Molecular Similarity Analysis as Tool to Predict Environmental Properties and Prioritize Research among Emerging Contaminants in the Environment

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Chenxi Li

May

2013

APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Chenxi Li

AUTHOR

The dissertation has been read and approved by the examining committee:

Lisa M. Colosi

Advisor

Roseanne Ford

Teresa B. Culver

James Smith

Wu-seng Lung

Accepted for the School of Engineering and Applied Science:

James H. Ayl

Dean, School of Engineering and Applied Science

May

2013

Molecular Similarity Analysis as Tool to Predict Environmental Properties and Prioritize Research among Emerging Contaminants in the Environment

A Dissertation

Presented to the

Faculty of the School of Engineering and Applied Science at the

University of Virginia

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in

Civil and Environmental Engineering

By

Chenxi Li

April 2013

APPROVAL SHEET

The Dissertation is Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy in Civil and Environmental Engineering

____

Chenxi Li (Author)

This dissertation has been read and approved by the examining Committee:

____

Lisa M. Colosi, Ph.D. (Advisor)

____

Teresa B. Culver, Ph.D. (Committee Chair)

____

James A. Smith, Ph.D. (Committee)

____

Roseanne M. Ford, Ph.D. (Committee)

____

Wu-Seng Lung, Ph.D. (Committee)

Accepted for the School of Engineering and Applied Science:

____

Dean, School of Engineering and Applied Science

April 2013

Abstract

The very large number of emerging contaminants entering the water supply makes it desirable to assess their environmental fate and behavior without direct measurements. Structure-property prediction models are promising in this regard. Although traditional, regression-based structure-property relationships have been proven to be accurate for prediction of some parameters, these regression models are either too narrowly-defined to be of practical benefit for more than one small class of chemicals, or so broad that data scarcity compromises their predictive accuracy. Quantitative molecular similarity assessment (QMSA) is one particularly appealing alternative to traditional, regression-based models. QMSA models are based on the assumption that "similar" molecules behave "similarly", such that parameters of interest for a target chemical can be computed based on parameter values for structurally similar chemicals. This is an evolving technique, which tends to be particularly appealing for applications in which predictive accuracy may be hampered by limited data availability.

This research has three main objectives: 1) demonstrating that QMSA models can accurately predict environmental parameters of interest for highly diverse chemical classes, then measuring fundamental fate and transport parameters for several key emerging contaminants, to externally validate QMSA hypotheses; 2) applying measured fate parameters to predict the fate of emerging contaminants in WWTPs; and 3) assessing the extent to which QMSA can help prioritize among unmeasured chemicals and determine which additional measurements will result in maximally increased model accuracy.

The results of this research are promising. Virtual experiments showed that valid

QMSA models have been created for accurate prediction of three distinct environmental parameters: *in vitro* estrogenicity, sorption distribution coefficient K_d, and pseudo first-order biodegradation rate constant k_b. Laboratory experiments for this research have focused on the measurement of K_d and k_b for three highly-prescribed pharmaceutical (i.e., metformin, fluconazole, and benazepril), in wastewater obtained from a municipal wastewater treatment plant. Measured K_d and k_b values are consistent with QMSA model predictions; furthermore, incorporation of these two parameters into simple mass balance models accurately predicts the effluent concentrations of studied emerging contaminants in WWTPs, demonstrating the usefulness of both QMSA and simple mass balance models. Finally we showed that three proposed QMSA-based prioritization approaches affords better improvement in QMSA estimation of property values among the remaining unmeasured compounds than random selection. This criterion provides additional information on selection of unmeasured emerging contaminants in terms of improving QMSA models' accuracy.

Acknowledgements

Here I would like take a moment to express my sincere gratitude to everyone has helped me during five years.

First I would like to thank my family, my dad, Songjie, my mom, Xiulan, and my wife, Yun. Although cannot be with you for most of time within these five years, I can always feel supports from you. Without supports from you, I will never have a chance to come abroad and finish my PhD degree.

I would like to thank my advisors, Professor Lisa Colosi for everything that you have done to help me on my PhD academics. Without your consistent guidance and motivation, I will never be able to get my PhD. I greatly appreciate your innovative directions on my research and always positive encouragement to me whenever I met problems. The research skills I've learned from you will definitely be a great asset for me in my future career.

I would also like to thank my other committee members, Professor Teresa Culver, Professor Roseanne Ford, Professor James Smith, and Professor Wu-Seng Lung for taking your precious time to help with my PhD academics. Your guidance and suggestions have greatly improved the quality of my PhD research.

I would also like to thank all of my friends and fellow graduates here in UVa. It was great experience to work with you. You guys make my life easier here at UVa. I would like to thank Charles, Karl, Yongli, Shane, Shee, Xiaowei, Dong, Shibo, Brian, Zhiyuan, Dianjun, Bo, Shanshan, Carly, Lydia, Han, R.D. and John. Thank you very much!

Finally, I would also like to thanks to all the others who have helped me on my PhD academics: Mr. James Danberg, Administration at Department of Civil and Environmental Engineering for your help on my experimental design, Mr. Gary Philips, Wastewater Manager at Rivanna Water and Sewer Authority for your arrangement for wastewater sampling and Mr. Paul Macek, Technical Support Specialist at Shimadzu Scientific Instruments for you help on setting up HPLC and LC-MS.

Table of Contents

Chapter 1 – Introduction

1.1 Definition, sources, and adverse effects of emerging contaminants

The categorization "emerging contaminants" refers to a large number of unregulated contaminants in the environment which are either recently developed or are recently detected because of advancements of analytical instrument (Richardson and Ternes, 2011; Petrović and Barcelo, 2006). There are a wide range of chemicals which can be categorized as emerging contaminants. They mainly include: endocrine disrupting chemicals (EDCs), hormones, pharmaceuticals and personal care products (PPCPs), pesticides, veterinary products, surfactants compounds, plasticizers, various industrial additives, food additives, and engineered nano-materials (Lapworth et al., 2012; Richardson and Ternes, 2005; Petrović et al., 2003).

Emerging contaminants enter into environment via a number of pathways. The major pathway is discharge of wastewater effluents from municipal wastewater treatment plants (WWTPs) (Heberer et al., 2004; Koplin et al., 2002). Emerging contaminants are discharged from domestic wastewaters or hospital wastewaters, entering into WWTPs, and are eventually released to natural water bodies. They can also enter into groundwater through septic tank systems (Swartz et al., 2006) or landfill systems (Holm et al., 1995). Finally, they can enter into soils via land-application of sludge from WWTPs as fertilizers (Ternes et al. 2004) or livestock manure (Watanabe et al., 2010).

Emerging contaminants have received increasing attention since they may cause adverse effects on human or aquatic species, even when they present at very low concentrations. For example, EDCs are found to be a group of chemicals which can disrupt the normal functioning of the endocrine system resulting in: reproductive abnormalities (Fry and Toone, 1981; Fry et al., 1987), population declines and reproductive disorders (Gibbs et al., 1991), feminization (Jobling et al., 1998), and more significant for humans, decreases in male sperm count and increases in a variety of cancers and reproductive malfunctions (Michael, 2001). Antibiotics, a sub-group of PPCPs, also show ecotoxicity to aquatic organisms (Sanderson et al., 2004) and contribute to development antibiotic-resistant bacteria in aquatic systems (Gilliver et al., 1999; Smith et al., 1999). The polar emerging contaminants are highly soluble in water and therefore are difficult to removed by typical WWTP processes (Knepper et al., 1999). This group of chemicals includes acidic pharmaceuticals, acidic pesticides, and acidic metabolites of non-ionic surfactants. Perhaps the most challenging characteristic of emerging contaminants is that they will cause continuous exposure for humans and animals due to their continuous usages and introduction into environment. Therefore their adverse effects, if exist, will never be stopped (Petrović et al., 2003).

1.2 WWTPs and their functions to remove emerging contaminants

The importance of WWTPs for removing emerging contaminants has been well documented in the scientific literature (Bolong et al., 2009; Petrović et al., 2003). Existing WWTPs are generally designed to regulate traditional environmental parameters, such as five-day biochemical oxygen demand (BOD), the chemical oxygen demand (COD), nitrogen/phosphorus concentration, and suspended solids (SS) (Carballa et al., 2004 and Salgado et al., 2012). In contrast, they were not specifically designed to remove low-concentration organic contaminants. Thus, they may vary widely in their ability to efficiently remove emerging contaminants when they are operated as originally designed (Bolong et al., 2009).

In WWTPs, there are three significant mechanisms responsible for removal of emerging contaminants, i.e. volatilization, sorption, and biodegradation (Pomiès et al., 2013; Joss et al., 2006). These mechanisms are discussed in more detail in the following paragraphs.

1.2.1 Volatilization

Removal of emerging contaminants via volatilization is highly dependent on their Henry's law constants and WWTP operating conditions, such as extent of aeration, agitation, atmospheric pressure, and temperature. Volatilization can be described by two distinct processes, i.e. stripping and surface volatilization (Pomiès et al., 2013). Stripping is the any water movement process in WWTPs during which chemicals are removed from wastewater and released into atmosphere via head loss or air bubbles, while surface volatilization is defined as release of chemicals from wastewater into quiescent or wind driven processes (Mihelcic et al., 1993). In WWTPs, stripping can be described as the following equation (Cowan et al., 1993):

$$\frac{dC_t}{dt} = -\frac{Q_{air} \times H \times C_t}{V \times R \times T} \tag{1.1}$$

where C_t is dissolved emerging contaminant concentration at time t (μg/L); t is the time (h); Q_{air} is the air flow rate (L/h); H is the Henry's Law constant (L·Pa/mol); R is the gas constant (L·Pa/(K·mol)); V is the volume (L); T is the temperature (K).

However, previously published literature indicates that volatilization is not a significant removal process for emerging contaminants in WWTPs. For example, Lee et al. (1998) concluded that stripping was important relative to total removal in WWTPs only if $H^{'}$ was higher than 0.8 and only if the chemicals are not biodegradable. The EU project POSEIDON concluded that musk fragrances with $H^{'}$ greater than 0.005 are slightly volatile (Poseidon, 2005). Results from Byrns (2001) suggested volatilization was not significant for chemicals with H lower than 0.1 Pa·L/mol after he studied a large number of VOCs, PAHs and pesticides. For most of emerging contaminants, including PPCPs and EDCs, their dimensionless Henry's Law constants, $H^{'}$ are usually less than 0.005 (Schwarzenbach et al., 2003), making their potential volatilization negligible.

1.2.2 Sorption

Sorption is believed to be the most significant removal process for emerging contaminants in WWTPs. It is because sorption can happen in any part of a WWTP where suspended solids are present For some persistent emerging contaminants, sorption could be the single most important removal mechanism (Simonich et al., 2002). The sorption process can be written as the following equilibrium equation:

$$\frac{dC_t}{dt} = -k_{sor} \times C_t \times X_{ss} + k_{des} \times C_{s,t} \tag{1.2}$$

where k_{sor} is the sorption kinetic constant (L/(g·h)); k_{des} is the desorption kinetic constant (h^{-1}); X_{ss} is the suspended solid concentration (g/L); $C_{s,t}$ is the sorbed emerging contaminant concentration in wastewater (μg/L).

However, in most scenarios, sorption can be assumed to reach equilibrium instantaneously because sorption processes are much faster the biodegradation (Parker et al., 1994). Therefore a simplified parameter, K_d, the sorption distribution coefficient, is widely used to describe the WWTP sorption (Ternes et al. 2004). It can be expressed as the following equation:

$$K_d = \frac{k_{des}}{k_{sor}} = \frac{C_t \times X_{ss}}{C_{s,t}} \tag{1.3}$$

This parameter encapsulates the distribution of a chemical between aqueous and suspended solid phase. K_d has been well demonstrated as one of the most significant parameters responsible for chemicals' removal since it can be used to accurately predict the sorption process in WWTPs (Schwarzenbach, 2003). However some other parameters which have indirect implications with sorption; for example, octanol–water distribution coefficient (K_{ow}), but this parameter may not accurately quantify sorption for emerging contaminants. This is because specific electrostatic interactions exist for sorption of certain pharmaceuticals (fluorochinolones) which have very high K_d but extremely low K_{ow} (Golet et al., 2003). This has also been reflected on the increasing research interests on measurement of K_d in WWTPs in recent years (Golet et al. 2003; Artola-Garicano et al. 2003; Clara et al. 2005; Radjenovic et al. 2009; Barron et al. 2009; Ottmar et al. 2010; Stevens-Garmon et al. 2011).

1.2.3 Biodegradation

Biodegradation comprises a series of biological kinetic processes controlled by WWTP microorganisms (such as bacteria and fungi), during which emerging contaminants are partially removed or completely eliminated from the aqueous phase (EUR 20418 EN/2). Biodegradation can be described by two different types of models, i.e. a pseudo first-order type (Dionisi et al., 2008; Byrns, 2001) or a Monod type model (Plosz et al., 2010). The pseudo first-order type can be written as the following equation:

$$\frac{dC_t}{dt} = -k_b \times C_t \times X_{ss} \tag{1.4}$$

where k_b is the pseudo first-order biodegradation rate constant (L/(g·h)).

The Monod type model comprises the following form:

$$\frac{dC_t}{dt} = -\frac{1}{Y} \times \mu_{max} \times \frac{C_{o,t}}{C_{o,t}+K_o} \times \frac{C_t}{C_t+K} \times X_{ss} \tag{1.5}$$

where Y is the conversion yield; μ_{max} is the bacteria maximum growth rate (h^{-1}); $C_{o,t}$ is the oxygen concentration at time t (mg/L); K_o is the oxygen half saturation coefficient (mg/L); K is the emeriging contaminant half saturation coefficient (µg/L).

In literatures, the pseudo first-order biodegradation rate constant (k_b) has been widely and conveniently used as the parameter for prediction of biological transformation of emerging contaminants in different WWTPs (Salgado et al., 2012; Thompson et al., 2011; Joss at al., 2006; Monteith et al., 1995; Melcer et al., 1994; Cowan et al., 1993). Therefore in this study, k_b was selected as the parameter to quantify biodegradation efficiency.

1.3 Quantitative Structure-Property Relationship (QSPR) for environmental property estimation

It has been well documented that emerging contaminants are widely detected in the US drinking water supply (USEPA, 2006), and there are over 8,400,000 different commercially available compounds worldwide (Muir and Howard, 2006). Since a significant fraction of these compounds are unregulated and may cause unknown or adverse impacts when ingested by humans or animals, the US Toxic Substances Control Act (TSCA) Inventory currently contains more than 82,000 chemical compounds (Muir and Howard, 2006). Not surprisingly, the amount of work required to obtain essential environmental modeling parameters for all of these chemicals can overwhelm the capacity of regulatory agencies. In response to a similar crisis overseas, the European Union utilized the extensive Registration Evaluation, and Authorization of CHemicals (REACH) Program to evaluate the toxicity of some 20,000–30,000 chemicals (EC, 2006). One shared component of the REACH and TSCA directives holds significant emphasis on obtaining essential environmental information of emerging contaminants via validated quantitative structure-property relationships (QSPRs) to predict environmental fate and behavior for classes of chemicals rather than measurement of individual compounds (Saliner et al, 2005; Basak et al, 2002).

Traditional QSPRs are widely applied for prediction of physicochemical properties, biological activities, or toxicity for a class of chemicals on the basis of molecular descriptors (Basak et al, 2002). These type of models are usually created using multiple linear regressions. The target environmental properties are predicted by linear combinations of all possible independent variables, i.e. descriptors (Yangali-Quintanilla et al., 2010). Although a well-validated structure-property model can significantly reduce the time and expense required to screen various chemicals of regulatory interest, and such models have been used increasingly over the last several decades to fill in data needs related to ecological effects (Cronin et al, 2003), a well-defined application domain is usually essential to define the applicability of QSPRs (Tropsha et al., 2003).

Another common criticism of traditional, regression-based structure-property relationships is related to their scope. In general, these models are either too narrowly-defined to be of practical benefit for more than one small class of chemicals (Cronin et al, 2003) or so broad that data scarcity compromises their predictive accuracy (Basak et al, 2002). Further, when larger quantities of data are available, regression-based QSPR models must be re-built based to incorporate new information. These and other criticisms of regression-based structure-property relationships have led to interest in alternative structure-property modeling.

Quantitative molecular similarity analysis (QMSA) is one promising alternative to traditional QSPRs for addressing the shortcomings referenced above. The fundamental hypothesis of QMSA is that “similar” chemicals will exhibit similar environmental properties. Therefore, rather than developing a series of regression-based QSPRs, the QMSA approach focuses on finding what compounds are most similar to the target chemical for which property values are desired. Regression-based QSPRs usually split compounds into small, discrete classes exhibiting good correlation with selected molecular descriptors. A series of linear regressions is then used to fit the properties to these molecular descriptors. However, QMSA seeks to maximize data utility by pooling all available measured compounds into a single dataset. A vast quantity of molecular information comprising empirically measured or theoretically calculated descriptors is then assembled for all of the chemicals in the pool. This information is culled through various processes, and the several remaining critical similarity features are incorporated into a similarity assessment algorithm to determine which *k* compounds within the training dataset are most similar to the target chemical. Ultimately, the target's output property is estimated from some linear combination of the measured property values associated with its *k* nearest (i.e., most chemically similar) neighbors (*k*-NN) (Basak and Grunwald, 1995; Basak et al, 2002; Gute et al, 2004).

To date, QMSAs models have been used to accurately predict various parameters related to environmental risk assessment, including: mutagenicity (Basak and Grunwald, 1995), hepatoxicity (Gute et al, 2004), and skin sensitization plus various other REACH-relevant toxicity endpoints (Estrada et al, 2004). Still, QMSA models have not been widely used to predict environmental fate and behavior parameters, and it is unknown how much additional data is needed to create valid QMSA models for emerging contaminants.

As discussed before, WWTPs serve as the most significant environmental end point for removal of emerging contaminants. And sorption as well as biodegradation will be the two most important processes responsible for removal of emerging contaminants in WWTPs. Therefore K_d and k_b are crucially important parameters for understanding the fate of emerging contaminants in WWTPs. Although there are some efforts to apply QSPR for estimation of K_d and k_b, these studies have used either a machine learning technique (e.g., artificial neural network), which tends to be a “black box” for the user and makes it very difficult for them to understand the model structure, or they didn’t perform no comprehensive model validation steps to ensure those models are accurate (Dickenson et al., 2010).

1.4 Hypothesis and Objectives

This study will evaluate three main hypotheses, each with several sub-hypotheses and objectives, as noted below. The results from part of Hypothesis 1 have already been published in *Separation and Purification Technology*. The results from of Hypothesis 3 have been accepted for publication in *SAR and QSAR in Environmental Research*. It is also proposed that the experimental results from Hypothesis 1 (as pertaining to measuring and predicting the sorption distribution coefficient, K_d and pseudo first-order biodegradation rate constant, k_b of three highly prescribed pharmaceuticals to wastewater sludge), will yield another two papers. These two manuscripts will be targeted for publication in a traditional environmental engineering journals.

Hypothesis 1: QMSA models can be used to accurately predict environmental engineering parameters of interest.

Objective 1: Establish QMSA models for three environmental parameters of interest (*in vitro* estrogenicity, sorption distribution coefficient - K_d, and pseudo first-order biodegradation rate constant - k_b) and validate model predictions using leave-one-out (LOO) cross-validation procedures; compute the cross-validation coefficients (q^2) and compare predictions with literature values or laboratory measurements.

Hypothesis 2: K_d and k_b of selected emerging contaminants can be useful for estimation their effluent concentrations out from WWTPs

Objective 2: Use measured K_d and k_b values of selected compounds in Hypothesis 1 and develop a simple mass balance model to predict the effluent concentrations and compare them with measured effluent concentrations reported in literatures.

Hypothesis 3: Prioritizations evaluated by QMSA are critical factors for prioritizing among unmeasured chemicals and determining which additional measurements will result in maximally increased model accuracy.

Hypothesis 3a: Addition of chemicals exhibiting high representativeness (as parameterized using summation of intermolecular distances among all unmeasured chemicals) to the pool of “measured” data will increase the accuracy of QMSA models. In contrast, addition of chemicals exhibiting low representativeness to the pool of “measured” data will decrease the QMSA model accuracy.

Objective 3a: Compare the predictive abilities of QMSA models incorporating “new” measurements as selected using a similarity-based “high representativeness” or “low representativeness” approach or a random (arbitrary) selection approach.

Hypothesis 3b: Addition of chemicals exhibiting low redundancy (as parameterized using summation of intermolecular distances between all measured chemicals and each chemical in the unmeasured pool) to the pool of “measured” data will increase the accuracy of QMSA models.

Objective 3b: Compare the predictive abilities of QMSA models incorporating “new” measurements as selected using a similarity-based “redundancy” approach or a random (arbitrary) selection approach.

Hypothesis 3c: Addition of chemicals exhibiting both high representativeness and low redundancy to the pool of "measured" data will maximally increase the accuracy of QMSA models.

Objective 3c: Compare the predictive abilities of QMSA models incorporating "new" measurements as selected using an "intersection" approach, which incorporates consideration of both representativeness and redundancy approach, or other selection approaches evaluated in this study.

Figure 1.1 illustrates the general flow of the six chapters in this dissertation.

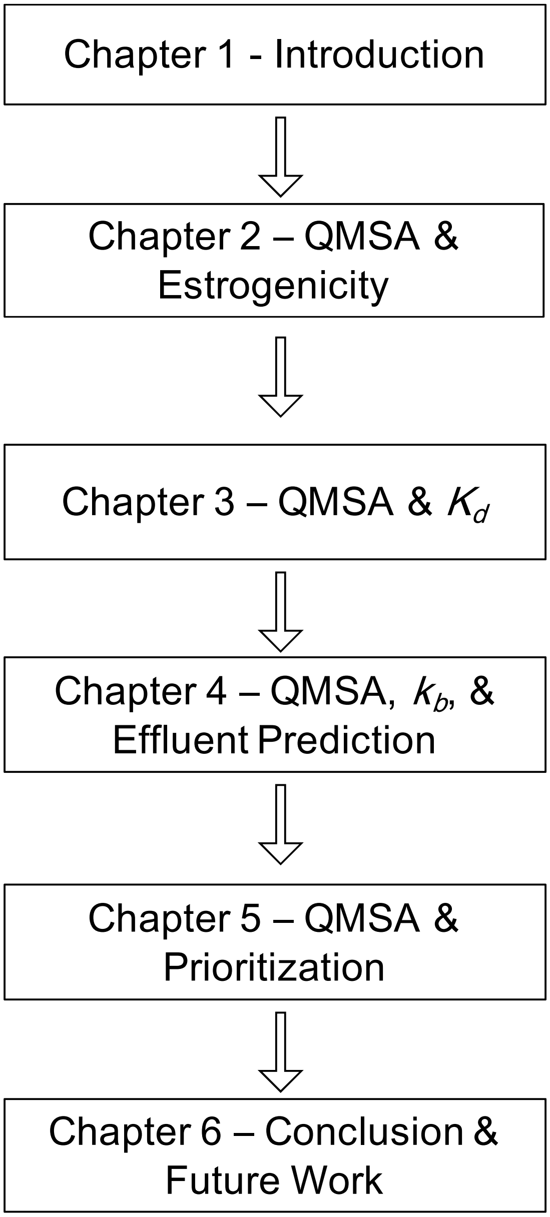
Chapter 1 - Introduction
Chapter 2 – QMSA & Estrogenicity
Chapter 3 – QMSA & K_d
Chapter 4 – QMSA, k_b, & Effluent Prediction
Chapter 5 – QMSA & Prioritization
Chapter 6 – Conclusion & Future Work

Figure 1.1 Flowchart of research focuses of six chapters in this study.

Chapter 1 covers the basic information and motivation on why we perform this research. In Chapter 2, QMSA will be first discussed and examined on its ability to predict one basis environmental engineering parameter, i.e. *in vitro* estrogenicity. The idea of this chapter is to demonstrate that QMSA can be a useful tool for environmental predictions since previously QMSA has not been applied in this area. Chapter 2 will only perform internal validation, i.e. examine QMSA's ability to accurately recreate property estimates for previously measured compounds, using the estrogenicity dataset. In Chapter 3, QMSA will be comprehensively applied to a K_d dataset, with application of both internal and external validation. External validation will require us to measure K_d of some chemicals which have never been measured before. Finally, QMSA's application domain will also be discussed to tackle our predictions to a reasonable range. Moving forward, Chapter 4 will examine if QMSA can be used to predict a more complicated fate parameter, k_b. Again both internal and external validation will performed. After that, K_d and k_b measurements for three selected emerging contaminants will be incorporated into a simple mass balance model that predicts WWTP effluent concentrations. This will be our final goal because it represents the best understanding of emerging contaminant fate in WWTPs. Chapter 5 will discuss more conceptual benefit of QMSA related to prioritization of unmeasured chemicals, In particular, we will try to show that QMSA procedures can be used to determine which chemicals should be measured first to best improve prediction accuracy for all other unmeasured chemicals. Finally, in Chapter 6, the main conclusions from this study will be summarized.

1.5 References

Artola-Garicano, I., Borkent, J.L.M., Hermens, W., Vaes H.J. (2003) Removal of two polycyclic musks in sewage treatment plants: freely dissolved and total concentrations. *Environ. Sci. Technol.* 37(14) , 3111–3116.

Barron, L., Havel, J., Purcell, M., Szpak, M., Kelleher, B., and Paull, B. (2009) Predicting sorption of pharmaceuticals and personal care products onto soil and digested sludge using artificial neural networks. *Analyst,* 134, 663-670.

Basak, S.C., Gute, B.D., Mills, D. (2002) Quantitative molecular similarity analysis (QMSA) methods for property estimation: a comparison of property-based, arbitrary, and tailored similarity spaces. *SAR QSAR Environ. Res.* 13, 727–742.

Basak, S.C., Grunwald, G.D. (1995) Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity-based study. *Chemosphere.* 31, 2592-2546.

Bolong, N., Ismail, A.F., Salim, M.R., Matsuura T. (2009) A review of the effects of emerging contaminants in wastewaters and options for their removal. *Desalination.* 239, 229–246.

Byrns G. (2001) The fate of xenobiotic organic compounds in wastewater treatment plants. *Water Res.* 35(10), 2523–2533.

Carballa, M., Omil, F., Lema, J.M., Llompart, M., Garcia-Jares, C., Rodriguez I., et al. (2004) Behavior of pharmaceuticals, cosmetics and hormones in a sewage treatment plant. *Water Res*, 38 (12), 2918–2926.

Clara, M., Kreuzinger, N., Strenn, B., Gans, O., Kroiss H. (2005) The solids retention time-a suitable design parameter to evaluate the capacity of wastewater treatment plants to remove micropollutants. *Water Res.* 39, 97–106.

Cowan, C.E., Caprara, R.J., White, C.E., Merves, M.L., Gullotti, M.J. (1993) A model for predicting the fate of 'down-the-drain' consumer product ingredients in United States rivers. *Water Environment Federation - 66th annual conference & exposition.* 351–358.

Cronin, M.T.D., Waler, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D., Worth, A.P. (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemicals, *Environ. Health Perspect.* 111, 1376–1390.

Dickenson, E., Drewes, J., Stevens-Garmon, J., Khan, S., McDonald J. (2010) Evaluation of QSPR Techniques for Wastewater Treatment Processes. *Water Environment Research Foundation*, Alexandria, VA.

Dionisi, D., Bornoromi, L., Mainelli, S., Majone, M., Pagnanelli, F., Papini, M.P. (2008) Theoretical and experimental analysis of the role of sludge age on the removal of adsorbed micropollutants in activated sludge processes. *Industrial Engineering Chemical Research.* 47, 6775–6782.

EUR 20418 EN/2. Institute for Health and Consumer Protection, European Chemicals Bureau, European Commission. EUR 20418 EN/2. (2003) Technical Guidance Document on Risk Assessment Part II.

European Commission (EC) (2006) Registration, evaluation, and authorization of chemicals (REACH) program. EC No. 1907/2006.

Fry, D.M., Toone, C.K., Speich, S.M., Peard R.J. (1987) Sex ratio skew and breeding patterns of gulls: demographic and toxicological considerations. *Stud. Avian. Biol.* 10, 26–43.

Fry, D.M., Toone. C.K. (1981) DDT-induced feminization of gull embryos. *Science*. 213, 922–924.

Gibbs, P.E., Pascoe, P.L., Bryan G.W. (1991) Tributyltin-induced imposex in stenoglossan gastropods: pathological effects on the female reproductive system. *Comp. Biochem. Physiol.* 100C, 231–235.

Gilliver, M. A., Bennett, M., Begon, M., Hazel, S., M., Hart, C. A. (1999) Enterobacteria: Antibiotic resistance found in wild rodents . *Nature.* 401, 233−234.

Golet, E.M., Xifra, I., Siegrist, H., Alder, A.C., Giger, W. (2003) Environmental exposure assessment of fluoroquinolone antibacterial agents from sewage to soil. *Environ. Sci. Technol.* 37(15), 3243–3249.

Gute, B.D., Balasubramanian, K., Giess, K.T., Basak, S.C. (2004) Prediction of halocarbon toxicity from structure:ahierarchicalQSARapproach. *Environ.Toxicol.Pharm.* 16, 121–129.

Heberer, T., Mechlinski, A., Fanck, B., Knappe, A., Massmann, G., Pekdeger, A., Fritz B. (2004) Field studies on the fate and transport of pharmaceutical residues in bank filtration. *Ground Water Monitoring and Remediation*. 24, 70–77.

Holm, J.V., Rügge, K., Bjerg, P.L., Christensen H. (1995) Occurrence and distribution of pharmaceutical organic compounds in the groundwater downgradient of a landfill (Grindsted, Denmark). *Environmental Science and Technology*. 28, 1415–1420.

Jobling, S., Nolan, M., Tyler, C.R., Brighty, G., Sumpter J.P. (1998) Widespread sexual disruption in wild fish. *Environ. Sci. Technol.* 32, 2498–2506.

Joss, A., Zabczynski, S., Göbel, A., Hoffmann, B., Löffler, D., McArdell, C.S., Ternes, T.A., Thomsen, A., Siegrist, H. (2006) Biological degradation of pharmaceuticals in municipal wastewater treatment: Proposing a classification scheme. *Water Research*. 40(8), 1686-1696.

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 U.S. streams, 1999–2000 - a national reconnaissance. *Environ Sci Technol.* 36, 1202–1211.

Knepper, T.P., Sacher, F., Lange, F.T., Brauch, H.J., Karrenbrock, F., Roerden O. et al. (1999) Detection of polar organic substances relevant for drinking water. *Waste Management.*19, 77–99.

Lapworth, D.J., Baran, N., Stuart, M.E., Ward R.S. (2012) Emerging organic contaminants in groundwater: A review of sources, fate and occurrence. *Environmental Pollution.* 163, 287–303.

Lee, H.B., Peart, T. (1998) Occurrence and Elimination of Nonylphenol Ethoxylates and Metabolites in Municipal Wastewater and Effluents. *Water Qual. Res. J. Can.* 33, 389-402.

Melcer, H., Bell, J. P., Thompson, D. J., Yendt, C. M., Kemp, J., Steel, P. (1994) Modeling Volatile Organic Contaminants' Fate in Wastewater Treatment Plants. *J. Environ. Eng.* 120(3), 588–609.

Mihelcic, J.R., Lueking, D.R., Mitzell, R.J., Stapleton, J.M. (1993) Bioavailability of sorbed- and separate-phase chemicals. *Biodegradation.* 4, 141-153.

Monteith, H.D., Parker, W.J., Bell, J.P., Melcer, H. (1995) Modeling the fate of pesticides in municipal wastewater treatment. Water Environment Research. 67(6), 964-970.

Muir, D.C.G., Howard, P.H. (2006) Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ. Sci. Technol.* 40, 7157-7166.

Ottmar, K., Colosi, L.M., and Smith, J.A. (2010) Sorption of Statin Pharmaceuticals to Wastewater-Treatment Biosolids, Terrestrial Soils, and Freshwater Sediment*. J. Environ. Eng.* 136, 256–264.

Parker, W.J., Monteith, H.D., Melcer, H. (1994) Estimation of anaerobic biodegradation rates for toxic organic compounds in municipal sludge digestion. *Water Res.* 28, 1779–1789.

Petrovic, M., Barceló, D. (2006) Application of liquid chromatography/quadruple time-of-flight mass spectrometry (LC-QqTOF-MS) in the environmental analysis. *Journal of Mass Spectrometry.* 41, 1259–1267.

Petrovic, M., Gonzalez, S., Barcelo, D. (2003) Analysis and removal of emerging contaminants in wastewater and drinking water. *Trends Anal. Chem.* 22, 685−696.

Plosz, B.G., Leknes, H., Thomas, K.V. (2010) Impacts of competitive inhibition, parent compound formation and partitioning behavior on the removal of antibiotics in municipal wastewater treatment. *Environ. Sci. Technol.* 44, 734-742.

Pomiès, M., Choubert, J.M., Wisniewski, C., Coquery, M. (2013) Modelling of micropollutant removal in biological wastewater treatments: A review. *Science of The Total Environment.* 443, 733–748.

Poseidon, 2005. POSEIDON—Final Report, http://www.eu-poseidon.com.

Radjenovic, J., Petrovic, M., Barceló, D. (2009) Fate and distribution of pharmaceuticals in wastewater and sewage sludge of the conventional activated sludge (CAS) and advanced membrane bioreactor (MBR) treatment. *Water Res.*, 43, 831–841.

Richardson, S.D., Ternes, T.A. (2011) Water analysis: emerging contaminants and current issues. *Analytical Chemistry.* 83, 4614–4648.

Richardson, S.D., Ternes, T.A. (2005) Water analysis: emerging contaminants and current issues. *Anal. Chem.* 77, 3807−3838.

Salgado, R. Marques, R. Noronha, J. Carvalho, G. Oehmen, A. Reis M. (2012) Assessing the removal of pharmaceuticals and personal care products in a full-scale activated sludge plant. *Environ Sci Pollut Res*, 19 (5), 1818–1827.

Saliner, A.G., Patlewicz, G., Worth, A.P. (2005) A similarity-based approach for chemical category classification. EUR21867/EN.

Sanderson, H., Brain, R.A., Johnson, D.J., Wilson, C.J., Solomon K.R. (2004) Toxicity classification and evaluation of four pharmaceuticals classes: antibiotics, antineoplastics, cardiovascular, and sex hormones. *Toxicology.* 203(1–3), 27–40.

Schwarzenbach, R.P., Gschwend, P.M., Imboden D.M. (2003) Environmental Organic Chemistry. 0-471-35750-2Wiley, Hoboken, New Jersey, USA.

Simonich, S.L., Federle, T.W. Eckhoff, W.S. Rottiers, A. Webb, S. Sabaliunas, D. De Wolf, W. (2002) Removal of fragrance materials during US and European wastewater treatment. *Environ. Sci. Technol.* 36, 2839–2847.

Smith, V.H., Tilman, G.D., Nekola J.C. (1999) Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environmental Pollution.* 100, 179–196.

Stevens-Garmona, J., Drewesa, J.E., Khanb, S.J., McDonaldb, J.A., Dickensona E.R.V. (2001) Sorption of emerging trace organic compounds onto wastewater sludge solids. *Water Research.* 45(11), 3417–3426

Swartz, W.C.H., Reddy, S., Benotti, M.J., Yin, H.F., Barber, L.B., Brownawell, B.J., Rudel R.A. (2006) Steroid estrogens, nonylphenol ethoxylate metabolites, and other wastewater contaminants in groundwater affected by a residential septic system on Cape Cod, MA. *Environmental Science and Technology.* 40, 4894–4902.

Ternes, T.A., Herrmanna, N., Bonerza, M., Knackerb, T., Siegristc, H., Joss, A. (2004) A rapid method to measure the solid–water distribution coefficient (K_d) for pharmaceuticals and musk fragrances in sewage sludge. *Water Research*. 38, 4075–4084.

Thompson, J., Eaglesham, G., Mueller J. (2011) Concentrations of PFOS, PFOA and other perfluorinated alkyl acids in Australian drinking water. *Chemosphere*. 83, 1320–1325.

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 & Combinatorial Science.* 22(1), 69–77.

Watanabe, N., Bergamaschi, B.A., Loftin, K.A., Meyer, M.T., Harter T. (2010) Use and environmental occurrence of antibiotics in freestall dairy farms with manure forage fields. *Environmental Science and Technology.* 44, 6591–6600.

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 Research.* 44, 373–384.

Chapter 2 – Application of QMSA as a tool for estimation of environmental engineering parameters of emerging contaminants

2.1 Introduction

As discussed in Chapter 1, QMSA is an alternative approach to traditional, regression-based QSAR models. Rather than directly applying linear regression on all possible molecular descriptors, it uses molecular descriptors as a ruler to measure the similarities among chemicals. Estimated property values for the "target" chemical are made based on the properties of its nearest neighbors. This is an evolving technique, which tends to be particularly appealing for applications in which predictive accuracy may be hampered by limited data availability (Basak and Gute, 1995). To date, QMSA models have been used to accurately predict various parameters related to environmental risk assessment, including: mutagenicity (Basak and Gute, 1995), hepatoxicity (Gute et al., 2004), and skin sensitization, plus various other REACH-relevant toxicity endpoints (Estrada et al., 2004). However, there is still no clear evidence that QMSA can be used to predict relevant environmental engineering parameters.

Thus, this chapter has one objective: to validate the use of QMSA in predicting one important environmental engineering parameter of interest, i.e. estrogencity of emerging contaminants. To achieve this, we need to demonstrate that a QMSA model can accurately predict estrogenicity of studied emerging contaminants.

2.2 Experimental

2.2.1 Data Sources

Estrogenicity data were taken from a paper by Nishihara et al. (2000). Measurements were collected using a yeast two-hybrid (Y2H) *in vitro* estrogenicity assay for 517 compounds in municipal and industrial wastes. Since the large majority of these estrogenicity measurements were reported using "less than" or "greater than" instead of exact numeric values, only 55 of the 517 compounds measured by Nishihara et al. could be utilized in this study. Additional Y2H estrogenicity measurements were taken from: Hayakawa et al. (2007); Kameda et al. (2008); Kawamura et al. (2003); and Nakano et al. (2001). In total, estrogenicity measurements were procured for N = 81 unique compounds including plasticizers, pesticides, herbicides, surfactants, steroid hormones, and other classes. Values were in units of REC10 (10% relative effective concentration); i.e., the concentration of a test chemical eliciting 10% of the response associated with a standard solution of 10E-7 M 17β-estradiol. The common logarithm transformation was applied to all estrogenicities, since these values spanned several orders of magnitudes.

2.2.2 Molecular Descriptors

Several hundred molecular descriptors were calculated using MolconnZ (Edusoft LC) and SYBYL v. 7.3 (Tripos Inc) for the 81 chemical structures evaluated in this study. These parameters included information on molecular connectivity, shape, quantitative characteristics, etc. To neutralize the dramatic differences in scale among computed molecular descriptors, all values were rescaled using a two-step process similar to that of Basak et al. (2003). First, a constant coefficient was added to all values of each molecular descriptor type such that no adjusted property value was less than or equal to zero. Then, all adjusted values were transformed via the natural logarithm. Molecular descriptors for which all compounds exhibited the same rescaled value were removed from the data set, resulting in a total of 193 usable molecular descriptors.

Because many of the descriptors computed by MolconnZ were highly correlated, three statistical approaches were used to select which of the 193 indices were most appropriate for inclusion in a QMSA model. This was necessary for several reasons: 1) use of so much chemical information significantly reduces calculation speeds; 2) many indices are highly inter-correlated and thus offer redundant chemical information; and, 3) it was expected that use of a smaller number of representative indices may provide deeper insight about how chemical structure drives environmental behavior.

The first general approach that was used for data reduction was Principal Component Analysis (PCA). Here, PCs (principle components) are composite "descriptors" formed via linear combinations of all available indices. The same number of PCs is formed as there are unique indices to work with; however, for this research, only statistically significant PCs (eigenvalues $\lambda \geq 1$) were retained in the final QMSA models. These retained PCs were used in two ways. For one scenario ("PCA"), the entire linear combination comprising each PC was used to make QMSA predictions. For the other scenario ("iPCA"), only the best correlated descriptor within each PC was used for QMSA. The best correlated descriptor was determined based on the absolute value of correlation coefficient associated with each PC. All PCA (and iPCA) calculations were performed using PASW Statistics 18 (SPSS Inc.).

The third statistical approach for data reduction was ridge regression (iRR), which is specifically designed to overcome multi-colinearity among predictors (Crisp et al., 1998). Similar to ordinary least-squares multiple linear regression, in iRR, all indices are utilized as independent predictors to form linear combination predictions of the environmental parameter of interest. However, a ridge parameter "beta", β is introduced to reduce multi-colinearity and redundancy among available chemical information. A detailed description of iRR calculations and their interpretation are provided by Basak et al. (2003). All iRR calculations were performed using Matlab (Mathworks Inc., Version 7.0).

2.2.3 Similarity Computation

Similarity between pairs of molecules was computed using the method of Euclidean Distances in an n-dimensional space, according to Eq. 2.1.

$$ED_{ij} = [\sum_{k=1}^{n}(X_{ik} - X_{jk})^2]^{1/2} \qquad (2.1)$$

Here, ED_{ij} is the Euclidean Distance between molecules i and j; n is the number of statistically significant principal components; and X_{ik} and X_{jk} are values of descriptor k, respectively (Gute et al. 2001). Pairs of molecules separated by smaller ED values are said to be more similar than pairs of molecules separated by larger distances, Two molecules with ED = 0 must have exactly identical chemical structures.

2.2.4 Selection of Nearest Neighbors, Property Estimation, and Model Validation

The k-nearest neighbor (*k*-NN) method was employed for estimation of a probe chemical's desired property. This was done in three steps, First, compute the EDs between probe chemical and each compound in the estrogenicity dataset, Second, identify which k compounds within the estrogenicity dataset exhibit the smallest EDs from the probe. Finally, compute the estimated estrogenicity for probe chemical using the arithmetic average of the estrogenicity values for the k chemicals that are most similar to the probe chemical.. In this study, tested k values ranged from 1 – 10.

The leave-one-out (LOO) cross validation procedure was used to assess the accuracy of QMSA predictions (Hawkins et al. 2003). In this procedure, one compound at a time is removed the dataset, and the $N - 1$ remaining compounds are used to predict its value. This is done systematically N times. The resulting N predictions are then compared with measured values for their respective compounds to enable computation of the cross validation coefficient (q^2) and the standard prediction residual error sum of squares (S_{PRESS}). Following literature precedent, two slightly different variations of q^2 coefficients were computed; i.e., the so-called "naïve" and "true" q^2 coefficients. The difference between these two hinges on whether descriptor selection occurs before or after assignment of the training subset. It has been demonstrated that selection of one single set of molecular descriptors prior to all cross-validation via iterative model fitting results in higher q^2 estimates (the so-called "naïve q^2") compared to iterative selection of different molecular descriptors for each cross-validation model fitting (as is done in during computation of the so-called "true

q^2"). Since both coefficients are currently reported in the literature (Hawkins et al. 2004; Basak et al. 2009; Roy and Das 2010, Li et al. 2009; Zhou et al. 2010), both were computed for this investigation. The reader is referred to Kraker et al. (2007) for a more detailed explanation of the differences between naïve and true q^2. S_{PRESS} is the standard prediction residual error sum of squares, such that smaller values of this parameter are indicative of better model accuracy.

2.3 Results and Discussion

2.3.1 Indices Selection for Development of PCA, iPCA, and iRR QMSA Models.

Table 2.1 summarizes the list of indices (i.e., molecular descriptors) selected for use in both types of PCA QMSA models and also the ridge regression (iRR) QMSA model.

PCA revealed that only 15 PCs should be retained on the basis of statistical significance (eigenvalue ≥ 1). These were used for PCA QMSA modeling, such that less than 10% of the total available chemical information (193 indices) was required to generate QMSA predictions. For each of the 15 retained PCs, correlation analysis was used to determine which index was best correlated with each retained PC, based on highest absolute value of correlation coefficient (R). The best-correlated indices were retained for use in so-called "iPCA" QMSA modeling. For cases in which two or more PCs had the same top-correlated index, the second-most correlated index from one PC was also retained. In this way, a total of 15 most-correlated or second-most correlated indices were selected for iPCA modeling of the estrogenicity dataset. Ridge regression was also performed for estrogenicity dataset. In order to be consistent with iPCA QMSA modeling, only the top 15 most significant indices selected by RR were retained for "iRR" modeling.

Table 2.1. Indices generated from iPCA and iRR data reduction for both datasets. PCA-selected indices were used to construct both PCA and iPCA QMSA models for estrogenicity. RR-selected indices were used to construct iRR QMSA models for estrogenicity. Each set of descriptors comprises all of the chemical information required to generate a QMSA model for each type of environmental parameter.

iPCA	iRR
Wp	nelem
nXc4	SsssCH
k3	SssCH2
etyp22	Tg
etyp14	nd1
SsOH	**SssO***
SaasC	molweight
n3Pad34	naasC
ishape	etyp24
etyp12	muldiam
mulrad	nssCH2
SHCsatu	SHBd
SssO*	dXvp4
n4Pae13	Hmax
SssssC	Hmaxpos

* Single asterisks denote which index was selected by both iPCA and iRR approaches for use in QMSA modeling to predict in vitro estrogenicity.

From Table 2.1, use of the iPCA or iRR indices selection approaches resulted in highly distinct sets of molecular descriptors for estrogenicity QMSA modeling. There was only one common index among the lists of 15 indices selected by iPCA and iRR; namely, SssO.

2.3.2 Internal Validation of PCA, iPCA, and iRR QMSA Models.

Results from leave-one-out (LOO) validation experiments are presented in Figure 2.1. All three models (PCA, iPCA, iRR) exhibit excellent predictive abilities. In all three cases, highest validation coefficient (q^2), highest correlation coefficient (R), and lowest S_{PRESS} were achieved for the use of k = 1 nearest neighbor. As k increases, q^2 and R decrease while S_{PRESS} increases. All three changes are statistical indications of decreased predictive accuracy when larger numbers of nearest neighbors are used to make parameter predictions.

To reiterate from the previous paragraph, all QMSA models for the selected environmental parameters exhibit excellent accuracy. The observed magnitudes of q^2, R, and S_{PRESS} are comparable to the best reported literature values for similar types of predictive models (Asikainen et al., 2004; Waller 2004). In particular, the k = 1 PCA modeling approach seems to outperform all other models (k = ≥ 2, iPCA or iRR). It possesses the highest values of q^2 (0.84) and R (0.92), and the lowest S_{PRESS} (0.64) among all three estrogenicity models.

The superior performance of PCA compared to iPCA and iRR is perhaps not unexpected, since PCA incorporates the largest amount of chemical information among the three approaches. In particular, PCA employs all principal components with eigenvalues equal or greater than 1. In contrast, iPCA and iRR models only retain a small portion of the available indices and therefore may be missing some important similarity information. Notably, iRR models seem to be more powerful than their corresponding iPCA models. This is consistent with reports by other authors (Gute and Basak, 2006).

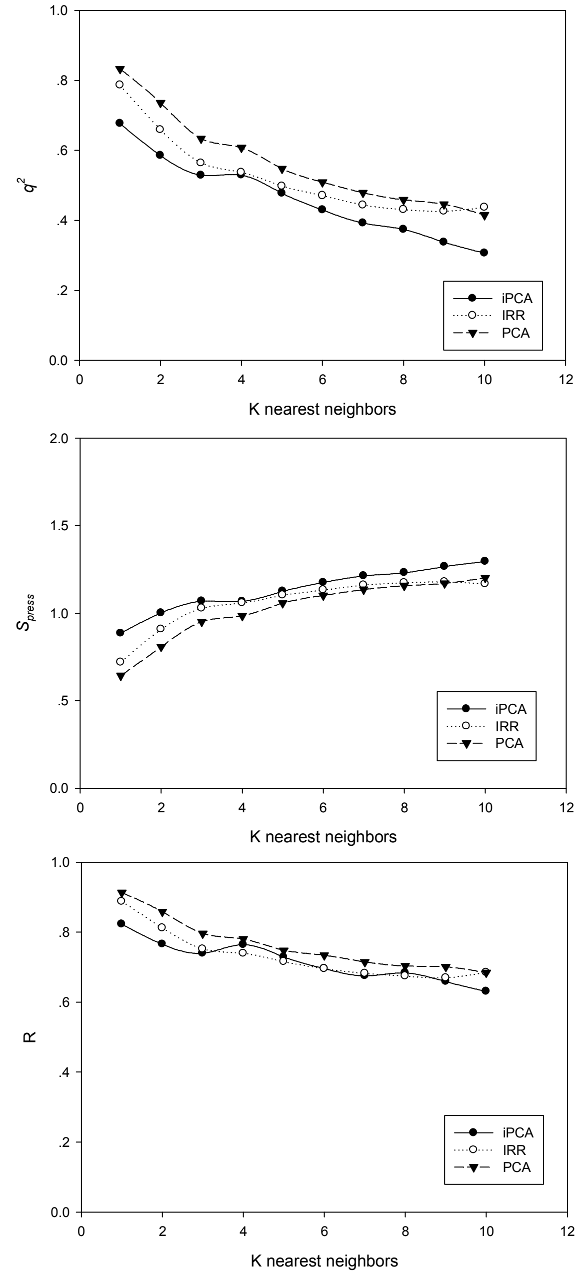

Figure 2.1. LOO validation statistics for three indices selection approaches (PCA, iPCA, and iRR), using k nearest neighbors ranging from k = 1-10, for QMSA modeling of *in vitro* estrogenicity. From top, statistics are q^2, R, and S_{PRESS}.

Table 2.2 shows the results of true q^2 and naive q^2 calculations. As evident in Table 2.2, cross-validation coefficients observed for the estrogenicity data set were 0.45-0.84 for naïve q^2 and 0.35-0.50 for true q^2. These indicate good model accuracy, as naïve q^2 values are on par with or slightly higher than naive true q^2 values observed by Asikainen et al. (2004) for prediction of *in vitro* estrogenicity using a related, but slightly different similarity technique (consensus *k*-NN structure-activity relationships). Similarly, true q^2 magnitudes are consistent with those reported by Kraker et al. (2007) for prediction of juvenile hormone activity (0.31-0.57).

Also evident in Table 2.2, naïve and true q^2 values exhibit different trends with increasing *k*. Naïve q^2 decreases dramatically with increasing *k* above $k = 1$ for the data utilized in this study; in contrast, true q^2 decreases with increasing k on the range 1-4, then decreases for $k = 5$, and holds roughly steady for $k = 6$-10. Of these two trends, the latter is more similar to a previously published QMSA study by Basak et al. (1995), in which increasing k over the range 1-10 mediated increasing q^2, followed by a decrease for $k > 10$.

Table 2.2. Cross-validation coefficients for leave-one-out (LOO) evaluation of a quantitative similarity assessment model to predict estrogenicity for $N = 81$ chemicals. “Naïve” and “true” designations refer to whether or not molecular descriptors were selected before or after subset data was removed from sample set, as described in Section 3.2.4. *k* was evaluated over the range 1-10 “nearest neighbors”.

k	1	2	3	4	5	6	7	8	9	10
Naïve q^2	0.84	0.75	0.69	0.67	0.61	0.57	0.54	0.51	0.50	0.45
True q^2	0.40	0.44	0.46	0.50	0.43	0.41	0.40	0.38	0.40	0.35

Improvement in q^2 with increasing k has generally been observed in datasets that are smaller and more structurally homogenous than the estrogenicity dataset used in this study; however, the naïve q^2 values observed here are not significantly less than naïve q^2 values in other QMSA studies or other structure-property studies in general. For this particular dataset, it is assumed that QMSA predictions utilizing small k (1-3) result in reasonably accurate predictions because they do not require averaging across a large number of dissimilar chemical structures. This is critical in so far as it suggests that QMSA, when implemented using a small k, can be an appropriate prediction tool for datasets encompassing a broadly diverse array of chemical structures, even despite scarce data availability (i.e. "nuggets"). Thus, emerging contaminants would seem well suited to QMSA predictions for emerging contaminants because these chemicals are both diverse in structure and poorly characterized to date.

Another interesting observation about the estrogenicity predictions summarized in Table 2.2, and about QMSA predictions in general, is related to the cases of extreme values. Because QMSA predictions rely on arithmetic averaging of property values for a compound's k nearest neighbors, compounds that are most dissimilar from the rest of a dataset are difficult to model accurately. In this study, four compounds were particularly problematic with respect to predictive accuracy. These include: two synthetic estrogens comprising the *most strongly estrogenic* substances in this dataset, 17α-ethynylestradiol (log REC10 = 9.70) and diethylstilbestrol (logREC10 = 9.70); and two polycyclic aromatic hydrocarbons (PAHs) comprising the *most weakly* *estrogenic* substances in this dataset, 3-hydroxybenz(a)-anthracene (log REC10 = 2.38) and 2-hydroxy-chrysene (log REC10 = 2.38).

Coincidentally, each of the chemicals with an extreme value had the same exact estrogenicity as one other chemical in this dataset. Still, not all of the QMSA models generated in this study could correctly identify which chemical should be most similar to the strongest or weakest estrogens. Additionally, LOO cross-validation trials in which one very strong or very weak chemical was removed from the training set made it virtually impossible to assign an accurate estrogenicity for the remaining very strong or weak chemical. This is because all of the other values were so much less extreme. This is clearly a shortcoming of QMSA models based on PCA; however, despite this weakness, excellent q^2 values were achieved for this highly diverse training set.

2.4 Conclusions

In this chapter, we examined the potential application of QMSA models for evaluation of a key environmental engineering parameter, *in vitro* estrogenicity, among the rapidly expanding number of emerging contaminants Since this particular environmental endpoint has been widely measured over the last twenty-five years, this study comprises a “retrospective” validation of QMSA prioritization using an existing, relatively rich data set. It has been demonstrated that QMSA can accurately predict environmental information for large, highly diverse classes of chemical structures. This is made evident by the good q^2 values (0.84) achieved for prediction of *in vitro* estrogenicity measurements (Table 2.2). Thus, it seems likely that efficient use of accurate QMSA models could deemphasize the need for labor-intensive, time-consuming analytical measurements to support regulatory agendas related to emerging contaminants.

2.5 References

Asikainen, A.H., Ruuskanen, J., Tuppurainen, K.A. (2004) Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. *Environ. Sci. Technol.* 38, 6724-6729.

Basak, S.C., Gute, B.D., Mills, D., Hawkins, D.M. (2003) Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: A comparison of arbitrary vs. tailored similarity spaces. *J. Mol. Struc. (Theochem)*. 622, 127-145.

Basak, S.C., Grunwald, G.D. (1995) Predicting mutagenicity of chemicals using topological and quantitative chemical parameters: A similarity-based study. *Chemosphere*, 31, 2592-2546.

Crisp, T.M., Clegg, E.D., Cooper, R.L., Wood W.P. et al. (1998) Environmental endocrine disruption: an effects assessment and analysis. *Environ Health Perspective.* 106(Suppl 1), 11-56.

Estrada, E., Patlewicz, G., Gutierrez, Y. (2004) From knowledge generation to knowledge archive: A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitization. *J. Chem. Info. Comp. Sci.*, 44, 688-698.

Gute, B.D. and Basak, S.C. (2006) Optimal neighbor selection in molecular similarity: Comparison of arbitrary versus tailored prediction spaces. *SAR QSAR Environ. Res.* **17**, 37–51.

Gute, B.D., Balasubramanian, K., Giess, K.T., Basak, S.C. (2004) Prediction of halocarbon toxicity from structure: A hierarchical QSAR approach. *Environ. Toxicol. Pharma.*, 16, 121-129.

Gute, B.D., Grunwald, G.D., Mills, D., Basak, S.C. (2001) Molecular similarity based estimations of properties: A comparison of structure spaces and property spaces, *SAR QSAR Environ. Res.* 11, 363-382.

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

Hayakawa, K. Onoda, Y. Tachikawa, C. Hosoi, S. Yoshita, M. Chung, S.W. Kizu, R. Toriba, A. Kameda, T. Tang, N. (2007) Estrogenic/antiestrogenic activities of polycyclic aromatic hydrocarbons and their monohydroxylated derivatives by yeast two-hybrid Assay, *J. Health Sci.*, 53, 562-570.

Kamamura, Y., Ogawa, Y., Nishimura, T., Kikuchi, Y., Nishikawa, J., Nishihara, T., Tanamoto, K. (2003) Estrogenic activities of UV stabilizers used in food contact plastics and benzophenone derivatives tested by the yeast two-hybrid assay. *J. Health Sci.* 49, 205-212.

Kameda, T., Akiyama, A., Toriba, A., Tachikama, C., Yoshita, M., Tang, N., Hayakawa, K. (2008) Evaluation of endocrine disrupting activities of monohydroxylated derivatives of 1-nitropyrene by yeast two-hybrid assay. *J. Health Sci.*, 54, 118-122.

Kraker, J.J., Hawkins, D.M., Basak, S.C., Natarajan, R., Mills, D. (2007) Quantitative Structure–Activity Relationship (QSAR) modeling of juvenile hormone activity: Comparison of validation procedures, *Chemometrics and Intelligent Laboratory Systems*, 87(1), 33-42.

Nakano, S., Nagao, Y., Kobayashi, T., Tanaka, M., Hirano, S., Nobuhara, Y., Yamada, T. (2002) Problems with methods used to screen estrogenic chemicals by yeast two-hybrid assays. *J. Health Sci.* 48, 83-88.

Nishihara, T., Nishikawa, J., Kanayama, T., Dakeyama, F., Saito, K., Imagawa, M., Takatori, S., Kitagawa, Y., Hori, S., Utsumi, H. (2000) Estrogenic activities of 517 chemicals by yeast two-hybrid assay. *J. Health Sci.*, 46, 282-298.

Waller, C. (2004) A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J. Chem. Inf. Comput. Sci.*, 44 (2), 758–765

Chapter 3 – Molecular Similarity Analysis as a Tool to Predict Sorption Distribution Coefficients Emerging Contaminants

3.1 Introduction

As discussed in Chapter 1, extensive production and application of a wide range of unregulated chemicals, usually referred to as "emerging contaminants", makes these compounds widely detected in the environment. It has been demonstrated that wastewater treatment plants (WWTPs) are the most significant source of emerging contaminants, including estrogens and pharmaceutical compounds, into natural water bodies (Kolpin et al., 2002; Daughton, 2004; Richardson and Ternes, 2005). This is because municipal WWTPs in the US have been designed to remove biochemical oxygen demand (BOD) and nutrients (nitrogen, phosphorus) rather than emerging contaminants. As a result, WWTP removal efficiencies for estrogens, pharmaceuticals, and other compounds can be quite low or vary widely from plant to plant (Heberer, 2002; Golet et al., 2003; Wick et al. 2009; Carballa et al. 2004). This inability to consistently remove emerging contaminants at high efficient poses potential safety risks to the US drinking water supply.

The two main types of treatment at a municipal WWTP include sorption and biodegradation. These treatments are not completely ineffective for removal of emerging contaminants, but it is not presently understood how well each type of treatment mediates removal of estrogens, pharmaceuticals, and other emerging contaminants. Thus there would be great value in understanding fate and transport parameters pertaining to these two removal processes. This mechanistic understanding could ultimately enable estimation of how these compounds will behave within the

WWTP and in the natural environmental following effluent discharge or land-application of biosolids (Ternes et al. 2004; Barron et al. 2009). This chapter will focus on use of QMSA modeling to estimate sorption distribution coefficient (K_d), which is an important parameterization of sorption efficacy.

Traditional sorption investigations have focused on measurement of the sorption distribution coefficient (K_d) (Ternes et al. 2004). This parameter encapsulates the distribution of a chemical between aqueous and suspended solid phase. K_d has been well demonstrated as one of the most significant parameters responsible for removal of organic chemicals, since it can be used to accurately predict the sorption process in WWTPs (Schwarzenbach, 2003). This has also been reflected on the increasing research interests on measurement of K_d in WWTPs in recent years (Golet et al. 2003; Artola-Garicano et al. 2003; Clara et al. 2005; Radjenovic et al. 2009; Barron et al. 2009; Ottmar et al. 2010; Stevens-Garmon et al. 2011).

The increasingly large number of emerging contaminants requiring evaluation makes it impractical, if not impossible, to obtain direct measurements of the K_d parameter for all compounds of interest. This approach would be too expensive and time-consuming. Therefore many environmental regulation agencies worldwide, such as US Environment Protection Agency (EPA) and European Environmental Agency (EEA), encouraged people obtaining such information by applying Quantitative Structure-Property Relationships (QSPR) (EC, 2006).

In Chapter 2, we demonstrated that QMSA can potentially serve as a powerful tool to predict an environmental engineering parameter, i.e. *in vitro* estrogenicity. Despite this, there has been almost no research on its sorption and biodegradation parameters in wastewater to date.

Therefore, this study has three objectives: 1) provide K_d of three emerging contaminants which have never been measured before; 2) develop a QMSA model that generates accurate estimations of sorption distribution coefficient (K_d) for sludge-water systems simulating sorption in primary and secondary treatment in WWTPs; and 3) provide a K_d estimation list for 223 unmeasured emerging contaminants studied in previous literatures. It is expected that K_d estimations generated from this model will be valuable for environmental regulators or researchers to obtain K_d of unmeasured emerging contaminants in WWTPs. It is even more valuable for understanding and ultimately predicting WWTP removal of emerging contaminants. Thus, QMSA-derived estimations are benchmarked against existing measurements and estimations from other models. Additionally, laboratory measurements for three highly-prescribed but previously unevaluated pharmaceutical compounds are presented as external validation of model accuracy.

3.2 Experimental

3.2.1 Data Sources.

Distribution coefficient (K_d) data were taken from various literature sources for municipal wastewater treatment plant (WWTP) biosolids ("sludge") as sorbent for pharmaceuticals, steroid estrogens, fragrance materials, biocides and some other classes of emerging contaminants, namely: Barron et al (2009); Ternes et al (2004); Feng et al (2010); Andersen et al (2005); Carballa et al (2007); Kupper et al (2006); Wick et al (2009); Simonich et al (2002); Zhao et al (2008); and Ottmar et al (2010). For chemicals which had more than one reported K_d value, the arithmetic mean was taken for all available values. All told, K_d measurements for 80 different chemicals were collected from the literature. As with the estrogenicity dataset, the common logarithm transformation was applied to reduce significant variability. To sum up, there are 81 measured chemicals in estrogenicity dataset and 80 measured chemicals in K_d dataset, respectively.

The total pool of chemical structures used in this study included not only those compounds that had been previously measured for estrogenicity and/or K_d but also all other emerging contaminants referenced in the papers noted above plus additional structures from a list of 200 top-prescribed generic pharmaceuticals (Verispan VONA, 2007). The total number of chemical structures evaluated, *N*, including previously measured and unmeasured chemicals, was 303.

3.2.2 Molecular Descriptors.

Several hundred molecular descriptors were calculated for the 303 chemical structures in the QMSA pool. This was done using MolconnZ (Edusoft LC) and SYBYL v. 7.3 (Tripos Inc). Computed descriptors included information on molecular connectivity, shape, quantum characteristics, etc. To neutralize dramatic differences in scale among the various computed molecular descriptors, all values were rescaled using a three-step process modified from Basak et al. (2003). First, a constant was added to all values of each molecular descriptor such that all values were greater than zero. Second, molecular descriptors for which all compounds exhibited the same rescaled value were removed from the data set. Finally, the 303 values of each molecular descriptor (the vector corresponding to all compounds in the training data set) were normalized by dividing each value by Eculidean length of the vector such that the Eculidean length of the normalized vector was one. The total number of usable molecular descriptors ("indices") resulting from this process was 193.

Three statistical approaches were used to identify which of the 193 indices were most appropriate for inclusion in a particular QMSA model. This was necessary because many of the molecular descriptors encapsulated redundant chemical information and because use of excessively large molecular descriptor sets makes for undesirably long calculation times. The first general approach for data reduction was Principal Component Analysis (PCA). In this analysis, PCs (principle components) are composite "descriptors" formed via linear combinations of all available indices. The same number of PCs are formed as there are unique indices to work with;

however, for this research, only statistically significant PCs (eigenvalues $\lambda \geq 1$) were retained. These retained PCs were used in two ways. For one scenario ("PCA"), the entire linear combination comprising each PC was used to make QMSA estimations. For the other scenario ("iPCA"), only the best correlated descriptor within each PC was used for QMSA. The best correlated descriptor was determined based on the absolute value of correlation coefficient associated with each PC. All PCA (and iPCA) calculations were performed using PASW Statistics 18 (SPSS Inc.).

The third statistical approach for data reduction was ridge regression (iRR), which is specifically designed to overcome multi-collinearity among predictors (Crisp et al, 1998). Similar to ordinary least-squares multiple linear regression, in RR, all indices are utilized as independent predictors to form linear combination estimations of the environmental parameter of interest. However, a ridge parameter "beta", β is introduced to reduce multiconllinearity and redundancy among available chemical information. A detailed description of RR calculations and their interpretation are provided by Basak et al (2003). All RR calculations were performed using Matlab (Mathworks Inc., Version 7.0).

3.2.3 Similarity Computation, Property Estimation, and Internal Validation.

The same procedure was performed according to Chapter 2, Section 2.2.3.

3.2.4 Selection of Nearest Neighbors, Property Estimation, and Model Validation.

The same procedure was performed according to Chapter 2, Section 2.2.4.

3.2.5 Measuring K_d for Three Selected Prioritization Compounds.

Laboratory measurements were required to for external validation of the QMSA model generated in this study; i.e., to demonstrate that the model can make accurate K_d estimations for compounds not included in its training set. Metformin, fluconazole and benazepril were chosen as the test chemicals since they appear to be among top 200 most prescribed pharmaceuticals in US (www.pharmacytimes.com) but none of their K_d have been reported before. Experimental procedures are summarized in the following paragraphs.

3.2.5.1 Materials and Chemical Reagents.

12 L Wastewater samples were collected from the secondary aerations basins at the Charlottesville WWTP. Samples were then transferred into three 4-L Erlenmeyer flasks after immediately transported to the laboratory. A synthetic wastewater solution (Ottmar, 2010) was then pumped into each reactor at approximately 1.5 mL/min. The composition of synthetic wastewater stock solution was adopted from Pholchan et al. (2008). The outlets of all three flasks were connected to a 10-L bucket, to collect overflow from each reactor. All outflow (effluent and sludge) collected during the first month was discarded, to ensure that all sludge used for sorption experiments was

initially drug-free. The outflow was also examined by HPLC to ensure no relevant chemicals can be detected. Sludge samples for use in each sorption experiment were collected from the reactors after measurement of total suspended solids (TSS) according to Standard Method 2540D (Eaton, 1995). Sludge samples were then autoclaved at 120 °C for 30 min and dosing with 2% (w/v) sodium azide to completely inactivate biological activity. The selected drug compounds were purchased from Fisher Scientific, Inc. Stock solutions were prepared by dilution in deionized water (DI) generated by NANOpure ultrapure water system (USA).

3.2.5.2 Kinetic Batch Experiments.

Preliminary batch kinetic experiments were performed to estimate equilibrium time for sorption of each test compounds onto WWTP. These preliminary experiments were also helpful for determining what amount of sludge should be used for K_d measurement experiments. For each test compound, three 15-mL glass centrifuge tubes were used as batch reactors. Each reactor received 5 mL of compound stock solution plus enough sludge solution to completely fill the test tube (~10 mL). The resulting concentration of the test compound was 500 μg/L. Sludge concentrations were 3.2 g/L, 3.0 g/L, and 6.5 g/L for metformin, fluconazole, and benazepril, respectively. Two types of controls were also: a positive control, comprising only compound stock solution and DI in each tube, without WWTP solids; and a negative control, comprising only autoclaved sludge solution and DI in each tube, without test compound. All controls were prepared in duplicate. pH was monitored throughout the experimental in all samples and controls for all three test compounds. All measurements were roughly 7.0.

All centrifuge tubes were sealed with Teflon-lined caps and transferred into a rotating horizontal shaker for incubation at 20 ºC. Samples were then collected at pre-determined times. For metformin, these times were t = 30 min, 5 h, 9 h, 20 h, 80 h, 120 h, and 288 h. For fluconazole, sampling times were t = 20 min, 5 h, 72 h, 120 h, and 150h. For benazepril, sampling times were t = 20 min, 3 h, 26 h, 96 h, and 168 h. For each sampling time, the three tubes were removed from the shaker and centrifuged at 3,500 rpm for 30 min at the incubation temperature. 600-μL aliquots of supernatant were then collected from each tube using a 1-mL syringe and passed through a 0.1-μm syringe filter (Millipore Inc.) to remove particulates. Approximately 200 μL of filtered supernatant was then transferred into glass vials containing 200-μL vial inserts. Following each sample collection, the centrifuged reactors were mixed to resuspend the autoclaved sludge. They were then returned to the incubating shaker until the next sampling interval.

3.2.5.3 Equilibrium Batch Sorption Isotherm Experiments.

Equilibrium batch isotherm experiments were designed for each test compound on the basis of preliminary results from the kinetic batch experiments. In order to obtain sorption isotherms, a series of concentrations was prepared to investigate sorption behaviors with different chemical concentration. For metformin, a series of eight solutions was created in DI water: 100, 200, 500, 1,000, 2,000, 5,000, 8,000, and

10,000 μg/L. For both fluconazole and benazepril, a series of seven solutions was created in DI water: 100, 200, 500, 1,000, 2,000, 5,000, and 10,000 μg/L. For each concentration, certain volume of chemcial was then loaded into triplicate 15-mL glass centrifuge tubes. Roughly 1 mL of autoclaved sludge solution followed by dosed with 3.0 mg/L sodium azide was also added to each tube, such that the final solids concentrations in each experiment were 5.1 g/L for metformin, 5.0 g/L for fluconazle, and 6.8 g/L for benazepril, All tubes were then incubated on a rotating shaker at 20 ºC for 7 d. Afterwards, all tubes were centrifuged and filtered using the same protocol referenced above. Drug concentrations were then analyzed via high pressure liquid chromatography (HPLC).

3.2.5.4 HPLC Measurement.

Drug concentrations were analyzed using a Shimadzu LC-20AB high performance liquid chromatograph (HPLC) with a diode array detector (DAD) set to 232 nm. Fluconazole and benazepril were analyzed by using the Agilent 1100 series HPLC with a DAD set to 210 nm and 235 nm, respectively. An Agilent (Santa Clara, CA) C-18 column was used for all of three compounds for chromatographic separation. Mobile phase was a mixture of buffer solution and acetonitrile (ACN). The buffer solutions used for analysis were 0.01 M sodium phosphate diabasic + 0.01 sodium dodecyl sulfate, adjusted to pH of 7, 0.01 M potassium phosphate monabaic, adjusted to pH 7, and 0.01 M sodium phosphate monabasic for metformin, fluconazole, and benazepril, respectively. Gradient methods were used for these three compounds. The detailed gradient procedure is listed in Table 3.1 – 3.3. Injection volume was set to 50 µL for all of three compounds. The retention times and detection limits for metformin, fluconzole, and benazepril were 18.8 min and 20 µg/ L, 19.0 min and 40 µg/L, and 20.4 and 40 µg/L, respectively.

Table 3.1. HPLC mobile phase gradient for analysis of metformin

Time (min)	Flow rate (mL/min)	Buffer percentage	ACN percentage
0-12	2.00	82	18
12-22	2.00	Linear decrease	Linear increase
22	2.00	73	27
22-23	2.00	Linear increase	Linear decrease
23	2.00	82	18
23-25	2.00	82	18

Table 3.2. HPLC mobile phase gradient or analysis of fluconazole

Time (min)	Flow rate (mL/min)	Buffer percentage	ACN percentage
0-2	0.45	95	5
2-24	0.45	Linear decrease	Linear increase
24	0.45	50	50
24-29	0.45	Linear increase	Linear decrease
29	0.45	95	5
29-35	0.45	95	5

Table 3.3. HPLC mobile phase gradient or analysis of benazepril

Time (min)	Flow rate (mL/min)	Buffer percentage	ACN percentage
0-8	0.40	83	17
8-24	0.40	Linear decrease	Linear increase
24	0.40	43	57
24-27	0.40	Linear increase	Linear decrease
27	0.40	83	17
27-35	0.40	83	17

3.3 Results and Discussion

As discussed before, the main objective of this study is to demonstrate the usefulness of QMSA on estimation of one of the most significant engineering property for emerging contaminants in WWTPs, i.e, sorption distribution coefficient. To achieve this goal, internal validation was first performed to test if QMSA is statistical significant for K_d dataset. After that, K_d of three chemical were measured and compared with the estimated values by QMSA to make sure QMSA is able to accurately estimate external chemicals. Finally, a list of estimated K_d values of all 303 chemicals was generated by QMSA to provide reference of K_d for other unmeasured chemicals.

3.3.1 Indices Selection for Development of PCA, iPCA, and iRR QMSA Models.

PCA models use PCs to construct similarity measurement. The PCs are the composite "descriptors" formed via linear combinations of all available indices. Although they have advantages on incorporating as much useful information as possible from a large number of molecular descriptors, the PCs themselves are usually complex and present a "black box" for people who are trying to understand it. Compared to PCA models, iPCA and iRR only extract a very small number of molecular descriptors to construct similarity measurement, which can be potentially convenient to understand significant molecular descriptors and molecular structural characteristics responsible for accurate prediction of the property of interest.

Table 3.4. Indices generated from iPCA and RR data reduction for the K_d dataset. PCA-selected indices were used to construct both PCA and iPCA QMSA models for K_d dataset. RR-selected indices were used to construct iRR QMSA models for K_d dataset. Each set of descriptors comprises the only chemical information required to generate a QMSA model for K_d dataset.

iPCA	RR
dX0	Xvp10
dXvp3	Xc4
mulrad	Xch5
ishape	Xch6
Pf	Xvc4
WT	Xvch5
Redundancy*	Xvch6
Qsv	dX1
Qv	dXv0
nwHBa	knotp
etyp12	knotpv
etyp33	IDCbar
n2Pag11	IDWbar
n2Pag12	totop
n4Pae12	Redundancy*
nHCsatu	SssCH2*
naasC	SsssCH
SssCH2*	SssssC

* Single asterisks denote which two indices were selected by both iPCA and iRR approaches for use in QMSA modeling to predict K_d.

For the K_d dataset, 18 statistically significant PCs were selected for PCA QMSA modeling, and 18 top-correlated indices were retained for iPCA QMSA modeling. Ridge regression was also applied to the same 193 indices. The 18 most-influential indices were retained for iRR QMSA to predict $K_{d.}$

Table 3.4 summarizes the list of indices (molecular descriptors) selected for use in both types of iPCA QMSA models and also the ridge regression (iRR) QMSA model. From Table 3.4, use of the iPCA or iRR indices selection approaches resulted in highly distinct sets of molecular descriptors for the same dataset. There were two common indices (SssCH2 and Redundancy) among the lists of 18 indices selected by iPCA and iRR. It should be noted that ridge regression was usually considered as a tailored QMSA model since it selected molecular descriptors by taking the target properties into account (Basak et al., 2009). Therefore molecular descriptors selected by RR should have some connections with the target properties, i.e. K_d in this case. The most influential descriptor selected by iRR was SssCH2, which is the sum of E-States for methylenes of a molecular. Surprisingly many authors also reported this type of indices had considerable contribution on QSAR model development and were very useful in evaluating molecular similarity (Hall, 1995). In contrast, iPCA models only gave a rank of 16th for SssCH2 out of 18 top correlated molecular descriptors. That probably because iPCA models extract molecular descriptors from the overall structural characteristics of all molecules but iRR will only extract those related to the target property.

3.3.2 Internal Validation of PCA, iPCA, and iRR QMSA Models.

As noted in Section 3.2.4, leave-one-out (LOO) validation was performed for each type of QMSA modeling for K_d dataset. The results from these experiments are presented in Figure 3.1. All three models exhibit excellent predictive abilities. In all three cases, highest validation coefficient (q^2), highest correlation coefficient (R), and lowest S_{PRESS} were achieved for k = 1 nearest neighbor. Usually q^2 above 0.5 is considered as a proof of highly predictive model (Golbraikh and Tropsha, 2002). Based on this criteria, tt can be seen that our q^2 for three models are generally very good for most of *k*. S_{PRESS} is the standard prediction residual error sum of squares, such that smaller values of this parameter are indicative of better model accuracy. As *k* increases, q^2 and *R* decrease while S_{PRESS} increases. All three changes are statistical indications of decreased predictive accuracy when larger numbers of nearest neighbors are used to make parameter predictions.

To reiterate from the previous paragraph, all QMSA models for the K_d exhibit excellent accuracy. The observed magnitudes of q^2, *R* and S_{PRESS} are comparable to the best reported literature values for similar types of predictive models (Asikainen et al, 2004; Waller 2004). In particular, the k = 1 PCA modeling approach seems to outperform all other models (k = ≥ 2, iPCA or iRR). It possesses the highest values of q^2 (0.82), and *R* (0.91), and lowest S_{PRESS} (0.41) among all three models. The second powerful models are iRR models. They generated the highest q^2 as 0.82 and lowest S_{PRESS} as 0.42, which is a little bit bigger than that of PCA models, when k = 1.

Similar to results in Chapter 2, PCA models seem most powerful of the three data reduction strategies evaluated in this study. In general this is perhaps because the incorporate the most chemical information, employing all principal components with eigenvalues equal or greater than 1. In contrast, iPCA and iRR models only retain a small portion of the available indices and therefore may be missing some important similarity information. However, in this study, the performance of iRR models seems to be comparable to PCA models even though they only used 18 individual molecular descriptors to construct the similarity measurement. Meanwhile iRR models are much simpler than PCA models and therefore 18 molecular descriptors selected by iRR models will have much more direct relationship with K_d. Notably, iRR models seem to be more powerful than their corresponding iPCA models. This is consistent with reports by other authors (Gute and Basak, 2006).

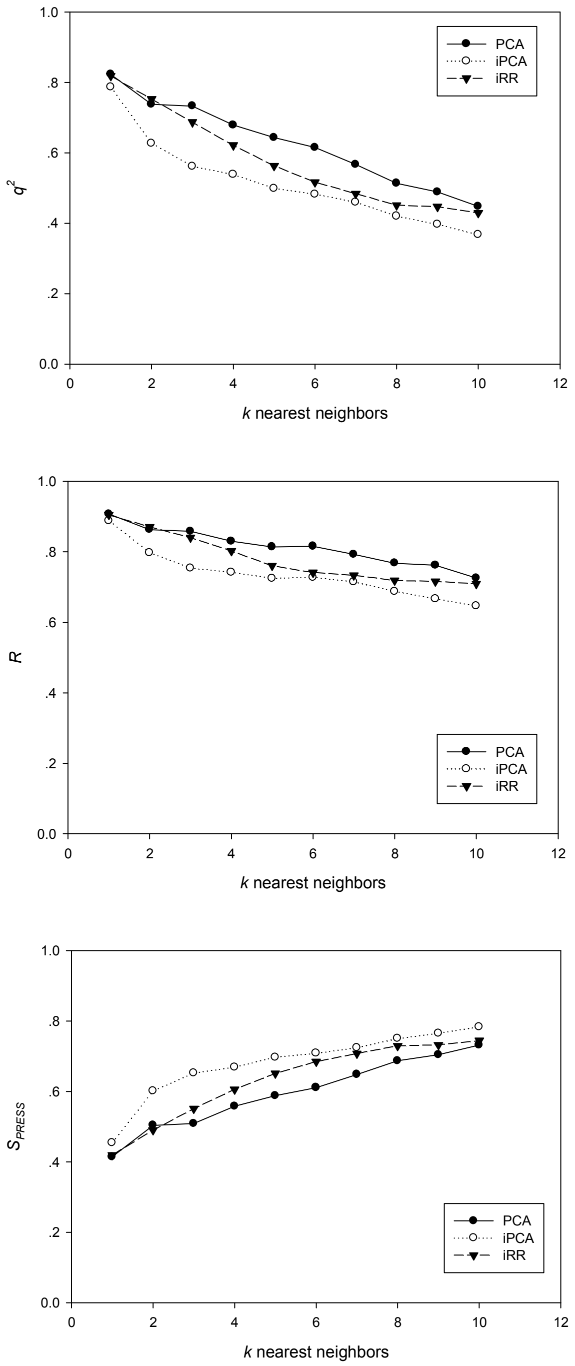

Figure 3.1. LOO validation statistics for three indices selection approaches (PCA, iPCA, and iRR) in QMSA modeling of sorption distribution coeffient (K_d). From top, statistics are q^2, R, and S_{PRESS}.

3.3.3 Measurement of K_d of three chemicals.

The other hypothesis of this study is as follows: QMSA models can be used to predict K_d of chemicals which have never been measured before. To test it, we need to measure K_d of selected chemicals and compare the measured K_d and predicted K_d. Therefore sorption coefficient K_d were measured for three test chemicals (metformin, fluconazole, and benazepril) in order to externally validate the accuracy of QMSA models for this parameter. This process is called external validation sometimes as in comparison to the internal validation we performed before. External validation is generally recommended for evaluation of QMSA (and other statistical modeling) predictions because it assesses the extent to which predictions can be accurately made for chemicals which were not part of the original training dataset.

Figure 3.2 – 3.4 show the results of K_d measurement of three chemicals. Figure 3.2A, 3.3A, and 3.4A depict the kinetic "pre-experiments", which were required to determine equilibrium time for sorption of three chemicals and also to assess what quantity of WWTP sludge is required to remove an appreciable (but not overly large) fraction of the starting metformin concentration (C_0). The duration of these "pre-experiments" were 288 h, 150 h, and 168 h for metformin, fluconazole, and benazepril, respectively. As evident from Figure 5.2A, 5.3A, and 5.4A, aqueous-phase concentration of three chemicals decreased over time. Results from positive controls (chemical + DI water without sludge; data not shown) suggest that this removal corresponds to sorption of the met onto the sludge since there were no reductions of chemical concentration in the absence of the sorbent. It also can be concluded that

metformin and benazepril reached their sorption equilibrium within 48 h, which were much quicker than fluconazole did. Fluconzole had a slow sorption rate. It reached its equilibrium after 120 h approximately.

According to previous research, the sorption kinetic can be described by the two site equilibrium/kinetic model as the following equation (Ottmar et al., 2010; Casey et al. 2003; Culver et al. 1997):

$$\frac{C_e}{C_0} = \frac{1+X_{ss}fK_d}{1+X_{ss}K_d} + (1 - \frac{1+X_{ss}fK_d}{1+X_{ss}K_d}) \times e^{(-\frac{1+X_{ss}K_d}{1+X_{ss}fK_d}\alpha t)} \tag{3.1}$$

Where f is the fraction of equilibrium sites and α is the mass-transfer-rate coefficient.

Measured metformin concentration (C_e) as a function of time (t) was fit to the exponential decay curve corresponding to Eq. 3.2:

$$\frac{C_e}{C_0} = 0.87 + 0.14 \times e^{-0.311t} \tag{3.2}$$

Similar procedures fit to fluconazole and benazepril data, we have fluconazole and benazepril as Eq. 3.3 and 3.4, respectively:

$$\frac{C_e}{C_0} = 0.76 + 0.24 \times e^{-0.017t} \tag{3.3}$$

$$\frac{C_e}{C_0} = 0.83 + 0.17 \times e^{-0.057t} \tag{3.4}$$

The regression coefficient (R^2) for these fits of Eq. 3.2, 3.3, and 3.4 were 0.96, 0.99, and 0.97, respectively, indicating very good agreement between the sorption kinetic model and our measurement. By looking at the graphs directly, the slope approaches zero at approximately 50 h, 100 h, and 120 h for metformin, fluconazole, and benazepril, respectively. To be more conservative, 120 h, i.e. 5 days was chosen as the equilibrium time for the all three chemicals' isotherm experiments designed to measure their K_d.

HPLC measurements provided equilibrium aqueous-phase metformin concentrations for each concentration. The mass of chemicals sorbed onto WWTP sludge (M_s, μg) was computed by subtraction of the equilibrium aqueous-phase concentration (C_e, μg/L) from each initial metformin concentration (C_0, μg/L) and multiplication by tube volume (V = 15 mL). Therefore, M_s can be written as:

$$M_s = (C_0 - C_e) \times V \tag{3.5}$$

This value was then divided by the known mass of sludge in each tube, to compute sorbed-phase metformin concentrations (S, μg/kg sorbent).

S values were plotted as a function of equilibrium aqueous-phase concentration (C_e) for comparison with the Freundlich adsorption isotherm equation (Eq. 3.6).

$$S = KC^n \tag{3.6}$$

Plots are depicted in Figure 3.2B, 3.3B, and 3.4B, wherein the regression curves are all linear (i.e., n = 0.95, 1.01, 1.05 with R^2 = 0.96, 0.99, and 0.99 for metformin, fluconazole, and benazepril, respectively). Thus, K is defined as the distribution coefficient (K_d). Notably, Figure 3.2B, 3.3B, and 3.4B exhibit very good linearity, such that the magnitudes of R^2 are as high as some of the best R^2 values reported in the literature. The resulting slopes were taken as K_d value for metformin, fluconazole, and benazepril as 48.6 L/kg, 65.8 L/kg, 32.6 L/kg, respectively. K_d values of three chemicals would be considered as low since the mean and median values of our 80 K_d dataset are 1495 L/kg and 105 L/kg, respectively. It can be expected that removal of three chemicals via sorption in wastewater treatment plants will be less significant than other emerging contaminants we know.

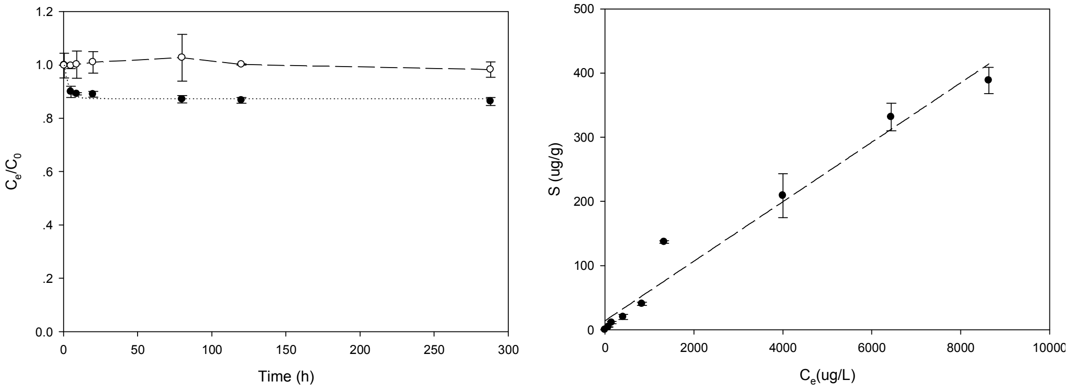

Figure 3.2. At left (4A), metformin concentration as a function of time during interaction with WWTP solids, for determination of equilibrium time. Dash line represents positive control change over time. At right (4B), sorbed-phase metformin concentration as a function of equilibrium aqueous-phase metformin concentration for determination of K_d.

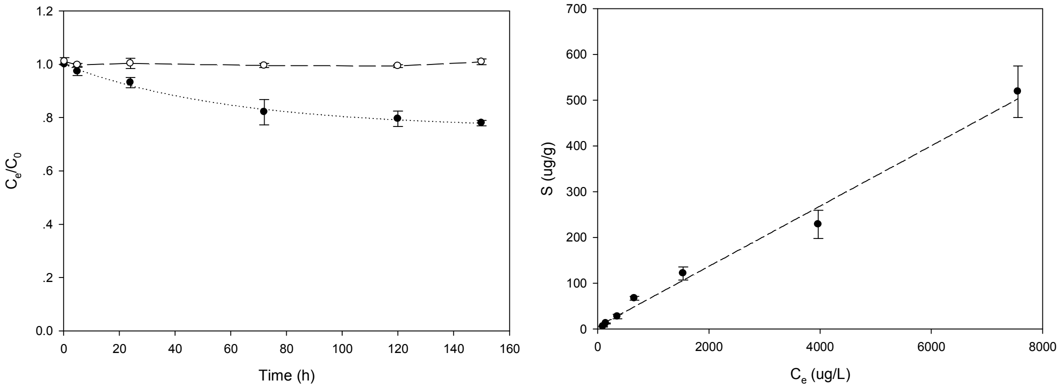

Figure 3.3. At left (5A), fluconazole concentration as a function of time during interaction with WWTP solids, for determination of equilibrium time. Dash line represents positive control change over time. At right (5B), sorbed-phase fluconazole concentration as a function of equilibrium aqueous-phase fluconazole concentration for determination of K_d.

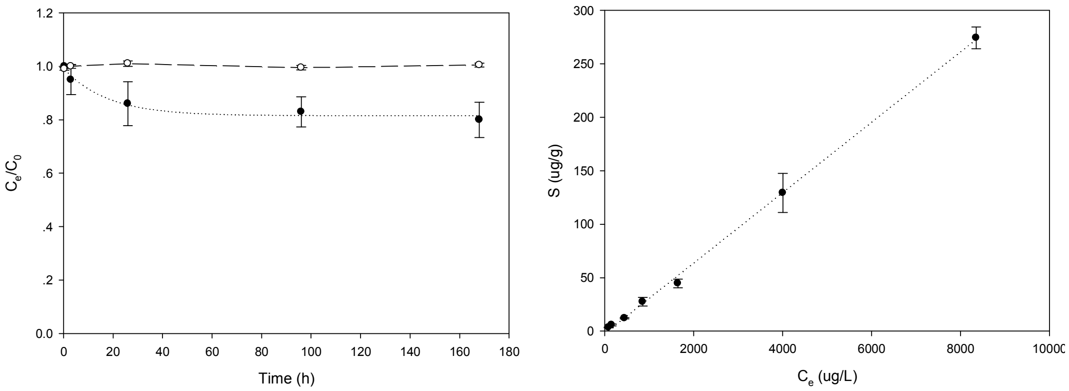

Figure 3.4. At left (6A), benazepril concentration as a function of time during interaction with WWTP solids, for determination of equilibrium time. Dash line represents positive control change over time. At right (6B), sorbed-phase benazepril concentration as a function of equilibrium aqueous-phase benazepril concentration for determination of K_d.

Measured K_d of three chemicals also allow us to back calculate the sorption-kinetic parameters, i.e. f and α. Table 3.5 summarized the calculated f and α for three chemicals. Calculated values of α for three chemicals suggested that they have relatively low sorption behaviors in wastewater compared to other rapid-sorption emerging contaminants (Ottmar et al., 2010). There seems no difference among values of f of three chemicals. That is probably because the same wastewater was used for all of three chemicals.

The low sorption extents (small K_d values) and slow sorption rates (small α values) of three studied chemicals will potentially make these chemicals very difficult removed within WWTPs by sorption processes. Then consider a scenario happened in a typical WWTP in United States with the average primary solid concentration around 0.0003 kg/L (Metcalf and Eddy, 1991), it can be expected that only approximately

1.4%, 1.9%, and 0.9% of total will adsorb onto the sludge. Therefore sorption process can only contribute little to remove these chemicals in WWTPs.

Table 3.5. Calculated kinetic sorption parameters for three chemicals.

Chemical	f	α (h^{-1})	K_d (L/kg)
Metformin	0.03	0.27	48.6
Fluconazole	0.04	0.01	65.8
Benazepril	0.03	0.05	32.6

3.3.4 Comparison between measured and predicted K_d of three chemicals.

Table 3.5 – 3.13 summarizes K_d predictions for three chemicals as computed using PCA, iPCA, and iRR indices selection. These predictions were computed using $k = 1 – 5$ measured nearest neighbors, and resulting values compared to the measured value from the laboratory experiments summarized above. To predict the K_d of three chemicals, the arithmetic averages of K_d of its nearest neighbors were taken as the predicted values when $k > 1$.

For metformin (Table 3.5 – 3.7), all of three types of QMSA models yield the excellent predictions. Especially for PCA and iRR models, their predictions were good enough even when k greater than 1. iPCA seems to be the least effective among three types of models. That is because it selected the second measured nearest neighbor which has a measured K_d (331.0 L/kg) greatly bigger than K_d of metfmorin. However, for k = 3 – 5. iPCA did make reasonable predictions. From the first five measured nearest neighbors provided by three types of models, we concluded that the PCA and iRR shared two same nearest neighbors out of five. Actually both PCA and iRR assigned 2,4-Dichlorophenol as the most nearest neighbor for metformin. K_d of 2,4-Dichlorophenol in wastewater were reported as 41.0 L/kg (Severtson and Banerjee, 1996) and 49.0 L/kg (Kennedy at al., 1992), compared to K_d of metformin as 48.6 L/kg measured in this study. Once again, similar to the results of internal validation, performance of iRR was comparable to PCA even though it only used 18 molecular descriptors.

For fluconazole (Table 3.8 – 3.10), the general predictive performances of three types of models were also excellent. Although the predicted K_d made by PCA and iRR are generally 1 – 3 times greater than measured one in this study, they are still in the same order of magnititude. In general performance of PCA (k = 1- 3) and iRR (k = 1- 5) would be acceptable. iPCA did the best predictions for fluconazole. K_d of all of the first five measured nearest neighbors selected by iPCA are very close to measured K_d of fluconazole. Especially for k =1 iPCA, it predicted K_d of fluconazole as 65.5 L/kg (Diclofenac) (Carballa et al., 2008) compared to 65.8 L/kg measured in this study.

For benzepril (Table 3.11 – 3.13), the general predictive performances of three models were the least accurate among three chemicals. Only PCA (k = 1) did the excellent prediction of 27.0 L/kg (warfarin) (Barron et al., 2009) compared to 32.6 L/kg measured in this study. The nearest measured neighbor for benazepril is warfarin as indicated by PCA. Warfarin was also selected by iPCA and iRR. However it is not the nearest measured one in iPCA and iRR.

The reasons why predictive performances of QMSA varied so much for three chemicals are probably as follow: 1) the relative similarity between measured neighbors and the target chemicals. In QMSA, the relative similarity was evaluated by *ED*. All the *EDs* between three target chemicals and their measured nearest neighbors were listed in Table 3.5 – 3.13. For all predictions, PCA models usually have the largest EDs and iRR models have the smallest EDs when evaluating the similarity between the same target chemical and its neighbors. This is because PCA took linear combinations of all 193 descriptors but iPCA and iRR only took 18 descriptors to establish the similarity space. It is interesting that iRR seems to have a much smaller ED measurement compared to iPCA even though both of them selected 18 molecular descriptors from 193. To determine how similar between two chemicals, only one type of QMSA models should be evaluated. For instance, the first measured nearest neighbor for metformin is 2,4-Dichlorophenol, which has an ED to metformin as 0.13 calculated by PCA. However, the first measured nearest neighbors for fluconazole and benazepril are furosemide and warfarin, respectively. The corresponding EDs to the target chemicals are 0.17 and 0.24 for fluconazole and benazepril, respectively. Therefore benazepril actually does not have the measured similar enough neighbor as metformin did in our current dataset. That probably explained why we have excellent predictions for metformin, but only the acceptable predictions for benazepril. 2) the experimental condition difference between literatures and this study. K_d would be greatly affected by experimental conditions, such as pH and temperature. It would be ideal to make predictions for K_d of three chemicals by using the literature data measured under the similar conditions as this study. In this study, experiments were performed at pH of 7 and temperature of 20 ºC. For metformin, its first measured nearest neighbor was identified as 2,4-Dichlorophenol by PCA and iRR. Interestingly two reported K_d values of 2,4-Dichlorophenol in wastewater were 41.0 L/kg and 49.0 L/kg in literatures, which were measured under pH 7 and 7.5, respectively. For fluconazole, its first measured nearest neighbor was identified as furosemide, diclofenac, and citalopram by PCA, iPCA, and iRR, respectively. No pH and temperature data can be found for furosemide (Stuer-Lauridsen et al., 2000). The pH for measuring diclofenac and citalopram were reported as approximately 6.5 (Carballa et al., 2008; Barron et al., 2009). Therefore they may not serve the better candidates for fluconazole as 2,4-Dichlorophenol did for metformin.

Table 3.5. PCA Model predictions for metformin K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 48.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between metformin and its neighbor in similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	2,4-Dichlorophenol	0.13	45.0	45.0
2	Paracetamol	0.15	19.0	32.0
3	APAP(Acetaminophen)	0.15	0.4	21.5
4	3,4-Dichlorophenol	0.16	96.0	40.1
5	4-Chlorophenol	0.18	18.0	35.7

Table 3.6. iPCA Model predictions for metformin K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 48.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between metformin and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Atenolol	0.06	37.7	37.7
2	Propranolol	0.07	331.0	184.3
3	Salicylic acid	0.07	23.0	130.6
4	Tramadol	0.08	47.0	109.7
5	2,4-Dichlorophenol	0.08	41.0	95.9

Table 3.7. iRR Model predictions for metformin K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 48.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between metformin and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	2,4-Dichlorophenol	0.01	45.0	45.0
2	3,4-Dichlorophenol	0.01	96.0	70.5
3	4-Chlorophenol	0.01	18.0	53.0
4	Paracetamol	0.01	19.0	44.5
5	APAP(Acetaminophen)	0.01	0.4	35.7

Table 3.8. PCA Model predictions for fluconazole K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 65.8 L/kg. *ED* is the Euclidean distance, which denotes the distance between fluconazole and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Furosemide	0.17	158.0	158.0
2	Diclofenac	0.18	65.5	111.8
3	Oxazepam	0.20	13.0	78.8
4	Triclocarban	0.20	2754.0	747.6
5	Nordiazepam	0.21	65.0	611.1

Table 3.9. iPCA Model predictions for fluconazole K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 65.8 L/kg. *ED* is the Euclidean distance, which denotes the distance between metformin and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Diclofenac	0.06	65.5	65.5
2	2,4-Dichlorophenol	0.07	41.0	53.3
3	Salicylic acid	0.07	23.0	43.2
4	3,4-Dichlorophenol	0.07	96.0	56.4
5	Flurbiprofen	0.07	65.0	58.1

Table 3.10. iRR Model predictions for fluconazole K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 65.8 L/kg. *ED* is the Euclidean distance, which denotes the distance between fluconazole and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Citalopram	0.02	282.0	282.0
2	Furosemide	0.02	158.0	220.0
3	Sulfamethoxazole	0.02	155.5	198.5
4	Indomethacin	0.02	121.0	179.1
5	Nifedipine	0.03	27.0	148.7

Table 3.11. PCA Model predictions for benazepril K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 32.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between benazepril and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Warfarin	0.24	27.0	27.0
2	Permethrin	0.24	5150.0	2588.5
3	Glibenclamide	0.24	239.0	1805.3
4	Indomethacin	0.27	121.0	1384.3
5	Ciprofloxacin HCl	0.27	471.0	1201.6

Table 3.12. iPCA Model predictions for benazepril K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 32.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between benazepril and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Permethrin	0.08	5150.0	5150.0
2	Warfarin	0.08	27.0	2588.5
3	Nortriptyline	0.08	600.0	1925.7
4	Simvastatin	0.08	866.5	1660.9
5	Amitriptyline	0.09	1049.0	1538.5

Table 3.13. iRR Model predictions for benazepril K_d based on its first five measured nearest neighbors. The laboratory-measured K_d value was 32.6 L/kg. *ED* is the Euclidean distance, which denotes the distance between benazepril and its neighbor in the similarity space. Note when $k = 1$, the predicted value was obtained by using its first nearest neighbor's K_d directly.

k	Nearest neighbor with known K_d	*ED*	Literature reported K_d	Predicted K_d
1	Ciprofloxacin HCl	0.02	471.0	471.0
2	Indomethacin	0.02	121.0	296.0
3	Warfarin	0.02	27.0	206.3
4	Glibenclamide	0.02	239.0	214.5
5	Loratidine	0.02	3321.0	835.8

3.3.5 Importance of ED on QMSA prediction accuracy

In order to identify the significance of *ED* values on QMSA estimation accuracy, the following test was performed. In this test, a threshold for *ED* was manually setup. For each target chemical, all its nearest neighbors with *ED*s less than or equal to the threshold were retained and considered as the neighbors used to estimate the K_d of this chemical. For target chemicals which don't have any nearest measured neighbors with *ED* less than or equal to the threshold, their first nearest neighbors were used for estimation. This process was done for all of 80 measured chemicals (77 from literatures and 3 from this study) for 80 times to calculate the internal validation coefficient q^2. A range of thresholds values from 0.03 to 0.20 was examined to find out the effects of thresholds on models' accuracy. The results were presented in Figure 5.5. It can be concluded that q^2 decreases with increasing threshold values of q^2. This conclusion is obviously reasonable since smaller *ED* threshold means neighbors selected for each target chemical in the dataset are more similar to the target chemical. However, in the real situation, since many chemicals usually don't have one or more measured neighbors with very small *ED*s, the threshold should therefore large enough to make sure every target chemical has at least one measured neighbor. From Figure 3.5, we recommend 0.13 as the *ED* threshold used to select nearest measured neighbors since q^2 drops dramatically when threshold greater than 0.13.

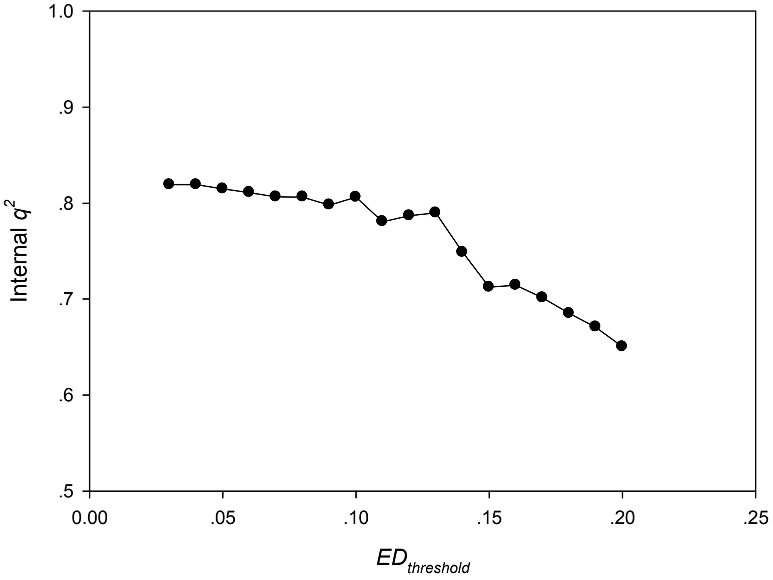

Figure 3.5. Internal cross validation coefficients of K_d dataset with different thresholds of *ED*

3.3.6 K_d Estimations for Unmeasured Chemicals

The final objective of this study is to provide estimated K_d values of other unmeasured chemicals. Therefore a “reliable” estimation list was provided for 63 unmeasured chemicals. The “reliable” chemicals are defined as those which have measured nearest neighbors *EDs* of which to them are less than 0.13. For each unmeasured chemical, its first five nearest measured neighbors were first selected. Only those with ED less than 0.13 were retained for predictions. This list could provide reference K_d values for these 63 un measured chemicals.

Table 3.14. QMSA predictions of K_d of 63 chemicals

Chemical No.	Chemical Name	Log (Predicted K_d)
23	MDMA (N-Methyl-3,4-methylenedioxyamphetamine)	2.33
34	Ranitidine	1.34
41	Sulfapyridine	2.19
43	Temazepam	1.44
49	Hydrocodone	1.15
66	Propoxyphene-N	3.71
67	Lorazepam	1.19
69	Clonazepam	1.35
72	Metoprolol Succinate	1.34
73	Cyclobenzaprine	2.46
74	Gabapentin	1.36
76	Citalopram HBR	2.45
79	Lovastatin	2.94
86	Allopurinol	1.65
89	Clonidine	2.33
90	Promethazine	3.27
96	Glyburid	2.38
103	Glipizide	2.38
105	Metronidazole	1.28
113	Mirtazapine	2.77
126	Clotrimazl	3.91

132	Felodipine	1.43
137	Methocarbamol	1.58
140	Phenazopyridine HCl	2.65
146	Aspirin	1.28
154	Baclofen	1.55
156	Phenobarbital	0.85
171	Hydralazine	2.33
172	Morphine Sulfate	1.08
191	Clopidogrel	2.85
192	Imipramine HCl	2.75
194	Hydromorphone HCl	1.08
196	Betaxolol	1.35
230	2-Hydroxy benzo[a]pyrene	1.98
232	3,4-Dichlorophenol	1.53
235	4,4'-Dihydroxybenzophenone	1.26
236	4,4'-Dihydroxybiphenyl	2.15
238	4-(branched)-Nonylphenol	1.26
239	4-Bromophenol	1.53
240	4-Chloro-3,5-xylenol	1.67
241	4-Chloro-3-methylphenol	1.37
243	4-Ethylphenol	1.37
245	4-Hydroxyacetophenone	1.28
247	4-Methylphenol	1.56
248	4-n-Butylphenol	1.28

251	4-n-Propylphenol	1.37
252	4-sec-Butylphenol	1.28
259	8-Hydroxy-3,4-dichlorodibenzofuran	1.56
265	Daidzin	1.11
270	Equol	2.03
272	Ethyl 4-hydroxybenzoate	1.55
274	Methyl 4-hydroxybenzoate	1.37
275	n-Butyl 4-hydroxybenzoate	2.33
276	n-Propyl 4-hydroxybenzoate	1.55
280	2-Hydroxy-4-methoxybenzophenone	1.92
281	2,2'-Dihydroxy-4-methoxybenzophenone	2.29
282	2-Hydroxybenzophenone	1.94
284	4-Hydroxybenzophenone	1.75
288	2-Hydroxy-5-methylbenzophenone	2.13
289	4-Hydroxy-4'-chlorobenzophenone	2.05
291	2,2',3,4'-tetrachlorobiphenyl	1.83
298	3-hydroxybenzo(k)fluoranthenes	1.98
299	10-hydroxybenzo(e)pyrenes	1.98

3.4 Conclusions

The results from this chapter provide important guidance related to the use of QMSA models for evaluation of significant environmental properties of the rapidly expanding number of emerging contaminants. First and foremost, this type of series of structural similarity models can accurately predict environmental information for large, highly diverse classes of chemical structures. This is proven by the excellent internal-validated q^2 values achieved for prediction of this K_d dataset and good external prediction results for K_d of three emerging contaminants which have never been measured before. Thus, it seems likely that efficient use of accurate QMSA models could deemphasize the need for labor-intensive, time-consuming analytical measurements to support regulatory agendas related to emerging contaminants. Second, the key success to make accurate external predictions by QMSA models depends on the relative similarity between the target chemicals and its nearest neighbors which have been measured before. The most ideal candidate to make accurate predictions for the target chemicals requires not only it is adequately similar to the target chemicals but also its properties of interest have been measured under the similar experimental conditions as that for the target chemicals. However, QMSA can still give predictions for unmeasured target chemicals by using their next measured nearest neighbors and it can be expected the predictive accuracy will be enhanced when more and more similar chemicals' property have been included. Finally, our measured K_d of three emerging contaminants also provided useful information of their sorption behaviors in wastewater. It can be expected that the removal of three chemicals by sorption process in WWTPs would be negligible.

The conclusions drawn from this chapter were made evident using K_d as a case study. Since this particular environmental endpoint is one of the most important parameters in WWTPs responsible for the removal of contaminants and it has been widely measured over the last twenty-five years, this study comprises a "retrospective" validation of QMSA prioritization using an existing, relatively rich data set. Future work will investigate the use of QMSA prioritization and subsequent formulation of prediction models for another significant environmental fate or behavior parameter in WWTPs that has been less widely studied than K_d; e.g. first-order biodegradation rate constants for prescription pharmaceuticals and personal care products. This property is more difficult to measure and measurements take more time and money. Therefore it is more desirable to get it by models. Additional work is also needed to figure out how to best-possible distribute different weight to different nearest neighbors based on their ED to the target chemicals such that prediction can be more accurate. In the following chapter, prediction of first-order biodegradation rate constants for these three chemicals made by QMSA will be examined to see if QMSA can predict this another important environmental engineering parameter responsible for removal of emerging contaminants in wastewater treatment plants.

3.5. References

Artola-Garicano, I., Borkent, J.L.M., Hermens, W., Vaes H.J. (2003) Removal of two polycyclic musks in sewage treatment plants: freely dissolved and total concentrations. *Environ. Sci. Technol.* 37(14) , 3111–3116.

Asikainen, A.H., Ruuskanen, J., Tuppurainen, K.A. (2004) Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. *Environ. Sci. Technol.* 38, 6724-6729.

Barron, L., Havel, J., Purcell, M., Szpak, M., Kelleher, B., Paull, B. (2009) Predicting sorption of pharmaceuticals and personal care products onto soil and digested sludge using artificial neural networks. *Analyst.* 134, 663-670.

Basak, S.C., Hawkins, D.M., Kraker, J.J. (2009) Quantitative Structure-Activity Relationship (QSAR) modeling of human blood: air partitioning with proper statistical methods and validation. *Chem. Biodivers.* 6, 487-502.

Basak, S.C., Gute, B.D., Mills, D., Hawkins, D.M. (2003) Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: A comparison of arbitrary vs. tailored similarity spaces. *J. Mol. Struc. (Theochem).* 622, 127-145.

Carballa, M., Fink, G., Omil, F., Lema, J.M., Ternes, T.A. (2008) Determination of the solid-water distribution coefficient (K_d) of pharmaceutical and personal care products (PPCPs) in digested sludge. *Water Res.* 42(1-2), 287–295.

Casey, F.X.M. (2003) Fate and transport of 17β-estradiol in soil–water systems. *Environ. Sci. Technol.* 37, 2400–2409.

Clara, M., Kreuzinger, N., Strenn, B., Gans, O., Kroiss H. (2005) The solids retention time-a suitable design parameter to evaluate the capacity of wastewater treatment plants to remove micropollutants. *Water Res.* 39, 97–106.

Culver, T.B., Hallisey, S.P., Sahoo, D., Deitsch, J.J., Smith J.A. (1997) Modeling the Desorption of Organic Contaminants from Long-Term Contaminated Soil Using Distributed Mass Transfer Rates. *Environ. Sci. Technol.* 31(6), 1581–1588.

Daughton, C.G. (2004) PPCPs in the environment: Future research—Beginning with the end always in mind. *Pharmaceuticals in the environmentsources, fate, effects and risks*. K. Kummerer, ed., Springer, New York, 463–495.

Eaton, A.D., Clesceri L.S., Greenberg A.E. (Eds.) (1995) Standard methods for the examination of water and wastewater (19th ed.). Washington, DC: American Public Health Association.

European Commission (EC) (2006) Registration, evaluation, and authorization of chemicals (REACH) program. EC No. 1907/2006

Golbraikh, A., Tropsha, A. (2002) Beware of q2! *Journal of Molecular Graphics and Modelling.* 20(4), 269–276.

Golet, E.M., Xifra, I., Siegrist, H., Alder, A.C., Giger, W. (2003) Environmental exposure assessment of fluoroquinolone antibacterial agents from sewage to soil. *Environ. Sci. Technol.* 37(15), 3243–3249.

Gute, B.D., Basak, S.C. (2006) Optimal neighbor selection in molecular similarity: comparison of arbitrary versus tailored prediction spaces. *SAR and QSAR in Environmental Research.* 17(1), 37-51

Hall , L.H., Kier, L.B. (1995) Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* 35(6), 1039–1045.

Heberer, T. (2002) Occurrence, fate, and removal of pharmaceutical residues in the aquatic environment, a review. *Toxicol. Lett.* 131, 5–17.

Kennedy, K.J., Lu, J., Mohn, W.W. (1992) Biosorption of chlorophenols to anaerobic granular sludge. *Water Research*. 26(8), 1085–1092.

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 U.S. streams, 1999–2000: A national reconnaissance. *Environ. Sci. Technol.* 366, 1202–1211.

Metcalf and Eddy (1991) Advanced wastewater treatment. In *Wastewater Engineering: Treatment, Disposal, Reuse*, Chap. 11. McGraw-Hill, Inc. New York, NY

Ottmar, K.J., Colosi, L.M., Smith, J.A. (2010) Development and application of a model to estimate wastewater treatment plant prescription pharmaceutical influent loadings and concentrations. *Bull. Environ. Contam. Toxicol.* 84, 507-512.

Pholchan, P., Jones, M., Donnelly, T., Sallis. P.J. (2008) Fate of estrogens during the biological treatment of synthetic wastewater in a nitrite-accumulating sequencing batch reactor. *Environ. Sci. Technol.* 42, 6141–6147.

Radjenovic, J., Petrovic, M., Barceló, D. (2009) Fate and distribution of pharmaceuticals in wastewater and sewage sludge of the conventional activated sludge (CAS) and advanced membrane bioreactor (MBR) treatment. *Water Res.*, 43, 831–841.

Richardson, S.D., Ternes. T.A. (2005) Water analysis: Emerging contaminants and current issues. *Anal. Chem.* 77(12), 3807–3838.

Schwarzenbach, P.M., Gschwend, D.M., (2003) Imboden Environmental Organic Chemistry Wiley Interscience, Hoboken, NJ.

Severtson, S.J., Banerjee, S. (1996) Sorption of chlorophenols to wood pulp. *Environ. Sci. Technol.* 30, 1961−1969.

Stevens-Garmona, J., Drewesa, J.E., Khanb, S.J., McDonaldb, J.A., Dickensona E.R.V. (2001) Sorption of emerging trace organic compounds onto wastewater sludge solids. *Water Research.* 45(11), 3417–3426

Ternes, T.A., Herrmanna, N., Bonerza, M., Knackerb, T., Siegristc, H., Joss, A. (2004) A rapid method to measure the solid–water distribution coefficient (K_d) for pharmaceuticals and musk fragrances in sewage sludge. *Water Research*. 38, 4075–4084.

Waller, C.L. (2004) A Comparative QSAR Study Using CoMFA, HQSAR, and FRED/SKEYS Paradigms for Estrogen Receptor Binding Affinities of Structurally Diverse Compounds. *J. Chem. Inf. Comput. Sci.* 44(2), 758–765.

Wick, A., Fink, G., Joss, A., Siegrist, H., Ternes, T. A. (2009) Fate of beta blockers and psycho active drugs in conventional wastewater treatment. *Water Res.* 43, 1060– 1074.

Chapter 4 – Molecular Similarity Analysis as a Tool to Predict Biodegradation Rate Constants for Emerging Contaminants in Wastewater

4.1 Introduction

In Chapter 3, we discussed the usefulness of QMSA to predict sorption distribution coefficient (K_d), an important environmental fate parameter, which can be used to predict removal of emerging contaminants in WWTPs. The internal accuracy of QMSA predictions for K_d was first examined by LOO, and the results indicated good internal accuracy, with $q^2 = 0.82$. $R = 0.91$, and S_{PRESS} as low as 0.41. External accuracy was then examined, by comparing newly measured K_d values for three emerging contaminants with their corresponding QMSA predictions. The results of this analysis show good consistency between measured values and QMSA-predicted values, especially for the PCA QMSA estimates.. Therefore, we demonstrated that QMSA can be used to accurately predict K_d.

In Chapter 3, it was noted that sorption and biodegradation are the two most important removal mechanisms for most emerging contaminants (Focazio et al., 2008; Glassmeyer et al., 2005). As such, it is desirable to measure or predict critical sorption and biodegradation parameters to understand the fate and behavior of emerging contaminants during wastewater treatment. Chapter 3 focused on sorption distribution coefficient; this chapter will focus on estimation of biodegradation rate constants.

Biodegradation, for the purposes of this dissertation, refers to a series of biological processes mediated by microorganisms (bacteria and fungi) in a WWTP setting, to transform, and thereby, eliminate aqueous phase emerging contaminants (EUR 20418 EN/2). Generally, the microorganisms "degrade" the emerging contaminants, that is, they break them down into smaller molecules, which can be consumed by the microorganism to produce energy and cellular biomass. The cellular biomass, which constitutes the microorganism themselves, can be separated from the treated effluent by sedimentation (Johnson and Sumpter, 2001). Previous studies on the biodegradation of emerging contaminants during typical, aerobic activated sludge treatment indicate that the removal of these compounds is highly variable. Some compounds are significantly removed (Martin Ruel et al., 2012; Choubert et al., 2011), while others pass through largely or completely unchanged and are therefore widely detected in the natural water bodies (Kahle et al., 2008; Lapertot et al., 2006; Muñoz and Guieysee, 2006).

Biodegradation has traditionally been described by two environmental engineering parameters: removal efficiency (Verlicchi et al., 2012), and biodegradation rate constant (Blair et al., 2013). Although removal efficiency, which encapsulates the difference in concentration between influent and effluent, as divided by either influent or effluent concentration, is more convenient to use, it is difficult to tell how much removal is caused by biodegradation alone. Furthermore, removal efficiency is highly correlated to WWTP operational parameters, such as hydraulic retention time (HRT), solids retention time (SRT), pH, temperature, etc (Verlicchi et al., 2012). This makes it difficult to apply removal efficiencies values from one study in one WWTP to any other scenario. In contrast, biodegradation rate constant (k_b) is a more generic parameter, which can be used to make biodegradation predictions in different WWTPs (Salgado et al., 2012; Thompson et al., 2011; Joss at al., 2006). As was discussed in Chapter 1, the very large (and increasing) number of emerging contaminants present in water systems makes it impractical to measure fate and behavior parameters for each individual compound; therefore, it is desirable to use numerical predictive tools, here QMSA, to estimate k_b (Dickenson et al., 2010).

Several recent studies have evaluated quantitative structure-activity relationship (QSAR) based prediction of k_b for emerging contaminants in WWTPs (Dickenson et al., 2010; Blair et al., 2013). These studies formulated QSAR models based on U.S. EPA's EPI Suite-BIOWIN modeling software. However, no comprehensive validation was performed in these studies, and the results were believed to be conservative (Dickenson et al., 2010). Additionally, many studies show that predictions made by BIOWIN are generally inaccurate (Tunkel et al. 2000; Yu et al. 2006). For example, Yu et al. (2006) applied BIOWIN 5 and concluded that models generally under predicted the likelihood of the biodegradation of Pharmaceuticals and Personal Care Products (PPCPs) in their dataset. However, BIOWIN 1 and 2 generally over predicted their biodegradation likelihood. Finally, this study did not attempt to incorporate their estimates of k_b into a larger WWTP mass balance model. Thus, the authors were unable to generate estimates of what typical effluent concentrations for these chemicals, even though this is nominally the goal of property estimation studies for emerging contaminants (Salgado et al., 2012; Wick et al., 2009; Joss et al., 2006; Joss et al., 2005).

This study has three objectives: 1) develop a QMSA model that generates accurate estimates of biodegradation rate constant (k_b) for sludge-water systems simulating secondary treatment in a typical municipal WWTP; 2) provide external validation of the k_B QMSA model by measuring pseudo first-order rate constants for three emerging contaminants which have never been measured before; and, 3) develop a simple mass balance model, incorporating K_d (from Chapter 3) and k_b (from this Chapter) as well as key operational parametrs, to predict effluent concentrations of the three selected emerging contaminants in a typical, representative WWTP.

4.2 Experimental

4.2.1 Data Sources

Pseudo first-order biodegradation rate constants (k_b) for 62 different chemicals were collected from literature sources: Majewsky et al. (2011), Li et al. (2010), Plosz et al. (2010), Suarez et al. (2010), Wick et al. (2009), Zeng et al. (2009), Maurer et al. (2007), Joss et al. (2006), Andreozzi et al. (2005), Urase et al. (2005) and Li et al. (2005). These studies report batch experiments using sludge-wastewater systems to mimic biodegradation process in WWTPs. In some instances, authors reported first-order biodegradation rate constants instead of pseudo-first order rate constants. For these date, the reported values were transformed to k_b by dividing by the total suspended solid concentrations (TSS) used in each study. For chemicals which had multiple reported k_b values, the arithmetic average of all reported values was used.

4.2.2 Molecular Descriptors

Molecular descriptors were computed using the procedure in Section 3.2.2.

4*.2.3 Similarity Computation, Property Estimation, and Internal Validation*

These calculations were performed according to the procedures in Section 3.2.3.

4.2.4 Selection of Nearest Neighbors, Property Estimation, and Model Validation

These calculations were performed according to the procedures in Section 3.2.4.

4.2.5 External Validation of the QMSA k_b Model

Laboratory experiments were used to measure k_b values for three previously unmeasured emerging contaminants. This was done to provide external validation of the QMSA k_b model, that is, to demonstrate that the model yields accurate k_b estimates for compounds that were not included in its original training set. As discussed in Section 3.2.5, metformin, fluconazole, and benazepril were chosen as the test chemicals, because they are among top 200 most-prescribed pharmaceuticals in the US (www.pharmacytimes.com); 3) their k_b values have been previously reported. Experimental procedures are summarized in the following paragraphs.

4.2.5.1 Materials and Chemical Reagents.

Necessary materials and chemical reagents are summarized in Section 3.2.5.1.

4.2.5.2 Batch Experiments for Determination of k_b

Batch experiments were used to measure biodegradation rate constants for the three selected emerging contaminants. Seven 1000-mL erlenmeyer flasks were used as reactors. These seven reactors consisted of: three regular reactors, two sorption controls, and two positive controls. For the three regular reactors, each contained 500 mL of a synthetic wastewater solution spiked with the target chemical. The detailed the recipe for the synthetic wastewater solution was obtained from Ottmar (2010).

The initial COD (chemical oxygen demand) of the synthetic wastewater was 1200 mg/L, and the initial concentration of the target chemical was 1 mg/L. Activated sludge biomass, from separate reactors, was also added to each flask, initially at a concentration of 1 g/L as total suspended solids (TSS). The sorption controls had the same contents as the regular reactors; however, the sludge in the sorption controls was first autoclaved at 120 °C for 4 h and then dosed with 2% (w/v) sodium azide to completely inactivate biological activity. The positive controls contained the same synthetic wastewater and target chemical concentrations, but without any sludge biomass.. All seven reactors were sparged with air to maintain appropriate dissolved oxygen concentrations greater than 2-3 mg/L at all times. A magnetic bar was also used to ensure complete mixing in each reactor. All reactors were monitored for pH, dissolved oxygen, and temperature every two days.

Timed sampling was performed to measure changes in the concentration of COD, TSS, and target emerging contaminant over time. For metformin, sampling times were t = 30 min, 6 h, 12 h, 21 h, 30 h, 42 h, 56 h, 66 h, 80 h, 104 h, and 134 h. For fluconazole, sampling times were t = 45 min, 6 h, 24.3 h, 49.5 h, 103.3 h, 144 h, and 220 h. For benazepril, sampling times were t = 50 min, 3.5 h, 6.5 h, 13.5 h, 20.3 h, 25.3 h, 41.5 h, 68.5 h, 95.5 h, 125.5 h, 147 h, and 219.5 h. Just before every sample was collected, make-up water (DI) was added to each reactor to compensate for evaporative losses. For each sampling time, a 5-mL aliquot was collected from each reactor. These were filtered through a pre-weighed 0.45-μm Whatman fiberglass filter, housed within a Buchner funnel system. The filters were used to measure TSS concentration according to Standard Method 2540D (Eaton et al., 1995). Then, 0.1-1 mL of each filtrate was transferred into a CHEMetrics COD reagent vial. These were used to measure COD concentration according to Standard Method 5220D. Another ~ 2 mL of each filtrate was transferred into glass vials for measurement of the target chemical concentration using an Agilent Technologies 1200 Series high pressure liquid chromatography (HPLC) system.

For the biodegradation experiments involving fluconazole and benazepril, approximately 2 mL of synthetic wastewater stock solution was added to each regular reactor every day to return the COD concentration to its initial value (1200 mg/L). This ensured that the sludge biomass always had an ample amounts of nutrients and energy source (i.e., COD). For the metformin biodegradation experiment, the re-dosing with synthetic wastewater stock was not performed. As such, the only COD available to the sludge biomass was that which was originally loaded into each reactor at time t=0. For all three experiments (metformin, fluconazole, and benazepril), 1 mL of a 20% (w/v) sodium azide solution was added to the sorption control daily, to maintain complete inactivation of biological activity.

4.2.5.3 HPLC Measurement

HPLC procedures for measurement of each selected emerging contaminant (metformin, fluconazole, and benazepril) are summarized in Section 3.2.5.4.

4.2.5.4 QMSA Predictions for the Selected External Validation Chemicals

The PCA, iPCA, and iRR QMSA procedures summarized in Section 3.2.2 were used to predict k_b values for the three selected external validation compounds.

4.2.5.5 Mass Balance Modeling of the External Validation Chemicals in a WWTP

A mass balance model was used to predict the effluent concentrations of the selected external validation compounds in a typical municipal WWTP. Figure 4.1 shows an illustration of this simple model. Sorption and biodegradation are presumed to be the principal removal mechanisms for emerging contaminants within the plant.

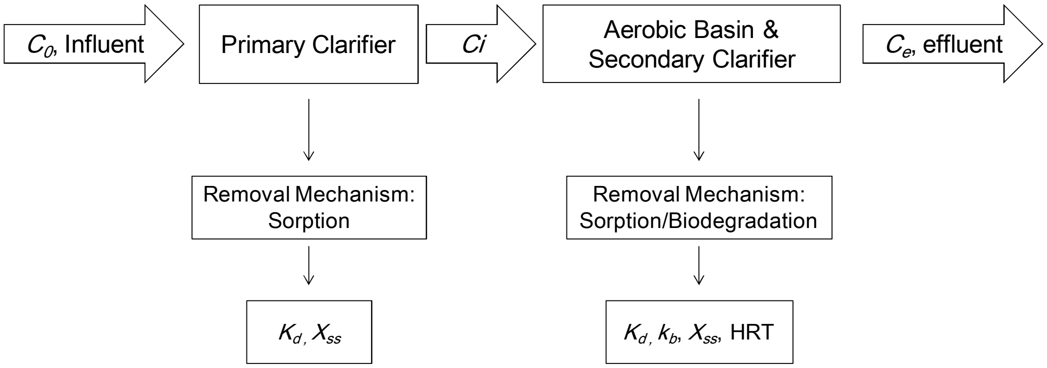
C_0, Influent
Primary Clarifier
Ci
Aerobic Basin & Secondary Clarifier
C_e, effluent
Removal Mechanism: Sorption
Removal Mechanism: Sorption/Biodegradation
K_d, X_{ss}
K_d, k_b, X_{ss}, HRT

Figure 4.1 The simple mass bass balance model for prediction three selected emerging contaminants' effluent concentrations out of WWTPs. C_0 is the influent concentration (μg/L); C_i is the effluent concentration from Primary Clarifier (μg/L); Ce is the effluent concentration from WWTPs (μg/L); X_{ss} represents the suspended solid concentrations in wastewater (g/L); HRT represents hydraulic retention time (h) in aerobic basin.

For the simplified mass balance model, it is assumed that the sorption distribution coefficient and sludge biomass concentrations in primary and secondary treatment, where sorption occurs, are known, constant values. We also assume sorption is the only removal mechanism for emerging contaminants in Primary Clarifier and sorption reaches equilibrium instantaneously. From Section 3.3.3, sorption can thus be written as follow in Primary Clarifier:

$$C_0 = (C_i + C_i K_d X_{ss}) \tag{4.1}$$

Here, C_i (μg/L) is aqueous phase concentration of the emerging contaminant out of the Primary Clarifier, and $C_iK_dX_{ss}$ is the sorbed concentration on the sludge (μg/L). C_0 is the influent concentration of the emerging contaminant (μg/L).

It is assumed that sorption and biodegradation are the two dominate removal mechanisms in Aerobic Basin and Secondary Clarifier. The whole processes are divided into two parts, first sorption and then biodegradation. This is reasonable because many studies demonstrated that sorption is fast compared to biodegradation, (Ternes et al., 2004; Wang and Grady, 1995). The sorption happened in Aerobic Basin and Secondary Clarified can be expressed as follow:

$$C_i = (C_t + C_t K_d X_{ss}) \tag{4.2}$$

where, C_t (μg/L) is aqueous phase concentration of the emerging contaminant at time *t,* and $C_tK_dX_{ss}$ is the sorbed concentration on the sludge (μg/L). C_i is the effluent concentration of the emerging contaminant out of Primary Clarifier (μg/L).

Likewise, it is also assumed that the pseudo first-order biodegradation rate constant and TSS concentrations in secondary treatment, where biodegradation occurs, are known, constant values. Thus, biodegradation can be expressed using the following pseudo first-order function:

$$\frac{dC_i}{dt} = -k_b X_{ss} C_t \tag{4.3}$$

Here, *t* is time (h), *k_b* is a pseudo first-order biodegradation rate constant (L/(kg·h)), and X_{ss} is the TSS concentration (kg/L). C_t is the concentration of each emerging contaminant at time *t* (µg/L).

Substituting Eq 4.2 into Eq 4.3, we have:

$$\frac{dC_t}{dt} = -\frac{k_b X_{ss}}{1+K_d X_{ss}} C_t \tag{4.4}$$

Then, integrating Eq 4.4, Eq 4.5 is obtained:

$$C_e = C_i e^{-HRT\frac{k_b X_{ss}}{1+K_d X_{ss}}} \tag{4.5}$$

Substituting Eq 4.1 into Eq 4.5, we have:

$$C_e = \frac{C_0}{(1+K_d X_{ss})} e^{-HRT\frac{k_b X_{ss}}{1+K_d X_{ss}}} \tag{4.6}$$

In above two equations, C_e is the effluent concentration of each emerging contaminant (µg/L), and HRT is the hydraulic retention time in the Secondary Treatment (Aerobic Basin and Secondary Clarifier) (h).

Eq 4.6 is the simple mass balance model (SM) for prediction of effluent concentration of emerging contaminants in WWTPs. However, since many assumptions and simplifications are required for this model to hold, a more detailed model was also used for comparison purposes. This complex model (CM) is based on Joss et al (2006). It has the following form:

$$C_e = C_i \times \frac{1+K_{d,p}X_{ss,p}}{1+K_{d,s}X_{ss.s}} \times \frac{1}{(1+R)\left[e^{HRT\frac{k_b X_{ss,s}}{(1+R)(1+K_d X_{ss,s})}}-1\right]+\frac{\left(1+K_{d,s}P_s\right)}{(1+K_{d,s}X_{ss,s})}} \qquad (4.7)$$

Here, C_i is the influent concentration (μg/L), C_e is the effluent concentration (μg/L), $K_{d,p}$ is the primary sorption distribution coefficient (L/kg), $K_{d,s}$ is the secondary sorption distribution coefficient (L/kg), $X_{ss,p}$ is the primary TSS concentration (kg/L), $X_{ss,s}$ is the secondary TSS concentration (kg/L), k_b is the pseudo first-order biodegradation rate constant (L/ (kg·h)), HRT is the hydraulic retention time (h) in secondary treatment, R is the ratio of $Q_{sludge\ recycle}$ to $Q_{wastewater}$ volumetric flow rates (dimensionless), and P_s is the specific sludge production per volume of wastewater treated (kg/L).

To test if the SM and CM models can accurately predict the effluent concentrations of emerging contaminants, predictions from each model were compared to several literature sources which have measured the influent and effluent concentrations of metformin and fluconazole in WWTPS. These studies are: Benotti and Brownawell (2007); Kahle et al. (2008). However, no available influent/effluent data in literatures can be found for benazepril.

4.3 Results and Discussion

In this chapter, the main objective is to demonstrate the usefulness of QMSA for prediction of a significant, fundamentaly engineering property, i.e, pseudo first-order biodegradation rate constant. Similar to Chapter 3, internal validation was first performed to test if the QMSA is internally accurate for the k_b dataset. The, external validity was examined by comparing QMSA predictions with measured k_b values that had heretofore not been measured.

4.3.1 Indices Selection for Development of PCA, iPCA, and iRR QMSA Models

For the k_b dataset, 18 statistically significant PCs were selected for use in PCA QMSA modeling, and 18 top-correlated indices were retained for iPCA QMSA modeling. Ridge regression was also applied to the same 193 indices. The 18 most-influential indices were retained for iRR QMSA to predict $k_{b.}$ Table 6.1 summaries the sets of indices retained for QMSA models.

From Table 4.1, and similar to the K_d dataset discussed in the last chapter, use of the iPCA or iRR indices selection approaches results in highly distinct sets of molecular descriptors for the same k_b dataset. There were two indices (SssCH2 and Pf) that appeared on the lists for both iPCA and iRR.

Table 4.1. Indices generated from iPCA and iRR data reduction for the k_b dataset. PCA-selected indices were used to construct both PCA and iPCA QMSA models. RR-selected indices were used to construct iRR QMSA models.

iPCA	iRR
dX0	nvx
dXvp3	nedges
mulrad	molweight
ishape	nX0
Pf*	nX1
WT	dXv0
Redundancy	dXv1
Qsv	dXv2
Qv	knotp
nwHBa	knotpv
etyp12	sumI
etyp33	Pf*
n2Pag11	IDWbar
n2Pag12	Si
n4Pae12	Gmax
nHCsatu	SssCH2*
naasC	SsssCH
SssCH2*	SssssC

* Single asterisks denote which two indices were selected by both iPCA and iRR for use in QMSA modeling to predict k_b.

4.3.2 Internal Validation of PCA, iPCA, and iRR QMSA Models.

Similar to Section 3.2.4, leave-one-out (LOO) validation was performed for each type of QMSA modeling. The results are illustrated in Figure 4.2. All three models exhibit excellent predictive abilities, as indicted by validation coefficient (q^2) values greater than 0.5 for all three models (Golbraikh and Tropsha, 2002). In all three cases, highest q^2 and R and lowest S_{PRESS} were achieved for k = 1. Similar to results in Chapter 2, PCA models seem to be the most powerful of the three data reduction strategies evaluated in this study. This approach yielded the highest q^2 (0.78),highest R (0.85), and lowest S_{PRESS} (0.45). his is perhaps because PCA models incorporate more chemical information than iPCA or iRR models, employing all principal components with eigenvalues equal or greater than 1. In contrast, iPCA and iRR models retain only a small portion of the available indices, and therefore may be missing some important similarity information. It also can be concluded that the iRR model performs better than the iPCA model for estimation of k_b, as indicated by the higher q^2 value of the iRR model.

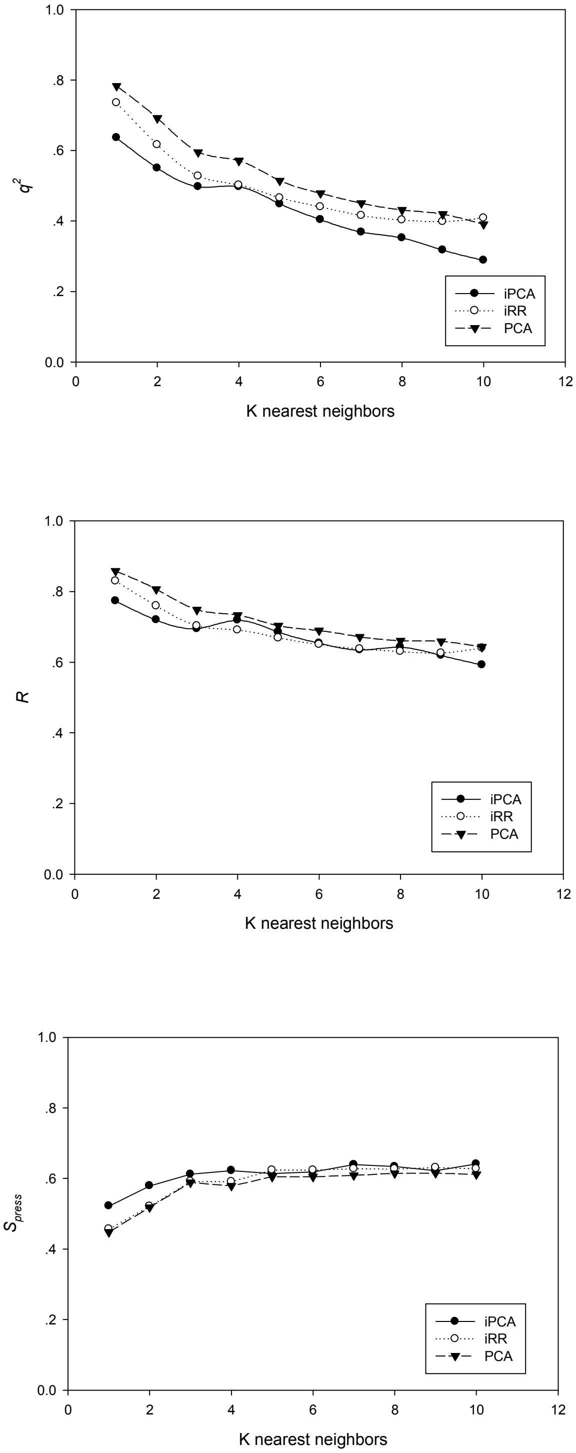

Figure 4.2. LOO validation statistics for QMSA models using three different indices selection approaches (PCA, iPCA, and iRR) to predict k_b From top, statistics are q^2, R, and S_{PRESS}.

Based on comparisons among multiple parameters, QMSA is less effective for estimation of k_b than estrogenicity (Chapter 2) or K_d (Chapter 3). This conclusion is based on comparison of highest q^2 values for each dataset. For k_b, the best q^2 value is 0.78, while q^2 values for the estrogenicity and K_d datasets were 0.84 and 0.82, respectively. This difference could be because the k_b dataset contains fewer measured chemicals (62 in total) compared to the other two datasets (81 for estrogenicity, and 80 for K_d). Regardless, the results from internal validation suggest that QMSA can serve as a powerful tool for prediction of the k_b parameter.

4.3.3 External Validation of the k_b QMSA Models

Another objective of this study was to show that QMSA models can be used to predict k_b of chemicals which have never been measured before; i.e., external validation. Thus, we measured k_b of three selected chemicals and compared the measured k_b values to their predicted k_b values The three compounds selected for use in external validation were metformin, fluconazole, and benazepril.

Figures 4.3 – 4.5 show the results of biodegradation experiments for the three selected chemicals. Figures 4.3A, 4.4A, and 4.5A show drug concentrations over time for at 5 days, which is much longer than the average WWTP retention time. As evident from Figures 4.3A, 4.4A, and 4.5A, aqueous-phase concentrations of the three selected chemicals decreased over time at different rates. Results from the positive controls indicate that the chemical concentrations did not change significant over time in the absence of sludge. Finally, the concentration profiles for the sorption controls are quite similar to the positive controls. This suggests that sorption could be very minimal for these three chemicals. This observation is consistent with previously published work that suggests that sorption of chemicals with K_d less than 300 L/kg is not significant (Joss et al., 2005). Recall from Chapter 3, that all three external validation chemicals exhibit K_d values that are less than 100 L/kg

Figure 4.3A also shows that metformin biodegradation seems to stop after 20 h. This observation could reflect the nature of the experimental design, whereby, COD was only dosed into the reactor at the very start of the experiment. There was no subsequent re-dosing. As such, the COD is mostly consumed within the same amount of time during which metformin removal is observed on Figure 4.3B This overlap between COD removal and emerging contaminant removal, is sometimes referred to as “co-metabolism” in the biodegradation literature, and it is a well-documented occurrence for emerging contaminant behavior in WWTPs (Ternes et al., 2004; Heberer, 2002; Ternes, 1998).

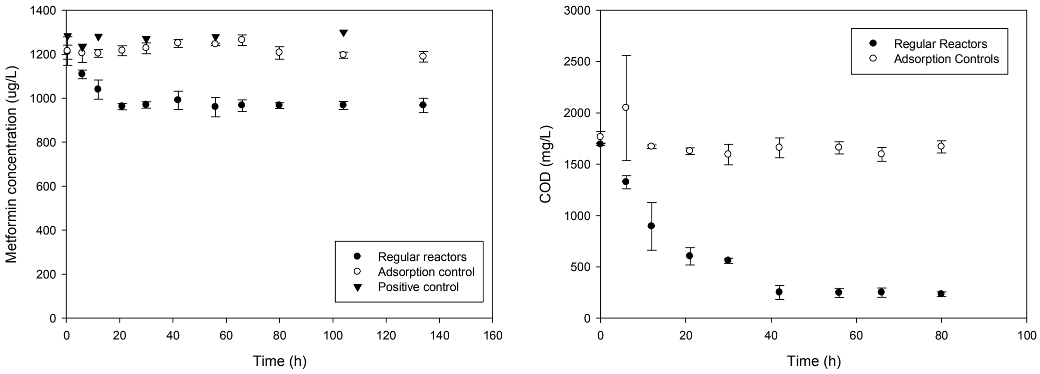

Figure 4.3. At left (4.3A), metformin concentration as a function of time during biodegradation experiment. At right (4.3B), corresponding COD concentrations as a function of time.

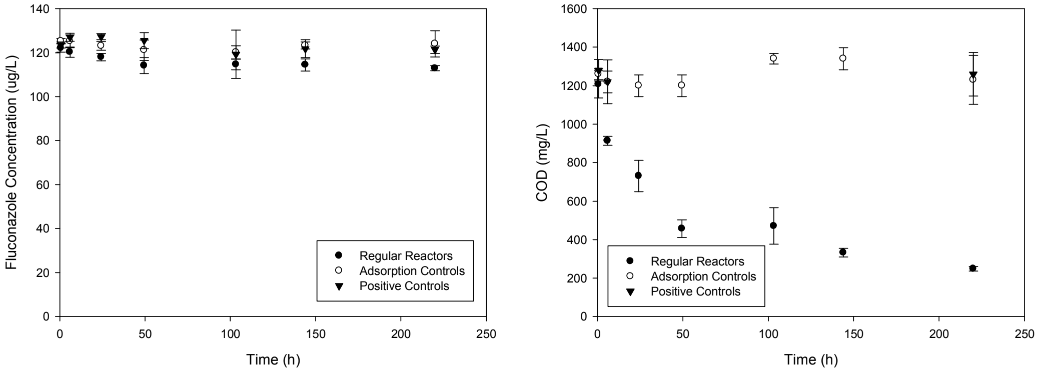

Figure 4.4. At left (4.4A), fluconazole concentration as a function of time during biodegradation experiment. At right (4.4B), corresponding COD concentrations as a function of time. Note synthetic wastewater was added to regular reactors on a daily basis

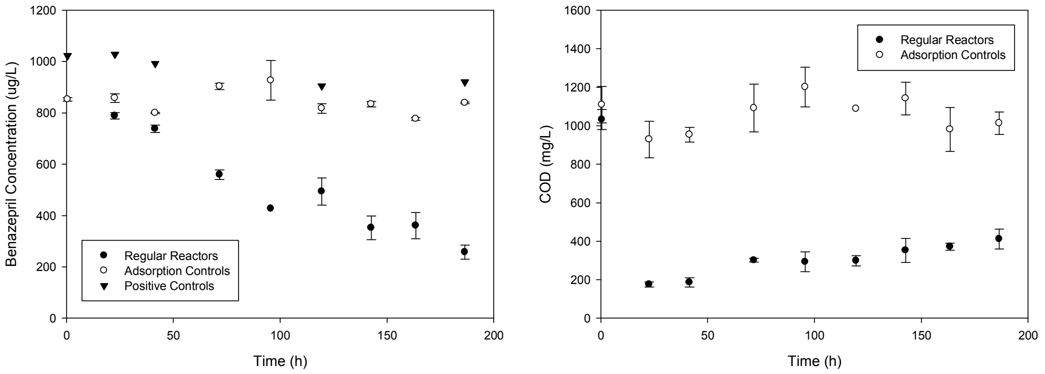

Figure 4.5. At left (4.5A), benazepril concentration as a function of time during biodegradation experiment. At right (4.5B), corresponding COD concentrations as a function of time. Note synthetic wastewater was added to regular reactors on a daily basis and therefore COD concentration showed some increases during the experiment.

Figures 4.3B, 4.4B, and 4.5B depict the change of COD over time for the three biodegradation experiments. We can conclude from these figures that there was minimal biological activity in the sorption controls, because none of these figures shows significant drop in COD over the measured timeframe. As such, the sorption controls can be used as references, to calculate what fraction of overall emerging contaminants removal in the regular reactors should be allocated to sorption versus biodegradation.

As discussed before, biodegradation of emerging contmainants has, in some cases, been well described by a pseudo first-order biodegradation function, as follows:

$$\frac{dC_t}{dt} = -k_b X_{ss} C_t \tag{4.8}$$

Integration of Eq 4.8, it yields:

$$\ln\frac{C_t}{C_0} = -k_b X_{ss} t \tag{4.9}$$

In Eq. 4.9, C_0 is the initial aqueous phase concentration (μg/L) of an emerging contaminant; C_t is the concentration at time t, assuming that the difference between C_0 and C_t is caused only by biodegradation (μg/L); k_b is a pseudo first-order biodegradation rate constant (L/(kg·h)); and X_{ss} is TSS concentration (kg/L).

Plotting Ln (C_t/C_0) against time for Eq. 4.9 yields a linear relationship with slope equal to $-k_bX_{ss}$. If we know the concentration change of each contaminant over time, we can use Eq. 8 to express decrease in concentration caused only by biodegradation:

$$C_t = \overline{C_{p,t}} - (\overline{C_{a,t}} - C_{r,t}) \tag{4.10}$$

where C_t is the decrease in emerging contaminant concentration caused by biodegradation (μg/L); $\overline{C_{p,t}}$ is the average concentration of the emerging contaminant in both positive controls at time t (μg/L); $\overline{C_{a,t}}$ is the average concentration of the emerging contaminant in both sorption controls at time t (μg/L); and, $C_{r,t}$ is the concentration in the average regular reactors at time t (μg/L). This formulation is plotted in Figures 4.6, 4.7, and 4.8 for each of the three selected external validation contaminants.

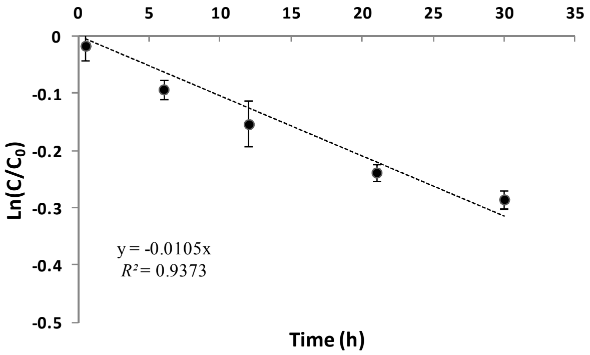

Figure 4.6. Calculation of k_b of metformin for the experiment without re-dosing of COD.

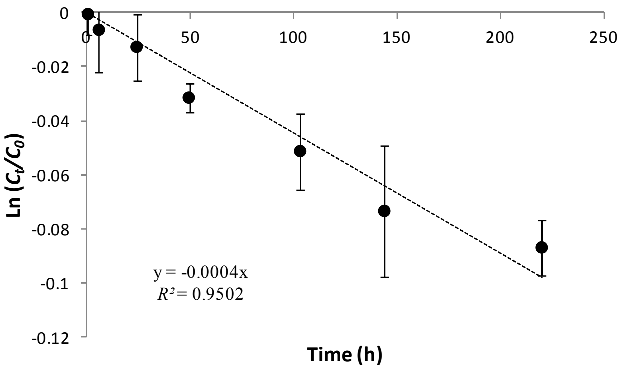

Figure 4.7. Calculation of k_b of fluconazole for the experiment with re-dosing of COD

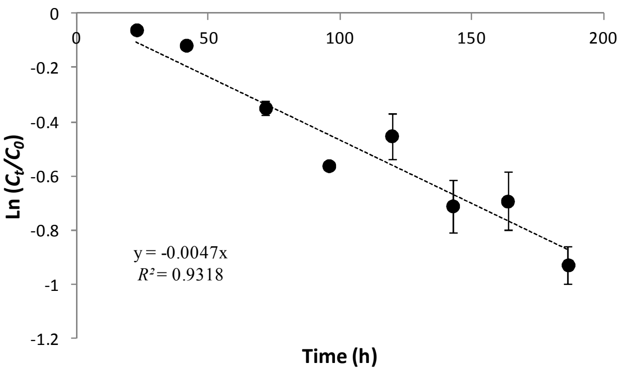

Figure 4.8. Calculation of k_b of benazepril for the experiment with re-dosing of COD.

Figures 4.6 – 4.8 show the results for the fitting of Eq. 4.10 to the data in each plot. It should be noted that only the first 30-h of data for metformin was used to make Figure 4.6., because it was observed that contaminant removal and COD removal stopped after 20 h during the metformin experiment, which did not utilize COD non-redosing. In all three figures, the data fit the linear models very well, as made evident via visual inspection and calculation of by high regression coefficients, R^2.

Another interesting, unexpected result from the fluconazole and benazepril biodegradation experiments, once it was decided that COD should be re-dosed into the reactors every day, is that the TSS concentrations in the regular reactors increases over the entire timeframe of the experiment. This did not occur during the biodegradation experiment for metformin, because there was no COD re-dosing TSS increased from 1 g/L to 3.4 g/L for benazepril, and from 1 g/L to 3.3 g/L for fluconazole. To account for this increase when computing k_b, the negative regression were divided by the average TSS concentration instead of the initial TSS concentration. Table 4.2 shows the calculated k_b and model fitting results from Eq. 4.9.

Table 4.2. Calculated k_b and model fitting parameters for three chemicals.

Chemical	k_b (L/(g·h))	Average TSS (g/L)	R^2
Metformin	1.05×10^{-2}	1.00	0.94
Fluconazole	1.86×10^{-4}	2.15	0.95
Benazepril	2.1×10^{-3}	2.20	0.93

4.3.4 Comparison of Measured and Predicted k_b for the Validation Chemicals

Table 4.3 – 4.5 summarizes k_b predictions for the three selected external validation chemicals, as computed using the PCA QMSA model. These predictions were computed using $k = 1 – 3$ measured nearest neighbors (which is slightly different than in Chapter 3), For k >1, the arithmetic average of the log k_b values of its nearest neighbors was computed and then log-transformed to yield the predicted log k_b value. Only PCA QMSA was used here because it was demonstrated as the most accurate data reduction approach in Section 3.3.2, and more importantly, because the size of k_b dataset is so much smaller than the K_d and estrogenicity datasets that use of less chemical information (as in iPCA or iRR) does not generate accurate estimates. Additionally, because the dataset is quite small, the "nearest" neighbor is frequently not that similar

Table 4.3. PCA Model predictions for metformin k_b based on its first three measured nearest neighbors. The laboratory-measured k_b value was 1.05×10^{-2} L/(g·h). *ED* is Euclidean distance, which denotes the distance between metformin and its neighbor in similarity space. Note when k = 1, the predicted value was obtained by using its first nearest neighbor's k_b directly.

k	Nearest neighbor with known k_b	*ED*	Literature reported k_b L/(g·h)	Predicted k_b L/(g·h)
1	2,4-Dichlorophenol	0.13	1.05×10^{-2}	1.05×10^{-2}
2	Paracetamol	0.15	2.87	0.17
3	4-Chlorophenol	0.18	0.13	0.16

Table 4.4. PCA Model predictions for fluconazole k_b based on its first five measured nearest neighbors. The laboratory-measured k_b value was 1.86×10^{-4} L/(g·h). *ED* is Euclidean distance, which denotes the distance between fluconazole and its neighbor in the similarity space. Note when k = 1, the predicted value was obtained by using its first nearest neighbor's k_b directly.

k	Nearest neighbor with known k_b	*ED*	Literature reported k_b L/(g·h)	Predicted k_b L/(g·h)
1	Diclofenac	0.18	8.30×10^{-4}	8.30×10^{-4}
2	Carbamazepine	0.24	4.16×10^{-4}	5.88×10^{-4}
3	Primidone	0.24	9.16×10^{-5}	3.16×10^{-4}

Table 4.5. PCA Model predictions for benazepril k_b based on its first three measured nearest neighbors. The laboratory-measured k_b value was 2.1×10^{-3} L/(g·h). *ED* is Euclidean distance, which denotes the distance between benazepril and its neighbor in the similarity space. Note when k = 1, the predicted value was obtained by using its first nearest neighbor's k_b directly.

k	Nearest neighbor with known k_b	*ED*	Literature reported k_b L/(g·h)	Predicted k_b L/(g·h)
1	Celiprolol	0.17	8.75×10^{-3}	8.75×10^{-3}
2	Diltiazem	0.19	1.01×10^{-2}	9.40×10^{-3}
3	Primidone	0.45	9.16×10^{-5}	2.00×10^{-3}

For metformin (Table 4.3), PCA-based QMSA yields excellent predictions for k_b. For k = 1, PCA gives the exactly same predicted value as was measured in the laboratory experiments. Therefore the prediction error (difference between predicted value and measured value divided by measured value) equals to zero. The model assigns 2,4-dichlorophenol as the nearest neighbor for metformin, which is same measured nearest neighbor selected by the K_d QSAR model that was discussed in Chapter 3. Published k_b values for 2,4-dichlorophenol include 0.80×10^{-2} L/(g·h) (Elkarmi et al., 2009) and 1.3×10^{-2} L/(g·h) (Tomei et al., 2012). Because these values are so close on either side of the k_b value that was measured for metformin during laboratory experiments, there was very little benefit to adding additional nearest neighbors. This is why the $k = 1$ model is better than the other two.

For fluconazole, the PCA-based QMSA predictions are generally acceptable, since predicted values and measured values are within the same order of magnitude. The best prediction was achieved for $k = 3$, which yields a prediction error of 71%. Fluconazole's first measured nearest neighbor is diclofenac, which was also found to be its second measured nearest neighbor in Chapter 3. However, diclofenac's k_b value (8.30×10^{-4} L/(g·h) (Fernandez-Fontaina et al, 2013) is almost four times greater than fluconazole's k_b value. Thus, the relative inaccuracy of the fluconazole predictions compared to the metformin predictions can be attributed to fluconazole's much larger degree of dissimilarity with its first measured nearest neighbor. The model is unable to yield a good prediction for fluconazole because there are no similar chemicals in the pool of previously measured structures. Ultimately, this occurs probably because biodegradation of fluconazole is very slow. In a previously published study, Kahle et al. (2008) concluded that fluconazole shows no biodegradation at all within first 24 h. The timeframe used our study was much longer 220 h (~9 days), yet we still saw only very limited fluconazole removal. Because the change in concentration was so small, the calculation of regression slope was very sensitive to slight variability among measured replicates. Thus, it is not unsurprising that the k_b fit for fluconazole is not very good. More broadly, it should be emphasized that QMSA will perform poorly when it comes to making predictions for chemicals exhibiting extreme property values; i.e., property value that are much greater or much less than the values of "most" other chemicals. This is because the chemicals exhibiting extreme property values cannot, by definition, have many similar neighbors.

As for fluconazole, the k = 3 PCA-based QMSA model gives the most accurate predictions for benazepril. The relative error of this estimate was 5%. However, this prediction incorporates benazepril's third measured nearest neighbor, primidone, which is much more dissimilar from benazepril compared to its two closest neighbors. This is evident from the ED values in Table 4.5. The inclusion of primidone's much smaller k_b value brings down the average of the first two nearer neighbors, resulting in a very good predicted value.

Generally speaking, the QMSA-based k_b predictions for the three selected external validation chemicals are very good, especially for metformin. These results are very similar to what was observed for K_d predictions using these same three chemicals in Chapter 3. This overlap in accuracy between K_d and k_b reflects the underlying structure of each dataset, whereby metformin has the nearest measured neighbor out of the three validation chemicals, such that it is best-predicted for both parameters of interest. To reiterate from Chapter 3, *EDs* magnitudes should served as a guide for selecting the best possible nearest neighbors.

4.3.5 Prediction of effluent concentrations of three selected chemicals in WWTPs.

As mentioned several times now, a key objective of studies measuring or predicting environmental fate parameters, such as K_d and k_b, is to predict effluent concentrations for real-world settings. Therefore, we used two mass balance models (SM and CM), to calculate effluent concentrations for the validation chemicals.

Table 4.6 summarizes the essential parameters required for estimating effluent concentrations of metformin and fluconazole using the SM and CM models. Since no literature data can be found for both influent and effluent concentrations of benazepril in WWTPs, mass balance calculation was not performed for benazepril. Several assumptions were required for the CM calculations. We assume the secondary TSS concentration is the same as the primary TSS concentration (i.e., $X_{ss,p} = X_{ss,s}$), and that the K_d of the secondary sludge is the same as the K_d of the primary sludge (i.e , $K_{d,s} = K_{d,p}$). Also because some of the literature studies that provided influent and effluent concentrations failed to provide WWTP operational parameters, some typical, default values were used (Joss et al., 2006; Kim et al., 2005).

Table 4.7 summarizes the predicted effluent concentrations arising from both models. The results are generally very good. Most values of absolute relative error (ARE) are less than 20% for the various predictions. The SM model seems to have better predictions, based on generally lower ARE values for SM compared to CM. The difference in accuracy between the two models is more significant for metformin, whereby SM gives ARE = 7%, compared to ARE = 31% for CM.

Table 4.6. WWTP operational parameters for effluent concentration calculation

SM model parameters[a]	Definition	Value	CM model parameters[a]	Definition	Value
X_{ss}	Total Suspended Solid	3 g/L	R	Sludge Recycle	2
R_t	Retention Time	24 h	P_s	Sludge Production	0.03 g/L
			X_{ss}	Total Suspended Solid	3 g/L
			R_t	Retention Time	24 h

[a] Values were taken from typical values reported in literatures (Joss et al., 2006; Kim et al., 2005)

Table 4.7. Prediction of effluent concentrations of metformin and fluconazole in WWTPs by SM and CM. WWTP operational parameters required in both models were taken from Table 4.6.

Chemical	Effluent (ng/L)	SM Predicted Effluent (ng/L)	SM ARE (%)[c]	CM Predicted Effluent (ng/L)	CM ARE (%)[d]
Metformin[a]	11000	11731	7	14483	31
Fluconazole[b]	28	28	0	33	20
Fluconazole[b]	83	73	12	87	5
Fluconazole[b]	52	74	43	89	72
Fluconazole[b]	48	50	5	61	26
Fluconazole[b]	39	38	3	46	17
Fluconazole[b]	63	32	49	39	38
Fluconazole[b]	35	90	157	108	209
Fluconazole[b]	42	35	17	42	1
Fluconazole[b]	36	26	27	32	12
Fluconazole[b]	67	45	33	54	20
Fluconazole[b]	49	49	0	33	20

[a] Benotti and Brownawell, 2007; [b] Kahle et al., 2008; [c] stands for Simple Model Absolute Relative Error, calculated by absolute difference between predicted and actual effluent concentration divided by actual effluent concentration; d stands for Complex Model Absolute Relative Error.

There are several important comments that can be made regarding the effluent concentration predictions arising from both models. First, in this study, we used default values for essential WWTP operational parameters because precise information could not be obtained from the literature studies themselves. However, in reality, these parameters may be very different from what we assumed here. For example, for fluconazole, influent and effluent samples were collected from ten

WWTPs in Switzerland (Kahle et al., 2008). The operational parameters may be also very different among these ten WWTPs. To test this idea, we performed sensitivity analysis for two different operational parameters, HRT and X_{ss}. The metformin data was used for this analysis. The HRT parameter was varied over the range 12 to 48 h, and the X_{ss} parameter was varied over the range 1 to 10 g/L.

Figures 4.9 and 4.10 show the results of the sensitivity analysis. As expected, these two parameters have significant impacts on ARE values for SM and CM predictions. For HRTs, approximately 26 h and 40 h are required to provide the exactly same predicted effluent concentration as the actual effluent concentration for SM and CM, respectively (Figure 4.9). These HRTs would be possible for some WWTPs (Oppenheimer et al., 2007). Additionally, X_{ss} seems to have a more prominent impact on model sensitivity. Only by changing X_{ss} from the default 3 g/L value to approximately 3.4 g/L, is it possible for the SM model to give ARE = 0. For CM, the X_{ss} must be changed to 5 g/L to yield ARE = 0. All of these X_{ss} values are typical for WWTPs (Wick et al., 2009); therefore it is highly desirable to obtain accurate, plant-specific operational parameters if valid effluent concentrations are desired.

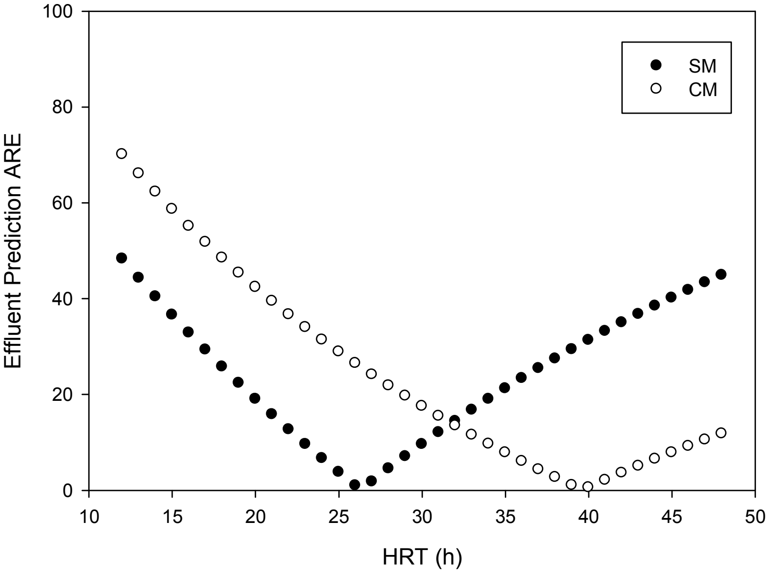

Figure 4.9. Model sensitivity analysis of metformin effluent prediction ARE on HRT.

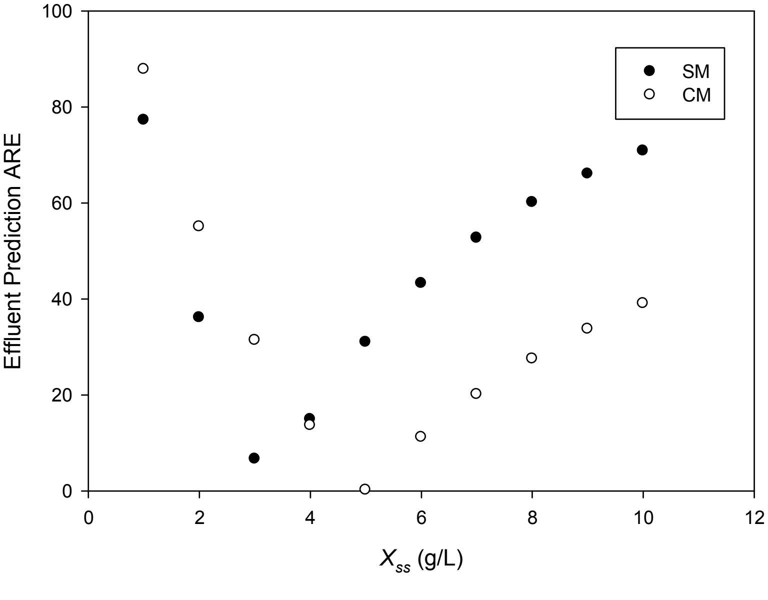

Figure 4.10. Model sensitivity analysis of metformin effluent prediction ARE on TSS.

Also related to Table 4.6 and 4.7, it has been demonstrated that many WWTP sampling studies fail to apply proper sampling strategies. As a result, short-term variation in influent and effluent concentrations may not be well captured. This will lead to the inaccurate estimation of effluent concentrations. Ort et al. (2010) reviewed 87 peer-reviewed articles and concluded that over 95% of these may not have utilized proper sampling campaigns and that the resulting sampling errors can potentially lead to wrong conclusions on effluent concentrations. They also recommend several specific measures to improve the quality of influent and effluent concentration data collected at a typical WWTP. Neither of the two articles we used as sources of information for influent/effluent concentrations of the external validation chemicals used the comprehensive sampling techniques recommended by Ort et al. Thus, it is highly possible that the measured influent/effluent concentrations in Table 4.7 not reflect actual real-time concentrations. This could also cause difference between our model predictions and measured concentrations. In the future, more comprehensive sampling of the three selected validation compounds could help evaluate the accuracy of our SM and CM predictions.

4.4 Conclusions

The results from this chapter provide additional, clear evidence that QMSA models can be used as powerful tools for prediction of significant environmental properties for the numerous emerging contaminants of interest. First and foremost, this type of structural similarity models can accurately predict k_b for large, highly diverse classes of chemical structures. This is proven by the excellent internal-validation q^2 values and the external prediction results for three emerging contaminants which have never been measured before. The choice of k_b as a test parameter for QMSA is especially relevant, because it is one of the most significant factors impacting the removal of emerging contaminants in WWTPs. It is also one of the most difficult factors to be measured in the laboratory, when considering costs and time invested for measurement.

The results from this chapter also demonstrate that the fate of two selected emerging contaminants, metformin and fluconazole, can be well predicted using both SM and CM. However, the values of key WWTP operational parameters may have a great impact on the accuracy of predicted effluent concentrations. Therefore it is essential to obtain WWTP-specific operational parameters to make calculation more accurate. To further test the model, it would be greatly preferable to perform more comprehensive sampling in the future.

4.5 References

Andreozzi, R., Cesaro, R., Marotta, R., Pirozzi. F. (2006) Evaluation of biodegradation kinetic constants for aromatic compounds by means of aerobic batch experiments. *Chemosphere*. 62(9), 1431–1436.

Benotti, M.J., Brownawell, B.J. (2007) Distributions of pharmaceuticals in an urban estuary during both dry- and wet-weather conditions. *Environ. Sci. Technol.* 41(16), 5795–5802.

Blair, B.D., Crago, J.P., Hedman, C.J., Treguer, R.J.F., Magruder, C., Scott Royer, L., Klaper, R.D. (2013) Evaluation of a model for the removal of pharmaceuticals, personal care products, and hormones from wastewater, *Science of The Total Environment.* 444(1), 515-521.

Carballa, M., Omil, F., Lema, J.M., Llompart, M., Garcia-Jares, C., Rodriguez I., et al. (2004) Behavior of pharmaceuticals, cosmetics and hormones in a sewage treatment plant. *Water Res*, 38 (12), 2918–2926.

Choubert, J. M., Martin Ruel, S., Esperanza, M., Budzinski, H., Miège, C., Lagarrigue C., Coquery M. (2011) Limiting the emissions of micro-pollutants: what efficiency can we expect from wastewater treatment plants? *Water Science & Technology*. 63(1), 57–65.

Dickenson, E., Drewes, J., Stevens-Garmon, J., Khan, S., McDonald J. (2010) Evaluation of QSPR Techniques for Wastewater Treatment Processes. *Water Environment Research Foundation*, Alexandria, VA.

Eaton, A.D., Clesceri L.S., Greenberg A.E. (Eds.) (1995) Standard methods for the examination of water and wastewater (19th ed.). Washington, DC: American Public Health Association.

Elkarmi, A.Z., Abu-Elteen, K.H., Atta, A.A., Abu-Sbitan, N.A. (2009). Biodegradation of 2,4-dichlorophenol originating from pharmaceutical industries. *Afri. J. Biotechnol.* 8(11), 2558 – 2564.

Fernandez-Fontaina, E., Pinho, I., Carballa, M., Omil, F., Lema, J.M. (2013) Biodegradation kinetic constants and sorption coefficients of micropollutants in membrane bioreactors. *Biodegradation.* 24(2), 165-177.

Focazio, M.J., Kolpin, D.W., Barnes, K.K., Furlong, E.T., Meyer, M.T., Zaugg, S.D., et al. (2008) A national reconnaissance for pharmaceuticals and other organic wastewater contaminants the United States — II untreated drinking water sources. *Sci Total Environ.* 402, 201–16.

Glassmeyer, S.T., Furlong, E.T. Kolpin, D.W. Cahill, J.D. Zaugg, S.D. Werner, S.L. Meyer M.T. Kryak., D.D. (2005) Transport of chemical and microbial compounds from known wastewater discharges: Potential for use as indicators of human fecal contamination. *Environmental Science & Technology.* 39:5157-5169.

Golbraikh, A., Tropsha, A. (2002) Beware of q2! *Journal of Molecular Graphics and Modelling.* 20(4), 269–276.

Heberer, T. (2002) Occurrence, Fate, and Removal of Pharmaceutical Residues in the Aquatic Environment: A Review of Recent Research Data. *Elsevier Toxicology Letters.* 131, 5-17.

EUR 20418 EN/2. Institute for Health and Consumer Protection, European Chemicals Bureau, European Commission. EUR 20418 EN/2. (2003) Technical Guidance Document on Risk Assessment Part II.

Johnson A.C., Sumpter, J.P. (2001) Removal of endocrine-disrupting chemicals in activated sludge treatment works. *Env. Sci. Techn.* 35(24), 4697–4703.

Joss, A., Zabczynski, S., Göbel, A., Hoffmann, B., Löffler, D., McArdell, C.S., Ternes, T.A., Thomsen, A., Siegrist, H. (2006) Biological degradation of pharmaceuticals in municipal wastewater treatment: Proposing a classification scheme. *Water Research*. 40(8), 1686-1696.

Joss, A., Keller, E., Alder, A., Göbel, A., McArdell, C.S., Ternes, T., Siegrist, H. (2005) Removal of pharmaceuticals and fragrances in biological wastewater treatment. *Water Research.* 39, 3139 – 3152.

Kahle, M., Buerge, I.J., Hauser, A., Muller, M.D., Poiger, T. (2008) Azole fungicides: occurrence and fate in wastewater and surface waters. *Environmental Science & Technology*, 42(19), 7193–7200.

Kim, S., Eichhorn, P., Jensen, J.N., Weber, A.S., Aga, D.S. (2005) Removal of antibiotics in wastewater: effect of hydraulic and solid retention times on the fate of tetracycline in the activated sludge process. *Environ. Sci. Technol.* 39, 5816–5823.

Lapertot, M., Pulgarin, C., Fernandez-Ibaznez, P., Maldonado, M.I., Perez-Estrada, L., Oller, I., Gernjak, W., and Malato, S. (2006). Enhancing biodegradability of priority substances (pesticides) by solar photo-Fenton. *Water Res.* 40, 1086.

Li, B., Zhang. T. (2010) Biodegradation and Adsorption of Antibiotics in the Activated Sludge Process. *Environmental Science & Technology*. 44(9), 3468-3473.

Li, F., Yuasa, A., Obara, A., Mathews. A.P. (2005) Aerobic batch degradation of 17-β estradiol (E2) by activated sludge: Effects of spiking E2 concentrations, MLVSS and temperatures. *Water Research*. 39(10), 2065–2075.

Muñoz, R., Guieysse, B. (2006) Algal-bacterial processes for the treatment of hazardous contaminants: a review. *Water Research.* 40: 2799-2815.

Majewsky, M., Gallé, T., Yargeau, V., Fischer. K. (2011) Active heterotrophic biomass and sludge retention time (SRT) as determining factors for biodegradation kinetics of pharmaceuticals in activated sludge. *Bioresource Technology*. 102(16), 7415–7421.

Martin Ruel, S., Choubert, J.-M., Budzinski, H., Miège, C., Esperanza M., Coquery M. (2012) Occurrence and fate of relevant substances in wastewater treatment plants regarding Water Framework Directive and future legislations. *Water Science & Technology*. 65(7), 1179–1189.

Maurer, M., Escher, B.I., Richle, P., Schaffner, C., Alder., A.C. (2007) Elimination of β-blockers in sewage treatment plants. *Water Research*. 41(7), 1614–1622.

Oppenheimer, J., Stephenson, R., Burbano, A., Liu, L. (2007) Characterizing the passage of personal care products through wastewater treatment processes. Water Environ Res. 79(13), 2564–2577.

Ort, C., Lawrence, M.G., Rieckermann, J., Joss, A. (2010) Sampling for Pharmaceuticals and Personal Care Products (PPCPs) and Illicit Drugs in Wastewater Systems: Are Your Conclusions Valid? A Critical Review. *Environ. Sci. Technol.* 44(16), 6024–6035.

Plosz, B.G., Leknes, H., Liltved, H., Thomas, K.V. (2010) Diurnal variations in the occurrence and the fate of hormones and antibiotics in activated sludge wastewater treatment in Oslo, Norway. *Sci Total Environ*. 408, 1915–1924.

Salgado, R. Marques, R. Noronha, J. Carvalho, G. Oehmen, A. Reis M. (2012) Assessing the removal of pharmaceuticals and personal care products in a full-scale activated sludge plant. *Environ Sci Pollut Res*, 19 (5), 1818–1827.

Suarez, S., Lema, J.M., Omil, F. (2010) Removal of pharmaceutical and personal care products (PPCPs) under nitrifying and denitrifying conditions. *Water Research.* 44, 3214 – 3224.

Ternes, T.A., Herrmanna, N., Bonerza, M., Knackerb, T., Siegristc, H., Joss, A. (2004) A rapid method to measure the solid–water distribution coefficient (K_d) for pharmaceuticals and musk fragrances in sewage sludge. *Water Research*. 38, 4075–4084.

Ternes, T.A. (1998) Occurrence of drugs in German sewage treatment plants and rivers. *Water Res.* 32(11), 3245–3260.

Thompson, J., Eaglesham, G., Mueller J. (2011) Concentrations of PFOS, PFOA and other perfluorinated alkyl acids in Australian drinking water. *Chemosphere*. 83, 1320–1325.

Tomei C.M., Annesini, M.C., Daugulis, A.J. 2,4-Dichlorophenol removal in a solid–liquid two phase partitioning bioreactor (TPPB): kinetics of absorption, desorption and biodegradation. *New Biotechnology.* 30(1), 44–50.

Tunkel, J., Howard, P.H., Boethling, R.S., Stiteler, W., Loonen, H. (2000) Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry Test Environ. *Toxicol. Chem.* 19, 2478–2485.

Urase, T., Kikuta. T. (2005) Separate estimation of adsorption and degradation of pharmaceutical substances and estrogens in the activated sludge process. *Water Research*. 39(7), 1289–1300.

Verlicchi, P., Al Aukidy, M., Zambello, E. (2012) Occurrence of pharmaceutical compounds in urban wastewater: Removal, mass load and environmental risk after a secondary treatment-A review. *Sci Total Environ.* 429, 123-155.

Wang, X., Grady, C.P.L. (1995) Effects of biosorption and dissolution on the biodegradation of di-n-butyl phthalate. *Water Environ. Res.* 67, 863–871.

Wick, A., Fink, G., Joss, A., Siegrist, H., Ternes, T.A. (2009) Fate of beta blockers and psycho-active drugs in conventional wastewater treatment. Water Research. 43(3), 1060-1074.

Yu, J.T., Bouwer, E.J., Coelhan, M. (2006) Occurrence and biodegradability studies of selected pharmaceuticals and personal care products in sewage effluent. *Agricultural Water Management*, 86 (1–2), 72–80.

Zeng, Q. Li, Y. Gu, G. Zhao, J. Zhang, C. Luan J. (2009) Sorption and Biodegradation of 17b-Estradiol by Acclimated Aerobic Activated Sludge and Isolation of the Bacterial Strain. *Environmental Engineering Science*. 26(4), 783-790.

Chapter 5 – Improving the Usefulness of Molecular Similarity-Based Chemical Prioritization Strategies

5.1 Introduction

In Chapter 2, 3, and 4, we demonstrated that QMSA can potentially serve as an efficient alternative to traditional, regression-based QSAR models for prediction of relevant environmental engineering parameters. Specifically, we showed that QMSA can be used to estimate *in vitro* estrogenicity with q^2 values as high as 0.84 (Section 2.3.2).

As discussed in Chapter 3 *ED* serves as a ruler to parameterize and quantify the similarities among chemicals. Chemicals which have smaller *ED* to the "target" are considered as the nearest neighbors for "target" chemicals. Therefore, it can be expected that the QMSA model accuracy would be highly dependent on the extent to which previously measured chemicals are similar to the "target" chemicals. Although it is good to find the nearest possible neighbors for each individual "target", we hypothesized that there should theoretically be some "representative" chemicals which are generally similar to many unmeasured chemicals. Identification and measurement of the desired property values for these chemicals could add valuable data to the QMSA model by inserting molecules that should be near neighbors to many other unmeasured chemicals.

Figure 5.1 depicts a two-dimensional (2-D) conceptualization of the scenario motivating the representativeness prioritization hypothesis articulated in the previous paragraph. For this hypothetical scenario, molecular similarity is projected into two dimensions. Thus, *n*-dimensional Euclidean Distances can be represented as the geometric (2-D) distance between two points. Numbered (unshaded) circles represent the limited number of chemicals for which a parameter of interest has been previously *measured* and reported in the scientific literature. Lettered (shaded) circles represent as yet *unmeasured* chemicals that are candidates for addition to the measured pool.

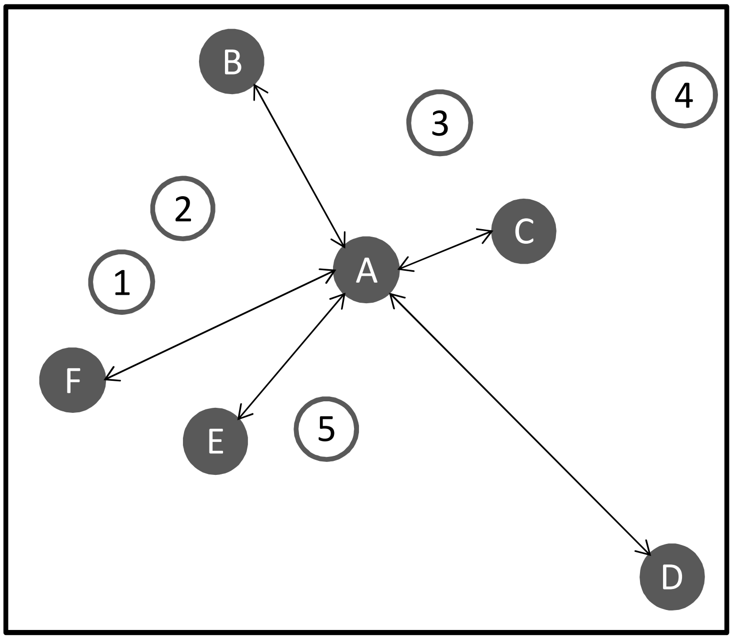

Figure 5.1. A two-dimensional depiction of the way in which QMSA is envisioned to prioritize among unmeasured compounds and subsequently streamline development of increasingly accurate models to predict environmental fate and behavior properties of emerging contaminants.

In Figure 5.1, there are eight candidate chemicals, A – F, that could be added to the existing data pool. In considering candidate D, it becomes clear that this compound will add practically zero predictive accuracy to a QMSA model, because it is completely dissimilar from the rest of the chemicals for which information is sought. This makes Candidate D a poor choice for additional research. Candidate E, which is also located towards the edge of the chemical space, and thus quite dissimilar from the other unmeasured compounds, could offer predictive insight into Chemicals F; however, this information is both somewhat redundant with and less helpful than information derived from Chemical 1 or Chemical 5. Thus, Candidate E would be a better choice than Candidate D, but is still suboptimal compared to other candidates. Using this same logic, Candidate A is clearly the best choice for addition to the pool of measured chemicals. This is because information about Candidate A will be directly useful in formulating predictions for Chemicals C, E, and, F. Since Candidate A is located almost directly at the center of the chemical space, it exhibits a "typical" or "representative" value. This means that information from Candidate A can also be used to make rough predictions about Chemicals B and D or any other new contaminants that come to light anywhere in this molecular space.

Numerically, it can be said that Candidate A is the optimal choice for addition to the measured data pool because it is minimizes the sum of two-dimensional distances with each other unmeasured point in the dataset. Analogously, maximally representative chemicals in an *n*-dimensional space are those that minimize the sum of Euclidean Distances between each unmeasured chemical and all other unmeasured compounds. This suggests that most representative unmeasured chemicals selected by QMSA should result in formulation of more accurate QMSA models than the random *ad hoc* approach which has been traditionally used to identify which chemicals should be measured in a laboratory or field experiment. Given the large amounts of time and money required to make laboratory measurements for emerging contaminants, a regulatory agency would like to ensure that resources devoted to elucidating the environmental fate and behavior of emerging contaminants are expended in an optimally efficient manner. Thus, there is clear incentive to prioritize candidates experimental testing based on representativeness, such that each new data point can help improve prediction accuracy for many other unmeasured chemicals.

A key question arising from the previously articulated "maximizing representativeness" argument is, "Should "representative" chemicals always be valuable to incorporate into a QMSA model? Figure 5.2 illustrates the conceptual model for another scenario that should be considered. In Figure 5.2, previously measured compounds are shown using un-shaded, numbered circles. Unmeasured compounds are shown using shaded, lettered circles. For this example, Chemical J would be selected on the basis of maximized representativeness. However, addition of Chemical J to the measured dataset is not ideal choice because it adds little new information that is not redundant with what could be inferred from chemicals 6-10. In contrast, Chemical G exhibits the least amount of redundancy among all possible candidate compounds to be measured. This is evident in 2D because Chemical G has the largest sum of distances with all previously measured chemicals

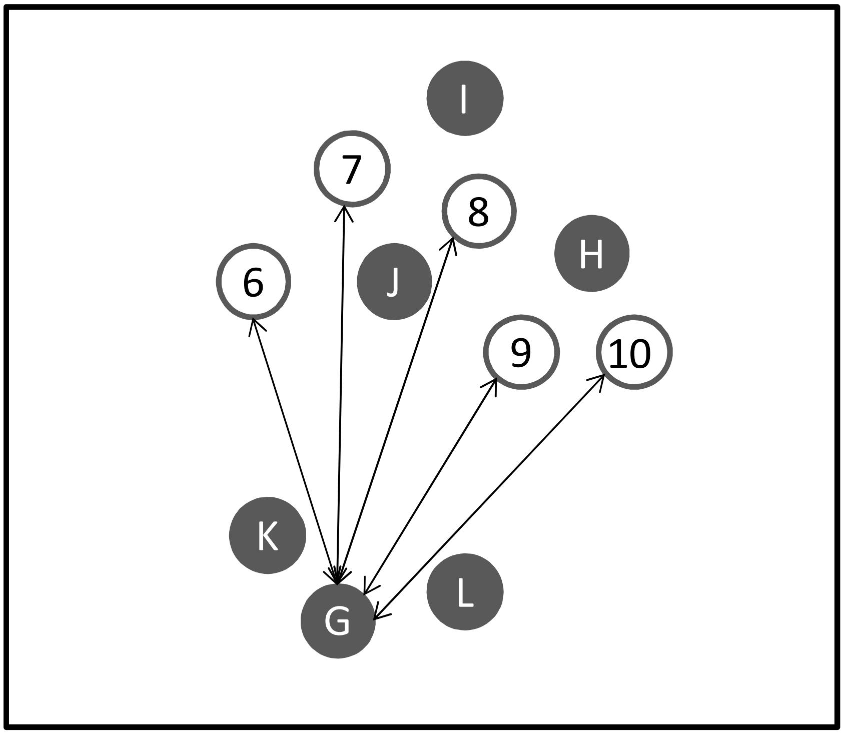

Figure 5.2. A 2-D illustration of the QMSA *redundancy minimization* principle. Previously measured chemicals are indicated using un-shaded, numbered circles. Previously unmeasured chemicals are indicated using shaded, lettered circles. Arrows indicate intermolecular distances between G and all *measured* chemicals.

In summary, Figures 5.1 and 5.2 purposefully depict different modeling scenarios (i.e., different sets of measured and unmeasured compounds); however, comparison of both panels together can be useful for illustrating that the *representativeness maximization* and *redundancy minimization* criteria are sometimes at odds with one another. Thus, the representativeness maximization and redundancy minimization criteria should be applied together to yield the best possible improvement in QMSA accuracy for estimation of all remaining unmeasured compounds.

In this chapter, we will have two objectives: 1) demonstrating the usefulness of QMSA-based *representativeness maximization* as a criterion for prioritizing among unmeasured compounds; and, 2) demonstrating the usefulness of QMSA-based *redundancy minimization* as a second criterion for prioritizing among unmeasured compound, in addition to representativeness maximization. We will also explore possible tradeoffs that can occur when both criteria are applied simultaneously. Both of the datasets from Chapter 2, *in vitro* estrogenicity and from Chapter 3, sorption distribution coefficient (K_d), are used to demonstrate the relative usefulness of both prioritization criteria. Finally, results for both datasets are then compared to one another to also illustrate how dataset composition impacts QMSA accuracy and the practical application of each prioritization criterion.

5.2 Experimental

5.2.1 Data Sources

Estrogenicity and sorption coefficient data were the same as those used in the Chapter 2 and Chapter 3, respectively. Multiple property values for the same chemical were averaged together into a single value. The common logarithm transformation was applied to all property values. All told, the final dataset contained K_d measurements for 80 different chemicals and estrogenicity measurements for 81 different chemicals.

The total pool of chemical structures used in this study included not only those compounds that had been previously measured for estrogenicity and/or K_d but also all other emerging contaminants referenced in the papers noted above plus additional structures from a list of 200 top-prescribed generic pharmaceuticals (Verispan VONA, 2007). The total number of chemical structures evaluated, N, including previously measured and unmeasured chemicals, was 303.

5.2.2 Molecular Descriptors and Similarity Computation

The same 193 molecular descriptors referenced in Chapter 2 were used in this study. Among the three possible types of QMSA models, only PCA models were used to examine each of the two hypothesized QMSA prioritization criteria. Molecular similarity, using the Euclidean Distances parameterization, was computed using the same procedure outlined in Section 3.2.3.

5.2.3 Prioritization hypothesis tests

Reiterating from Section 5.1, the two QMSA-based prioritization criteria to be evaluated in this study include, "*maximizing representativeness*" and "*minimizing redundancy*". Statistical hypothesis testing was used to formally evaluate the usefulness of these two prioritization criteria. In particular, the hypothesis tests were used to assess whether or not QMSA-based prioritization using either or both criteria results in more accurate QMSA models compared to random selection of compounds to be "measured" (i.e., added to the dataset). Mathematically, the representativeness and redundancy criteria were parameterized using the Euclidean Distances (*ED*s) referenced in Section 3.2.3. Since *ED* quantifies the extent of dissimilarity between two chemical structures, we characterized maximally representative (MRP) compounds as those exhibiting the smallest sum of *ED*s ($min\Sigma ED_n$) with all other measured and unmeasured chemicals. In contrast, the least redundant (LRD) compounds were characterized as those exhibiting the largest sum of *ED*s ($max\Sigma ED_m$) with all previously measured compounds (hence the subscript "*m*"), because larger ΣEDs indicate greater dissimilarity from the compounds for which data already exists. Finally, since we want to know how model accuracy is impacted when both prioritization criteria are applied at the same time, we characterized intersection (INT) compounds as those exhibiting both $min\Sigma ED_n$ (representativeness) rankings and $max\Sigma ED_m$ (redundancy) rankings.

The hypothesis tests used in this study were based on iterative application of a modified "leave many out" (LMO) cross-validation procedure put forth in previous studies (Li and Colosi, 2012; Hawkins et al., 2003). This procedure is summarized graphically in Figure 5.3. As indicated near the top of the figure, it was assumed that n*=30 “measured” property values were initially available based on previous measurement, and that 30 additional chemicals were to be selected for additional “measurement. Of these additional 30, $n^{**} = 29$ were selected based on maximized representativeness (MRP = $min\Sigma ED_n$) and 1 was selected using LRD, INT, or randomly (depending on the particular test being evaluated). Then, the $n^{*}+n^{**}+1$ “measured” chemicals were used as the training set to formulate a PCA QMSA model and estimate the properties of the remaining n-n^{*}-n^{**}-1 chemicals. This procedure was completed twice, once for each of the prioritization strategies being evaluated within a particular comparison, and the difference between the resulting q^2 values (Δ) was then computed. The statistical significance of the difference in accuracy between two prioritization strategies was evaluated using a paired t-test for 10,000 iterative trials. The symbol “μ_Δ” was used to connote the average value of Δ for one paired t-test. The null hypothesis for this test was that there is no difference in the accuracy (q^2) between the two prioritization strategies (i.e., H_0: $\mu_\Delta = 0$). The alternative hypothesis was that there is an appreciable difference in accuracy (q^2) between the two strategies (i.e., H_A: $\mu_\Delta \neq 0$). All P-values corresponded to two-sided tests.

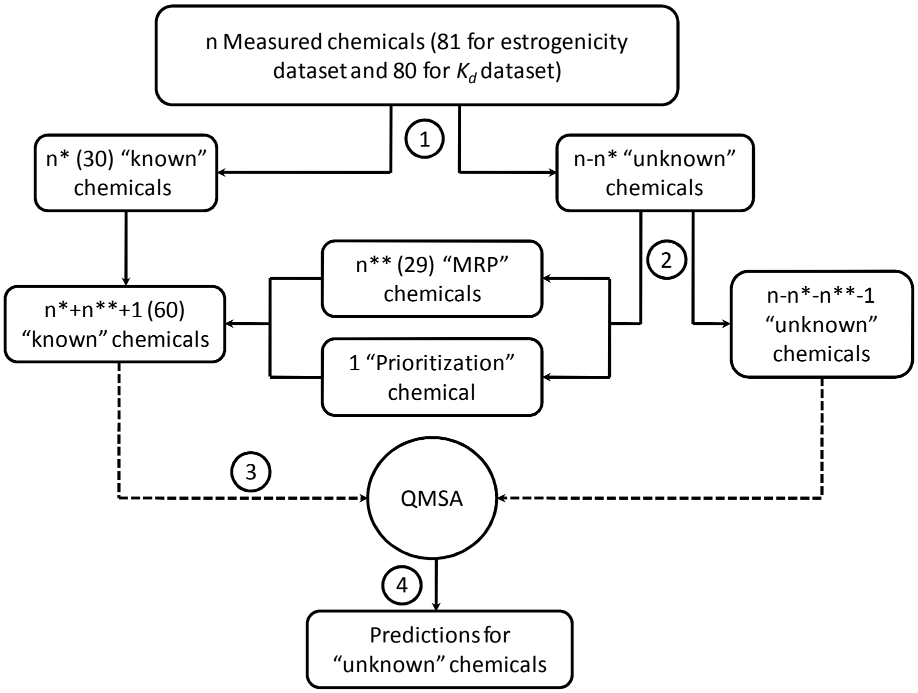
n Measured chemicals (81 for estrogenicity dataset and 80 for K_d dataset)
1
n* (30) "known" chemicals
n-n* "unknown" chemicals
2
n** (29) "MRP" chemicals
n*+n**+1 (60) "known" chemicals
1 "Prioritization" chemical
n-n*-n**-1 "unknown" chemicals
3
QMSA
4
Predictions for "unknown" chemicals

Figure 5.3. A visual summary of the iterative sampling procedure underlying the hypothesis tests used to compare compound selection strategies. Circled numbers refer to critical steps: 1) random assignment of training dataset compounds into one of two groups representing chemicals with "measured" or "unmeasured" property values; 2) use of a "prioritization" strategy, based on cumulative Euclidean Distances to select 30 additional compounds to be "measured": 29 of these were always based on maximized representativeness (MPR), but the remaining 1 was selected based on MRP, LRD, INT, or random; 3) use of the 60 "measured" compounds to build a QMSA model; and, 4) use of the resulting QMSA model to make predictions for the remaining "unknown" compounds.

5.2.3.1 "Representativeness" and "Redundancy Avoidance" versus Arbitrary Selection

For the first comparison, representativeness maximization was compared to arbitrary selection. Maximally "representative" compounds (MRP) were selected based on minimization of summed cumulative Euclidean Distances ($\sum ED$) between each unmeasured chemical and all other unmeasured compounds in the 302-compound pool. For arbitrary selection, compounds were added to the "measured" dataset based on selection using a random number generator.

Estrogenicity and K_d predictions produced using both the QMSA representativeness prioritization approach and the random prioritization approach were assessed using “leave many out” (LMO) cross-validation (Hawkins et al 2003). This sampling, prioritization, and validation process depicted in Figure 5.3 was repeated 10,000 times. For each iteration, the difference (Δ) between correlation coefficients for ΣED_n minimization prioritization (q^2_{MRP}) and random selection (q^2_{RAN}) was computed according to Eq. 5.1. Here, and throughout this chapter, q^2 corresponds to the so-called “naïve q^2” cross-validation coefficient referenced in last section.

$$\Delta = q^2_{\ MRP} - q^2_{RAN} \tag{5.1}$$

Ultimately 10,000 values of Δ were used in each paired t-test. The null hypothesis was $\mu_\Delta = 0$; i.e., there is no difference between random and MRP approaches for chemical prioritization. The alternative hypothesis was $\mu_\Delta \neq 0$; i.e, there is a statistically significant difference between random and MRP approaches for chemical prioritization. This t-test was repeated for various values of k, the number of nearest neighbors used to predict estrogenicity or K_d, ranging from 1 - 3. This was done to ensure the arbitrary selection of k did not artificially affect the outcome of this experiment.

The other experiment was also performed to show that prioritization of chemicals only by maximizing ΣED_m (we use LRD to represent it) should *decrease* the accuracy of QMSA models compared to random (arbitrary) prioritization. Sampling, was repeated 10,000 times. For each iteration, the difference (Δ) between correlation coefficients for the max ΣED_m prioritization (q^2_{LRD}) and random selection (q^2_{RAN}) was computed according to Eq. 5.2:

$$\Delta = q^2{}_{LRD} - q^2{}_{RAN} \quad (5.2)$$

10,000 values of Δ were used in the test. The null hypothesis was $\mu_\Delta = 0$; i.e., there is no difference between random and LRD approaches for chemical prioritization. The alternative hypothesis was $\mu_\Delta \neq 0$; i.e, there is a statistically significant difference between random and LRD approaches for chemical prioritization. This t-test was repeated for various *k* between 1 – 3 for estrogenicity and K_d.

5.2.3.2 Representativeness and Redundancy

A series of paired t-tests was used to assess whether simultaneous consideration of both "*representativeness*" and "*redundancy avoidance*" during chemical prioritization yields better QMSA model accuracy than other prioritization strategies, including random selection. Five types of paired t-tests were used for these experiments in this category: two involved picking "extreme" values of ΣED_m; and three involved picking midrange values of ΣED, at the "intersection" of high representativeness and low redundancy.

The first set of "extreme" paired t-tests was designed to assess whether simultaneous consideration of both "*representativeness*" and "*redundancy avoidance*" during chemical prioritization yields better QMSA model accuracy than arbitrary prioritization. Specifically, these tests compared q^2 values for QMSA models incorporating the *n*** unmeasured chemicals exhibiting smallest values of cumulative dissimilarity ($min\Sigma ED_n$) plus one unmeasured chemical exhibiting largest cumulative dissimilarity ($max\Sigma ED_m$) versus QMSA models incorporating *n*** randomly selected chemicals. Sampling was repeated for 10,000 trials. For each iteration, the difference

(Δ) between correlation coefficients for the "representativeness and redundancy" method ($q^2_{MRP+LRD}$) and the random method (q^2_{RAN}) was computed according to Eq. 5.3:

$$\Delta = q^2_{MRP+LRD} - q^2_{RAN} \tag{5.3}$$

The null and alternative hypotheses were the same as for other paired t-tests described in this section; as were values for *k*, *n**, and *n***.

The second set of "extreme" paired t-tests was designed to assess whether simultaneous consideration of both "*representativeness*" and "*redundancy avoidance*" during chemical prioritization yields better QMSA model accuracy than consideration of *just "representativeness"*. Specifically, these tests compared q^2 values for QMSA models incorporating the *n*** unmeasured chemicals exhibiting smallest values of cumulative dissimilarity ($min\Sigma ED_n$) plus one unmeasured chemical exhibiting largest cumulative dissimilarity ($max\Sigma ED_m$) versus QMSA models incorporating just *n*** chemicals exhibiting $min\Sigma ED_n$. Sampling was repeated for 10,000 trials. For each iteration, the difference (Δ) between correlation coefficients for the "*representativeness and redundancy*" method ($q^2_{MRP+LRD}$) and the "*representativeness only*" method (q^2_{MRP}) was computed according to Eq. 5.4:

$$\Delta = q^2_{MRP+LRD} - q^2_{MRP} \tag{5.4}$$

The null hypothesis was $\mu_{\Delta} = 0$; i.e., there is no difference between these two QMSA approaches for chemical prioritization. The alternative hypothesis was $\mu_{\Delta} \neq 0$; i.e, there is a statistically significant difference between these two QMSA approaches for chemical prioritization. The null and alternative hypotheses were the same as for other paired t-tests described in this section; as were values for *k*, *n**, and *n***.

The third type of paired t-tests, one of three for assessment of "intersection" chemicals, was designed to assess whether simultaneous consideration of both "*representativeness*" and "*redundancy avoidance*" during chemical prioritization yields better QMSA model accuracy than *random* prioritization. This is the same goal as the first type of paired t-test described in this section; therefore, almost the same test was used as was described involving Eq. 5.4. Both tests incorporated *n*** chemicals exhibiting $min\Sigma ED_n$; however, one chemical exhibiting midrange "*representativeness*" and "*redundancy avoidance*" was used in place of the $max\Sigma ED_m$ chemical. Sampling was repeated for 10,000 trials, each time collecting *n*** new chemicals from the MRP ranking plus one chemical at the intersection between MRP and LRD rankings. For each iteration, the difference (Δ) between correlation coefficients for the "intersection" method ($q^2_{MRP+INT}$) and the random method (q^2_{RAN}) was computed according to Eq. 5.5.

$$\Delta = q^2_{MRP+INT} - q^2_{RAN} \tag{5.5}$$

The null and alternative hypotheses were the same as for other paired t-tests described in this section; as were values for *k*, *n**, and *n***.

The fourth paired t-test required for assessment of simultaneous representativeness and redundancy considerations was the second of three tests assessing “intersection” prioritization strategies. This test compared the relative accuracy of QMSA models embodying the simultaneous “*representativeness*” and *“redundancy avoidance*” selection strategies referenced in Eq. 5.5 versus just “*representativeness*”. The difference (Δ) in accuracy between both methods was defined according to Eq. 5.6, wherein $q^2_{MRP+INT}$ and q^2_{MRP} are defined as in Eqs. 5.5 and 5.1, respectively. Sampling was repeated for 10,000 trials.

$$\Delta = q^2_{MRP+INT} - q^2_{MRP} \tag{5.6}$$

The null and alternative hypotheses were the same as for other paired t-tests described in this section; as were values for *k*, *n**, and *n***.

The fifth and final type of paired t-test required for assessment of simultaneous “*representativeness*” and “*redundancy avoidance*” considerations was the third of three tests assessing “intersection” prioritization strategies. This test compared the relative accuracy of QMSA models embodying INT strategy referenced in Eq. 5.6 versus that in Eqs. 5.3 and 5.4. The difference (Δ) in accuracy between both methods was defined according to Eq. 5.7, wherein $q^2_{MRP+INT}$ and $q^2_{MRP+LRD}$ are defined as noted above. Sampling was repeated for 10,000 trials.

$$\Delta = q^2_{MRP+INT} - q^2_{MRP+INT} \tag{5.7}$$

The null and alternative hypotheses were the same as for other paired t-tests described in this section; as were values for k, $n*$, and $n**$.

5.3.3 Recap of Paired t-Test Experiments for Evaluation of Prioritization Strategies.

Table 4.1 summarizes each of the t-test experiments used for comparison of two chemical prioritization strategies. It provides some explanation of the rationale behind each test and help differentiate between similar sounding tests and metrics.

Table 5.1. Summary of paired t-tests performed to evaluate various QMSA-based chemical prioritization strategies.

Test Name	Test Statistics	Comparison[a]
A	$\Delta = q^2_{MRP} - q^2_{RAN}$	MRP vs. RAN
B	$\Delta = q^2_{LRD} - q^2_{RAN}$	LRD vs. RAN
C	$\Delta = q^2_{MRP+LRD} - q^2_{RAN}$	MRP+ LRD vs. RAN
D	$\Delta = q^2_{MRP+LRD} - q^2_{MRP}$	MRP+ LRD vs. MRP
E	$\Delta = q^2_{MRP+INT} - q^2_{RAN}$	MRP+ INT vs. RAN
F	$\Delta = q^2_{MRP+INT} - q^2_{MRP}$	MRP+ INT vs. MRP
G	$\Delta = q^2_{MRP+INT} - q^2_{MRP+LRD}$	MRP+ INT vs. MRP+ LRD

[a] Abbreviations: MRP is most representative compounds ($min\Sigma ED_n$), RAN is random compounds, LRD is least redundant compounds ($max\Sigma ED_m$), INT is “intersection” compounds that are simultaneously most representative and least redundant.

5.3 Results and Discussion

5.3.1 Validating the Usefulness of "Representativeness" and "Redundancy Avoidance" during Prioritization among Unmeasured Chemicals to Improve Predictive Accuracy

The first objective of this chapter was to demonstrate that a QMSA-based approach is useful for prioritization among unmeasured chemicals. In particular, we were hoping to validate the following hypothesis: If only a limited number of property measurements are available, QMSA can be used to determine which "*maximally representative*" unmeasured chemicals will be most useful in developing a QMSA model that will make accurate predictions for the rest of the unmeasured chemicals. As noted in Section 5.2.3, representativeness was parameterized using ΣED_n, whereby compounds exhibiting $min\Sigma ED_n$ are said to be, on average, most "similar" to all other unmeasured compounds in the pool of 303 available structures.

A paired t-test was used to compare the accuracy of QMSA models formulated using two different prioritization strategies: representativeness maximization versus random selection. Table 5.2 summarizes test statistics (t^*) resulting from 10,000 sampling trials of this test for two types of datasets: estrogenicity and K_d. These tests were designated Tests A in Table 5.1.

Table 5.2. q^2 test statistics (t*) for hypothesis testing to compare use of representativeness maximization (MPR) versus random selection (RAN), for the estrogenicity and K_d datasets. These results correspond to Tests A in Table 5.1. All t* values correspond to P-value less than 0.001 unless footnoted otherwise.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean q^2_{MRP}	Mean q^2_{RAN}	t^*	Mean q^2_{MRP}	Mean q^2_{RAN}
1	27	0.34	0.24	54	0.29	0.07
2	28	0.42	0.34	39	0.40	0.28
3	20	0.41	0.36	13	0.29	0.25
4	17	0.39	0.36	-2.6[a]	0.18	0.18

[a] Corresponding P-value is 0.009

All but one of the t* values in Table 5.2 correspond to P-values less than 0.001. Thus, these data indicate that there is a statistically significant difference between q^2_{MRP} and q^2_{RAN}, whereby $min\Sigma ED_n$ yields more accurate QMSA models than random selection of additional compounds to be measured. This is an indication that representativeness-based prioritization (as parameterized using $min\Sigma ED_n$) could make structure-property relationships a more useful and reliable tool for screening the environmental fate and behavior of widely diverse emerging contaminants.

Another observation about Table 5.2 is that the mean q^2_{MRP} and q^2_{RAN} values do not monotonically decrease with increasing k. This is in contrast to previous QMSA analyses with estrogenicity and K_d data, in which smaller k value gave better q^2 values (see Section 2.3.2 and 3.3.2). It's possible that this discrepancy arises from use of "r*epresentativeness*" in selecting among chemical candidates, because addition of chemicals that are most similar, on average, to every other chemical in the test dataset could neutralize the "nugget" effect associated with a heterogeneous, sparsely populated dataset.

Figure 5.4 depicts a subset of the chemical structures utilized in Figure 5.1, to demonstrate in 2-D why larger k might be more efficient for datasets with increasing numbers of "representative" compounds. In panel A, the error associated with prediction of Probe P using $k = 1$ is the distance between P and its nearest neighbor, $\overline{P4}$. The error associated with $k = 2$ is the distance between P and the average of Chemicals 2 and 4, $\overline{P\mu_{2-4}}$. Since $\overline{P4} < \overline{P\mu_{2-4}}$, $k = 1$ results in a better prediction than $k = 2$. Panel B represents the same pool of measured data plus one highly "representative", and thus centrally located, compound – Chemical B. The error associated with prediction of Probe P using $k = 1$ is the distance between P and its nearest neighbor, either Chemical F or Chemical 4 since $\overline{P4} \approx \overline{PF}$. Use of $k = 2$ requires averaging these two nearest neighbors. Since μF-4 < (μF ≈ μ4), use of $k = 2$ results in a more accurate prediction than k = 1 for the expanded dataset depicted Panel B. In this way, adding highly representative structures can improve a model's ability to make predictions for new contaminants by offering them a higher density of "nearer" molecular neighbors.

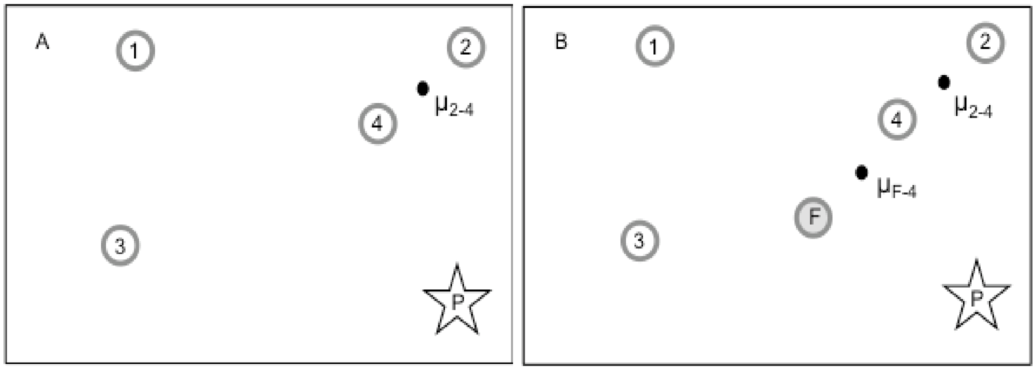

Figure 5.4. A two-dimensional (2-D) depiction of how representativeness-based QMSA prioritization among unmeasured chemicals could increase optimal k compared to $k = 1$.

In working with the available datasets for both selected environmental parameters and evaluating the extent of variability within each dataset, the effectiveness of representativeness prioritization in overcoming nugget heterogeneity was particularly pronounced for the estrogenicity measurements. This dataset was highly heterogeneous, and it contained a large number of values that were very disparate from the rest of the data. This is why increasing *k* from 1 to 2 in Table 4.2 mediates such a dramatic increase in *t**.

The other experiment was also performed to show that prioritization only considering LRD should decrease the accuracy of QMSA models compared to random (arbitrary) prioritization. This result can be expected since LRD prioritization tends to select "unmeasured" chemicals which are not similar to most of "measured" chemicals, which is a large portion (30 out of 81 or 80) in the whole datasets. The results which correspond to Tests B in Table 5.1, are summarized in Tables 5.3.

Table 5.3. q^2 test statistics (t^*) for hypothesis testing to compare use of representativeness minimization (LRP) versus random selection (RAN), for the estrogenicity and K_d datasets. These results correspond to Tests B in Table 5.1. All t^* values correspond to P-value less than 0.001 unless footnoted otherwise.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean q^2_{LRP}	Mean q^2_{RAN}	t^*	Mean q^2_{LRP}	Mean q^2_{RAN}
1	-84	-0.25	0.24	-4	0.05	0.07
2	-102	-0.02	0.34	-5	0.27	0.28
3	-106	0.09	0.36	-6	0.23	0.25
4	-115	-0.22	0.36	0.1[a]	0.17	0.18

[a] Corresponding P-value is 0.761

Several observations can be made about the results in Tables 5.3. First, LRD selection alone generally does not do a better job than random selection. This is evident from the large, negative *t** values exhibited for most tests. This indicates QMSA prioritization selection should focus on MRP chemicals initially. This is logically reasonable since redundancy would only be a problem when the dataset is large enough. Initially it is more important to select MRP chemicals to make the initial chemicals as similar to all possible chemicals as possible.

Second, the estrogenicity dataset is much more sensitive to LRD selection than the K_d dataset. The *t** values for estrogenicity dataset are one order of magnitude larger than those of the K_d dataset. This is probably because the estrogenicity dataset is more heterogeneous than the K_d dataset. Therefore, addition of LRD chemicals to the estrogenicity dataset actually provided the "target" with very dissimilar neighbors in estrogenicity dataset.

5.3.2 Validating the Simultaneous Usefulness of Representativeness and Redundancy Criteria during QMSA-Based Prioritization

From Section 5.3.1, "*representativeness*" is an important consideration during chemical prioritization; however, intuitively, it seems reasonable that "*redundancy avoidance"* should also be important. This section thus describes results from several experiments designed to probe how representativeness-based selected can be improved. All tests involve prioritization of some n compounds exhibiting $min\Sigma ED_n$

(to capture representativeness) plus one chemical picked at least partly on the basis of redundancy avoidance (as parameterized using *max* ΣED_m).

The first of these tests compared q^2 values for QMSA models incorporating the *n*** unmeasured chemicals exhibiting smallest values of cumulative dissimilarity (MRP) plus one unmeasured chemical exhibiting largest cumulative dissimilarity (LRD), versus QMSA models incorporating *n**+1* randomly selected chemicals. These tests were classified as "extremes" selection tests, because *min*ΣED_n was evaluated separately from *min*ΣED_n such that both extremes were represented. Although it may seem self-evident that application of representativeness and redundancy together should be preferable to random selection, because maximized representativeness (MRP) is already preferable to random selection, the mathematical parameterizations of the representativeness and redundancy criteria make it necessary to ensure that the application of both criteria together doesn't undermine the usefulness of MRP alone. Because these mathematical parameterizations are nearly opposite to each other, it was necessary to ensure that addition of redundancy minimization didn't invalidate the usefulness of representativeness maximization. These experiments were designated as Tests C in Table 5.1. Results are summarized in Table 5.4.

Table 5.4. q^2 test statistics (t*) for hypothesis testing to compare simultaneous use of representativeness maximization and redundancy minimization versus random selection, for the estrogenicity and K_d datasets. These results correspond to Tests C in Table 5.1. All t* values correspond to P-value less than 0.001 unless footnoted otherwise.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean $q^2{}_{MRP+LRD}$	Mean $q^2{}_{RAN}$	t^*	Mean $q^2{}_{MRP+LRD}$	Mean $q^2{}_{RAN}$
1	60	0.39	0.04	18	0.20	0.10
2	49	0.47	0.28	22	0.37	0.29
3	45	0.42	0.24	7	0.31	0.29
4	21	0.37	0.30	-1[a]	0.30	0.30

[a] Corresponding P-value is 0.271

Table 5.4 contains t* values for the hypothesis tests comparing MRP+LRD versus random prioritization, where LRD corresponds to the $max\Sigma ED_m$ parameterization. For these tests, 29 of the 30 chemical structures selected for addition to the measured data pool were from the low end of the ranked ΣED_n list. The other 1 chemical structure was from the high end of the ranked ΣED_m list. All P-values are less than 0.001 except k = 4 for the K_d dataset. Thus, the null hypothesis is generally rejected for both estrogenicity and K_d, and there is a statistically significant difference between the two approaches. These results therefore indicate that addition of the redundancy avoidance criterion, as parameterized using $max\Sigma ED_m$, does not undermine the helpfulness of the MRP criterion during QMSA-based prioritization, even though the mathematical parameterizations for these two criteria are nearly opposites of one another.

The second set of tests for simultaneous consideration of representativeness and redundancy were also categorized as an “extremes” test. These tests correspond to Tests D in Table 5.1, and results are summarized in Table 5.5. For both datasets, t*

values for all *k* correspond to *P*-values less than 0.001. Thus, these data indicate that MRP+LRD prioritization yields more accurate QMSA models than just MRP prioritization.

Table 5.5. q^2 test statistics (t*) for hypothesis testing to compare simultaneous use of representativeness maximization and redundancy minimization versus representativeness maximization only, for the estrogenicity and K_d datasets. These results correspond to Tests D in Table 5.1. All t* values correspond to P-value less than 0.001.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean $q^2{}_{MRP+LRD}$	Mean $q^2{}_{MRP}$	t^*	Mean q^2 $_{MRP+LRD}$	Mean $q^2{}_{MRP}$
1	52	0.38	0.26	13	0.20	0.18
2	40	0.44	0.37	21	0.36	0.34
3	30	0.41	0.37	18	0.31	0.30
4	16	0.37	0.35	8	0.30	0.29

The t* values for the estrogenicity dataset are consistently higher than those for the K_d dataset, when considering the same *k* value. The difference in trends for t* versus *k*, whereby t* decreases with increasing k for estrogenicity but increases then decreases for K_d, may reflect underlying differences in the relative density and heterogeneity of the estrogenicity and K_d datasets. For example, the estrogenicity dataset exhibits larger heterogeneity (larger ranges of measured property values) with significant clustering (spatial "nuggets"), whereby there are several groups of structures that are similar to each other but very different from the most "representative" structures in the rest of the 303-chemical pool. For all *k*, forcing the model to select an unmeasured chemical from one of the clusters, which tend to be poorly predicted because they are so dissimilar to the bulk of the more "representative" chemicals, will improve the model's ability to make predictions for other poorly-predicted chemicals. This likely improves the overall model accuracy, because the new chemical will fill an existing "gap" between the well-predicted clusters and the poorly-predicted outliers. In contrast, the K_d dataset tends is far more homogenous (smaller ranges of measured property values) with more even coverage (less clustering). This might make it less important and more difficult to avoid redundancy when using max ΣED_m, because there are fewer (if any) "gaps" to fill. It's currently unclear why this effect should decrease at larger k values; however, it might be possible to design artificial datasets (with various heterogeneity, coverage, and clustering patterns) to explore why this could be and explore this effect in depth.

5.3.3 Validating the Usefulness of Intersection Compounds during Intersection

Three types of experiments were designed to test the hypothesis that chemicals at the intersection between most representative ($min\Sigma ED_n$) and least redundant ($max\Sigma ED_m$) should be the best candidates for addition to the measured dataset. In particular, picking one set of compounds that exhibit midrange values of *both* representativeness and redundancy avoidance could be better than picking sets of compounds that exhibit one of either desirable attributes. This is why these tests are said to evaluate "intersection" rather than "extreme" selection strategies. These tests were designated as Tests E-G in Table 5.1.

Table 5.6. q^2 test statistics (t*) for hypothesis testing to compare INT selection versus random selection, for the estrogenicity and K_d datasets. These results correspond to Tests E in Table 5.1. All t* values correspond to P-value less than 0.001 unless footnoted otherwise.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean $q^2_{MRP+INT}$	Mean q^2_{RAN}	t^*	Mean $q^2_{MRP+INT}$	Mean q^2_{RAN}
1	44	0.26	0.04	16	0.18	0.10
2	40	0.37	0.21	22	0.35	0.27
3	36	0.37	0.24	10	0.32	0.29
4	15	0.34	0.30	3[a]	0.31	0.30

[a] Corresponding P-value is 0.002

The results in Table 5.6 correspond to Test E in Table 5.1. For these tests, compounds were selected from the so-called "intersection" (INT) of the $min\Sigma ED_n$ (representativeness) rankings and the $max\Sigma ED_m$ (redundancy) rankings. More specifically, 29 of 30 chemical structures selected for addition to the measured data pool were from the low end of the ranked ΣED_n list while the other 1 was from INT. This was done to avoid using mathematically opposite parameterizations for the two prioritization criteria of interest; in an attempt to mitigate the possible tradeoff associated with the seemingly contradictory criteria. All t* values in the lower half of Table 5.6 correspond to P-values less than 0.002. Thus, neither of the evaluated redundancy parameterizations undermines the usefulness of the maximized representativeness criterion during QMSA-based prioritization. For both the LRD and the INT approaches, the combined use of the representativeness and redundancy criteria together results in better QMSA accuracy than random selection of compounds to be measured. Again, k = 2 ensures that best model accuracy was achieved for both datasets.

The results in Table 5.7 corresponding to Test F. These results are less clear than the results in Table 5.6. Many *t** values correspond to statistically significant P-values which point to rejection of the null hypothesis (H_0), which suggests that there is some statistically significant difference in QMSA estimation accuracy for use of MRP+INT versus just MPR by itself. However, several of the statistically significant *t** values are less than zero, which means t hat the MRP + INT approach is sometimes better and sometimes worse than just the maximized representativeness approach. As such, there is no conclusive evidence to show that there is any difference between the two approaches. In light of these conflicting results, we cannot conclusively say that the MRP+INT is generally useful as a QMSA-based prioritization strategy.

Table 5.7. q^2 test statistics (t^*) for hypothesis testing to compare INT selection versus represnetativeness selection, for the estrogenicity and K_d datasets. These results correspond to Tests F in Table 5.1. All t^* values correspond to P-value less than 0.001 unless footnoted otherwise.

k	Estrogenicity Dataset			K_d Dataset		
	t^*	Mean $q^2_{MRP+INT}$	Mean q^2_{MRP}	t^*	Mean $q^2_{MRP+INT}$	Mean q^2_{MRP}
1	12	0.27	0.27	-2 [b]	0.18	0.18
2	5	0.37	0.37	14	0.38	0.37
3	-2[a]	0.37	0.37	24	0.32	0.30
4	-8	0.34	0.35	-8	0.30	0.31

[a] Corresponding P-value is 0.110; [b] Corresponding P-value is 0.046

Now the question arises, which prioritization method is better, LRD or INT? The inconclusive results in Table 5.6 make it worthwhile to explicitly evaluate the relative usefulness of the LRD and INT redundancy minimization parameterizations; however, it is emphasized that this is explored as a question of practical implementation rather than an investigation of generalized, theoretical usefulness. For this comparison, it was assumed that representativeness maximization and redundancy minimization are implemented together (i.e., MRP+LRD vs. MRP+INT). The following hypothesis set was used: H_0: μ_Δ = 0, there is no difference between the LRD versus INT parameterizations when using redundancy minimization and representativeness maximization; and H_0: $\mu_\Delta \neq 0$, there is some appreciable difference between the LRD versus INT parameterizations when using redundancy minimization and representativeness maximization. For these hypotheses, Δ was defined as $q^2_{MRP+LRD}$ minus $q^2_{MRP+INT}$, such that values greater than zero correspond to increased accuracy for MRP+LRD compared to MRP+INT. This test corresponds to Test G in Table 5.1 Results are summarized in Table 5.8.

From Table 5.8, it is very obvious that the MRP+LRD approach results in more increased accuracy for QMSA predictions compared to MRP+INT. All t^* values correspond to P-values less than 0.001. This is consistent with results from Table 5.5 and Table 5.7, wherein use of MPR+LRD is shown to improve QMSA accuracy compared to representativeness maximization for all tested k values but use of MRP+INT exhibited some increases and some decreases in QMSA accuracy.

Table 5.8. Test statistics (t*) for hypothesis testing to compare use of representativeness maximization with either LRD or INT parameterization of the redundancy criterion, for the estrogenicity and K_d datasets. These results correspond to Test G in Table 5.1.

k	Estrogenicity Dataset[a]			K_d Dataset[a]		
	t^*	Mean $q^2_{MRP+LRD}$	Mean $q^2_{MRP+INT}$	t^*	Mean $q^2_{MRP+LRD}$	Mean $q^2_{MRP+INT}$
1	16.1	0.377	0.343	45.9	0.192	0.173
2	58.1	0.465	0.372	19.8	0.356	0.339
3	75.4	0.416	0.304	15.8	0.307	0.297
4	89.6	0.369	0.249	6.2	0.294	0.291

Table 5.9 presents a recap of the paired t-tests performed in this chapter. To sum up, we conclude that "*representativeness maximization"* is a very important criterion when determining what new "unmeasured" chemicals should be added to a dataset. There is also some evidence that "*redundancy avoidance*" is an important prioritization criterion. We expect that redundancy avoidance will become increasingly important over time, once sufficient "representative" chemicals have been added to datasets of interest

Table 5.9. Recap of the results of hypothesis q^2 tests for comparisons among selected prioritization strategies.

Test	Comparison	Conclusion for Estrogenicity[a]	Conclusion for K_d [a]
A	MRP vs. RAN	MRP > RAN	MRP > RAN
B	LRD vs. RAN	LRD < RAN	LRD < RAN
C	MRP + LRD vs. RAN	MRP + LRD > RAN	MRP + LRD > RAN
D	MRP + INT vs. RAN	MRP + INT > RAN	MRP + INT > RAN
E	MRP + LRD vs. MRP	MRP + LRD > MRP	MRP + LRD > MRP
F	MRP + INT vs. MRP	MRP + INT > MRP	MRP + INT > MRP
G	MRP + LRD vs. MRP + INT	MRP + LRD > MRP + INT	MRP + LRD > MRP + INT

[a] based on tests achieving highest q^2 value for each prioritization strategy.

5.3.4 Additional Observations Regarding the Differences among Datasets

One common observation about Tables 5.2-5.9 is that, for hypotheses tests delivering statistically significant conclusions (i.e., P-values less than 0.05), t^* values tend to be larger for the estrogenicity dataset than for the K_d dataset. This indicates that the QMSA prioritization methods provided better prediction accuracy for the estrogenicity dataset than for K_d dataset, offering potentially improved enhancement of QMSA accuracy compared to what can be achieved for the K_d dataset.

The reason why QMSA-based prioritization achieves different improvement for the two datasets is probably related to underlying differences in the structure of the two datasets. As noted previously, the estrogenicity dataset exhibits larger heterogeneity with significant clustering compared to the relatively more homogenous K_d dataset. This is once again demonstrated from the cumulative probability functions plotted in Figure 5.5. In Figure 5.5, the measured chemicals in K_d dataset are clustered closer together in multi-dimensional space than those in estrogenicity dataset. This can be illustrated by comparison of the median ΣED values (i.e., 50% cumulative probability, the horizontal line in Figure 5.5) for each dataset: roughly 2600 for K_d versus roughly 2750 for estrogenicity. The datasets with smaller median ΣED values should be more clustered and therefore more homogenous than those with larger values. Therefore, we concluded that the estrogenicity dataset exhibits larger heterogeneity with significant clustering, whereby there are several groups of structures which are similar to each other but very different from the most representative structures in the rest of the 303-chemical pool.

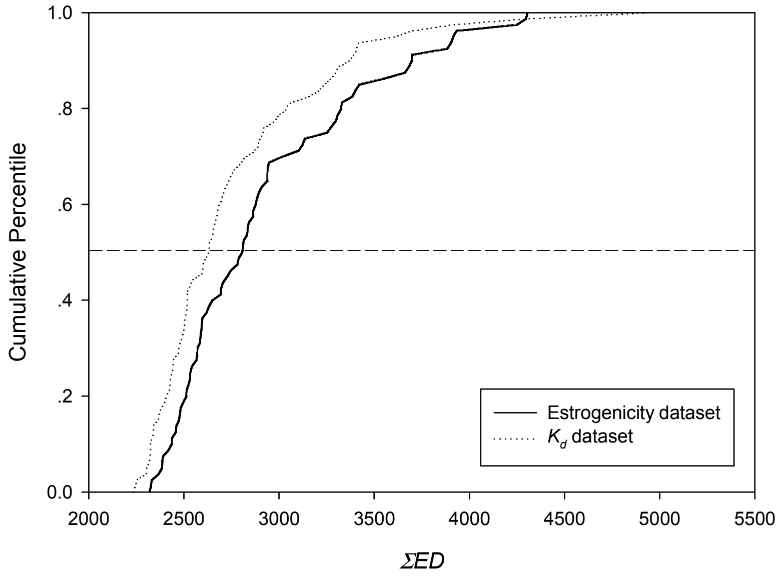

Figure 5.5. Cumulative probability analysis for both datasets based on *ΣED* of each chemical with respect to other 302 in the original pool. Medians for two datasets are indicated by the intersections between the dash line and cumulative probability curves.

Within the estrogenicity dataset, for all *k* values, forcing the model to select an unmeasured chemical from one of the clusters, which tend to be poorly predicted because they are so dissimilar to the many more "representative" chemicals, will improve the model's ability to make predictions for other poorly-predicted chemicals. This likely improves the overall model accuracy, because the new chemical will fill an existing "gap" between the well-predicted clusters and the poorly-predicted outliers. Therefore, redundancy avoidance should become a higher priority when considering incorporation of new chemicals for heterogeneous datasets. In comparably homogeneous sets, maximizing representativeness is the most important consideration.

5.3.4 Selection of External Validation Compounds

Three "unmeasured" chemicals within the 303 chemical pool were identified as good candidates for external validation of K_d predictions based on MRP, LRD, or INT prioritization. These are fluconazole, metformin, and benazepril, respectively. These chemicals were selected for measurement in order to validate QMSA model predictions. More specific, fluconazole is one of MRP chemicals in the 303 chemical dataset. Metformin is one of LRD chemicals and benazepril is one of INT chemicals. Therefore, once they are added to the K_d dataset, they will novel representative and/or non-redundant information compared to other, previously-measured chemicals. This is one of reasons why we chose these three selected chemicals for K_d and k_b measurements in previous two chapters. Another reasons include they are top 200 prescribed pharmaceuticals in U.S. and their mass loadings to WWTPs (Ottmar et al., 2010) are very high. However, K_d and k_b of these three have never been reported before.

5.4 Conclusions

In this chapter, we have demonstrated the usefulness of QMSA as a tool for selection of which “unmeasured” candidates should be prioritized for measurement to increase a QMSA model’s accuracy. These results suggest that QMSA is indeed a potentially powerful tool for prioritization among unmeasured chemicals. In particular, we showed that QMSA-based prioritization using some combination of maximized representativeness and minimized redundancy approaches affords better improvement in QMSA estimation of property values among the remaining unmeasured compounds than random selection. This trend was observed for both the estrogenicity dataset and the sorption distribution dataset, suggesting that perhaps the two-fold prioritization criteria could be used to improve estimation accuracy for a wide variety of environmental parameters. The results from this study could be construed as motivation for a change in the current paradigm for selection of which compounds should be evaluated in a particular research study; i.e., greater emphasis on a *modeling*-based selection strategy given that it is impossible to make individual *measurements* for each and every emerging contaminant of interest.

This chapter also provided a future research direction by identifying three chemicals that are high priorities for experimental evaluation. They are fluconazole, metformin, and benazepril. These three chemicals serve as a tool to examine the external validity of our QMSA models in subsequent chapters.

5.5 References

Andersen, H.R., Hansen, M., Kjolholt, J., Stuer-Lauridsen, F., Ternes, T., and Halling-Sorensen, B. (2005) Assessment of the importance of sorption for steroid estrogens removal during activated sludge treatment. *Chemosphere*, 61, 139-146.

Barron, L., Havel, J., Purcell, M., Szpak, M., Kelleher, B., and Paull, B. (2009) Predicting sorption of pharmaceuticals and personal care products onto soil and digested sludge using artificial neural networks. *Analyst,* 134, 663-670.

Carballa, M., Omil, F., Ternes, T., and Lema, J.M. (2007) Fate of pharmaceutical and personal care products (PPCPs) during anaerobic digestion of sewage sludge. *Water Research.* 41, 2139-2150.

Feng, Y., Zhang, Z., Gao, P., Su, H., Yu, Y., and Ren, N. (2010) Adsorption behavior of EE2 (17α-ethinylestradiol) onto the inactivated sewage sludge: Kinetics, thermodynamics and influence factors. *Journal of Hazardous Materials*. 175, 970-976.

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

Kupper, T., Plagellat, C., Brändli, R.C., de Alencastro, L.F., Grandjean, D., and Tarradellas, J. (2006) Fate and removal of polycyclic musks, UV filters and biocides during wastewater treatment. *Water Research*. 40, 2603-2612.

Li C., and Colosi, L.M. (2012) Molecular similarity analysis as tool to prioritize research among emerging contaminants in the environment. *Separation and Purification Technology*. 84, 22-28.

Ottmar, K.J., Colosi, L.M., Smith, J.A. (2010b) Development and application of a model to estimate wastewater treatment plant prescription pharmaceutical influent loadings and concentrations. *Bull. Environ. Contam. Toxicol.* 84, 507-512.

Pharmacy Times. (2013) Top 200 Drugs of 2011. Last accessed in April 21, 2013. http://www.pharmacytimes.com/_media/_pdf/Top_200_Drugs_2011_Total_Rx.pdf

Simonich, S.L., Federle, T.W., Eckhoff, W.S., Rottiers, Webb, A., Sabaliunas, S. D., and de Wolf, W. (2002) Removal of fragrance material during US and European wastewater treatment. *Environ. Sci. Technol.* 36, 2839–2847.

Ternes, T.A., Herrmann, N., Bonerz, M., Knacker, T., Siegrist, H., and Joss, A. (2004) A rapid method to measure the solid–water distribution coefficient (K_d) for pharmaceuticals and musk fragrances in sewage sludge. *Water Research*, 38, 4075-4084.

Wick, A., Fink, G., Joss, A., Siegrist, H., and Ternes, T.A. (2009) Fate of beta blockers and psycho-active drugs in conventional wastewater treatment. *Water Research*. 43, 1060-1074.

Zhao, J., Li, Y., Zhang, C., Zeng, Q., and Zhou, Q. (2008) Sorption and degradation of bisphenol A by aerobic activated sludge. *J. Hazard. Mater.* 155, 305–311.

Chapter 6 – Conclusions and Future Work

6.1 Conclusions

This research comprehensively investigated the applicability of QMSA as a tool for estimation several important environmental engineering parameters that have not been previously evaluated by QMSA before. The overall objective was not only to develop a series of accurate predictive tools for property estimation but also to apply the generated property estimates to characterize the fate of three important emerging contaminants in representative WWTP systems. Here we recap the three hypotheses proposed in Chapter 1:

Hypothesis 1: *QMSA models can be used to accurately predict environmental engineering parameters of interest.*

The first hypothesis was demonstrated as true in Chapters 2, 3, and 4. In Chapter 2, we first demonstrated that QMSA can accurately predict environmental information for large, highly diverse classes of chemical structures. This is made evident by the good q^2 values (0.84) achieved for prediction of *in vitro* estrogenicity measurements. Thus, it seems likely that efficient use of accurate QMSA models could deemphasize the need for labor-intensive, time-consuming analytical measurements to support regulatory agendas related to emerging contaminants. In Chapter 3, this hypothesis is proven by the excellent internal-validated q^2 values (as high as 0.82) achieved for prediction of this K_d dataset and good external prediction results for K_d of three priority emerging contaminants (fluconazole, metformin, and benazepril) which have never been measured before. In Chapter 4, the results once again provide clear evidence that QMSA models can be used to predict k_b for large, highly diverse classes of chemical structures. This is proven by the excellent internal-validated q^2 values (as high as 0.78) achieved for prediction and external prediction results for k_b of three emerging contaminants which have never been measured before.

Hypothesis 2: *K_d and k_b of selected emerging contaminants can be useful for estimation their effluent concentrations out from WWTPs*

The second hypothesis was demonstrated in Chapter 4. The results from Chapter 4 demonstrate that the fate of two selected emerging contaminants can be well predicted using either a simple mass balance model or a slightly more complex model from previously published literature. Data in Chapter 4 also provides additional indication that WWTP operational parameters may have a great impact on effluent concentrations. Therefore it is essential to obtain WWTP-specific operational parameters in order to generate accurate calculations.

Hypothesis 3: *Prioritizations evaluated by QMSA are critical factors for prioritizing among unmeasured chemicals and determining which additional measurements will result in maximally increased model accuracy.*

The last hypothesis is demonstrated in Chapter 5, where we have clearly demonstrated the usefulness of QMSA as a tool for selection best possible “unmeasured” candidates. Those best possible candidates, once measured, can be expected to increase QMSA model’s accuracy. More specificcally, Chapter 5 provides us three QMSA selection criteria, i.e. "*representativeness*", "*redundancy avoidance*", and "*intersection*". All of these are demonstrated as effective tools for selection of chemicals that can improve QMSA models' accuracy compared to random selection. However, the relative effectiveness among these three is highly dependent on the structure of individual investigated dataset.

6.2 Future work

In general, this research demonstrates that QMSA can be applied for a broader prospective, i.e. to predict essential environmental engineering parameters. However, more research would be beneficial and highly essential for future consideration. First, current QMSA relies on the averaging of property values among several measured nearest neighbors for the target chemical. This can sometimes be very problematic, since this procedure considers all nearest neighbors to be equally relevant. Therefore a more detailed evaluation of different nearest neighbor based on their individual *ED*s to the target chemical should be investigated. The possible distance related algorithms which can be applied in this scenario could be Inverse Distance Weighting or Kriging. Second, more quantitative parameters should be introduced to tell whether a dataset is homogenous or heterogeneous. This is particular important when considering what are the best possible unmeasured chemicals to be measured to improve QMSA's accuracy. As discussed in Chapter 5, structural difference among the two case study datasets lead to selection of different QMSA prioritization methods for each dataset. Third, in this research, QMSA was selected as the primary predictive models and PCA coupled with *k*NN algorithm were the primary statistical methods. In future, it is very good to know the predictive performance of traditional regression-based QSPRs on the studied datasets. It is also interesting to test if the model accuracy could be further improved by some other advanced machine learning techniques, such as Artificial Neural Network coupled with *k*NN. Finally, in Chapter 4, we illustrated the significance of the sampling campaign on elimination of actual sampling variation. To further test the two models used in Chapter 4, it would be greatly preferable to perform the comprehensive sampling campaign in the future.