0	7	0.22	-0.051	-0.08	5
17	Sulfamethazine	8.30×10^{9}	1.02	2	0	6	0.312	-0.117	-0.33	6
18	Sulfamethizole*	7.90×10^{9}	1.03	2	0	6	0.32	-0.185	-0.209	6

19	Sulfamethoxazole	8.50×10^{9}	1.03	2	0	6	0.32	-0.162	-0.406	6
20	Sulfamerazine	7.80×10^{9}	1.02	2	0	6	0.313	-0.128	-0.382	6
21	Bisphenol A	6.90×10^{9}	0.99	0	0	2	0.338	-0.041	-0.251	7
22	Iohexol	3.21×10^{9}	1.03	3	0	12	0.449	-0.055	-0.083	8
23	Iopromide	3.34×10^{9}	1.03	3	0	11	0.436	-0.058	0.164	8
24	Iopamidol	3.42×10^{9}	1.03	3	0	11	0.451	-0.058	0.161	8
25	2,3,5-Triiodbenzoic acid	9.70×10^{9}	1.03	1	0	2	0.558	-0.145	-0.433	8
26	3-Acetamino benzoic acid	5.40×10^{9}	1.03	2	0	4	0.46	0.192	-0.232	8
27	Chlortetracycline	5.20×10^{9}	1.03	5	0	10	0.259	-0.025	0.132	8
28	Oxytetracycline	5.63×10^{9}	1.03	5	0	11	0.266	-0.056	0.097	8
29	Doxycycline	7.58×10^{9}	1.03	5	0	10	0.262	-0.031	0.008	8
30	Trimethoprim	8.34×10^{9}	1.01	0	0	7	0.308	-0.009	-0.366	8
31	Atrazine	3.17×10^{9}	1.01	0	0	5	0.451	0.034	0.108	8
32	Diclofenac*	9.29×10^{9}	1.02	1	0	3	0.322	-0.069	-0.444	8
33	Ibuprofen	5.97×10^{9}	0.99	1	0	2	0.416	0.356	0.038	8
34	Naproxen	7.53×10^{9}	1.01	1	0	3	0.328	0.198	-0.133	8
35	2,6-Dinitrotoluene	1.50×10^{9}	1.07	4	0	4	0.532	0.228	-0.202	9
36	2,4-Dinitrotoluene*	1.40×10^{9}	1.07	4	0	4	0.512	0.228	-0.208	9
37	EPTC	4.80×10^{9}	0.98	1	0	2	0.764	0.053	0.18	9
38	Prometon	2.80×10^{9}	1	0	0	6	0.442	0.581	0.092	9
39	Linuron	6.40×10^{9}	1.04	1	0	4	0.424	-0.13	-0.215	9

40	Diuron	7.40×10^{9}	1.02	1	0	3	0.44	-0.109	-0.179	9
41	Cyclonite	1.10×10^{9}	1.12	6	0	9	0.498	0.582	-0.044	9
42	Molinate	6.90×10^{9}	0.99	1	0	2	0.468	0.067	0.171	9
43	Nitrobenzene	3.90×10^{9}	1.04	2	0	2	0.558	0.242	-0.242	9
44	Terbacil*	7.40×10^{9}	1.02	3	0	4	0.572	-0.058	0.403	9
45	Chlortoluron	6.90×10^{9}	1.01	1	0	3	0.44	-0.05	-0.224	10
46	Isoproturon*	7.90×10^{9}	0.99	1	0	3	0.407	0.572	-0.115	10
47	Dibromomethane	9.00×10^{7}	1.05	0	0	0	2.042	0.5	0.078	11
48	Dichloromethane	9.00×10^{7}	1.08	0	0	0	2.042	0.5	0.018	11
49	Trichloromethane*	5.40×10^{7}	1.15	0	0	0	1.592	0.333	0.063	11
50	Tribromomethane*	1.30×10^{8}	1.09	0	0	0	1.592	0.333	0.174	11
51	1,1,2-Trichloroethane	1.30×10^{8}	1.08	0	1	0	1.106	-0.583	0.139	11
52	1,2-Dichloropropane*	3.80×10^{8}	1.02	0	1	0	1.106	-0.583	0.16	11
53	1,2-Dibromo-3-chloropropane*	3.20×10^{8}	1.03	0	2	0	0.955	-0.543	0.043	11
54	2-Bromoethanol	3.50×10^{8}	1.02	0	1	1	1.089	-0.388	0.01	11
55	1,1,1-Trichloro-2-methyl-2-propanol*	2.70×10^{8}	1.05	0	0	1	1.201	0.42	0.21	11
56	Aldicarb	8.10×10^{9}	1.01	2	0	4	0.682	-0.249	0.296	11
57	Dalapon	7.30×10^{7}	1.09	1	0	2	1.159	-0.121	0.205	11
58	Lindane	5.80×10^{8}	1.07	0	0	0	0.605	-0.333	0.545	11
59	beta-Cyclocitral	7.42×10^{9}	0.98	2	0	1	0.607	0.031	0.295	12
60	Geosmin	7.80×10^{9}	0.97	0	0	1	0.484	-0.064	0.569	12

61	3-Hexen-1-ol*	7.45×10^{9}	0.98	1	0	1	0.666	-0.067	0.066	12
62	belta-ionone*	7.79×10^{9}	0.97	2	0	1	0.489	-0.031	0.706	12
63	2-Isopropyl-3-methoxypyrazine	4.91×10^{9}	1	0	0	3	0.547	0.037	0.093	12
64	2,6-Nonadienal	1.05×10^{10}	0.98	3	0	1	0.552	-0.028	0.135	12
65	1-Penten-3-one	4.71×10^{9}	0.99	2	0	1	0.955	-0.28	0.016	12
66	2,6-Di-tert-butyl-4-methylpenol*	3.20×10^{9}	0.97	0	0	1	0.544	-0.015	0.386	12
67	2,4,6-Tribromoanisole	3.74×10^{9}	1.03	0	0	1	0.568	-0.184	-0.331	12
68	2,4,6-Trichloroanisole*	5.10×10^{9}	1.05	0	0	1	0.568	-0.202	-0.055	12
69	Carbamazepine	8.80×10^{9}	1	2	0	3	0.327	0.468	0.118	13
70	Diazepam*	7.20×10^{9}	1	2	0	3	0.3	-0.029	-0.314	13
71	Azithromycin	2.90×10^{9}	1	1	0	14	0.26	-0.114	0.383	14
72	Tylosin	8.20×10^{9}	1.01	5	0	18	0.174	-0.099	0.639	14
73	Ciprofloxacin	4.10×10^{9}	1.02	3	0	7	0.236	-0.05	-0.127	14
74	Lincomycin	8.50×10^{9}	1.01	1	0	8	0.301	0.03	0.381	14
75	Cephalexin	8.50×10^{9}	1.02	4	0	7	0.237	-0.029	0.18	14
76	Amikacin	7.20×10^{9}	1.03	1	0	18	0.235	-0.244	0.415	14
77	Roxithromycin*	5.40×10^{9}	1.01	2	0	17	0.253	-0.088	0.753	14
78	Acetochlor*	6.30×10^{9}	1	1	1	3	0.482	0.005	-0.254	15
79	Propachlor*	4.60×10^{9}	1	1	1	2	0.498	-0.006	-0.16	15
80	Metolachlor*	6.70×10^{9}	1	1	1	3	0.498	0.003	-0.152	15
81	Butachlor	7.40×10^{9}	0.99	1	1	3	0.442	0.005	-0.165	15

82	Acebutolol	4.60×10^{9}	1	2	0	6	0.356	-0.17	0.029	15
83	Metoprolol	7.30×10^{9}	0.99	0	0	4	0.322	-0.191	-0.04	15
84	Tris(2-butoxyethyl) phosphate	1.03×10^{10}	1	1	0	7	0.559	-0.175	0.283	16
85	Tributyl phosphate	6.40×10^{9}	0.99	1	0	4	0.668	-0.184	0.437	16
86	Tris(2-chloroethyl) phosphate	5.60×10^{8}	1.05	1	3	4	0.734	-0.337	0.082	16
87	Tris(2-chloroisopropyl) phosphate	1.98×10^{8}	1.03	1	3	4	0.798	-0.189	0.361	16
88	17belta-Estradiol*	1.41×10^{10}	0.98	0	0	2	0.261	-0.012	0.171	17
89	Parathion	9.70×10^{9}	1.03	3	0	5	0.405	-0.071	-0.397	16
90	4-Chloro-3,5-dinitrobenzoic acid	3.30×10^{8}	1.12	5	0	6	0.517	-0.064	-0.081	18
91	1-Chloro-2,4-dinitrobenzene	8.20×10^{8}	1.11	4	0	4	0.512	-0.087	-0.161	18
92	1,3-Dinitrobenzene	1.10×10^{9}	1.09	4	0	4	0.504	0.189	-0.212	18
93	2,4-Dinitrophenol	2.30×10^{9}	1.1	4	0	5	0.512	0.078	-0.222	18
94	3-Nitrophenol	5.00×10^{9}	1.06	2	0	3	0.544	0.054	-0.251	18
95	2-Nitrophenol	5.90×10^{9}	1.06	2	0	3	0.57	0.054	-0.256	18
96	4-Nitrophenol	6.20×10^{9}	1.06	2	0	3	0.525	0.054	-0.244	18
97	3-Nitrotoluene	8.20×10^{9}	1.02	2	0	2	0.544	0.266	-0.226	18
98	4-Nitrotoluene*	8.60×10^{9}	1.02	2	0	2	0.525	0.266	-0.237	18
99	Equilenin*	1.70×10^{10}	0.99	1	0	2	0.261	-0.012	-0.081	19
100	Butylated hydroxyanisole	7.40×10^{9}	0.99	0	0	2	0.532	-0.067	0.033	19
101	Fenoterol*	3.90×10^{9}	1.01	0	0	5	0.248	-0.138	-0.287	19
102	Tetracycline	8.20×10^{9}	1.03	5	0	10	0.26	-0.05	0.163	19

103	Triclosan	6.00×10^{9}	1.04	0	0	2	0.318	-0.117	-0.331	19
104	Phenol	6.10×10^{9}	1	0	0	1	0.643	-0.125	-0.277	19
105	17alpha-Ethinylestradiol	4.90×10^{9}	0.98	0	0	2	0.254	-0.038	0.171	19
106	Gemfibrozil	7.10×10^{9}	0.99	1	0	3	0.361	0.152	0.041	19
107	Methicillin*	1.00×10^{10}	1.03	3	0	8	0.25	-0.144	-0.168	19
108	Benzo[a]pyrene*	9.40×10^{8}	0.98	0	0	0	0.251	1	-0.226	19
109	Clofibric acid	5.20×10^{9}	1.02	1	0	3	0.451	-0.09	-0.067	19
110	Trifluralin	1.30×10^{9}	1.06	4	0	8	0.499	0.424	-0.079	19
111	Methoxychlor	3.90×10^{9}	1.02	0	0	2	0.319	0.602	-0.355	19
112	Butylbenzyl phthalate*	4.00×10^{9}	1	2	0	4	0.267	0.271	-0.308	19
113	Iomeprol	2.50×10^{9}	1.03	3	0	11	0.442	-0.058	0.117	19
114	Dicamba	3.50×10^{9}	1.06	1	0	3	0.562	-0.121	-0.091	19
115	Dicofol	3.70×10^{9}	1.04	0	0	1	0.363	0.279	-0.343	19
116	Di(2-ethylhexyl) phthalate*	3.40×10^{8}	0.99	2	0	4	0.348	0.304	0.407	19
117	Hexachlorobenzene*	2.40×10^{8}	1.13	0	0	0	0.605	-0.333	0.038	19
118	Pyrene	1.40×10^{9}	0.98	0	0	0	0.314	1	0	19

* Validation set compounds (n = 29).

Reference: (1) Razavi *et al.*, 2009; (2) Song *et al.*, 2009; (3) Song *et al.*, 2008a; (4) Song *et al.*, 2008b; (5) Santoke *et al.*, 2009; (6) Mezyk *et al.*, 2007; (7) Peller *et al.*, 2009; (8) Cooper *et al.*, 2010; (9) Elovitz *et al.*, 2008; (10) Benitez *et al.*, 2007; (11) Haag and Yao 1992; (12) Peter and von Gunten 2007; (13) Huber *et al.*, 2003; (14) Dodd *et al.*, 2006; (15) Benner *et al.*, 2008; (16) Watts and Linden 2009; (17) Rosenfeldt and Linden 2004; (18) Einschlag *et al.*, 2003; (19) Chapter 4.

Model 1 fits the training set compounds well as shown by the high adjusted R^2 value (R^2_{adj} = 0.823), and the prediction error represented by *RMSE* (0.204) is small. In addition, Q^2 is 0.773 indicating good robustness and internal predictivity, and the high predictive R^2 (R^2_{pred} = 0.772) and relative small error (*RMSEP* = 0.329) indicates good external predictivity. All seven descriptors are statistically significant at the 0.95 confidence level. The MLR method assumes that the molecular descriptors are independent from each other. Multicollinearity occurs when two or more descriptors are highly correlated and it is difficult to reliably estimate their individual regression coefficients (Eriksson *et al.*, 2003). MLR can be applied in QSPR studies if multicollinearity among variables is small. To detect the multicollinearity, the pairwise correlation and variance inflation factor (*VIF*) were calculated. First of all, no high correlation pairs are found (Table 6.2). However, the pairwise correlation among two descriptors is limiting in general. It is possible though that a linear dependence exists among three or more descriptors. *VIF* can be used to detect and quantify the correlation among a descriptor and all the remaining descriptors in the model (Roy and Roy, 2009). A *VIF* of 1 for a specific descriptor means that there is no correlation between this descriptor and the remaining descriptors, and a *VIF* exceeding 10 is a sign of serious multicollinearity. As shown in Table 6.3, the *VIF* of all the selected descriptors are very small (close to 1) i.e., much smaller than the cut-off value of 10, indicating that multicollinearity is not an issue in this descriptor set.

Table 6.2 Correlations of selected molecular descriptors.

	Me	*nDB*	*nCH2RX*	*nHAcc*	*Vindex*	*MATS2m*	*Mor27p*
Me	1						
nDB	0.171	1					
nCH2RX	0.073	-0.086	1				
nHAcc	-0.129	0.469	-0.129	1			
Vindex	0.300	-0.262	0.281	-0.507	1		
MATS2m	-0.036	-0.192	-0.294	-0.345	0.055	1	
Mor27p	-0.195	0.015	0.120	-0.140	0.149	-0.176	1

Table 6.3 Model properties of the selected molecular descriptors.

Descriptor	Std. coefficient	Prob. level	*VIF*	Correlation to logk_{OH}
Me	-0.490	<0.0001	1.33	-0.56
nDB	0.137	0.0129	1.44	0.27
nCH2RX	-0.285	<0.0001	1.23	-0.42
nHAcc	0.154	0.0155	1.93	0.59
Vindex	-0.344	<0.0001	1.61	-0.73
MATS2m	-0.186	0.0007	1.40	-0.15
Mor27p	-0.200	0.0001	1.21	-0.18

The predicted vs. the measured logk_{OH} values are shown in Figure 6.1(a). It shows that model 1 works well for most of the training set compounds as the predicted values are very close to the measured values. However, the k_{OH} of compound #57 (Dalapon) is substantially over predicted and appears to be an outlier. This was confirmed by the outlier analysis in section 6.3.2.

It is very common in data analysis and statistical modeling applications that a small proportion of observations are far from the rest of the data. Such data or even a single outlier can distort the regression results by pulling the least square fit too much in their direction, thereby impacting the regression coefficients, and limiting the ability to understand the data. Therefore, Model 1 can be further improved by removing compound #57. The new model is shown in Equation 6.11 (Model 2). The R^2_{adj} value increases to 0.846 and *RMSE* decreases to 0.178. More importantly, the predictive power is substantially improved (R^2_{pred} value increases from 0.772 to 0.858, and *RMSEP* drops from 0.329 to 0.255). For all compounds, the predicted logk_{OH} values are very close to the experimentally measured logk_{OH} and no other outliers are apparent in Figure 6.1(b).

$$\log k_{OH} = 16.451 - 6.932 \times Me + 0.159 \times nDB - 0.679 \times CH2RX + 0.401 \times nHAcc - 0.460 \times Vindex - 0.363 \times MATS2m - 0.362 \times Mor27p \quad \textbf{(6.11)}$$

n_{training} = 88, R^2 = 0.859, R^2_{adj} = 0.846, Q^2 = 0.804, F(7,80) = 69.384 (p < 0.0001), *RMSE* = 0.178

$n_{\text{validation}}$ = 28, R^2_{pred} = 0.858 , *RMSEP* = 0.255

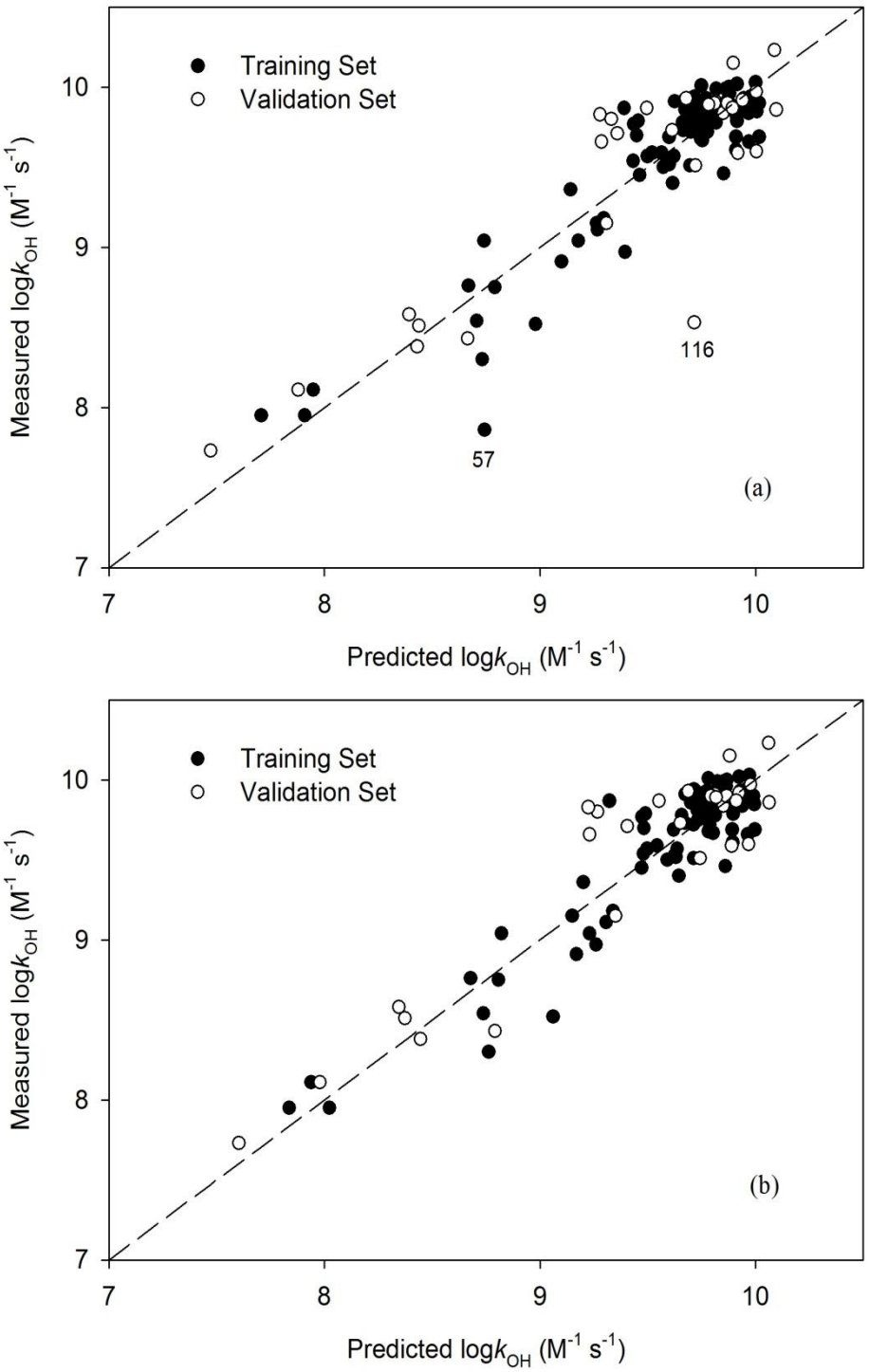


Figure 6.1 A plot of predicted logk_{OH} values vs. measured logk_{OH} (a) Model 1, (b) Model 2 (outliers removed: #57 Dalapon in the training set, and #116 DEHP in the validation set).

The models are then validated both internally and externally. The leave-one-out cross-validation method was used for internal validation. For Model 1, the Q^2 value is high (0.773) indicating the good robustness and internal predictivity of the model. When applying this model to an external validation set, the R^2_{pred} value (0.772) is high as well indicating very good predictive power of the model, and the prediction error for the validation set (*RMSEP* = 0.329) is small. However, an outlier compound #116 (DEHP) is also found in the validation set which is far away from the regression line (as shown in Figure 6.1(a)). Model 2 is obtained by removing outlier compounds from training set (#57) and validation set (#116). As a result, the internal validation Q^2 value is increased to 0.804, the external validation R^2_{pred} value is increased to 0.858, and *RMSEP* is reduced to 0.255. These statistical values indicate that model 2 is an excellent model for predicting k_{OH} values.

6.3.2 Applicability Domain and Outliers Detection

A Williams plot is drawn to show the applicability domain of Model 1 (Figure 6.2). The applicability domain is established by a squared area within ± 3 standard deviations of the standardized residual and a leverage threshold h^*. A value of 3 for a standardized residual is commonly used as a cut-off value for acceptable predictions and compounds with standardized residuals > 3 standard deviation units are considered outliers (Gramatica and Papa 2005). In terms of leverage, a compound with $h_i > h^*$ diverges in structure from most compounds in the training set and seriously influences the regression performance. But a compound with a high leverage value is not necessarily an outlier because its standardized residual may be small. First, the leverage value of 8 training set compounds are higher than the leverage threshold ($h^* = 0.27$) indicating influencing structural features. However, these compounds in the training set fit the model well, thus they stabilize the model and make it more precise. Five validation set compounds were far from the centroid of the descriptor space ($h_i > h^*$), they are trichloromethane (#49), 1,2-dichloropropane (#52), 1,2-dibromo-3-chloropropane (#53), acetochlor (#78), and hexachlorobenzene (#117), but the model still shows good predictivity for these compounds. However, predictions of compounds with high leverage values should be used with great caution. Second, the analysis of the applicability domain confirms the presence of outliers. Compounds #57 (Dalapon) in the training set and #116 (DEHP) in the validation set are identified as outliers (> 3 standard deviation). As shown in Model 2, after

removing the two outliers, model performance is improved as was shown by internal and external validation.

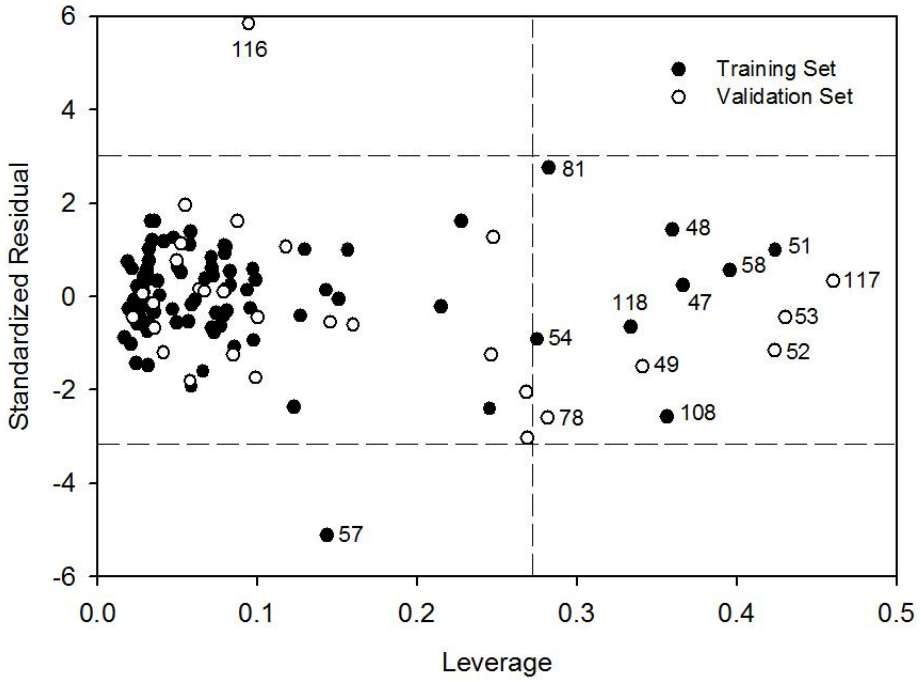


Figure 6.2 Williams plot of the entire data set for model 1 (h^* = 0.27).

6.3.3 Mechanistic Implications of the Descriptors in the QSPR model

As shown in Table 6.3, the mean atomic Sanderson electronegativity (*Me*) descriptor is the main contributor to the logk_{OH} because of the highest standardized coefficient, and this descriptor is negatively correlated to logk_{OH}. Electronegativity is the tendency or power of an atom (or a functional group) to attract electrons. The greater the electronegativity of an atom the greater is its desire to withhold its electrons (i.e. less likely to donate its electrons). For a molecule with a high mean electronegativity, a very high energy is required to remove the electrons thereby making the hydroxyl radical induced electron transfer difficult (Sanderson 1983). In addition, electronegativity is related to the average of the highest occupied molecular orbital energy (HOMO) and the lowest unoccupied molecular orbital energy (LUMO) (Zhan *et al.*, 2003). Quantum-chemical descriptors such as HOMO and LUMO have been used in predicting the reactivity of compounds in ozonation and hydroxyl radical reactions (Gramatica *et al.*, 2004). In a similar study (Kusic *et al.*, 2009) the main contribution

to the hydroxyl degradation rate is given by the HOMO energy parameter. However, HOMO and LUMO energy descriptors are not used in this study because they were found not to be significant in the preliminary analysis (Appendix D).

Descriptors *Mor*27*p*, *MATS*2*m*, and *Vindex* are related to the topological structure of a molecule. The 3D-MoRSE descriptors *Mor*27*p* is weighted by atomic polarizability. This descriptor is highly sensitive to the 3-dimensional molecular structure and polarizability. It is known that polarizability is related to chemical reactivity of a molecule. Hence, the polarizability weighting descriptor *Mor*27*p* confirms the significance of molecular polarity and polarizability for reactivity. *MATS*2*m* is a 2D autocorrelation descriptor weighted by molecular mass, giving information on the distribution of molecular mass along the topological structure. Similarly, molecular weight was found important in predicting the hydroxyl radical rate constants of aromatic compounds in water (Kusic *et al.*, 2009).

The discrete descriptors *nDB* (number of double bonds) and *nHAcc* (number of acceptor atoms for H-bonds) have positive indices, while all other descriptors have negative indices. The positive coefficient of *nDB* can be explained by hydroxyl radical addition to double bonds. The functional group *nHAcc* can positively affect the H-atom abstraction during hydroxyl radical reaction by withdrawing electrons from the C-H bond. The functional group descriptor *nCH*2*RX* represents alkyl halide (primary) substructures. The halogens (Cl, Br, and I) are electron withdrawing groups, making the C atom electrophilic and prone to attack by nucleophiles. They are therefore less likely to be attacked by hydroxyl radicals which are excellent electrophiles. This explains the negative coefficient of *nCH*2*RX*.

The model developed in this study is applicable to a wide range of micropollutants with diverse structures, and can be used to provide reliable estimation of k_{OH} for many micropollutants when experimental data are not available. It is therefore useful for the water industry when assessing the removal efficiency of unknown micropollutants during AOPs, i.e. screening micropollutants of interest, and providing an estimate of AOP feasibility. In addition, this model can provide input to the R_{ct} model (Elovitz and von Gunten 1999) which together with a QSPR model for k_{O3} prediction (Chapter 5) can be used to assess the removal of micropollutants in natural water during ozonation treatment. When dealing with the prediction for unknown compounds, first, users are expected to check if the compounds fall into the applicability domain by calculating the leverage. Predictions made for compounds outside of the applicability domain should be used with great caution. The

second step is to calculate the molecular descriptors by using the DRAGON software. And finally, calculate the logk_{OH} using the developed QSPR model.

6.4 Conclusions

The k_{OH} values in the aqueous phase are important for assessing the removal efficiency of micropollutants during advanced oxidation processes. A QSPR model for the prediction of aqueous phase k_{OH} values was successfully developed in this study. A data set including 118 micropollutants with diverse structures were collected from the literature and divided into the training set (n = 89) and validation set (n = 29). Multiple linear regression was then used to develop a QSPR model based on the training set. The model was then externally validated with the validation set. The leverage approach (Williams plot) was used to determine the applicability domain of the QSPR model and to identify outliers in the training set and validation set. The developed QSPR model provides a valuable tool for the prediction of k_{OH} values of a wide range of micropollutants.

1. A seven-variable model was developed using the training set. The main contribution to the rate constant was obtained from the mean atomic Sanderson electronegativity descriptor *Me*. In addition, model descriptors were also related to polarizability, double bonds, H-bond acceptors, etc., which can be explained by the H-atom abstraction and OH-addition mechanisms of the radical reaction.
2. The performance of the QSPR model was assessed in terms of goodness-of-fit, robustness and predictivity (using a validation set). The model fitted the training set very well as seen in the adjusted R^2 = 0.823; the cross validated Q^2 = 0.773 and R^2_{pred} = 0.772, all indicating good robustness and predictivity.
3. One outlier compound was identified in the training set (Dalapon) and one in the validation set (DEHP). By removing these two outliers, the QSPR model was further improved as indicated by higher R^2_{adj} (0.846), Q^2 (0.804), and R^2_{pred} (0.858), and lower *RMSE* (0.178) and *RMSEP* (0.255).

Chapter 7 QSPR Models Application in Natural Waters for Assessing Removals of Micropollutants during Ozonation

The R_{ct} model developed by Elovitz and von Gunten (1999) can be used to predict the percentage removal of micropollutants from natural water during ozonation if the k_{O3} and k_{OH} of the target micropollutants are known. This chapter will focus on the application of QSPR models developed in Chapter 5 (for k_{O3}) and Chapter 6 (for k_{OH}) under natural water conditions by assessing the removal efficiency of micropollutants during ozonation using the R_{ct} model developed by Elovitz and von Gunten (1999). The predicted k_{O3} and k_{OH} by QSPR models were used as input for the R_{ct} model to predict the removal, and then predicted removal values were compared with reported values.

Outline: QSPR models developed previously can be used to estimate the rate constants k_{O3} and k_{OH} of untested micropollutants, and the R_{ct} model developed by Elovitz and von Gunten (1999) can be used to assess the removal efficiency of micropollutants during ozonation if that their rate constants are known. Therefore, the combination of QSPR and R_{ct} models are useful in the evaluation of the removal efficiency of untested micropollutants from natural water by ozonation. To demonstrate the applicability of this approach, sixteen micropollutants were collected from reported ozonation studies using a number of different water sources and known R_{ct} values. The k_{O3} and k_{OH} of these collected micropollutants were estimated by QSPR models, and the predicted percentage removals were calculated based on reported R_{ct} values. These estimated removals were then compared with the reported values which were determined experimentally. The methods to increase the removal were discussed based on a case study for geosmin, a taste and odour compound. In addition, the sources of error of the prediction were also discussed. The results show that the combination of R_{ct} with QSPR models can provide reliable estimations for most of the selected micropollutants and can be used as a tool for initial assessment and estimation of ozonation system.

Keywords: ozone, hydroxyl radical, rate constant, percentage removal, QSPR model, R_{ct} model.

7.1 Introduction

Ozonation is commonly applied in drinking water treatment for disinfection, oxidation, taste and odor control, and color removal. Molecular ozone is unstable in water and decomposes gradually into hydroxyl radicals. Molecular ozone reacts selectively with micropollutants with functional groups such as amines, phenols and double bonds, while hydroxyl radicals react less selectively and more rapidly with various micropollutants. Therefore, ozone (direct oxidation) and hydroxyl radicals (indirect oxidation) pathways have to be considered simultaneously when assessing the overall effect of ozonation on micropollutants (von Gunten 2003).

Concentrations of ozone and hydroxyl radicals are needed to estimate the overall effect of ozonation. Ozone can be easily monitored via the Indigo method or a spectrophotometer (Eaton *et al.*, 2005). In contrast, hydroxyl radicals are very difficult to measure directly because of their high reactivity and their very low steady-state concentrations in water. Therefore R_{ct}, which is defined as the ratio of hydroxyl radical exposure (i.e., oxidant concentration integrated over the reaction time) to the molecular ozone exposure during the ozonation process (Equation 7.1), was developed as an indirect way to measure hydroxyl radicals (Elovitz and von Gunten, 1999). For a given water source, R_{ct} can be experimentally determined by monitoring the decrease of a probe compound, *p*CBA (*para*-chlorobenzoic acid), during ozonation (Equation 7.2) which reacts rapidly with hydroxyl radicals ($k_{OH,pCBA} = 5\times10^9$ $M^{-1}s^{-1}$) but very slowly with molecular ozone ($k_{O3,pCBA} = 0.15$ $M^{-1}s^{-1}$).

$$R_{ct} = \frac{\int [OH]dt}{\int [O_3]dt} \qquad \textbf{(7.1)}$$

$$\ln\left(\frac{[pCBA]_t}{[pCBA]_0}\right) = -k_{OH,pCBA}\int [OH]dt = -k_{OH,pCBA}R_{ct}\int [O_3]dt \qquad \textbf{(7.2)}$$

Where $k_{OH,pCBA}$ is the second-order rate constant of *p*CBA with hydroxyl radicals, $[pCBA]_0$ and $[pCBA]_t$ is the initial concentration and the remaining concentration at time *t*, respectively.

After an initial phase (seconds), the R_{ct} value remains constant for the rest of the ozonation process (Buffle *et al.*, 2006). In natural waters, R_{ct} values were reported in the range of $10^{-10} – 10^{-7}$, depending on the water matrix (Elovitz *et al.*, 2000). Generally, R_{ct} values increase with enhanced hydroxyl radical formation from ozone decomposition at increased pH and temperature; and R_{ct} values decrease with increases in hydroxyl radical scavengers such as bicarbonate ions (i.e., increased alkalinity) (Elovitz *et al.*, 2000). The R_{ct} value for a given water source is relatively easy to determine

experimentally but difficult to predict. However, a model was developed to predict R_{ct} of surface water using a few water quality parameters (DOC, pH, UV_{254}, the ratio of UV_{210} over UV_{254}) and treatment condition (H_2O_2/O_3 mass ratio). A very high coefficient of correlation ($R^2 = 0.92$) was obtained, and the predictivity was validated for MIB oxidation which closely matched the published data (Vincent *et al.,* 2010). However, the error associated with the R_{ct} prediction was very large for waters with low pH (5.6) and high pH values (8.1), and the temperature effect on R_{ct} was not considered in this model.

For any given water source, after the R_{ct} value has been determined, the removal efficiency of micropollutants as a function of ozone exposure can be assessed as long as their k_{O3} and k_{OH} values are known (Equation 7.3).

$$\ln\left(\frac{[P]_t}{[P]_0}\right) = -\left(k_{OH}\int[OH]dt + k_{O3}\int[O_3]dt\right) = -\left(k_{OH}R_{ct} + k_{O3}\right)\int[O_3]dt \qquad \textbf{(7.3)}$$

However for many emerging micropollutants, especially endocrine disrupting chemicals (EDCs) and pharmaceuticals and personal care products (PPCPs), the reaction rate constants k_{O3} and k_{OH} are not available. In addition, the experimental determination of these rate constants is time consuming and expensive. QSPR models are therefore very useful to predict the rate constants of a large number of untested compounds based on their structural features. QSPR models on k_{O3} and k_{OH}, for a wide range of micropollutants with diverse structures, were developed in Chapters 5 and 6, respectively. Combining the QSPR models with the R_{ct} model, it is possible to assess the removal efficiency of untested micropollutants during ozonation in a particular natural water provided it's R_{ct} value is known.

The objective of this Chapter is therefore to explore the applicability of the developed QSPR models in combination with the R_{ct} models for the assessment of the removals of micropollutants during ozonation. Information with respect to target micropollutant removals, R_{ct} values of given water, and ozone exposures were collected from the literature. The predicted removals of these micropollutants obtained by using the QSPR models and reported R_{ct} values, were then compared with their reported removals.

7.2 Materials and Methods

7.2.1 Data Set

Micropollutants used in this study were collected from the literature and included taste and odour compounds, acetamide herbicides, phenyl-urea herbicides, fuel additives, and pesticides. Their experimentally determined k_{O3} and k_{OH} values, as well as the R_{ct} values of the water sources are summarized in Table Appendix E.1.

7.2.2 Rate Constant Prediction by QSPR Models

QSPR models used for predicting the k_{O3} and k_{OH} were developed in Chapter 5 and Chapter 6, respectively. The model based on average molecular weight and phenolic functional group (Equation 5.8) was used for k_{O3} prediction, and Equation 6.10 was used for k_{OH} predictions. The prediction error and the 95% confidence interval for the predictions on new compounds (i.e., prediction interval) were calculated by Equations 7.4 and 7.5, respectively (Table 7.1).

$$s_{pred} = \sqrt{MSE\left(1 + x_p'\left(X'X\right)^{-1}x_p\right)} \quad \textbf{(7.4)}$$

$$\hat{y}_i \pm t_{\alpha/2,n-p-1}\sqrt{MSE\left(1 + x_p'\left(X'X\right)^{-1}x_p\right)} \quad \textbf{(7.5)}$$

Where *MSE* is the mean squared error of the training set compounds; X is the descriptor matrix of the training set; and x_p is the descriptors vector for the new compound; $\hat{y}_i$ is the predicted rate constant in log scale; $t_{\alpha/2,n\text{-}p\text{-}1}$ is the two-sided student's *t*-distribution with (*n*-*p*-1) degrees of freedom at 100(1-α) percent confidence level, where *n* is the number of training set compounds and *p* is the number of descriptors involved in the model (Neter *et al.*, 1983). The prediction interval takes into account both the error from the fitted model and the error associated with the new compound.

7.2.3 Calculation of the Ozone Exposure

The ozone exposure (integration of the ozone residual concentration over time) is also needed for predicting the removal of micropollutants. Ozone exposure was reported in Peter and von Gunten (2007). However, this is not the case for the other studies. Instead, ozone exposure can be calculated with the reported ozone decomposition rate (Appendix E.1).

The ozone decomposition in natural waters usually involves two phases, a fast initial decrease of ozone (in the order of seconds), and a second phase in the order of minutes to hours which can be modeled with first-order kinetics (Equation 7.6).

$$\ln\left(\frac{[O_3]_t}{[O_3]_0}\right) = -kt \qquad \textbf{(7.6)}$$

Where t is the contact time (s), k is the decomposition rate (s^{-1}), $[O_3]_0$ and $[O_3]_t$ are the initial ozone concentration and the remaining concentration at time t, respectively. The ozone exposure is calculated by integrating the ozone residual concentration over time (Equation 7.7). The calculated ozone exposure values are shown in Appendix E.1.

$$\int [O_3]dt = \frac{[O_3]_0}{k}\left(1 - e^{-kt}\right) \qquad \textbf{(7.7)}$$

The reactor hydraulics is important for delivering a certain ozone dose, but we are not providing any details on reactor hydraulics in this study.

7.2.4 Calculation of the Percentage Removal

The percentage removal (%R) of a micropollutant P is calculated using the R_{ct} concept (Equation 7.8).

$$\%R = 100\left(1 - \frac{[P]}{[P]_0}\right) = 100\left(1 - \exp\left[-\left(k_{OH}R_{ct} + k_{O3}\right)\int [O_3]dt\right]\right) \qquad \textbf{(7.8)}$$

The prediction errors in k_{O3} and k_{OH} directly impact the calculation of %R. To assess the error involved in the calculation of %R, the interval associated with the %R calculation is therefore calculated using the prediction interval associated with the k_{O3} and k_{OH} (Table 7.1).

Table 7.1 The calculation results of predicted rate constant, percentage removal, prediction error, and prediction interval.

No.	Compound	k_{O3}				k_{OH}				%R	
		logk_{O3}	s_{pred}	PI	h	logk_{OH}	s_{pred}	PI	h	%R	Interval
1	2,6-Nonadienal	4.58	0.77	(2.96, 6.19)	0.49	9.91	0.22	(9.48, 10.35)	0.053	100	(99, 100)
2	1-Penten-3-one	4.50	0.75	(2.92, 6.07)	0.42	9.75	0.23	(9.53, 10.20)	0.127	100	(98, 100)
3	Belta-cyclocitral	4.61	0.78	(2.98, 6.25)	0.53	9.79	0.22	(9.57, 10.23)	0.067	100	(99, 100)
4	Isoproturon	4.43	0.74	(2.88, 5.97)	0.36	9.88	0.22	(9.65, 10.33)	0.110	100	(61, 100)
5	Chlortoluron	4.01	0.69	2.57, 5.44)	0.17	10.14	0.23	(9.91, 10.60)	0.169	100	(70, 100)
6	2-Isopropyl-3-methoxypyrazine	-0.03	0.69	(-1.48, 1.42)	0.20	9.60	0.22	(9.38, 10.04)	0.073	55	(38, 90)
7	Diuron	-0.51	0.67	(-1.91, 0.88)	0.11	10.02	0.23	(9.80, 10.47)	0.114	74	(55, 98)
8	Atrazine	-0.22	0.68	(-1.64, 1.20)	0.15	9.57	0.22	(9.35, 10.01)	0.075	66	(47, 96)
9	Butachlor	-0.04	0.69	(-1.49, 1.41)	0.20	9.36	0.24	(9.12, 9.84)	0.276	27	(16, 66)
10	Acetochlor	-0.12	0.69	(-1.55, 1.31)	0.17	9.23	0.24	(8.99, 9.71)	0.262	20	(12, 56)
11	Linuron	-0.54	0.67	(-1.94, 0.85)	0.11	9.85	0.22	(9.63, 10.29)	0.079	59	(42, 92)
12	Propachlor	-0.20	0.68	(-1.62, 1.22)	0.15	9.19	0.24	(8.95, 9.67)	0.258	19	(11, 52)
13	2-Methylisoborneol	0.25	0.73	(-1.27, 1.78)	0.32	9.83	0.22	(9.61, 10.28)	0.089	75	(56, 98)
14	MTBE	0.34	0.74	(-1.21, 1.90)	0.38	9.49	0.24	(9.25, 9.96)	0.219	25	(13, 90)
15	Geosmin	0.27	0.73	(-1.26, 1.79)	0.33	9.60	0.24	(9.36, 10.07)	0.232	28	(17, 69)
16	2,4,6-Tribromoanisole	-1.51	0.76	(-3.09, 0.08)	0.43	9.62	0.22	(9.40, 10.07)	0.096	57	(39, 90)

logk_{O3} and logk_{OH} values were calculated by QSPR models (Equation 5.8 and Equation 6.10, respectively), prediction error (s_{pred}) and 95% prediction interval (PI) were done by Equations 7.4 and 7.5, respectively, h is the leverage value.

7.3 Results and Discussion

Once the R_{ct} value for a given natural water source has been determined using the probe compound *p*CBA, and k_{O3} and k_{OH} are known, the removal of a micropollutant *P* can be modeled and predicted (Equation 7.8). The theoretical relationship between the percentage removal and the rate constants k_{O3} and k_{OH} is shown in Figure 7.1. The percentage removal increases with increasing k_{O3} and k_{OH} values. The ozone rate constant has a more pronounced effect on the percentage removal than the hydroxyl radical rate constant because ozone pathway is more important than the hydroxyl radical pathway when k_{O3} is larger than $k_{OH}R_{ct}$, and vice versa. Considering the fact that k_{OH} values are mostly in the range of 10^8-10^{10} M^{-1} s^{-1} and R_{ct} ranges from 10^{-10}-10^{-7}, the product of k_{OH} and R_{ct} is usually less than 100. Therefore, the percentage removals are low for ozone-resistant compounds, and high percentage removals can be achieved for ozone-reactive (>100 M^{-1} s^{-1}) compounds. For example, over 80% removal can be achieved for a compound with k_{O3} = 100 $M^{-1}s^{-1}$, only 20-30% removal can be achieved for a compound with k_{O3} = 10 $M^{-1}s^{-1}$. In addition, it can be found that the percentage removal also increases with the increasing R_{ct} and ozone exposure (Equation 7.8). Therefore, for a particular micropollutant (i.e. k_{O3} and k_{OH} are constant), higher removal can be achieved by increasing ozone exposure (higher concentration, longer contact time, or both) or by switching to AOPs (R_{ct} increases). For example, adding H_2O_2 into ozone (O_3/H_2O_2 AOP) increases the hydroxyl radicals production, thereby increasing the R_{ct}. According to Vincent *et al.* (2001), the logarithm scale R_{ct} is linearly related to the H_2O_2/O_3 mass ratio.

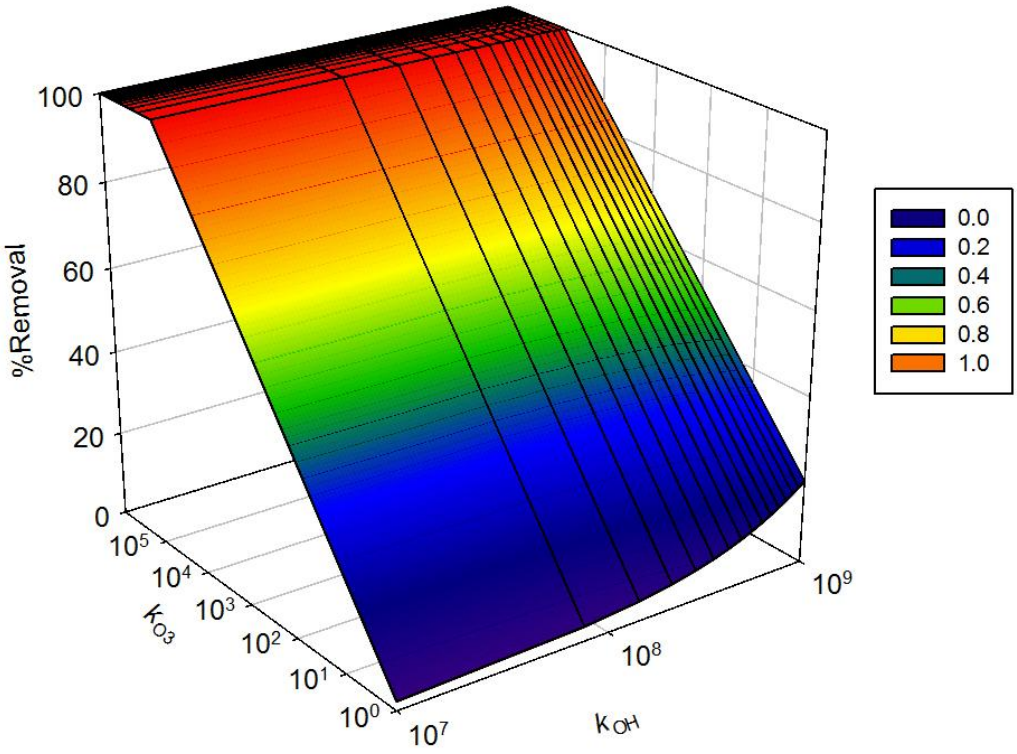
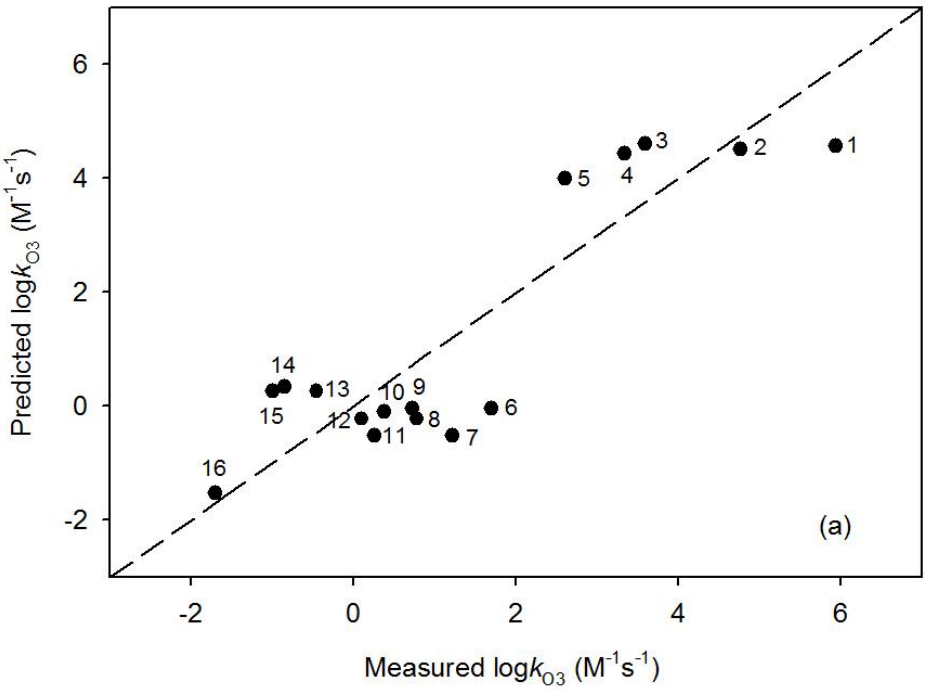


Figure 7.1 The theoretical relationship between the percent removal (%R) and the rate constants (k_{O3} and k_{OH}). Assume the R_{ct} value is 10^{-8} and the ozone exposure is 0.02 Ms.

A group of micropollutants with reported percentage removals, R_{ct} values, and ozone exposure were collected from the literature (Table 7.1). R_{ct} values were in the range of $10^{-9} – 10^{-7}$, and most commonly in the order of 10^{-8}. Their k_{O3} values were in the range of $0.02 – 8.7 \times 10^5$ $M^{-1}s^{-1}$ and k_{OH} in the range of $1.9 \times 10^9 – 8.95 \times 10^9$ $M^{-1}s^{-1}$.

Before predicting rate constants, first, the leverage values of all the selected micropollutants were calculated. All of the compounds fall into the applicability domain of QSPR models, except that the leverage value of butachlor (0.276) is higher than the warning leverage ($h^* = 0.270$) of the k_{OH} model indicating that the predicted k_{OH} of butachlor should be used with caution. Secondly, the k_{O3} and k_{OH} values of these micropollutants were predicted using the developed QSPR models (details in Chapters 5 and 6). As shown in Figure 7.2(a), the predicted k_{O3} values by QSPR models are fairly close to the experimentally determined values for most of the micropollutants. However, a few micropollutants scattered away from the regression line such as chlortoluron (#5) and 2-isopropyl-3-methoxypyrazine (#6). As shown in Figure 7.2(b), nearly all the QSPR model predicted k_{OH} values are close to their measured k_{OH} except for three compounds, butachlor (#9), acetochlor (#10), and propachlor (#12)

which were slightly underestimated. However, the differences between predicted and actual values are relative small.



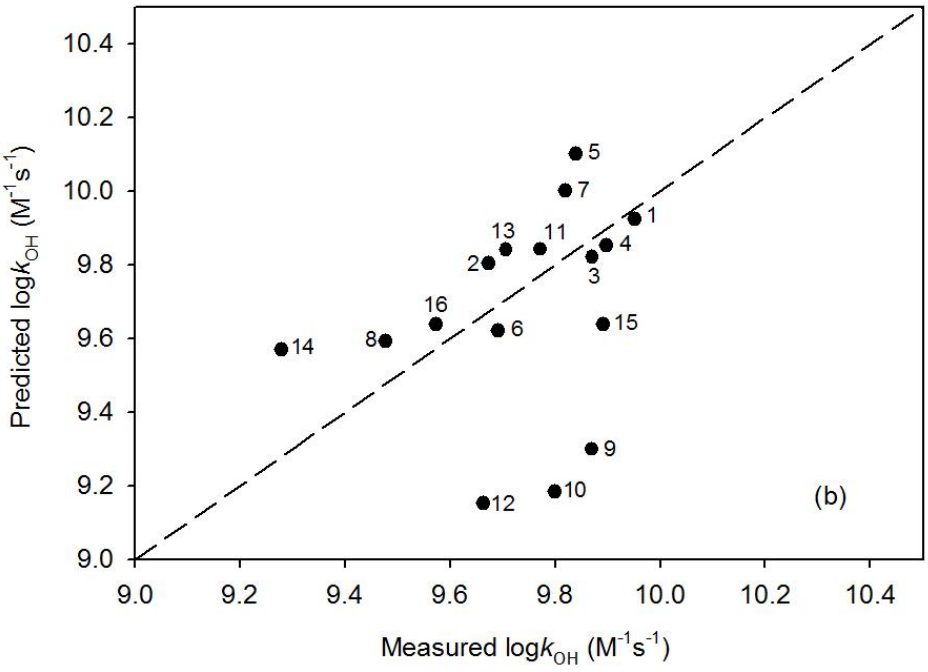


Figure 7.2 Predicted second-order rate constants vs. their experimental determined values, (a) ozone rate constants, (b) hydroxyl radical rate constants. The numbers of micropollutants are shown in Table 7.1.

The percentage removals of these micropollutants were then calculated (Table 7.1) and plotted against the observed removals (Figure 7.3). Most of the micropollutants were found close to the theoretical line indicating that the predictions of the percentage removal of these compounds agree well with the experimental data. However, three compounds (acetochlor, propachlor, and butachlor) were underestimated because of the error associated with their k_{OH} predictions. Overall, the results show that the developed QSPR models can provide reliable estimation of the k_{O3} and k_{OH}. Although the selected micropollutants are structurally diverse; the combination of R_{ct} with QSPR models provides a possible approach to estimate the percentage removals before experimentation.

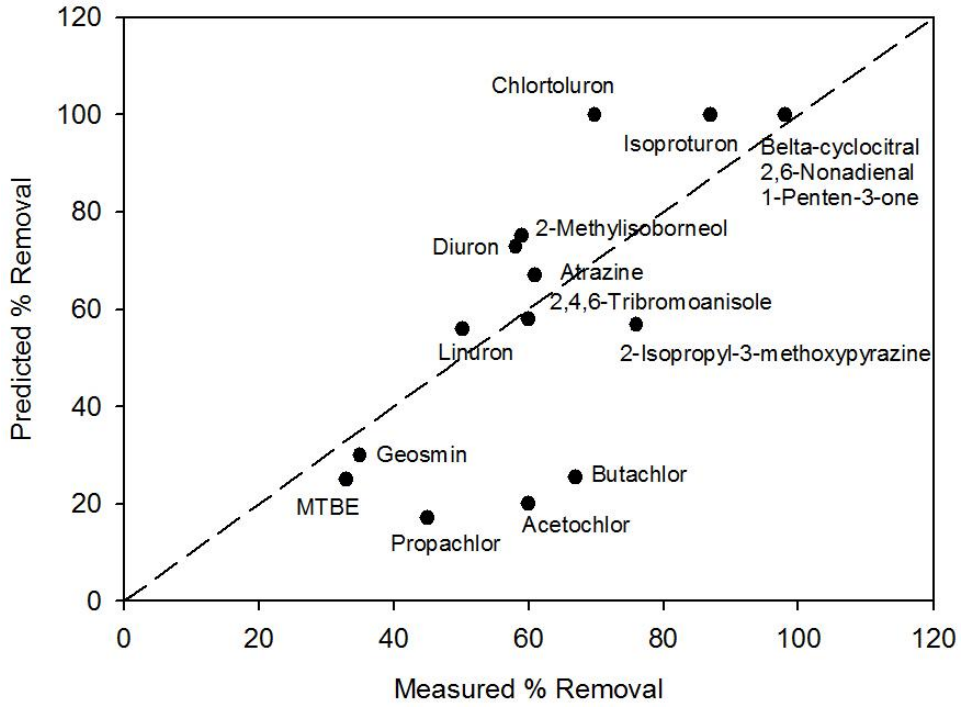
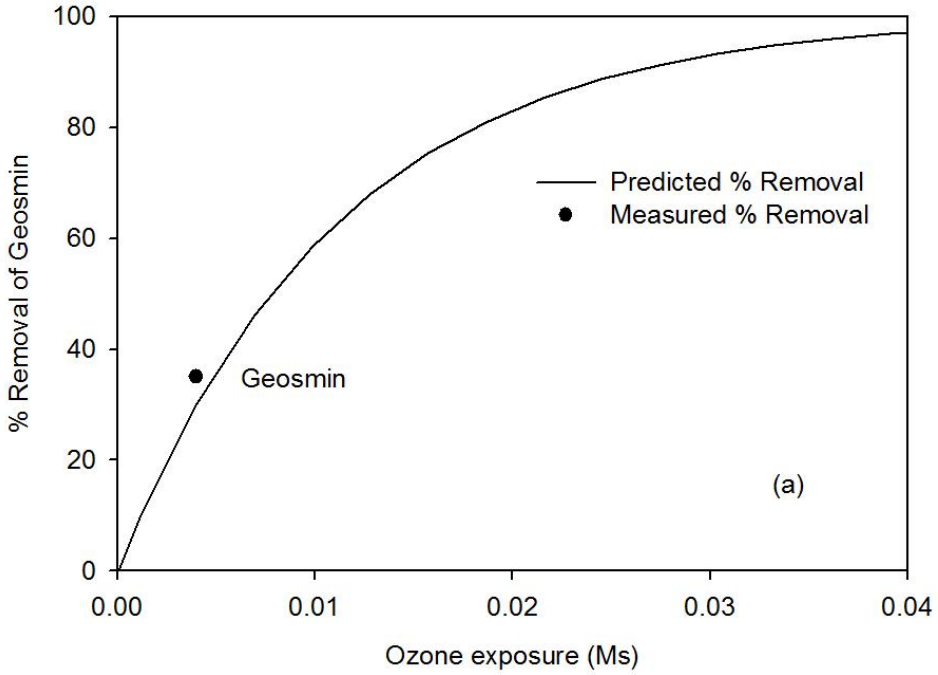


Figure 7.3 Predicted percentage removal vs. measured percentage removal.

Actions can be taken if the removals of micropollutants do not reach the treatment goal. For compounds with low removals, a few options are available to improve their removal: increasing the ozone exposure (not feasible for compounds with extremely low reactivity), adding H_2O_2 which accelerates hydroxyl radicals generation and thereby increasing R_{ct}, or considering other treatment

processes. For example, one of the taste and odour compounds, geosmin, shows low removal at about 35% (Figure 7.3) under the conditions given in the literature (Peter and von Gunten 2007). Its removal can theoretically be improved to over 80% if the ozone exposure increases from 0.004 to 0.02 Ms, or the R_{ct} increases from 2×10^{-8} to 10^{-7} (Figure 7.4).



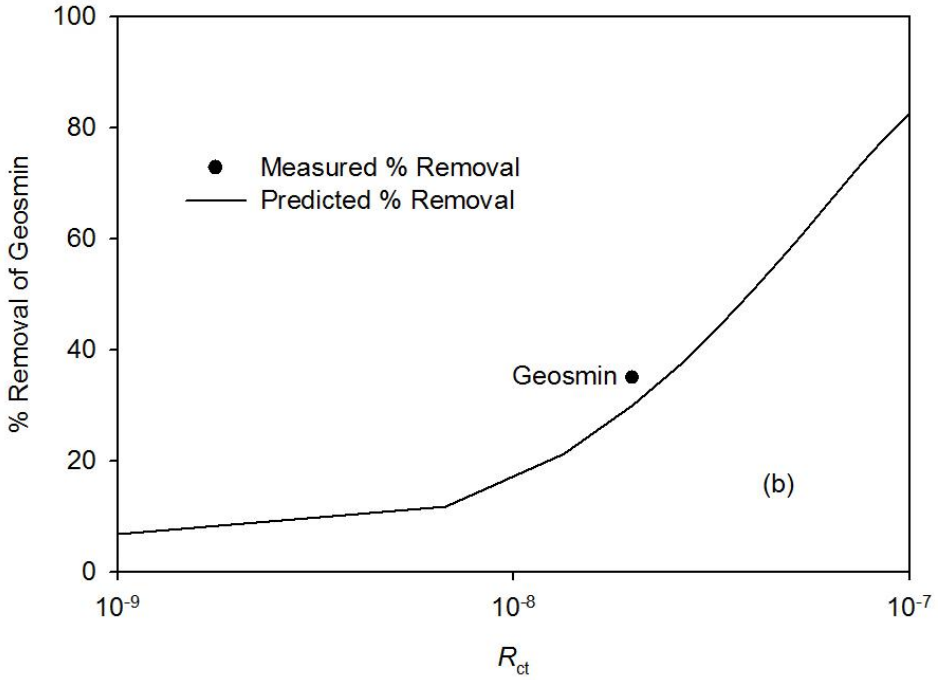


Figure 7.4 Predicted percentage removal of geosmin in the relationship with (a) ozone exposure, (b) R_{ct} values. The measured percentage removal was obtained from Peter and von Gunten (2007), and the prediction curve was obtained using the QSPR model predicted k_{O3} and k_{OH} of geosmin.

When applying the approach for prediction, the uncertainty of the prediction should be kept in mind. The overall uncertainty of the prediction is determined by the extent of all possible errors. These errors can be associated with, but are not limited to, the rate constant prediction, measurement of R_{ct} and ozone concentration, etc. Neumann *et al.* (2009) discussed the uncertainty associated with applying the R_{ct} model in pilot-scale reactors in detail and found that the source of uncertainty for predicting the removal of micropollutants largely depends on their reactivity, i.e. rate constants. For ozone-resistant micropollutants (for example, $k_{O3} < 10\ M^{-1}s^{-1}$), R_{ct} is the most influential factor which can explain most of the variance, whereas for micropollutants reacting fast with ozone (for example, $k_{O3} > 100\ mM^{-1}s^{-1}$), k_{O3} and reactor hydraulics are important sources of uncertainty. In this study, three compounds (acetochlor, propachlor, and butachlor) show relatively large errors when predicting their removals (Figure 7.3). They are all ozone-resistant compounds as indicated by their low k_{O3} values. Therefore, the direct pathway (i.e. oxidation with molecular ozone) is less important than the hydroxyl radical pathway. Thus the error in predicting k_{OH} is largely reflected in the removal estimation. On the other hand, chlortoluron is a fast-reacting compound and its predicted logk_{O3} differed by 1.4 from the measured logk_{O3}.This explains the difference between measured (70%) and predicted removal (99.8%).

Overall, the combination of R_{ct} with QSPR models is useful for estimating the removal efficiency of unknown micropollutants in natural water by ozonation treatment. It can be used as a tool for initial estimation and assessment for a given treatment goal. Water treatment professionals can apply this tool to determine conditions required to achieve a certain treatment goal i.e. a certain % removal of micropollutants during ozonation. First of all, the R_{ct} value of a given water source can be determined by monitoring the probe compound *p*CBA. Second, identify the micropollutants of interest, e.g., certain toxic compounds which are commonly found in the given water source based on previous knowledge or experience, or a group of representative compounds for screening purposes. Third, the k_{O3} and k_{OH} of compounds of interest can be estimated by the QSPR models. And finally, estimate the percent removal at a given level of ozone exposure using the R_{ct} model. These results will be useful in determining further course of action i.e. if ozonation is in principle a viable option

and should be followed up on, or if alternative treatment technologies including AOPs should be considered.

7.4 Conclusions

The R_{ct} model developed by Elovitz and von Gunten (1999) is useful in assessing the removal efficiency of micropollutants in natural water during ozonation. But the model input parameters k_{O3} and k_{OH} of the target micropollutants are not available for many emerging micropollutants, such as EDCs and PPCPs. With the QSPR models developed in this thesis which can provide estimations of k_{O3} and k_{OH} for untested micropollutants, R_{ct} model can be used for estimating the removal efficiency of many micropollutants, even though their rate constants are unknown. Sixteen micropollutants were collected from reported studies in natural waters as well as reported R_{ct} values. Their k_{O3} and k_{OH} were estimated by the QSPR models, and the predictive removal were then calculated and compared with the experimentally determined removals. The following conclusions can be drawn.

1. The QSPR models can provide reliable estimations on k_{O3} and k_{OH} for most of the selected micropollutants as the predicted rate constant are close to the experimentally determined values. Relative large errors were observed on k_{O3} of chlortoluron and 2-isopropyl-3-methoxypyrazine, and k_{OH} of butachlor, acetochlor, and propachlor.
2. A case study was conducted on geosmin which was not well removed under the conditions given in the literature. It was found that geosmin removal can theoretically be improved to over 80% if the ozone exposure increases from 0.004 to 0.02 Ms, or the R_{ct} increases from 2×10^{-8} to 10^{-7}.
3. The k_{O3} prediction is the main source of error for ozone-reactive compounds, and k_{OH} for ozone-resistant compounds.
4. The combination of the published R_{ct} model with QSPR models developed in this thesis for k_{O3} and k_{OH} prediction provides a valuable approach for estimating the removal efficiency of many micropollutants, even though their rate constants are unknown.

Chapter 8
Summary, Conclusions and Recommendations

8.1 Summary of the Thesis

The overall goal of this research was to develop reliable QSPR models which link the rate constants of a wide range of micropollutants in their reaction with ozone and hydroxyl radicals to their structural characteristics. The rate constants of numerous untested micropollutants can then be predicted without experimentation. Furthermore, the secondary objective was to assess the removal efficiency of ozonation and advanced oxidation processes. The percentage removal of micropollutants in natural waters during ozonation and AOPs can be estimated by combining the predicted rate constants with the existing models such as R_{ct} and $R_{OH,UV}$.

To develop QSPR models and explore their applications in natural waters, this research consisted of five major phases. The first phase included a literature review of the QSPR methodology and existing QSPR applications in ozonation and advanced oxidation processes (Chapter 2). In the first part of the literature review, the key elements of QSPR model development were reviewed, including selection of training set, selection of molecular descriptors, statistical methods for modeling, model evaluation, model validation, and applicability domain. The second part of the literature review focused on QSPR studies modeling rate constants of micropollutants in ozonation and advanced oxidation processes. Finally, knowledge gaps and research needs were identified and discussed.

The second phase put forth a systematic statistical approach for the selection of representative compounds from a large compound pool (Chapter 3). First, this approach identified and collected a pool of micropollutants based on reported occurrence in water and wastewater, and availability of treatment studies. Second, suitable molecular descriptors which link the structural characteristics of micropollutants to mechanisms of water treatment processes were identified. A relative small set of structural representative micropollutants (22 micropollutants) were then selected using principal component analysis and experimental design. Selected compounds cover the entire chemical domain in a well-balanced and efficient manner. The selected compounds served as training set for subsequent QSPR model development (Chapter 5).

The third phase of this research involved extensive laboratory analysis to determine the rate constants of the selected micropollutants (selected in Chapter 3) in their reactions with ozone and hydroxyl radicals (Chapter 4). Three methods (compound monitoring, ozone monitoring, and

competition kinetics) were used to determine the ozone rate constants, and competition kinetics was used to measure the hydroxyl radical rate constants. The results were in good agreement with literature data where available. The experimentally determined rate constants were used as model input for QSPR model development (Chapter 5 and 6).

The fourth phase included the development and validation of QSPR models for predicting the ozone rate constants (Chapter 5), and hydroxyl radical rate constants (Chapter 6). For the ozone rate constant QSPR models, the experimentally determined rate constants (from Chapter 4) were used as the training set; models were developed using piecewise linear regression, and the models were externally validated using data collected from the literature. For the hydroxyl radical rate constants QSPR models, experimental data and literature data were pooled together and then divided into training set and validation set. DRAGON descriptors were used to describe the chemical structure of the compounds. A very good model was developed using multiple linear regression.

The last phase of this research included an example of QSPR model (developed in this research) applications in natural water to assess the removal efficiency of micropollutants during ozonation (Chapter 7). Combined with the existing R_{ct} model, QSPR model predictions were shown to be suitable for providing an initial assessment of the removal efficiency of untested compounds during real-world treatment.

This research project was more complex than expected, and challenge were numerous and difficult to overcome. A few challenges throughout the development of the QSPR models were:

1. It is a challenge to develop QSPR models widely applicable to many structural diverse compounds. Most studies used a group of structurally relative homogeneous compounds as the training set. The similarity of the compounds generally ensures fairly high predictive power of the developed QSPR models. However, the applicability of these QSPR models is limited to a small range of compounds which are structural similar to the training set compounds. Training set selection is therefore very important because it determines the applicability of the QSPR model. A set of heterogeneous compounds with diverse structures is preferred when the goal of the QSPR model is to predict a wide range of compounds. To ensure the training set compounds were heterogeneous and limited in number, a systematic selection approach was modified and applied in this research (Chapter 3).

2. It is still a challenge to address both the non-dissociating compounds and dissociating compounds in a single QSPR model for rate constant prediction. Neutral and ionic species can react differently with oxidants such as ozone. Therefore, existing models mainly focus on the neutral species. This research is the first attempt to use the predominant species at neutral pH for the calculation of molecular descriptors, therefore expanding the applicability of QSPR models to dissociating compounds.

3. Selection and calculation of molecular descriptors were challenging. Numerous molecular descriptor were available, however, a set of descriptors with clear physical meanings which can aid in the interpretation of the mechanisms of the various treatment technologies, were not available. This thesis successfully identified a list of 12 descriptors which can relate the structural features to treatment mechanisms. Furthermore, many software packages (free or commercial) are available but have limitations. A number of software packages have to be used to calculate all the descriptors of interest.

4. The laboratory analysis was more complex than expected. Determination of the rate constants of 22 representative micropollutants was challenging and time-consuming. Several different methods had to be used to determine the ozone rate constants. Instrument methods (HPLC-PDA) had to be developed for all the micropollutants investigated. The quality of the analysis had to be carefully controlled to obtain reliable data. As a consequence the laboratory analysis took longer than expected.

5. Development of QSPR models involved advanced knowledge of statistical modeling techniques. Misuse or missing one or more key elements for QSPR modeling can lead to incorrect or poor models with low predictive power. The original plan was to use partial least squares (PLS) regression to build QSPR models, however, models developed with PLS regressions were not statistically satisfactory. Various modeling techniques were tested and compared with each other. Piecewise linear regression and multiple linear regression were finally applied to develop models for predicting ozone rate constants and hydroxyl radical rate constants, respectively.

8.2 Summary of Findings and Conclusions

A systematic selection approach (Chapter 3) which identifies representative micropollutants solely based on their physico-chemical and structural properties relevant in individual water treatment processes was modified and applied and the following conclusions can be drawn:

1. Physico-chemical properties (i.e. molecular descriptors) of micropollutants determine to a large extent their removal from drinking water. A set of 22 molecular descriptors which are relevant to the removal mechanisms of individual treatment processes (i.e. coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration) was identified. Only descriptors with clear physical meanings were included.
2. A systematic statistical approach combining principal component analysis and experimental design was modified and applied to a pool of heterogeneous micropollutants and their molecular descriptors. Principal component analysis summarized the variation in this original multivariate dataset and extracted latent variables, the principal components. D-optimal onion design was applied to these principal components to select structural representative compounds.
3. To demonstrate the applicability of the selection approach, it was applied to a pool of 182 micropollutants and two sets of 22 representative micropollutants were selected. The first set is suitable for experimental studies of a range of water treatment processes (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration) whereas the second set can be used for studying oxidation processes. The small number of selected micropollutants (22 out of 182) provided very good coverage over the entire property space and thus represented the original micropollutant pool well.
4. Maximum information on treatability of compounds with very diverse structures can be obtained with a minimum amount of experimental study when using the selected compounds, therefore making treatment studies more cost effective.
5. The selection approach presented here is flexible and can be customized to fit individual needs by for example reducing the number of compounds, applying it to other processes such as adsorption and/or membrane filtration, or studying other classes of micropollutants by re-defining the compound pool.

In Chapter 4, laboratory analysis was conducted to determine the rate constants of selected micropollutants in the reaction with ozone and hydroxyl radicals, from which we can make the following conclusions:

1. Three different methods had to be used to determine k_{O3} (at pH 7 and 20-22 ℃) because of the wide range of rate constants and the limitations of each method. The competition kinetics method was satisfactory to determine k_{OH} values (at pH 7 and 20-22 ℃) of all selected micropollutants in which the hydroxyl radicals were produced by UV/H_2O_2 since k_{OH} values varied over a comparatively small range.

2. For the majority of the micropollutants investigated k_{O3} and k_{OH} were not reported. Data provided herein are thus filling this data gap, and provide valuable information for modeling and design of ozonation and AOP treatment.

3. The k_{O3} values determined in this study ranged from 10^{-2} to 10^{7} M^{-1} s^{-1}. In general, compounds with activated aromatic rings such as a phenolic moiety, anisole, or aniline moiety show high reactivity (~10^{4} to 10^{7} M^{-1} s^{-1}) toward ozone with the exception of the chlorine substituted compound dicamba (~10^{-2} M^{-1} s^{-1}). Polycyclic aromatic hydrocarbons show moderate reactivity (~10^{3} to 10^{4} M^{-1} s^{-1}) and compounds with deactivated aromatic rings such as phthalate, organochlorine compounds, and X-ray contrast media show moderate to very low reactivity (~10^{-2} to 10^{2} M^{-1} s^{-1}) toward ozone. Saturated aliphatic compound such as organophosphorus compounds have a low reactivity (<10 $M^{-1}s^{-1}$) towards ozone as well. The general trend of micropollutant reactivity with ozone can be explained by the micropollutant structures and the electrophilic nature of ozone reactions.

4. All compounds are highly reactive toward hydroxyl radicals as shown by their high k_{OH} values confirming that the hydroxyl radicals are relatively non-selective oxidants.

5. For compounds with low reactivity toward ozone, ozonation treatment could be insufficient for removing them from drinking water, therefore hydroxyl radicals based treatment techniques such as O_3/H_2O_2 or UV/H_2O_2 are recommended.

QSPR models for predicting the rate constants of micropollutants in the reaction with ozone were developed in Chapter 5. We can draw the following conclusions:

1. QSPR models were developed with a set of 22 selected representative micropollutants as the training set, and a set of pre-selected molecular descriptors. Preliminary modeling with

stepwise MLR, partial least squares (PLS) regression, and principal component regression (PCR) failed to develop satisfactory models.

2. With a pre-defined breakpoint ($\log k_{O3} = 2.00\ M^{-1} s^{-1}$), the models developed by piecewise linear regression (PLR) show significant better results ($R^2 > 0.9$). In addition, the piecewise linear regression models were externally validated using data ($n = 33$) collected from literature, indicating good predictive power as shown by their high predictive R^2 (> 0.8).

3. A linear discriminant analysis (LDA) was carried out to classify the compounds into one of the two groups, high-reactive and low-reactive compounds. The resulting discriminant function shows great classification ability in both training set and validation set. With this function, new compounds can be easily classified into one of the two defined groups, and then predicted by the PLR models accordingly.

4. The applicability domains of the models were defined using the Williams plot approach based on leverage, so that the applicability of the models can be determined for new compounds. Predictions made for compounds outside of the applicability domain should be used with great caution.

5. Overall, the PLR-LDA approach provides the means to model the ozone rate constants of various, structural diverse compounds. The predicted k_{O3} is an indication of compound reactivity and therefore provides an initial assessment whether a compound can be treated with ozone at all. When combining the predicted k_{O3} with the R_{ct} model the percentage removal of these compounds in natural water can be assessed for varying ozonation conditions.

QSPR models for predicting the rate constants of micropollutants in the reaction with hydroxyl radicals were developed in Chapter 6, from which we can make the following conclusions:

1. A seven-variable model was developed using the training set. The main contribution to the rate constant was obtained from the mean atomic Sanderson electronegativity descriptor *Me*. In addition, model descriptors were also related to polarizability, double bonds, H-bond acceptors, etc. The importance of these descriptors can be explained by the H-atom abstraction and OH-addition mechanisms of the radical reaction.

2. The performance of the QSPR model was assessed by goodness-of-fit, robustness and predictivity (using validation set). The model fitted the training set very well as seen in the

adjusted R^2 = 0.823; the cross validated Q^2 = 0.773 and predictive R^2 = 0.772, all indicating good robustness and predictivity.

3. Outlier compounds were identified, one in the training set (Dalapon) and one in the validation set (DEHP). By removing the two outliers, the QSPR model was further improved as indicated by higher adjusted R^2 (0.846), Q^2 (0.804), and predictive R^2 (0.858).

The combination of the R_{ct} model with QSPR models to predict k_{O3} and k_{OH} values provides a possible approach for assessing the removal efficiency of many micropollutants, even though their rate constants are unknown. The applicability of this combination was explored in Chapter 7. We can draw the following conclusions:

1. The QSPR models can provide reliable estimations on k_{O3} and k_{OH} for most of the 16 selected micropollutants as the predicted rate constant are close to the experimentally determined values. Relative large errors were observed on k_{O3} of chlortoluron and 2-isopropyl-3-methoxypyrazine, and k_{OH} of butachlor, acetochlor, and propachlor.
2. A case study was conducted on geosmin which was not well removed under the condition given in the literature. It is found that its removal can theoretically be improved to over 80% if the ozone exposure increases from 0.004 to 0.02 Ms, or the R_{ct} increases from 2×10^{-8} to 10^{-7}.
3. The k_{O3} prediction is the main source of error for ozone-reactive compounds, and k_{OH} for ozone-resistant compounds.
4. The combination of the published R_{ct} model with QSPR models developed in this thesis for k_{O3} and k_{OH} prediction provides a valuable approach for estimating the removal efficiency of many micropollutants, even though their rate constants are unknown.

8.3 Future Directions and Implications for the Water Treatment Community

This research developed a systematic compound selection approach for water treatment screening studies. It is useful to identify a small group of compounds representing a large compound pool for the ease of the laboratory analysis and modeling, especially for emerging contaminants as it is impossible to experimentally study all of these contaminants. The future directions for this approach could be:

- Select representative compounds set for other treatment processes such as activate carbon adsorption, membrane filtration, biofiltration, etc. In addition, verify the representativeness of the selected compounds by conducting pilot-scale and full-scale studies. The ultimate goal is to recommend a representative compound lists to the water industry and obtain wide recognition.
- To aid in regulatory decisions. For example, US EPA is now reviewing the Contaminant Candidate List 3, from which contaminants may be identified for regulation based on grouping compounds. The selection approach introduced here can be adapted to identify similarity and dissimilarity among those contaminants, by grouping or clustering and identifying indicator compounds (e.g. worse case scenarios).
- The identified molecular descriptor set can be further improved as more knowledge becomes available. Better descriptors are needed to describe the electron status of the compounds especially for dissociating compounds.

This research developed QSPR models for predicting the rate constant of micropollutants in the reaction with ozone and hydroxyl radicals. The future direction for the QSPR modeling application for the water treatment could be:

- Improve the model application of pH-dependent property predictions. Ideally, a model which can be used at the entire pH range encountered in natural water is needed.
- Verify the applicability of the model by applying it together with the R_{ct} model in pilot scale and full-scale experiments.
- Assess the removal efficiency of micropollutants in UV/H_2O_2 AOP in natural water by combining the QSPR models with the $R_{OH,UV}$ model.
- Develop QSPR models for other oxidation and related processes in water treatment, e.g., chlorination, UV photolysis. Further, develop QSPR models for other water treatment processes such as activated carbon adsorption and membrane filtration.

Appendix A

Supplementary Material for Chapter 3

Selection of Representative Emerging Micropollutants for Drinking Water Treatment Studies: A Systematic Approach

Diffusivity Calculations

Liquid phase diffusivity of organic compound in water can be estimated by the formula below (Gnielinski, 1979), in which V_b can be estimated using the Tyn and Calus method (Reid *et al.,* 1977). V_c values are calculated by Molecular Modeling Pro Plus (MMP+) software (ChemSW Inc.) applying the Joback and Reid (1987) method.

$$Df = \frac{13.26 \times 10^{-5}}{\mu_l^{1.14} V_b^{0.589}} \quad \textbf{(A.1)}$$

$$V_b = 0.285 \times V_c^{1.048} \quad \textbf{(A.2)}$$

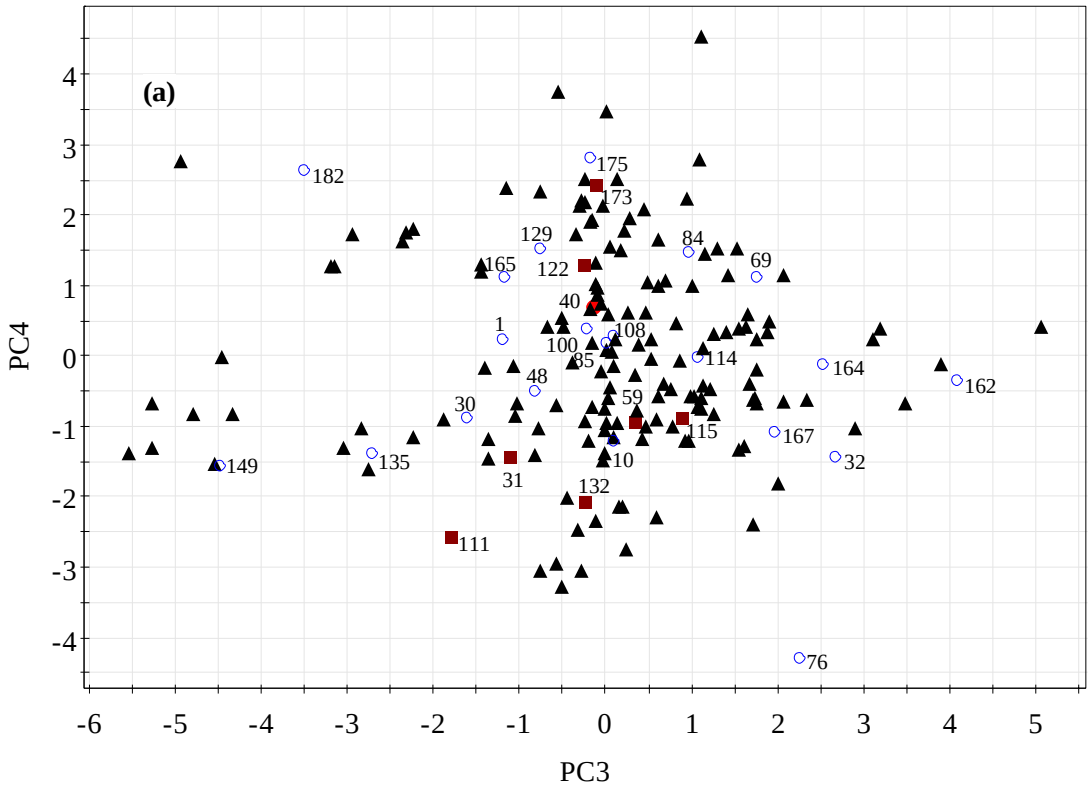
Where:

Df the diffusivity of organic compound in water (cm^2/s)

μ_l the viscosity of water (centipoise), μ_w = 1.003 centipoise = 1.003×10^{-3} Pa·s (20°C)

V_b the molar volume at the boiling point temperature (cm^3/mol)

V_c the critical volume (cm^3/mol)


(a)
PC4
PC3
4
3
2
1
0
-1
-2
-3
-4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
182
175
173
129
84
69
165
122
40
1
108
100
85
114
164
162
48
30
59
115
167
149
135
10
32
31
132
111
76

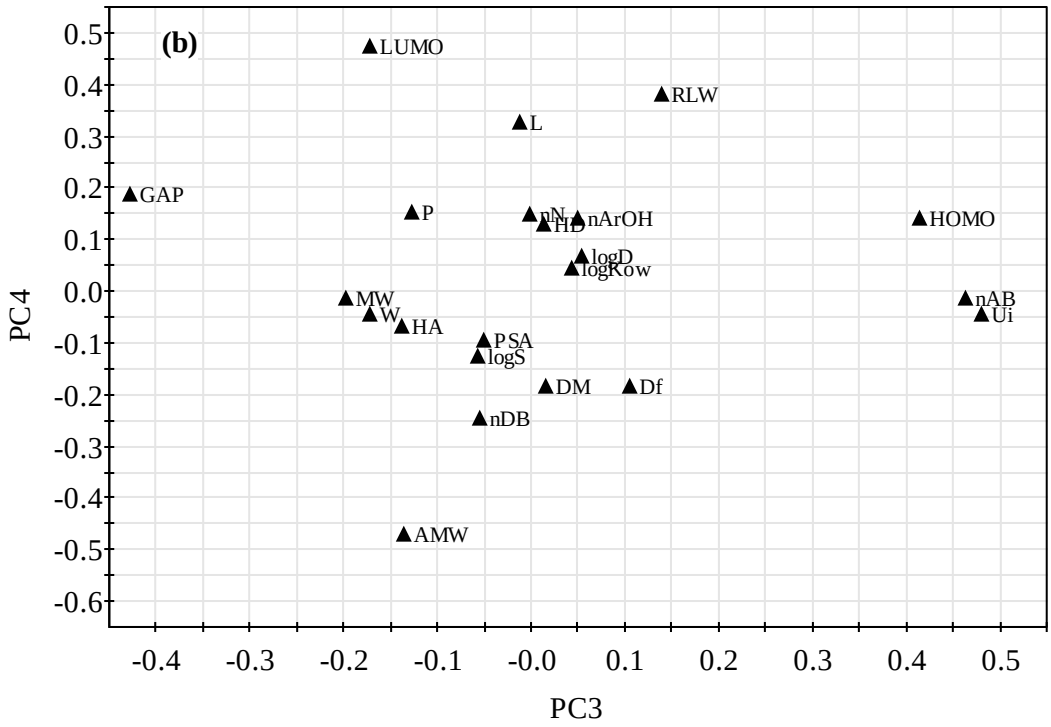
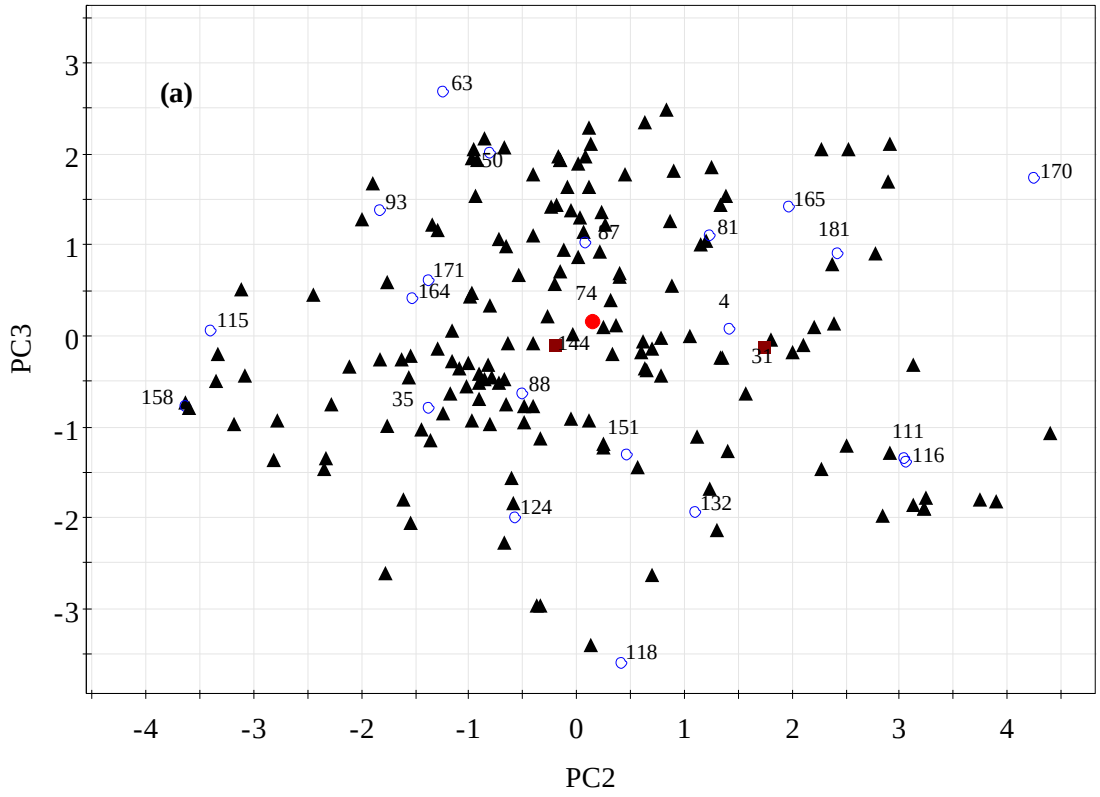


Figure A.1 PCA analysis for treatment set 1 (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration). (a) Score plot of principle component three (PC3) and four (PC4). D-optimal onion design applied (3 layers) for compound selection. Black triangles represent compounds not selected, blue circles represent selected compounds, the red dot represents the center compound and purple boxes represent compounds selected to replace similar compounds; (b) Loading plot of PC3 and PC4. The meaning of the abbreviations can be found in the main manuscript Table 3.2.


(a)
63
150
93
165
81
87
181
171
164
74
4
115
144
31
158
35
88
151
111
116
124
132
118
PC3
PC2

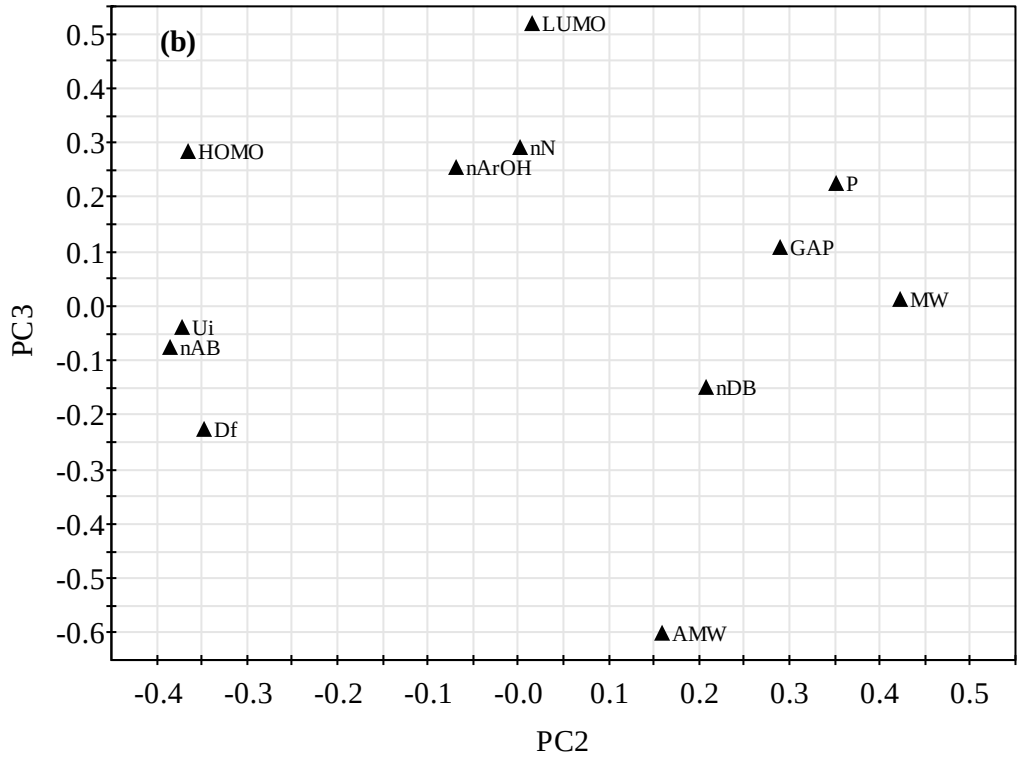

(b)
LUMO
HOMO
nN
nArOH
P
GAP
MW
Ui
nAB
nDB
Df
AMW
PC3
PC2

Figure A.2 PCA analysis for treatment set 2 (oxidation processes). (a) Score plot of principle component two (PC2) and three (PC3). D-optimal onion design applied (3 layers) for training set selection. Black triangles represent compounds not selected, blue circles represent selected compounds, the red dot represents the center compound and purple boxes represent compounds selected to replace similar compounds; (b) Loading plot of PC2 and PC3. The meaning of the abbreviations can be found in the main manuscript Table 3.2.

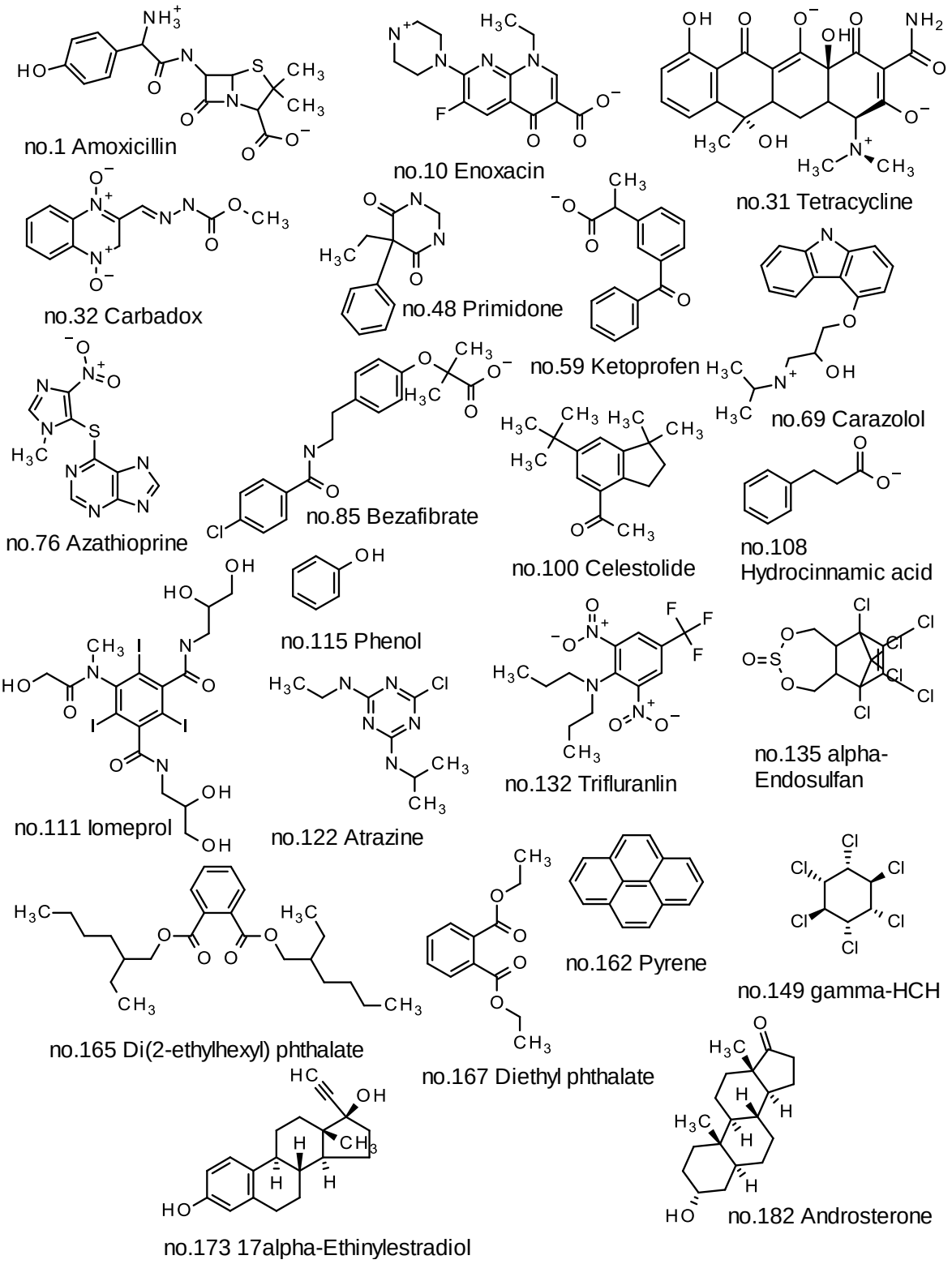


Figure A.3 Structures of micropollutants selected for water treatment set 1 (coagulation/flocculation, oxidation, activated carbon adsorption, and membrane filtration).

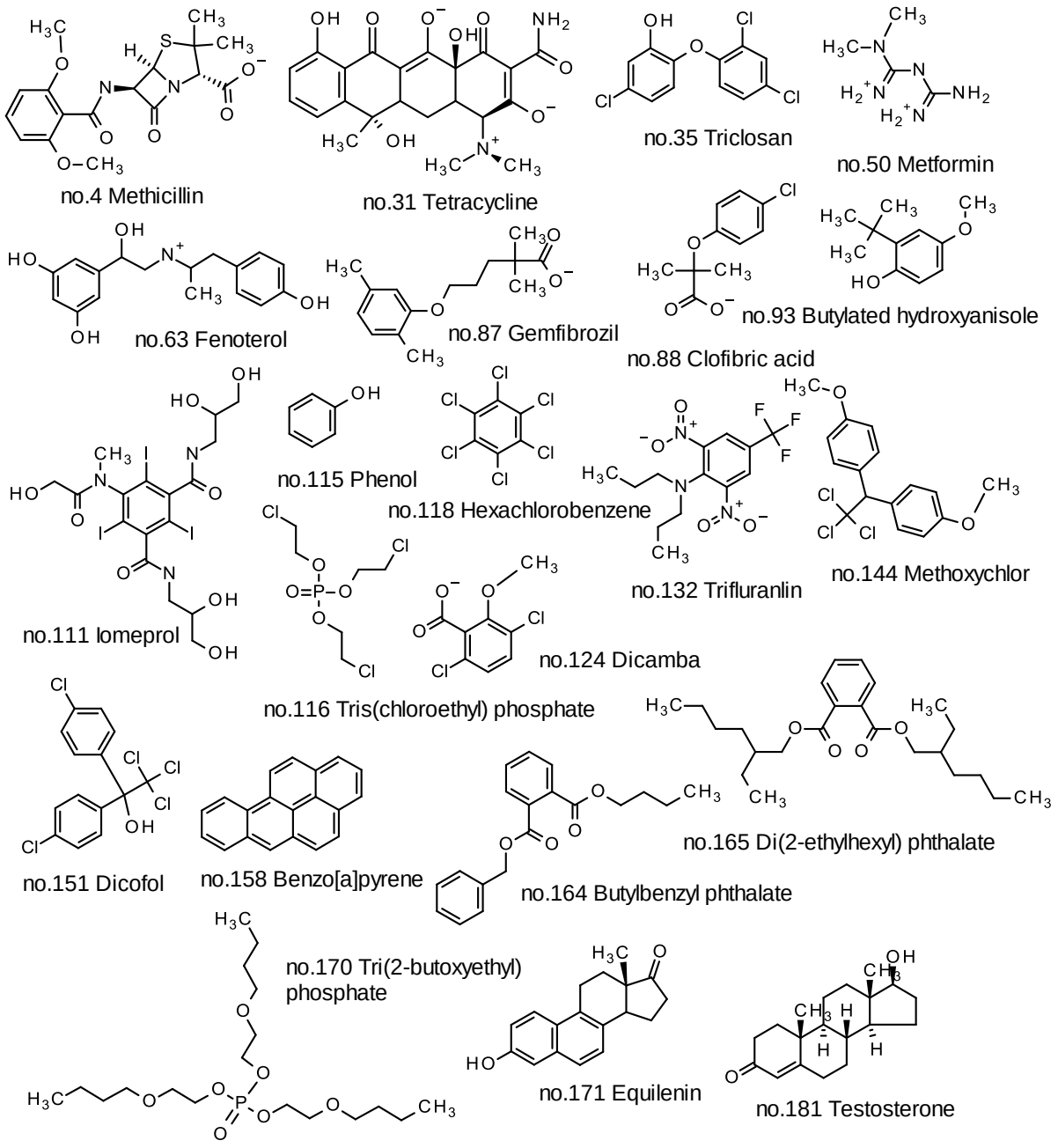


Figure A.4 Structures of micropollutants selected for water treatment set 2 (oxidation processes).

Table A.1 Micropollutants included in this study and their occurrence in water (*n* = 182).

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Reference
			Veterinary and Human Antibiotics (*n* = 36)			
1	Amoxicillin	61336-70-7	365.40	β-lactam	Max. in STP [b] effluents: 0.12 μg/L	Andreozzi *et al.*, 2004
2	Cloxacillin	61-72-3	435.88	β-lactam	Not detected in STP effluents	Hirsch *et al.*, 1999
3	Dicloxacillin	3116-76-5	470.33	β-lactam	Not detected in STP effluents	Hirsch *et al.*, 1999
4	Methicillin	61-32-5	380.41	β-lactam	Not detected in STP effluents	Hirsch *et al.*, 1999
5	Penicillin G	61-33-6	334.39	β-lactam	Not detected in STP effluents	Hirsch *et al.*, 1999
6	Sultamicillin	76497-13-7	594.65	β-lactam	Not detected in STP effluents	Cokgor *et al.*, 2004
7	Clindamycin	18323-44-9	424.98	Macrolide	Max. in surface waters: 1.1 μg/L	Batt *et al.*, 2005
8	Lincomycin	154-21-2	406.54	Macrolide	Max. in surface waters: 0.73 μg/L	Kolpin *et al.*, 2002
9	Ciprofloxacin	85721-33-1	331.35	Quinolone	Max. in surface waters: 0.03 μg/L	Kolpin *et al.*, 2002
10	Enoxacin	74011-58-8	320.32	Quinolone	Max. in STP effluents: 0.03 μg/L	Andreozzi *et al.*, 2003
11	Enrofloxacin	93106-60-6	359.40	Quinolone	STP effluents: 0.10 μg/L	Batt *et al.*, 2005
12	Levofloxacin	100986-85-4	361.37	Quinolone	Detected in STP effluents	Yasojima *et al.*, 2006
13	Lomefloxacin	98079-51-7	351.35	Quinolone	Max. in STP effluents: 0.32 μg/L	Andreozzi *et al.*, 2003
14	Norfloxacin	70458-96-7	319.33	Quinolone	Max. in surface waters: 0.12 μg/L	Kolpin *et al.*, 2002
15	Ofloxacin	83380-47-6	361.37	Quinolone	Max. in STP effluents: 0.2 μg/L	Nakata *et al.*, 2005
16	Sulfacetamide	144-80-9	214.24	Sulfonamide	Max. in STP effluents: 0.151 μg/L	Miao *et al.*, 2004
17	Sulfachlorpyridazine	80-32-0	284.72	Sulfonamide	Not detected in STP effluents	Adams *et al.*, 2002
18	Sulfadiazine	68-35-9	250.27	Sulfonamide	Max. in STP effluents: 0.019 μg/L	Miao *et al.*, 2004
19	Sulfadimethoxine	122-11-2	310.33	Sulfonamide	Max. in surface waters: 0.06 μg/L	Kolpin *et al.*, 2002
20	Sulfamerazine	127-79-7	264.30	Sulfonamide	Detected in STP effluents	Heberer 2002
21	Sulfamethazine	57-68-1	277.34	Sulfonamide	Max. in surface waters: 0.22 μg/L	Kolpin *et al.*, 2002
22	Sulfamethoxazole	723-46-6	253.28	Sulfonamide	Max. in surface waters: 1.9 μg/L	Kolpin *et al.*, 2002
23	Sulfamethizole	144-82-1	270.32	Sulfonamide	Max. in surface waters: 0.13 μg/L	Kolpin *et al.*, 2002
24	Sulfamoxole	729-99-7	267.30	Sulfonamide	Not detected in STP effluents	Adams *et al.*, 2002
25	Sulfapyridine	144-83-2	249.29	Sulfonamide	Max. in STP effluents: 0.228 μg/L	Miao *et al.*, 2004
26	Sulfathiazole	72-14-0	255.31	Sulfonamide	Not detected in STP effluents	Adams *et al.*, 2002
27	Sulfisoxazole	127-69-5	267.30	Sulfonamide	Max. in STP effluents: 0.034 μg/L	Miao *et al.*, 2004

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Ref.
28	Chlortetracycline	57-62-5	478.89	Tetracycline	Max. in surface waters: 0.69 μg/L	Kolpin *et al.*, 2002
29	Doxycycline	564-25-0	462.46	Tetracycline	Not detected in STP effluents	Miao *et al.*, 2004
30	Oxytetracycline	79-57-2	460.44	Tetracycline	Max. in surface waters: 0.34 μg/L	Kolpin *et al.*, 2002
31	Tetracycline	60-54-8	444.44	Tetracycline	Max. in surface waters: 0.11 μg/L	Kolpin *et al.*, 2002
32	Carbadox	6804-07-5	262.22	Others	Not detected in STP effluents	Adams *et al.*, 2002
33	Chloramphenicol	56-75-7	323.13	Others	Max. in surface waters: 0.06 μg/L Max. in STP effluents: 0.56 μg/L	Ternes *et al.*, 2003
34	Spectinomycin	1695-77-8	332.25	Others	Not detected in STP effluents	Adams *et al.*, 2002
35	Triclosan	3380-34-5	289.55	Others	Max. in surface waters: 2.3 μg/L	Kolpin *et al.*, 2002
36	Trimethoprim	738-70-5	290.32	Others	Max. in surface waters: 0.71 μg/L	Kolpin *et al.*, 2002
	Prescription and Nonprescription Drugs (*n* = 53)					
37	Acetaminophen	103-90-2	151.16	Analgesic	Max. in surface waters: 10 μg/L	Kolpin *et al.*, 2002
38	Acetylsalicylic acid	50-78-2	180.16	Analgesic	Max. in surface waters: 0.34 μg/L Max. in STP effluents: 1.5 μg/L	Ternes 1998
39	Codeine	76-57-3	299.37	Analgesic	Max. in surface waters: 1.0 μg/L	Kolpin *et al.*, 2002
40	Hydrocodone	125-29-1	299.37	Analgesic	Detected in STP effluents	Heberer 2002
41	Phenazone	60-80-0	188.23	Analgesic	Max. in surface waters: 0.95 μg/L Max STP effluents: 0.41 μg/L	Ternes 1998
42	Cimetidine	51481-61-9	252.34	Antacid	Max. in surface waters: 0.58 μg/L	Kolpin *et al.*, 2002
43	Ranitidine	66357-35-5	314.40	Antacid	Max. in surface waters: 0.01 μg/L	Kolpin *et al.*, 2002
44	Buspirone	36505-84-7	385.51	Anti-anxiety agent	Not detected in STP effluents	Calza *et al.*, 2004
45	Diazepam	439-14-5	284.74	Anti-anxiety agent	Max. in STP effluents: 0.04 μg/L	Ternes 1998
46	Warfarin	81-81-2	308.33	Anticoagulant	Not detected in STP effluents	Kolpin *et al.*, 2002
47	Carbamazepine	298-46-4	236.27	Anticonvulsant	Max. in surface waters: 1.1 μg/L Max. in STP effluents: 6.3 μg/L	Ternes 1998
48	Primidone	125-33-7	218.25	Anticonvulsant	Not detected in STP effluents	Ternes *et al.*, 2002
49	Fluoxetine	54910-89-3	309.33	Antidepressant	Max. in surface waters: 0.012 μg/L	Kolpin *et al.*, 2002
50	Metformin	657-24-9	129.16	Anti-diabetic	Max. in surface waters: 0.15 μg/L	Kolpin *et al.*, 2002
51	Diphenhydramine	58-73-1	255.36	Antihistamine	Detected in raw water samples	Stackelberg *et al.*, 2004
52	Diltiazem	42399-41-7	414.52	Antihypertensive	Max. in surface waters: 0.049 μg/L	Kolpin *et al.*, 2002

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Ref.
53	Enalaprilat	76420-72-9	348.40	Antihypertensive metabolite	Max. in surface waters: 0.046 μg/L	Kolpin *et al.*, 2002
54	Aminopyrine	58-15-1	231.30	Anti-inflammatory	Max. in STP effluents: 0.43 μg/L	Andreozzi *et al.*, 2003
55	Fenoprofen	31879-05-7	242.27	Anti-inflammatory	Max. in STP effluents: 0.28 μg/L	Andreozzi *et al.*, 2003
56	Flurbiprofen	5104-49-4	244.26	Anti-inflammatory	Max. in STP effluents: 0.34 μg/L	Andreozzi *et al.*, 2003
57	Ibuprofen	15687-27-1	206.28	Anti-inflammatory	Max. in surface waters: 1.0 μg/L	Kolpin *et al.*, 2002
58	Indomethacin	53-86-1	357.79	Anti-inflammatory	STP effluents: 0.10 μg/L	Ternes *et al.*, 2003
59	Ketoprofen	22071-15-4	254.28	Anti-inflammatory	Max. in surface waters: 0.12 μg/L Max STP effluents: 0.38 μg/L	Ternes 1998
60	Naproxen	22204-53-1	230.26	Anti-inflammatory	Max. in surface waters: 0.39 μg/L Max. in STP effluents: 0.52 μg/L	Ternes 1998
61	Crotamiton	483-63-6	203.28	Antipruritic	Max. in STP effluent: 0.365 μg/L	Nakada *et al.*, 2006
62	Clenbuterol	37148-27-9	277.19	β_2-sympathomimetics	Max. in surface waters: 0.050 μg/L Max. in STP effluents: 0.08 μg/L	Ternes 1998
63	Fenoterol	13392-18-2	312.41	β_2-sympathomimetics	Max. in surface waters: 0.061 μg/L Max. in STP effluents: 0.060 μg/L	Ternes 1998
64	Salbutamol	18559-94-9	239.31	β_2-sympathomimetics	Max. in surface waters: 0.035 μg/L Max. in STP effluents: 0.17 μg/L	Ternes 1998
65	Terbutaline	23031-25-6	225.29	β_2-sympathomimetics	Max. in STP effluents: 0.12 μg/L	Ternes 1998
66	Atenolol	29122-68-7	266.34	β-blocker	STP effluents: 0.36 μg/L	Ternes *et al.*, 2003
67	Betaxolol	63659-18-7	307.43	β-blocker	Max. in surface waters: 0.028 μg/L Max. in STP effluents: 0.19 μg/L	Ternes 1998
68	Bisoprolol	66722-44-9	325.45	β-blocker	Max. in surface waters: 2.9 μg/L Max. in STP effluents: 0.37 μg/L	Ternes 1998
69	Carazolol	57775-29-8	298.38	β-blocker	Max. in surface waters: 0.11 μg/L Max. in STP effluents: 0.12 μg/L	Ternes 1998
70	Celiprolol	56980-93-9	379.50	β-blocker	STP effluents: 0.28 μg/L	Ternes *et al.*, 2003
71	Metoprolol	37350-58-6	267.37	β-blocker	Max. in surface waters: 2.2 μg/L Max. in STP effluents: 2.2 μg/L	Ternes 1998
72	Nadolol	42200-33-9	309.40	β-blocker	Max STP effluents: 0.06 μg/L	Ternes 1998

73	Propranolol	525-66-6	259.34	β-blocker	Max. in surface waters: 0.59 μg/L Max. in STP effluents: 0.29 μg/L	Ternes 1998
74	Sotalol	3930-20-9	272.36	β-blocker	STP effluents: 1.32 μg/L	Ternes *et al.*, 2003
75	Timolol	26839-75-8	316.42	β-blocker	Max. in surface waters: 0.01 μg/L Max. in STP effluents: 0.07 μg/L	Ternes 1998
76	Azathioprine	446-86-6	277.26	Cytostatic drug	Not detected in STP effluents	Rey *et al.*, 1999
77	Cyclophosphamide	50-18-0	261.09	Cytostatic drug	Max. in STP effluent: 0.020 μg/L	Ternes, 1998
78	Cytarabine	147-94-4	243.22	Cytostatic drug	Not detected in STP effluents	Rey *et al.*, 1999
79	Daunorubicin	20830-81-3	527.53	Cytostatic drug	Not detected in STP effluents	Castegnaro *et al*, 1997
80	Doxorubicin	23214-92-8	543.53	Cytostatic drug	Not detected in STP effluents	Castegnaro *et al*, 1997
81	Epirubicin	56420-45-2	543.53	Cytostatic drug	Not detected in STP effluents	Castegnaro *et al*, 1997
82	Idarubicin	58957-92-9	497.50	Cytostatic drug	Not detected in STP effluents	Castegnaro *et al*, 1997
83	Ifosfamid	3778-73-2	261.09	Cytostatic drug	Max. in STP effluent: 2.9 μg/L	Ternes 1998
84	Methotrexate	59-05-2	454.44	Cytostatic drug	Not detected in STP effluents	Rey *et al.*, 1999
85	Bezafibrate	41859-67-0	361.82	Lipid regulator	Max. in surface waters: 3.1 μg/L Max. in STP effluent: 4.6 μg/L	Ternes 1998
86	Fenofibrate	49562-28-9	360.84	Lipid regulator	Max. in STP effluents: 0.03 μg/L	Ternes 1998
87	Gemfibrozil	25812-30-0	250.34	Lipid regulator	Max. in surface waters: 0.79 μg/L	Kolpin *et al.*, 2002
88	Clofibric acid	882-09-7	214.65	Metabolites of lipid regulator	Max. in surface waters: 0.55 μg/L Max. in STP effluent: 1.6 μg/L	Ternes 1998
89	Fenofibric acid	42017-89-0	318.84	Metabolites of lipid regulator	Max. in surface waters: 0.28 μg/L Max. in STP effluent: 1.2 μg/L	Ternes 1998
	Personal Care Products (*n* = 23)					
90	2,6-Di-tert-butyl-p-benzoquinone	719-22-2	220.31	Antioxidant	Max. in surface waters: 0.46 μg/L	Kolpin *et al.*, 2002
91	2,6-Di-tert-butylphenol	128-39-2	206.33	Antioxidant	Max. in surface waters: 0.11 μg/L	Kolpin *et al.*, 2002
92	Butylated hydroxy toluene	128-37-0	220.35	Antioxidant	Max. in surface waters: 0.1 μg/L	Kolpin *et al.*, 2002
93	Butylated hydroxyanisole^	25013-16-5	180.25	Antioxidant	Max. in surface waters: 0.2 μg/L	Kolpin *et al.*, 2002
94	1,7-Dimethylxanthine	611-59-6	180.17	Caffeine metabolite	Max. in surface waters: 3.1 μg/L	Kolpin *et al.*, 2002
95	1,4-	106-46-7	147.00	Deodorizer	Max. in surface waters: 4.3 μg/L	Kolpin *et al.*, 2002

	Dichlorobenzene*[#]					
96	4-Nonylphenol	104-40-5	220.35	Detergent metabolite	Max. in surface waters: 40 μg/L	Kolpin *et al.*, 2002
97	4-tert-Butylphenol	98-54-4	150.22	Detergent metabolite	Max. in surface waters: 0.13 μg/L	Brossa *et al.*, 2005
98	4-tert-Octylphenol	140-66-9	206.33	Detergent metabolite	Max. in surface waters: 0.19 μg/L	Brossa *et al.*, 2005
99	Acetophenone	98-86-2	120.15	Fragrance	Max. in surface waters: 0.41 μg/L	Kolpin *et al.*, 2002
100	Celestolide	13171-00-1	244.38	Musk fragrance	Max. in surface waters: 0.008 μg/L	Winkler *et al.*, 1998
101	Galaxolide	1222-05-5	258.40	Musk fragrance	Max. in surface waters: 0.152 μg/L	Winkler *et al.*, 1998
102	Musk ketone	81-14-1	294.31	Musk fragrance	Max. in surface waters: 0.010 μg/L	Winkler *et al.*, 1998
103	Tonalide	21145-77-7	258.40	Musk fragrance	Max. in surface waters: 0.088 μg/L	Winkler *et al.*, 1998
104	Cotinine	486-56-6	176.22	Nicotine metabolite	Max. in surface waters: 0.90 μg/L	Kolpin *et al.*, 2002
105	Caffeine	58-08-2	194.19	Stimulant	Max. in surface waters: 6.0 μg/L	Kolpin *et al.*, 2002
106	Salicyclic acid	69-72-7	138.12	Stimulant	Max. in surface waters: 4.1 μg/L Max. in STP effluents: 0.14 μg/L	Ternes 1998
107	Benzophenone	119-61-9	182.22	Sunscreen	Reclaimed wastewater: 0.993 μg/L	Loraine *et al.*, 2006
108	Hydrocinnamic acid	501-52-0	150.18	Sunscreen	Max. in raw drinking water: 20.3 μg/L	Loraine *et al.*, 2006
109	Octyl methoxycinnamate	5466-77-3	290.40	Sunscreen	Max. in raw drinking water: 5.61 μg/L	Loraine *et al.*, 2006
110	Oxybenzone	131-57-7	228.25	Sunscreen	Reclaimed wastewater: 0.84 μg/L	Loraine *et al.*, 2006
111	Iomeprol	78649-41-9	777.09	X-ray contrast media	STP effluents: 2.3 μg/L	Ternes *et al.*, 2003
112	Iopamidol	62883-00-5	777.09	X-ray contrast media	Max. in STP effluents: 15 μg/L	Ternes *et al.*, 2000
	Other Wastewater-Related Compounds (*n* = 59)					
113	5-Methyl-1H-benzotriazole	136-85-6	133.15	Anticorrosive	Max. in surface waters: 2.4 μg/L	Kolpin *et al.*, 2002
114	4-Methyl phenol	106-44-5	108.14	Disinfectant	Max. in surface waters: 0.54 μg/L	Kolpin *et al.*, 2002
115	Phenol	108-95-2	94.11	Disinfectant	Max. in surface waters: 1.3 μg/L	Kolpin *et al.*, 2002
116	Tris(chloroethyl) phosphate	115-96-8	285.49	Fire retardant	Max. in surface waters: 0.54 μg/L	Kolpin *et al.*, 2002
117	Tris(1,3-dichloroisopropyl) phosphate	13674-87-8	430.91	Fire retardant	Max. in surface waters: 0.16 μg/L	Kolpin *et al.*, 2002
118	Hexachlorobenzene[#]	118-74-1	284.78	Fungicide	Max. in surface waters: 0.14 μg/L	Brossa *et al.*, 2005
119	Thiabendazole	148-79-8	201.25	Fungicide	Detected in raw water samples	Stackelberg *et al.*, 2004
120	2,4-Dichloro-phenoxyacetic acid*[#]	94-75-7	221.04	Herbicide	Max. in surface waters: 2.67 μg/L	Grover *et al.*, 1997

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Ref.
121	Acetochlor^	34256-82-1	269.771	Herbicide	Max. in surface waters: 0.0015 μg/L	Xue *et al.,* 2006
122	Atrazine*#	1912-24-9	215.69	Herbicide	Max. in surface waters: 0.16 μg/L	Brossa *et al.,* 2005
123	Bromoxynil*	1689-84-5	276.915	Herbicide	Max. in surface waters: 0.33 μg/L	Grover *et al.,* 1997
124	Dicamba*	1918-00-9	221.04	Herbicide	Max. in surface waters: 11.2 μg/L	Grover *et al.,* 1997
125	Diuron*^	330-54-1	233.10	Herbicide	Max. in surface waters: 0.06 μg/L	Brossa *et al.,* 2005
126	MCPA*	94-74-6	200.621	Herbicide	Max. in surface waters: 1.97 μg/L	Grover *et al.,* 1997
127	Metolachlor*^	51218-45-2	283.797	Herbicide	Max. in surface waters: 0.027 μg/L	Xue *et al.,* 2006
128	Pentachlorophenol*#	87-86-5	266.34	Herbicide	Max. in surface waters: 0.03 μg/L	Brossa *et al.,* 2005
129	Prometone	1610-18-0	225.29	Herbicide	Max. in finished water: 0.096 μg/L	Stackelberg *et al.,* 2004
130	Simazine*#	122-34-9	201.66	Herbicide	Max. in surface waters: 0.016 μg/L	Loos *et al.,* 2007
131	Triallate	2303-17-5	304.662	Herbicide	Max. in surface waters: 0.87 μg/L	Grover *et al.,* 1997
132	Trifluralin*	1582-09-8	335.282	Herbicide	Max. in surface waters: 0.11 μg/L	Grover *et al.,* 1997
133	Alachlor#	15972-60-8	269.771	Insecticide	Max. in surface waters: 0.0057 μg/L	Xue *et al.,* 2006
134	Aldrin*	309-00-2	364.91	Insecticide	Max. in surface waters: 0.11 μg/L	Brossa *et al.,* 2005
135	α-Endosulfan	959-98-8	406.92	Insecticide	Max. in surface waters: 1.60 μg/L	Brossa *et al.,* 2005
136	Carbaryl*	63-25-2	201.22	Insecticide	Max. in surface waters: 0.1 μg/L	Kolpin *et al.,* 2002
137	Carbazole	86-74-8	167.21	Insecticide	Detected in raw water samples	Stackelberg *et al.,* 2004
138	Chlorpyrifos*	2921-88-2	350.58	Insecticide	Max. in surface waters: 0.31 μg/L	Kolpin *et al.,* 2002
139	α-Chlordane#	5103-71-9	409.78	Insecticide	Max. in surface waters: 0.1 μg/L	Kolpin *et al.,* 2002
140	Deltamethrin	52918-63-5	505.205	Insecticide	Max. in surface waters: 0.0063 μg/L	Xue *et al.,* 2006
141	Diazinon*	333-41-5	304.34	Insecticide	Max. in surface waters: 0.35 μg/L	Kolpin *et al.,* 2002
142	Dieldrin*	60-57-1	380.91	Insecticide	Max. in surface waters: 0.21 μg/L	Kolpin *et al.,* 2002
143	Heptachlor#	76-44-8	373.321	Insecticide	Max. in surface waters: 0.0085 μg/L	Xue *et al.,* 2006
144	Methoxychlor*#	72-43-5	345.652	Insecticide	Max. in surface waters: 0.022 μg/L	Xue *et al.,* 2006
145	Methyl parathion	298-00-0	263.20	Insecticide	Max. in surface waters: 0.01 μg/L	Kolpin *et al.,* 2002
146	N,N-diethyltoluamide	134-62-3	191.27	Insecticide	Max. in surface waters: 1.1 μg/L	Kolpin *et al.,* 2002
147	α-HCH	319-84-6	290.831	Pesticide	Max. in surface waters: 0.018 μg/L	Xue *et al.,* 2006
148	β-HCH	319-85-7	290.831	Pesticide	Max. in surface waters: 0.061 μg/L	Xue *et al.,* 2006
149	γ-HCH#	58-89-9	290.831	Pesticide	Max. in surface waters: 0.12 μg/L	Xue *et al.,* 2006
150	δ-HCH	319-86-8	290.831	Pesticide	Max. in surface waters: 0.0046 μg/L	Xue *et al.,* 2006

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Ref.
151	Dicofol	115-32-2	370.49	Pesticide	Max. in surface waters: 0.0026 μg/L	Xue *et al.*, 2006
152	Nitrofen	1836-75-5	284.098	Pesticide	Max. in surface waters: 0.0023 μg/L	Xue *et al.*, 2006
153	*p,p*'-DDD	72-54-8	320.045	Pesticide	Max. in surface waters: 0.0021 μg/L	Xue *et al.*, 2006
154	*o,p*'-DDT	789-02-6	354.49	Pesticide	Max. in surface waters: 0.161 μg/L	Xue *et al.*, 2006
155	*p,p*'-DDT	50-29-3	354.49	Pesticide	Max. in surface waters: 0.030 μg/L	Xue *et al.*, 2006
156	*p,p*'-DDE	72-55-9	318.03	Breakdown product of DDT	Max. in surface waters: 0.06 μg/L	Brossa *et al.*, 2005
157	Anthracene	120-12-7	178.23	PAH	Max. in surface waters: 0.11 μg/L	Kolpin *et al.*, 2002
158	Benzo[a]pyrene*[#]	50-32-8	252.31	PAH	Max. in surface waters: 0.24 μg/L	Kolpin *et al.*, 2002
159	Fluoranthene	206-44-0	202.26	PAH	Max. in surface waters: 1.2 μg/L	Kolpin *et al.*, 2002
160	Naphthalene	91-20-3	128.17	PAH	Max. in surface waters: 0.08 μg/L	Kolpin *et al.*, 2002
161	Phenanthrene	85-01-8	178.23	PAH	Max. in surface waters: 0.53 μg/L	Kolpin *et al.*, 2002
162	Pyrene	129-00-0	202.26	PAH	Max. in surface waters: 0.84 μg/L	Kolpin *et al.*, 2002
163	Bisphenol A	80-05-7	228.29	Plasticizer	Max. in surface waters: 12 μg/L	Kolpin *et al.*, 2002
164	Butylbenzyl phthalate	85-68-7	312.365	Plasticizer	Max. in STP effluent: 0.29 μg/L	Vethaak *et al.*, 2005
165	Di (2-ethylhexyl) phthalate[#]	117-81-7	390.562	Plasticizer	Max. in STP effluent: 2.4 μg/L	Vethaak *et al.*, 2005
166	Di-n-butyl phthalate	84-74-2	278.347	Plasticizer	Max. in STP effluent: 0.84 μg/L	Vethaak *et al.*, 2005
167	Diethyl phthalate	84-66-2	222.24	Plasticizer	Max. in surface waters: 0.42 μg/L	Kolpin *et al.*, 2002
168	Dimethyl phthalate	131-11-3	194.187	Plasticizer	Max. in STP effluent: 0.32 μg/L	Vethaak *et al.*, 2005
169	Triphenyl phosphate	115-86-6	326.29	Plasticizer	Max. in surface waters: 0.22 μg/L	Kolpin *et al.*, 2002
170	Tris (2-butoxyethyl) phosphate	78-51-3	398.48	Plasticizer	Max. in surface waters: 6.7 μg/L	Kolpin *et al.*, 2002
	Steroids and Hormones (*n* = 12)					
171	Equilenin^	517-09-9	266.34	Estrogen replacement	Max. in surface waters: 0.278 μg/L	Kolpin *et al.*, 2002
172	Equilin^	474-86-2	268.35	Estrogen replacement	Max. in surface waters: 0.147 μg/L	Kolpin *et al.*, 2002
173	17α-Ethinylestradiol^	57-63-6	296.41	Ovulation inhibitor	Max. in surface waters: 0.831 μg/L	Kolpin *et al.*, 2002
174	19-Norethisterone^	68-22-4	298.42	Ovulation inhibitor	Max. in surface waters: 0.872 μg/L	Kolpin *et al.*, 2002

No.	Compound	CASRN [a]	MW	Use/origin	Environmental occurrence	Ref.
175	Mestranol^	72-33-3	310.44	Ovulation inhibitor	Max. in surface waters: 0.407 μg/L	Kolpin *et al.*, 2002
176	17α-Estradiol^	57-91-0	272.39	Reproductive hormone	Max. in surface waters: 0.074 μg/L	Kolpin *et al.*, 2002
177	17β-Estradiol^	50-28-2	272.39	Reproductive hormone	Max. in surface waters: 0.2 μg/L	Kolpin *et al.*, 2002
178	Estriol^	50-27-1	288.39	Reproductive hormone	Max. in surface waters: 0.051 μg/L	Kolpin *et al.*, 2002
179	Estrone^	53-16-7	270.37	Reproductive hormone	Max. in surface waters: 0.112 μg/L	Kolpin *et al.*, 2002
180	Progesterone	57-83-0	314.47	Reproductive hormone	Max. in surface waters: 0.199 μg/L	Kolpin *et al.*, 2002
181	Testosterone	58-22-0	288.43	Reproductive hormone	Max. in surface waters: 0.214 μg/L	Kolpin *et al.*, 2002
182	Androsterone	53-41-8	290.44	Urinary steroid	Max. in surface waters: 0.214 μg/L	Kolpin *et al.*, 2002

[a] Chemical Abstracts Service Registry Number. [b] STP represents sewage treatment plant. * Contaminants regulated by Guidelines for Canadian Drinking Water Quality (Health Canada). [#] Contaminants regulated by National Primary Drinking Water Regulations (US EPA). ^ Contaminants suggested by the Third Contaminant Candidate List (CCL3, US EPA).

Table A.2 Selected molecular descriptors and their calculation methods.

Abbreviations	Molecular Descriptors	Calculation Method
MW	Molecular weight (g/mol)	E-Dragon
AMW	Average molecular weight (g/mol)	E-Dragon
nDB	Number of double bonds	E-Dragon
nAB	Number of aromatic bonds	E-Dragon
nN [a]	Number of primary and secondary amines	E-Dragon
nArOH	Number of phenolic group (aromatic hydroxyls)	E-Dragon
Ui	Unsaturation index	E-Dragon
logK_{ow}	Log octanol-water partition coefficient	E-Dragon
logS	Log water solubility (mol/L)	E-Dragon
logD	Log distribution coefficient at pH = 7	Marvin
P	Polarizability at pH = 7	Marvin
PSA	Polar surface area at pH = 7 ($Å^2$)	Marvin
L	Molecular length (Å)	MMP+
W	Molecular width (Å)	MMP+
RLW	Ratio of molecular length and width	MMP+
HA	Hydrogen bond acceptor	MMP+
HD	Hydrogen bond donor	MMP+
Df [b]	Diffusivity (cm^2/s)	MMP+
HOMO	The energy of highest occupied molecular orbital (eV)	HyperChem
LUMO	The energy of lowest unoccupied molecular orbital (eV)	HyperChem
GAP	HOMO-LUMO energy gap (eV)	HyperChem
DM	Dipole moment (debye)	HyperChem

[a] nN is the sum of the numbers of primary and secondary amines. [b] Diffusivity was calculated from the critical volume (V_c) of the organic compound using formulas provided by Gnielinski (1978) and Reid *et al.*, (1977); The critical volume was calculated with the MMP+ software. Quantum-mechanical parameters (HOMO, LUMO, and GAP) and dipole moments were estimated for neutral species. All the other calculations are based on the major species present at a drinking water pH range of 5.5~8.5. For HyperChem calculations, the AM1 method was used for geometry optimization.

Table A.3 Molecular descriptor values for all compounds in the pool (n = 182).

Comp.	logK_{ow}	logS	logD	MW	AMW	nDB	nAB	nArOH	nN	Ui	HOMO
1	-0.9	-3.3	-3.8	365.5	8.31	3	6	1	1	3.32	-9.329
2	2.63	-3.9	-1.2	434.9	9.45	3	11	0	0	3.91	-9.373
3	3.43	-4.2	-0.6	469.4	10.2	3	11	0	0	3.91	-9.124
4	1.42	-2.8	-2.7	379.5	8.43	3	6	0	0	3.32	-9.242
5	1.57	-3	-2.3	333.4	8.34	3	6	0	0	3.32	-9.295
6	1.55	-3.3	-1	594.7	8.5	7	6	0	1	3.81	-9.111
7	1.76	-2.1	0.38	425.1	7.08	1	0	0	0	1	-8.907
8	0.5	-1.1	-1.3	406.6	6.67	1	0	0	0	1	-8.594
9	0.2	-3.9	-1.4	331.4	7.89	3	6	0	1	3.32	-8.818
10	0.21	-3.4	-1.7	320.4	8.01	3	6	0	1	3.32	-8.987
11	2	-2.7	0.88	358.4	7.63	3	6	0	0	3.32	-8.799
12	1.12	-2.4	0.07	360.4	8.01	3	6	0	0	3.32	-8.827
13	0.99	-4	-1.1	351.4	7.99	3	6	0	1	3.32	-8.943
14	0.57	-3.8	-1.5	319.4	7.79	3	6	0	1	3.32	-8.808
15	1.12	-2.4	0.07	360.4	8.01	3	6	0	0	3.32	-8.817
16	-0.4	-1.5	-1.2	213.3	9.27	3	6	0	1	3.32	-9.172
17	0.97	-3.3	0.41	284.8	10.55	2	12	0	1	3.91	-9.403
18	0.25	-2.6	0.13	250.3	9.27	2	12	0	1	3.91	-9.18
19	1.08	-3.1	0.97	310.4	8.87	2	12	0	1	3.91	-9.074
20	0.44	-2.9	0.26	264.3	8.81	2	12	0	1	3.91	-9.168
21	0.43	-3.1	0.39	278.4	8.44	2	12	0	1	3.91	-9.141
22	0.8	-2.1	0.14	252.3	9.34	2	11	0	1	3.81	-9.134
23	0.13	-2.3	-0.6	269.4	10.36	2	11	0	1	3.81	-9.151
24	1.04	-3	0.25	267.3	8.62	2	11	0	1	3.81	-8.793
25	0.84	-3	1	249.3	8.9	2	12	0	1	3.91	-9.098
26	0.63	-2.1	0.18	254.3	10.6	2	11	0	1	3.81	-9.084
27	0.82	-2.5	-0.1	266.3	8.88	2	11	0	1	3.81	-9.096
28	-0	-4.2	-6	477.9	8.69	5	6	1	0	3.59	-9.186
29	0.01	-4	-6.3	443.5	8.06	5	6	1	0	3.59	-9.273
30	-0.2	-3.7	-7.6	459.5	8.2	5	6	1	0	3.59	-9.492
31	-0.3	-3.8	-6.4	443.5	8.06	5	6	1	0	3.59	-9.191
32	0.04	-3.2	-0.7	262.3	9.04	4	11	0	0	4	-8.485
33	1.15	-2.9	0.88	323.2	10.1	3	6	0	0	3.32	-10.39
34	0.31	-2.2	-3.9	334.4	6.82	1	0	0	2	1	-9.425
35	5.53	-4.7	4.9	289.5	12.06	0	12	1	0	3.7	-8.768
36	1.26	-2.7	0.92	290.4	7.45	0	12	0	2	3.7	-8.736
37	0.51	-1.6	0.91	151.2	7.56	1	6	1	0	3	-8.462
38	1.23	-1.7	-2	179.2	8.96	2	6	0	0	3.17	-8.136

Comp.	logK_{ow}	logS	logD	MW	AMW	nDB	nAB	nArOH	nN	Ui	HOMO
39	0.23	-4.8	-0.8	300.4	6.83	1	6	0	0	3	-8.471
40	0.19	-5.3	0.35	300.4	6.83	1	6	0	0	3	-8.555
41	1.18	-0.6	1.22	188.3	7.24	2	6	0	0	3.17	-9.013
42	0.44	-3.2	-0.3	252.4	7.65	1	5	0	2	3	-8.817
43	-1.6	-4.4	-0.1	315.5	7.17	3	5	0	2	3.17	-9.008
44	0.42	-3.7	1.06	386.6	6.44	2	6	0	0	3.17	-8.795
45	2.63	-4.4	3.08	284.8	8.63	2	12	0	0	3.91	-9.246
46	3.04	-4.1	1.99	307.3	8.09	3	12	0	0	4	-9.284
47	2.1	-3.2	2.77	236.3	7.88	2	12	0	0	3.91	-8.611
48	0.62	-2.3	1.12	218.3	7.28	2	6	0	0	3.17	-9.658
49	2.44	-6	1.5	310.4	7.57	0	12	0	1	3.7	-9.508
50	-0.9	-0.3	-5.7	131.2	5.96	2	0	0	2	1.59	-9.014
51	-1	-6.5	1.79	256.4	6.25	0	12	0	0	3.7	-9.203
52	-0.7	-6.2	1.53	415.6	7.42	2	12	0	0	3.91	-7.938
53	1.14	-4.3	-4.1	347.4	7.24	3	6	0	1	3.32	-9.217
54	0.94	-1	1.15	231.3	6.8	2	6	0	0	3.17	-8.355
55	3.6	-3.8	0.73	241.3	7.78	1	12	0	0	3.81	-8.994
56	3.79	-4.4	1.37	243.3	8.11	1	12	0	0	3.81	-9.208
57	3.76	-4	1.71	205.3	6.42	1	6	0	0	3	-9.657
58	4	-5	0.49	356.8	8.92	2	16	0	0	4.25	-8.694
59	3.04	-4.3	0.64	253.3	7.92	2	12	0	0	3.91	-9.777
60	3.43	-4	0.25	229.3	7.64	1	11	0	0	3.7	-8.651
61	2.7	-2.8	3.09	203.3	6.35	2	6	0	0	3.17	-9.077
62	0.02	-4.6	-0.2	278.2	7.73	0	6	0	2	2.81	-8.759
63	-0.7	-4.4	-0.1	304.4	6.92	0	12	3	1	3.7	-8.844
64	-2.2	-2.9	-1.7	240.4	6.16	0	6	1	1	2.81	-9.27
65	-1.4	-2.7	-1.2	226.3	6.29	0	6	2	1	2.81	-9.164
66	-1.5	-3.7	-2.1	267.4	6.37	1	6	0	1	3	-9.189
67	1.66	-5.4	-0	308.5	5.93	0	6	0	1	2.81	-8.916
68	1.58	-4.8	-0.4	326.5	5.94	0	6	0	1	2.81	-8.871
69	1.03	-5.6	0.15	299.4	6.65	0	15	0	1	4	-8.489
70	1.11	-4.3	-1.1	380.6	6.24	2	6	0	1	3.17	-8.74
71	0.44	-4.3	-0.8	268.4	5.96	0	6	0	1	2.81	-8.917
72	-1	-3.3	-1.8	310.5	6.21	0	6	0	1	2.81	-9.415
73	0.31	-5.1	0.02	260.4	6.35	0	11	0	1	3.59	-8.522
74	-1.2	-3.7	-2.5	273.4	7.01	2	6	0	1	3.17	-9.239
75	-0.2	-3.4	-1.3	317.5	6.9	0	5	0	1	2.59	-8.721
76	0.84	-2.4	1.16	277.3	10.67	2	15	0	0	4.17	-8.912
77	0.76	-1.2	0.1	261.1	9	1	0	0	0	1	-10.61

Comp.	logK_{ow}	logS	logD	MW	AMW	nDB	nAB	nArOH	nN	Ui	HOMO
78	-2.2	-0.7	-2.8	243.3	8.11	3	0	0	1	2	-9.379
79	0.15	-3.4	2.32	528.6	7.77	3	12	2	1	4	-9.084
80	-0.1	-3.2	1.5	544.6	7.89	3	12	2	1	4	-7.331
81	-0.1	-3.2	1.5	544.6	7.89	3	12	2	1	4	-8.993
82	0.12	-3.4	0.29	498.6	7.79	3	12	2	1	4	-9.178
83	0.57	-1.2	0.1	261.1	9	1	0	0	0	1	-10.42
84	0.77	-3.2	-3.4	452.5	8.54	3	17	0	2	4.39	-8.931
85	4.17	-5.5	0.97	360.8	8.2	2	12	0	0	3.91	-9.482
86	4.86	-5.7	5.28	360.9	7.84	2	12	0	0	3.91	-9.775
87	4.21	-4.4	1.85	249.4	6.39	1	6	0	0	3	-9.122
88	3.05	-2.7	-0.4	213.7	8.9	1	6	0	0	3	-9.461
89	4	-5.1	0.98	317.8	8.83	2	12	0	0	3.91	-9.638
90	3.79	-3.3	3.88	220.3	6.12	4	0	0	0	2.32	-10.42
91	4.9	-3.8	4.76	206.4	5.58	0	6	1	0	2.81	-8.831
92	5.25	-4.2	5.27	220.4	5.51	0	6	1	0	2.81	-8.657
93	3.25	-2.3	3.06	180.3	6.22	0	6	1	0	2.81	-8.527
94	-0.6	-1.3	-0.6	180.2	8.58	2	5	0	0	3	-9.125
95	3.46	-3	3.18	147	12.25	0	6	0	0	2.81	-9.523
96	6.09	-5.3	5.74	220.4	5.51	0	6	1	0	2.81	-8.858
97	3.47	-2.4	3.21	150.2	6.01	0	6	1	0	2.81	-8.898
98	5.29	-3.9	4.69	206.4	5.58	0	6	1	0	2.81	-8.885
99	1.65	-2	1.53	120.2	7.07	1	6	0	0	3	-9.915
100	5.53	-5.4	4.67	244.4	5.82	1	6	0	0	3	-9.202
101	5.52	-5.5	4.72	258.4	5.74	0	6	0	0	2.81	-9.013
102	3.28	-4.8	3.98	294.3	7.55	5	6	0	0	3.59	-10.25
103	5.7	-5.6	4.96	258.4	5.74	1	6	0	0	3	-9.213
104	0.39	-0.2	0.21	176.2	7.05	1	6	0	0	3	-9.678
105	-0.2	-1.3	-0.6	194.2	8.09	2	5	0	0	3	-8.964
106	1.01	-0.6	-1.5	137.1	9.14	1	6	1	0	3	-9.461
107	3.03	-3.7	3.43	182.2	7.59	1	12	0	0	3.81	-9.849
108	1.99	-1.9	-0.2	149.2	7.46	1	6	0	0	3	-9.523
109	5.62	-5.8	5.38	290.4	6.18	2	6	0	0	3.17	-8.949
110	3.35	-3.3	3.55	228.3	7.87	1	12	1	0	3.81	-9.274
111	-2.5	-3.1	-1.5	777.1	14.66	3	6	0	0	3.32	-8.863
112	-2.3	-3	-2	777.1	14.66	3	6	0	0	3.32	-8.101
113	1.6	-1.2	1.81	133.2	7.83	0	10	0	0	3.46	-9.22
114	1.95	-0.7	2.18	108.2	6.76	0	6	1	0	2.81	-8.88
115	1.39	-0.3	1.67	94.12	7.24	0	6	1	0	2.81	-9.115
116	1.36	-1.6	2.11	285.5	10.98	1	0	0	0	1	-11.52

Comp.	logK_{ow}	logS	logD	MW	AMW	nDB	nAB	nArOH	nN	Ui	HOMO
117	3.26	-3.7	4.28	430.9	12.31	1	0	0	0	1	-11.37
118	5.7	-6.2	5.6	284.8	23.73	0	6	0	0	2.81	-9.912
119	2.47	-3.2	2.24	201.3	9.58	0	15	0	0	4	-8.652
120	2.69	-2.7	-0.9	220	12.22	1	6	0	0	3	-9.352
121	3.17	-3	3.5	269.8	7.1	1	6	0	0	3	-9.34
122	2.7	-3.9	2.2	215.7	7.7	0	6	0	2	2.81	-9.422
123	2.86	-3.7	1.39	275.9	21.22	0	6	1	0	3	-9.778
124	2.68	-2.9	-0.8	220	12.22	1	6	0	0	3	-9.553
125	2.92	-3.2	2.53	233.1	9.71	1	6	0	0	3	-8.849
126	2.16	-2.4	-0.9	199.6	9.51	1	6	0	0	3	-9.17
127	3.37	-3.4	3.45	283.8	6.92	1	6	0	0	3	-8.536
128	4.75	-4.8	2.95	265.3	22.11	0	6	1	0	2.81	-9.574
129	2.8	-3.7	2.23	225.3	6.44	0	6	0	2	2.81	-9.386
130	2.48	-3.2	1.78	201.7	8.07	0	6	0	2	2.81	-9.355
131	4.41	-4.6	3.8	304.7	9.52	2	0	0	0	1.59	-9.241
132	5.09	-5.8	4.6	335.3	8.6	4	6	0	0	3.46	-9.984
133	3.02	-3.1	3.59	269.8	7.1	1	6	0	0	3	-7.019
134	5.9	-6.7	4.73	364.9	14.03	2	0	0	0	1.59	-9.606
135	4.32	-3.2	2.6	406.9	16.28	2	0	0	0	1.59	-10.08
136	2.45	-3.3	2.46	201.2	7.74	1	11	0	0	3.7	-8.538
137	3.69	-3.4	3.09	167.2	7.6	0	15	0	0	4	-8.341
138	5.15	-5.4	4.78	350.6	12.09	1	6	0	0	3	-9.88
139	6.02	-6.8	5.27	409.8	17.07	1	0	0	0	1	-10.06
140	6.13	-5.6	5.74	505.2	10.75	2	12	0	0	4	-9.125
141	4.45	-3.9	4.19	304.4	7.61	1	6	0	0	3	-10.02
142	4.98	-6.3	3.95	380.9	14.11	1	0	0	0	1	-9.937
143	5.83	-6.5	4.78	373.3	16.97	2	0	0	0	1.59	-10.02
144	5.12	-6.8	4.93	345.7	9.6	0	12	0	0	3.7	-8.973
145	2.97	-3.7	2.6	263.2	10.12	3	6	0	0	3.32	-10.42
146	2.1	-2.1	2.5	191.3	6.17	1	6	0	0	3	-9.285
147	3.94	-4.7	4.35	290.8	16.16	0	0	0	0	0	-11.03
148	3.94	-4.7	4.35	290.8	16.16	0	0	0	0	0	-11.73
149	3.94	-4.7	4.35	290.8	16.16	0	0	0	0	0	-11.04
150	3.94	-4.7	4.35	290.8	16.16	0	0	0	0	0	-11.57
151	5.59	-6.6	5.56	370.5	12.78	0	12	0	0	3.7	-9.742
152	4.88	-5.5	4.62	284.1	11.36	2	12	0	0	3.91	-9.467
153	6.15	-7.5	6.11	320	11.43	0	12	0	0	3.7	-9.543
154	6.6	-8.1	6.46	354.5	12.66	0	12	0	0	3.7	-9.58
155	6.29	-8	6.46	354.5	12.66	0	12	0	0	3.7	-9.587

Comp.	logK_{ow}	logS	logD	MW	AMW	nDB	nAB	nArOH	nN	Ui	HOMO
156	6.22	-6.6	6.11	318	12.23	1	12	0	0	3.81	-9.038
157	4.56	-5.6	3.95	178.2	7.43	0	16	0	0	4.09	-8.123
158	6.39	-8.3	5.27	252.3	7.89	0	24	0	0	4.64	-7.922
159	5.04	-6.3	4.28	202.3	7.78	0	19	0	0	4.32	-8.63
160	3.33	-3.3	2.96	128.2	7.12	0	11	0	0	3.59	-8.711
161	4.55	-5.7	3.95	178.2	7.43	0	16	0	0	4.09	-8.617
162	5.19	-6.9	4.28	202.3	7.78	0	19	0	0	4.32	-8.132
163	3.81	-3.4	4.04	228.3	6.92	0	12	2	0	3.7	-8.829
164	4.54	-5	5.03	312.4	7.26	2	12	0	0	3.91	-7.49
165	7.07	-6.6	8.03	390.6	5.92	2	6	0	0	3.17	-9.339
166	4.53	-4.7	4.63	278.4	6.63	2	6	0	0	3.17	-7.664
167	2.6	-2.8	2.69	222.3	7.41	2	6	0	0	3.17	-7.714
168	1.96	-2.3	1.98	194.2	8.09	2	6	0	0	3.17	-7.774
169	4.16	-4	5.09	326.3	8.59	1	18	0	0	4.32	-9.433
170	3.31	-3.3	3.94	398.5	6.13	1	0	0	0	1	-10.57
171	4.32	-4.7	4.3	266.4	7.01	1	11	1	0	3.7	-8.566
172	3.8	-4.3	3.9	268.4	6.71	2	6	1	0	3.17	-8.874
173	3.63	-4.6	3.81	296.4	6.44	0	6	1	0	3	-8.833
174	2.72	-4.7	3.13	298.5	6.22	2	0	0	0	2	-10.03
175	3.89	-4.9	3.96	310.5	6.34	0	6	0	0	3	-8.743
176	3.57	-4.1	3.75	272.4	6.19	0	6	1	0	2.81	-8.815
177	3.57	-4.1	3.75	272.4	6.19	0	6	1	0	2.81	-8.839
178	2.54	-3.4	2.67	288.4	6.41	0	6	1	0	2.81	-8.858
179	4.03	-4.8	4.31	270.4	6.44	1	6	1	0	3	-8.893
180	3.58	-4.8	4.15	314.5	5.93	3	0	0	0	2	-10.05
181	2.99	-3.9	3.37	288.5	5.89	2	0	0	0	1.59	-10.03
182	3.71	-4.7	3.77	290.5	5.7	1	0	0	0	1	-10.09

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
1	35.8	163	4.45	14.61	10.16	1.44	2.4	1.53	4.31×10^{-6}	9.24	-0.094
2	41.3	141	5.74	14.30	11.73	1.22	2.1	0.45	3.75×10^{-6}	9.09	-0.282
3	43.3	141	3.05	16.78	11.96	1.40	2.1	0.41	3.64×10^{-6}	7.98	-1.149
4	37.2	133	3.56	12.04	10.53	1.14	2.2	0.41	4.00×10^{-6}	9.07	-0.173
5	33	115	4.74	16.11	8.59	1.88	2	0.48	4.26×10^{-6}	9.23	-0.067
6	56.5	218	3.78	18.18	9.80	1.85	3	0.96	3.08×10^{-6}	7.7	-1.415
7	42.7	129	2.77	15.22	11.30	1.35	2.4	1.25	3.59×10^{-6}	9.2	0.2908
8	41.4	149	4.02	17.01	11.22	1.52	2.7	1.58	3.65×10^{-6}	9.08	0.4844
9	32.9	80.3	9.13	14.66	10.65	1.38	1.5	0.48	4.38×10^{-6}	8.15	-0.669
10	31.3	93.2	7.64	13.53	9.57	1.41	1.6	0.44	4.45×10^{-6}	8.14	-0.85
11	36.6	66.9	9.27	16.17	10.75	1.50	1.7	0.13	4.10×10^{-6}	8.14	-0.657

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
12	35.8	76.2	7.42	14.99	10.22	1.47	1.8	0.084	4.20×10^{-6}	8.08	-0.75
13	34.3	80.3	6.15	15.36	10.39	1.48	1.5	0.44	4.19×10^{-6}	8.09	-0.858
14	31.7	80.3	9.04	14.32	10.29	1.39	1.5	0.48	4.42×10^{-6}	8.14	-0.667
15	35.5	76.2	8.86	13.64	10.03	1.36	1.8	0.084	4.20×10^{-6}	8.09	-0.728
16	20.3	94.8	6.08	12.23	7.24	1.69	1.8	0.64	5.50×10^{-6}	8.68	-0.495
17	26.4	104	10.3	14.93	8.08	1.85	1.1	1.05	4.78×10^{-6}	8.72	-0.678
18	24.3	104	7.2	13.36	8.42	1.59	1.4	1.12	5.00×10^{-6}	8.68	-0.505
19	29.4	122	7.92	15.41	10.24	1.51	1.5	1.01	4.42×10^{-6}	8.65	-0.429
20	26.1	104	7.39	13.58	8.61	1.58	1.4	1.06	4.75×10^{-6}	8.69	-0.479
21	27.8	104	7.38	14.55	10.03	1.45	1.4	1.01	4.54×10^{-6}	8.7	-0.443
22	24.6	104	6.53	14.05	8.22	1.71	1.7	0.68	5.02×10^{-6}	8.66	-0.475
23	24.7	132	6.93	14.01	7.30	1.92	1.6	0.64	4.94×10^{-6}	8.41	-0.737
24	25.6	104	7.48	13.76	7.66	1.80	1.2	0.97	4.78×10^{-6}	8.12	-0.676
25	25.5	93.5	6.49	13.54	8.68	1.56	1.2	1.13	4.96×10^{-6}	8.71	-0.384
26	24.2	119	6.34	13.41	8.53	1.57	1.7	0.73	5.18×10^{-6}	8.49	-0.591
27	26.4	104	6.45	13.76	8.10	1.70	1.7	0.64	4.78×10^{-6}	8.45	-0.648
28	45.5	188	1.85	14.67	10.12	1.45	4.5	1.97	3.62×10^{-6}	8.24	-0.95
29	42	188	6.64	16.69	10.41	1.60	4.4	1.96	3.71×10^{-6}	8.45	-0.825
30	43.7	209	3.98	15.90	11.25	1.41	4.9	2.34	3.68×10^{-6}	8.65	-0.845
31	43.5	188	7.41	15.18	11.36	1.34	4.5	2.01	3.72×10^{-6}	8.24	-0.956
32	25.2	102	3.66	15.73	8.09	1.95	0.9	0.59	5.12×10^{-6}	6.96	-1.522
33	27.8	115	4.58	14.73	8.76	1.68	1.5	1.12	4.51×10^{-6}	9.11	-1.285
34	33.5	139	2.53	13.81	10.29	1.34	2.1	1.76	4.32×10^{-6}	9.6	0.1777
35	27.2	29.5	1.97	14.21	8.33	1.71	0.5	0.63	5.22×10^{-6}	8.23	-0.54
36	30	107	2.17	13.98	10.55	1.32	1.7	1.12	4.44×10^{-6}	8.73	-0.007
37	15.8	49.3	4.55	11.26	6.61	1.70	0.9	0.82	7.01×10^{-6}	8.75	0.2836
38	16.5	66.4	1.55	10.34	8.31	1.24	1.4	0.15	6.11×10^{-6}	7.18	-0.957
39	32.6	43.1	2.97	12.58	9.26	1.36	0.8	0.7	4.40×10^{-6}	8.87	0.4019
40	32.6	40	3.86	12.96	10.07	1.29	0.6	0.27	4.39×10^{-6}	8.82	0.2664
41	20.9	23.6	4.45	11.68	7.82	1.49	0.2	0.23	5.88×10^{-6}	8.88	-0.134
42	25.9	114	9.44	12.41	9.68	1.28	1.4	0.83	4.58×10^{-6}	8.89	0.0738
43	34.6	113	8.8	15.21	8.79	1.73	0.8	0.88	4.13×10^{-6}	8.71	-0.297
44	43.9	70.8	3.35	19.55	8.99	2.18	0.9	0.33	3.71×10^{-6}	8.93	0.1302
45	30.4	32.7	3.14	12.51	10.63	1.18	0.5	0.32	4.59×10^{-6}	8.64	-0.605
46	31.4	66.4	3.62	14.21	10.65	1.33	1.4	0.34	4.47×10^{-6}	8.26	-1.021
47	27	46.3	3.53	12.07	9.02	1.34	0.7	0.87	5.07×10^{-6}	8.15	-0.458
48	23.1	58.2	3.41	10.96	8.91	1.23	1	0.71	5.24×10^{-6}	9.82	0.1614
49	30.8	25.8	4.62	12.82	10.25	1.25	0.3	0.69	4.26×10^{-6}	9.35	-0.157
50	14	92.5	0.35	9.74	6.87	1.42	1.5	2.97	7.20×10^{-6}	10	1.003

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
51	30.4	13.7	1.99	13.43	9.44	1.42	0.2	0.56	4.46×10^{-6}	9.48	0.2809
52	43.8	85.6	7.29	18.24	12.39	1.47	0.7	0.49	3.62×10^{-6}	7.36	-0.576
53	35.1	117	5.11	16.12	10.21	1.58	2.6	0.68	3.96×10^{-6}	9.74	0.5183
54	25.7	26.8	3.83	12.59	8.61	1.46	0.4	0.19	5.12×10^{-6}	8.32	-0.036
55	25.7	49.4	2.18	14.23	7.26	1.96	1.2	0.34	4.87×10^{-6}	9	0.0051
56	24.8	40.1	3.21	13.68	7.22	1.89	1.1	0.3	4.87×10^{-6}	8.69	-0.514
57	23.3	40.1	4.97	12.96	7.50	1.73	1.1	0.15	5.00×10^{-6}	9.56	-0.097
58	36.1	71.4	2.51	16.74	10.15	1.65	1.5	0.27	4.04×10^{-6}	8.08	-0.616
59	26.4	57.2	2.84	11.66	9.27	1.26	1.4	0.34	4.69×10^{-6}	9.27	-0.508
60	24.4	49.4	2.51	13.80	7.55	1.83	1.3	0.23	4.99×10^{-6}	8.25	-0.402
61	24.1	20.3	3.34	12.77	9.14	1.40	0.3	0.22	5.03×10^{-6}	9.36	0.2815
62	28.7	62.9	2.77	12.74	9.50	1.34	1	1.3	4.60×10^{-6}	8.65	-0.114
63	32.1	97.5	2.92	16.00	9.26	1.73	1.8	2.19	4.79×10^{-6}	8.94	0.0913
64	26.7	77.3	2.78	13.54	8.48	1.60	1.5	1.62	4.94×10^{-6}	9.34	0.0687
65	25	77.3	1.48	11.52	8.51	1.35	1.4	1.64	5.50×10^{-6}	9.23	0.062
66	30.5	89.2	4.57	17.11	8.32	2.06	1.3	1.33	4.45×10^{-6}	9.27	0.0786
67	36.7	55.3	2.4	14.77	10.59	1.39	0.7	0.84	3.96×10^{-6}	9.33	0.4149
68	39.1	64.5	1.26	19.77	9.72	2.04	0.9	0.84	3.83×10^{-6}	9.33	0.4545
69	34.4	61.9	0.96	12.99	10.03	1.30	0.8	1.23	4.15×10^{-6}	8.22	-0.269
70	43.4	95.5	5.45	18.59	10.02	1.85	1.4	1.09	3.53×10^{-6}	8.66	-0.078
71	32.2	55.3	2.45	18.03	7.92	2.28	0.7	0.84	4.31×10^{-6}	9.33	0.413
72	35.4	86.5	3.34	17.07	9.88	1.73	1.4	1.47	4.09×10^{-6}	9.62	0.2034
73	30.5	46.1	1.22	14.32	9.25	1.55	0.6	0.95	4.44×10^{-6}	8.22	-0.306
74	29.8	91.4	5.08	14.56	9.18	1.59	1	1.21	4.44×10^{-6}	8.49	-0.751
75	33.6	113	2.73	12.69	9.81	1.29	1.3	0.69	4.30×10^{-6}	8.43	-0.292
76	25.5	143	8.53	9.99	9.67	1.03	0.9	0.46	4.96×10^{-6}	7.1	-1.808
77	23.7	51.4	0	11.35	9.53	1.19	2	0.23	5.70×10^{-6}	11.4	0.7485
78	21.5	129	5.17	12.38	8.72	1.42	2.3	1.57	5.57×10^{-6}	9.21	-0.169
79	54.8	187	5.76	16.54	14.19	1.17	3.1	2.13	3.40×10^{-6}	7.7	-1.384
80	55.7	208	6.7	16.65	13.55	1.23	3.6	2.5	3.37×10^{-6}	6.69	-0.637
81	55.8	208	2.59	17.71	14.08	1.26	3.6	2.5	3.37×10^{-6}	7.64	-1.358
82	51.2	178	4.87	16.11	13.43	1.20	2.9	2.16	3.53×10^{-6}	7.48	-1.702
83	23.7	51.4	5.38	11.19	10.46	1.07	2	0.23	5.70×10^{-6}	11.1	0.6645
84	44.8	216	4.41	22.84	9.17	2.49	4.5	1.46	3.46×10^{-6}	8.13	-0.798
85	36.7	78.5	3.89	15.60	10.93	1.43	1.8	0.56	3.88×10^{-6}	9.06	-0.42
86	38.3	52.6	4.78	16.72	9.86	1.70	0.7	0.31	3.84×10^{-6}	9.04	-0.737
87	28.7	49.4	0.7	14.66	8.33	1.76	1.3	0.11	4.47×10^{-6}	9.42	0.2973
88	20.3	49.4	2.83	12.05	7.35	1.64	1.3	0.15	5.52×10^{-6}	9.34	-0.126
89	31.8	66.4	2.67	16.38	8.27	1.98	1.5	0.31	4.28×10^{-6}	8.9	-0.739

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
90	25.2	34.1	2.18	11.44	8.82	1.30	0.5	0.083	4.78×10^{-6}	9.08	-1.341
91	25.2	20.2	1.39	11.88	7.99	1.49	0.4	0.5	5.05×10^{-6}	9.31	0.4837
92	27.3	20.2	1.43	11.57	9.40	1.23	0.4	0.47	4.80×10^{-6}	9.17	0.5152
93	20.8	29.5	0.4	11.76	7.64	1.54	0.6	0.51	5.85×10^{-6}	8.87	0.3455
94	16.1	67.2	3.91	9.76	8.15	1.20	0.9	0.34	6.60×10^{-6}	8.74	-0.385
95	14.2	0	1.4×10^{-5}	9.80	6.35	1.54	0	0.15	7.30×10^{-6}	9.31	-0.216
96	27.5	20.2	1.37	18.61	6.59	2.83	0.4	0.54	4.71×10^{-6}	9.31	0.4505
97	18.3	20.2	1.37	10.25	7.39	1.39	0.4	0.54	6.44×10^{-6}	9.37	0.4671
98	25.6	20.2	1.35	11.94	8.59	1.39	0.4	0.54	5.05×10^{-6}	9.36	0.4742
99	14.1	17.1	2.85	9.41	6.68	1.41	0.2	0.19	7.06×10^{-6}	9.56	-0.358
100	29.7	17.1	3.16	11.64	10.49	1.11	0.2	0.077	4.37×10^{-6}	9.12	-0.086
101	31.5	9.23	1.72	12.41	8.87	1.40	0.2	0.077	4.28×10^{-6}	9.42	0.4027
102	28.7	109	2.17	11.94	9.63	1.24	0.2	0	4.24×10^{-6}	8.9	-1.348
103	31.6	17.1	3.17	12.68	10.01	1.27	0.2	0.077	4.20×10^{-6}	9.06	-0.15
104	19.1	33.2	2.22	10.65	7.80	1.36	0.5	0.18	5.96×10^{-6}	9.47	-0.204
105	17.9	58.4	3.57	9.77	8.86	1.10	0.7	0.057	6.28×10^{-6}	8.62	-0.349
106	12.6	60.4	2.21	8.53	7.05	1.21	1.5	0.54	8.03×10^{-6}	8.87	-0.59
107	22	17.1	2.58	11.60	7.92	1.47	0.2	0.37	5.61×10^{-6}	9.22	-0.627
108	15.4	40.1	1.68	10.97	6.51	1.68	1.1	0.18	6.33×10^{-6}	9.84	0.3178
109	34.7	35.5	2.37	19.49	8.67	2.25	0.4	0.23	4.01×10^{-6}	8.33	-0.619
110	24	46.5	2.89	13.96	7.90	1.77	0.8	0.7	5.37×10^{-6}	8.79	-0.481
111	55.3	180	6.12	14.79	13.33	1.11	4.3	2.21	3.25×10^{-6}	7.6	-1.259
112	54.4	188	8.95	15.26	13.44	1.14	4.5	2.49	3.23×10^{-6}	6.53	-1.575
113	13.6	41.6	3.77	9.29	7.01	1.32	0.2	0.43	6.87×10^{-6}	8.9	-0.324
114	11.9	20.2	1.31	9.01	6.83	1.32	0.4	0.54	8.44×10^{-6}	9.31	0.4264
115	9.74	20.2	1.23	7.78	6.71	1.16	0.4	0.58	9.66×10^{-6}	9.51	0.3978
116	24.4	54.6	2.11	13.38	9.97	1.34	1.9	0	5.78×10^{-6}	11.4	-0.101
117	34.4	54.6	1.52	13.48	11.43	1.18	2.1	0	4.39×10^{-6}	11.2	-0.183
118	21.8	0	1.4×10^{-2}	9.84	9.00	1.09	0	0	5.58×10^{-6}	8.87	-1.041
119	21	69.8	4.77	11.94	7.50	1.59	0.6	0.53	5.79×10^{-6}	7.85	-0.803
120	18.9	49.4	3.23	11.18	7.39	1.51	1.3	0.12	5.86×10^{-6}	9.04	-0.312
121	28.8	29.5	3.75	11.95	11.51	1.04	0.5	0.11	4.46×10^{-6}	9.55	0.2142
122	22.3	62.7	3.67	13.60	8.40	1.62	1	0.51	5.12×10^{-6}	9.47	0.0452
123	18.1	46.9	2.31	9.71	8.86	1.10	1	0.081	6.50×10^{-6}	8.89	-0.888
124	18.6	49.4	2.17	9.97	8.73	1.14	1.2	0.077	5.86×10^{-6}	8.88	-0.671
125	22.2	32.3	4.77	13.16	7.98	1.65	0.6	0.4	5.38×10^{-6}	8.77	-0.074
126	18.8	49.4	0.29	12.55	7.78	1.61	1.3	0.12	5.81×10^{-6}	9.15	-0.021
127	30.4	29.5	2.55	12.51	11.43	1.09	0.5	0.11	4.30×10^{-6}	8.66	0.1265
128	20.1	23.1	1.24	9.83	9.03	1.09	0.9	0	6.17×10^{-6}	8.6	-0.977

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
129	25	72	1.76	12.55	9.43	1.33	1.1	0.51	4.78×10^{-6}	9.72	0.3339
130	20.6	62.7	3.01	12.45	9.56	1.30	1	0.51	5.38×10^{-6}	9.46	0.1073
131	28.4	45.6	2.83	14.21	9.39	1.51	0.3	0	4.50×10^{-6}	8.69	-0.548
132	28.8	94.9	3.86	13.17	10.76	1.22	0.3	0.084	4.21×10^{-6}	8.45	-1.53
133	28.7	29.5	2.28	11.50	10.25	1.12	0.5	0.11	4.46×10^{-6}	6.5	-0.515
134	30.8	0	2.92	10.63	8.90	1.19	0.2	0.052	4.49×10^{-6}	9.34	-0.269
135	31.9	54.7	3.75	12.71	8.49	1.50	0.2	0	4.37×10^{-6}	9.53	-0.547
136	21.1	38.3	2.56	11.46	9.15	1.25	0.6	0.52	5.42×10^{-6}	8.25	-0.29
137	18.8	15.8	1.2	10.78	7.12	1.51	0.3	0.57	5.95×10^{-6}	8.23	-0.108
138	30.5	82.5	6.71	13.69	9.80	1.40	1.3	0.041	4.89×10^{-6}	8.12	-1.757
139	32.6	0	1.12	11.03	9.36	1.18	0.2	0	4.38×10^{-6}	9.53	-0.529
140	44.3	59.3	3.61	17.97	9.93	1.81	0.6	0.39	3.53×10^{-6}	8.79	-0.332
141	31.4	95.4	4.53	13.92	9.59	1.45	1.4	0.043	4.63×10^{-6}	8.43	-1.593
142	31.6	12.5	2.31	10.94	8.93	1.23	0.3	0	4.47×10^{-6}	9.52	-0.415
143	29.6	0	1.79	11.41	9.84	1.16	0.2	0.054	4.60×10^{-6}	9.56	-0.461
144	34	18.5	2.9	14.81	9.52	1.56	0.3	0.31	4.21×10^{-6}	8.75	-0.222
145	22.9	115	3.44	13.34	8.30	1.61	1.1	0.16	5.83×10^{-6}	8.32	-2.099
146	22.1	20.3	3.13	12.11	8.73	1.39	0.3	0.15	5.21×10^{-6}	9.22	-0.068
147	23.3	0	1.71	10.04	9.26	1.08	0.3	0	5.37×10^{-6}	10.9	-0.154
148	23.6	0	0	9.38	9.23	1.02	0.3	0	5.37×10^{-6}	11.8	0.1114
149	23.1	0	2.4	9.81	8.81	1.11	0.3	0	5.37×10^{-6}	10.9	-0.15
150	23.4	0	1.8	9.86	9.06	1.09	0.3	0	5.37×10^{-6}	11.7	0.0928
151	33.1	20.2	2	12.44	11.97	1.04	0.5	0.69	4.32×10^{-6}	9.18	-0.561
152	26	55.1	4.61	14.63	9.26	1.58	0.1	0.28	4.90×10^{-6}	8.18	-1.286
153	30.3	0	0.59	13.42	10.63	1.26	0.1	0.3	4.51×10^{-6}	9.22	-0.325
154	32	0	2.03	12.76	10.43	1.22	0	0.3	4.36×10^{-6}	9.12	-0.461
155	32.4	0	1.08	14.13	9.93	1.42	0	0.3	4.36×10^{-6}	9.07	-0.517
156	30.8	0	0.18	13.61	9.22	1.48	0	0.31	4.53×10^{-6}	7.9	-1.137
157	20.6	0	0	11.53	8.10	1.42	0	0.38	5.60×10^{-6}	7.28	-0.84
158	29.1	0	0.04	13.43	8.97	1.50	0	0.46	4.59×10^{-6}	6.81	-1.111
159	23.2	0	0.24	10.87	9.12	1.19	0	0.38	5.23×10^{-6}	7.7	-0.929
160	14.6	0	1.2×10^{-4}	9.21	7.54	1.22	0	0.3	6.76×10^{-6}	8.45	-0.265
161	20.4	0	0.02	11.29	7.58	1.49	0	0.38	5.60×10^{-6}	8.21	-0.408
162	22.9	0	8.7×10^{-4}	11.21	8.80	1.27	0	0.38	5.23×10^{-6}	7.24	-0.889
163	25.3	40.5	2.11	12.16	8.88	1.37	0.9	1.09	5.45×10^{-6}	9.23	0.3972
164	33.8	52.6	3.8	13.90	11.02	1.26	0.6	0.34	4.06×10^{-6}	6.97	-0.517
165	46.1	52.6	5.31	15.76	13.78	1.14	0.6	0.15	3.29×10^{-6}	9.24	-0.102
166	31.4	52.6	4.22	12.77	9.57	1.33	0.6	0.15	4.23×10^{-6}	7.05	-0.616
167	23.2	52.6	4.11	11.87	8.97	1.32	0.6	0.15	5.07×10^{-6}	7.05	-0.662

Comp.	P	PSA	DM	L	W	RLW	HA	HD	Df	GAP	LUMO
168	19	52.6	2.84	10.40	8.67	1.20	0.6	0.15	5.70×10^{-6}	7.07	-0.703
169	32.6	54.6	2.57	14.65	10.62	1.38	1.6	0.56	4.73×10^{-6}	9.15	-0.288
170	46.4	82.3	2.63	23.48	14.84	1.58	2.2	0	3.66×10^{-6}	11	0.4338
171	29.9	37.3	3.05	13.92	8.04	1.73	0.6	0.59	4.71×10^{-6}	8.15	-0.421
172	30.5	37.3	2.94	13.48	8.57	1.57	0.7	0.54	4.66×10^{-6}	9.18	0.3089
173	34.3	40.5	1.31	14.14	9.13	1.55	1.1	1.02	4.32×10^{-6}	9.23	0.4017
174	34.6	37.3	2.64	15.94	9.23	1.73	0.9	0.55	4.12×10^{-6}	10	-0.025
175	36.6	29.5	1.5	16.40	9.51	1.72	0.8	0.62	4.01×10^{-6}	9.22	0.472
176	31.8	40.5	1.28	12.72	8.53	1.49	0.9	0.84	4.57×10^{-6}	9.23	0.4189
177	31.8	40.5	1.17	13.49	7.95	1.70	0.9	0.84	4.57×10^{-6}	9.24	0.3972
178	32.8	60.7	0.57	13.67	8.88	1.54	1.3	1.17	4.51×10^{-6}	9.24	0.3791
179	31.1	37.3	3.37	12.62	8.75	1.44	0.6	0.51	4.61×10^{-6}	9.24	0.3441
180	37.1	34.1	2.16	15.22	8.45	1.80	0.5	0.039	3.91×10^{-6}	10	-0.022
181	33.9	37.3	2.7	14.27	8.28	1.72	0.7	0.37	4.17×10^{-6}	10	-0.008
182	34.5	37.3	3.22	13.52	8.70	1.55	0.6	0.33	4.14×10^{-6}	11.1	0.9798

Table A.4 Tanimoto coefficients for water treatment set 1.

Compound	No.	1	10	31	32	48	59	69	76	85	100	108	111	115	122	132	135	149	162	165	167	173	182
Amoxicillin	1	1																					
Enoxacin	10	0.20	1																				
Tetracycline	31	0.18	0.22	1																			
Carbadox	32	0.16	0.17	0.13	1																		
Primidone	48	0.37	0.26	0.26	0.21	1																	
Ketoprofen	59	0.26	0.31	0.37	0.19	0.40	1																
Carazolol	69	0.21	0.29	0.25	0.21	0.27	0.37	1															
Azathioprine	76	0.16	0.17	0.08	0.15	0.13	0.09	0.14	1														
Bezafibrate	85	0.32	0.23	0.16	0.16	0.32	0.22	0.27	0.10	1													
Celestolide	100	0.23	0.24	0.28	0.19	0.42	0.48	0.25	0.09	0.23	1												
Hydrocinnamic acid	108	0.29	0.42	0.26	0.25	0.50	0.50	0.32	0.16	0.29	0.38	1											
Iomeprol	111	0.24	0.23	0.19	0.16	0.27	0.22	0.20	0.11	0.27	0.23	0.20	1										
Phenol	115	0.29	0.20	0.21	0.30	0.35	0.30	0.32	0.13	0.28	0.32	0.50	0.19	1									
Atrazine	122	0.15	0.32	0.12	0.10	0.20	0.10	0.16	0.22	0.11	0.10	0.14	0.10	0.11	1								
Trifluralin	132	0.17	0.31	0.17	0.27	0.22	0.20	0.25	0.24	0.23	0.21	0.26	0.23	0.25	0.12	1							
Alpha-Endosulfan	135	0.16	0.20	0.21	0.12	0.25	0.23	0.21	0.09	0.16	0.23	0.30	0.11	0.18	0.10	0.14	1						
Gamma-HCH	149	0.10	0.14	0.13	0.09	0.17	0.16	0.11	0.06	0.10	0.16	0.21	0.07	0.09	0.10	0.08	0.32	1					
Pyrene	162	0.24	0.30	0.26	0.21	0.33	0.52	0.41	0.09	0.24	0.42	0.42	0.21	0.35	0.11	0.22	0.21	0.17	1				
DEHP	165	0.18	0.16	0.17	0.15	0.22	0.24	0.28	0.07	0.18	0.24	0.22	0.16	0.21	0.08	0.16	0.18	0.12	0.22	1			
DEP	167	0.24	0.22	0.20	0.21	0.33	0.35	0.36	0.09	0.24	0.36	0.35	0.21	0.35	0.11	0.22	0.21	0.13	0.33	0.57	1		
EE2	173	0.22	0.20	0.45	0.16	0.32	0.29	0.21	0.07	0.22	0.34	0.33	0.14	0.28	0.08	0.17	0.22	0.26	0.28	0.21	0.21	1	
Androsterone	182	0.13	0.14	0.34	0.07	0.24	0.15	0.12	0.05	0.11	0.29	0.19	0.08	0.10	0.08	0.07	0.22	0.26	0.14	0.17	0.14	0.50	1

The calculations of Tanimoto Coefficient are based on the maximum common substructure (MCS) developed by Cao *et al.*, (2008). A free web tool ChemMine was used for the calculation (http://chemmine.ucr.edu/iframe/similarity).

Table A.5 Tanimoto coefficients for water treatment set 2.

Compound	No.	4	31	35	50	63	87	88	93	111	115	116	118	124	132	144	151	158	164	165	170	171	181
Methicillin	4	1																					
Tetracycline	31	0.19	1																				
Triclosan	35	0.31	0.16	1																			
Metformin	50	0.12	0.11	0.04	1																		
Fenoterol	63	0.26	0.21	0.21	0.1	1																	
Gemfibrozil	87	0.24	0.19	0.35	0.04	0.24	1																
Clofibric acid	88	0.23	0.18	0.48	0.05	0.23	0.39	1															
BHA	93	0.24	0.28	0.36	0.05	0.32	0.35	0.42	1														
Iomeprol	111	0.22	0.19	0.14	0.11	0.28	0.17	0.15	0.19	1													
Phenol	115	0.24	0.21	0.41	0.07	0.29	0.39	0.5	0.54	0.19	1												
TCEP	116	0.08	0.07	0.11	0.05	0.09	0.10	0.12	0.13	0.07	0.17	1											
HCB	118	0.17	0.15	0.38	0.05	0.20	0.25	0.37	0.32	0.16	0.46	0.13	1										
Dicamba	124	0.31	0.21	0.36	0.05	0.28	0.35	0.42	0.44	0.22	0.54	0.17	0.47	1									
Trifluralin	132	0.18	0.17	0.18	0.14	0.18	0.21	0.19	0.24	0.23	0.25	0.06	0.21	0.24	1								
Methoxychlor	144	0.19	0.32	0.27	0.03	0.29	0.26	0.30	0.36	0.16	0.33	0.09	0.22	0.31	0.19	1							
Dicofol	151	0.17	0.33	0.23	0.04	0.26	0.23	0.26	0.38	0.16	0.29	0.13	0.28	0.32	0.19	0.71	1						
Benzo(a)pyrene	158	0.17	0.29	0.19	0.04	0.22	0.27	0.21	0.32	0.19	0.29	0.06	0.23	0.27	0.19	0.19	0.38	1					
BBP	164	0.18	0.17	0.38	0.03	0.21	0.37	0.28	0.29	0.17	0.25	0.09	0.21	0.33	0.18	0.19	0.23	0.23	1				
DEHP	165	0.16	0.17	0.22	0.03	0.18	0.35	0.24	0.24	0.16	0.21	0.08	0.18	0.28	0.16	0.17	0.20	0.20	0.59	1			
TBEP	170	0.10	0.09	0.13	0.03	0.09	0.19	0.18	0.15	0.08	0.10	0.38	0.06	0.11	0.07	0.09	0.10	0.10	0.17	0.15	1		
Equilenin	171	0.17	0.39	0.23	0.01	0.29	0.27	0.26	0.38	0.19	0.35	0.10	0.23	0.27	0.19	0.32	0.33	0.48	0.23	0.20	0.10	1	
Testosterone	181	0.13	0.35	0.11	0.03	0.2	0.23	0.15	0.27	0.12	0.14	0.08	0.12	0.15	0.12	0.21	0.22	0.22	0.14	0.20	0.11	0.40	1

The calculations of Tanimoto Coefficient are based on the maximum common substructure (MCS) developed by Cao *et al.*, (2008). A free web tool ChemMine was used for the calculation (http://chemmine.ucr.edu/iframe/similarity).

Appendix B

Supplementary Material for Chapter 4

Table B.1 HPLC methods

Compound	Mobile phase	Flow	Quantitation	MDL (μM)
17α-Ethinylestradiol	70% A: 30% B	1 ml/min	220 nm	0.047
Benzo(a)pyrene	95% A: 5% B	1 ml/min	294 nm	0.007
Butylated hydroxyanisole	70% A: 30% B	1 ml/min	226 nm	0.023
Butylbenzyl phthalate	85% A: 15% B	1 ml/min	220 nm	0.16
Clofibric acid	70% A: 30% B	1 ml/min	224 nm	0.04
Di(2-ethylhexyl) phthalate	95% A: 5% B	1 ml/min	220 nm	0.12
Dicamba	55% A: 45% B	1 ml/min	220 nm	0.02
Dicofol	90% A: 10% B	1 ml/min	229 nm	0.026
Equilenin	70% A: 30% B	1 ml/min	229 nm	0.05
Fenoterol	20% A: 80% B	1 ml/min	220 nm	0.03
Gemfibrozil	90% A: 10% B	1 ml/min	220 nm	0.07
Hexachlorobenzene	95% A: 5% B	1 ml/min	216 nm	0.053
Methicillin	55% A: 45% B	1 ml/min	210 nm	0.057
Methoxychlor	90% A: 10% B	1 ml/min	226 nm	0.063
Phenol	55% A: 45% B	1 ml/min	270 nm	0.12
Pyrene	95% A: 5% B	1 ml/min	239 nm	0.008
Sulfamethoxazole	35% A: 65% B	1 ml/min	268 nm	0.04
Tetracycline	30% A: 70% B	1 ml/min	270 nm	0.067
Triclosan	85% A: 15% B	1 ml/min	220 nm	0.06
Trifluralin	85% A: 15% B	1 ml/min	273 nm	0.10
*p*CBA	70% A: 30% B	1 ml/min	238 nm	0.037

A: Methanol, B: 10 mM H_3PO_4 buffer (pH = 2). MDL: method detection limit.

Appendix C
Supplementary Material for Chapter 5

Modeling Ozone Reaction Rate Constants of Micropollutants Using Quantitative Structure–Property Relationships

Statistical Methods

1. Multiple Linear Regression (MLR)

MLR method is among the most widely used modeling methods in QSPR studies, which models a dependent variable (property to be predicted) as a linear combination of independent variables (molecular descriptors) with the regression coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \quad \textbf{(C.1)}$$

Where β_1, β_2, … β_p are regression coefficients and β_0 is the constant, x_{i1}, x_{i2}, … x_{ip} are molecular descriptors of the i'th compound, y_i is the property to be predicted, and ε_i represents the residuals. Least squares method which finds the smallest possible residual sum of square (sum of squared differences between the true y values and expected y values calculated by the model) may be used to calculate the regression coefficients and constant. When the property and molecular descriptors are standardized to have means of zero and standard deviation of one, the equation can be written in the matrix form:

$$Y = XB + E \quad \textbf{(C.2)}$$

Where Y is the matrix of property, X is the matrix of molecular descriptors, B is the matrix of regression coefficients, and E represents the matrix of residuals.

A few parameters such as the squared multiple correlation coefficient (R^2), adjusted R^2, and variance ratio (F) are used to judge the statistical qualities of the equations. R^2 is defined as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad \textbf{(C.3)}$$

Where y_i and $\hat{y}_i$ are observed and calculated property values, respectively, while $\bar{y}$ is the mean of the observed property. R^2 is a measure of how well a regression model fits a data set. R^2 ranges from 0 to 1, the closer the value of R^2 to 1, the better the regression model describe the observed data. The value of R^2 depends on the number of compounds (n) and number of descriptors (p), therefore another statistical parameter can be used, called adjusted R^2 (R^2_{adj}).

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1} \qquad \textbf{(C.4)}$$

Adjusted R^2 shows similar information as R^2 but adjusted by the number of compounds and number of descriptors. Because of the inflation of R^2 with the number of independent variables, adjusted R^2 is a more appropriate and meaningful parameter to compare models with different numbers of independent variables.

The dispersion of the observed dependent variable about the regression line (surface) can be assessed by the value of the standard error of estimate s. Larger value of s means worse statistical fit of the model and less reliability of the prediction.

$$s = \sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-p-1}} \qquad \textbf{(C.5)}$$

The statistical significance of a regression equation can be assessed by means of the Fisher (F) statistic. A regression model is considered to be statistically significant if the F value is greater than a tabulated value for the chosen level of significance (typically 95% level) and the corresponding degrees of freedom.

$$F = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2 / p}{\sum\left(y_i - \hat{y}_i\right)^2 / \left(n-p-1\right)} \qquad \textbf{(C.6)}$$

MLR assumes the linear relationships between molecular descriptors and the property to be predicted, and the molecular descriptors are mathematically independent. In practice, if the data set shows multicollinearity among the predictor variables, the model will be ill-conditioned and the calculated regression coefficients will be unstable and uninterpretable, for example, coefficients with the wrong sign may be found or the coefficients are much larger than expected (Eriksson *et al.,* 2003).

MLR is satisfactory applied in QSPR studies if the main problem of the multicollinearity among variables is solved.

2. Principal Component Regression (PCR)

Another regression-based method is PCR. The first step of conducting PCR is principal component analysis (PCA). PCA is used to summarize the information residing in the initial multivariate data (compounds and their descriptors). PCA results in several new variables, known as latent variables or principal components (PCs), which capture the major pattern of the initial dataset. By using PCA, the information of data can be represented by a few PCs and can be displayed graphically (Jackson 1991). Eriksson *et al.* (2006) identified PCA as the most suitable technique for dimensionality reduction and generation of orthogonal latent variables, and it has been applied successfully in a number of studies (e.g. Knekta *et al.*, 2004; Papa *et al.*, 2007; Harju *et al.*, 2002; Kitti *et al.*, 2003). A reduced set of variable is much easier to analyze and interpret. As a common procedure to avoid the influence of the unit of variables, data for PCA are usually pre-processed by means of mean-centering and scaling to unit variance. PCA then decomposes the *X*-matrix into the product of two matrices, the score matrix *T* and the loading matrix *P*, plus a residuals matrix *E*. The product of score matrix and loading matrix *TP'* is used to model the initial data matrix *X* (Wold *et al.,* 1987).

$$X = TP' + E \qquad \textbf{(C.7)}$$

The number of principal components (i.e., the number of columns in score matrix and the number of rows in loading matrix), is determined by cross-validation. In the next step, the first few significant principal components are used as predictor variables in a MLR. Because the principal components are mathematically independent, multicollinearity among original variables is no longer a problem as it is in MLR.

$$y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + ... + \beta_q PC_q + \varepsilon \qquad \textbf{(C.8)}$$

Where *y* is the property to be predicted, β_1, β_2, … β_q are regression coefficients and β_0 is the constant, PC_1, PC_2, … PC_q are principal components extracted by PCA, *q* is the number of significant principal component, and ε represents the residuals. The regression model can be found by using the

usual MLR algorithm, and same statistical parameters as used in MLR can be applied to assess the quality of models.

3. Partial Least Squares Regression (PLS)

PLS is a recently developed regression methods, which can be seen as a generalization of multiple linear regression (Eriksson *et al.,* 2006). PLS is a projection method which finds new variables (latent variables) which are linear combinations of the original variables and orthogonal, and also well correlated to the dependent variable(s). The dependent variable(s) can be a single property (e.g. rate constant) or multiple responses (e.g. toxicity determined in several testing systems). Similar with PCA method, but PCA works with *X* matrix while PLS works with two matrices, *X* and *Y*, respectively. In other words, the difference between PCR and PLS is that PLS finds the latent variables and the regression coefficients at the same time. The PLS projects the matrix *X* into a lower-dimensional hyper-plane which is a good summary of *X*, and several latent variables are introduced to describe the positions of projected data. Latent variables are used to correlate the values of *Y*. If response matrix *Y* contains multiple responses, *Y* will be summarized by another projected lower-dimensional hyper-plane. The number of significant dimensions in PLS is determined with cross-validation. Models will be set up between latent variables of *X* and *Y*.

$$Y = TB + E \qquad \textbf{(C.9)}$$

Where *T* is the matrix containing the scores of the extracted latent variables, (*T* = *XF*, where *F* is the matrix of the loadings of the original variables in the principal component scores, and *X* is the matrix of mean-centered molecular descriptors) *B* is the matrix of the PLS regression coefficients, and *E* is the unexplained variance in *Y*. The PLS regression model can be presented in terms of the original molecular descriptors.

$$Y = XQ + E \qquad \textbf{(C.10)}$$

Where *Q* = *FB* is the matrix of regression coefficient for original descriptors. In addition, the original mean-centered descriptors can also be expressed by the latent variables.

$$X = TP' + K \qquad \textbf{(C.11)}$$

Where *P* is the loading matrix, and *K* is the unexplained part of *X*. A detailed tutorial of PLS method with some good examples can be found in literature (Wold *et al.*, 2001). Unlike MLR, PLS

can work well when data are strongly correlated since the extracted latent variables are orthogonal and limited in number.

A commonly used procedure to build PLS models is cross-validation. Cross-validation is a statistical tool for assessing the predictive ability of a model. It is obtained by removing one (leave-one-out) or many (leave-many-out) chemicals from the dataset, developing the models on the remaining chemicals and using that to predict the activity of the chemicals removed. All the chemicals will be removed in turn once and only once (Eriksson *et al.,* 2003). Predicted residual error sum of squares (*PRESS*) can be calculated for that model.

$$PRESS = \sum (y_i - \hat{y}_{i/i})^2 \quad \textbf{(C.12)}$$

Note that this equation looks similar to the residual sum of squares given in MLR but different. Here the $\hat{y}_{i/i}$ is the calculated dependent variable from a model developed without that data point. Cross-validation ensures the resulting model contains the optimum number of components, and the model is built based on the ability to predict the data rather than to fit the data.

4. Piecewise Linear Regression (PLR)

Piecewise linear regression is similar to the MLR method. Instead of fitting an overall equation, PLR separates the compound set into two or more groups and fits submodel in each group. The advantage of PLR over MLR is that it can approximate the nonlinear phenomena of the data into local linearity. In addition, PLR is a simple enough modeling method comparing to nonlinear techniques. From a practical and scientific point-of-view, it is always desirable to develop the simplest model which has a satisfactory predictive power. However, the challenging problem of applying PLR is to find a meaningful breakpoint which splits the dataset into two or more subsets. In complex cases where it is difficult to find the breakpoint, quasi-Newton algorithm can be used to search the breakpoint or multiple breakpoints (Molina *et al.,* 2008). In the case of two pieces model where both the intercept and slope switches, the PLR equation can be written in the following form.

$$\hat{y} = (b_{10} + b_{11}x_1 + b_{12}x_2 + ... + b_{1p}x_p)(y \le C) + (b_{20} + b_{21}x_1 + b_{22}x_2 + ... + b_{2p}x_p)(y > C) \quad \textbf{(C.13)}$$

Where b_{ij} is the coefficient of j'th molecular descriptor for the i'th submodel. The expressions ($y \le C$) and ($y > C$) are logical conditions that equal to 0 if the condition is false and to 1 if true.

5. Linear Discriminant Analysis (LDA)

LDA is a typical statistical method for developing classification models which classify the data into two of more pre-defined categories or groups, and can be used to predict the group membership of new observations (Worth and Cronin 2003). In QSPR study, for scientific or regulatory purpose, chemicals may be classified into categories according to the kinds of toxicity they exhibit, toxic mechanisms (i.e., mode of action) (Spycher *et al.*, 2004), reactivity to certain reactant, etc. LDA provides a convenient means of labeling compounds, for example, a compound could be labeled as toxic or non-toxic, negative or positive, reactive or non-reactive according to analytical systems applied. In situations such as several different mechanisms dominant the entire data set and/or the variation of data is too large, which makes it difficult to develop reasonable quantitative prediction models, modelers may have to apply classification methods to divide data into groups. In addition, because of the limitation of instrument or analytical technique, compounds with responses lower than detection limit are marked such as "not detected", "no effect", "not toxic", etc., it is difficult to include these compounds into the data set to build quantitative relationships.

In the simplest type of LDA, i.e., two-group case, a linear discriminant function (LDF) can be obtained to maximize the distance between the means of the calculated values of the function in the two groups.

$$LDF = b_1 x_1 + b_2 x_2 + ... + b_p x_p + b \qquad \textbf{(C.14)}$$

where b is a constant, and b_1, b_2, … b_p are the regression coefficient for p variables. Once the discriminant function is finalized, the classification function can be derived to determine the group membership of each compound by calculating the classification scores for each compound for each group, by applying the formula:

$$S_i = w_{i1} x_1 + w_{i2} x_2 + ... + w_{ip} x_p + c_i \qquad \textbf{(C.15)}$$

where the subscript i denotes the respective group, the subscripts 1, 2, ..., p denote the p variables, c_i is a constant for the i'th group, w_{ij} is the weight for the j'th variable in the computation of the classification score for the i'th group, x_j is the observed value for the respective case for the j'th variable, S_i is the resultant classification score.

To test the overall statistical significance of the discrimination the parameter Wilks'λ is used. Wilks'λ ranges from 0 to 1, with 0 meaning perfect discrimination and 1 meaning all groups means are the same.

$$Wilks'\lambda = \det(W)/\det(T) \qquad \textbf{(C.16)}$$

Where det(*W*) is the within-groups variance/covariance matrix, and det(*T*) is the total variance/covariance matrix. The significance of Wilks'λ can be tested using Fisher *F* test. In addition, the significance of each variable added to the model can be assessed by partial Wilks'λ.

To calculate the probability that a given compound belongs to a given group, the Mahalanobis distance is used. This is the distance of the compound from the group centroid in the multidimensional space defined by the molecular descriptors. The group centroid is the point in the multivariate space with coordinates equal to the means of all variables. A compound is classified to the group to which it has the smallest distance.

Table C. 1 Data set used in the QSPR modeling.

No.	Compound (ref.)	p*Ka*	*MW*	*AMW*	*nDB*	*nAB*	*nArOH*	*nN*	*Ui*	*HOMO*	*LUMO*	*GaP*	*P*	log*Df*	logk_{O3}
						Training Set									
1	Equilenin	9.8	266.36	7.01	1	11	1	0	3.7	-8.57	-0.42	8.15	29.93	-5.33	7.00
2	Butylated hydroxyanisole	10.6	180.27	6.22	0	6	1	0	2.807	-8.53	0.35	8.87	20.76	-5.23	6.52
3	Fenoterol	8.6	304.4	6.92	0	12	3	1	3.7	-8.84	0.09	8.94	32.13	-5.32	6.45
4	Tetracycline	9.7	443.47	8.06	5	6	1	0	3.585	-9.19	-0.96	8.24	43.46	-5.43	6.40
5	Triclosan	8.1	289.54	12.06	0	12	1	0	3.7	-8.77	-0.54	8.23	27.16	-5.28	6.40
6	Phenol	9.9	94.12	7.24	0	6	1	0	2.807	-9.12	0.40	9.51	9.74	-5.02	6.26
7	17α-Ethinylestradiol	10.4	296.44	6.44	0	6	1	0	3	-8.83	0.40	9.23	34.26	-5.36	6.26
8	Gemfibrozil	4.4	249.36	6.39	1	6	0	0	3	-9.12	0.30	9.42	28.67	-5.35	5.69
9	Methicillin	2.8	379.45	8.43	3	6	0	0	3.322	-9.24	-0.17	9.07	37.24	-5.40	4.59
10	Benzo[a]pyrene	NA	252.32	7.89	0	24	0	0	4.644	-7.92	-1.11	6.81	29.1	-5.34	3.88
11	Clofibric acid	3.4	213.65	8.9	1	6	0	0	3	-9.46	-0.13	9.34	20.34	-5.26	3.70
12	Trifluralin	NA	335.32	8.6	4	6	0	0	3.459	-9.98	-1.53	8.45	28.76	-5.38	3.28
13	Methoxychlor	NA	345.66	9.6	0	12	0	0	3.7	-8.97	-0.22	8.75	34.02	-5.38	2.40
14	Tri(2-butoxyethyl)phosphate	NA	398.54	6.13	1	0	0	0	1	-10.57	0.43	11.00	46.42	-5.44	0.66
15	Butylbenzyl phthalate	NA	312.39	7.26	2	12	0	0	3.907	-7.49	-0.52	6.97	33.82	-5.39	0.15
16	Metformin	10.3	131.22	5.96	2	0	0	2	1.585	-9.01	1.00	10.02	14	-5.14	0.08
17	Tris(chloroethyl)phosphate	NA	285.5	10.98	1	0	0	0	1	-11.52	-0.10	11.42	24.4	-5.24	-0.10
18	Iomeprol	NA	777.12	14.66	3	6	0	0	3.322	-8.86	-1.26	7.60	55.3	-5.49	-0.70
19	Dicamba	2.0	220.03	12.22	1	6	0	0	3	-9.55	-0.67	8.88	18.58	-5.23	-1.00
20	Dicofol	NA	370.48	12.78	0	12	0	0	3.7	-9.74	-0.56	9.18	33.09	-5.36	-1.00
21	Di(2-ethylhexyl)phthalate	NA	390.62	5.92	2	6	0	0	3.17	-9.34	-0.10	9.24	46.06	-5.48	-1.00
22	Hexachlorobenzene	NA	284.76	23.73	0	6	0	0	2.807	-9.91	-1.04	8.87	21.75	-5.25	-2.00
						Validation Set									
23	Bisphenol A [a]	9.8	228.31	6.92	0	12	2	0	3.7	-8.83	0.40	9.23	25.3	-5.26	6.43
24	Sulfamethoxazole	6.2	252.3	9.34	2	11	0	1	3.807	-9.13	-0.47	8.66	24.58	-5.30	6.30
25	Estrone [a]	10.3	270.4	6.44	1	6	1	0	3	-8.89	0.34	9.24	31.12	-5.34	6.26

26	17beta-Estradiol [a]	10.3	272.42	6.19	0	6	1	0	2.807	-8.84	0.40	9.24	31.76	-5.34	6.23
27	Estriol [a]	10.3	288.42	6.41	0	6	1	0	2.807	-8.86	0.38	9.24	32.83	-5.35	6.22
28	Amoxicillin [b]	7.4	365.45	8.31	3	6	1	1	3.322	-9.33	-0.09	9.24	35.84	-5.37	6.18
29	4-Nonylphenol [a]	10.3	220.39	5.51	0	6	1	0	2.807	-8.86	0.45	9.31	27.49	-5.33	6.15
30	Lincomycin [c]	8	406.61	6.67	1	0	0	0	1	-8.59	0.48	9.08	41.44	-5.44	5.83
31	Carbamazepine [d]	NA	236.29	7.88	2	12	0	0	3.907	-8.61	-0.46	8.15	26.95	-5.30	5.48
32	Trimethoprim [e]	7.2	290.36	7.45	0	12	0	2	3.7	-8.74	-0.01	8.73	30.04	-5.35	5.43
33	Naproxen [b]	4.2	229.27	7.64	1	11	0	0	3.7	-8.65	-0.40	8.25	24.41	-5.30	5.41
34	Enrofloxacin [c]	6.7	358.43	7.63	3	6	0	0	3.322	-8.80	-0.66	8.14	36.56	-5.39	5.18
35	Pyrene	NA	202.26	7.78	0	19	0	0	4.322	-8.13	-0.89	7.24	22.86	-5.28	4.56
36	Ciprofloxacin [c]	8.7	331.38	7.89	3	6	0	1	3.322	-8.82	-0.67	8.15	32.85	-5.36	4.28
37	Phenant1ene [f]	NA	178.24	7.43	0	16	0	0	4.087	-8.62	-0.41	8.21	20.43	-5.25	4.20
38	Penicillin G [c]	2.7	333.42	8.34	3	6	0	0	3.322	-9.30	-0.07	9.23	33.02	-5.37	3.68
39	Metoprolol [b]	NA	268.42	5.96	0	6	0	1	2.807	-8.92	0.41	9.33	32.19	-5.37	3.15
40	Bromoxynil [g]	3.9	275.9	21.22	0	6	1	0	3	-9.78	-0.89	8.89	18.1	-5.19	3.06
41	Bezafibrate [d]	3.8	360.84	8.2	2	12	0	0	3.907	-9.48	-0.42	9.06	36.74	-5.41	2.77
42	Ibuprofen [d]	4.9	205.3	6.42	1	6	0	0	3	-9.66	-0.10	9.56	23.27	-5.30	0.98
43	Alachlor [e]	NA	269.8	7.1	1	6	0	0	3	-7.02	-0.52	6.50	28.74	-5.35	0.58
44	Diazepam [d]	3.4	284.76	8.63	2	12	0	0	3.907	-9.25	-0.60	8.64	30.38	-5.34	-0.12
45	Dimethyl phthalate [e]	NA	194.2	8.09	2	6	0	0	3.17	-7.77	-0.70	7.07	19.01	-5.24	-0.70
46	Diethylphthalate [e]	NA	222.26	7.41	2	6	0	0	3.17	-7.71	-0.66	7.05	23.19	-5.29	-0.85
47	Gamma-HCH [e]	NA	290.82	16.16	0	0	0	0	0	-11.04	-0.15	10.89	23.09	-5.27	-1.40
48	2,4-D [e]	2.6	221.04	11.63	1	6	0	0	3	-9.35	-0.31	9.04	18.9	-5.23	0.36
49	Atrazine [e]	1.6	215.69	7.7	0	6	0	2	2.807	-9.42	0.05	9.47	22.3	-5.29	0.78
50	Simazine [e]	2.0	201.66	8.07	0	6	0	2	2.807	-9.36	0.11	9.46	20.6	-5.27	0.68
51	Acetochlor [h]	NA	269.8	7.1	1	6	0	0	3	-7.03	-1.58	5.45	29.12	-5.37	0.38
52	Cyanazine [i]	12.9	240.73	8.3	0	6	0	2	3	-9.03	0.037	9.06	23.96	-5.34	0.87

53	Iopromide [d]	NA	791.15	14.13	3	6	0	0	3.322	-7.72	-1.81	5.91	55.03	-5.52	-0.10
54	Metolachlor [h]	NA	283.83	6.92	1	6	0	0	3	-8.19	-1.43	6.76	30.95	-5.39	0.04
55	Propachlor [h]	NA	211.71	7.56	1	6	0	0	3	-9.42	-0.78	8.65	22.97	-5.30	-0.05

MW represents molecular weight, *AMW* is average molecular weight, *nDB* is number of conjugated double bonds, *nAB* is number of isolated double bonds, *nArOH* is number of phenolic group, *nN* is number of primary and secondary amines, *Ui* is unsaturation index, *HOMO* is highest occupied molecular orbital, *LUMO* is lowest unoccupied molecular orbital, *GaP* represents *HOMO-LUMO* gap, and *P* is polarizability, log*Df* is diffusivity in logarithm form.

Literature cited: [a] Deborde *et al.,* 2005; [b] Benitez *et al.,* 2009; [c] Dodd *et al.,* 2006; [d] Huber *et al.,* 2003; [e] Yao and Haag 1991; [f] Butkovic *et al.,* 1983; [g] Cheme-Ayala *et al.,* 2010; [h] Acero *et al.,* 2003; [i] Benitez *et al.,* 1994. Pyrene and Sulfamethoxazole are measured data. For some compounds, the apparent rate constant at pH = 7 were not reported, but were calculated using the elementary absolute rate constant of each species (k_1, k_2) as follows:

$$k_{app} = k_1 \frac{10^{-pH}}{10^{-pKa} + 10^{-pH}} + k_2 \frac{10^{-pKa}}{10^{-pKa} + 10^{-pH}} \quad \textbf{(C.17)}$$

Table C. 2 Intercorrelation matrix.

	log*MW*	log*AMW*	*nDB*	*nAB*	*nArOH*	*Ui*	*HOMO*	*LUMO*	*GaP*	*P*	log*Df*
log*MW*	1.00										
log*AMW*	0.29	1.00									
nDB	0.25	-0.19	1.00								
nAB	0.13	0.18	-0.33	1.00							
nArOH	-0.25	-0.28	-0.42	0.27	1.00						
Ui	0.21	0.10	-0.33	0.86*	0.19	1.00					
HOMO	-0.11	-0.29	-0.25	0.56	0.31	0.71	1.00				
LUMO	-0.54	-0.62	-0.05	-0.46	0.21	-0.55	-0.04	1.00			
GaP	-0.24	-0.14	0.16	-0.71	-0.12	-0.88*	-0.81	0.63	1.00		
P	0.92*	-0.02	0.27	0.08	-0.14	0.17	0.02	-0.31	-0.20	1.00	
log*Df*	-0.92*	0.03	-0.31	-0.18	0.26	-0.29	-0.07	0.40	0.29	-0.94*	1.00

*Correlation coefficients higher than 0.85. It shows that log*MW* is highly correlated with *P* and log*Df*, and n*AB* is correlated with *Ui*, *Ui* is correlated with *GaP*.

Table C. 3 Classification results.

No.	Compound	Group	*X*1*A*	*ATS*3*m*	*GATS*2*v*	*MATS*6*p*	*GATS*3*v*	*Class*
			Training Set					
1	Equilenin	1	0.42	3.74	0.59	-0.17	0.81	1.61
2	Butylated hydroxyanisole	1	0.46	3.03	0.75	-0.11	0.99	1.51
3	Fenoterol	1	0.45	3.51	0.91	0.27	0.95	2.70
4	Tetracycline	1	0.42	4.49	0.8	-0.06	1.01	1.02
5	Triclosan	1	0.45	3.75	1.09	-0.2	1.32	1.41
6	Phenol	1	0.49	1.9	0.88	0	1.4	2.27
7	17α-Ethinylestradiol	1	0.42	3.9	0.72	-0.12	0.88	1.89
8	Gemfibrozil	1	0.46	3.34	0.68	1	1.42	2.18
9	Methicillin	1	0.44	4.07	0.71	0.06	0.89	0.87
10	Benzo[a]pyrene	1	0.41	3.71	0	1	0	3.39
11	Clofibric acid	1	0.46	3.28	0.97	-0.25	0.92	1.87
12	Trifluralin	1	0.46	3.83	0.68	0.67	1.05	1.20
13	Methoxychlor	1	0.45	3.83	0.71	-0.01	0.69	1.04
14	Tri(2-butoxyethyl)phosphate	-1	0.51	3.67	0.8	-0.13	1.06	-2.17
15	Butylbenzyl phthalate	-1	0.47	3.59	0.67	-0.13	1.4	-1.21
16	Metformin	-1	0.51	2.34	0	0	1.8	-4.34
17	Tris(chloroethyl)phosphate	-1	0.51	3.48	0.46	0.03	1.68	-4.14
18	Iomeprol	-1	0.47	5.2	0.62	-0.17	0.85	-3.63
19	Dicamba	-1	0.47	3.58	0.58	0	1.04	-0.69
20	Dicofol	-1	0.44	4.18	0.94	-0.53	2	-1.52
21	Di(2-ethylhexyl)phthalate	-1	0.48	3.78	0.7	-0.18	1.45	-2.14
22	Hexachlorobenzene	-1	0.46	4.52	0	0	0	-2.80
			Validation Set					
23	Bisphenol A	1	0.44	3.34	0.7	-0.35	0.7	1.70
24	Sulfamethoxazole	1	0.44	3.52	0.82	0.07	1.15	1.99
25	Estrone	1	0.42	3.74	0.59	-0.17	0.81	1.61
26	17beta-Estradiol	1	0.42	3.74	0.59	-0.17	0.81	1.61
27	Estriol	1	0.42	3.84	0.61	-0.22	0.74	1.53
28	Amoxicillin	1	0.43	3.99	0.69	0.12	0.75	1.81
29	4-Nonylphenol	1	0.49	2.87	0.89	-0.04	1	1.09
30	Lincomycin	1	0.45	4.12	1.04	0.03	0.84	1.82
31	Carbamazepine *	1	0.44	3.58	0.37	-0.25	0.88	-0.39
32	Trimethoprim *	1	0.46	3.67	0.77	-0.38	0.92	-0.07
33	Naproxen	1	0.45	3.37	0.52	-0.29	0.75	0.35
34	Enrofloxacin	1	0.43	3.99	0.86	-0.28	0.86	1.55
35	Pyrene	1	0.42	3.43	0	1	0	3.43

36	Ciprofloxacin	1	0.43	3.91	0.85	-0.34	0.89	1.48
37	Phenanthrene	1	0.43	3.14	0	1	0	3.48
38	Penicillin G	1	0.43	3.83	0.61	0.14	0.71	1.87
39	Metoprolol	1	0.48	3.21	1.04	-0.4	0.72	1.36
40	Bromoxynil	1	0.47	4.01	0.62	1.87	1.48	1.85
41	Bezafibrate	1	0.45	3.81	0.84	-0.42	0.99	0.27
42	Ibuprofen	-1	0.47	3.06	0.4	-0.18	0.85	-0.56
43	Alachlor	-1	0.48	3.6	0.45	-0.07	1.11	-2.10
44	Diazepam *	-1	0.44	3.78	0.7	0	0.84	1.35
45	Dimethyl phthalate	-1	0.48	3.19	0.51	-0.53	1.14	-2.06
46	Diethylphthalate	-1	0.48	3.27	0.75	-0.11	1.14	-0.22
47	Gamma-HCH	-1	0.46	4.52	0	0	0	-2.80
48	2,4-D	-1	0.47	3.37	0.77	-0.76	0.52	-0.06
49	Atrazine	-1	0.47	3.12	0.56	-0.11	1.39	-0.84
50	Simazine	-1	0.48	3.08	0.49	-0.06	1.43	-1.52
51	Acetochlor	-1	0.48	3.57	0.64	-0.02	1.15	-1.21
52	Cyanazine	-1	0.47	3.31	0.64	-0.19	1.30	-0.63
53	Iopromide	-1	0.47	5.20	0.62	-0.18	0.88	-3.57
54	Metolachlor	-1	0.48	3.65	0.73	-0.31	0.90	-0.92
55	Propachlor *	-1	0.48	3.27	0.93	0.06	1.19	1.06

* Wrong cases for +1 group (high-reactive) are carbamazepine, and trimethoprim; Wrong cases for -1 group (low-reactive) are diazepam, and propachlor. The class values are calculated using equation 5.10.

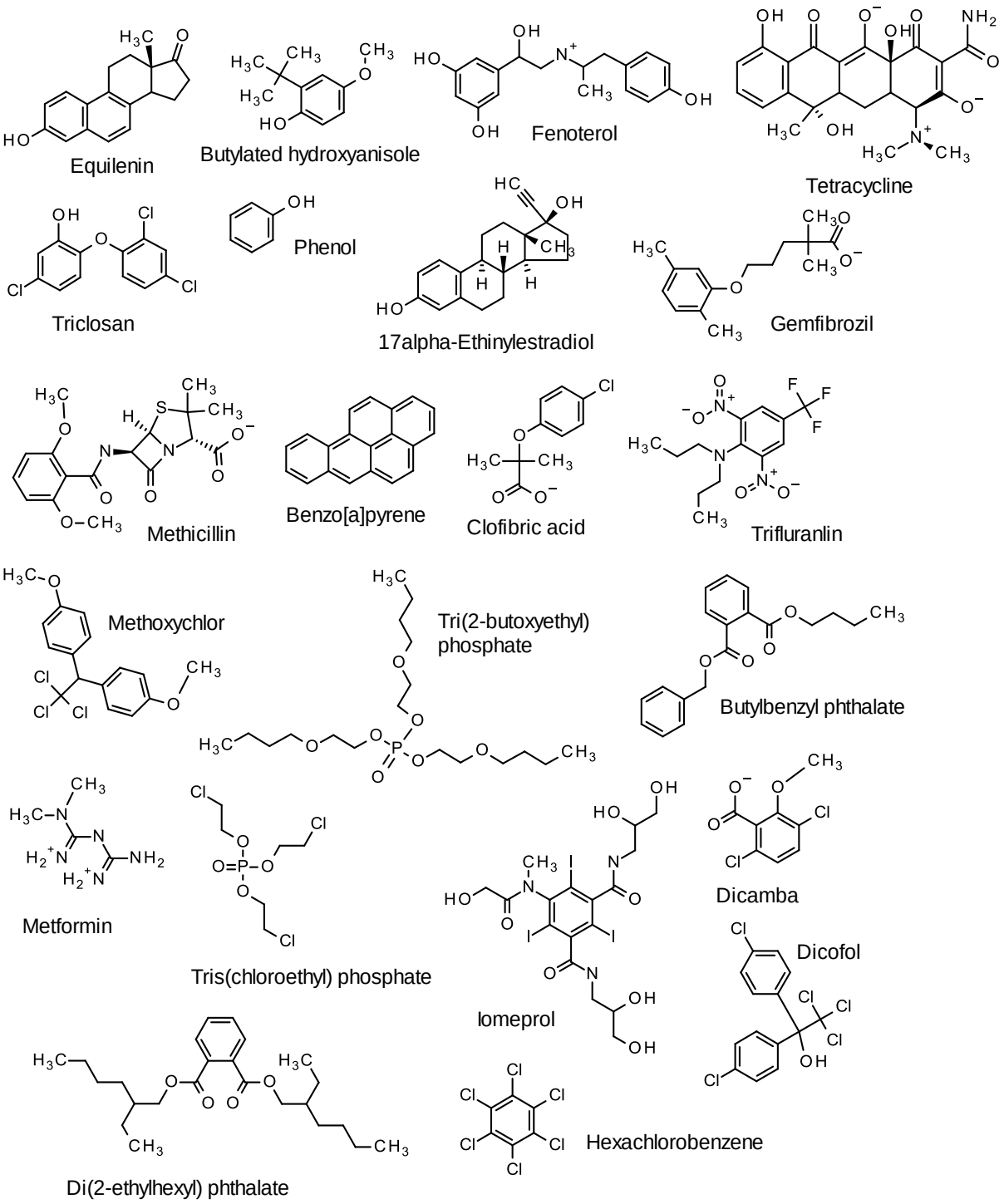


Figure C. 1 Chemical structure of compounds (at pH = 7) in the training set.

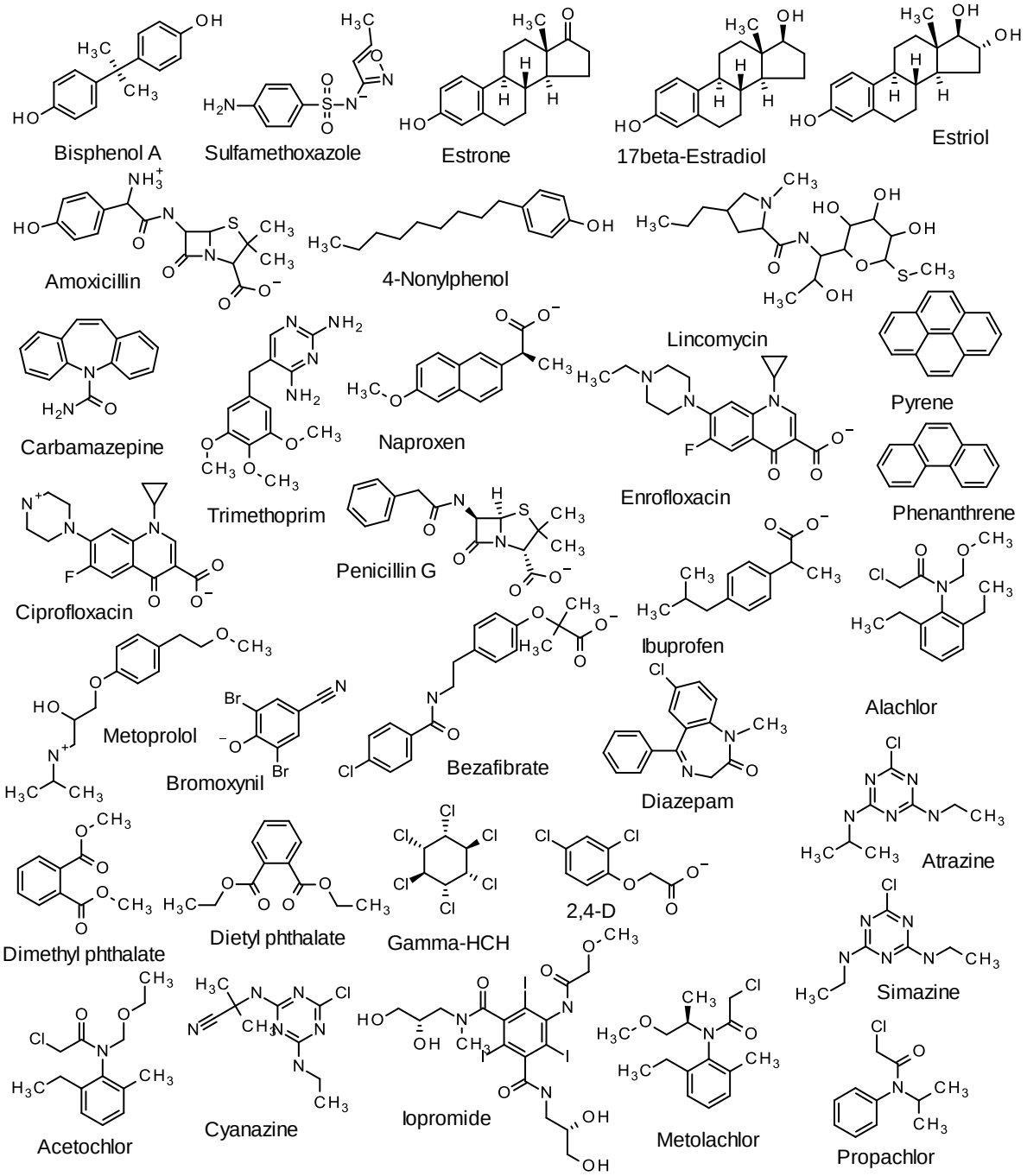


Figure C. 2 Chemical structure of compounds (at pH = 7) in the validation set.

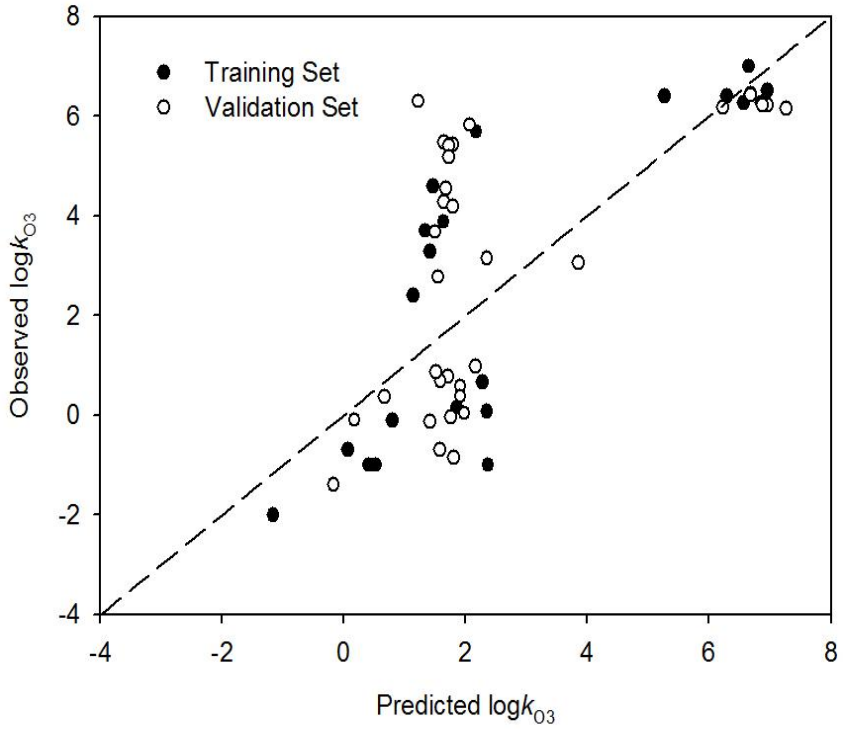


Figure C. 3 Plot of observed logk_{O3} vs. predicted logk_{O3} calculated by MLR model.

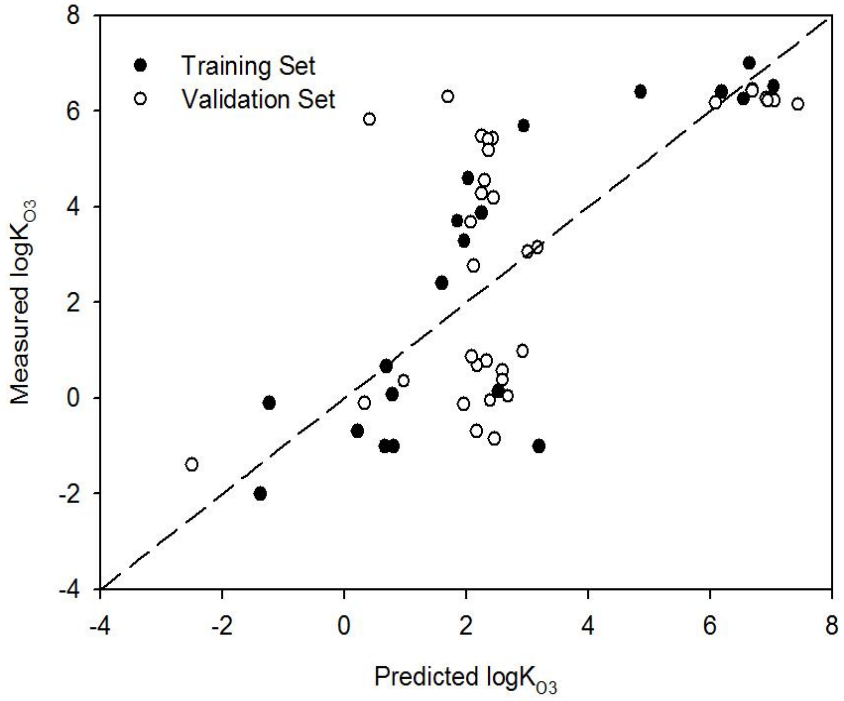


Figure C. 4 Plot of observed logk_{O3} vs. predicted logk_{O3} calculated by PLS model.

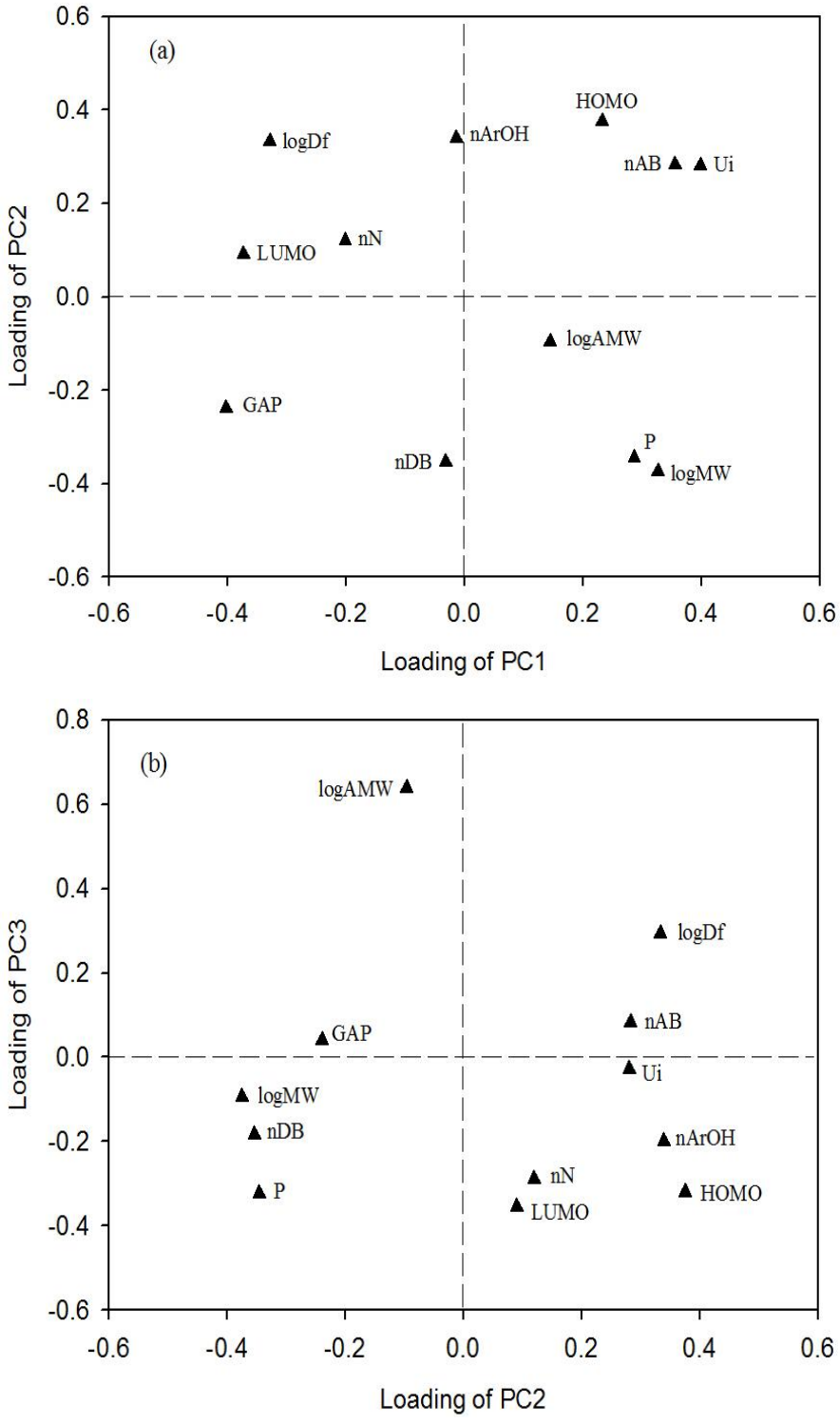


Figure C. 5 Loading plot of the principal components extracted for PCR model (a) PC1 vs. PC2, (b) PC2 vs. PC3.

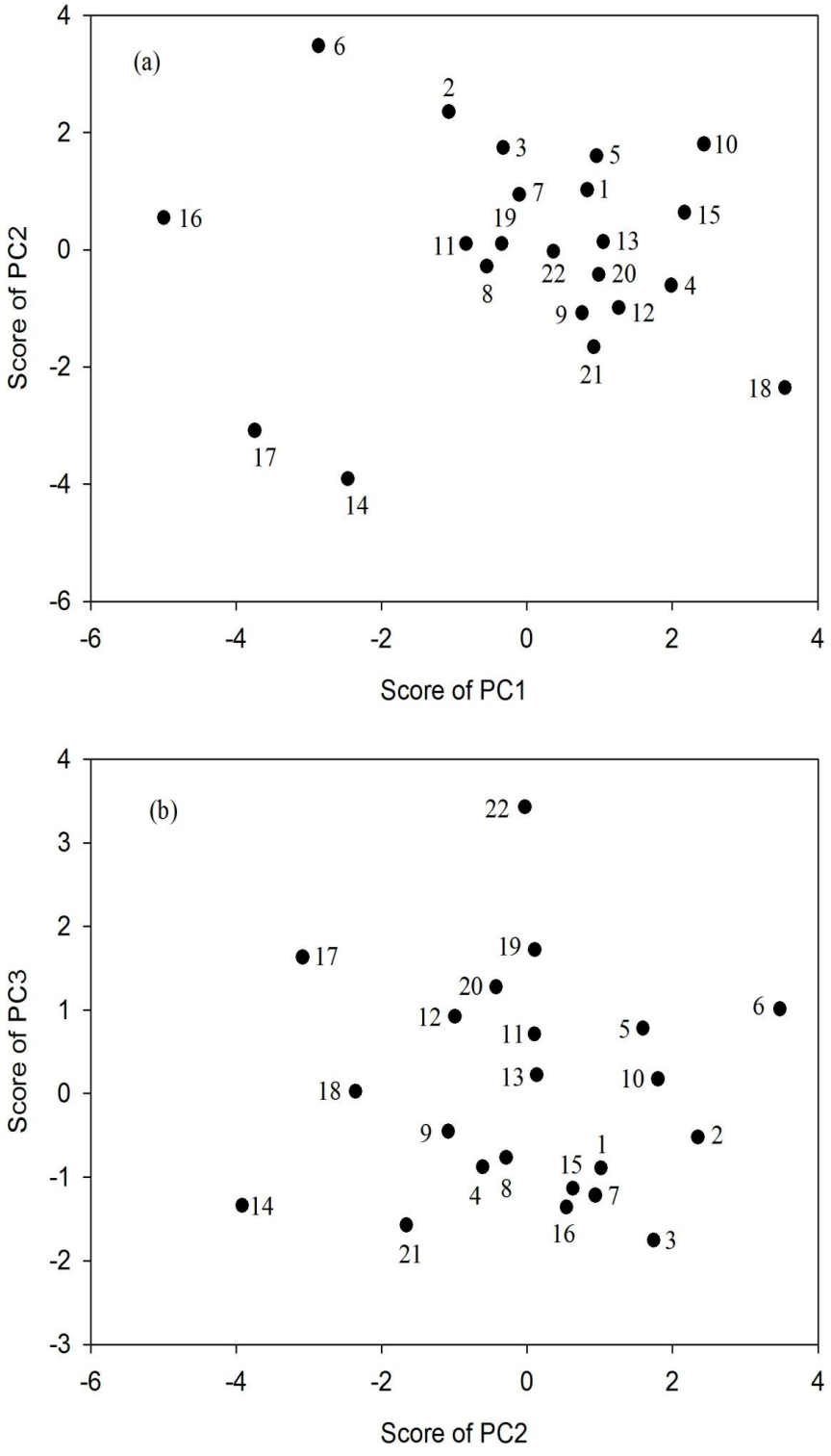


Figure C. 6 Score plot of the principal components extracted for PCR model (a) PC1 vs. PC2, (b) PC2 vs. PC3.

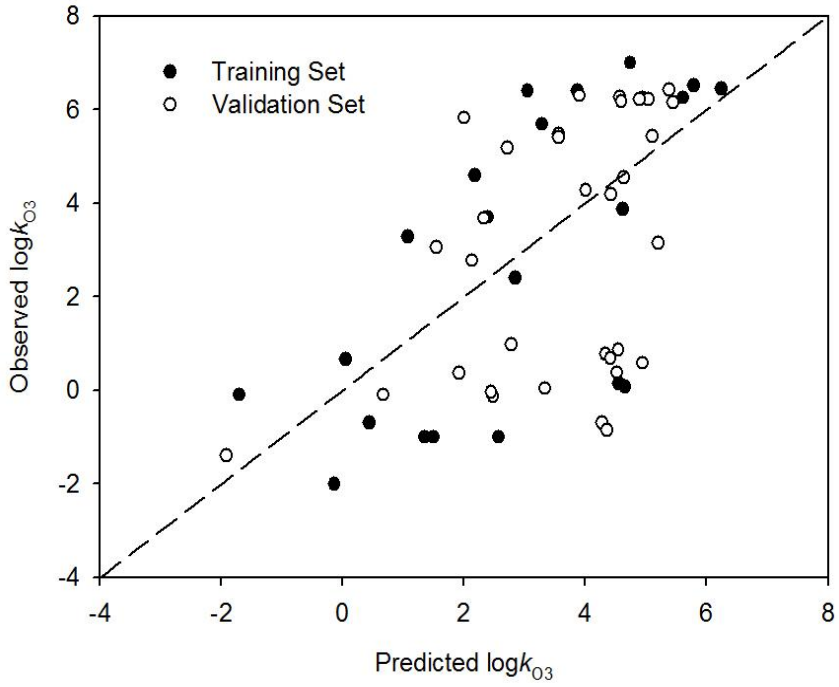


Figure C. 7 Plot of observed logk_{O3} vs. predicted logk_{O3} calculated by PCR model.

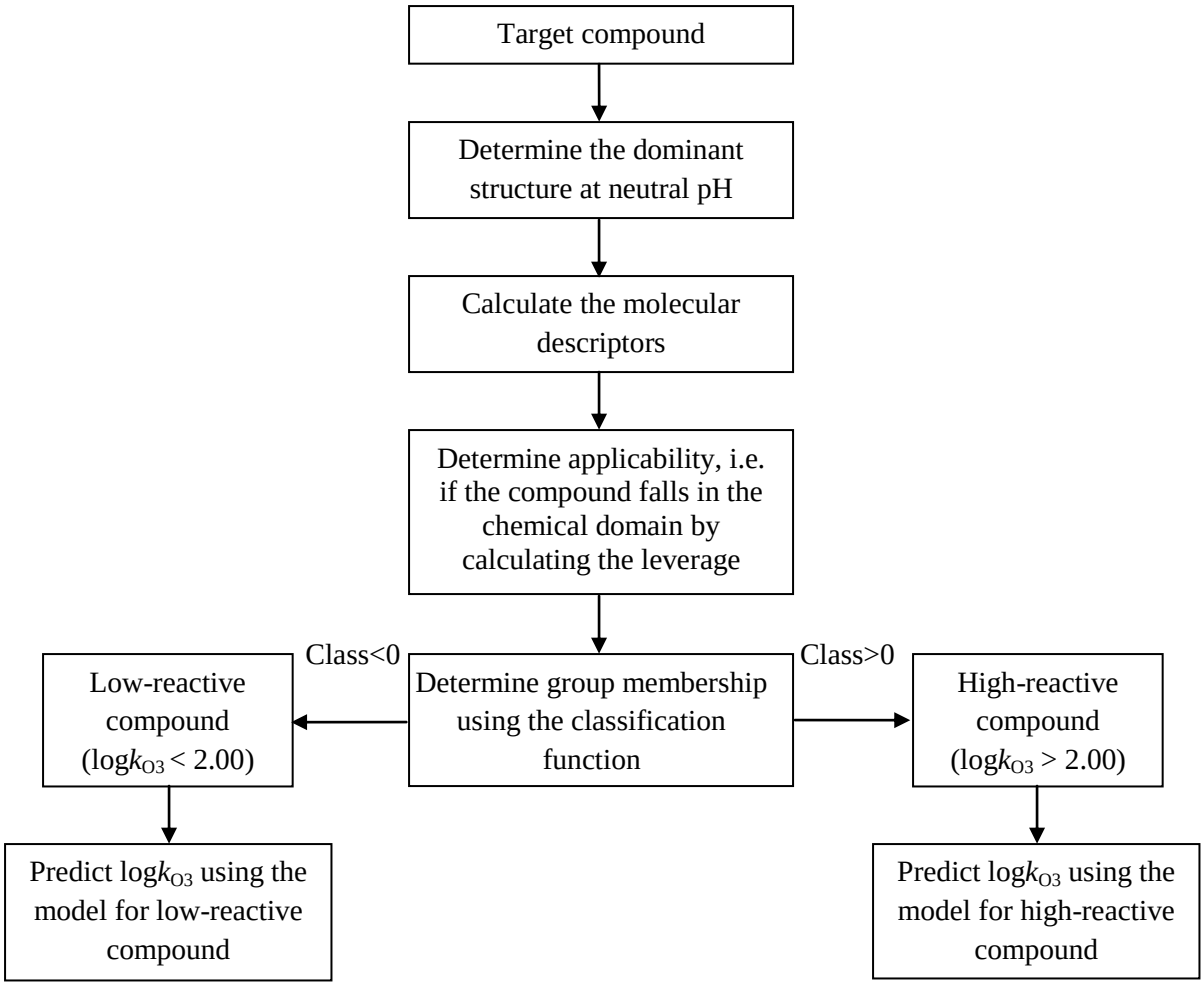


Figure C. 8 The processes of applying the model to predict k_{O3} of a new compound.

Appendix D QSPR Modeling of Hydroxyl Radical Rate Constants Using Selected Micropollutants and Molecular Descriptors

Introduction

Twenty-two micropollutants with diverse structures were selected as representative compounds from a compound pool of 182 micropollutants using a systematic statistical approach (principal component analysis followed by D-optimal onion design), and 12 molecular descriptors were identified to describe the structural features related to the reactivity in ozonation and advanced oxidation processes. The selected micropollutants and descriptors can be used to develop QSPR models because the selection is based on linking the compound properties to oxidation reaction (details shown in Chapter 3). The rate constants of these selected micropollutants in the reaction with ozone (k_{O3}) and hydroxyl radical (k_{OH}) were then experimentally determined at neutral pH (Chapter 4). QSPR models for k_{O3} were successfully developed using a piecewise linear regression – linear discrimination analysis (PLR-LDA) approach using the experimental data of the selected micropollutants, and then externally validated with a validation set including micropollutants collected from the literature (Chapter 5). Following the above studies, the objective of this study is to develop QSPR models for predicting k_{OH} using the selected micropollutants and molecular descriptors.

Materials and Methods

Data set

The training set included the selected 22 micropollutants and their k_{OH} values which were experimentally determined previously (Chapter 4). However because of the unavailability of the instrument, k_{OH} values of two micropollutants, tris(2-butoxyethyl) phosphate and tris(2-chloroethyl) phosphate, were not determined. Instead, reliable literature data (Watts and Linden 2009) were used for these two micropollutants. A validation set including 33 micropollutants (literature data) was used to validate the QSPR model externally (Table D.1).

The selected 12 molecular descriptors were calculated using various software packages (details shown in Chapter 5). Before modeling, the rate constants k_{OH} were converted to log scale, number of functional groups counts were converted to categorical variables with “1” representing presence, and “0” representing the absence of the functional group. Molecular weight (*MW*), average molecular weight (*AMW*), and diffusivity (*Df*) were converted to log scale.

Table D.1 Data set used for QSPR modeling.

Micropollutant	*nDB*	*nArOH*	*Ui*	*LUMO*	[1] logk_{OH}	[2] logk_{OH}	Reference
			Training Set				
17α-Ethinylestradiol	0	1	3	0.402	9.69	9.86	Chapter 4
Benzo[a]pyrene	0	0	4.644	-1.111	8.97	9.15	Chapter 4
Butylated hydroxyanisole	0	1	2.807	0.346	9.87	9.77	Chapter 4
Butylbenzyl phthalate	1	0	3.907	-0.517	9.6	9.84	Chapter 4
Clofibric acid	1	0	3	-0.126	9.72	9.80	Chapter 4
Di(2-ethylhexyl) phthalate	1	0	3.17	-0.102	8.53	NA	Chapter 4
Dicamba	1	0	3	-0.671	9.54	9.48	Chapter 4
Dicofol	0	0	3.7	-0.561	9.57	9.18	Chapter 4
Equilenin	1	1	3.7	-0.421	10.23	10.17	Chapter 4
Fenoterol	0	1	3.7	0.0913	9.59	9.89	Chapter 4
Gemfibrozil	1	0	3	0.297	9.85	10.04	Chapter 4
Hexachlorobenzene	0	0	2.807	-1.041	8.38	8.64	Chapter 4
Iomeprol	1	0	3.322	-1.259	9.4	9.24	Chapter 4
Methicillin	1	0	3.322	-0.173	10	9.87	Chapter 4
Methoxychlor	0	0	3.7	-0.222	9.59	9.38	Chapter 4
Phenol	0	1	2.807	0.398	9.79	9.80	Chapter 4
Pyrene	0	0	4.322	-0.889	9.15	9.18	Chapter 4
Tris(2-butoxyethyl) phosphate	1	0	1	0.434	10.01	9.52	Watts and Linden, 2009
Tris(2-chloroethyl) phosphate	1	0	1	-0.101	8.75	9.21	Watts and Linden, 2009
Tetracycline	1	1	3.585	-0.956	9.91	9.83	Chapter 4
Triclosan	0	1	3.7	-0.540	9.78	9.53	Chapter 4
Trifluralin	1	0	3.459	-1.530	9.11	9.12	Chapter 4
			Validation Set				
17β-Estradiol	0	1	2.807	0.397	10.15	9.80	Rosenfeldt and Linden, 2004
Acetochlor	1	0	3	0.214	9.80	9.99	Benner *et al.*, 2008
Amoxicillin	1	1	3.322	-0.0940	9.84	10.25	Song *et al.*, 2008b
Atenolol	1	0	3	0.0786	9.85	9.91	Song *et al.*, 2008a
Atrazine	0	0	2.807	0.0452	9.50	9.26	Cooper *et al.*, 2010
Bezafibrate	1	0	3.907	-0.420	9.90	9.90	Razavi *et al.*, 2009
Bisphenol A	0	1	3.7	0.397	9.84	10.07	Peller *et al.*, 2009
Carbamazepine	1	0	3.907	-0.458	9.94	9.88	Huber *et al.*, 2003
Chlortetracycline	1	1	3.585	-0.950	9.72	9.83	Cooper *et al.*, 2010
Ciprofloxacin	1	0	3.322	-0.669	9.61	9.58	Dodd *et al.*, 2006
DEET	1	0	3	-0.0684	9.69	9.83	Song *et al.*, 2009
Diazepam	1	0	3.907	-0.605	9.86	9.79	Huber *et al.*, 2003

Diuron	1	0	3	-0.0743	9.87	9.83	Elovitz *et al.*, 2008
Doxycycline	1	1	3.585	-0.825	9.88	9.90	Cooper *et al.*, 2010
Enrofloxacin	1	0	3.322	-0.657	9.90	9.59	Santoke *et al.*, 2009
Gamma-HCH	0	0	0	-0.150	8.76	8.31	Haag and Yao 1992
Ibuprofen	1	0	3	-0.0975	9.78	9.81	Cooper *et al.*, 2010
Iopamidol	1	0	3.322	-1.575	9.53	9.05	Cooper *et al.*, 2010
Levofloxacin	1	0	3.322	-0.750	9.88	9.53	Santoke *et al.*, 2009
Lincomycin	1	0	1	0.484	9.93	9.55	Dodd *et al.*, 2006
Lomefloxacin	1	0	3.322	-0.858	9.91	9.47	Santoke *et al.*, 2009
Metolachlor	1	0	3	0.127	9.83	9.94	Benner *et al.*, 2008
Metoprolol	0	0	2.807	0.413	9.86	9.48	Song *et al.*, 2008a
Naproxen	1	0	3.7	-0.402	9.88	9.85	Cooper *et al.*, 2010
Norfloxacin	1	0	3.322	-0.667	9.82	9.58	Santoke *et al.*, 2009
Oxytetracycline	1	1	3.585	-0.845	9.75	9.89	Cooper *et al.*, 2010
Penicillin G	1	0	3.322	-0.067	9.90	9.93	Song *et al.*, 2008b
Propranolol	0	0	3.585	-0.306	10.03	9.29	Song *et al.*, 2008a
Sulfamerazine	1	0	3.907	-0.479	9.89	9.86	Mezyk *et al.*, 2007
Sulfamethazine	1	0	3.907	-0.443	9.92	9.88	Mezyk *et al.*, 2007
Sulfamethizole	1	0	3.807	-0.737	9.90	9.68	Mezyk *et al.*, 2007
Sulfamethoxazole	1	0	3.807	-0.475	9.93	9.84	Mezyk *et al.*, 2007
Trimethoprim	0	0	3.7	-0.007	9.92	9.50	Cooper *et al.*, 2010

[1] measured logk_{OH} values from Chapter 4 or collected from literature; [2] predicted logk_{OH} values using Equation D.5 (Model 3)

Modeling Method

Multiple linear regression (MLR) is often used in QSPR to identify a linear relationship between a property to be predicted and a set of molecular descriptors. However, when some of the assumptions are invalid (e.g., occurrence of outliers, non-normality, multicollinearity), the ordinary least-square estimation can perform poorly (Ho and Naugher 2000). It is very common in data analysis and statistical modeling applications that a small proportion of observations are far from the rest of the data. Such data or even a single outlier can distort the regression results by pulling the least square fit too much in their direction thereby impacting the regression coefficients, and limiting our ability to understand the data.

Alternatively, robust regression works with less restrictive assumptions and provides much better regression coefficient estimates when outliers are present in the data. The robust regression techniques fit a model that describes the information in the majority of the data (Hample *et al.*, 1986). It can be used to detect outliers and to provide reliable results in the presence of outliers.

One of the most common general methods of robust regression is *M*-estimation introduced by Huber (1964).

The performance of the models are evaluated by R^2, R^2_{adj}, Q^2, R^2_{pred}, *RMSE*, and *RMSEP* (details shown in Chapter 5 and 6).The data analysis and modeling were carried out using the software NCSS 2007 (NCSS, Kaysville, Utah, US). The dependent variable, the rate constant for reaction with the hydroxyl radical (k_{OH}), was transformed to its logarithm ($\log k_{OH}$). The discrete independent variables, such as the functional groups counts, and atom-centered fragments were converted to categorical variable with two categories ("0" represents absence and "1" represents presence).

Applicability Domain and Outlier Identification

The applicability domain is the chemical space defined by the properties of the training set. Predictions for new compounds falling within this space are expected to be reliable. One of the commonly used methods to determine the applicability domain of a QSPR model is to determine the leverage of each compound. Leverage indicates the compound's distance from the centroid of *X* (where *X* is the descriptor matrix):

$$h_i = x_i^T \left(X^T X \right)^{-1} x_i \qquad \textbf{(D.1)}$$

Where x_i is the descriptor vector of the considered compound and *X* is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as:

$$h^* = 3p / n \qquad \textbf{(D.2)}$$

Where *n* is the number of training set compounds, and *p* is the number of descriptors in the model plus one. The leverage values are then plotted against standardized residuals for each compound, i.e. Williams plot. The applicability domain is established by a squared area within $\pm$ 3 standard deviations and a leverage threshold h^*. Thus, compounds with standardized residuals > 3 standard deviation units and h_i > leverage threshold h^* are considered as outliers. However, a high leverage training set compound with small residual is not necessarily an outlier, even though it has been excluded from the applicability domain (Gramatica *et al.,* 2004).

Results and Discussion

Multiple linear regression (forward) was used to analyze the data set. As a result, three molecular descriptors, LUMO energy (*LUMO*), number of double bonds (*nDB*), and number of phenolic group (*nArOH*) were found to be significant, as shown in Equation D.3 (Model 1).

$$\log k_{OH} = 9.255 + 0.239 \times LUMO + 0.515 \times nArOH + 0.315 \times nDB \qquad \textbf{(D.3)}$$

$n_{training}$ = 22, R^2 = 0.397, R^2_{adj} = 0.297, Q^2 = 0.174, F(3,18) = 3.951 (p = 0.025), *RMSE* = 0.372.

However, a very low R^2_{adj} value (0.297) indicated a very poor model fit, and a low Q^2 value (0.174) indicating that the model was over-fitted and not robust. In addition, compounds hexachlorobenzene (HCB), tris(2-chloroethyl) phosphate (TCEP), and di(2-ethylhexyl) phthalate (DEHP) were far from the regression line indicating the presence of potential outliers (Figure D.1). In such situation, robust regression can be used to improve the estimation.

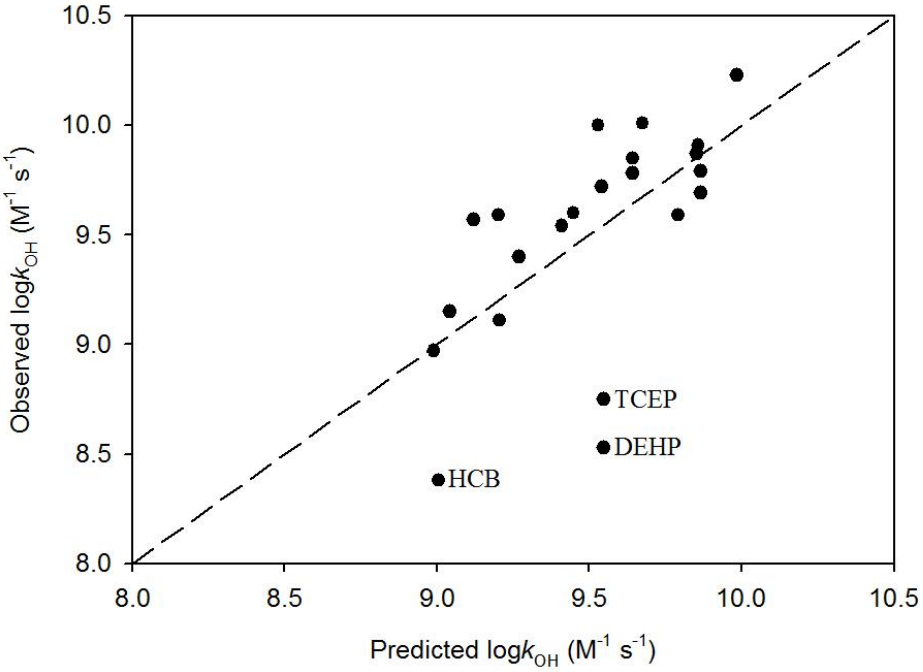


Figure D.1 The predicted logk_{OH} vs. the observed logk_{OH} for Model 1.

As a result of the robust regression, an additional descriptor (unsaturation index, *Ui*) was found to be significant, and the R^2_{adj} was increased from 0.297 to 0.584, and *RMSE* decreased from 0.372 to 0.351, as shown in Equation D.4 (Model 2).

$$\log k_{OH} = 8.333 + 0.518 \times LUMO + 0.394 \times nArOH + 0.535 \times nDB + 0.303 \times Ui \qquad \textbf{(D.4)}$$

$n_{training}$ = 22, R^2 = 0.664, R^2_{adj} = 0.585, Q^2 = 0.081, F (4,17) = 8.403 (p = 0.0006), *RMSE* = 0.351

However, the cross-validated Q^2 value of this model was very low (0.081) indicating the very low robustness of the model (i.e., the model was very much over-fitted as the random error was largely modeled instead of the real variation). In addition, the compound DEHP was still located far from the regression line in the plot of predicted logk_{OH} against measured logk_{OH} (as shown in Figure D.2), however, the prediction of HCB and TCEP was improved mainly because of the additional descriptor *Ui*. Overall, robust MLR was able to improve the model fit, but provide little help in improving the predictive power of the model.

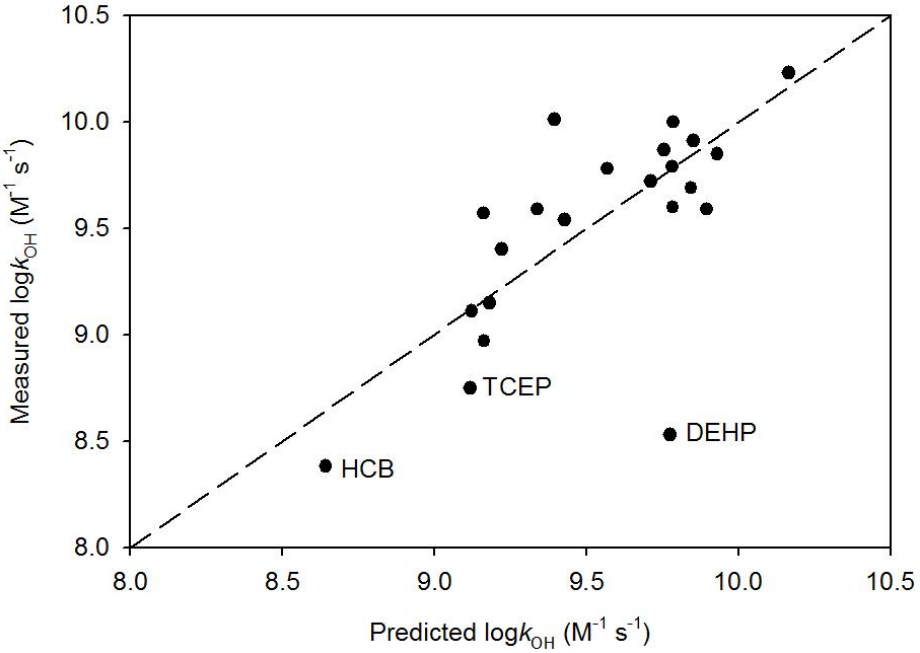


Figure D. 2 The predicted logk_{OH} vs. the observed logk_{OH} for the robust MLR model (Model 2).

Next, the leverage approach (i.e., Williams plot) was used to establish the applicability domain and identify the outliers. As a result, the leverage of all the compounds was less than the warning leverage h^* indicating that no compound is structurally anomalous among the training set. DEHP was identified as an outlier (Figure D.3) because of the large prediction error (> 3 standard deviation). The errors associated with TCEP and HCB were less than 3 units of standard deviation and therefore these two compounds were not considered outliers.

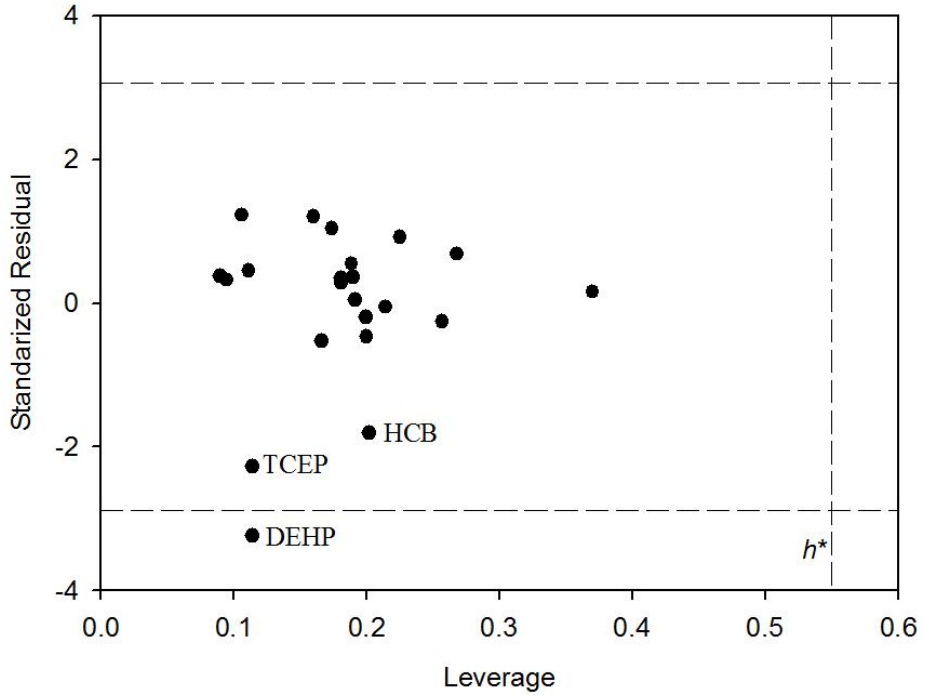


Figure D.3 The William plot for the MLR model D.1.

After removing the outlier compound DEHP, an MLR analysis was conducted on the remaining 21 compounds, as shown in Equation D.5 (Model 3). Similarly, four descriptors were found to be significant. Compared with Model 1 (with outlier), the R^2_{adj} of the model was increased from 0.297 to 0.702, the cross-validated Q^2 value was increased from 0.174 to 0.446, and the prediction error *RMSE* decreased from 0.372 to 0.230. When compared to Model 2 (robust MLR model), Model 3 is also considered better because of higher R^2_{adj}, higher Q^2, and lower *RMSE*. However, the Q^2 value is still lower than 0.5 which is considered an acceptable level (Fan *et al.*, 2001) although some disagree (Golbraikh and Tropsha 2002b).

$$\log k_{OH} = 8.398 + 0.578 \times LUMO + 0.335 \times nArOH + 0.574 \times nDB + 0.299 \times Ui \quad \textbf{(D.5)}$$

$n_{training}$ = 21, R^2 = 0.731, R^2_{adj} = 0.663, Q^2 = 0.440, F (4,16) = 10.85 (p = 0.0002), *RMSE* = 0.229

$n_{validation}$ = 33, R^2_{pred} = 0.368, *RMSEP* = 0.275

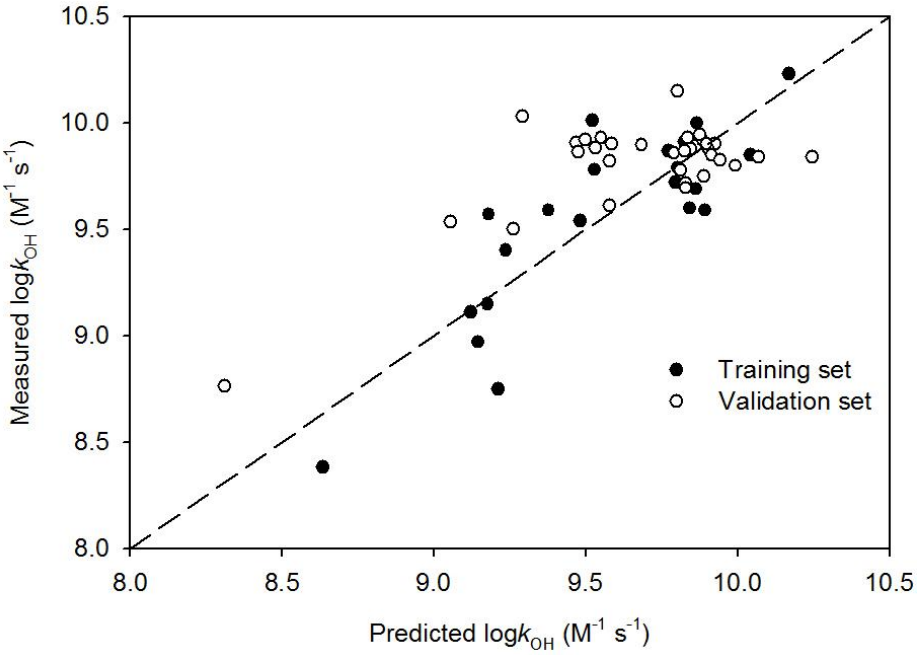


Figure D. 4 The predicted logk_{OH} vs. the observed logk_{OH} for the MLR Model 3.

Finally Model 3 was then externally validated using 33 compounds (validation set) collected from the literature (Table D.1), but the R^2_{pred} value is quite low (0.368) indicating low predictive power of this model (Figure D.4). Overall, QSPR models for k_{OH} prediction were unsatisfactory. Especially since the primary purpose of QSPR modeling is to develop predictive models, the model developed in this study can hardly be used to predict k_{OH} of new, unknown compounds due to the low R^2_{pred}. Unlike ozone, hydroxyl radicals are relatively non-selective oxidants which were indicated by a very small range of k_{OH} values, i.e., the k_{OH} of different micropollutants are very close to each other. In such a situation, the random error associated with the rate constant can have a large impact over the variation of rate constant. It is likely that the selected 12 descriptors are not sensitive enough to capture the structural features which related to hydroxyl radical reactions. In addition, it is likely that the model was not well trained with such a small training set. The model can be improved by increasing the number of compounds and finding better molecular descriptors.

Conclusions

QSPR models for predicting hydroxyl radical rate constants were developed using 22 selected micropollutants and 12 molecular descriptors (details of the selection process were shown in Chapter 3).

Model 1 developed by multiple linear regression showed poor fitting (low R^2) to the training set because of the presence of outliers. Model 2 developed by robust regression showed better fitting (higher R^2) but no improvement in predictivity (lower Q^2). By removing the outlier DEHP, as identified by Williams plot, the best model (Model 3) was obtained (R^2_{adj} = 0.663) but it was still not satisfactory (Q^2 < 0.5). Furthermore, Model 3 was validated with an external data set showing poor predictive power (R^2_{pred} = 0.368). In conclusion, additional compounds and better molecular descriptors are needed to improve the QSPR model for k_{OH} prediction.

Appendix E Supplementary Material for Chapter 7

Table E.1 Data set collected from literature and the calculation of ozone exposure.

#	Compound	k_{O3} (M^{-1} s^{-1})	k_{OH} (M^{-1} s^{-1})	%*Removal*	R_{ct}	O_3 dose (M)	*t* (min)	*k* (s^{-1})	O_3 exposure (Ms)
1	2,6-Nonadienal [a]	8.7×10^{5}	8.95×10^{9}	98	5×10^{-8}				4×10^{-3} #
2	1-Penten-3-one [a]	5.9×10^{4}	4.71×10^{9}	98	5×10^{-8}				4×10^{-3} #
3	Belta-cyclocitral [a]	3.9×10^{3}	7.42×10^{9}	98	5×10^{-8}				4×10^{-3} #
4	Isoproturon [b]	2.2×10^{3}	7.90×10^{9}	87	2.6×10^{-7}	3.0×10^{-5}	10	6.1×10^{-2}	4.9×10^{-4} *
5	Chlortoluron [b]	403.6	6.90×10^{9}	70	2.6×10^{-7}	3.0×10^{-5}	10	6.1×10^{-2}	4.9×10^{-4} *
6	2-Isopropyl-3-methoxypyrazine [a]	50.2	4.91×10^{9}	76	5×10^{-8}				4×10^{-3} #
7	Diuron [b]	16.6	6.60×10^{9}	58	2.6×10^{-7}	3.0×10^{-5}	10	6.1×10^{-2}	4.9×10^{-4} *
8	Atrazine [c]	6	3×10^{9}	61	1.5×10^{-8}	4.2×10^{-5}	30	2.2×10^{-3}	1.9×10^{-2} *
9	Butachlor [d]	5.32	7.40×10^{9}	67	1.9×10^{-8}	4.2×10^{-5}	20	6.1×10^{-3}	6.9×10^{-3} *
10	Acetochlor [d]	2.39	6.30×10^{9}	60	1.9×10^{-8}	4.2×10^{-5}	20	6.1×10^{-3}	6.9×10^{-3} *
11	Linuron [b]	1.8	5.90×10^{9}	50	2.6×10^{-7}	3.0×10^{-5}	10	6.1×10^{-2}	4.9×10^{-4} *
12	Propachlor [d]	1.24	4.60×10^{9}	45	1.9×10^{-8}	4.2×10^{-5}	20	6.1×10^{-3}	6.9×10^{-3} *
13	2-Methylisoborneol [a]	0.35	5.09×10^{9}	59	5×10^{-8}				4×10^{-3} #
14	MTBE [e]	0.14	1.90×10^{9}	33	4×10^{-9}	4.2×10^{-5}	10	7.7×10^{-4}	2.0×10^{-2} *
15	Geosmin [a]	0.1	7.80×10^{9}	35	2×10^{-8}				4×10^{-3} #
16	2,4,6-Tribromoanisole [a]	0.02	3.74×10^{9}	60	5×10^{-8}				4×10^{-3} #

Reference and water sampling location: [a] Peter and von Gunten 2007, Lake Zurich, Swiss; [b] Benitez *et al.*, 2007, reservoir "Peña del Aguila", Spain; [c] Acero *et al.*, 2000, River Seine, France; [d] Acero *et al.*, 2003, Zujar and Orellana reservoirs, Spain; [e] Acero *et al.*, 2001, Lake Murten, Swiss. *t* is the contact time (min) and *k* is the ozone decay rate constant (s^{-1}), # Ozone exposure (integration of the ozone residual concentration over time) values were reported, * ozone exposure values were calculated by Equation 7.7.

References

Acero, J. L., Benitez, F. J., Real, F. J., Maya, C. (2003) Oxidation of acetamide herbicides in natural waters by ozone and by the combination of ozone/hydrogen peroxide: Kinetics study and process modeling. Ind. Eng. Chem. Res. 42, 5762-5769.

Acero, J. L., Stemmler, K., von Gunten, U. (2000) Degradation kinetics of atrazine and its degradation products with ozone and OH radicals: A predictive tool for drinking water treatment. Environ. Sci. Technol. 34, 591-597.

Acero, J. L., von Gunten, U. (2001) Characterization of oxidation processes: Ozonation and the AOP O_3/H_2O_2. J. Am. Water Works Assoc. 93, 90-100.

Adams, C., Wang, Y., Loftin, K., Meyer, M. (2002) Removal of antibiotics from surface and distilled water in conventional water treatment processes. J. Environ. Eng. ASCE 128(3), 253-260.

Andersen, C. M. and Bro, R. (2010) Variable selection in regression-a tutorial. J. Chemometr. 24(11-12), 728-737.

Andersson, P. M., Sjostrom, M., Wold, S., Lundstedt, T. (2000) Comparison between physicochemical and calculated molecular descriptors. J. Chemometr. 14(5-6), 629-642.

Andreozzi, R., Raffaele, M., Nicklas, P. (2003) Pharmaceuticals in STP effluents and their solar photodegradation in aquatic environment. Chemosphere. 50, 1319-1330.

Andreozzi, R., Caprio, V., Ciniglia, C., de Champdore, M., Lo Giudice, R., Marotta, R., Zuccato, E. (2004) Antibiotics in the environment: occurrence in Italian STPs, fate, and preliminary assessment on algal toxicity of amoxicillin. Environ. Sci. Technol. 38, 6832-6838.

Andre, C. D. S., Narula, S. C., Elian, S. N., Tavares, R. A. (2003) An overview of the variables selection methods for the minimum sum of absolute errors regression. Stat. Med. 22(13), 2101-2111.

Andresen, J. and Bester, K. (2006) Elimination of organophosphate ester flame retardants and plasticizers in drinking water purification. Water Res. 40(3), 621-629.

Atkinson, R. (2000) Atmospheric oxidation. In Boethling, R. S., Mackay, D., eds, Handbook of Property Estimation Methods for Chemicals. Environmental and Health Sciences. CRC, Boca Raton, FL, USA.

Atkinson, R. (1988) Estimation of gas-phase hydroxyl radical rate constants for organic-chemicals. Environ. Toxicol. Chem. 7(6), 435-442.

Atkinson, R. (1987) A structure-activity relationship for the estimation of rate constants for the gas-phase reactions of OH radicals with organic compounds. Int. J. Chem. Kinet. 19(9), 799-828.

Bader, H. and Hoigné, J. (1981) Determination of ozone in water by the indigo method. Water Res. 15(4), 449-456.

Batt, A.L., Aga, D.S. (2005) Simultaneous analysis of multiple classes of antibiotics by ion trap LC/MS. Anal. Chem. 77, 2940-2947.

Bellona, C., Drewes, J. E., Xu, P., Amy, G. (2004) Factors affecting the rejection of organic solutes during NF/RO treatment - a literature review. Water Res. 38(12), 2795-2809.

Benitez, F. J., Acero, J. L., Real, F. J., Roldan, G. (2009) Ozonation of pharmaceutical compounds: rate constants and elimination in various water matrices. Chemosphere 77, 53-59.

Benitez, F. J., Beltran-Heredia, J., Gonzalez, T. (1994) Degradation by ozone and UV-radiation of the herbicide cyanazine. Ozone-Sci. Eng. 16, 213-234.

Benitez, F. J., Real, F. J., Acero, J. L., Garcia, C. (2007) Kinetics of the transformation of phenyl-urea herbicides during ozonation of natural waters: Rate constants and model predictions. Water Res. 41(18), 4073-4084.

Benner, J., Salhi, E., Ternes, T., von Gunten, U. (2008) Ozonation of reverse osmosis concentrate: Kinetics and efficiency of beta blocker oxidation. Water Res. 42(12), 3003-3012.

Benner, J., Ternes, T. A. (2009) Ozonation of metoprolol: elucidation of oxidation pathways and major oxidation products. Environ. Sci. Technol. 43(14), 5472-5480.

Benotti, M. J., Trenholm, R. A., Vanderford, B. J., Holady, J. C., Stanford, B. D., Snyder, S. A. (2009) Pharmaceuticals and endocrine disrupting compounds in US drinking water. Environ. Sci. Technol. 43(3), 597-603.

Bolton, J. and Linden, K. (2003) Standardization of methods for fluence (UV dose) determination in bench-scale UV experiments. J. Environ. Eng. -ASCE 129(3), 209-215.

Box, G. E. P. and Draper, N. R. (1987) Empirical Model-Building and Response Surfaces. Wiley, New York.

Brasquet, C., Subrenat, E., LeCloirec, P. (1997) Selective adsorption on fibrous activated carbon of organics from aqueous solution: Correlation between adsorption and molecular structure. Water Sci. Technol. 35(7), 251-259.

Brossa, L., Marce, R.A., Borrull, F., Pocurull, E. (2005) Occurrence of twenty-six endocrine-disrupting compounds in environmental water samples from Catalonia, Spain. Environ. Toxicol. Chem. 24, 261-267.

Brown, H. C., Okamoto, Y. (1958) Substituent constants for aromatic substitution. J. Am. Chem. Soc. 80, 4979.

Buffle, M. O., Schumacher, J., Salhi, E., Jekel, M., von Gunten, U. (2006) Measurement of the initial phase of ozone decomposition in water and wastewater by means of a continuous quench-flow system: Application to disinfection and pharmaceutical oxidation. Water Res. 40, 1884-1894.

Bundy, M. M., Doucette, W. J., McNeill, L., Ericson, J. F. (2007) Removal of pharmaceuticals and related compounds by a bench-scale drinking water treatment system. J. Water Supply Res. Technol.-AQUA 56(2), 105-115.

Burden, F. R., Ford, M. G., Whitley, D. C., Winkler, D. A. (2000) Use of automatic relevance determination in QSAR studies using Bayesian neural networks. J. Chem. Inf. Comput. Sci. 40(6), 1423-1430.

Burden, F. R. and Winkler, D. A. (1999) Robust QSAR models using Bayesian regularized neural networks. J. Med. Chem. 42(16), 3183-3187.

Butkovic, V., Klasinc, L., Orhanovic, M., Turk, J., Gusten, H. (1983) Reaction-rates of polynuclear aromatic-hydrocarbons with ozone in water. Environ. Sci. Technol. 17(9), 546-548.

Buxton, G. V., Greenstock, C. L., Helman, W. P., Ross, A. B. (1988) Critical review of rate constants for reactions of hydrated electrons, hydrogen-atoms and hydroxyl radicals (OH/O^-) in aqueous solution. J. Phys. Chem. Ref. Data 17(2), 513-886.

Calza, P., Medana, C., Pazzi, M., Baiocchi, C., Pelizzetti, E. (2004) Photocatalytic transformations of sulphonamides on titanium dioxide. Appl. Catal. B-Environ. 53, 63-69.

Canonica, S. and Tratnyek, P. G. (2003) Quantitative structure-activity relationships for oxidation reactions of organic chemicals in water. Environ. Toxicol. Chem. 22(8), 1743-1754.

Cao, Y., Jiang, T., Girke, T. (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics 24, 366-374.

Castegnaro, M., DeMeo, M., Laget, M., Michelon, J., Garren, L., Sportouch, M. H., Hansel, S. (1997) Chemical degradation of wastes of antineoplastic agents. 2. Six anthracyclines: idarubicin, doxorubicin, epirubicin, pirarubicin, aclarubicin, and daunorubicin. Int. Arch. Occup. Environ. Health. 70, 378-384.

Cheme-Ayala, P., El-Din, M. G., Smith, D. W. (2010) Kinetics and mechanism of the degradation of two pesticides in aqueous solutions by ozonation. Chemosphere 78, 557-562.

Cleuvers, M. (2004) Mixture toxicity of the anti-inflammatory drugs diclofenac, ibuprofen, naproxen, and acetylsalicylic acid. Ecotoxicol. Environ. Saf. 59(3), 309-315.

Colborn, T., vom Saal, F. S., Soto, A. M. (1993) Development effects of endocrine-disrupting chemicals in wildlife and humans. Environ. Health Perspect. 101, 378-384.

Cokgor, E. U., Alaton, I. A., Karahan, O., Dogruel, S., Orhon, D. (2004) Biological treatability of raw and ozonated penicillin formulation effluent. J. Hazard. Mater. 116, 159-166.

Comerton, A. M., Andrews, R. C., Bagley, D. M., Hao, C. (2008) The rejection of endocrine disrupting and pharmaceutically active compounds by NF and RO membranes as a function of compound, a water matrix properties. J. Membr. Sci. 313(1-2), 323-335.

Cooper, W. J., Snyder, S. A., Mezyk, S. P., Peller, J. R., Nickelsen, M. G. (2010) Reaction rates and mechanisms of advanced oxidation processes for water reuse. WateReuse Foundation Project Number: WRF-04-017. WateReuse Foundation, Alexandria, VA, USA.

Crittenden, J. C., Sanongraj, S., Bulloch, J. L., Hand, D. W., Rogers, T. N., Speth, T. F., Ulmer, M. (1999) Correlation of aqueous-phase adsorption isotherms. Environ. Sci. Technol. 33(17), 2926-2933.

Cronin, M., Walker, J. D., Jaworska, J. S., Comber, M., Watts, C., Worth, A. P. (2003) Use of (Q)SARs in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances. Environ. Health Perspect. 111, 1376-1390.

Cronin, M. T. D. and Schultz, T. W. (2003) Pitfalls in QSAR. J. Mol. Struc. Theochem 622(1-2), 39-51.

Crosina, Q., Peldszus, S., Huck, P. M. (2006) The degradation efficiency of select PPCPs using UV and UV/H_2O_2 - degradation kinetics and application to pretreated water. Proceedings of the AWWA Water Quality Technology Conference and Exposition (WQTC). Denver, Colorado, USA.

Daughton, C. G., Ternes, T. A. (1999) Pharmaceuticals and personal care products in the environment: Agents of subtle change? Environ. Health Perspect. 107, 907-938.

De Aguiar, P. F., Bourguignon, B., Khots, M. S., Massart, D. L., PhanThanLuu, R. (1995) D-optimal designs. Chemometr. Intellig. Lab. Syst. 30(2), 199-210.

Deborde, M., Rabouan, S., Duguet, J. P., Legube, B. (2005) Kinetics of aqueous ozone-induced oxidation of some endocrine disruptors. Environ. Sci. Technol. 39(16), 6086-6092.

Dodd, M. C., Buffle, M. O., von Gunten, U. (2006) Oxidation of antibacterial molecules by aqueous ozone: Moiety-specific reaction kinetics and application to ozone-based wastewater treatment. Environ. Sci. Technol. 40(6), 1969-1977.

Dunn, W. J. (1989) Quantitative structure-activity relationships (QSAR). Chemometr. Intell. Lab. 6(3), 181-190.

Dutot, A. L., Rude, J., Aumont, B. (2003) Neural network method to estimate the aqueous rate constants for the OH reactions with organic compounds. Atmos. Environ. 37(2), 269-276.

Eaton, A. D., Clesceri, L. S., Rice, E. W., Greenberg, A. E. (2005). Standard methods for the examination of water and wastewater. 21st Edition. American Public Health Association. U.S.

Einschlag, F. S. G., Carlos, L., Capparelli, A. L. (2003) Competition kinetics using the UV/H2O2 process: a structure reactivity correlation for the rate constants of hydroxyl radicals toward nitroaromatic compounds. Chemosphere 53(1), 1-7.

Elovitz, M. S., Shemer, H., Peller, J. R., Vinodgopal, K., Sivaganesan, M., Linden, K. G. (2008) Hydroxyl radical rate constants: comparing UV/H_2O_2 and pulse radiolysis for environmental pollutants. J. Water Supply Res. T. 57(6), 391-401.

Elovitz, M. S. and von Gunten, U. (1999) Hydroxyl radical ozone ratios during ozonation processes. I The R_{ct} concept. Ozone-Sci. Eng. 21(3), 239-260.

Elovitz, M. S., von Gunten, U., Kaiser, H. P. (2000) Hydroxyl radical/ozone ratios during ozonation processes. II. The effect of temperature, pH, alkalinity and DOM properties. Ozone Sci. Eng. 22, 123-150.

Eriksson, L., Andersson, P. L., Johansson, E., Tysklind, M. (2006) Megavariate analysis of environmental QSAR data. Part I - A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Mol. Divers. 10(2), 169-186.

Eriksson, L., Arnhold, T., Beck, B., Fox, T., Johansson, E., Kriegl, J. M. (2004) Onion design and its application to a pharmaceutical QSAR problem. J. Chemometr. 18(3-4), 188-202.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M., Gramatica, P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111(10), 1361-1375.

Eriksson, L. and Johansson, E. (1996) Multivariate design and modeling in QSAR. Chemometr. Intellig. Lab. Syst. 34(1), 1-19.

Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. (2001) Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. J. Med. Chem. 44, 3254-3263.

Fent, K., Weston, A. A., Caminada, D. (2006) Ecotoxicology of human pharmaceuticals. Aquat. Toxicol. 76(2), 122-159.

Gnielinski, V. (1978) Gleichungen zur Berechnung des Warme und Stoffaustausches in durchstromten ruhenden Kugelschuttungen bei mittleren und grossen Pecletzahlen. Verf. Tech, 12, 363-366.

Gramatica, P. (2007) Minireview: Principles of QSAR models validation: Internal and external. QSAR Comb. Sci. 26, 694-701.

Gramatica, P., Pilutti, P., Papa, E. (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. J. Chem. Inf. Comp. Sci. 44, 1794-1802.

Gramtica, P., Papa, E. (2005) An update of the BCF QSAR model based on theoretical molecular descriptors. QSAR Comb. Sci. 24, 953-960.

Gramatica, P., Pilutti, P., Papa, E. (2003) QSAR prediction of ozone tropospheric degradation. QSAR Comb. Sci. 22(3), 364-373.

Golbraikh, A., Shen, M., Xiao, Z. Y., Xiao, Y. D., Lee, K. H., Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. 17(2), 241-253.

Golbraikh, A. and Tropsha, A. (2002a) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J. Comput. Aided Mol. Des. 16(5-6), 357-369.

Golbraikh, A. and Tropsha, A. (2002b) Beware of q^2! J. Mol. Graph. Model. 20(4), 269-276.

Grover, R., Waite, D. T., Cessna, A. J., Nicholaichuk, W., Irvin, D. G., Kerr, L. A., Best, K. (1997) Magnitude and persistence of herbicide residues in farm dugouts and ponds in the Canadian prairies. Environ. Toxicol. Chem. 16, 638-643.

Guha, R. and Jurs, P. C. (2005) Determining the validity of a QSAR model - A classification approach. J. Chem. Inf. Model. 45(1), 65-73.

Gurol, M. D. and Nekouinaini, S. (1984) Kinetic-behavior of ozone in aqueous-solutions of substituted phenols. Ind. Eng. Chem. Fund. 23(1), 54-60.

Haag, W. R. and Yao, C. C. D. (1992) Rate constants for reaction of hydroxyl radicals with several drinking water contaminants. Environ. Sci. Technol. 26(5), 1005-1013.

Hammett, L. P. (1937) The effect of structure upon the reactions of organic compounds benzene derivatives. J. Am. Chem. Soc. 59, 96-103.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986) Robust Statistics. The Approach based on Influence Functions. John Wiley and Sons, New York.

Hansch, C. and Leo, A. (1995) Exploring QSAR: fundamentals and application in chemistry and biology. American Chemical Society, Washington, DC.

Hansch, C. and Gao, H. (1997) Comparative QSAR: Radical reactions of benzene derivatives in chemistry and biology. Chem. Rev. 97(8), 2995-3059.

Hansch, C., Leo, A., Taft, R. W. (1991) A survey of Hammett substituent constants and resonance and field parameters. Chem. Rev. 91(2), 165-195.

Harju, M., Andersson, P. L., Haglund, P., Tysklind, M. (2002) Multivariate physico-chemical characterization and quantitative structure-property relationship modeling of polybrominated diphenyl ethers. Chemosphere. 47(4), 375-384.

Hawkins, D. M., Basak, S. C., Mills, D. (2003) Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci. 43(2), 579-586.

Heberer, T. (2002) Occurrence, fate, and removal of pharmaceutical residues in the aquatic environment: a review of recent research data. Toxicol. Lett. 131(1-2), 5-17.

Hirsch, R., Ternes, T., Haberer, K., Kratz, K. L. (1999) Occurrence of antibiotics in the aquatic environment. Sci. Total Environ. 225, 109-118.

Ho, K., Naugher, J. (2000) Outlier lie: An illustrative example of identifying outliers and applying robust models. Multiple Linear Regression Viewpoints 26(2), 2-6.

Hoigné, J. and Bader, H. (1983a) Rate constants of reactions of ozone with organic and inorganic-compounds in water .1. Non-dissociating organic-compounds. Water Res. 17(2), 173-183.

Hoigné, J. and Bader, H. (1983b) Rate constants of reactions of ozone with organic and inorganic-compounds in water .2. Dissociating organic-compounds. Water Res. 17(2), 185-194.

Hu, J. Y., Morita, T., Magara, Y., Aizawa, T. (2000) Evaluation of reactivity of pesticides with ozone in water using the energies of frontier molecular orbitals. Water Res. 34(8), 2215-2222.

Huber, M. M., Canonica, S., Park, G. Y., von Gunten, U. (2003) Oxidation of pharmaceuticals during ozonation and advanced oxidation processes. Environ. Sci. Technol. 37(5), 1016-1024.

Huber, P. J. (1964) Robust Estimation of a Location Parameter. Annals of Mathematical Statistics 35, 73-101.

Huber, M. M., Gobel, A., Joss, A., Hermann, N., Loffler, D., Mcardell, C. S., Ried, A., Siegrist, H., Ternes, T. A., von Gunten, U. (2005) Oxidation of pharmaceuticals during ozonation of municipal wastewater effluents: A pilot study. Environ. Sci. Technol. 39(11), 4290-4299.

Huerta-Fontela, M., Galceran, M. T., Ventura, F. (2011) Occurrence and removal of pharmaceuticals and hormones through drinking water treatment. Water Res. 45, 1432-1442.

Ikehata, K., Naghashkar, N. J., Ei-Din, M. G. (2006) Degradation of aqueous pharmaceuticals by ozonation and advanced oxidation processes: A review. Ozone-Sci. Eng. 28(6), 353-414.

Ikehata, K. and El-Din, M. G. (2005) Aqueous pesticide degradation by ozonation and ozone-based advanced oxidation processes: A review (Part I). Ozone-Sci. Eng. 27(2), 83-114.

Jackson, J. E. (1991) A Users Guide to Principal Components. John Wiley & Sons, Inc., New York,

Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T. (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: A review. ATLA Altern. Lab. Anim. 33(5), 445-459.

Jiang, J. L., Yue, X. A., Chen, Q. F., Gao, Z. (2010) Determination of ozonation reaction rate constants of aromatic pollutants and QSAR study. Bull. Environ. Contam. Toxicol 85, 568-572.

Jin, X., Peldszus, S., Huck, P. M. (2009) Optimized selection strategy to identify representative emerging contaminants for removal studies involving oxidation processes. Proceedings of the AWWA Water Quality Technology Conference and Exposition (WQTC), Seattle, Washington, USA.Karelson, M., Lobanov, V. S., Katritzky, A. R. (1996) Quantum-chemical descriptors in QSAR/QSPR studies. Chem. Rev. 96(3), 1027-1043.

Jin, X. and Peldszus, S. (2012) Selection of representative emerging micropollutants for drinking water treatment studies: A systematic approach. Sci. Total Environ. 414(1), 653-663.

Joback, K. G., Reid, R. C. (1987) Estimation of pure-component properties from group-contributions. Chem. Eng. Commun. 57, 233-243.

Kim, T., Drewes, J. E., Summers, R. S., Amy, G. (2007) Solute transport model for trace organic neutral and charged compounds through nanofiltration and reverse osmosis membranes. Water Res. 41(17), 3977-3988.

Kitti, A., Harju, M., Tysklind, M., van Bavel, B. (2003) Multivariate characterization of polycyclic aromatic hydrocarbons using semi-empirical molecule orbital calculations and physical data. Chemosphere 50(5), 627-637.

Klassen, N. V., Marchington, D., Mcgowan, H. C. E. (1994) H2O2 determination by the I3- method and by KMnO4 titration. Anal. Chem. 66(18), 2921-2925.

Knekta, E., Andersson, P. L., Johansson, M., Tysklind, M. (2004) An overview of OSPAR priority compounds and selection of a representative training set. Chemosphere 57(10), 1495-1503.

Kolpin, D. W., Furlong, E. T., Meyer, M. T., Thurman, E. M., Zaugg, S. D., Barber, L. B., Buxton, H. T. (2002) Pharmaceuticals, hormones, and other organic wastewater contaminants in US streams, 1999-2000: A national reconnaissance. Environ. Sci. Technol. 36(6), 1202-1211.

Kusic, H., Rasulev, B., Leszczynska, D., Leszczynski, J., Koprivanac, N. (2009) Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: A QSAR study. Chemosphere 75(8), 1128-1134.

Kwok, E. S. C. and Atkinson, R. (1995) Estimation of hydroxyl radical reaction rate constants for gas-phase organic compounds using a structure reactivity relationship - an update. Atmos. Environ. 29(14), 1685-1695.

Kümmerer, K. (2001) Drugs in the environment: emission of drugs, diagnostic aids and disinfectants into wastewater by hospitals in relation to other sources - a review. Chemosphere 45(6-7), 957-969.

Leardi, R., Boggia, R., Terrile, M. (1992) Genetic algorithms as a strategy for feature-selection. J. Chemometr. 6(5), 267-281.

Lei, H. X. and Snyder, S. A. (2007) 3D QSPR models for the removal of trace organic contaminants by ozone and free chlorine. Water Res. 41(18), 4051-4060.

Leonard, J. T. and Roy, K. (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb. Sci. 25(3), 235-251.

Loos, R., Wollgast, J., Huber, T., Hanke, G. (2007) Polar herbicides, pharmaceutical products, perfluorooctanesulfonate (PFOS), perfluorooctanoate (PFOA), and nonylphenol and its

carboxylates and ethoxylates in surface and tap waters around Lake Maggiore in Northern Italy. Anal. Bioanal. Chem. 387(4), 1469-1478.

Loraine, G. A., Pettigrove, M. E. (2006) Seasonal variations in concentrations of pharmaceuticals and personal care products in drinking water and reclaimed wastewater in Southern California. Environ. Sci. Technol. 40, 687-695.

Luehrs, D. C., Hickey, J. P., Nilsen, P. E., Godbole, K. A., Rogers, T. N. (1996) Linear solvation energy relationship of the limiting partition coefficient of organic solutes between water and activated carbon. Environ. Sci. Technol. 30, 143-152.

Magnuson, M. L., Speth, T. F. (2005) Quantitative structure - property relationships for enhancing predictions of synthetic organic chemical removal from drinking water by granular activated carbon. Environ. Sci. Technol. 39(19), 7706-7711.

Makdissy, G., Peldszus, S., McPhail, R., Huck, P. M. (2007) Towards a mechanistic understanding of the impact of fouling on the removal of EDCs/PPCPs by nanofiltration membranes. Proceedings to the AWWA Water Quality Technology Conference and Exposition (WQTC). Charlotte, NC, USA.

Matter, H. (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. J. Med. Chem. 40, 1219-1229.

Metivier-Pignon, H., Faur, C., Le Cloirec, P. (2007) Adsorption of dyes onto activated carbon cloth: Using QSPRs as tools to approach adsorption mechanisms. Chemosphere 66(5), 887-893.

Mezyk, S. P., Neubauer, T. J., Cooper, W. J., Peller, J. R. (2007) Free-radical-induced oxidative and reductive degradation of sulfa drugs in water: Absolute kinetics and efficiencies of hydroxyl radical and hydrated electron reactions. J. Phys. Chem. A 111(37), 9019-9024.

Miao, X. S., Bishay, F., Chen, M., Metcalfe, C. D. (2004) Occurrence of antimicrobials in the final effluents of wastewater treatment plants in Canada. Environ. Sci. Technol. 38, 3533-3541.

Minakata, D., Li, K., Westerhoff, P., Crittenden, J. (2009) Development of a group contribution method to predict aqueous phase hydroxyl radical (HO·) reaction rate constants. Environ. Sci. Technol. 43(16), 6220-6227.

Molina, E., Estrada, E., Nodarse, D., Torres, L. A., Gonzalez, H., Uriarte, E. (2008) Quantitative structure-antibacterial activity relationship modeling using a combination of piecewise linear regression-discriminant analysis (I): Quantum chemical, topographic, and topological descriptors. Int. J. Quantum Chem. 108(10), 1856-1871.

Monod, A. and Doussin, J. F. (2008) Structure-activity relationship for the estimation of OH-oxidation rate constants of aliphatic organic compounds in the aqueous phase: alkanes, alcohols, organic acids and bases. Atmos. Environ. 42(33), 7611-7622.

Monod, A., Poulain, L., Grubert, S., Voisin, D., Wortham, H. (2005) Kinetics of OH-initiated oxidation of oxygenated organic compounds in the aqueous phase: new rate constants, structure-activity relationships and atmospheric implications. Atmos. Environ. 39(40), 7667-7688.

Munoz, F. and von Sonntag, C. (2000) The reactions of ozone with tertiary amines including the complexing agents nitrilotriacetic acid (NTA) and ethylenediaminetetraacetic acid (EDTA) in aqueous solution. J. Chem. Soc. Perk. T. 2(10), 2029-2033.

Nakata, H., Kannan, K., Jones, P. D., Giesy, J. P. (2005) Determination of fluoroquinolone antibiotics in wastewater effluents by liquid chromatography-mass spectrometry and fluorescence detection. Chemosphere. 58, 759-766.

Nakada, N., Tanishima, T., Shinohara, H., Kiri, K., Takada, H. (2006) Pharmaceutical chemicals and endocrine disrupters in municipal wastewater in Tokyo and their removal during activated sludge treatment. Water Res. 40, 3297-3303.

Neter, J., Wasserman, W., Kutner, M. H. (1983) Applied Linear Regression Models. Irwin, Homewood, IL.

Netzeva, T.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.; Gramatica, P.; Jaworska, J.; Kahn, S.; Klopman, G.; Marchant, C.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D.; van de Sandt, J.; Tong, W.; Veith, G.; Yang, C. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. ATLA-Altern. Lab. Anim. 33, 155-173.

Neumann, M. B., Gujer, W., von Gunten, U. (2009) Global sensitivity analysis for model-based prediction of oxidative micropollutants transformation during drinking water treatment. Water Res. 43, 997-1004.

NIST (2002) NDRL/NIST solution kinetics database on the web, NIST standard reference database 40, a compilation of kinetics data on solution-phase reactions. http://kinetics.nist.gov/solution/index.php.

Oberg, T. (2005) A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application, and prediction. Atmos. Environ. 39(12), 2189-2200.

Olsson, I. M., Gottfries, J., Wold, S. (2004) D-optimal onion designs in statistical molecular design. Chemometr. Intellig. Lab. Syst. 73(1), 37-46.

Pablo Pocostales, J., Sein, M. M., Knolle, W., von Sonntag, C., Schmidt, T. C. (2010) Degradation of ozone-refractory organic phosphates in wastewater by ozone and ozone/hydrogen peroxide (peroxone): The role of ozone consumption by dissolved organic matter. Environ. Sci. Technol. 44(21), 8248-8253.

Papa, E., Fick, J., Lindberg, R., Johansson, M., Gramatica, P., Andersson, P. L. (2007) Multivariate chemical mapping of antibiotics and identification of structurally representative substances. Environ. Sci. Technol. 41(5), 1653-1661.

Park, J. S., Choi, H., Cho, J. (2004) Kinetic decomposition of ozone and para-chlorobenzoic acid (*p*CBA) during catalytic ozonation. Water Res. 38, 2285-2292.

Peller, J. R., Mezyk, S. P., Cooper, W. J. (2009) Bisphenol A reactions with hydroxyl radicals: Diverse pathways determined between deionized water and tertiary treated wastewater solutions. Res. Chem. Intermediat. 35(1), 21-34.

Pereira, V. J., Weinberg, H. S., Linden, K. G., Singer, P. C. (2007) UV degradation kinetics and modeling of pharmaceutical compounds in laboratory grade and surface water via direct and indirect photolysis at 254 nm. Environ. Sci. Technol. 41(5), 1682-1688.

Peres, J. A., Dominguez, J. R., Beltran-Heredia, J. (2010) Reaction of phenolic acids with Fenton-generated hydroxyl radicals: Hammett correlation. Desalination 252(1-3), 167-171.

Peter, A., von Gunten, U. (2007) Oxidation kinetics of selected taste and odor compounds during ozonation of drinking water. Environ. Sci. Technol. 41(2), 626-631.

Razavi, B., Song, W., Cooper, W. J., Greaves, J., Jeong, J. (2009) Free-radical-induced oxidative and reductive degradation of fibrate pharmaceuticals: Kinetic studies and degradation mechanisms. J. Phy. Chem. A 113(7), 1287-1294.

Regnery, J. and Puettmann, W. (2010) Occurrence and fate of organophosphorus flame retardants and plasticizers in urban and remote surface waters in Germany. Water Res. 44(14), 4097-4104.

Reid, R. C., Prausnitz, J. M., Sherwood, T. K. (1977) The properties of gases and liquids. McGraw-Hill Book Company, USA.

Ren, X., Lee, Y. J., Han, H. J., Kim, I. S. (2008) Effect of tris-(2-chloroethyl)-phosphate (TCEP) at environmental concentration on the levels of cell cycle regulatory protein expression in primary cultured rabbit renal proximal tubule cells. Chemosphere 74(1), 84-88.

Rey, R. P., Padron, A. S., Leon, L. G., Pozo, M. M., Baluja, C. (1999) Ozonation of cytostatics in water medium. Nitrogen bases. Ozone Sci. Eng. 21, 69-77.

Roche, P. and Prados, M. (1995) Removal of pesticides by use of ozone or hydrogen peroxide ozone. Ozone-Sci. Eng. 17(6), 657-672.

Rosenfeldt, E. J., Linden, K. G. (2004) Degradation of endocrine disrupting chemicals bisphenol A, ethinyl estradiol, and estradiol during UV photolysis and advanced oxidation processes. Environ. Sci. Technol. 38(20), 5476-5483.

Rosenfeldt, E. J. and Linden, K. G. (2007) The $R_{OH,UV}$ concept to characterize and the model UV/H_2O_2 process in natural waters. Environ. Sci. Technol. 41(7), 2548-2553.

Roy, K. and Roy, P. P. (2009) Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. Eur. J. Med. Chem. 44(7), 2913-2922.

Sanderson, R. T. (1983) Electronegativity and bond energy. J. Am. Chem. Soc. 105(8), 2259-2261.

Santoke, H., Song, W., Cooper, W. J., Greaves, J., Miller, G. E. (2009) Free-radical-induced oxidative and reductive degradation of fluoroquinolone pharmaceuticals: Kinetic studies and degradation mechanism. J. Phys. Chem. A. 113(27), 7846-7851.

Shemer, H., Sharpless, C. M., Elovitz, M. S., Linden, K. G. (2006) Relative rate constants of contaminant candidate list pesticides with hydroxyl radicals. Environ. Sci. Technol. 40(14), 4460-4466.

Snyder, S. A. (2008) Occurrence, treatment, and toxicological relevance of EDCs and pharmaceuticals in water. Ozone Sci. Eng. 30(1), 65-69.

Snyder, S. A., Adham, S., Redding, A. M., Cannon, F. S., DeCarolis, J., Oppenheimer, J., Wert, E. C., Yoon, Y. (2007) Role of membranes and activated carbon in the removal of endocrine disruptors and pharmaceuticals. Desalination 202(1-3), 156-181.

Song, W., Chen, W., Cooper, W. J., Greaves, J., Miller, G. E. (2008a) Free-radical destruction of beta-lactam antibiotics in aqueous solution. J. Phys. Chem. A. 112(32), 7411-7417.

Song, W., Cooper, W. J., Mezyk, S. P., Greaves, J., Peake, B. M. (2008b) Free radical destruction of beta-blockers in aqueous solution. Environ. Sci. Technol. 42(4), 1256-1261.

Song, W., Cooper, W. J., Peake, B. M., Mezyk, S. P., Nickelsen, M. G., O'Shea, K. E. (2009) Free-radical-induced oxidative and reductive degradation of N,N'-diethyl-m-toluamide (DEET): Kinetic studies and degradation pathway. Water Res. 43(3), 635-642.

Spycher, S., Nendza, M., Gasteiger, J. (2004) Comparison of different classification methods applied to a mode of toxic action data set. QSAR Comb. Sci. 23(9), 779-791.

Stackelberg, P. E., Furlong, E. T., Meyer, M. T., Zaugg, S. D., Henderson, A. K., Reissman, D. B. (2004) Persistence of pharmaceutical compounds and other organic wastewater contaminants in a conventional drinking-water treatment plant. Sci. Total Environ. 329(1-3), 99-113.

Suarez, S., Dodd, M. C., Omil, F., von Gunten, U. (2007) Kinetics of triclosan oxidation by aqueous ozone and consequent loss of antibacterial activity: Relevance to municipal wastewater ozonation. Water Res. 41(12), 2481-2490.

Ternes, T. A., Hirsch, R. (2000) Occurrence and behavior of X-ray contrast media in sewage facilities and the aquatic environment. Environ. Sci. Technol. 34, 2741-2748.

Ternes, T. A. (1998) Occurrence of drugs in German sewage treatment plants and rivers. Water Res. 32(11), 3245-3260.

Ternes, T. A., Stuber, J., Herrmann, N., McDowell, D., Ried, A., Kampmann, M., Teiser, B. (2003) Ozonation: a tool for removal of pharmaceuticals, contrast media and musk fragrances from wastewater? Water Res. 37(8), 1976-1982.

Ternes, T. A., Meisenheimer, M., McDowell, D., Sacher, F., Brauch, H. J., Gulde, B. H., Preuss, G., Wilme, U., Seibert, N. Z. (2002) Removal of pharmaceuticals during drinking water treatment. Environ. Sci. Technol. 36(17), 3855-3863.

Thorpe, K. L., Cummings, R. I., Hutchinson, T. H., Scholze, M., Brighty, G., Sumpter, J. P., Tyler, C. R. (2003) Relative potencies and combination effects of steroidal estrogens in fish. Environ. Sci. Technol. 37(6), 1142-1149.

Todeschini, R., Consonni, V., Mauri, A., Pavan, M. (2005) The remote version of the well known DRAGON software: E-DRAGON. http://michem.disat.unimib.it/chm/Help/edragon/index.html. Accessed on December 2011.

Todeschini, R. and Consonni, V. (2000) Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany.

Topliss, J. G. and Edwards, R. P. (1979) Chance factors in studies of quantitative structure-activity-relationships. J. Med. Chem. 22, 1238-1244.

Tropsha, A., Gramatica, P., Gombar, V. K. (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22(1), 69-77.

van der Bruggen, B., Schaep, J., Wilms, D., Vandecasteele, C. (1999) Influence of molecular size, polarity and charge on the retention of organic molecules by nanofiltration. J. Membr. Sci. 156(1), 29-41.

Vethaak, A. D., Lahr, J., Schrap, S. M., Belfroid, A. C., Rijs, G. B.J., Gerritsen, A., de Boer, J., Bulder, A. S., Grinwis, G. C. M., Kuiper, R. V., Legler, J., Murk, T. A. J., Peijnenburg, W., Verhaar, H. J. M., de Voogt, P. (2005) An integrated assessment of estrogenic contamination and biological effects in the aquatic environment of The Netherlands. Chemosphere. 59, 511-524.

Vincent, S., Kotbi, A., Barbeau, B. (2010) Predicting hydroxyl radical activity and trace contaminants removal in ozonated water. Ozone Sci. Eng. 32, 244-251.

von Gunten, U. (2003) Ozonation of drinking water: Part I. Oxidation kinetics and product formation. Water Res. 37, 1443-1467.

Wang, Y., Chen, J., Li, X., Zhang, S., Qiao, X. (2009) Estimation of aqueous-phase reaction rate constants of hydroxyl radical with phenols, alkanes and alcohols. QSAR Comb. Sci. 28(11-12), 1309-1316.

Westerhoff, P., Yoon, Y., Snyder, S., Wert, E. (2005) Fate of endocrine-disruptor, pharmaceutical, and personal care product chemicals during simulated drinking water treatment processes. Environ. Sci. Technol. 39(17), 6649-6663.

Watts, M. J., Linden, K. G. (2008) Photooxidation and subsequent biodegradability of recalcitrant tri-alkyl phosphates TCEP and TBP in water. Water Res. 42(20), 4949-4954.

Watts, M. J. and Linden, K. G. (2009) Advanced oxidation kinetics of aqueous trialkyl phosphate flame retardants and plasticizers. Environ. Sci. Technol. 43(8), 2937-2942.

Winkler, M., Kopf, G., Hauptvogel, C., Neu, T. (1998) Fate of artificial musk fragrances associated with suspended particulate matter (SPM) from the River Elbe (Germany) in comparison to other organic contaminants. Chemosphere. 37, 1139-1156.

Wold, S., Josefson, M., Gottfries, J., Linusson, A. (2004) The utility of multivariate design in PLS modeling. J. Chemometr. 18(3-4), 156-165.

Wold, S., Sjostrom, M., Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. Chemometr. Intellig. Lab. Syst. 58(2), 109-130.

Wold, S., Esbensen, K., Geladi, P. (1987) Principal component analysis. Chemometr. Intellig. Lab. Syst. 2(1-3), 37-52.

Worth, A. P. and Cronin, M. T. D. (2003) The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. J. Mol. Struc. –Theochem. 622, 97-111.

Wu, C., Shemer, H., Linden, K. G. (2007) Photodegradation of metolachlor applying UV and UV/H_2O_2. J. Agric. Food Chem. 55(10), 4059-4065.

Xu, L. and Zhang, W. J. (2001) Comparison of different methods for variable selection. Anal. Chim. Acta 446(1-2), 477-483.

Xue, N. D., Xu, X. B. (2006) Composition, distribution, and characterization of suspected endocrine-disrupting pesticides in Beijing GuanTing Reservoir (GTR). Arch. Environ. Contam. Toxicol. 50, 463-473.

Yang, L., Wang, P., Jiang, Y., Chen, J. (2005) Studying the explanatory capacity of artificial neural networks for understanding environmental chemical quantitative structure-activity relationship models. J. Chem. Inf. Model. 45, 1804-1811.

Yangali-Quintanilla, V., Sadmani, A., McConville, M., Kennedy, M., Amy, G. (2010) A QSAR model for predicting rejection of emerging contaminants (pharmaceuticals, endocrine disruptors) by nanofiltration membranes. Water Res. 44(2), 373-384.

Yangali-Quintanilla, V., Verliefde, A., Kim, T., Sadmani, A., Kennedy, M., Amy, G. (2009) Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes. J. Membr. Sci. 342(1-2), 251-262.

Yao, C. C. D. and Haag, W. R. (1991) Rate constants for direct reactions of ozone with several drinking-water contaminants. Water Res. 25(7), 761-773.

Yasojima, M., Nakada, N., Komori, K., Suzuki, Y., Tanaka, H. (2006) Occurrence of levofloxacin, clarithromycin and azithromycin in wastewater treatment plant in Japan. Water Sci. Technol. 53, 227-233.

Yu, Z., Peldszus, S., Huck, P. M. (2008) Adsorption characteristics of selected pharmaceuticals and an endocrine disrupting compound - naproxen, carbamazepine and nonylphenol - on activated carbon. Water Res. 42(12), 2873-2882.

Yu, Z., Peldszus, S., Huck, P. M. (2009a) Adsorption of selected pharmaceuticals and an endocrine disrupting compound by granular activated carbon. 1. Adsorption capacity and kinetics. Environ. Sci. Technol. 43(5), 1467-1473.

Yu, Z., Peldszus, S., Huck, P. M. (2009b) Adsorption of selected pharmaceuticals and an endocrine disrupting compound by granular activated carbon. 2. Model prediction. Environ. Sci. Technol. 43(5), 1474-1479.

Zhan, C. G., Nichols, J. A., Dixon, D. A. (2003) Ionization potential, electron affinity, electronegativity, hardness, and electron excitation energy: Molecular properties from density functional theory orbital energies. J. Phys. Chem. A. 107, 4184-4195.

Zimbron, J. A. and Reardon, K. F. (2005) Hydroxyl free radical reactivity toward aqueous chlorinated phenols. Water Res. 39(5), 865-869.