Predicting Biological Removal of Contaminants in Wastewater Treatment: QSBR Modeling

A Thesis

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Charles Burgis

December

2012

APPROVAL SHEET

The thesis

is submitted in partial fulfillment of the requirements

for the degree of

Master of Science

AUTHOR

This thesis has been read and approved by the examining committee:

Dr. Lisa M. Colosi

Thesis advisor

Dr. Wu-Seng Lung

Dr. Teresa B. Culver

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

December

2012

Abstract

Contaminant fate and transport models are a highly desirable alternative to direct measurement of environmental behavior for the very large number of so-called “emerging contaminants”. In particular, it is desirable to estimate what fraction of emerging contaminant loading is removed by conventional wastewater treatment plant processes, so that the loading of organic wastewater contaminants (OWCs) into the environment can be assessed. In this thesis, we focused on prediction of biodegradation rate constants for removal of OWCs during activated sludge treatment. A Quantitative Structure-Biodegradability Relationship (QSBR) modeling approach was used to predict pseudo first-order biodegradation rate constants (k_b) based on molecular descriptors from commercially available computational chemistry software. A training dataset comprising 65 previously measured molecular structures was collected from nine different literature sources. This data was then used to create four QSBR models using varying molecular subsets and their associated descriptors. Internal validation statistics indicate that the overall QSBR model (comprising all compounds in the training set) achieves less predictive ability (R^2 =0. 49) than three smaller QSBR models (R^2 = 0.97, 0.88, and 0.90) that were created from smaller subsets of the same dataset. External validation was performed via direct measurement of three highly-prescribed, previously unevaluated pharmaceutical OWCs: metformin, benazepril, and warfarin. Their respective k_b values were 0.0105, 0.0033, and ≈ 0 L/g-h. Of the four QSBR models, the general, all-encompassing model delivered highly accurate predictions for metformin Logk_b (measured = -1.98 versus predicted = -2.03) and benazepril Logk_b (measured = -2.48 versus predicted = -2.29), while warfarin was best estimated using one of the smaller subset QSBRs. Analysis of external validation results also indicated that the diversity of the molecules comprising each model’s underlying dataset should be used to assess each model’s application domain before the model is used to make predictions for unmeasured

compounds. Other results from the QSBR models, including identification of which molecular descriptors are best correlated with biodegradation rate constant, offer new information on particular contaminants of concern in surface and groundwater systems and the nature of WWTP biodegradation reactions. Future work will focus on comparisons between QSBR and other parameter modeling approaches.

Acknowledgements

This work was completed as a collaborative effort with Professor Lisa M. Colosi and Chenxi Li. Professor Colosi, your willingness to discuss questions (even trivial ones) in detail and your careful analysis of our work made this research experience much more enjoyable and successful. I greatly appreciate your effort and interest in this project over the past year and a half. And Chenxi, thanks for teaching me so much during all of the hours we spent together in the lab. Of course, thanks also to my family and friends. Your support keeps me going. With all of the great people that I’ve known and the many opportunities that I’ve been given, I feel pretty lucky.

Table of Contents

Page

Chapter 1 - Introduction

1.1 Emerging Contaminants

Water, arguably our planet's greatest resource, once seemed to be in limitless supply. In recent times, however, as humans have begun to consume our planet's natural resources at ever increasing rates, many have come to realize that our supply of clean, fresh water is indeed finite, and in need of protection. If future generations of people hope to enjoy the luxury of clean, inexpensive fresh water that many have come to take for granted today, more needs to be done to preserve the quantity and quality of our water resources. Adequate protection requires understanding of our water systems, their surrounding environments and the ways in which we affect their health. The lifestyles of developed and developing countries have placed more of a burden on worldwide water resources than most realize. The first task at hand, tracing out all the ways we affect the quantity and quality of fresh water on this planet, is an enormous task in itself.

With more people on Earth than ever before, using and disposing of an increasingly diverse array of chemical substances, accounting for potential pollutants in our water systems has become very challenging. Part of this challenge is the wide variety of historically unrealized, recently detected, and largely unregulated class of contaminants known as "emerging contaminants". Speaking generally, emerging contaminants are both naturally occurring and man-made chemicals and/or microbial contaminants with a wide variety of uses, arising from a wide variety of sources (USGS, 1012). They are categorized as "emerging" because they have not been traditionally considered pollutants. This definition is understandably vague as it includes a huge variety of contaminants, most of which are not well understood in natural and engineered aquatic environments. Emerging contaminants may include pharmaceuticals, personal care products, fragrances, household cleaning agents, plasticizers, pesticides, steroids, industrial products, veterinary products, food additives and other substances, the majority of which are unregulated throughout the world (Lapworth et al., 2012; Focazio et al., 2008; Daughton and

Ternes, 1999; Muir, 2006). While some of these emerging contaminants have just recently been created, others have been in use for a long time, but have only recently been detected. Recent improvements in analytical techniques have led to an increased frequency of detection of these compounds throughout a wide variety of our natural and man-made water systems (Kolpin et al., 2002). Many emerging contaminants have been detected at varying concentrations in surface water, groundwater, wastewater, and even in some drinking water sources (Lapworth et al., 2012; Focazio et al., 2008; Kolpin et al., 2002). Because of the vague nature of the term "emerging contaminants", other slightly more specific terms are often used in scientific literature to better identify this group of contaminants. These terms include "organic wastewater contaminants" (OWCs), "Emerging Organic Contaminants" (EOCs), or just "pharmaceuticals and personal care products" (PPCPs).

Due to a lack of research, few generalizations can be made regarding overall human health and environmental effects of low levels of emerging contaminants in various water systems. However, specific case studies have highlighted some of the adverse impacts of certain specific emerging contaminants. For example, one subclass of emerging contaminants are endocrine disrupting compounds (EDCs), which encompasses compounds such as 17β-estradiol (E1), 17α-estradiol (EE1) and alkylphenol polyethoxylates (APEO). These chemicals have been detected in surface and ground waters around the world (Rudel, 1998; Cargouët et al., 2004; Lozano et al., 2012). These EDCs are capable of mimicking natural hormones and interacting with animal estrogen receptors. As a result, they are associated with several negative health effects, such as decreased fertility, feminization/defeminization, and alterations to immune functions (Colborn, 1993). While many of these EDCs are only found in nanogram per liter levels in most surface waters, there is much concern that even very small concentrations of EDCs in water could cause harm to humans, aquatic wildlife, and the environment (Cargouët et al. 2004). In fact, endocrine disrupting effects of estrogens have been reported in surface waters with concentrations as low as the nanogram per liter range (Joss, 2006). Additionally, many lipophilic organic contaminants are found to bioaccumulate within an organism and/or biomagnify up the food chain, placing certain organisms, particularly those at the top of certain food chains, at an elevated risk (Lozano et al., 2012; McLachlan et al., 2011; Gray, 2002). As a result, even very low concentrations of a contaminant in surface waters may gradually build up in certain “high risk” organisms.

While the known and/or suspected effects of certain emerging contaminants like EDCs are one cause for concern, there may also be less perceptible or even imperceptible consequences of exposure to emerging contaminants. Daughton and Ternes warn against such imperceptible effects in a 1999 paper in *Environmental Health Perspectives*, arguing that, “A major concern is not necessarily acute effects to nontarget species (effects amenable to monitoring once they are understood), but rather the manifestation of perhaps imperceptible effects that can accumulate over time to ultimately yield truly profound changes-those whose causes would be obscured by time and that would not be distinguishable from natural events.” In other words, even minute changes caused by continuous exposure to certain emerging contaminants could accumulate over time, possibly very gradually over a long enough time to mask such changes all together. Because these contaminants have infiltrated so many of our fresh water systems, the mystery of the unknown effects of many emerging contaminants is troubling.

The newfound recognition of these compounds has resulted in an increase in research regarding the fate, transport, and toxicity of pharmaceuticals and personal care products (PPCP). Currently in the US, the US EPA and USGS are investigating detection, sources, fate and transport, and ecological effects of PCCPs (US EPA, 2012; USGS, 2012). The EU has also begun major research projects focused on inexpensive methods to lower PPCP loading into from wastewater treatment plants, such as project “Poseidon” (Ternes 2004). Furthermore, a variety of studies (including this thesis) are aimed at modeling the fate and transport of emerging contaminants.

1.2 Fate and Transport of Emerging Contaminants

1.2.1 Sources

Emerging contaminants enter the water supply via point sources and non-point sources. Major point sources include wastewater treatment plant (WWTP) discharges, which frequently contain excreted or improperly disposed of pharmaceuticals, their metabolites and transformation products, and personal care products such as soaps, shampoos and fragrances from washing, bathing and showering (Ternes et al., 2004). WWTP effluents may also contain residuals from food additives and household cleaning agents, and these may mix together with PPCPs as they travel into and out of a WWTP system. Many emerging contaminants undergo only partial removal during municipal wastewater treatment, and, as a result, they may completely or partially pass though the plant unchanged and exit with the plant effluent. From these surface waters, contaminants may travel into groundwater and/or enter into drinking water sources.

Other emerging contaminants may enter into freshwater systems as non-point sources. Examples of non-point sources include the land application of pesticides and the release of veterinary pharmaceuticals from animal feedlot operations. Additionally, land-application of contaminated WWTP biosolids may constitute an additional non-point source, as compounds desorb from the biomass and are released into the environment. Sorption onto sewage sludge is often significant for personal care products due to their high lipophilicity (Ternes et al., 2004). Pesticides are applied in heavy doses in agricultural practices and can easily be carried into surface or ground water via runoff. Similarly, animal waste containing veterinary pharmaceuticals from animal feedlots may overflow or leak from waste storage structures or be applied to land directly as manure (Kolpin et al., 2002). Though these non-point sources are significant, WWTPs (point-sources) are considered the principal point of entry for emerging contaminants into the environment.

With so many different contaminants travelling and transforming through often complicated water systems from disparate sources, sophisticated techniques are generally required to model and/or map the fate and transport of various compounds in the environment. Because of large mass loadings of diverse emerging contaminants entering surface waters though WWTPs, understanding the fate and behavior of these chemicals during typical wastewater treatment processes has become a key research focus. Such a step is the focus of numerous analytical, kinetic, mechanistic and modeling studies (Kolpin, 2002; Johnson, 2001; Joss, 2006; Okey, 1996); it is also the focus of this thesis.

1.2.2 Wastewater Treatment Plants and Emerging Contaminants

In most developed countries the process of wastewater treatment includes a variety of steps designed to improve wastewater quality with respect to certain regulatory standards. According to the US EPA, as of 2004, municipal wastewater treatment plants served 75% of the nation's people, with the rest of the country's population utilizing septic or other onsite systems (EPA, 2009). Regulatory standards based on the 1972 Clean Water Act (CWA) set limits on the biochemical oxygen demand (BOD) and total suspended solids (TSS) of WWTP effluents (EPA, 2007). These regulations have been put in place to protect WWTP receiving waters. Because of such regulations, WWTPs are designed with the primary purpose of eliminating certain nutrients and organic matter from wastewater in an effort to reduce BOD. In contrast, WWTPs are not designed to remove the comparatively low concentrations of currently unregulated emerging contaminants that are also present in the wastewater.

A simplified summary of a typical WWTP is summarized in Figure 1.1. As displayed in the figure, the first step includes the removal of wastewater solids by settling in the primary clarifiers. Solids with densities greater than water settle to the bottom of the primary clarifiers, while plastics and greases with densities less than water rise to the top. These floaters and sinkers are removed from the rest of the wastewater and either reused as primary sludge or disposed of (USGS, 2012). In the second step, the wastewater flows into activated sludge ("secondary") basins where nutrients and organic components are broken down biologically by sludge bacteria. Essentially, the activated sludge bacteria are grown in aerated conditions and fed a constant supply of wastewater, leading to the decomposition or biodegradation of many nutrients and organic components (EPA, 2009). Also included in the secondary treatment step is another round of settling though a secondary clarifier, such that the secondary sludge can be removed and recycled. A variety of tertiary treatment steps may then be implemented to further remove harmful bacteria and potentially other suspended contaminants. These include sand or activated carbon filtration, possibly followed by disinfection by chlorination. Following these steps, the treated wastewater effluent is released into a receiving body of water (generally a river, lake or ocean).

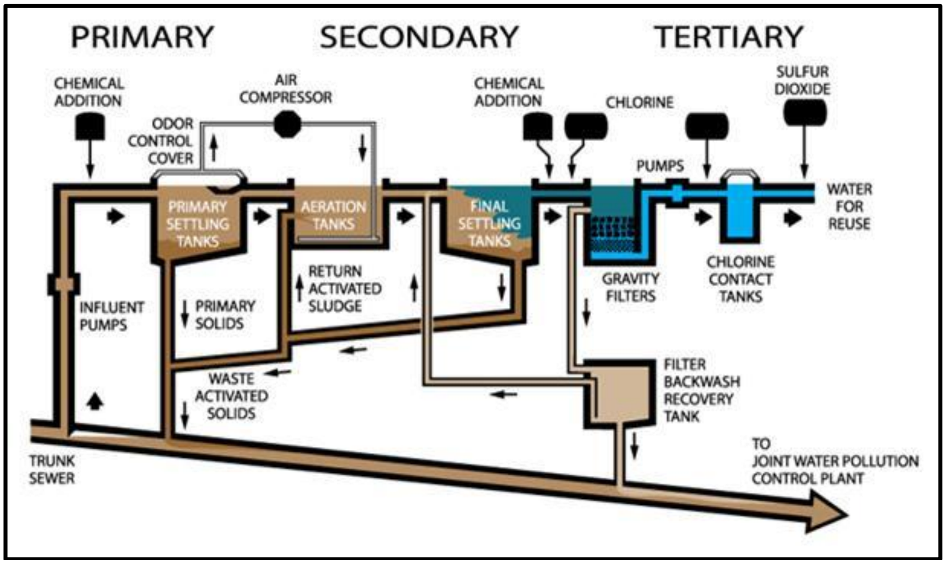


Figure 1.1: A simplified summary of the wastewater treatment process. Though many WWTPs follow this scheme, there are a wide variety of other WWTP systems. This image is just one common example. Image from http://www.lacsd.org/wastewater/wwfacilities/moresanj.asp.

1.2.3 Removal of OWCs in WWTPs

While emerging contaminants found in wastewater are not specifically targeted for removal by WWTPs, there are several ways they may be removed by certain of the above treatment processes. Organic wastewater contaminants (OWCs) may be transformed or removed from wastewater by the following general mechanisms: sorption, biodegradation, volatilization, hydrolysis and photolysis. The most important of these reactions are sorption and biodegradation. Absorption involves the attraction of a molecule's hydrophobic aliphatic and aromatic groups to the lipophilic cell membrane of microorganisms and fat/grease fractions of the sludge (Ternes et al., 2004). Adsorption results from electrostatic attractions between positively charge functional groups of an organic compound (e.g., amino groups) and the negatively charged surfaces of sludge microorganisms. Sorption of OWCs may occur in both the primary and secondary treatment stages of waste water treatment, but secondary sludge is generally better suited to achieve more significant removal. While sorption can remove OWCs from wastewater, it leads to sludge contaminated with active compounds, which may still pollute land/water environments when applied as fertilizer or stored in landfills. Sorption can be modeled in wastewater by the following linear equation (Equation 1.1), where C_S is the quantity of a compound sorbed per liter of wastewater, K_d is the sorption constant, X_{ss} is the concentration of suspended solids in the wastewater, and C_d is the dissolved concentration of the compound in wastewater (Joss et al., 2006; Ternes et al., 2004).

$$\textbf{Equation 1.1}: \quad C_S = K_d \times X_{SS} \times C_d$$

If K_d, X_{ss}, and C_d are known, this equation can be used to predict the removal of a compound in the wastewater treatment process by sorption. While X_{ss} and C_d may be determined experimentally with relative ease, the calculation of K_d (sorption constant) is unique for each compound. Thus, a model capable of predicting the K_d values of individual OWCs would be of value.

Biodegradation (i.e., biotransformation) comprises another potential removal mechanism for OWCs during wastewater treatment. Here, a dissolved organic molecule is broken down by WWTP bacteria, generally during the activated sludge stage of secondary treatment. Because OWCs generally occur in wastewater at relatively low concentrations ($<10^{-4}$ g/L), degradation frequently occurs as a result of co-metabolism alongside other, higher concentration wastewater nutrients. This is because the sludge bacteria require a larger concentration of organic carbon than can be met using OWCs alone (Ternes, 1998; Heberer, 2002; Ternes et al., 2004). This co-metabolism scenario frequently results in only partial biotransformation of OWCs; however, another possible biodegradation scenario involves mixed substrate growth where bacteria use the trace OWC as a carbon/energy source and may completely mineralize it (Ternes et al., 2004). This could result in more complete removal of the OWC.

Both of the biodegradation reactions occurring during activated sludge treatment can be modeled using a "pseudo first order" differential equation (Equation 1.2), where $\frac{dC}{dt}$ is the rate of change in dissolved concentration of the OWC over time t, k_b is the pseudo first order rate constant, X_{ss} is the concentration of suspended solids in the wastewater (used here as an approximation for MLVSS), and C is the dissolved concentration of the OWC.

$$\textbf{Equation 1.2}: \quad \frac{dC}{dt} = -k_b \times X_{SS} \times C$$

Solving for C as a function of t yields the following expression (Equation1.3), where C(t) is the dissolved concentration of the OWC at time t.

$$\textbf{Equation 1.3}: \quad C(t) = C_0 \times e^{-k_b \times X_{SS} \times t}$$

These pseudo first order equations assume that the degradation of the OWC proceeds at a rate that is proportional to the OWC and suspended solids concentrations. The $\frac{dC}{dt}$ term accounts for any degradation of the compound, partial or complete. These differential equations with their stated assumptions cannot differentiate between mechanisms of degradation. It's also important to note that when using this equation to model a real WWTP activated sludge system there are inflows and outflows to consider. Thus an OWC only has a limited amount of time (based on the hydraulic retention time of the reactor) to be degraded by sludge until it may exit the reactor with the out flow. Just like the K_d sorption coefficient, the k_b biodegradation rate constant is unique for each compound. The k_b values are dependent upon characteristics of each OWC molecule and the activated sludge and they are difficult, expensive, and time-consuming to measure. Because of the limited amount of k_b data currently available today and the huge, ever changing pattern of OWC compounds, it would be valuable to have a numerical model for predicting, rather than directly measuring, k_b values for OWCs of interest.

The rate of OWC biotransformation depends on both environmental factors and molecular properties of the OWC. Both types of information are encapsulated within a k_b rate constant. Examples of environmental factors that influence OWC degradation rate include: sludge redox conditions (aerobic, anaerobic or denitrifying) and sludge age. These factors affect the microbial population diversity of the sludge, which, in turn, affects the degradation rates of OWCs. In general, the rate of OWC biodegradation tends to increase in older, more diverse, more acclimated sludges (Ternes et al., 2004). Regarding molecular properties, a wide variety of shape and size indices and also certain structural, electronic, and other chemical properties can affect the rate of OWC degradation. Specific information on these so called “molecular descriptors” affecting OWC rates of biodegradation will be discussed in greater detail in the next section. In fact, investigating which molecular descriptors are correlated with OWC rates of degradation is one of the primary focuses of this study.

To summarize, WWTPs are important sources of OWCs into surface waters. While not designed for their treatment, these plants have the potential to remove trace organic contaminants by sorption, biodegradation, hydrolysis, volatilization and photolysis, primarily during secondary treatment. A future goal for WWTPs should be nearly complete or complete removal of OWCs from wastewater. To achieve this task, greater knowledge of WWTP OWC removal is needed, in particular rates of sorption and biodegradation processes. This study presents a method for predicting k_b biodegradation rate constants of OWCs in WWTP activated sludge.

1.3 QSBR Modeling

1.3.1 Modeling Benefits

As of 2006, it was estimated that only 240,000 of roughly 8,400,000 commercially available substances worldwide had been inventoried and/or regulated (Muir, 2006). Over the past 30 years in the US alone, as many as 100,000 pharmaceutical, cosmetic, food additive and pesticide compounds have been registered for commercial use (Muir, 2006). Because of the huge number of compounds and the difficulty of the laboratory methods used to measure k_b values, it is not feasible to calculate k_b rate constants for every OWC of interest. Instead, a model capable of predicting k_b values based on OWC molecular descriptors is a much more desirable option.

A model capable of predicting k_b based on molecular descriptors would have several principal benefits. First, accurate estimation of OWC k_b values would improve fate and transport assessment of OWCs in WWTPs; identifying which OWCs pass through WWTP processes, and aiding in identification of particular compounds of concern. Second, better estimation would also enable better modeling of OWC fate and transport in receiving surface waters. Third, the identification of key molecular descriptors affecting degradation rate constants could illuminate the processes and mechanisms by which OWCs are removed during biological treatment, potentially revealing information that could be used to improve current treatment methods. Finally, key molecular descriptor information could suggest why certain contaminants biodegrade quickly and why others take longer. This knowledge might eventually enable chemical manufacturers to design more readily biodegradable compounds.

1.3.2 Quantitative Structure-Biodegradability Relationships

Several existing “quantitative structure-biodegradability relationships” (QSBRs) have attempted to predict the biodegradability of OWCs based on their molecular characteristics (Okey and Stensel, 1996; Papa et al., 2007; Yang et al. 2006; Hongwei et al, 2006; Hao et al., 2009). These models are one subset of more general “quantitative structure-activity relationships” (QSARs), which are used to predict a wide variety of chemical behaviors based on known molecular properties. The process of creating a QSAR is relatively simple. First, a dataset of compounds with known values of the desired property is collected. Next, suspected molecular descriptors of importance to the predicted property are calculated for each compound in the dataset. Various forms of regression analysis are then used to correlate molecular descriptors and the known values of the predicted property. Finally, if successful, a model equation will be developed, which will predict the property of interest as a function of the most pertinent molecular descriptors. Such QSAR models are generally validated both internally, using statistical parameters such as R^2 or q^2, and externally by measuring the property of interest for a new, previously unmeasured compound and comparing the predicted and measured values to each other.

Table 1.1 displays a summary of previously published QSBR studies, including the key molecular descriptors reported in each investigation. All studies were attempting to model the biodegradation of one or more OWCs, but each paper had a slightly different focus. Thus, there is some variation in the types of compounds included within Table 1.1. Also, there is some significant variability in sludge conditions among the surveyed studies.

Table 1.1: A summary of previously published QSBR studies. Descriptors listed in bold text were found to be especially well correlated with biodegradation rates. Asterisk (*) indicates one study in which OWC biodegradation was modeled for environmental conditions without sludge.

QSBR Studies				
Authors	Year	Key Molecular Descriptors	Compounds of Interest	Sludge Condition
Okey and Stensel	1996	**heteroatom** and carboxyl group presence, **size**, **charge** and complexity	wide array of organics	aerobic
Yang et al.	2006	**total energy**, **molecular refractivity**, EHOMO, LogP, Gibbs free energy	aromatic compounds	anaerobic
Yang et al.	2006	**EHOMO**, 2nd order molecular connectivity index	nitrogenous compounds	anaerobic
Papa et al.	2007	**size**, **aromatic bonds**, HOMO-LUMO gap, atomic energy, heat of formation, LogP	antibiotics	no sludge*
Hao et al.	2009	2nd and 4th order connectivity indicies	nonylphenol Isomers	aerobic

Despite the differences in format for the studies summarized with Table 1, many papers reported similar sets of molecular descriptors to be important for prediction of OWC biodegradation. For instance, in three of the five studies “EHOMO” (energy of the highest occupied molecular orbital) or “EHOMO-ELUMO gap” (the difference in energy between the highest occupied molecular orbital and the lowest unoccupied molecular orbital) were reported as relevant to modeling biodegradation. Also, “LogP” and “molecular size” were each reported in two of the five studies, with molecular size being listed twice as a highly correlated descriptor. While the models summarized in Table 1 possess varying degrees of predictive ability, the occurrence of several same descriptors in each suggests that these molecular characteristics strongly impact biodegradation rate.

Identification of molecular descriptors that are highly correlated to OWC biodegradation can reveal information about the processes of activated sludge biodegradation, because the best-correlated descriptors can provide information about which reactions comprise the rate-determining step during biodegradation. Biodegradation rates can be affected by microbial uptake/transport rates or OWC binding to enzyme active sites (Parsons and Govers, 1990). If microbial uptake/transport rate is the rate determining step, a compound's ability to diffuse though lipid-rich cell membranes would likely be of importance, such that polarity (as parameterized using the Log of the octanol water partitioning coefficient - LogP) may affect be highly correlated with biodegradation rate (Parsons and Govers, 1990). Alternatively, if enzyme binding or transformation is the rate determining step, properties pertaining to the OWC's electronic structure and steric hindrance (e.g., EHOMO, ELUMO, EHOMO-ELUMO gap, total energy, molecular size) would be highly correlated with biodegradation rates (Parsons and Govers, 1990). Molecular electronics and sterics are portrayed by molecular descriptors such as (weight or surface area) and certain connectivity indices. In this way, identification of molecular descriptors that are highly correlated with biodegradation rate constants can reveal information about the nature of the biodegradation mechanism itself.

While some authors have achieved reasonable success in development of accurate QSBR for biodegradation rates of certain OWC subset classes, it has been suggested that QSBR may be unsuitable for predicting "all" OWC biodegradation rates. Thus, most models exhibiting high degrees of internal and external validation to date have used a small dataset of structurally related compounds. In contrast, studies focusing on larger, more diverse sets of OWCs tend to exhibit smaller internal validation coefficients; e.g., a study by Okey and Stensel (Table 1.1), which achieved R^2 = 0.72 for 131 compounds. Parsons and Govers (1990) have summarized this dilemma as follows: "In general, these relationships are limited to structurally related compounds. There is no general relationship between biodegradability and chemical structure.” Though there is reason to doubt the ability of general QSBR models for large and diverse molecular sets, other literature studies suggest QSBR may be effective at modeling smaller, more structurally similar groups of compounds (Yang et al., 2006, Papa et al., 2007, Hao et al., 2009).

1.4 Study Objectives

The primary objective of this study was to create a QSBR model capable of predicting OWC biodegradation rates during activated sludge treatment, based on OWC molecular descriptors. Achievement of this objective is expected to aid in the advancement of fate and transport models for mapping OWCs in natural and engineered water systems. Molecular descriptors of importance were also identified, in an effort to improve understanding of reactions underpinning biological OWC removal. Finally, the predictive ability and statistical validity of the resulting QSBR model was compared to other predictive modeling approaches, to understand the advantages and disadvantages of each approach.

Chapter 2 - Methodology

2.0 Overview of Methodologies

The paragraphs of this chapter summarize formulation and validation of a quantitative structure-biodegradation relationship (QSBR). The formulation of the QSBR is characterized by three distinct steps. First, a dataset of biodegradation rate constants (k_b) was compiled from a variety of literature sources. Second, desired molecular descriptors were calculated for each compound represented within the k_b dataset, using molecular modeling software. Finally, the regression analysis was performed using statistical software. Following the formation of the model, both *internal* and *external* validations were performed.

2.1 Compilation of the k_b Dataset

Degradation rate constants (k_b) were collected for removal of OWCs during secondary (i.e., biological) municipal wastewater treatment were gathered from scientific literature. Though emerging contaminants were the focus of this study, all organic molecules were considered suitable for inclusion in the training dataset; provided the authors were directly measuring the change in concentration of a compound over time in aerated activated sludge from a WWTP. Papers reporting k_b values derived changes in concentration of chemical oxygen demand (COD) or estimated from biodegradation model were not included. As a result, all values included in the k_b dataset correspond to studies using aerated semi-batch reactor setups with activated sludge and dosing of a target OWC into synthetic wastewater. The total suspended solids (TSS) concentrations were recorded for all experiments; however, detailed characterizations of the sludges were not required. In all studies, liquid chromatography with or without mass spectrometry was used to quantify the change in OWC concentration over time. Based on these criteria, k_b values were collected for 65 different organic compounds from nine different papers: Majewsky et al. (2011), Li 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). For situations in which multiple

k_b values were found for the same compound, the arithmetic average of all values was computed, such that a single value could be used in the k_b dataset.

Of the 65 k_b values used in the dataset, all were reported as either first order or pseudo-first order rate constants. The first order k_b rate constants were converted to pseudo first order rate constants by dividing by the reported TSS used in each experiment. TSS values were assumed to be constant in time for all experiments. Following conversion to pseudo first order, the log of each value was taken. The log values of the 65 k_b compounds were used in subsequent regression models. The k_b data set is displayed in Appendix A.

2.2 Molecular Descriptor Selection and Calculation

Selected molecular descriptors were calculated for each of the 65 compounds corresponding to measured k_b values in the training dataset, and for three additional compounds required for external validation (see Section 2.4). These molecular descriptors were chosen from a large group of descriptors that had been previously used in QSBR studies (Okey and Stensel, 1996; Papa et al., 2007; Yang et al., 2006; Hao et al., 2009). Descriptors that had been shown to be best correlated with k_b rate constants in previous studies were selected for use in this study. Calculated molecular descriptors included: total molecular energy, energy of the highest occupied molecular orbital (EHOMO), energy of the lowest unoccupied molecular orbital (ELUMO), dipole strength, Gibbs free energy, heat of formation, the log of the octanol/water partitioning coefficient (LogP), molecular refractivity, accessible surface area, molecular surface area, and molecular weight. Appendix B summarizes these descriptors and their calculated values for each of the 65 compounds in the k_b dataset plus the three selected external validation compounds.

All molecular descriptors were calculated using the molecular modeling software ChemBio3D Ultra 13.0 (CambridgeSoft: Cambridge, UK). Computational models of the molecular structures for all evaluated chemicals were first built using the molecular modeling software; by entering the compound name, the IUPAC name, or the SMILE file into the ChemBio3D software. The software then recognized the molecule names and generated their associate structures. Compounds with specific stereochemistry were generated in such specific configurations; however, only one specific stereo isomer was generated for racemic compounds. The molecular structures were then geometrically optimized using molecular mechanics optimization (MM2) until energy convergence within 0.1Kcal/mol was achieved. Following geometric optimization, molecular structures underwent a semi-empirical Austin Model 1 (AM1) quantum optimization using the GAMESS computational chemistry interface within ChemBio3D. This procedure was also carried until energy convergence was within 0.1kcal/mol.

Molecular descriptors for the quantum-optimized structures were computed using a variety of computational chemistry interfaces available within the ChemBio3D suite. Total energy, dipole strength and the molecular surfaces required for calculation of EHOMO and ELUMO were computed using an *ab initio* Hartree Fock method within the GAMESS interface. Once molecular surfaces were calculated, the HOMO and LUMO of each molecule were plotted along with their associated energies. Gibbs free energy, heat of formation, LogP, and molecular refractivity were estimated using the ChemProp Pro interface within ChemBio3D. Accessible area, molecular area, and molecular weight were calculated using the ChemProp Standard interface. Calculated molecular descriptors are shown in Appendix B.

2.3 QSBR Regression Analysis

Calculated values for the selected molecular descriptors, and their associated k_b values, were entered into Minitab 16 for statistical analysis (Minitab Inc.: State College, PA), specifically regression analysis. Minitab's "Best Subsets" regression functionality was used to assess preliminary multiple linear regression (MLR) models for relationships between $\text{Log}k_b$ values in the training dataset and all computed molecular descriptors. The Best Subsets function displays the best two regressions (based on R^2) of each number of descriptors (in this case from one to ten). Because 15 of the 65 compounds had at least one missing descriptor value (due to limitations of the molecular modeling software), only 50 compounds could be evaluated using the best subset functionality. The molecular descriptors which were most frequently identified by the Best Subsets functionality reports were retained for future use: total energy, EHOMO, ELUMO, dipole magnitude, Gibbs free energy, heat of formation, LogP, accessible area, molecular area and molecular weight. In contrast, molecular descriptors which were least frequently identified by the Best Subsets functionality were discarded from the analysis. Of the 65 molecules represented in the training dataset, 58 molecules exhibited complete sets of the ten retained descriptors noted above. Subsequently, the Best Subset regression function was re-run, using these 58 molecules and their chosen descriptors to factor in the eight added molecules. This procedure was then repeated for subgroups of the training dataset, comprising sets of compounds taken from the three largest papers collected during compilation of the k_b dataset: Wick et al. (2009), Andreozzi et al. (2005) and Urase et al. (2005).

Following completion of best subset regression analyses for the entire training dataset and the data subsets from each of the three papers noted above, one preferred MLR model was chosen for each group of molecules. These four models were selected based on a variety of internal validation statistical parameters: R^2 (coefficient of determination), adjusted R^2 (adjusted to account for an increase in the number of descriptors), Mallows' Cp (compares the precision and bias of each best subset model against the model including all of the other descriptors), and S (standard distance from regression line in units of response) (see Appendix E for internal validation parameter equations). Generally, the "best" MLR model for each dataset was the one that required the fewest molecular descriptors while still achieving good R^2. Best-fit coefficients for each of the four MLRs selected based on the best-subset analyses were calculated using the "regression" functionality within Minitab. Standard errors for each coefficient were also computed, and the diversity among independent variables (molecular descriptors) and dependent variables (k_b values) was characterized via calculation of range, mean, and standard deviation for each parameter.

2.4 External Validation

External validation was performed by directly measuring the k_b values of three previously unmeasured OWCs: metformin, benazepril and warfarin. Measured values were then compared with model-predicted k_b values. These three compounds were chosen based on their high loading rates in wastewater, their lack of literature k_b measurement and their availability in our lab due to other research group experiments (Ottmar, 2010).

Reagents for the validation compounds were acquired from Fisher Scientific Inc. (Waltham, Ma). Activated sludge was obtained from the aerobic sludge basins of the Charlottesville WWTP in March of 2011 and immediately transported to the laboratory. The sludge was then divided into three 4-L Erlenmeyer flasks, each with a vacuum side spout. Each flask was aerated to maintain DO levels around 5 mg/L, using three small aquarium pumps. Synthetic wastewater was delivered continuously at approximately 1.5 mL/min, using a peristaltic pump. The synthetic wastewater was prepared based on a formulation from the Organization for Economic Cooperation and Development (OECD) (Pholchan et al., 2008), with the addition of several micronutrient salts (Jefferson et al., 2000). The exact formula for an 80x stock solution is displayed in Appendix C. The resulting mixture was diluted to 500 ml with deionized water, autoclaved at 120 °C for 20 minutes, and then diluted 80 times with deionized water. The activated sludge was allowed to grow for over a month before experiments began, with excess sludge dripping out of the vacuum side spouts. This process resulted in a gradual flushing out of background contaminants from the bioreactors (Ottmar, 2010).

The biodegradation rate constants (k_b) for three external validation compounds were measured separately using a batch reactor setup. For each compound, three 1-L Erlenmeyer flasks were prepared containing activated sludge, an initial dose of synthetic wastewater, the target compound, and deionized water in such a proportion as to create 500 mL of solution exhibiting 1-g/L TSS concentration (assumed to be equal to MLVSS and X_{ss}), ≈1200 mg/L initial COD concentration, and 1 mg/L of the target compound. Sludge concentrations (as approximated by TSS) may have been lower than average activated sludge TSS levels, but this did not affect rate calculations due to the use of pseudo-first order rate constants. Two 1-L Erlenmeyer flasks were used as sorption controls, containing the exact same total volume and concentrations of sludge, COD, and the target compound; but with sludge that had been autoclaved at 120 °C for 4 h. To ensure that the sludge in the sorption control reactors was indeed completely deactivated (such that it could not mediate biotransformation of the target OWC), 20 mg of sodium azide was added to each reactor. Two additional 1-L Erlenmeyer flasks were used as positive control reactors, containing only the initial dose of synthetic wastewater, the target compound, and deionized water. The COD and target compound concentrations of these two reactors were the same as the other five, just without any sludge. Table 2.1 summarizes the contents of the seven reactors used in these experiments. Both the metformin experiment and one of the benazepril experiments were only given a single, initial dose of synthetic wastewater solution, while the other benazepril and warfarin experiments were re-dosed with the same initial amount of synthetic wastewater daily.

Aeration was achieved using aquarium pumps, as noted above. Reactors were placed on magnetic stir plates and stirred at a constant rate to ensure consistent, complete mixing. Make-up water was used to neutralize evaporative losses and maintain a constant volume in each reactor.

Table 2.1: Biodegradation batch reactor setup. Initial conditions in each of the seven batch reactors are displayed. The "number" column states the number of reactors of a certain type.

Reactor	Number	Total Volume (mL)	[Sludge] (g/L)	[COD] (mg/L)	[Target Compound] (mg/L)
Experimental	3	500	1	≈ 1200	1
Sorption Control	2	500	1 (dead)	≈ 1200	1
Positive Control	2	500	0	≈ 1200	1

Once aeration and mixing had been initiated in each reactor, the reactors were left to equilibrate for about 20 min. Following this brief equilibration period, samples were taken from each reactor at semi-regular intervals until all reactors were observed to approach constant concentrations of the target OWC. Because the three target compounds reached equilibrium concentrations at different times, the total number and frequency of samples collected differed among the three external validation experiments. At selected sampling times, 5-mL aliquots were pipetted out of each reactor. The samples were then filtered through either a 0.45 -µm Buchner funnel system or a 0.45 -µm syringe filter. After filtration, a portion of each sample was used to analyze COD concentration, using a commercial kit. The remainder of each sample was refrigerated until HPLC analysis could be performed.

Concentrations of the OWCs in each sample were analyzed using an Agilent Technologies 1200 Series HPLC system. HPLC methods were slightly different for each of the four target compounds. A summary of each method is presented in Table 2.2. Calibration curves were generated using concentration standards over the range 0.4 - 1.6 mg/L for metformin, 0.25 - 2 mg/L for benazepril and 0.1 - 2 mg/L for warfarin. Best-fit calibration equations were then used to compute sample OWC concentrations based on measured peak areas of each of the samples.

Table2.2: Summary of metformin, benazepril and warfarin HPLC methods.

Metformin		
Flow Rate (mL/min.)	0.3	
Injection Volume (µL)	50	
Column Brand/type	Phenominex Hypersil C18	
Column Dimentions (mm)	250 x 2	
Wavelength (nm)	232	
Retention Time (min.)	22.6	
Gradient Method		
Time (min.)	% H_20	% Acetonitrile
0 to 3	85	15
3 to 22	85-60	15-40
22 to 24	60-85	40-15
Benazepril		
Flow Rate (mL/min.)	0.8	
Injection Volume (µL)	50	
Column Brand/type	Phenominex EnviroSep-PP	
Column Dimentions (mm)	125 x 3.2	
Wavelength (nm)	245	
Retention Time (min.)	11.8	
Gradient Method		
Time (min.)	% H_20	% Acetonitrile
0 to 6	80	20
6 to 12	80-60	20-40
12 to 16	60-80	40-20
Warfarin		
Flow Rate (mL/min.)	1.2	
Injection Volume (µL)	70	
Column Brand/type	Phenominex EnviroSep-PP	
Column Dimentions (mm)	125 x 3.2	
Wavelength (nm)	310	
Retention Time (min.)	9.9	
Gradient Method		
Time (min.)	% H_20	% Acetonitrile
0 to 5	80	20
5 to 12	80-65	20-35
12 to 15	65-80	35-20

Plotting the concentrations of each selected OWC in the seven reactors against time, k_b rates for each compound were computed. Reductions in target compound concentrations in the three regular reactors arising from sorption, volatilization, and hydrolysis were subtracted from apparent removal, using appropriate controls, to identify what fraction of removal corresponded to biotransformation. This method is shown in Equation 2.1.

Equation 2.1: Isolating the biodegradation fraction of apparent removal

$$Isolated\ Biodeg.(C) = [Ave. Positive\ Controls\ (t)] - ([Ave. Sorption\ Controls\ (t)] - [Regular\ Reactor\ x\ (t)])$$

The change in target compound concentration resulting directly from biodegradation over time was determined by plotting the resulting "isolated biodegradation", i.e., C(t) values in each regular reactor sample, against time. The initial concentration of the target compound in each reactor (C_0) was computed as the average concentration in the first collected sample of the positive control. Assuming pseudo-first order k_b rates and a constant sludge concentration of 1 g/L, Equation 1.3 was used to solve for k_b, yielding Equation 2.2.

Equation 1.3: Pseudo-first order solution (assuming constant X_{ss})

$$C(t) = C_0 e^{-K_b \times X_{ss} \times t}$$

Equation 2.2: pseudo-first order k_b rate (constant X_{ss})

$$k_b = \frac{-\ln(\frac{C}{C_0})}{t \times X_{ss}} = \frac{k_{b1}}{X_{ss}}$$

By plotting ln[C(t)/C_0] vs. t and calculating the negative slope of the best-fit linear regression line, a first order (k_{b1}) rate constant was computed for each of the four external validation OWCs. First order k_{b1} rate constants were divided by TSS concentration (X_{ss}) for conversion into pseudo-first order rate constants (k_b). Equation 2.2 also shows this relationship between first order k_{b1} and pseudo-first order k_b. For comparison, the first order (k_{b1}) differential equation as well as solutions for C(t) and k_{b1} are listed in Equations 2.3, 2.4 and 2.5. First order k_{b1}s do not utilize X_{ss} (TSS) in their differential equations and are in units of time^{-1}. Only pseudo-first order k_bs (units of $\frac{Volume}{(mass)(time)}$) were used in QSBR modeling in this study.

Equation 2.3: First order differential equation

$$\frac{dC}{dt} = K_{b1} \times C$$

Equation 2.4: First order solution

$$C(t) = C_0 \times e^{-K_{b1} \times t}$$

Equation 2.5: First order k_{b1} rate

$$k_{b1} = \frac{-\ln(\frac{C}{C_0})}{t}$$

For re-dosing experiments, wherein suspended sludge concentration (X_{ss}) was not constant over time, pseudo-first order rates (k_b) were generated by dividing derived first order rates (k_{b1}) by the average TSS concentration (X_{ss}) over the experiment. This assumption was only needed for comparing the biodegradation rates calculated in the benazepril dosing and re-dosing experiments.

Chapter 3 - Results and Discussion

3.1 Biodegradation Rate Constant (k_b) Data Set

In total, 82 different k_b values were collected for 65 unique compounds from nine different literature sources. These 82 values, as well as their literature sources, are displayed in Appendix A. This table reveals interesting information regarding literature trends in WWTP biodegradation for OWCs. Though the search for k_b literature was fairly extensive, in total only nine papers were found that met the desired search criteria (see Methods, 2.1). This data is relatively sparse and seemingly delocalized, as no major review papers were discovered. Of these nine sources, the oldest dates back to only 2005, suggesting the use of semi-batch reactors to conduct OWC sludge biodegradation kinetics experiments is relatively recent. Additionally, the comparison of duplicate k_b measurements reveals a surprising amount of variation between different papers measuring the same compound. For example, in Appendix A there are five different k_b values from three different sources for 17β- estradiol. Three of the measurements are from the same source (Li et al., 2005), and these are in relative agreement with each other (1.86, 1.53 and 1.95 L/g-h); however, they are quite different from the other two reported values (1.16 L/g-h) (Zeng et al., 2009) and 0 L/g-h (Urase et al., 2005) This type of variability is evident for several other compounds as well. Sulfamethoxazole, naproxen, ibuprofen, gemfibrozil, fenoprofen and diclofenac exhibit up to one order of magnitude difference among duplicate k_b values.

Variability among reported k_b measurements from different sources for the same compound may indicate differences in experimental methodology and/or experimental error, even within carefully selected sources employing similar semi-batch reactor experimental setups. There is little information in the literature describing standard methods for sludge biodegradation kinetic experiments. The general experimental criteria described in the Methods (section 2.1) may not be specific enough to capture differences in experimental parameters such as sludge preparation (age, origin, aeration level, temperature, biodiversity, etc.), synthetic wastewater composition, initial dose of the target OWC, and other initial conditions of the bioreactors. Also, there may be subtle differences in rate calculations among the nine studies. Procedures for calculating first order or pseudo first order rates are often reported in little detail within these literature sources. These general inconsistencies/errors in experimental procedures may account for the often large degree of variability apparent among k_b measurements of the same compound from different authors. These variations and data scarcity are two major difficulties associated with modeling OWC biodegradation kinetics for a wide variety of compounds at present.

3.2 Molecular Descriptor Data Set

As described in Methods Section 2.2, a chosen set of molecular descriptors were calculated for each of the 65 compounds in the k_b data set. These molecular descriptors were also computed for the three selected validation compounds. All calculations were performed using the molecular modeling software ChemBio3D Ultra 13.0 (CambridgeSoft: Cambridge, UK). Appendix B displays the calculated values of the selected molecular descriptors, including: total molecular energy (Total E), energy of the highest occupied molecular orbital (EHOMO), energy of the lowest unoccupied molecular orbital (ELUMO), EHOMO-ELUMO energy gap (ELUMO-EHOMO), dipole strength, Gibbs free energy (Gibbs), heat of formation (HoF), the log of the octanol/water partition coefficient (LogP), molecular refractivity (MR), accessible surface area (A Area), molecular surface area (M Area) and molecular weight (MW). Of the 65 compounds, 15 exhibited one or more missing descriptors from the ChemProp Pro interface (Gibbs, HoF, LogP and MR). Appendix B also includes the log (base 10) of the literature k_b value for each compound. Where it was necessary to take the average of several values for the same compound, the mean value was computed before the log transformation was applied.

One other potential source of error in the computation of molecular descriptors for inclusion in the QSBR modeling, which has been largely ignored in the literature, is the simplification of racemic mixtures into one specific stereo isomer. For example, the psychotropic agent Doxepin (one of the 65 compounds in the k_b data set) occurs as a racemic mixture of *cis* and *trans* conformal isomers, but only one isomer was built in ChemBio3D for calculation of molecular descriptors. Several other compounds in the k_b and validation datasets also occur as racemic mixtures of isomers. It could be problematic to assign only one structure for a racemic mixture, because many isomers exhibit different properties, potentially including biodegradation rate constant. Because of this, simplifying racemic mixtures into one specific isomer could affect the accuracy of a QSBR model. In a 2004 paper, Berset et al. attempted to display the enantioselectivity of polycyclic musk removal in WWTP sludge. In short, the study was performed by measuring the "enantiomeric ratio" (the ratio of the concentration of one isomer over another) of pairs of enantiomers/diastereomers before and after treatment by sewage sludge. While most sets of isomers exhibited little change in enantiomeric ratio before and after treatment, there were several racemic mixtures that did display preferential degradation of a particular isomer. This suggests that the removal of certain isomer mixtures by WWTP sludge could be stereoselective. These isomeric effects are generally overlooked in the OWC WWTP removal literature. For simplification purposes, mostly pertaining to software limitations, these effects have been also overlooked again in this study. But this could be an interesting and important area of future work.

3.3 Regression Analysis and QSBR Formulation

Regression analysis for Logk_b (dependent variable) and selected molecular descriptors (independent variable) was performed using Minitab 16 statistical software (Minitab Inc.: State College, PA). As described in Methods Section 2.3, 50 compounds with complete descriptor value sets were evaluated using the "Best Subsets" regression, to evaluate all possible multiple linear regression models (MLRs). Following this initial analysis, the descriptor "Molecular Refractivity" (MR) was dropped from further regression analysis, because it occurred infrequently among the output of the preliminary best subsets analyses. It also displayed a high degree of correlation with Molecular Area (M Area) (Pearson

Correlation value = 0.956), and MR values were unavailable for eight of the compounds in Appendix A. Dropping MR freed up these eight compounds for use in a subsequent Best Subsets analysis.

Results from the Best Subsets following exclusion of the Molecular Area descriptor are depicted in Figure 3.1. This output gives an overview of what combinations of the ten descriptors in various MLRs give the best internal validity (R^2) without actually computing regression equations. Each row of the Figure 3.1 represents one MLR model incorporating the stated number and combination of selected molecular descriptors. The predictive accuracy of each MLR model was analyzed using several statistical parameters: R^2, Adjusted R^2 (R^2 (adj)), Mallows' Cp , and S, as described in Methods Section 2.5. Adjusted R^2 and Mallows' Cp are particularly useful in differentiating among multiple possible MLR candidates, because adjusted R^2 accounts for the number of descriptors used with the model, and Mallows' Cp compares the precision and bias of each best subset model with a model including all other descriptors. Due to its relatively high values for both R^2 and R^2 (adj), and its minimization of Mallows' Cp and S (standard distance from regression line in units of response), a MLR model with seven descriptors was chosen as the "best" MLR model. These seven descriptors include: Total E, EHOMO, Dipole, HoF, A Area, M Area and MW.

58 Complete Compounds

Best Subsets Regression: Logkb versus 10 Descriptors

```
Response is Logkb
58 cases used
                                         T
                                         o     D       A M
                                         t E E i G
                                         a H L p i   L A A
                                         l O U o b H o r r
                       Mallows             M M l b o g e e M
Vars  R-Sq  R-Sq(adj)       Cp        S  E O O e s F P a a W
   1  11.1        9.6     29.8  0.73859                X
   1  10.2        8.6     30.7  0.74268                  X
   2  19.8       16.9     23.6  0.70786          X     X
   2  19.3       16.3     24.1  0.71039          X       X
   3  34.3       30.6     12.0  0.64701    X     X     X
   3  33.0       29.3     13.2  0.65325    X     X X
   4  39.1       34.5      9.4  0.62853  X X     X         X
   4  38.8       34.1      9.7  0.63024  X X       X       X
   5  41.7       36.1      8.9  0.62066  X X   X   X       X
   5  41.3       35.7      9.3  0.62291  X X     X       X X
   6  45.9       39.5      7.0  0.60388  X X     X     X X X
   6  45.8       39.4      7.1  0.60449  X X       X   X X X
   7  49.1       41.9      6.0  0.59182  X X   X   X   X X X
   7  47.5       40.2      7.5  0.60076  X X   X X     X X X
   8  49.7       41.5      7.5  0.59427  X X   X   X X X X X
   8  49.4       41.1      7.7  0.59589  X X X X   X   X X X
   9  50.2       40.8      9.0  0.59746  X X X X   X X X X X
   9  49.7       40.3      9.4  0.59996  X X   X X X X X X X
  10  50.2       39.6     11.0  0.60378  X X X X X X X X X X
```

Figure 3.1: Minitab "Best Subsets" regression function output for the 58 compounds in the logk$_b$ set. Each row represents a MLR with a certain number and combination of descriptors. The "Vars" column states the number of descriptors ("variables") and the "X" marks below a stated descriptor indicates the use of that descriptor within a particular MLR. The two highest R^2 values are shown for each number of variables (except for the MLR in which all ten variables are used together). The chosen "best" regression model is highlighted in green. Generated with Minitab 16 (Minitab Inc.: State College, PA).

As seen in Figure 3.1, the internally validation statistics (R^2, R^2 (adj), Mallows Cp and S) exhibit generally poor performance for all of the MLR subsets run. Even the chosen "best" MLR model, with seven descriptors, displays modest R^2 and R^2 (adj) values of 49.1% and 41.9%, respectively. The model's S value (0.59, as reported in units of logk$_b$) is similarly poor, considering that the logk$_b$ values for the 58 compounds ranges only from -2.48 to 0.58 with a standard deviation of about 0.77. Because the S value indicates the standard distance each datum falls from the regression line, reported k$_b$ values fall almost one standard deviation away from the regression line, on average, for even the best MLR model in Figure 3.1. The highest R^2 value in Figure 3.1 is 50.2%, which indicates that there is little correlation between the selected ten descriptors and Logk$_b$ values. Additionally, no single descriptor appears to be particularly well correlated with the 58 values of Logk$_b$, based on the very low R^2 values for both single descriptor MLR models: 11.% for A Area alone, and 10.2% for M Area alone. Because the Best Subset regression output lists the "best" two models of each number of variables, it can be concluded that the other eight descriptors are even less well-correlated with Logk$_b$ for the dataset of 58 OWCs.

Equation 3.1 displays the model QSBR equation corresponding to the "best" model MLR from Figure 3.1, with best-fit coefficients for seven molecular descriptors. Table 3.1 displays P-values for each descriptor coefficient, as well as their standard errors, for assessment of which descriptors have the most effect on Logk$_b$ prediction within the model.

Equation 3.1: "Best" QSBR model based on the 58-compound logk$_b$ data set

$$Logk_b = 5.254 - 0.00000029\, Total\ E + 0.4178\, EHOMO - 0.07669\, Dipole - 0.00173\, HoF - 0.01654\, A\ Area + 0.0315\, M\ Area - 0.01142\, MW$$

Table 3.1: Descriptor coefficients, standard errors and P values of the "Best" QSBR of the 58 compound logk$_b$ data set. All parameters generated by Minitab 16 (Minitab Inc.: State College, PA).

Descriptor	Coeff.	SE Coeff.	P
Constant	5.254	1.299	0.000
Total E	-2.9E-07	1E-07	0.005
EHOMO	0.4178	0.1092	0.000
Dipole	-0.07669	0.04282	0.079
HoF	-0.00173	0.000389	0.000
A Area	-0.01654	0.006589	0.015
M Area	0.0315	0.01175	0.010
MW	-0.01142	0.003706	0.003

Though the Best Subsets regression output for the 58-compound dataset indicates that A-Area and M-Area have the highest single-value correlations with Logk$_b$, the P-values in Table 3.1 indicate that several other variables are equally or more significant in predicting Logk$_b$ within the seven descriptor MLR. P-values ≤0.15 indicate a statistically significant relationship between a descriptor and Logk$_b$; therefore, all of the descriptors display statistical significance.

Despite the presence of many descriptors displaying statistically significant relationships with $Logk_b$ in the 58 compound “best” QSBR, internal validation statistics indicate that the “Best” QSBR model has low predictive ability. This model will be further tested in its ability to predict the $Logk_b$s of unknown compounds from the external validation set later in this chapter. The low predictive ability of the overall QSBR model is not surprising. One contributing factor may be the variability in $Logk_b$ data, as noted in Section 3.2. Experimental inconsistencies or errors could be affecting the reliability of the k_b data set and, in turn, lowering the predictive ability of the model. Additionally, as discussed in Introduction Section 1.3, effective QSBRs are generally thought to be limited to structurally similar compounds (Parsons and Govers, 1990). It seems likely that the “best” QSBR referenced above may have drawn from a k_b data set that is too structurally diverse to yield good predictive ability. That is, the molecular diversity of the 58 compounds in the $Logk_b$ data set may be too great to be captured by one overarching set of linear relationships between descriptors. This could be true, in part, because molecular descriptors may vary in their relative importance for predicting k_b among different classes of OWC molecules, and these relationships may display non-linear or seemingly discontinuous trends among the large and structurally diverse data set of the 58 OWCs. This hypothesized basis for k_b variability and the poor predictive performance of MLR models in Figure 3.1, was evaluated by attempting to build MLRs amongst more specific sub-classes of molecules grouped within individual k_b studies. This process illuminated more specific/continuous relationships between descriptors and $Logk_b$ values, as described below.

3.4 Subset QSBR Formation

In an effort to create additional QSBR models with better internal validation results for smaller subsets of the available k_b data, the overall dataset was broken in three groups. These three groups corresponded to measurements made within the three largest papers from Appendix A: a 2009 study from Wick et al. focusing on beta blockers and psycho-active drugs, examining 15 compounds in total; a

2005 study from Andreozzi et al. focusing on aromatic structural analogs, evaluating 20 compounds in total; and another 2005 paper from Urase et al. taking a more general focus on pharmaceuticals and estrogens, evaluating 15 compounds overall. Andreozzi et al. and Wick et al. focused on relatively specific, somewhat homogenous OWC classes; whereas, Urase et al. examined a more diverse set of compounds. A list of the compounds evaluated in each of these papers can be seen in Appendix A. Three subgroups of the overall Logk_b data set, corresponding to the sets of chemicals evaluated in each of the three papers referenced above, were analyzed using Minitab's Best Subsets regression functionality and the same ten molecular descriptors from Figure 3.1. The k_b values used in each analysis were unique to each paper; that is, there was no averaging of multiple k_b measurements from different sources. Output from the three Best Subset regression analyses are summarized in Figures 3.2, 3.3., and 3.4, wherein green highlighting is used to indicate the best QSBR model for subgroup. The corresponding regression equations for each data subset comprise Equations 3.2, 3.3., and 3.4. Best-fit coefficients for all molecular descriptors utilized in the regression equations are summarized in Tables 3.2, 3.3., and 3.4, which also identify P-values and standard errors for these coefficients.

Wick et al. (Water Research 43 (2009) 1060–1074)

Best Subsets Regression: Logkb versus Total E, EHOMO, ...

```
Response is Logkb
14 cases used
                                         T
                                         o     D       A M
                                         t E E i G
                                         a H L p i   L A A
                                         l O U o b H o r r
                       Mallows             M M l b o g e e M
Vars  R-Sq  R-Sq(adj)       Cp        S  E O O e s F P a a W
   1  19.1       12.4     73.5  0.60371    X
   1   7.8        0.1     85.2  0.64452  X
   2  23.2        9.3     71.3  0.61430    X         X
   2  21.7        7.5     72.9  0.62038    X   X
   3  51.0       36.4     44.6  0.51452          X X       X
   3  31.7       11.3     64.5  0.60755    X   X     X
   4  58.2       39.6     39.2  0.50130          X X X     X
   4  56.9       37.8     40.5  0.50879    X     X X       X
   5  84.1       74.2     14.4  0.32750    X     X X     X X
   5  73.3       56.6     25.6  0.42503    X     X X   X   X
   6  90.1       81.6     10.2  0.27686    X X   X X     X X
   6  89.5       80.5     10.8  0.28468    X     X X   X X X
   7  93.8       86.5      8.4  0.23671    X X   X X   X X X
   7  92.1       82.8     10.2  0.26751    X     X X X X X X
   8  96.1       89.9      8.0  0.20510    X X   X X X X X X
   8  93.8       84.0     10.4  0.25828  X X X   X X   X X X
   9  97.0       90.4      9.1  0.20007    X X X X X X X X X
   9  96.1       87.4     10.0  0.22917  X X X   X X X X X X
  10  97.1       87.4     11.0  0.22880  X X X X X X X X X X
```

Figure 3.2: Minitab “Best Subsets” regression function output for the “Wick et al.” $\log k_b$ subgroup. Each row represents a MLR with a certain number and combination of descriptors. The “Vars” column states the number of descriptors and the “X“ marks below a stated descriptor indicates the use of that descriptor in the MLR. The chosen “best” regression model is highlighted in green. Generated with Minitab 16 (Minitab Inc.: State College, PA).

From Figure 3.2, the best MLR model for the Wick et al. dataset utilizes eight of the ten molecular descriptors from Figure 3.1. The 14 compounds comprising their data set exhibit Logk_b values from -2.27 to -0.25, with a standard deviation of about 0.62. The eight descriptor MLR was chosen based on its adjusted R^2, Mallows’ Cp, and S values, with an emphasis on selecting models using as few descriptors as possible without sacrificing predictive ability. This MLR optimized adjusted R^2 and minimized S, while including only descriptors that were statistically significant (P values ≤ 0.15, Table 3.2). In general, the internal validation statistics indicate that many of the MLR subset models (certainly including the chosen MLR) arising from the Wick et al. dataset exhibit much higher predictive ability than those from the overall (58-compound) dataset. This is evident based on comparison of Figure 3.1

and Figure 3.2. Equation 3.2 and Table 3.2 display the QSBR equation and ANOVA output, respectively, for the best MLR model for the Wick et al dataset. Again using P ≤ 0.15 as a benchmark, all eight of the selected molecular descriptor are statistically significant within the regression equation.

Equation 3.2: “Best” QSBR model based on the Wick et al. compound Logk$_b$ data set.

$$Logkb = -12.4 - 2.87\ EHOMO - 0.250\ ELUMO + 0.0442\ Gibbs - 0.0365\ HoF$$
$$- 0.146\ LogP + 0.0129\ A\ Area - 0.0370\ M\ Area - 0.0840\ MW$$

Table 3.2: Descriptor coefficients, standard errors and P values of the Wick et al. Logk$_b$ data set “Best” QSBR. All parameters generated by Minitab 16 (Minitab Inc.: State College, PA).

Predictor	**Coeff.**	**SE Coeff.**	**P**
Constant	-12.4070	2.5590	0.005
EHOMO	-2.8680	0.4345	0.001
ELUMO	-0.2501	0.1096	0.071
Gibbs	0.0442	0.0051	0.000
HoF	-0.0365	0.0042	0.000
LogP	-0.1459	0.0844	0.144
A Area	0.0129	0.0047	0.040
M Area	-0.0370	0.0090	0.009
MW	-0.0840	0.0089	0.000

Andreozzi et al. (Chemosphere 62 (2006) 1431–1436)
Best Subsets Regression: Logkb versus Total E, EHOMO, ...

```
Response is Logkb
20 cases used
                                         T
                                         o     D       A M
                                         t E E i G
                                         a H L p i   L A A
                                         l O U o b H o r r
                       Mallows             M M l b o g e e M
Vars  R-Sq  R-Sq(adj)       Cp        S  E O O e s F P a a W
   1  47.7       44.8     26.6  0.62273                    X
   1  47.4       44.4     26.9  0.62498    X
   2  83.5       81.6     -0.6  0.35960          X         X
   2  81.1       78.8      1.4  0.38561            X       X
   3  86.5       83.9     -1.0  0.33614    X     X         X
   3  85.0       82.2      0.2  0.35373    X       X       X
   4  87.8       84.6     -0.1  0.32915  X X     X X
   4  86.8       83.2      0.8  0.34342    X     X     X   X
   5  88.1       83.9      1.7  0.33646  X X   X X X
   5  88.0       83.8      1.8  0.33795  X X     X X   X
   6  88.9       83.7      3.1  0.33841  X X   X X X     X
   6  88.9       83.7      3.1  0.33846  X X   X X X   X
   7  88.9       82.4      5.0  0.35164  X X   X X X   X   X
   7  88.9       82.4      5.1  0.35197  X X   X X X X X
   8  88.9       80.8      7.0  0.36710  X X   X X X X X   X
   8  88.9       80.8      7.0  0.36715  X X X X X X   X   X
   9  89.0       79.0      9.0  0.38416  X X   X X X X X X X
   9  88.9       78.9      9.0  0.38485  X X X X X X   X X X
  10  89.0       76.7     11.0  0.40494  X X X X X X X X X X
```

Figure 3.3: Minitab "Best Subsets" regression function output for the "Andreozzi et al." logk$_b$ subgroup. Each row represents a MLR with a certain number and combination of descriptors. The "Vars" column states the number of descriptors and the "X" marks below a stated descriptor indicates the use of that descriptor in the MLR. The chosen "best" regression model is highlighted in green. Generated with Minitab 16 (Minitab Inc.: State College, PA).

From Figure 3.3, the best MLR model for the Andreozzi et al. dataset uses four of the ten molecular descriptors in Figure 3.1. The 20 compounds comprising their data set exhibit Logk$_b$ values from about -2.48 to 0.58, with a standard deviation of about 0.82. The "best" MLR model (highlighted in green) was again chosen based on its values for adjusted R^2, Mallows' Cp, and S, in particular because it maximizes adjusted R^2 and minimizes S without using nearly as many descriptors as competing models. Here again, the internal validation statistics indicate that many of the MLR subset models, including the chosen "best", from the Andreozzi et al. subgroup exhibit much higher predictive ability than those from the overall 58-compound data set. For example, the range of Logk$_b$ values in this subgroup is just as large as that of the 58-compound group, but the S value of the chosen Andreozzi MLR is roughly half of that of the "best" 58-compound MLR. Equation 3.3 and Table 3.3 display the QSBR equation and

ANOVA output, respectively, for the best MLR model for the Andreozzi et al dataset. For this model, all four molecular descriptors exhibit P ≤ 0.15, thus they are helpful for explaining variation within Logk$_b$ for the selected subgroup of OWCs.

Equation 3.3: "Best" QSBR model based on the Andreozzi et al. compound Logk$_b$ data set.

$$Log\mathrm{k}b = 3.45 + 0.000002\,Total\ E + 0.266\,EHOMO - 0.00906\,Gibbs + 0.00775\,HoF$$

Table 3.3: Descriptor coefficients, standard errors and P values of the Andreozzi et al. Logk$_b$ data set "Best" QSBR. All parameters generated by Minitab 16 (Minitab Inc.: State College, PA).

Predictor	**Coeff.**	**SE Coeff.**	**P**
Constant	3.4467	0.8860	0.001
Total E	2.38E-06	5.60E-07	0.001
EHOMO	0.2661	0.1087	0.027
Gibbs	-0.0091	0.0019	0.000
HoF	0.0078	0.0019	0.001

Urase et al. (Water Research 39 (2005) 1289–1300)
Best Subsets Regression: Logkb versus Total E, EHOMO, ...

```
Response is Logkb
12 cases used
                                          T
                                          o     D       A M
                                          t E E i G
                                          a H L p i   L A A
                                          l O U o b H o r r
                       Mallows              M M l b o g e e M
Vars  R-Sq  R-Sq(adj)       Cp        S   E O O e s F P a a W
   1  20.1       12.1     22.8  0.51579                   X
   1  19.5       11.4     23.1  0.51785                 X
   2  29.4       13.8     21.2  0.51102                   X X
   2  28.9       13.1     21.4  0.51295     X             X
   3  51.7       33.7     14.6  0.44821               X   X X
   3  44.5       23.7     17.4  0.48050               X X   X
   4  63.6       42.9     12.0  0.41598       X       X   X X
   4  59.4       36.3     13.7  0.43934               X X X X
   5  77.2       58.3      8.8  0.35555   X         X X   X X
   5  76.2       56.3      9.2  0.36363   X         X X X   X
   6  82.9       62.4      8.6  0.33747     X X X     X   X X
   6  82.0       60.3      9.0  0.34664     X X X     X X   X
   7  90.0       72.4      7.9  0.28900   X X X       X X X X
   7  89.4       70.7      8.1  0.29772   X X X     X X   X X
   8  92.3       71.8      9.0  0.29214   X X X     X X X X X
   8  92.0       70.6      9.1  0.29854   X X X X X   X   X X
   9  94.8       71.3     10.0  0.29504   X X   X X X X X X X
   9  93.4       63.9     10.5  0.33078   X X X   X X X X X X
  10  97.4       71.5     11.0  0.29372   X X X X X X X X X X
```

Figure 3.4: Minitab "Best Subsets" regression function output for the "Urase et al." logk$_b$ subgroup. Each row represents a MLR with a certain number and combination of descriptors. The "Vars" column states the number of descriptors and the "X" marks below a stated descriptor indicates the use of that descriptor in the MLR. The chosen "best" regression model is highlighted in green. Generated with Minitab 16 (Minitab Inc.: State College, PA).

From Figure 3.4, the best MLR model for the Urase et al. dataset uses seven of the ten molecular descriptors from Figure 3.1. The 12 compounds comprising their data set exhibit a range of Logk$_b$ values from -2.31 to -0.69, with a standard deviation of about 0.51. The "best" MLR (highlighted in green) was again chosen based on its adjusted R^2, Mallows' Cp, and S values, specifically because it maximizes adjusted R^2 and minimizes S. As for the Wicks and Andreozzi data sets, the internal validation statistics for the Urase data set outperform those of the overall 58-compound group. Thus, the chosen MLR for the Urase et al. data set exhibits higher predictive ability than best MLR for the overall dataset. Equation 3.4 and Table 3.4 display the QSBR equation and ANOVA output, respectively, for the best MLR model for the Urase et al dataset. Again using $P \leq 0.15$ as a benchmark, all seven of the selected molecular descriptor are statistically significant within the regression equation.

Equation 3.4: “Best” QSBR model based on the Urase et al. compound Logk$_b$ data set.

$$Logkb = 5.93 - 0.000003\,Total\,E + 0.684\,EHOMO - 0.341\,ELUMO + 1.23\,LogP$$

$$- 0.0400\,A\,Area + 0.0879\,M\,Area - 0.0328\,MW$$

Table 3.4: Descriptor coefficients, standard errors and P values of the Urase et al. Logk$_b$ data set “Best” QSBR. All parameters generated by Minitab 16 (Minitab Inc.: State College, PA).

Predictor	**Coeff.**	**SE Coeff.**	**P**
Constant	5.9330	3.3980	0.156
Tot E	-2.90E-06	1.29E-06	0.088
EHOMO	0.6843	0.2851	0.074
ELUMO	-0.3413	0.1232	0.05
LogP	1.2267	0.2600	0.009
A Area	-0.0400	0.0161	0.068
M Area	0.0879	0.0274	0.033
MW	-0.0328	0.0082	0.016

Based on the statistical results for the three selected subset models, it appears that breaking down one general QSBR model into several smaller QSBRs increases model fitness for prediction of Logk$_b$. As initially hypothesized, this effect may be due to two major factors. First, modeling Logk$_b$ data one experiment at a time removes variation in k$_b$ measurements across different studies. Because each of the three sub-models uses k$_b$ rate data from only one paper, variations in experimental procedures and rate calculations are minimized. This creates more consistent data, making it easier to fit each smaller regression model. The formation of several sub-models also appears to improve the models’ internal accuracy and simplicity by breaking down the large group of diverse molecules into more specific molecular classes. Each of the three sub-models draws from Logk$_b$ data set that contains less heterogeneous molecular structures than the overall 58-compound dataset. Additionally, it should be noted that the three sub-models identify different statistically significant molecular descriptors. This indicates that the nature of the relationship between molecular structure and biodegradation (as parameterized using Logk$_b$) is somewhat dependent on the group of molecules being evaluated. Finally, the number of predictors used in each of the three sub-models suggests that more specific subclasses of molecules can be accurately predicted using fewer descriptors. For example, the “best” QSBR from the Andreozzi et al. subgroup (the most specific subgroup) only uses four statistically significant descriptors, to achieve an adjusted R^2 of 87.8%. In contrast, the “best” QSBR from the Urase et al. subgroup (the most diverse subgroup) requires seven statistically significant descriptors to achieve a similar R^2 value of 90%. The diversity of a model’s k_b data set will be discussed in further detail in Section 3.6, which pertains to external validation.

3.5 Experimental Measurement of k_b

With one general QSBR model and three QSBR sub-models having undergone internal validation, it was desirable to assess each model’s ability to make accurate k_b predictions for compounds that had not been included in the original training data sets. In other words, it was desirable to perform external validation. The three molecules selected for measurement include: metformin, benazepril and warfarin. Appendix D displays the molecular structures of these three compounds. The k_b values for these OWCs were measured according to the procedures described in Methods Section 2.5. The principal results from these experiments comprise plots of target compound concentration over time, reactor COD over time, and k_b rate calculations for each of the three external validation compounds. These are summarized in the following paragraphs.

3.5.1 Metformin

Metformin was selected for study because of its high loading rate in wastewater, its lack of literature k_b measurement and its availability due to other research group experiments (Ottmar, 2010). Its k_b value was measured using a series of six batch reactors (one of which was positive control), as described in Section 2.5. The concentrations of metformin in each reactor are plotted as a function of time in Figure 3.5. As might be expected based on its low sorption, volatilization, and hydrolysis potentials, metformin concentrations are relatively constant in the two sorption control reactors (a1 and a2) and in the positive control reactor (PC). The experimental reactors with activated sludge, however, display a decrease in metformin concentration over time. This decrease appears to reach equilibrium after roughly 30 hours. COD concentrations in the experimental reactors, plotted in Figure 3.6, show a similar trend, whereby they initially decrease and then achieve equilibrium after about 30-40 h. In this metformin rate experiment, the same initial dose of synthetic wastewater, the main source of COD, was supplied to each reactor. No additional carbon source was provided during the 140-h. The similarity in metformin and COD concentrations (Figures 3.5 and 3.6, respectively) in the experimental reactors suggests that there is some correlation between COD and metformin concentrations. For example, metformin concentration may reach equilibrium due to the low level of COD in the experimental reactors starting around 30 h. If so, there could be some co-metabolic relationship between metformin and COD; whereby, the abundant COD is the primary substrate and the metformin is a secondary substrate. Because this hypothesis could not be directly evaluated from the metformin data displayed in Figure 3.5 or 3.6, a slightly different experimental protocol was used in subsequent benazepril experiments (Section 3.5.2).

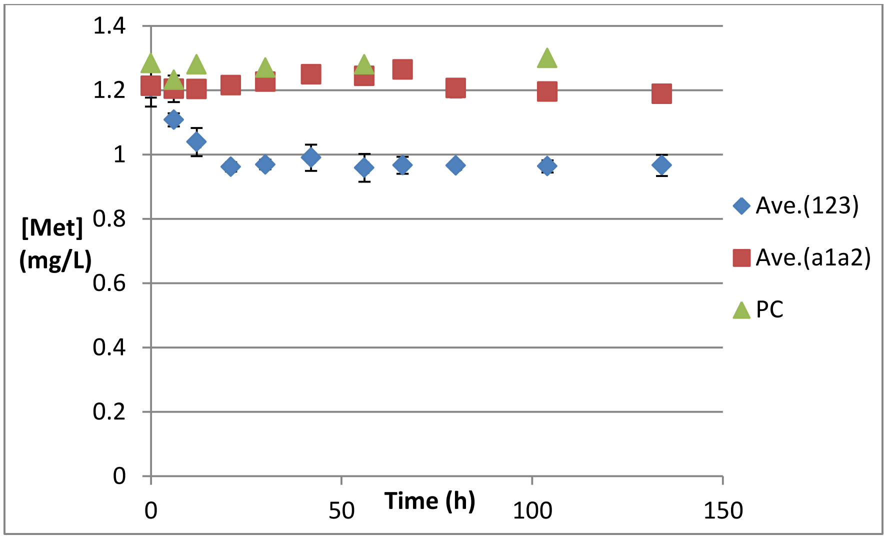


Figure 3.5: Averaged metformin concentrations in experimental (regular) reactors, sorption controls, and one positive control over time. Average values of the three experimental reactors (123) are displayed in blue, the two sorption controls (a1a2) are in red, and the positive control (PC) is in green. Error bars display plus and minus one standard deviation of each set of reactor concentrations at one time.

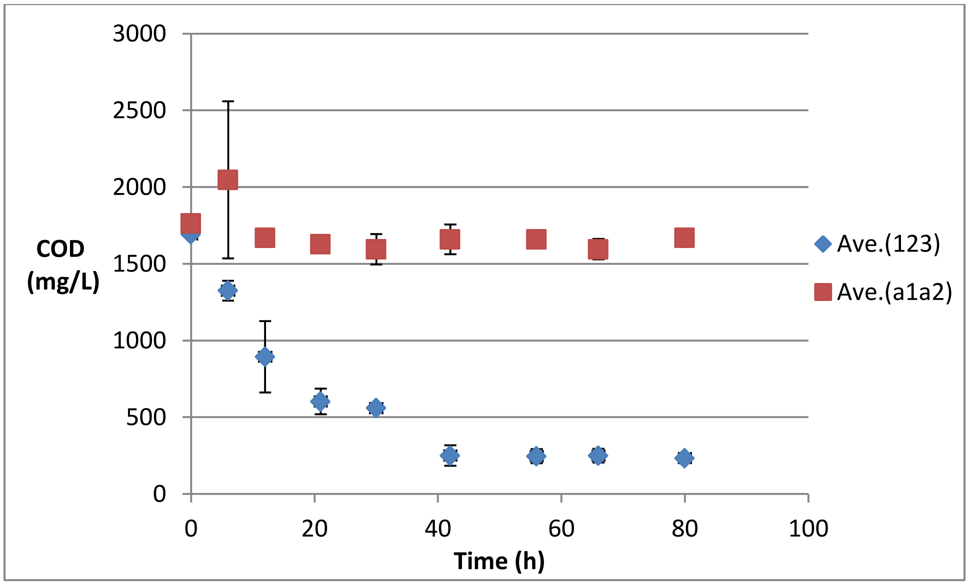


Figure 3.6: Average COD (chemical oxygen demand) concentrations over time in experimental (regular) reactors, sorption controls, and a positive control for a biodegradation kinetics experiment in which all reactors are only initially dosed with synthetic wastewater (COD) and target OWC. Again, the three experimental reactors (123) are displayed in blue and the two sorption controls (a1a2) are in red. Error bars display plus and minus one standard deviation of each set of reactor concentrations at one time.

Regression calculations for determination of the metformin rate constant (k_b) are displayed in Figure 3.7. These calculations are documented in Methods Section 2.5. The rate constant for a first order model was approximated to be 0.0105 h^{-1}, based on the negative slope of the best-fit regression line shown in Figure 3.7. Because the suspended sludge concentration was constant over the experiment, at about 1 g/L, this first order rate has the same magnitude as a pseudo-first order rate constant; however, the units would be k_b = 0.0105 L/g-h. Only the first five samples were used in this rate calculation as the experimental reactors reached equilibrium concentrations after the fifth sample.

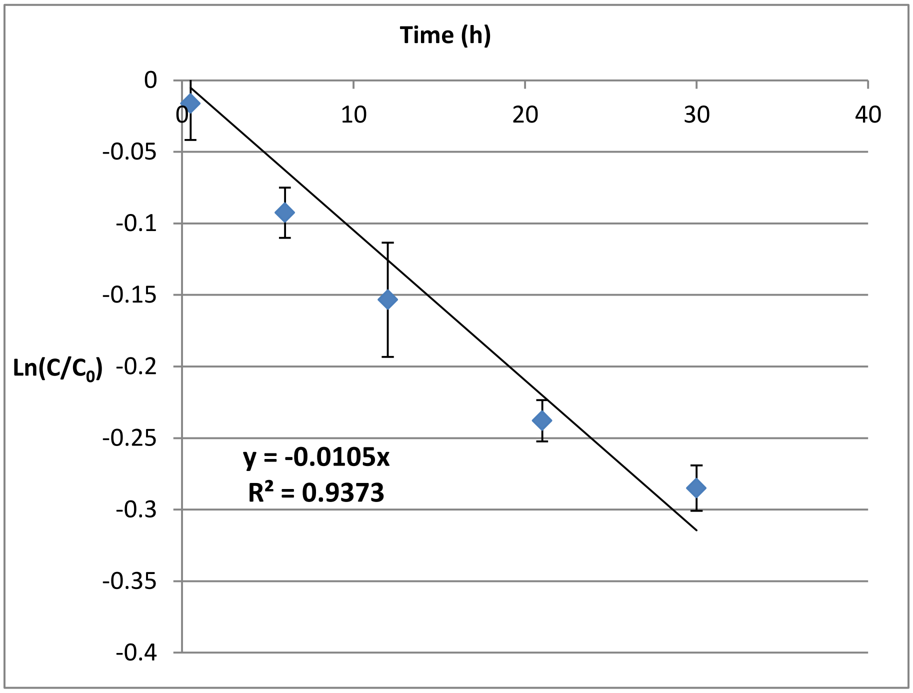


Figure 3.7: Calculation of biodegradation rate constant, k_b, for metformin based on best fit regression analysis. Only the pre-equilibrium points, corresponding to samples collected within the first 30 h, were used in this rate calculation. Error bars display plus and minus one standard deviation of the underlying data from each point.

3.5.2 Benazepril

Benazepril was selected for inclusion in this study because of its high prescription rank in wastewater, its lack of literature k_b measurement and its availability. Multiple experiments were performed to determine k_b for this drug. In particular, biodegradation experiments were performed with and without daily re-dosing of COD (as referenced in Section 3.5.1). Both types of experiments utilized six reactors: three experimental (regular) reactors containing benazepril-spiked synthetic wastewater and activated sludge; two sorption controls containing benazepril-spiked synthetic wastewater and sludge that had been thermally and chemically deactivated; and one positive control, which was just benazepril in deionized water.

Figure 3.8 shows average concentrations of benazepril in the six reactors, and Figure 3.9 shows average COD concentrations in the experimental reactors and the duplicate sorption controls. As was the case for the metformin experiments, the sorption and positive control reactors show little change in concentration over time. In contrast to the metformin experiment, there was slightly less correlation between low COD concentration and the onset of equilibrium in benazepril concentration. COD concentrations in the experimental reactors reach equilibrium after roughly 50 h, whereas the benazepril concentration does not do so until after nearly 100 h. The equilibrium concentration of benazepril in the experimental reactors appears to be 0.5 or 0.6 mg/L.

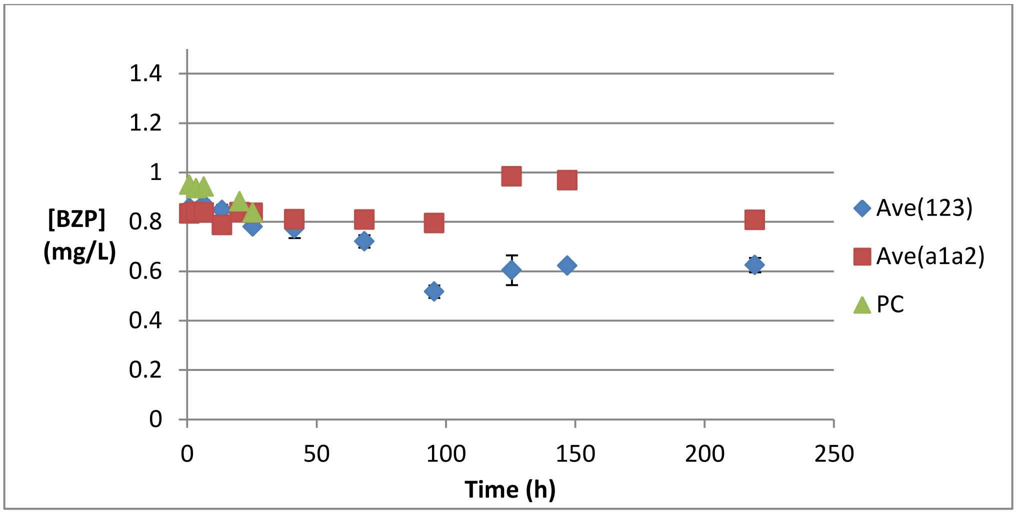


Figure 3.8: Average benazepril concentrations over time in experimental (regular) reactors, sorption controls, and a positive control for a biodegradation kinetics experiment in which all reactors are only dosing with synthetic wastewater (COD) and target OWC one time (at the beginning). Average values of the three experimental reactors (123) are displayed in blue, the two sorption controls (a1a2) are in red, and the positive control (PC) is in green.

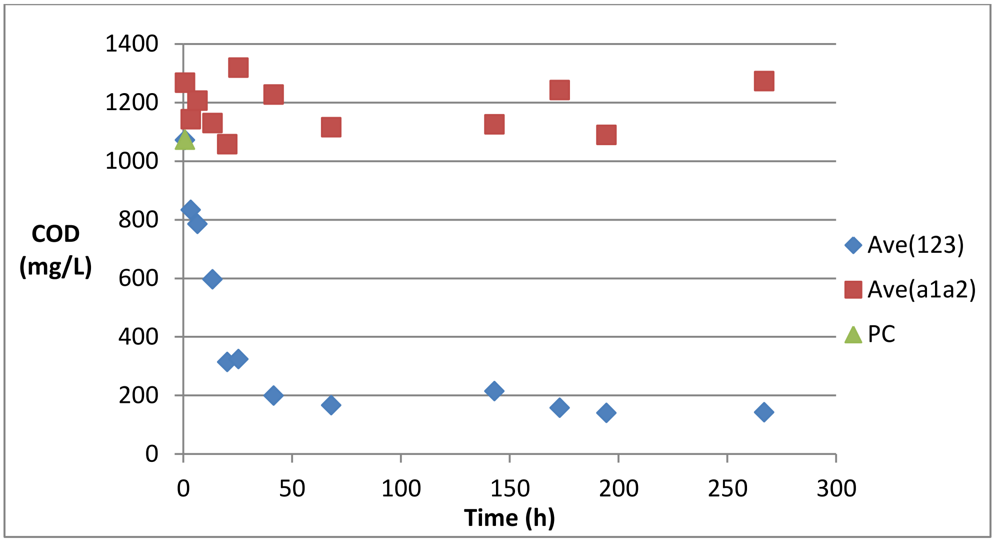


Figure 3.9: Average COD (chemical oxygen demand) concentrations over time in experimental (regular) reactors, sorption controls, and a positive control for a biodegradation kinetics experiment in which all reactors are only initially dosed with synthetic wastewater (COD) and target OWC. The COD concentration of the PC was measured only at the beginning of the experiment; thereafter, it was assumed to remain constant.

Regression calculations for determination of the benazepril rate constant (k_b) in an experiment without re-dosing of COD are displayed in Figure 3.10. The best-fit value was approximately 0.0033 h^{-1}, based on the negative slope of the regression line. Again, the suspended sludge concentration was constant over the experiment at about 1 g/L, so the magnitude of the first order rate constant is the same as the magnitude of the pseudo-first order rate constant, which is k_b = 0.0033 L/g-h.

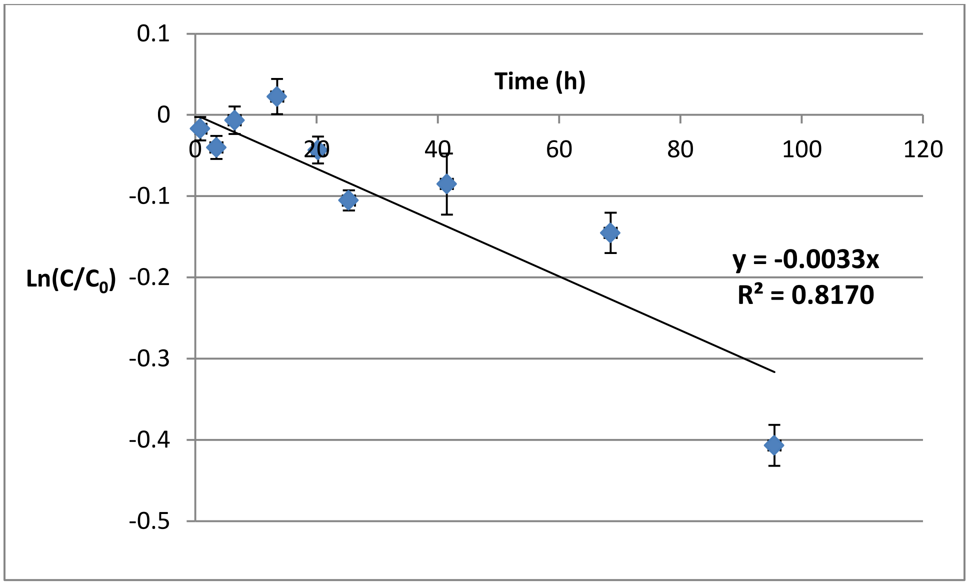


Figure 3.10: Calculation of the benazepril biodegradation rate constant (k_b) for experiments without re-dosing on COD. Only the pre-equilibrium points (within the first 100 h) were used in this rate calculation.

A second benazepril experiment was completed using all of the same experimental conditions as the first, except with daily re-dosing back to the initial COD concentration for the three experimental reactors (123). Figure 3.11 shows average benazepril concentrations over time, and Figure 3.12 shows corresponding average COD concentrations. These COD measurements correspond to samples collected at the end of each ~24 h re-dosing cycle. As was the case for the metformin experiment and the benazepril experiment without re-dosing, benazepril concentrations in the sorption controls and the positive control reactors remained relatively steady throughout the experiment. Unlike the metformin experiment and the benazepril experiment without re-dosing, the three experimental reactors showed relatively steady decreases in benazepril concentration throughout the entire experiment, without ever achieving equilibrium. After nearly 200 h, the average benazepril concentration in the experimental reactors is had 0.25 mg/L, which is roughly half of what was achieved in the benazepril experiment without re-dosing (0.6 mg/L after 200 h). Based on this comparison, the daily re-dosing of COD appears to extend and improve benazepril removal over time.

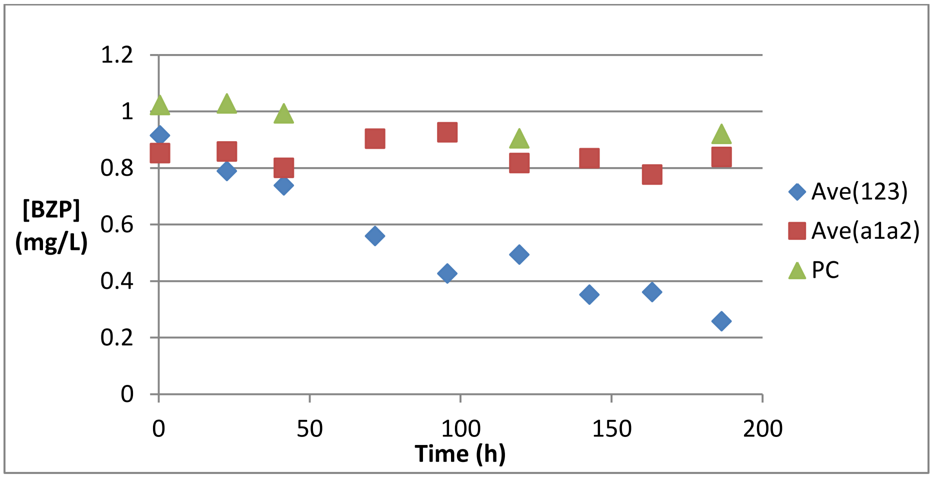


Figure 3.11: Average benazepril concentrations in experimental (regular) reactors (123), sorption controls (a1a2), and a positive control (PC) over time during the re-dosing experiment.

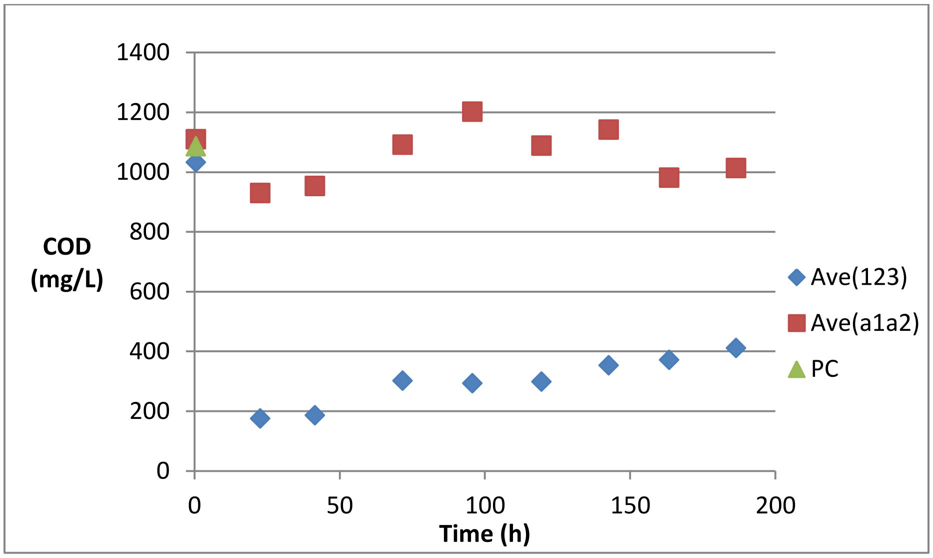


Figure 3.12: Average COD concentrations during a biodegradation experiment involving daily COD re-dosing. COD measurements of the regular reactors (blue) were always measured near the end of the 24- h re-dosing cyle. Regular reactors were re-dosed to approximately their initial lCOD evel with synthetic wastewater after each blue measurement was taken. The initial COD of the PC is displayed and was assumed to remail constant.

Regression calculations for determination of the benazepril rate constant (k_b) for the experiment with COD re-dosing are displayed in Figure 3.13. The negative slope of the best-fit regression line is about 0.007 h^{-1} if a first order rate model is assumed. Unlike the metformin experiment and the benazepril experiment without re-dosing, the suspended sludge (X_{ss}) concentration increased from 1 to 3.4 g/L during the experiment; therefore, it is not appropriate to assume that the first order rate constant and the pseudo-first order rate constant have the same magnitude. Ideally, a functional expression would be used to model the change in suspended sludge concentration over time. This function could be substituted into Equation 2.3 (Methods) and integrated to solve for a pseudo-first order k_b value. For simplicity, this approach was not used in this study. Instead, the negative slope of the regression line was divided by the average suspended sludge concentration, 2.2 g/L. This yields a pseudo-first order rate constant of approximately 0.0032 L/g-h. Intriguingly, this value is quite close to the pseudo-first order rate value computed from the benazepril experiment without COD re-dosing, 0.0033 L/g-h. This could suggest that sludge concentration affects the rate of biodegradation in the experiments (as expected by a pseudo-first order rate). Additionally, comparison of the non-re-dosing and re-dosing experiments suggests that greater per cent reductions can be achieved if COD is continuously supplied (re-dosed) to the microbial biomass.

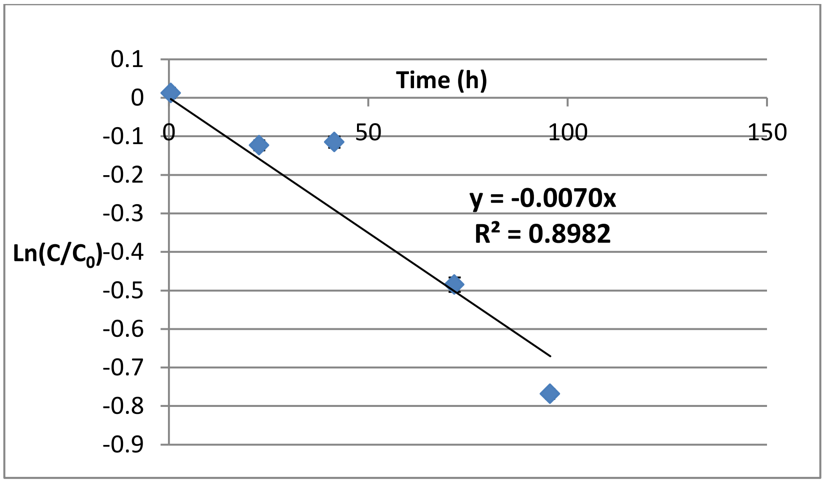


Figure 3.13: Calculation of the benazepril biodegradation rate constant (k_b) for experiments with re-dosing on COD. Only data points within the first 100 h were used in this regression, for consistency with the benazepril experiment without COD re-dosing.

3.5.3 Warfarin

Warfarin was selected for inclusion in this study because of its high prescription rank, its lack of literature k_b measurement and its availability in our lab (IMS Institute, 2011). The biodegradation experiment for determination of warfarin k_b was performed using a series of seven batch reactors (three experimental reactors, two sorption controls, and two positive controls) and the daily COD re-dosing scheme described for benazepril in Section 3.5.2. Figure 3.14 depicts warfarin concentration over time in this experiment, though there is very little change in any of these concentrations over nearly 170 h. Because the biodegradation contribution to apparent removal is isolated via subtraction of the sorption control and positive control concentrations (Methods, 2.1), it is difficult to analyze the warfarin data in Figure 3.14 using this method. Thus, the k_b of warfarin was reported as 0 L/g-h. This presents practical difficulties for external validation, since the log transformation was applied to all k_b values prior to use in the predictive model and the log of zero is negative infinity. This issue will be addressed in the following section.

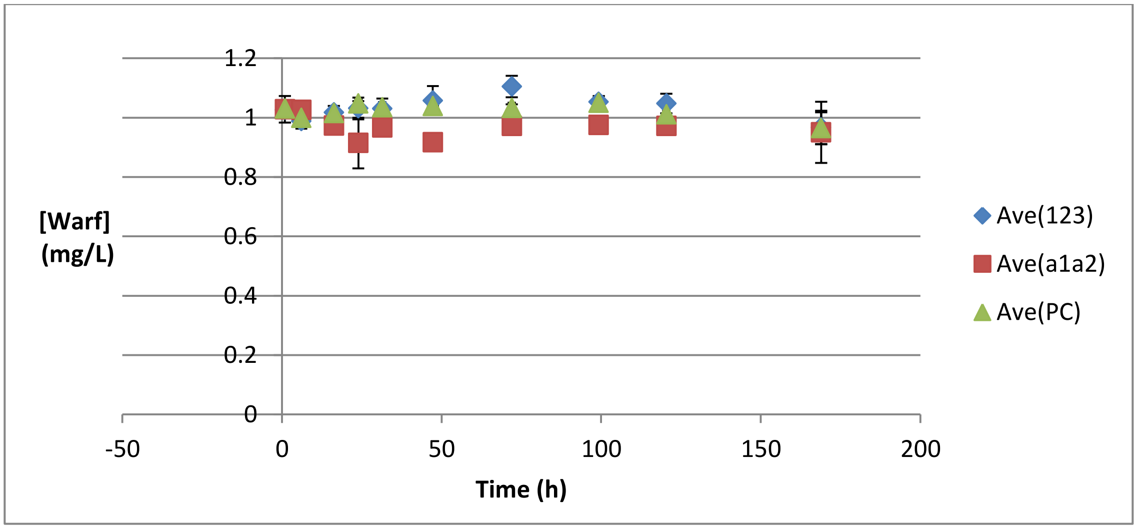


Figure 3.14: Averaged warfarin concentrations in experimental (regular – 123), sorption controls (a1a2), and positive control reactors over time in a daily COD re-dosing experiment. Error bars display plus and minus one standard deviation of each set of reactor concentrations at one time.

3.6 Applying the QSBR Models to the External Validation k_b Values

Molecular descriptors of the three external validation compounds (metformin, benazepril and warfarin) are displayed in Appendix B. These values were used in conjunction with the newly measured k_b values, to evaluate the four QSBR models described in Section 3.4. For benazepril, only the k_b value from the biodegradation experiment without re-dosing was used.

Table 3.5 displays measured and predicted Logk_b values for the three validation compounds. Predictions correspond to all four different QSBR models from Section 3.4. Colored highlighting indicates which of the four predictions is closest to the measured value. The overall 58-compound model gives the closest prediction for metformin and benazepril; however, warfarin is best predicted by the Wick et al. subset model. The predictions of each model vary greatly. For example, metformin predictions range from -6.1 to 14.3, though both of these values are way outside of the range of Logk_b values actually included in the 58-compound dataset (i.e., -2.48 to 0.58). Despite this, the closest predictions for each compound are actually quite close to their corresponding measured Logk_b values.

It is interesting that two of the most accurate predictions come from the overall 58-compound QSBR. Based on internal validation statistics, this model exhibits the worst fit for the Logk_b data, but it generates Logk_b predictions for metformin and benazepril that are remarkably close to the measured values. The Andreozzi QSBR generates a fairly close prediction for metformin as well. The Wick QSBR yields the lowest Logk_b prediction for warfarin, which makes it most accurate given that the warfarin k_b was approximated as roughly 0 L/g-h.

Table 3.5: Logk_b predictions for four validation compounds, as made using four QSBR models from Section 3.4. Corresponding measured values are also presented. Colored highlighting indicates the best model prediction for each compound.

	Model Logk_b Predictions					
Compound	58 General	Wick et al.	Andreozzi et al.	Urase et al.	Measured k_b (L/gh)	Measured Logk_b
Metformin	-2.03	14.61	-2.30	-6.09	0.0105	-1.98
Benezepril	-2.29	-9.26	-4.97	-4.92	0.0033	-2.48
Warfarin	-1.20	-9.93	-2.17	-2.11	0.0000	Log(0)

As displayed by Table 3.5, QSBR model predictions can be relatively accurate or extremely inaccurate, depending on which model is chosen. This raises the question: how does one know, *a priori*, which model should be used when a prediction is needed? This question can be partially answered by looking at each of the four models and quantifying the range and diversity of descriptors for their constituent compounds. Analyzing how well the molecular descriptors of an unmeasured compound match the range and diversity of the data within a given QBSR's training data set can give insight into the ability of that model to make an accurate prediction for the selected chemical of interest. In other words, QSBR models that attempt to predict Logk_b for molecules that are similar, based on descriptor values, to the underlying molecules that the model was created from should generally yield better predictions than QSBRs that were created based on molecules that are different from the chemical of interest. This hypothesis is tested in Figure 3.15, wherein it is demonstrated that the general 58 and Andreozzi et al. QSBRs are "best" (have the most similar underlying descriptors) for prediction of metformin and the general 58 QSBR is best for prediction of benazepril. However, the model that matched best with warfarin (general 58, based on Figure 3.15) is inconsistent with the model that made the closest prediction for warfarin (Wick et al.).

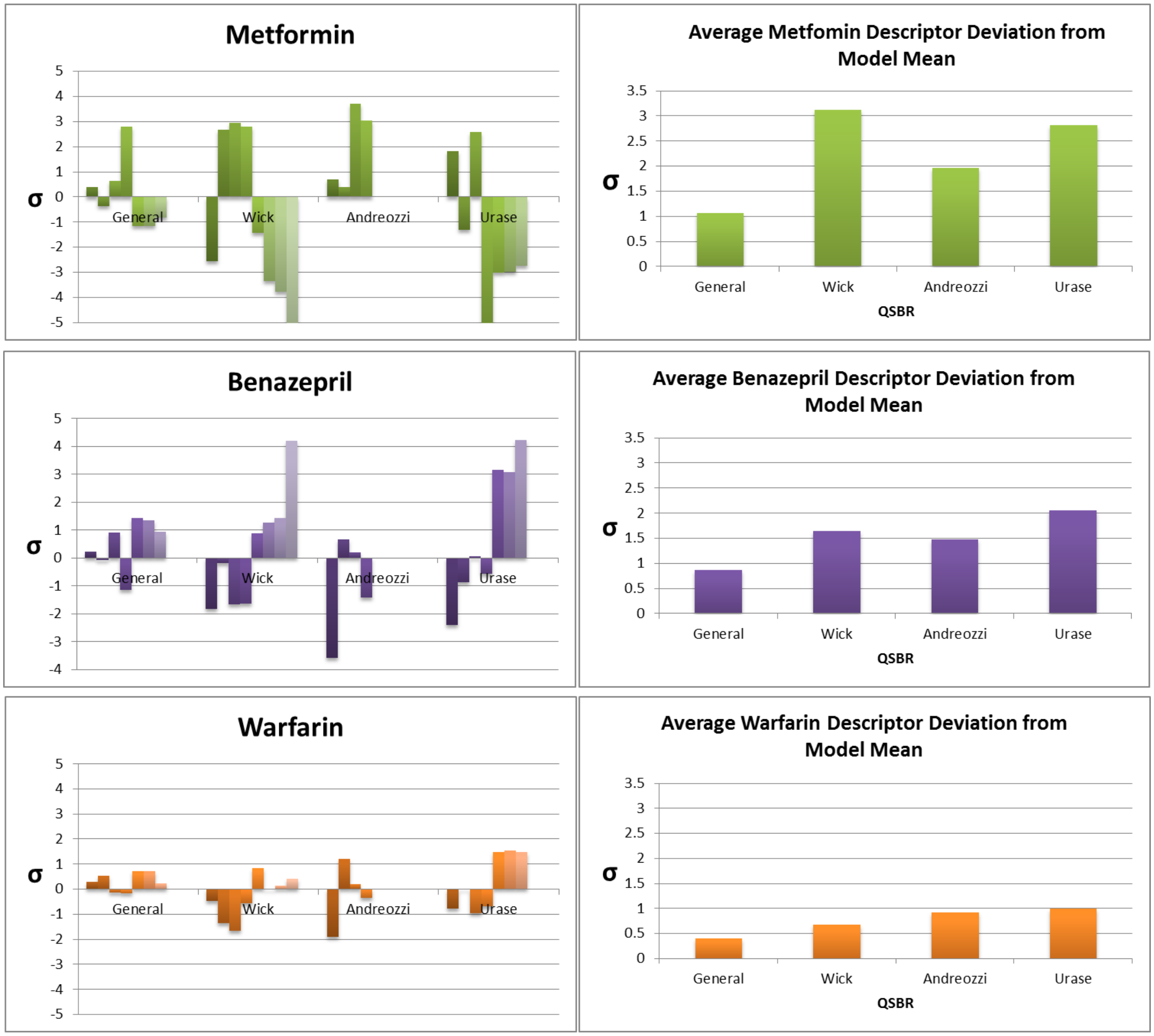


Figure 3.15: Validation compounds (metformin, benazepril and warfarin) are matched to the training data sets which were used to construct four different QSBR models. "General" corresponds to the general-58 compound QSBR, Wick symbolizes the Wick et al. QSBR, Andreozzi corresponds to the Andreozzi et al. QSBR, and Urase stands for the Urase et al. QSBR. The three graphs on the left show the distance in standard units that each descriptor of a listed validation compound falls from that descriptor's average within each model's underlying data ((validation compound descriptor value – average descriptor value within a model)/standard deviation of descriptor in a model). The three graphs on the right show the addition of the absolute values of the distances for each of the four corresponding models in the graphs to the left divided by the number of descriptor used in each model.

The graphs in Figure 3.15 show which of the four models have the most similar descriptors to each of the three validation compounds. The model with the lowest average deviation from the validation compound value is said to be the "most similar" model, and thus might be expected to be the most appropriate model to use when making a prediction. As previously stated, the general 58 appears to be the most similar to all three of the validation compounds. This is expected, as the general 58 model has the largest standard deviation and contains all of the other three subset models within it. The general 58 model did make the most accurate prediction for metformin and benazepril, but not for warfarin. Multiple assumptions should be addressed here. First, this type of average similarity analysis assumes that all of the descriptors in each model are of equal importance, but, as displayed in Tables 3.1 through 3.4 different descriptors have different levels of importance in each model. Also, just because a model is the most similar to a prospective compound in need of prediction does not mean that it will make the best prediction. The similarity analysis is really only an assessment of whether or not it is reasonable to use a model to make a prediction, or in other words whether a compound falls within the application domain of a model. This application domain may be an agreed upon level of model similarity that a prospective compound must fall within in order to be logically predicted. And, even if a compound does lie within the application domain of a model, the model's internal validation parameters are still important indicators of its predictive ability.

If we say that a validation compound must fall within about one standard deviation of the mean of a model's descriptors to be within its application domain, then metformin and benazepril only fall within the application domain of the general 58 model (metformin is actually just over 1 standard deviation away from the general 58's mean, but we'll say that's close enough, Figure3.15). So, based on these application domain criteria, it only makes sense to predict the logk$_b$s of these two compounds with the general 58 model. However, using the same criteria, warfarin falls within the application domain of all four models, each creating a different logk$_b$ prediction (from -1.2 to -9.9). While the large range of these predictions is slightly unsettling, it is noteworthy that the best prediction was made by the model (Wick et al.) with the best internal validation parameters (R^2 = 96). In short, a model's underlying descriptor similarity to a predicted compound is important in determining whether the model is appropriate for making a prediction, but other statistics (like internal validation parameters) should also be considered when assessing the accuracy of a prediction.

Results indicate that an appropriate QSBR must not only possess favorable internal validation statistics, but it must also possess similarity of training data set molecular descriptors to those of the compound that will be predicted. For the four QSBRs created in this study, the three subset models achieved much higher internal validation coefficients than the overall 58-compound QSBR, but they can only logically make predictions for compounds that possess similar descriptor values to their less diverse, respective molecular data set. On the other hand, the overall 58-compound QSBR achieved low internal validation coefficients, but, because it draws from a more diverse underlying molecular data set, it can make predictions for a wider array of compounds. This concept will be further addressed in the next chapter, as part of a larger comparison between traditional, regression-based QBSR and a novel prediction technique referred to as “quantitative molecular similarly assessment” (QMSA).

Chapter 4 –Implications and Conclusions

4.1 Implications

4.1.1 First Order vs. Pseudo-First Order Rate Models

In the OWC WWTP biodegradation literature, both first order and pseudo-first order rate models are used by various authors. As discussed in Section 3.1, there are several, common inconsistencies among the types of experiments and calculations used to determine k_b. The comparative results of the benazepril k_b experiments with and without COD re-dosing are highly suggestive of a pseudo-first order differential equation describing compound biodegradation in WWTP sludge (Equation 2.2). Re-dosing COD in the second benazepril experiment steadily increased the suspended sludge concentration of the three regular bioreactors. The observed biodegradation rate of benazepril (not considering suspended sludge concentration) in the re-dosing experiment was around twice that of the rate in the single-dosed experiment. Because the only difference between these two experiments was the re-dosing of COD, it seems very likely that suspended sludge concentration proportionally affects the rate of compound biodegradation. Thus, a pseudo-first order biodegradation differential equation is more appropriate than a first order rate constant.

4.1.2 Descriptors of Importance

The results of the four QSBR models created in this study, particularly their statistically-significant P-values, indicate that different molecular descriptors are important for predicting biodegradation rates for different classes of molecules. Some descriptors appear in the four QSBRs more frequently than others; e.g., EHOMO appears in all four models; and Total E, HoF, A Area, M Area and MW appear in three of the four. The importance of certain descriptors in each model can give insight into biodegradation reaction mechanisms taking place during activated sludge treatment. For instance, in two of the four QSBRs (general 58 and Andreozzi et al.), the descriptor LogP is not used. Furthermore, in these two models steric and electronic descriptors like EHOMO, A-Area, M -Area and MW are used.

The simultaneous absence of LogP and presence of steric and electronic descriptors in these models suggests that the rate-determining step during biodegradation is likely enzyme binding or transformation rate and not microbial uptake/transport (Parsons and Govers, 1990). LogP is statistically significant, based on P value ≤.15, in the Urase et al. and Wick et al. QSBRs; in addition to electronic and steric descriptors EHOMO ELUMO, A-Area, M- Area, and MW (all of which have statistically significant P values). Because both types of descriptors are statistically significant, it is more difficult to draw conclusions about what step of the biodegradation process is rate-limiting. It could be microbial uptake/transport, enzyme binding, or enzymatic transformation rate. This information suggests differences in the rate determining step of WWTP biodegradation for different groups of compounds.

4.1.3 QSBR and QMSA

Internal validation results for the four QSBR models created in this study affirm the notion that general QSBRs, as created from large heterogeneous sets of molecules, are less able to fit molecular descriptors to Logk$_b$ data (by multiple linear regression) than specific QSBRs created from smaller, more homogeneous sets of molecules. However, external validation results demonstrate the importance of matching training set characteristics to target compound characteristics. The overall 58-compound QSBR draws from a diverse Logk$_b$ data set, but it cannot be expected to consistently deliver accurate predictions, based on internal validation statistics. The three smaller, subset QSBRs exhibit better internal validation performance, but they possess relatively smaller descriptor diversity, making them only able to deliver accurate predictions within their narrow application domains.

It seems that the most accurate way to predict k$_b$ for a wide variety of OWCs with QSBRs would be to create many subset QSBRs to cover as many molecular classes as possible. However, the feasibility of this practice is questionable. Creating such models would be laborious due to the necessity of obtaining k$_b$ data sets (that don't currently exist) for each specific molecular class of interest. Additionally, if the many QSBRs were created, a series of calculations would need to be performed to determine the best subset model to use (based on the descriptors of the compound to be predicted) each time a prediction was desired. It would be desirable to use just one model instead of many.

These considerations for QSBR models (or QSARs in general) make it such that an alternative predictive method could be desirable. One such alternative may be quantitative molecular similarity analysis (QMSA). This class of models is used to predict an unknown property for a compound, based on some linear combination of the previously measured molecular descriptors of several "similar" compounds; whereby, similarity is quantitatively parameterized in n dimensions corresponding to *n* molecular descriptors (Li and Colosi, 2011) The key difference between QSAR models and QMSA models is that QSAR models make predictions based on the descriptors and properties of the entire dataset, while QMSAs define a "radius of interest" around a particular compound of interest. Then, only previously measured compounds within that radius are used to make a prediction of the compound of interest. In this way, QMSA is similar to having several different specific QSBR models nested within one model. Additionally, QMSA can be used to optimize efficiency in measuring the k_b (or other properties), by identifying compounds whose k_b data would add the most diversity to the underlying k_b data set. Taken together, these advantages and the somewhat lackluster QSBR results from this study could suggest that QMSA may be a favorable alternative method for predicting biodegradation rates of OWCs.

4.2 Conclusions

Contaminant fate and transport models are a highly desirable alternative to direct measurement of environmental behavior for the increasingly large number of organic wastewater contaminants. In particular, it is desirable to estimate what fraction of OWC contaminant loading is removed by conventional wastewater treatment plant processes, so that subsequent loading into the environment can be assessed. The QSBR models created in this study are capable of predicting biodegradation rate constants for specific compounds within the crucial activated sludge stage of wastewater treatment. Tying together significant descriptors used in previous QSBR studies, these four models display the capabilities of both general and specific QSBRs. Statistics indicate that the overall QSBR model (comprising all compounds in the training set) achieves less predictive ability (R^2 =0. 49) than the three smaller QSBR models (R^2 = 0.97, 0.88, and 0.90) that were created from smaller subsets of the same dataset. The favorable statistical performance of the smaller subset models is expected to be a result of both the isolation of k_b measurements by specific papers (removing experimental differences across papers in the general model) and the focus on more homogeneous molecular groups. However, external validation demonstrates highly accurate predictions from the overall QSBR for metformin logk_b (measured = -1.98 versus predicted = -2.03) and benazepril logk_b (measured = -2.48 versus predicted = -2.29). Warfarin was best estimated using the Wick et al. smaller subset QSBR. These three new experimentally derived k_b values add to the body of literature, while also serving to test each model. Findings suggest that while a QSBR may have high internal validation, it can only be expected to make predictions within its application domain, the extent of which is based on the diversity of its underlying descriptors. While the four QSBRs display considerable potential in making accurate k_b predictions, their greater contribution may be serving as stepping stones toward the creation of future models.

References

R. Andreozzi, R. Cesaro, R. Marotta, F. Pirozzi.

“Evaluation of biodegradation kinetic constants for aromatic compounds by means of aerobic batch experiments *Chemosphere*”, Volume 62, Issue 9, (2006), 1431–1436.

J.D. Berset , T. Kupper , R. Etter, J. Tarradellas.

“Considerations about the enantioselective transformation of polycyclic musks in wastewater, treated wastewater and sewage sludge and analysis of their fate in a sequencing batch reactor plant” *Chemosphere* 57 (2004), 987–996.

M. Cargouët, D. Perdiz, A. Mouatassim-Souali, S. Tamisier-Karolak, Y. Levi.

“Assessment of river contamination by estrogenic compounds in Paris area (France)” *Science of The Total Environment*, Volume 324(2004), Issues 1–3, Pages 55–66.

T. Colborn, F. S. vom Saal, A. M. Soto.

“Developmental effects of endocrine-disrupting chemicals in wildlife and humans.” *Environ Health Perspect*. 101 (1993), 5, 378–384.

C.G. Daughton, T.A. Ternes.

“Pharmaceuticals and personal care products in the environment: Agents of subtle change?” *Environ Health Perspectives*, 107 (1999), pp. 907–938.

USEPA.

"Pharmaceuticals and Personal Care Products (PPCPs) in Water." *Home*. N.p., n.d. Web. 01 Dec. 2012. <http://water.epa.gov/scitech/swguidance/ppcp/index.cfm>.

M. J. Focazio, D. W. Kolpin, K. K. Barnes, E. T. Furlong, M. T. Meyer, S. D. Zaugg, L. B. Barber, M. E. Thurman.

“A national reconnaissance for pharmaceuticals and other organic wastewater contaminants in the United States — II) Untreated drinking water sources” *Science of The Total Environment*, Volume 402(2008), Issues 2–3, Pages 201–216.

J. S. Gray.

“Biomagnification in marine systems: the perspective of an ecologist” *Marine Pollution Bulletin*, Volume 45 (2002), Issues 1–12, Pages 46–52.

T. Heberer.

“Occurrence, fate, and removal of pharmaceutical residues in the aquatic environment: a review of recent research data” *Toxicology Letters*, Volume 131(2002), Issues 1–2, Pages 5–17.

R. Hao, J. Li, Y. Zhou, S. Cheng, Y. Zhang.

“Structure–biodegradability relationship of nonylphenol isomers during biological wastewater treatment process” *Chemosphere*, Volume 75 (2009), Issue 8, Pages 987–994.

IMS Institute for Healthcare Informatics. “The Use of Medicines in the United States: Review of 2011.” IMS Institute for Healthcare Informatics Web Site. http://www.imshealth.com/ims/Global/Content/Insights/IMS%20Institute%20for%20Healthcare%20Informatics/IHII_Medicines_in_U.S_Report_2011.pdf..

A. C. Johnson and J. P. Sumpter.
“Removal of Endocrine-Disrupting Chemicals in Activated Sludge Treatment Works” *Environmental Science and Technology*, Vol. 35(2001), no. 24.

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

D.W. Kolpin, E.T. Furlong, M.T. Meyer, E.M. Thurman, S.D. Zaugg, L.B. Barber, H.T. Buxton.
“Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999–2000 — a national reconnaissance” *Environ Sci Technol*, 36 (2002), pp. 1202–121.

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

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

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

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

N. Lozano, C. P. Rice, J. Pagano, L. Zintek, L. B. Barber, E. W. Murphy, T. Nettesheim, T. Minarik, H. L. Schoenfuss.
“Concentration of organic contaminants in fish and their biological effects in a wastewater-dominated urban stream” *Science of The Total Environment*, Volume 420 (2012), Pages 191–201.

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

M. Maurer, B.I. Escher, P. Richle, C. Schaffner, A.C. Alder.

“Elimination of β-blockers in sewage treatment plants” *Water Research*, Volume 41, Issue 7, (2007), Pages 1614–1622.

M. S. McLachlan, G. Czub, M. MacLeod, and J. A. Arnot.
“Bioaccumulation of Organic Contaminants in Humans: A Multimedia Perspective and the Importance of Biotransformation” *Environmental Science & Technology* 45 (2011), 1, 197-202.

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

R. W. Okey, H.D. Stensel.
“A QSAR-based biodegradability model—A QSBR” *Water Research*, Volume 30 (1996), Issue 9, , Pages 2206–2214.

K. J. Ottmar, L. M. Colosi, J. A. Smith.
“Development and Application of a Model to Estimate Wastewater Treatment Plant Prescription Pharmaceutical Influent Loadings and Concentrations”
Bulletin of Environmental Contamination and Toxicology (2010), 84:507–512.

E. Papa, J. Fick, R. Lindberg, M. Johansson, P. Gramatica, and P. L. Andersson.
“Multivariate Chemical Mapping of Antibiotics and Identification of Structurally Representative Substances” *Environmental Science & Technology* 41 (2007), 5, 1653-1661.

J.R. Parsons, H.A.J. Govers.
“Quantitative structure-activity relationships for biodegradation”
Ecotoxicology and Environmental Safety, Volume 19 (1990), Issue 2, Pages 212–227.

Ternes TA, Joss A, Siegrist H.
“Scrutinizing pharmaceuticals and personal care products in wastewater treatment.” *Environ Sci Technol.*(2004),38(20):392A-399A. Review.

T. A. Ternes.
“Occurrence of drugs in German sewage treatment plants and rivers” *Water Research*, Volume 32 (1998), Issue 11, Pages 3245–3260.

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

USGS.
"Research Projects - Emerging Contaminants in the Environment." *Research Projects - Emerging Contaminants in the Environment*. N.p., n.d. Web. 01 Dec. 2012.
<http://toxics.usgs.gov/regional/emc/index.html>.

A. Wick, G. Fink, A. Joss, H. Siegrist, T. A. Ternes.

“Fate of beta blockers and psycho-active drugs in conventional wastewater treatment” *Water Research*, Volume 43, Issue 4, (2009), Pages 1060–1074H.

Yang, Z. Jiang, S. Shi.
“Aromatic compounds biodegradation under anaerobic conditions and their QSBR models” *Science of The Total Environment*, Volume 358 (2006), Issues 1–3, Pages 265–276.

H. Yang, Z. Jiang, S. Shi, “Biodegradability of nitrogenous compounds under anaerobic conditions and its estimation” *Ecotoxicology and Environmental Safety*, Volume 63 (2006), Issue 2, Pages 299–305.

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

"Secondary Treatment Standards." *EPA* -. N.p., n.d. Web. 01 Dec. 2012. <http://cfpub.epa.gov/npdes/techbasedpermitting/sectreat.cfm>.

"EU-Project Poseidon." *EU-Project Poseidon*. N.p., n.d. Web. 01 Dec. 2012. <http://poseidon.bafg.de/servlet/is/2884/>.

"Wastewater Treatment and Water Reclamation." *LACSD Website* -. N.p., n.d. Web. 01 Dec. 2012. <http://www.lacsd.org/wastewater/wwfacilities/moresanj.asp>.

"3. Wastewater Treatment." *3. Wastewater Treatment*. N.p., n.d. Web. 01 Dec. 2012. <http://www.fao.org/docrep/t0551e/t0551e05.htm>.

"A Visit to a Wastewater-treatment Plant:Primary Treatment of Wastewater." *Wastewater-treatment Plant Visit, USGS Water-Science School*. N.p., n.d. Web. 01 Dec. 2012. <http://ga.water.usgs.gov/edu/wwvisit.html>.

Appendix A: All 82 collected literature k_b values grouped by paper. Green highlighting indicates compounds that have multiple k_b measurements in the data set. Of these 82 values, 65 are unique compounds. All displayed biodegradation rates are pseudo-first order.

Compound	Paper	Kb (L/(gh)) (pseudo first order)
Caffeine	Majewsky et al. (2011)	1.76500
Carbamazepine	Majewsky et al. (2011)	0.00850
Diclofenac	Majewsky et al. (2011)	0.02700
Paracetamol	Majewsky et al. (2011)	1.03450
Sulfamethoxazole	Majewsky et al. (2011)	0.27600
Cefalexin	Li et al. (2010)	0.11765
Sulfadiazine	Li et al. (2010)	0.00724
Sulfamethoxazole	Li et al. (2010)	0.00498
Atenolol	Wick et al. (2009)	0.06250
Betaxolol	Wick et al. (2009)	0.25000
Bisoprolol	Wick et al. (2009)	0.02938
Celiprolol	Wick et al. (2009)	0.00875
Codeine	Wick et al. (2009)	0.19792
DHH	Wick et al. (2009)	0.00667
Dihydrocodeine	Wick et al. (2009)	0.07500
Doxepin	Wick et al. (2009)	0.02021
Methadone	Wick et al. (2009)	0.00875
Metoprolol	Wick et al. (2009)	0.01563
Morphine	Wick et al. (2009)	0.56250
Nordiazepam	Wick et al. (2009)	0.00542
Oxycodon	Wick et al. (2009)	0.00875
Propranolol	Wick et al. (2009)	0.01708
Sotalol	Wick et al. (2009)	0.01729
17b estradiol	Zeng et al. (2009)	1.16333
Atenolol	Maurer et al. (2007)	0.02875
Metoprolol	Maurer et al. (2007)	0.02417
Propranolol	Maurer et al. (2007)	0.01625
Sotalol	Maurer et al. (2007)	0.01208
ATH	Joss et al. (2006)	0.06667
Bezafibrate	Joss et al. (2006)	0.10625
Clofibric acid	Joss et al. (2006)	0.02292
DAMI	Joss et al. (2006)	0.14167
Fenofibric acid	Joss et al. (2006)	0.37500
Fenoprofen	Joss et al. (2006)	0.50000
Gemfibrozil	Joss et al. (2006)	0.33333
Ibuprofen	Joss et al. (2006)	1.16667
Iohexol	Joss et al. (2006)	0.08750
Iomeprol	Joss et al. (2006)	0.05833
Iopromide	Joss et al. (2006)	0.08542
Ioxithalamic acid	Joss et al. (2006)	0.01875
N4-acetyl-sulfamethoxazole	Joss et al. (2006)	0.28125
Naproxen	Joss et al. (2006)	0.06042
Paracetamol	Joss et al. (2006)	2.87500
Piracetam	Joss et al. (2006)	0.14167
2,3-Dimethylphenol	Andreozzi et al. (2005)	0.81000
2,6-Dichlorophenol	Andreozzi et al. (2005)	0.08333
2-Chloro-4-nitrophenol	Andreozzi et al. (2005)	0.02333
3,4-Dihydroxybenzoic acid	Andreozzi et al. (2005)	1.48333
3-Nitrobenzoic acid	Andreozzi et al. (2005)	0.20333
4-Chlorophenol	Andreozzi et al. (2005)	0.13000
4-Hydroxybenzoic acid	Andreozzi et al. (2005)	2.41667
4-Nitroaniline	Andreozzi et al. (2005)	0.12000
4-Nitrophenol	Andreozzi et al. (2005)	0.11000
Benzene	Andreozzi et al. (2005)	0.67000
Benzoic acid	Andreozzi et al. (2005)	3.80000
Nitrobenzene	Andreozzi et al. (2005)	0.08000
o-Cresol	Andreozzi et al. (2005)	0.56000
Phenol	Andreozzi et al. (2005)	2.44000
1,3-Dinitrobenzene	Andreozzi et al. (2005)	0.01333
2,4-Dinitrophenol	Andreozzi et al. (2005)	0.01000
3,5-Dinitrobenzoic acid	Andreozzi et al. (2005)	0.00333
Gallic acid	Andreozzi et al. (2005)	0.43667
Ortophthalic acid	Andreozzi et al. (2005)	0.38667
Terephthalic acid	Andreozzi et al. (2005)	0.15667
17a ethynilestradiol	Urase et al. (2005)	0.00503
17b estradiol	Urase et al. (2005)	0.00000
Benzophenone	Urase et al. (2005)	0.13631
Bisphenol A	Urase et al. (2005)	0.01435
Carbamazepine	Urase et al. (2005)	0.01127
Clofibric acid	Urase et al. (2005)	0.00638
Diclofenac	Urase et al. (2005)	0.00000
Estrone	Urase et al. (2005)	0.04288
Fenoprofen	Urase et al. (2005)	0.06008
Gemfibrozil	Urase et al. (2005)	0.01953
Ibuprofen	Urase et al. (2005)	0.07548
Indomethacin	Urase et al. (2005)	0.20315
Ketoprofen	Urase et al. (2005)	0.02291
Naproxen	Urase et al. (2005)	0.00488
Propyphenazone	Urase et al. (2005)	0.01127
17b estradiol	Li et al. (2005)	1.86286
17b estradiol	Li et al. (2005)	1.52941
17b estradiol	Li et al. (2005)	1.95402

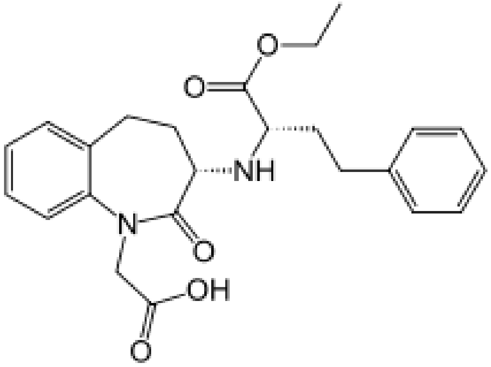
Appendix B: The 12 calculated molecular descriptors of each of the 65 unique compounds from the k_b data set, as well as the three external validation compounds. All descriptors were calculated using ChemBio3D Ultra 13.0 (CambridgeSoft: Cambridge, UK). Blacked-out cells indicate descriptors that were unable to be calculated due to software limitations and green cells indicate the three external validation compounds.

		GAMESS	GAMESS	GAMESS	GAMESS	GAMESS	ChemPropPro	ChemPropPro	ChemPropPro	ChemPropPro	ChemPropStandard	ChemPropStandard	ChemPropStandard
Compound	LogKb	Total E (Kcal/mol)	EHOMO (eV)	ELUMO (eV)	ELUMO-EHOMO	Dipole (Debye)	Gibbs (kJ/Mol)	HoF (KJ/Mol)	LogP	MR (cm^3/Mol)	A Area (Angs^2)	M Area (Angs^2)	MW
2,3-Dimethylphenol	-0.09151	-239389.8932	-8.329	3.918	12.247	1.132028	-35.36	-160.7	2.307	37.835	309.82	141.259	122.1644
2,6-Dichlorophenol	-1.07918	-763846.3515	-9.197	2.882	12.079	2.860887	-76.06	-150.9	3.36	37.362	301.359	134.991	163.00136
2-Chloro-4-nitrophenol	-1.63202	-604191.6955	-9.811	1.345	11.156	5.219265	-142.79	-296.27	2.056	■	316.194	144.148	174.5612514
3,4-Dihydroxybenzoic acid	0.171239	-354457.2269	-8.781	2.357	11.138	3.376815	-532.68	-656.17	0.86	36.205	313.37	141.811	154.12014
3-Nitrobenzoic acid	-0.69179	-387966.0977	-10.669	1.034	11.703	3.41168	-311.73	-474.13	1.183	■	325.913	149.383	168.1262914
4-Chlorophenol	-0.88606	-477259.8036	-8.681	3.328	12.009	2.402338	-54.5	-123.69	2.426	32.557	282.542	122.367	128.5563
4-Hydroxybenzoic acid	0.383217	-307745.7058	-9.16	2.479	11.639	1.89273	-378.06	-478.86	1.454	34.51	302.467	134.456	138.12074
4-Nitroaniline	-0.92082	-305239.3598	-8.675	2.073	10.748	7.739074	90.21	-69.43	1.013	■	279.279	119.265	139.1314314
4-Nitrophenol	-0.95861	-317607.805	-9.708	1.665	11.373	5.482523	-121.23	-269.06	1.437	■	293.527	129.912	140.1161914
1,3-Dinitrobenzene	-1.87506	-397825.5706	-11.187	0.542	11.729	5.078088	-54.9	-264.33	1.786	■	318.092	145.65	170.1217428
17a-ethynilestradiol	-2.29813	-574723.7221	-8.135	3.756	11.891	1.469536	283.58	-72.83	4.004	87.087	456.832	229.91	296.40336
17b estradiol (E2)	0.114586	-527501.4468	-8.149	3.74	11.889	1.371953	49.16	-338.69	3.909	79.618	424.203	210.633	272.38196
2,4-Dinitrophenol	-2	-444530.6742	-10.63	0.822	11.452	6.818212	-209.52	-441.64	1.397	■	325.001	150.681	186.1211428
3,5-Dinitrobenzoic acid	-2.47712	-514895.3387	-11.573	0.291	11.864	3.733186	-400.02	-646.71	1.143	■	361.582	172.027	214.1312428
Atenolol	-1.3408	-546902.7091	-8.222	3.912	12.134	3.928415	-50	-427.56	0.22	73.504	493.226	232.393	266.33608
ATH	-1.17609	-13625419.1	-9.002	1.583	10.585	7.554397	■	■	■	■	559.716	309.726	663.92765
Benzene	-0.17393	-143960.3515	-9.198	4.022	13.22	0.009683	121.68	80.83	2.058	26.058	245.905	99.919	78.11184
Benzoic acid	0.579784	-261032.6245	-9.656	2.239	11.895	2.424832	-223.44	-301.55	1.862	32.816	289.862	126.515	122.12134
benzophenone	-0.86546	-357533.3106	-9.303	1.642	10.945	3.022329	154.48	48.83	3.293	56.634	386.932	186.31	182.2179
Betaxolol	-0.60206	-609950.4231	-8.737	3.658	12.395	1.788967	1.9	-490.75	2.441	88.638	650.075	341.308	307.42776
Bezafibrate	-0.97367	-964037.397	-9.115	2.372	11.487	2.482826	-202.42	-604.56	3.771	95.461	635.51	335.224	361.8194
Bisoprolol	-1.53202	-657418.0442	-8.407	3.882	12.289	1.106033	-166.29	-701.05	1.938	92.151	615.607	296.349	325.44304
bisphenol A	-1.843	-453691.9787	-7.967	3.615	11.582	2.805083	-6.16	-243.24	3.32	68.207	395.663	188.455	228.28634
Caffeine	0.246745	-422017.0054	-8.441	2.698	11.139	4.141279	■	■	0.08	49.283	378.629	183.355	194.1906
carbamazepine	-2.00512	-473468.2194	-7.404	2.145	9.549	3.97864	■	■	2.926	70.351	439.573	222.099	236.26858
cefalexin	-0.92942	-921229.8605	-9.176	2.376	11.552	4.159613	5.42	-421.64	-0.667	88.979	566.065	301.529	347.38888
Celiprolol	-2.05799	-773388.117	-8.353	2.107	10.46	8.252825	-18.58	-644.81	0.907	106.413	638.267	324.078	379.49372
clofibric acid	-1.83416	-667397.8657	-9.128	3.226	12.354	1.790061	-321.9	-531.65	2.617	52.617	400.855	195.818	214.64554
Codeine	-0.70352	-606987.9345	-7.967	2.308	10.275	3.855022	238.9	-211.61	1.451	84.604	480.5	255.459	299.36424
DAMI	-0.84873	-13710433.13	-8.273	2.31	10.583	4.674992	-248.89	-686.09	1.716	124.344	618.022	337.346	719.04921
DHH	-2.17609	-567643.3688	-8.918	3.331	12.249	1.986921	120.79	-159.91	1.081	73.444	462.445	239.421	270.28326
diclofenac	-1.86967	-1036123.394	-8.055	2.827	10.882	3.261606	-15.45	-221.92	4.117	75.461	488.053	250.307	296.14864
Dihydrocodeine	-1.12494	-607800.9128	-7.721	4.095	11.816	5.438543	208.94	-269.39	1.633	83.641	493.13	264.202	301.38012
Doxepin	-1.69447	-537259.2196	-8.1	2.848	10.948	1.802136	453.24	119.33	3.722	89.126	554.594	289.341	279.37614
estrone	-1.36776	-526782.8804	-8.218	3.713	11.931	2.55226	71.1	-303.82	4.398	78.796	423.904	210.034	270.36608
Fenofibric acid	-0.42597	-880978.0229	-9.484	1.507	10.991	1.436535	-289.1	-563.65	3.826	83.192	545.646	283.388	318.7516
fenoprofen	-0.55278	-499688.8	-9.021	3.29	12.311	1.790414	-160.74	-379.11	3.639	68.181	481.163	241.52	242.26986
Gallic acid	-0.35985	-401171.097	-8.914	2.228	11.142	2.423909	-687.3	-833.48	0.47	37.899	322.732	147.98	170.11954
gemfibrozil	-0.75343	-502596.3829	-8.488	3.896	12.384	6.513803	-277.5	-630.58	4.411	71.819	461.413	219.237	250.33338
ibuprofen	-0.20686	-407178.4695	-8.797	3.601	12.398	2.027036	-187.43	-447.42	3.51	60.732	457.17	228.34	206.28082
indomethacin	-0.69217	-962575.8758	-7.918	1.689	9.607	2.42996	-82.69	-405.98	3.58	94.608	593.717	314.42	357.78764
Iohexol	-1.05799	-13947267.82	-9.066	2.24	11.306	12.545857	-616.82	-1224.39	0.373	144.98	681.127	399.569	821.13785
Iomeprol	-1.23408	-13851837.61	-9.145	2.151	11.296	6.995656	-494.4	-1025.6	0.726	134.428	684.724	398.887	777.08529
Iopromide	-1.06846	-13876194.35	-9.069	2.132	11.201	10.828122	-454.16	-1026.23	1.088	139.179	692.15	402.44	791.11187
Ioxithalamic acid	-1.727	-13555599.54	-9.172	2.025	11.197	5.326285	-310.11	-608.14	3.196	102.893	539.97	286.409	643.93955
ketoprofen	-1.64004	-523325.738	-9.392	1.838	11.23	3.233595	-176.24	-380.11	3.312	72.516	489.593	248.843	254.28056
Methadone	-2.05799	-586709.9928	-8.672	3.186	11.858	3.497052	333.02	-62.79	4.553	97.504	570.18	306.107	309.44518
Metoprolol	-1.70124	-537627.4546	-8.157	3.912	12.069	3.685789	-84.11	-501.63	2.177	76.696	591.311	301.677	267.3639
Morphine	-0.24988	-582636.4505	-7.993	2.292	10.285	4.223999	190.49	-224.59	0.73	79.835	457.176	240.118	285.33766
N4-acetyl-sulfamethoxazole	-0.55091	-825072.8508	-9.377	2.111	11.488	9.40982	■	■	0.574	72.877	511.282	257.42	295.31432
naproxen	-1.48613	-476055.6505	-8.198	2.099	10.297	6.607728	-184.55	-415.4	3.18	64.854	403.171	190.531	230.25916
Nitrobenzene	-1.09691	-270895.4473	-10.185	1.481	11.666	5.381458	33.39	-91.75	1.826	■	282.169	122.858	124.1167914
Nordiazepam	-2.26627	-760057.0986	-8.911	1.815	10.726	2.99722	395.36	194.58	2.74	74.915	426.701	207.604	270.71364
o-Cresol	-0.25181	-215031.2088	-8.382	3.801	12.183	1.355943	-34.15	-128.59	1.993	32.793	287.614	125.686	108.13782
Ortophthalic acid	-0.41266	-378090.26	-10.102	1.735	11.837	3.867824	-568.56	-683.93	0.79	39.575	332.413	152.14	166.13084
Oxycodon	-2.05799	-653707.445	-8.271	3.222	11.493	5.341919	88.57	-371.51	0.36	84.157	500.449	269.158	315.36364
Paracetamol	0.291091	-319714.6234	-7.909	3.377	11.286	3.69304	-84.81	-261.44	0.285	40.834	301.827	131.066	151.16256
Phenol	0.38739	-190671.7727	-8.571	3.811	12.382	1.658276	-32.94	-96.48	1.503	27.752	258.377	107.717	94.11124
Piracetam	-0.84873	-306814.1134	-9.849	4.506	14.355	4.29332	-18.28	-229.15	-1.385	35.061	326.972	150.765	142.1558
Propranolol	-1.77815	-513115.2614	-7.854	2.51	10.364	1.410289	135.96	-198.98	3.346	76.825	466.576	225.065	259.34344
propyphenazone	-1.94825	-452005.9948	-8.163	3.099	11.262	5.024429	355.1	24.3	2.107	69.923	473.847	242.757	230.30552
Sotalol	-1.83305	-746954.1252	-8.672	3.409	12.081	5.935361	■	■	-0.06	71.264	520.988	266.151	272.3638
sulfadiazine	-2.14027	-717303.0652	-8.686	2.737	11.423	9.007624	■	■	0.003	64.201	403.497	190.951	250.277
sulfamethoxazole	-0.85236	-730360.5348	-9.031	2.674	11.705	10.086124	■	■	0.863	64.495	448.545	220.38	253.27764
Terephthalic acid	-0.80502	-378077.2796	-10.569	0.615	11.184	0.297374	-568.56	-683.93	2	39.575	325.862	150.191	166.13084
Warfarin		-641653.9105	-8.479	2.08	10.559	3.404464	-163	-427.13	2.974	86.639	528.37	275.578	308.32794
Benezepril		-878831.1199	-8.983	2.945	11.928	5.914534	-164.1	-699.99	3.053	115.497	617.927	326.063	424.48952
Metformin		-268363.9793	-9.251	5.055	14.306	5.249738	656.62	425.56	0.148	36.111	291.535	124.957	129.16364

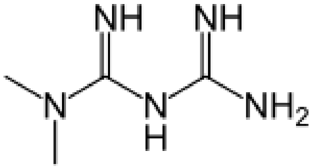
Appendix C: Synthetic wastewater formula based on a method from the Organization for Economic Cooperation and Development (OECD) (Pholchan et al., 2008), with the addition of several micronutrient salts (Jefferson et al., 2000).

Compound	**Mass Added**
Peptone	32 g
Beef extract	22 g
Urea	6.0 g
NaCl	1.4 g
$CaCl_2(2H_2O)$	0.8 g
K_2HPO_4	5.6 g
$MgSO_4(7H_2O)$	0.40 g
$FeSO_4(7H_2O)$	0.20 g
$MnCl_2(4H_2O)$	0.18 g
$CuSO_4(5H_2O)$	0.21 g
$Al(SO_4)_3(13H_2O)$	1.02 g
$ZnSO_4(7H_2O)$	0.22 g
$(NH_4)6Mo_7O_{24}(4H_2O)$	0.10 g
$CoC_4H_6O_4(4H_2O)$	2.11 g

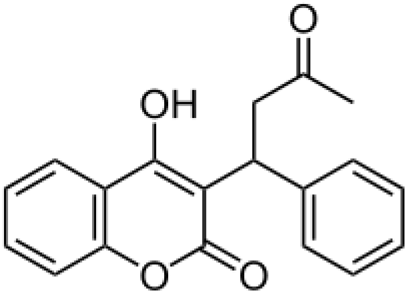
Appendix D: Molecular structures of the three external validation compounds as well as their functions. All three images gathered from Wikipedia.org.



Benazepril
Usage: treatment for high blood pressure



Metformin
Usage: antidiabetic drug



Warfarin
Usage: anticoagulant (previously used as a pesticide)

Appendix E: Internal validation parameters used in the Minitab “Best Subsets Regression” function and their associated equations (Minitab Inc.: State College, PA).

R^2

(coefficient of determination)

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

$$SS_{err} = \sum_i (y_i - f_i)^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$y_i = data\ set\ value$$

$$f_i = modelled\ value$$

Adjusted R^2

(adjusted to account for an increase in the number of descriptors)

$$Adjusted\ R^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{(n - p - 1)}\right]$$

n= sample size

p= number of model terms (including constant)

Mallows’ Cp

(compares precision and bias of each subset model against the model including all of the descriptors)

$$Mallows'Cp = \frac{SSE_P}{MSE_m} - (n - 2p)$$

$$SSE_p = \sum_{i=1}^{n} \left(y_i - f_{pi}\right)^2$$

$$f_{pi} = modelled\ value\ from\ p\ variable\ regression$$

$$MSE_m = \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n - p}$$

$$\widehat{y_i} = vector\ of\ n\ descriptors$$

S

(standard distance from regression line in units of response)

$$S = \frac{s}{\sqrt{n}}$$

$$s = sample\ standard\ deviation$$