7-2-2011

Quantitative structure-property relationships for predicting chlorine demand and disinfection byproducts formation in drinking water

Gebhard Luilo

Gebhard B. Lulls
Candidate

Chemistry and Chemical Biology
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

_____, Chairperson

_____

_____

_____

_____

_____

_____

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS FOR PREDICTING CHLORINE DEMAND AND DISINFECTION BYPRODUCTS FORMATION IN DRINKING WATER

By

Gebhard Balthazar Luilo

B.Sc. (Ed.), University of Dar es Salaam, 1998

M.Sc. (Environmental Science), University of Dar es Salaam, 2002

DISSERTATION

Submitted in Partial Fulfillment of

Requirements for the Degree of

Doctor of Philosophy

Chemistry

The University of New Mexico,

Albuquerque, New Mexico

May 2011

DEDICATION

This dissertation is in memory of Elizabeth, Kevin (Lucius) and Fausta.

ACKNOWLEDGEMENT

I heartily acknowledge Dr. Stephen Cabaniss, my advisor and dissertation chair, for his guidance during the research phase and write up of this dissertation. I also thank my committee members, Dr. Wei Wang, Dr. Deborah Evans and Dr. Andrew Schuler, for their valuable recommendations. Gratitude is extended to the Department of Chemistry and Chemical Biology and NSF-DEB for the funding to pursue this research. I also thank Dr. David Reckhow (University of Massachusetts) for inspiring the project and providing some of the data used in this dissertation.

To Prof. Pius Mbawala and Ms. Ruth Mollel (Former Permanent Secretary of the Ministry of Science, Technology and Higher Education, Tanzania) for their joint efforts to secure flight ticket to US for me; Prof. George Malekela and Mr. Kinemo Kihomano for moral and financial support; Ms. Joyce Mori and Prof. Tolly Mbwette of the Open University of Tanzania for administrative support while preparing for my departure to the US. Though a small word of thanks is not enough for your support, I do thank all of you from the bottom of my heart.

To Nancy and Harold Delaney, Kim and Steve Willard, Joseph Sempombe and Regina Mtei, and Peter Moyo for the company. And finally to Vestina Luilo, Erick and Mr. & Mrs. Venant Romanus, your love is the greatest gift of all that I will always cherish.

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS FOR PREDICTING CHLORINE DEMAND AND DISINFECTION BYPRODUCTS FORMATION IN DRINKING WATER

By

Gebhard Balthazar Luilo

B.Sc. (Ed.), Chemistry & Biology, University of Dar es Salaam, 1998

M.Sc. (Environmental Science), University of Dar es Salaam, 2002

Ph.D. (Chemistry), University of New Mexico, 2011

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Chemistry

The University of New Mexico

Albuquerque, New Mexico

May, 2011

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS FOR PREDICTING CHLORINE DEMAND AND DISINFECTION BYPRODUCTS FORMATION IN DRINKING WATER

By

Gebhard Balthazar Luilo

B.Sc. (Ed.), Chemistry & Biology, University of Dar es Salaam, 1998
M.Sc. (Environmental Science), University of Dar es Salaam, 2002
Ph.D. (Chemistry), University of New Mexico, 2011

ABSTRACT

Models are important tools for designing or redesigning water treatment processes and technologies to minimize disinfection byproducts (DBPs) formation without compromising disinfection efficiency. Empirical models, which are the most common, are based on bulk water quality parameters that vary with time and space. These parameters may not always have linear relationships with chlorine demand and DBPs formation which make structure-based models more attractive to study. In this dissertation, Quantitative Structure-Property Relationship (QSPR) models which make use of structural properties of individual molecules were developed using experimental data obtained from the literature. The amounts are reported in moles of chlorine (HOCl) consumed or DBP formed per mole of a compound (Cp). The QSPRs were derived by multiple

linear regression of chlorine demand or DBPs on a set of significant constitutional descriptors. The QSPRs were also tested for predictive power using cross validation and external validation for which the criteria were: $R_c^2 > 0.6$, $q^2 > 0.5$, $0.85 \leq k \leq 1.15$ and $R_t = (R_i^2-R_o^2)/R_i^2 < 0.1$.

The eight descriptor QSPR for HOCl demand had good statistics of fit (R_c^2 = 0.86 and SDE = 1.24 mol-HOCl/mol-Cp, N = 159) and also showed high predictive power on cross validation data (q^2_{LMO} = 0.86, $RMSE_{LMO}$ = 1.21 mol-Cl_2/mol-Cp) and external validation data (q^2_{ext} = 0.88, $RMSE_{LMO}$ = 1.17 mol-HOCl/mol-Cp). The QSPR also met all the criteria for QSPR predictive power and was robust. This model was integrated with AlphaStep model of natural organic matter (NOM) so as to estimate chlorine demand of surface waters. The predicted chlorine demand was 27.55 µmol-HOCl/mg-C which is comparable to 27-33 µmol-HOCl/mg-C reported for surface waters.

The 4 descriptor QSPR for total organic halide (TOX) formation had R_c^2 = 0.72 and SDE = 0.43 mol-Cl/mol-Cp. The Leave-One-Out validation of the QSPR (q^2_{LOO} = 0.60, RMSE = 0.5 mol-Cl/mol-Cp, N = 49) and external validation (q^2_{Ext} = 0.67, RMSE = 0.48 mol-Cl/mol-Cp, N = 12). These statistics showed that the QSPR had high predictive power and also was robust. Results from integration of the QSPR with AlphaStep predicted TOX in surface water to be 183.6 µmol-Cl/mg-C which comparable 170-298 µg-Cl/mol-Cp for the experimental TOX formation measured for whole dissolved organic matter.

Trichloromethane (TCM) and trichloroacetic (TCAA) were the two specific DBPs studied. The QSPR for TCM formation had three descriptors and statistics of fit were R_c^2 = 0.97 and SDE = 0.08 mol-TCM/mol-Cp and was validated by LMO data and external data. The results showed that LMO cross validation (q^2_{LMO} = 0.94, RMSE = 0.09 mol-TCM/mol-Cp, N = 90) and external validation (q^2_{Ext} = 0.94, RMSE = 0.08 mol-TCM/mol-Cp, N = 27) met criteria of predictive power and was therefore robust. The model prediction of 0.33 mol-TCM/mol-Cp was higher than 0.13 mol-TCM/mol-Cp observed for tannic acid. The QSPRs for predicting TCAA formation were developed but none of them met all the criteria for predictive power and were therefore not robust. The relationship between predicted TCAA and experimental data was too weak to be useful. This implies that TCAA formation has insignificant linear relationship with constitutional descriptors and it may better be predicted by QSPRs derived from non-linear algorithms. A major drawback of the constitutional descriptors is that they cannot explain electronic or steric effects. It is not easy to explain the differences in electron density and steric effects when same number of substituents occupy different position relative each other in aromatic ring (e.g., catechol vs. quinol). Use of geometrical descriptors (e.g., molecular volume, solvent accessible area), quantum-chemical descriptors (e.g., dipole moment, polarizability) or electrostatic descriptors (e.g., partial charge, polarity index) is recommended.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACN = Number of aliphatic carbons bonded to amines

AD = Applicability domain

AdjR_c^2 = Adjusted coefficient of determination for calibration

ArED = Total number of strong donors (OH and NH_2) attached to aromatic ring

ArED:C = Total number of strong donors (OH and NH_2) attached to aromatic ring per carbon

ArOH = Number phenols in a molecule

ArOH:C = Number of phenols per carbon

ArORact = # Alkoxy groups (OCH_3 and OC_2H_5) bonded to the aromatic ring without strong electron donors (NH_2 and OH)

ArORnact = # Alkoxy groups (OCH_3 and OC_2H_5) bonded to the aromatic ring with strong electron donors (NH_2 and OH)

ArOR:C = Sum of alkoxy groups (OCH_3 and OC_2H_5) bonded to the aromatic ring per carbon

AS = Number of aliphatic sulfur

BDCM = Bromodichloromethane

Chla = Chlorophyll a

CI = Carbonyl index

CORH:C = Total number of aldehyde and ketone groups per carbon in molecule

Cp = Compound

D = UV radiation dosage

DBPs = Disinfection byproducts

DCAA = Dichloroacetic acid

DFBETAS = Difference in beta standardized

DFFITS = Difference in fit standardized

D_i = Cook's distance

DOC = Dissolved organic carbon

DOM = Dissolved organic matter

EDCORH = The difference in number of strong electron donors per carbon and number of aldehydes and ketones per carbon in a molecule.

Eq = Equation

Eqs = Equations

EU = European Union

FA = Fulvic acid

g = Gram

GC/MS = Gas chromatography mass spectrometry

H:C = Atomic hydrogen to carbon ratio

HA = Humic acid

HAA5 = Total of mass concentrations (µg/L) of five species of HAAs

HAA6 = Total of mass concentrations (µg/L) of six species of HAAs

HAA9 = Total of mass concentrations (µg/L) of nine species of HAAs

HAAs = Haloacetic acids

HANs = Haloacetonitriles

HNMs = Halonitromethanes

$HOCl_{dem}$ = Chlorine demand

k = Slope of regression equation

k_i = Slope of regression equation with intercept

k_o = Slope of regression equation through origin

LFER = Linear free energy relationship

LMO = Leave-Many-Out

LMO_{cv} = Leave-Many-Out cross validation

logH:C = Logarithm of atomic hydrogen to carbon ratio

LOO = Leave-One-Out

LOO_{cv} = Leave-One-Out cross validation

MBD = Mean bias deviation (%)

MCL = Maximum contaminant level

mg = Milligram

MLR = Multiple linear regression

mmol = Millimole

mol = Mole

MS = Microsoft

MW = Molecular weight

NOM = Natural organic matter

NTU = Nephelometeric turbidity unit

O:C = Atomic oxygen to carbon ratio

OTactC = Number of doubly activated carbon in 1,3-disubstituted aromatic ring (e.g., 3-chlorophenols, 1,3-dihydroxybenzene)

q^2 = R square of cross validation and external validation

QSAR = Quantitative structure activity relationship

QSPR = Quantitative structure property relationship

QSTR = Quantitative structure toxicity relationship

RAI = Ring activation index

R_c^2 = Coefficient of determination for calibration

R_i^2 = R squared of regression with intercept

RMSE = Root Mean Square Error

RMSE = Root mean square error

R_o^2 = R squared of regression through origin

R_t = Ratio of R_i^2- R_o^2 to R_i^2

SDE = Standard deviation of estimation (regression)

SDR = Standardized residuals

sqrtArOH = Square root of number of phenols

sqrtHeA = Square root of number of heteroatoms

sqrtRAI = Square root of number of ring activation index

Stdev = Standard deviation

SUVA = Specific UV absorption

TBM = Tribromomethane

TCAA = Trichloroacetic acid

$TCAA_f$ = Trichloroacetic acid formation

TCM = Trichloromethane

TCM_f = Trichloromethane formation

THM4 = Total mass concentrations (µg/L) of the four trihalomethanes

THMs = Trihalomethanes

TOC = Total organic carbon

TOX = Total organic halides

TOX_f = Total organic halides formation

TTHM = Total mass concentrations (µg/L) of trihalomethanes

UV = Ultraviolet radiation

UV_{254} = UV absorbance at 254 nm

V = Volt

ε_j = Standard error of coefficient of a descriptor

µg = Microgram

µmol = Micromole

β_j = Coefficient of a descriptor

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1. Water Quality Supply and Health

Freshwater is a very important resource for sustaining our socio-economic development (*1*). However, natural water sources are rarely in pristine conditions due to microbiological and chemical contamination. Microbiological contamination of drinking water causes increasing cases of morbidity and mortality rates. Data on outbreaks of waterborne diseases outbreaks worldwide show that developing countries are more impacted than developed countries and that is due to poor access to sanitary and water supply services (*2,3,4*). For example, there were 39 outbreaks of waterborne diseases in US from 1999 - 2000 (*5*), 116 outbreaks in Sweden from 1980-1999 (*6*). Africa reported about 118,349 cases cholera and 5,853 deaths, Americas reported 17,760 cases and 225 deaths, Asia reported 11,293 cases and 196 deaths, Europe reported 18 cases and 1 death whereas Oceania reported 5 cases and zero deaths (*7*).

Africa emerges as the most impacted region of the world accounting for 80.3% and 93.3% of all cases and deaths reported to WHO in 1997 respectively. A glimpse of waterborne disease cases in Africa for 1990s indicated that there were 1,931 cases of cholera reported in refugee camps in Malawi in 1990 (*8*) and 5,600 cases in Nigeria in 1996 (*9*) and Tanzania reported about 40,249 cases in year 1997 (*7*). The global statistics of cholera outbreaks between 1995-2005

show that cholera outbreaks are still a global phenomenon and Africa is the most impacted continent (*10*). Thus, drinking water treatment using chemical disinfectants or UV radiation is required for water supply companies in order to protect the public health.

1.2. Drinking Water Disinfection

There are different methods with which pathogenic microbes can be eliminated from drinking water. The well known methods are use of chlorine (Cl_2), ozone (O_3), chlorine dioxide (ClO_2), potassium permanganate ($KMNO_4$), chloroamine (NH_2Cl) and UV radiation (*2,11*). These disinfectants will be discussed briefly in terms of their chemistry, advantages and disadvantages of using them as primary disinfectants. These oxidants inactivate bacteria but there are no clear mechanisms known and reported in literature. However, there are three proposed mechanisms by which pathogens get inactivated during water treatment (*11*).

i. The oxidants may destroy or impair cell wall of pathogens by attacking its cell constituents,
ii. Oxidants may enter the cells where they interfere with energy-providing metabolic process by making enzyme involved in metabolic process non-functional,
iii. Oxidants may enter the cells of pathogens where they interfere with synthesis of important proteins, amino acids, co-enzymes or cell wall which impair their growth.

However, the mechanism of action of disinfectant may vary depending on the type of oxidant employed and type of microorganisms. Microbes differ in their cell organization and mechanism of biosynthesis of important proteins or enzymes for their growth and energy-production and therefore one or more of these mechanisms will be involved.

1.2.1. Chlorination

Chlorine was first used as a water disinfectant in London in the 1850s following a cholera outbreak and it was first used to combat waterborne diseases in the US in 1908 (*12*). It is still the major disinfectant in most water treatment plants in the US and Europe (*13*) and is a reliable and cost effective disinfectant in developing countries (*2*). Chlorine is available commercially as chlorine gas or hypochlorite salts mostly as $Ca(OCl)_2$ and NaOCl. These chlorine sources have to be dissolved in water in order to produce the reactive oxidizing species.

Chlorine

Gaseous chlorine is dissolves in water to form hypochlorous acid (HOCl) and aqueous HCl (Eq. 1.1). The hypochlorous acid (pKa = 7.5) dissociates in water to give hypochlorite and hydronium ions (Eq. 1.2).

$$Cl_2(g) + H_2O(l) \rightarrow HOCl\ (aq) + HCl(aq) \quad K_{eq} = 3\times10^{-5} \text{ ..... (Eq. 1.1)}$$

$$HOCl(aq) + H_2O(l) \rightarrow H_3O^+\ (aq) + OCl^-(aq) \quad K_a = 3\times10^{-8} \text{ ..... (Eq. 1.2)}$$

The rates of formation of OCl^- and HOCl are pH dependent and at pH below 7.5 HOCl predominates over OCl^- species (*14,15*). Figure 1.1 provides information on the distribution of these two oxidizing species as function of increasing pH in

water. Predominance of HOCl over OCl^- at water treatment pH conditions is advantageous to achieving the goals of disinfecting drinking water from microbiological contaminants.

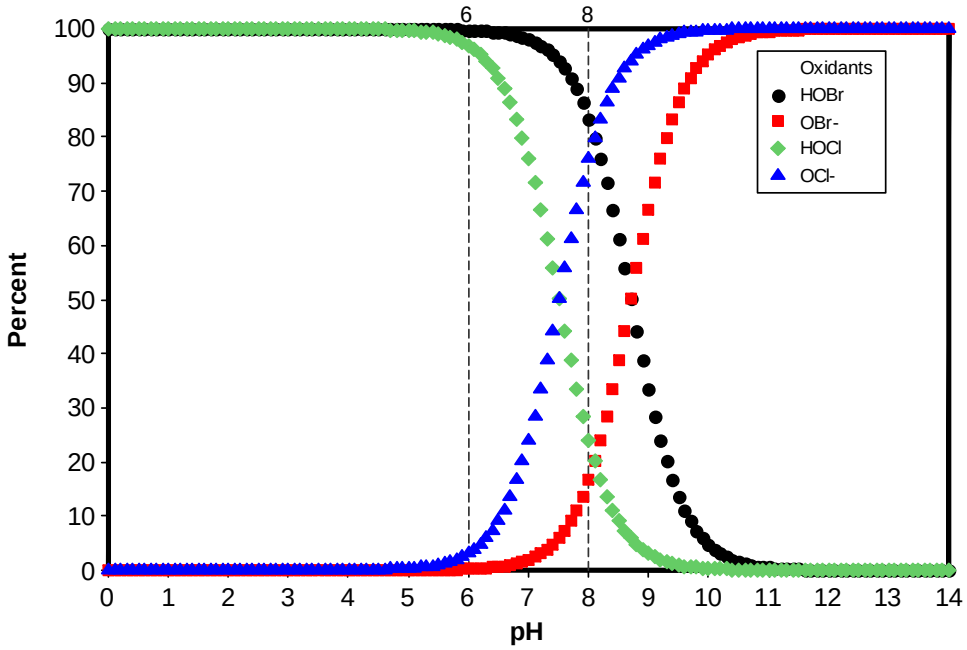


Figure 1.1. Distribution of HOCl and $HOBr^-$ species at various pH levels

First, HOCl has a reduction potential of 1.49 V that makes it a stronger oxidant than hypochlorite ion whose potential is 0.9 V (*16*). Second, neutral nature of HOCl is an added advantage in that OCl^- will face stronger electrostatic repulsion than HOCl on the surface of microbial pathogens. This is because most microbial pathogens, particularly bacteria and viruses, have negatively charged surfaces (*17,18,19*) and the same property has been utilized to separate bacteria by capillary electrophoresis (*20,21,22*). Thus, it is expected that HOCl will penetrate the cell wall more easily than OCl^-.

Hypochlorite

Sodium hypochlorite and calcium hypochlorite are the second most commonly used disinfectants in drinking water treatment systems (*14*). They dissolve in water to produce HOCl and OH^- (Eqs 1.3 & 1.4). The hydroxide ions produced in hydrolysis of these salts raise the pH of water above normal water treatment pH. That may affect the biocidal potency of HOCl. Thus, hypochlorites may work better with slightly acidic water than neutral water.

$$NaOCl(aq) + H_2O(l) \rightarrow HOCl(aq) + NaOH(aq) \quad \text{(Eq. 1.3)}$$

$$Ca(OCl)_2(s) + 2H_2O(l) \rightarrow 2HOCl(aq) + Ca(OH)_2(aq) \quad \text{(Eq. 1.4)}$$

Drinking water sources may contain bromide at concentrations from trace to 0.5 mg/L and desalinized water may have up to 1.0 mg/L (*23*) most of which comes from erosion of rock salts and degradation of methyl bromide used agriculturally (*24*). Bromide may also come from hypochlorite salts used to disinfect drinking water (*25,26*). Thus, hypochlorous acid reacts with bromide ions to give HOBr (pKa = 8.7) based on Equations 1.5 and 1.6. Figure 1.1 shows HOBr would be a better disinfectant than HOCl at around neutral drinking water treatment conditions because the former dissociates less than the latter (*27,28*). Thus, the presence of HOBr (E^o = 1.33 V) and HOCl (E^o = 1.46 V) together in water would increase microbial inactivation on one hand and increase amounts of disinfection byproducts formation on the other (*24*).

$$HOCl(aq) + Br^-(aq) \rightarrow BrCl(aq) + OH^-(aq) \quad \text{(Eq. 1.5)}$$

$$HO^{-}(aq) + BrCl(aq) \rightarrow HOBr(aq) + Cl^{-}(aq)$$ …….. (Eq. 1.6)

Chlorine is the most widely used disinfectant in water treatment and that may be due to the following attributes of chlorine (*2,11*): It has wide spectrum of biocidal effects; it leaves residual in water to take care of microbes in distribution system; it is cheap and it has proven to be effective in improving water quality over a century (*2*). The draw backs of chlorine are that it reacts with dissolved organic matter (DOM) to give chlorinated disinfection byproducts; chlorine gas requires strict handling and protective gears as it is toxic when released to air and NaOCl is corrosive; high doses of chlorine gives bad odor and taste to treated water (*11*). It is not very effective at eliminating *Cryptosporidium parvum* and *Giardia lamblia* (*2*).

1.2.2. Ozonation

Ozone is the disinfectant that came into use in Europe in the late 19th century and was first introduced in US in 1987 (*11*) and since then its use has grown gradually. Solubility of pure ozone in water at 20 ^{o}C is about 570 mg/L while the amount of ozone used for water treatment does not exceed 14% and hence its typical level in water treatment plants ranges between < 0.1 and 1 mg/L (*11*). Ozone chemistry is associated with two reaction pathways. The first is through direction oxidation of substrate by ozone and is prominent in acidic conditions; and the second is through ozone decomposition into hydroxy radicals that occurs mostly at high pH (*11*). The second is by reaction of ozone with

bromide ions in water to produce HOBr (Eq. 1.7) and OBr^-, formed from dissociation of HOBr, reacts with ozone to form BrO_3^- (Eq. 1.8).

$$O_3\ (g) + Br^-\ (aq) + H_2O(l) \rightarrow HOBr(aq) + OH^-(aq) + O_2(g) \quad \text{..... (Eq. 1.7)}$$

$$O_3(g) + OBr^-(aq) \rightarrow BrO_3^-(aq) + O_2(g) \quad \text{..... (Eq. 1.8)}$$

The advantages of ozone as disinfectant are: it is more effective disinfectant than chlorine because it attacks *Giardia* and *Cryptosporidium* species; it removes color and odor from water; it requires very short contact time, produces no halogenated DBPs in bromine free water and its activity is independent of pH of water (*11*). The disadvantages of ozone are that: it has high investment cost of infrastructure, it requires high operator skills and energy as it has to be prepared on site; it corrosive and toxic; it is decays very quickly at high pH and temperature and it leaves no residual to protect the water in distribution system (*11*).

1.2.3. Chlorine dioxide

Chlorine dioxide was first used in water treatment in 1950s and 40 years later 700 to 900 public water systems use it (*29*). Chlorine dioxide is a neutral molecule which oxidizes substrate by one-electron transfer mechanism and gets reduced to ClO_2^- as given by Equation 1.9 (*30,31*). This is the most predominant reaction in water treatment systems.

$$ClO_2(aq) + e^- \rightarrow ClO_2^-(aq) \quad E^o = 0.954\,V \quad \text{..... (Eq. 1.9)}$$

The dosage of chlorine dioxide in water ranges from 0.07 to 2.0 mg/L depending on water quality conditions though the standard is 1.0 mg/L (*11*). Although chlorine dioxide is not widely used in water treatment it has some advantages as an alternative disinfectant. It is more effective than chlorine in inactivating *Giardia* and *Cryptosporidium* species; removes odor and taste from decaying algae and vegetation; biocidal effect is not pH dependent; it is easier to generate than ozone and provides residual (*11*). Some of the disadvantages of chlorine dioxide are: it must be made on site and therefore requires skilled operators and expensive equipment; it decomposes easily when exposed to light and high temperature; it may produces noxious in some water treatment plants and produces chlorite and chlorate as byproducts which are expensive to measure (*11*).

1.2.4. Potassium permanganate

Potassium permanganate is supplied as crystalline solid and its solution, which is purple in color, by mixing it with water because its solubility in water at 20 ^{o}C is 6.0 mg/L (*11*). It is a good oxidizing agent under acidic conditions and basic conditions (Eqs 1.10 & 1.11) (*31*). These reactions release heat.

$$MnO_4^-(aq) + 4H^+(aq) + 3e^- \rightarrow MnO_2(s) + 2H_2O(l) \quad E^o = 1.68\,V \quad \text{.....(Eq. 1.10)}$$

$$MnO_4^-(aq) + 2H_2O(l) + 3e^- \rightarrow MnO_2(s) + 4OH^-(aq) \quad E^o = 0.60\,V \quad \text{..... (Eq. 1.11)}$$

There are a few advantages of potassium permanganate worth mentioning. It removes odor and taste from water; it is easy to store, transport

and apply to water; it works well with certain group of viruses; and controls nuisance organisms (*11*). Potassium permanganate has also some drawbacks that include: long contact time is required; it leaves pink color to the water; it require proper handling as it is toxic and irritant to skin and mucous membrane; over dosing may cause detrimental health effects (*11*).

1.2.5. Chloroamine

Chloroamine was identified as having oxidizing power in the early 20th century and was used regularly up to 1940s (*11*). Monochloroamine is the most predominant species in water at water treatment pH conditions. The species is produced in situ during chlorination of water with high ammonia levels (*31*) based on Equation 1.13. The chloroamine can also produced ex-situ by using the same reaction and maintain chlorine to ammonia ratio of 3:1 to 5:1 (*11*).

$$HOCl(aq) + NH_3(aq) \rightarrow NH_2Cl\ (aq) + H_2O(l) \quad K_a = 3\times10^{-8} \quad \text{..... (Eq. 1.13)}$$

The merits of chloroamines are that: they are not reactive to dissolve organic materials; chloroamine residual have long life time in water; no odors and taste problems; it not expensive and easy to make (*11*). The demerits of chloramines are that: it has less oxidizing power; it has to be generated on site; excess ammonia leads to nitrification problems; monochloramines lose oxidizing power at high pH (*11*).

1.2.6. Ultraviolet radiation

Ultraviolet radiation is a band of electromagnetic radiation waves, located between X-rays and visible regions of light spectrum, has a wavelength ranging

from 100 to 400 nm and the most potent biocidal wavelength range is 240-280 nm (*11*). The UV radiation dosage, D, is a product of light intensity, I (mW/cm^2) and exposure time, t, in seconds (i.e., D = I*t). Thus, its mechanism is unique in that it inactivates pathogens by triggering a series of photochemical reactions that leads to disruptions of essential molecules in their body (*11*).

The advantages of UV-radiation are that: It produces no chlorinated DBPs; it eliminates spores forming virus and bacteria; it requires short contact time; it is easy to operate and require minimum skills of operation when UV lamps are used (*11*). .The shortcomings of using UV radiation as a disinfectant are that: its inactivation efficiency requires water to have low UV absorbing organics and inorganics, low turbidity, low coloring materials; microbial aggregation or clumping limit efficiency of UV radiation (*11*).

1.2.7. Efficacy of water chlorination

When pros and cons of the alternative disinfectants are compared to the conventional disinfectant (chlorine), chlorine is the most preferred because it provides low capital, operating and maintenance costs (*32,33*), low technical skills required for operating and handling, and low price of chlorine (*2,11,33*). The most recent studies showed that chlorine tablets and sodium hypochlorite were reliable and cost effective water disinfectants. Application of these disinfectants to treat water at household level saved millions of lives of displaced people in tsunami and earthquake stricken areas of Indonesia and other neighboring countries (*34,35*).

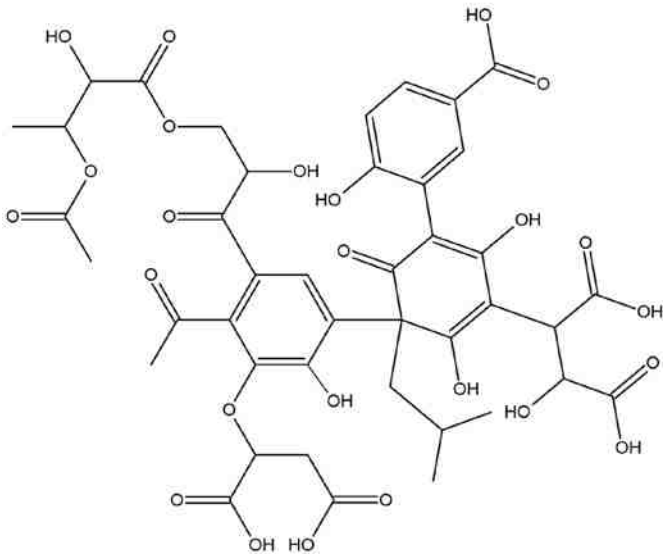
The use of chlorine disinfectant has proven effective in protecting the public health as evidenced by the decline in mortality and morbidity from waterborne diseases in US and Europe since introduction of chlorine in water treatment systems (*5,12*) and other countries in the world (*2*) and it remains the disinfectant of choice for most countries (*36*). However, little attention was paid to the fate of chemical disinfectants in water and their public health consequences until early 1970's.

1.3. Disinfection Byproducts Formation

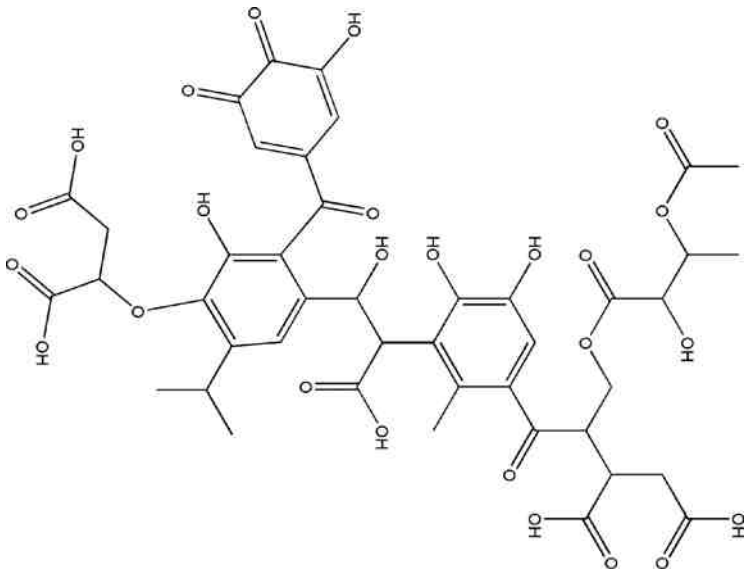
Natural organic matter (NOM) in fresh water is a heterogeneous mixture of organic materials of allochthonous, autochthonous and anthropogenic origins and the yellow-brown color of water is associated with high levels of NOM (*37*). NOM is operationally divided into dissolved organic matter (DOM), which is a mixture of carbonaceous materials that pass though a 0.45 μm pore filter and what remains on the filter is particulate organic matter (POM) (*37,38*). The major components of DOM are humic materials (fulvic acid, humic acid and humin) and non-humic materials (amino acids, and lipids, etc.) at various levels (*16*). In this context fulvic acid refers to a fraction of DOM soluble in water at pH<2; humic acid is the fraction of DOM that is insoluble in water only at pH<2 and humin is a DOM fraction that insoluble in water at all pH values (*38*). Nonetheless, components of DOM are not completely removed by conventional water treatment processes (*39,40,41*).

It is usually stated in the literature that chlorine or any alternative disinfectant reacts with DOM which may suggest that it is a single molecule.

Rather it implies that a disinfectant reacts with thousands of individual molecules of diverse chemical structures and molecular sizes to form known and unknown disinfection byproducts. With advances in spectroscopic methods attempts have been be made to characterize the DOM or its fractions. FTIR studies of DOM have shown that components of DOM have functional groups such as alcohols, carbonyls (aldehyde and ketones), amines, acids, etc (*42,43*). Arnold et al (*44*) used compound specific isotopic analysis of model compounds to predict the functional groups in surfaces water. Size exclusion chromatography was used to characterize molecular weights of components of freshwater DOM and molecular weights varied from less than 500 to more than 30,000 (*45,46*). Some studies have reported use computational methods to model the structures of DOM (*47*). Based on elemental analyses and spectroscopic information of DOM, structures of fulvic acid shown in Figure 1.2 (*48*) and humic acid (*49*) have been proposed. However, there is no study that has comprehensively characterized the structures of DOM and no consensus has been reached on proposed structures reported in the literature. Thus, studies of reaction of disinfectants with DOM are mainly based on integration of bench scale experiments using model compounds with raw water laboratory experiments or water treatment plant data. Tannic acid is a naturally occurring compound in plants (*50*) and its molecular structure is shown by Figure 1.3 (*51*) whereas commercially available tannic acid, such as corilagin, have slightly different structure from the natural one (*50*).



FA-2 (MW = 948)



FA-1 (MW = 960)

Figure 1.2. Proposed structures of fulvic acid

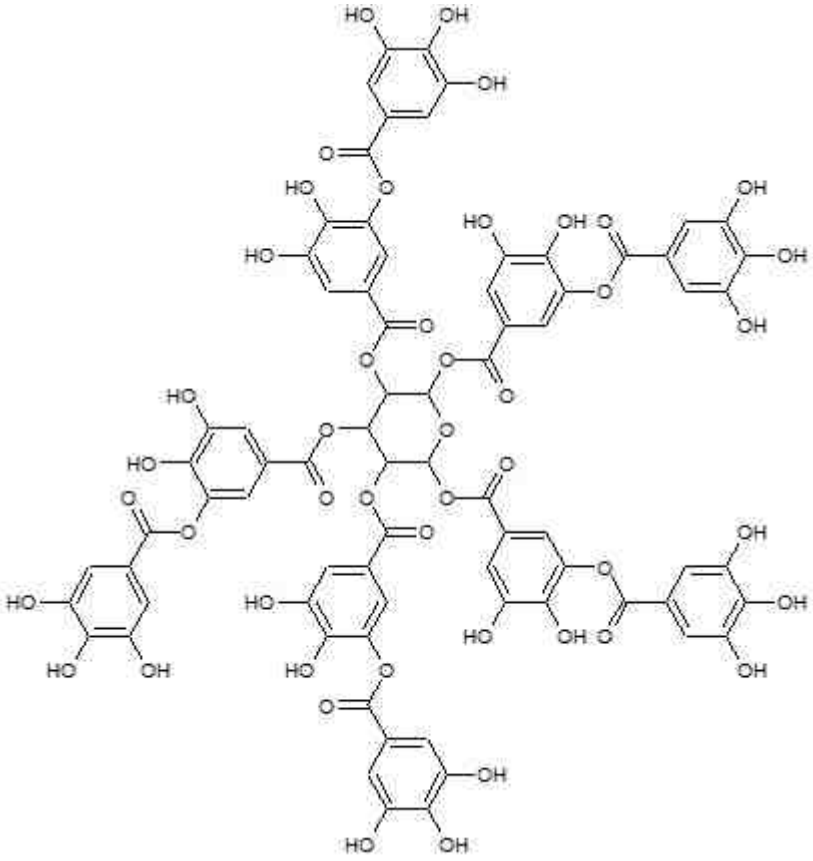


Figure 1.3. Structures of tannic acid (MW = 1701)

Chlorine, being a non-selective oxidant, will also react with traces of dissolved organic matter (DOM) in water to produce various species of chlorinated disinfection byproducts (DBPs), many of which are unknown (*52*). The reaction of DOM with chlorine in water is mostly through electrophilic substitution reactions for aromatic compounds (*16*) and electrophilic addition or elimination in aliphatic compounds and haloform reaction in the case of simple ketones or β-diketones (*53*).

Trihalomethanes were the first class of DBPs to be detected in potable drinking water in the US (*54,55*). The detection of chloroform in water sparked further research on the fate of chlorine in water which brought to light ten classes

of DBPs (*52*). Recent advances in analytical techniques have led to identification of new halogenated disinfection byproducts such as furanones, halonitromethanes, haloaldehydes, and haloketones (*56,57*). Trihalomethanes (THMs), haloacetic acids (HAAs), haloacetonitriles (HAN), halonitromethames (HNMs) and haloketones emanate from the chlorination process whereas ozonation process yields carboxylic acids, aldoketones and aldehydes (*52*). Chloroamination process produces nitrosamine and cyanogen halides whereas chlorine dioxide disinfectant yields oxyhalides (*52*). Of the ten disinfection byproducts, THMs and HAAs are of public health concern (*52,58*).

1.3.1. Trihalomethanes

Trihalomethanes are a group of chlorinated organic compounds formed in water when chlorine reacts with natural organic matter. In the presence of trace amounts of Br^- in water low levels of HOBr will also be present (Eqs 1.5 & 1.6) whereas in Br^- rich water HOBr levels will be higher than HOCl. The conjugate acid (neutral) forms are usually favored at water treatment pH of less than 7.5 (*15*). Thus, trihalomethanes formed in Br- rich water will have higher levels of brominated byproducts than chlorinated byproducts and account for over 90% of total THMs produced (*59*). Trihalomethanes represent the total mass concentrations (µg/L) of four species of trihalogenated methanes (denoted as THM4 or TTHM) formed from chlorination of DOM in water contaminated with bromide and the 4 species are: trichloromethane ($CHCl_3$), bromodichloromethane ($CHBrCl_2$), dibromochloromethane ($CHBr_2Cl$), tribromomethane ($CHBr_3$) (*60*). The maximum contaminant level, which is the highest allowable concentration of

a regulated contaminant in water delivered consumers of potable water supply, for THM4 is 80 µg/L in US (*61*). There are different precursors of DBPs in DOM matrices and based on results from studies using model compounds and raw water samples aromatic 1,3-dihydroxy substituted benzene (e.g., resorcinols) and β-diketones (e.g., acetylacetone) have been identified as key functional groups for DBPs formation (Arnold et al. 2008). A proposed mechanism for formation of chloroform (TCM) from chlorination of methyl ketone (*16*) and β-diketone is given in Figure 1.4 (*62*). The halogenation is a multistep process and both halogenated and non-halogenated products are possible.

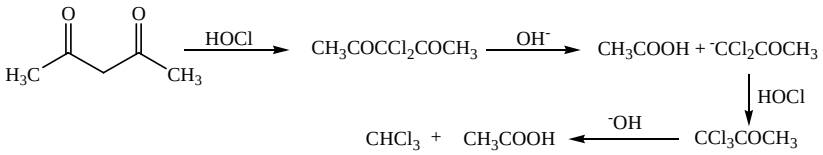


Figure 1.4. Proposed pathway for formation of TCM from ketones (*62*)

If all possible species of halomethanes are considered there are five other species of halomethanes, in addition to the four species above, namely CH_3Cl_1, CH_2Cl_2, $CH_2B_1Cl_1$, CH_3Br_1 and CH_2Br_2. If the water is free of bromide ions, only three possible species of chloromethanes (CH_3Cl_1, CH_2Cl_2, and $CHCl_3$) will be produced during chlorination of water.

1.3.2. Haloacetic acids

Haloacetic acids (HAAs) are a group of halogenated organic acids that formed from the reaction of chlorine and naturally occurring organic substances such as fulvic acids, humic acids and amino acids (*16,63*). The HAAs of significance in disinfected water are chloroacetic acid, dichloroacetic acid,

trichloroacetic acid, as well as some brominated forms (*15*). Some of the reaction pathways for the formation HAAs from aspartic acids (*64*) and cyanoethanoic acid (*65*) have been proposed. Since cyanoethanoic acid is the end product proposed for chlorination of aspartic acid, Peters et al. (*65*) used it as precursor to follow the reaction products. Figure 1.4 shows the reaction pathways for the formation of trichloroacetic acid from alanine based on proposed reaction mechanism for chlorination of a primary amino acid (*63*).

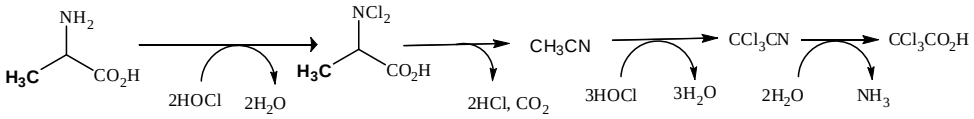


Figure 1.5. Proposed trichloroacetic acid formation from alanine

In water free of bromine, only three haloacetic acid species are possible which are CH_2Cl_1COOH, $CHCl_2COOH$ and CCl_3COOH. But in the presence of bromine there are also CH_2Br_1COOH and $CHBr_2COOH$, and the total mass concentrations of the five haloacetic acid species is denoted by HAA5 (*60*). The HAA5 is being regulated and its maximum contaminant level in drinking water is 60 µg/L in US (*61*), When an acronym, HAA6, is used in the literature it represents the total concentrations of haloacetic acid, i.e., HAA5 and $CHBr_1Cl_1COOH$ (*60,66*). There are also four other species of haloacetic acids that may be formed in presence of Br- ions namely CBr_2Cl_1COOH, CBr_1Cl_2COOH and CBr_3COOH. The total mass concentration of the nine species of HAAs (µg/L) is denoted as HAA9 (*60*).

1.3.3. Total organic halides (TOX)

Total organic halide (or total organic halogens) is term that refers to all organic compounds that contain covalently bonded chlorine, bromine and iodide formed from halogenation of water samples (*67*). TOX is comprised mostly of organic chlorides and organic bromines and in rare instances organic iodides when water has at least traces of iodine. Disinfection of drinking water using chlorinating agents produces the chlorinated disinfection byproducts discussed above but there are still many unidentified disinfection byproducts at the present time (*67,68*). Thus, measurement of TOX is used as a surrogate for toxic potential of disinfection byproducts formed from chlorination of drinking water (*39,67*). It has been reported that TOX varies linearly with activated aromatic content, UV absorbance and dissolved organic carbon (*69*). Nonetheless, TOX is currently not regulated in any country in the world (*67*).

1.3.4. Impacts of THMs and HAAs on public health

Studies conducted in animals have shown that the THMs and HAAs have toxic effects. Bromodichloromethane (BDCM) has been shown to reduce sperm motility in rats consuming 39 mg/kg of body weight per day in drinking water. BDCM and tribromomethane (TBM) induce tumors of the large intestine in rats (*58*). HAAs are carcinogenic and their effects appear to be limited to the liver and at high doses whereas dichloroacetic acid (DCAA) and trichloroacetic acid (TCAA) have tumorgenic effects in cell division and cell death (*58*). Some epidemiological studies conducted in Norway (*70*), in Chesapeake USA (*71*) and North Carolina USA (*72*) and in Nova Scotia, Canada (*73*) linked THMs

consumption to birth and urinary track defects in newborns, spontaneous miscarriages in women and stillbirth respectively. But studies elsewhere concluded that there is insufficient evidence to link THMs in water to the reproductive problems reported in previous studies (*74,75*). Currently there is insufficient evidence of carcinogenicity of THMs in humans (*76*) with the exception of one meta analysis study that showed that some association between drinking chlorinated water and risk to bladder cancer in men and women (*77*). Despite the insufficiency in toxicological data, these compounds have to be regulated based on precautionary principle because very little is known of their synergistic effects in the body (*78*). For example, the EU Directive 1998 has set maximum contaminant levels (MCL) for TTHMs at 150 µg/L in 2003 (*79*) and the US regulation sets MCL at 80 µg/L for THM and 60 µg/L for HAA5 (*61*)**.**

1.3.5. Challenges facing water supply authorities

The water authorities have the responsibility of supplying water to the public that meets requirements of the law. There are different ways the water treatment plant can achieve the goal of reducing DBPs in finished water. However, the waters supply companies may incur unnecessary costs if they treat water without prior knowledge of amounts of DBP precursors and the expected amounts of DBPs produced. The methods of acquiring such information should be fast, cheap, less sophisticated and reliable. This goal could be achieved by developing predictive models based on chlorine consumptions or formation of the THMs and HAAs during chlorination of drinking water.

1.4. Current Modeling Practices

Modeling refers to the use of mathematical equations to calculate the behavior of a natural system (*80*) and such equations are used to predict the effects of a set of independent variables on the dependent variable in the system. Current modeling practices in drinking water disinfection are primarily kinetic modeling and empirical modeling. The models were developed using either lab experimental data (using model compounds, river water, water from distribution system) or water treatment plant data. They have been used to predict either chlorine demand or formation of disinfection byproducts from model compounds or water samples obtained from water treatment plants or distribution systems.

1.4.1. Kinetic models

Kinetic models are a set of differential equations that simulate the disappearance of HOCl and or formation THMs and HAAs formation based on water quality parameters. Gallard and von Gunten (*81***)** and Gang et al. (*82*) have monitored TTHM and HAA9 formation and developed kinetic models based on bulk water parameters. Gallard and von Gunten (*83*) derived a rate equation for chlorine demands of phenols. They found out that the extent of formation of chloroform increases with time and there is a linear relationship between chloroform production and chlorine demand of phenols. On the hand, Norwood *et al. (84)* studied the rate of consumption of chlorine and formation of chloroform from resorcinol and phenolic model compounds with acid functional groups. The reactivity of resorcinol was relatively higher than that of other phenolic compounds under the same reaction conditions.

Mechanistic kinetic modeling monitors the disappearance and formation of intermediate products from a known precursor. de Laat et al. (*62*) monitored acetylacetone, 3,5-dichlorophenol and resorcinol using GC/MS. They found out that chlorination of precursor molecule is a multi-step process leading to formation of more than one chlorination byproduct. However, they did not derive any rate equations for the disappearance or the rate of formation of individual disinfection byproducts. Boyce and Hornig (*85*) performed mechanistic kinetic study to determine formation of products other than chloroform and bromoform from dihydroxyphenolic precursors at pH 4, 7 and 10. They found out that chlorination of the ring is usually followed by ring cleavage from which chloroform and other products are produced by elimination or addition process. They also used mass spectrometry to analyze the structure of the intermediate and final products. However, they did not report the rate of disappearance of precursor or rate of formation products.

Mechanistic kinetic models requires knowledge of the structure of each molecule and monitoring of the reaction intermediates over time, followed by derivation of a rate equation for each elementary reaction and finally solving the simultaneous equations to obtain the overall rate equations (*86*). The advantages are that it can provide information on rate of reaction, flexibility with reaction conditions and may provide reliable predictions of chlorine consumption or disinfection byproducts formation. The disadvantage is that it requires a lot of computation time when there is a large number of molecules in the database (*86*)

and we do not know the molecular structure of components of dissolved organic matter.

1.4.2. Empirical models

Empirical models relate $HOCl_{dem}$ or disinfection byproducts formation to bulk water quality parameters like pH, ultraviolet absorption, temperature, turbidity and dissolved organic carbon, etc. Although there are different approaches to empirical modeling, the discussion will be restricted to linear regression approaches which are within the scope of this research work.

Simple linear regression models

A simple linear regression model relates chlorine demand and DBPs formation to a single water quality parameter. It assumes that the relationship between the dependent variable (y) and independent variable (x) is linear with a y-intercept (b) as represented by Equation 1.14.

$$y = \beta x + b \quad \text{..... (Eq. 1.14)}$$

Reckhow et al. (*87*) studied chlorination of humic and fulvic acids at pH 7.0. The chlorine demand, THM formation potential (THMFP) and TOX formation were proportional to the concentration of activated aromatic carbons ([ActAr-R]), measured in µmol/mg-C, for both fulvic and humic acids (Eqs 1.15, 1.16 & 1.17). However, the authors did not directly measure activated aromatic content of humic or fulvic acid, rather they estimated it using probability with the assumption that aromatic ring substituted with OH and NH_2 are the most reactive. This approach worked better for fulvic acids than humic acids and the equations could

not account for chlorine demand or DBPs formation from deactivated aromatic compounds or aliphatic compounds.

$$HOCl_{dem} = a + 7.9[ActAr - R] \quad \text{..... (Eq. 1.15}$$

$$THMFP = 10 + 17[ActAr - R] \quad \text{..... (Eq. 1.16)}$$

$$TOX^{corr} = 50 + 85[ActAr - R] \quad \text{..... (Eq. 1.17)}$$

Total organic carbon (TOC) or dissolved organic carbon (DOC) has been reported to be a good indicator of chlorine demand and DBPs formation particularly in bench scale experiments (*88*). Simple linear regression was used to model DBPs formation from TOC using annual average data from 85 conventional water treatment plants in Pennsylvania (*89*). They found that annual average TOC did not show significant correlation with TTHM/HAA5. When plant-specific average TOC and TTHM/HAA5 data were used, they found a good correlation between the two parameters in one out of four plants with the highest average annual TOC (~ 2.3 - 3.1 mg/L) and in two out four plants with the lowest mean TOC (~ 0.6 - 0.8 mg/L). They arrived to a general conclusion that TOC is not a good indicator for THM or HAA formation (*89*).

Multiple linear regression models

Multiple linear regression (MLR) models for prediction of chlorine demand and disinfection byproducts formation have been reported. The chlorine demand or disinfection byproduct is a function of a linear combination of two or more water quality parameters given by Equation 1.18. β_j is the coefficient of j^{th}

descriptor, b is y-intercept (for regression through origin, $b = 0$) and ε_j is the error

$$y = \sum_{j=1}^{j=n} (\beta_j x_j + ... + \beta_n x_n) + b + \varepsilon \quad \text{..... (Eq. 1.18)}$$

Lekkas and Nikolaou (*90*) reported empirical models for prediction of THMs and HAAs from chlorination of raw water rich in bromide ions using bench scale experiments. They found that logarithm of THM (logTHM) was related to pH, time and chlorine dose (R^2 = 0.87, N = 192) whereas logHAA was a function of pH, Br, chlorine dose and time (R^2 = 0.52, N = 192). However, THM and HAA models behaved well with external data set at 95% confidence interval particularly for HAA.

Other authors have also reported MLR models for formation of THMs and HAAs from NOM at water treatment plants. Obolensky and Singer (*91*) generated models for prediction of DBPs formation. The THM4 (in log units) was a linear combination of turbidity, bromide, time, TOC, UV, Cl_2 dose, Cl_2 residual, alkalinity, temperature, and pre-chlorination dose. The model gave R^2 of 0.707 which indicates a good linear relationship but that could have been attributed to over fitting as N was 741 and the model's predictive power was not tested using external data. Golfinopoulos *et al.* (*92***)** developed MLR model for formation of TTHM with chlorophyll a (Chla), temperature, pH, chlorine, bromide, and sampling time (summer). But the residual plots showed that some points had very large residuals and the predictive power of the model was not tested against the external data.

Other authors have also reported the application of multiple linear regression models to predict disinfection byproducts formation using field water samples. THM formation databases from three water treatment plants were used to derive MLR THM formation models using pH, UV, TOC, DOC, Br-, chlorine dose and contact time as descriptors (*93*). The MLR models for prediction of DBPs formation in distribution systems using $KMnO_4$, residual chlorine, pH, temperature and contact time as descriptors have been reported (*94*). The MLR models from these two models were verified by independent data and there was a good agreement between predicted THM and observed THM. Nonetheless, some MLR models for DBPs formation derived using water treatment plant data reported in literature have low coefficients of determination (R^2) ranging from 0.35 to 0.62 (*92,95*). This implies that the models failed to explain about 40-60% of the variances in DBP formation. These observations show that regression models based on water quality parameters require rigorous testing using independent data from different water treatment plants in order determine model robustness and predictive power. The most common norm is that researchers calibrate the model and test its predictive power using same calibration data (internal validation) and sometimes autocorrelation analysis is not done to check for redundant descriptors. It also rarely involves use of external data to test the model predictive power and defining applicability domain in order to determine influential training data and prediction outliers (*96*).

Although empirical models using bulk water quality parameters as descriptors have been widely used in predicting chlorine consumption or DBPs

formation, they do have some drawbacks. Firstly, empirical approach treats DOM as a single**,** average entity while it is actually a mixture of organic molecules with different chemical structures (*97*). Secondly, the structure of the molecule is the key to reactive behavior of DOM towards HOCl. Types of functional groups and their arrangement in a molecule are expected to strongly affect reactivity towards chlorine during chlorination process (*62,66,88*). But this aspect of the molecule is not considered in development of the empirical models. Thirdly, the empirical models are derived by varying the water quality conditions (e.g. pH, turbidity, DOC, UV_{254}) though most of these are optimized prior to addition of chlorinating agent at water treatment plant. An alternative approach of developing predictive models based on molecular structure must be explored and that brings the discussion to QSPR modeling.

1.5. Quantitative Structure-Property Relationships (QSPRs)

Quantitative Structure-Activity Relationship (QSAR) is an alternative method of predicting biological behavior of molecules based on their chemical structures (*98,99*). The term was evolved in drug design and discovery fields where it is used to screen biological activity of drugs with similar chemical structures or screen a number of structural analogues to narrow down to those chemicals with potential biological activities (*100*). Since its evolution the application of QSAR principles have been extended beyond drug design and discovery. The term is, therefore, coined slightly different depending on the field of science in which the method is used. In medical sciences, where structure of molecules is related to toxicity molecules the term quantitative structure-toxicity

relationship (QSTR) is normally used (*101,102,103*). In physical and environmental sciences professions where physico-chemical property of molecules is related to the chemical structures of the molecules, the term quantitative structure-property relationship (QSPR) is frequently used interchangeably with QSAR. Regardless of slight differences in the terminology they all follow the same basic principles and protocols of model calibration and validation. However, in this work quantitative structure-property relationship would be used throughout unless otherwise mentioned.

1.5.1. Principles of quantitative structure-property relationships

A quantitative structure-property relationship is a modeling approach that attempts to relate structure of molecules to the property being measured. Early QSAR is accredited to the work of Hammett who correlated electronic properties of organic acids and bases with equilibrium constants and reactivity (*104*), termed Linear Free Energy Relationships (LFER). Hansch introduced lipophilicity descriptor that is represented by octanol-water partitioning coefficient (*105*). The use of structural properties to model chemical or physical property is more attractive in modeling because reactivity of a molecule with another molecule under optimized reaction conditions depends primarily on the molecular structures of the reactants (*106*). In the water treatment it is known that dissolved organic matter react with chlorine under the optimized water treatment conditions to form disinfection byproducts (Eq. 1.19).

$$DOM + HOCl(aq) \rightarrow DBPs \quad \text{..... (Eq. 1.19)}$$

The optimal water quality conditions, as recommended by WHO, for effective terminal chlorination include: 6 ≤ pH ≤ 8, turbidity ≤ 5 NTU, contact time ≥ 30 minutes and chlorine residual ≥ 0.5 mg/L (*107,108*). These parameters are essentially kept more or less constant (i.e., optimum levels) at the water treatment plant. Under such a reaction scenario only the molecular structure of individual molecules in DOM matrix is the most important (*106*) and that is why application of QSPR becomes of interest in drinking water treatment.

There are different algorithms for developing QSPRs in the literature and the most commonly used algorithm is multiple linear regression (*109*). Since quantitative structure relationships (QSAR and QSPR) are an extension of linear free energy modeling approach, linear regression algorithms (simple or multiple linear regressions) are the most convenient option for derivation of the models. Therefore the algorithm follows the fundamental assumptions of multiple linear regressions, with some flexibility (*95*):

i. The effect of each independent variable (x_i) on dependent variable (y) is linear and additive. Since bulk water quality parameters vary with time and space, they are not always linearly related to chlorine demand or DBPs formation which violates the assumption. However, the violation is usually ignored in regression techniques in water treatment system;

ii. Independent variables used in regression are assumed to be free of errors. This is not always true because there are some uncertainties in measurement of independent variables under laboratory conditions.

Correlation analysis of variables or determination the inflation factor may be used to test for independence of variables.

The model developed by multiple linear algorithms must also be analyzed for its adequacy by checking if it satisfies the key assumptions on model residuals (*95*):

i. Normal distribution of residuals, tested using normal score plots,
ii. Constant variance of the residuals, tested using residual plots,
iii. Independence of residuals, tested by using a plot of residual against observed data or predicted data,
iv. Average of residuals must be near zero and is tested by calculating mean of residuals.

1.5.2. Important steps in developing QSPRs

There two most commonly used steps in QSPR modeling which are data preparation and model generation. Additional steps that are highly recommended, though not commonly used in the past, include model validation using cross validation and external data and defining model applicability domain (*96,110*). These steps are elaborated as follows:

i. The data preparation involves lab or literature data collection for the target property to be modeled and calculation of descriptors to be used. The descriptors can be calculated manually for constitutional descriptors or using special software for other descriptors (typological, quantum-chemical, etc).
ii. Selection of significant descriptors to be used for calibration of the model because not all descriptors generated will be statistically significant. It is

important to check for redundant descriptors by performing multi-colinearity analysis of selected descriptors.

iii. Model generation involves establishment of statistical relationship between the target property (y) and the list of significant descriptors (x).

iv. Model validation which involves assessment of model robustness and predictive power using internal and external data.

v. Definitions of the applicability domains, AD (in this context AD is the range of values within which a model is calibrated and prediction is reliable) of the model which assess influential points in model calibration data and reveal which molecules in the external data are predicted due to extrapolation of the model. An influential data point, in this context, refers to value of an observation that changes coefficients of model descriptors or fits when the data point is omitted from the calibration data set.

Since QSPR is calibrated with a set of experimental data from compounds of known structures, it becomes easier to predict properties of a new compound of similar structure without performing any experiment (*111*)**.** This is done by calculating the descriptors and plugging them into the QSPR equation which saves time and resources. There are several QSPRs or QSARs derived using linear regression algorithms reported in literature (*112,113,114*). However, not all of models reported in literature have been evaluated for predictive power (using validation data), presence of redundant descriptors and applicability domain (AD). A few studies reported in literature are used as examples to emphasize the importance as two steps are rarely performed. Lack of model external validation

and AD analysis may raise question over robustness of most models to predict the property of interest using different molecules or systems.

The QSPRs for prediction of refractive index of polymers and surface tension of non-surfactants have been reported (*115,116*). However authors did not go farther to analyze the predictive power of the model using external validation data set. QSPR has also been used to predict soil sorption coefficient of organic pollutants, aqueous solubility of drug-like organic compounds; chromatographic retention time and boiling point of organic compounds and UV absorption intensities of organic molecules (*117-120*). The model predictive power in these studies was validated by cross validation and external data.

QSAR has been successfully used to predict toxicities and biodegradation of anilines and phenols (*121*) whereas QSTR was used to predict toxicities of aliphatic compounds and organic compounds (*122,123*). The models in these studies were not tested for predictive power using external data. Heat of formation and enthalpy of formation of organic compounds (*124,125*) and polarographic wave half life of benzonoids (*126*) have been predicted by QSPR but the models were not evaluated for their predictive power using external data as well. Of the studies reviewed here only Katritzky et al. (*120*) went further at defining applicability domain of the QSPR which is rare in literature though it is highly recommended (*96,110, 127,128*)

Model internal and external validations provide information on model stability and to what extent it can be used to predict the property of interest using new data. Applicability domain helps to identify which data points in calibration

data set are outliers or influential and therefore affect the model fits and coefficients of descriptors. Thus, high model calibration R^2 and high R^2 of validation (q^2) have to be interpreted with caution because they may be high due to over fitting of the model or presence of a few data points that were outliers (i.e., a data point with a value far from other data) or both outliers and influential.

1.6. Statement of the Research Problem

Despite numerous applications of QSPR in environmental studies, we are unaware of any attempts to predict $HOCl_{dem}$ or disinfection byproducts formation using this approach. This approach is not a substitute for traditional experimental measurements (e.g., routine jar test), rather it destined to enable predictions of HOCl demand and disinfection byproducts due to potential changes in the DOM such as changing pre-treatment methods or land use within the watershed. A predictive model has been developed which provides composition data for thousands of molecules in a DOM mixture (*129,130*). What is lacking is a rapid and quantitative method to predict HOCl demand from the molecular information. In this work experimental data from the literature on $HOCl_{dem}$ and disinfection byproducts formation from small molecules are used to calibrate and validate a QSPR that predicts $HOCl_{dem}$ and disinfection byproducts based solely on constitutional descriptors. The novel MLR QSPRs that will be used to predict $HOCl_{dem}$ and disinfection byproducts formation from model DOM structures which have never reported before. The QSPRs for $HOCl_{dem}$ and TOX formation will be interfaced with the AlphaStep model of NOM for prediction chlorine demand and DBPs formation in surface waters used in water treatment facilities or watershed

catchment waters. It is anticipated that the models will be quick tools for screening the chlorine demand of water and potential production of DBPs formation. This would not only save time and resources but also minimize excessive formation of disinfection byproducts.

1.7. Statement of Goals and Objectives

The main goals of this research are: to develop Quantitative Structure-Property Relationships (QSPRs) that will predict chlorine demand and formation of THMs and HAAs from the precursor model compounds and determine the performance of the QSPR models to predict the chlorine demand and THMs and HAAs formation of the test compounds.

The plan is envisaged to accomplish the following specific objectives:

1) Develop QSPRs to predict HOCl consumption by known precursor model compounds that include aromatic and aliphatic compounds.

2) Develop QSPRs for prediction of trichloromethane (TCM), trichloroacetic acid (TCAA) and total organic halide (TOX) formation from aromatic and aliphatic model compounds.

3) Evaluate the performance of QSPRs to predict HOCl demand, TCM, TCAA and TOX formation using external data and determine the applicability domains.

4) Integrate the QSPRs for HOCl demand and TOX formation with AlphaStep model of natural organic matter in order to estimate HOCl demand and TOX formation from to chlorination of NOM in surface water.

1.8. Dissertation Organization

This dissertation is comprised of seven chapters. Chapter 1 explores different disinfecting techniques and efficacy of water chlorination and provides a brief overview of the principles of QSPR and potential application of QSPR in water treatment industry. The chapter closes with objectives of the current research work. Chapter 2 provides information on general research methodology used in this work. It covers data collection, descriptor generation and data splitting into training and external data sets, QSPR calibration and evaluation. Chapter 3 covers QSPR development for predicting chlorine demand (published in *Environmental Science and Technology, 2010,* 44(7), 2503-2508). Chapter 4 describes calibration and evaluation of the QSPR for predicting total organic halides (TOX), Chapter 5 covers QSPR development for prediction of trichloromethane (Submitted to *SAR & QSAR in Environmental Research*) and Chapter 6 discusses QSPRs for predicting trichloroacetic acid formation. Chapter 7 gives a summary the results from the four QSPRs and also covers the integration of QSPRs for chlorine demand and TOX formation with AlphaStep NOM model. The chapter also summarizes implications of descriptors and limitation of the QSPRs and also gives recommendations for future work.

1.9. References

1. van Hylckama, T.E.A. *water Resources.* In: Environment, resources, pollution and society, Murdoch, W.W. (ed.), Sinauer Associates Inc., Stamford, CT: 1971, pp. 135-155.

2. Galal-Gorchev, H. Chlorine in water disinfection. *Pure Appl. Chem*. **1996**, 68(9), 1731-1735.

3. Hermer, R. Water quality and health. *The Environmentalist*. **1999**. 19(1), 11-16.

4. Ballester, F.; Sunyer, J. Water and health: precaution must be guided for the health of the public. *J. Epidemiol. Community Health*. **2000**, 54, 729-730.

5. Lee S.H.; Levy D.A.; Craun, G.F.; Beach M.J.; Calderon R.L. Surveillance for waterborne-disease outbreaks-United States, 1999-2000. *Morbid. Mortal. Wkly Rep*. **2002**, 51(SS-8), 1-28.

6. Anderson, Y.; Bohan, P. *Disease surveillance and waterborne disease outbreaks.* In: Water quality guidelines, standards and health: Risk assessment and management for water related infectious diseases, Fewtrell, L.; Batram, J. (eds), WHO and IWA Publishers: London: 2000, pp 115-133.

7. WHO (World Health Organization). Cholera in 1997. *Wkly Epidemiol. Rec.* **1998**, 73(27), 201-208.

8. Swerdlow, D.L.; Malenga, G.; Begkoyian, G.; Nyangulu, D.; Toole, M.; Waldman, R.J.; Puhr, D.N.; Tauxe, R.V. Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. *Epidemiol. Infect.* **1997**, 118(3), 207-214.

9. Hutin, Y.; Luby, S.; Paquet, C. A large cholera outbreak in Kano City, Nigeria: the importance of hand washing with soap and the danger of street-vended water. *J. Water Health*, **2003**, 1(1): 45-52.

10. Griffith, D.C.; Kelly-Hope, L.A.; Miller, M.A. Review of reported cholera outbreaks worldwide. *Am. J. Trop. Med. Hyg.* **2006**, 75(5), 973–977.

11. USEPA (U.S. Environmental Protection Agency). *Alternative disinfectants and oxidants guidance manual.* United States Office of Water EPA 815-R-99-014: 1999

12. Christman, K. History of chlorine. Chlorine Chemistry Council, Waterworld : 1998. http://c3.oreg/chlorine_knowledge_center/hioistroy.htm

13. Chaidou, C.I.; Georgakilas, V.I.; Stalikas, C.; Saraçi, M.; Lahaniatis, E.S.; Formation of chloroform by aqueous chlorination of organic compounds. *Chemosphere*. **1999**, 39(4), 587-594.

14. White, G.C. *Handbook of chlorination and alternative disinfectants*. Volume 3. VanNostrand Reinhold Co. New York: 1992.

15. Nikolaou, A.D.; Kostopoulou, M.N.; Lekkas, T.D. Organic by-products of drinking water chlorination. *Global Nest: Int. J.* **1999**, 1(3), 143-156.

16. Larson, R.A.; Weber, E.J. *Reaction mechanisms in environmental organic chemistry*; Lewis Publishers: New York: 1994.

17. Wilson, W.W.; Wade, M.M.; Holman, S.C.; Champlin, F.R. Status of methods for assessing bacterial cell surface charge properties based on zeta potential measurements. *J. Microbiol. Methods*. **2001**, 43: 153-164.

18. Martinez, R.E.; Pokrovsky, O.S.; Schott, J.; Oelkers, E.H.J. Surface charge and zeta-potential of metabolically active and dead cyanobacteria. *Colloid. Interface Sci.* **2008**, 323(2), 317-325.

19. Soni. K.A.; Balasubramanian. A.K.; Beskok. A.; Pillai, S.D. Zeta potential of selected bacteria in drinking water when dead, starved, or exposed to minimal and rich culture media. *Curr. Microbiol.* **2008**, 56, 93–97.

20. Moon, B.G.; Kim, Y. Analysis of health-related microbes by capillary electrophoresis. *Bull. Korean Chem. Soc*. **2003**, 24(8), 1203-1206.

21. Moon, B.G.; Lee. Y-I.; Kang, S.H.; Kim, Y. Capillary electrophoresis of microbes. *Bull. Korean Chem. Soc*. **2003**, 24(1), 81-84

22. Rodriguez, M.A.; Lantz, A.W.; Armstrong, D.W. Capillary electrophoretic method for the detection of bacterial contamination. *Anal. Chem*. **2006**, 78(14), 4759-4767.

23. WHO (World Health Organization). *Bromide in drinking-water, Background document for development* of WHO guidelines for drinking-water quality: 2009. http://whqlibdoc.who.int/hq/2009/WHO_HSE_WSH_09.01_6_eng.pdf

24. USEPA (US Environmental Protection Agency). *Effect of bromide on chlorination byproducts in finished drinking water,* Environmental Protection Agency Project Summary, EPA/800/S2-91/036, Cincinnati, OH*:* 1991. nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=30003U2U.txt

25. Cooper, W.J.; Zika, R.G.; Steinhauer, M.S. Bromide oxidant interactions and THM formation: A literature review. *J. Am. Water Works Assoc.* **77**, 116-121.

26. Siddiqui, M.S.; Amy, G.L.; Murphy, B.D. Ozone enhanced removal of natural organic matter from drinking water sources. *Water Res.* **1997**, 31(12), 3098-3106

27. Upsher, F.J.; Fletcher, L.E. *Review of chlorine and organohalides and their significance to the Royal Australian Navy*. Technical Report, Aeronautical and Maritime Research Laboratory, Melbourne, Australia: 1996. http://dspace.dsto.defence.gov.au/dspace/handle/1947/3916

28. DOW (Dow Chemical Company). *FILMTEC™ membranes water chemistry and pretreatment: biological fouling prevention*. Tech Manual Excerpt, Form No. 609-02034-1004: 2004.

29. Hoehn, R.C. *Chlorine dioxide use in water treatment: Key issues*. Proceedings of the second international symposium on chlorine dioxide: Drinking water issues, Houston, TX: 1992.

30. Hoehn, R.C.; Rosenblatt, A.A.; Gates, D.J. *Considerations for chlorine dioxide treatment of drinking water.* Conference proceedings, AWWA water quality technology conference, Boston, MA: 1996.

31. Lide, D.L. (ed.), *CRC handbook of chemistry and physics,* 71st edn. CRC Press, Boca Raton, FL: 1990.

32. PNL (Pacific Northwest Laboratories). *Disinfection technologies for potable water and waste water treatment:* Alternatives to chlorine. Pacific Northwest Laboratories: 1998.

33. ACC (American Chemical Council). *The Benefits of Chlorine Chemistry in Water Treatment.* Report prepared by Whitfield & Associates for the Chlorine Chemistry Division of the American Chemistry Council: 2008.

34. Clasen, T.; Smith, L. *The Drinking water response to the Indian Ocean Tsunami including the role of household water treatment*. World Health Organization sustainable development and healthy environments: 2005

35. Gupta, S.K.; Suantio. A.; Gray. A.; Widyastuti. E.; Jain. N.; Rolos, R.; Hoekstra, R.M.; Quick, R. Factors associated with *E. coli* contamination of household drinking water among tsunami and earthquake survivors, Indonesia. *Am. J.Trop. Med. Hyg*. **2007**, 76(6), 1158-1166

36. WQHC (Water Quality and Health Council). *Drinking Water & Health Quarterly*, **2002**, 8(1).

37. Peuravuori, J.; Pihlaja, K. *Characterization of freshwater humic matter,* In: Handbook of water analysis, 2nd edn, Nollet, L.M.L. (ed.), CRC Press, Boca Raton, FL: 2007: pp 435-447.

38. Clesceri, L.S.; Greenberg, A.E.; Eaton, A.D. (eds). *Standard methods for the examination of water and wastewater,* 20th edn, APHA, Washington DC: 1998.

39. Ribas, F.; Frias, J.; Lucena, F.A new dynamic method for the rapid determination of the biodegradable dissolved organic carbon in drinking water. *J. Appl. Bacteriol.* **1991**, 71(4), 371-370

40. Volk, C., Wood, L.; Johnson, B.; Robinson, J.; Zhuc, H.W.; Kaplan, L. Monitoring dissolved organic carbon in surface and drinking waters. *J. Environ. Monit.* **2002**, 4(1), 43-47.

41. Volk. C.; Kaplan, L.A, Robinson, J., Johnson, B., Wood, L., Zhu, H.W., LeChevallier, M. Fluctuations of dissolved organic matter in river used for drinking water and impacts on conventional treatment plant performance. *Environ. Sci. Technol*. **2005**, 39(11), 4258-4264

42. Wershaw, R.L.; Leenheer, J.A.; Cox, L. *Characterization of dissolved and particulate natural organic matter (NOM) in Neversink Reservoir, New York*. Scientific Investigations Report 2005-5108. U.S. Geological Survey: 2005.

43. Kanokkantapong, V.; Marhaba, T.F.; Panyapinyophol, B.; Pavasant, P. FTIR evaluation of functional groups involved in the formation of haloacetic acids during the chlorination of raw water. *J. Hazard. Mater*. **2006**, 136, 188-196.

44. Arnold, W.A.; Bolotin, J.; von Gunten, U.; Hofstetter, T.B.. Evaluation of functional groups responsible for chloroform formation during water chlorination using compound specific isotope analysis. *Environ. Sci. Technol*. **2008**, 42, 7778-7785.

45. Pelekani, C.; Newcombe, G.; Snoeyink, V.L.; Hepplewhite, C.; Assemi, S.; Beckett R. Characterization of natural organic matter using high performance size exclusion chromatography. *Environ. Sci. Technol*. **1999**, 33 (16), 2807-2813.

46. Fiorentino, G.; Spaccini, R.; Piccolo, A. Separation of molecular constituents from a humic acid by solid-phase extraction following a transesterification reaction. *Talanta*. **2006**, 68, 1135–1142.

47. Alvarez-Puebla, R.A., Valenzuela-Calahorro C.; Garrido, J.J. Theoretical study on fulvic acid structure, conformation and aggregation: A molecular modeling approach. *Sci.Total Environ.* **2006**, 358, 243–254.

48. Leenheer, J.A.; Brown, G.K.; MacCarthy, P.; Cabaniss, S.,E. Models of metal binding structures in fulvic acid from the Suwannee River, Georgia**.** *Environ. Sci. Technol.* **1998**, *32*, 2410-2416.

49. Stevenson J. *Humus chemistry: genesis, composition, reactions*. John Wiley & Sons, New York: 1982.

50. Hayat, M. A. *Principles and techniques of electron microscopy: Biological applications*. 4th edn, Cambridge University Press, Cambridge, UK: 2000.

51. ChemBlink, *Online Database of Chemicals from Around the World*

http://www.chemblink.com/products/1401-55-4.htm

52. Crittenden, J.C.; Trussell, R.R.; Hand, D.W.; Howe, K.J.; Tchobanoglous, G. *Water treatment: Principles and design*, 2nd, edn, John Wiley & Sons Inc., New York: 2005.

53. USEPA (US Environmental Protection Agency). *Formation of halogenated organics by chlorination of water supplies.* USEPA, National Service Center for Environmental Publication (NEPIS), 1975. http://nepis.epa.gov/EPA/html/pubs/pubtitleORD.htm

54. Bellar, T.A.; Lichtenberg, J.J.; Kroner, R.C. The occurrence of organohalides in chlorinated drinking waters. *J. Am. Water Works Assoc*. **1974**, 66, 703-706.

55. Rook, J.J. Formation of haloforms during chlorination of natural waters. *Water Treat. Exam*. **1974**, 23, 234-243.

56. Onstad, G.D.; Weinberg, H.S.; Stuart, W.K. Occurrence of halogenated furanones in US drinking waters. *Environ. Sci. Technol*. **2008**, 42 (9), 3341-3348

57. Krasner, S.K.; Weinberg, H.S.; Richardson, S.D.; Pastor, S.J.; Chinn, R.; Sclimenti, M.J.; Onstad, G.D.; Thruston Jr, A.D. Occurrence of a New Generation of Disinfection Byproducts *Environ. Sci. Technol*. **2006**, 40 (23), 7175-7185.

58. IPCS (International Program on Chemical Safety)*. Disinfectants and disinfectant byproducts formation.* Environmental Health Criteria 216; WHO; Geneva: 2000.

59. Golfinopoulos, S.K.; Kostopoulou, M.N.; Lekkas, T.D. THM formation in the high-bromide water supply of Athens. *J. Environ. Sci. Health A*. **1996**, 31(1), 67-81

60. USEPA (US Environmental Protection Agency). *Disinfection Byproducts*: A Reference Resource. Information Collection Rule (ICR). http://www.epa.gov/envirofw/html/icr/gloss_dbp.html. (Accessed on October 2, 2010).

61. USEPA (US environmental Protection Agency). *National primary drinking water regulations: disinfectants and disinfection by-products*; Final rule, Federal Registry, 63:241:69390: 1998.

62. de Laat, J.; Merlet, N.; Doré, M. Chlorination of organic compounds: chlorine demand and reactivity in relationship to the trihalomethane formation. *Water Res.* **1982**, *16*, 1437-1441

63. Hureiki, L.; Croué, J-P.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res.* **1994**, *28*, 2521–2531.

64. Trehy, M.L.; Yost, R.A.; Miles, C.J. Chlorination byproducts of amino acids in natural waters. *Environ.Sci. Technol.* **1986**, 20(11), 1117-1122.

65. Peters, J.B.; de Leer, W.B.; de Galan, L. Chlorination of cyanoethanoic acid in aqueous medium. *Environ. Sci. Technol.* **1990**, 24(1), 81-86.

66. Bull, R.J.; Reckhow, D.A.; Rotello, V.; Bull, O.M.; Kim, J. *Use of toxicological and chemical models to prioritize DBP research*; AWWA Research Foundation: 2006.

67. Reckhow, D.A.; Hua, G.; Kim, J.; Hatcher, P.G.; Caccamise, S.A.L.; Sachdeva, R. *Characterization of total organic halogen produced during disinfection processes*. AWWA Research Foundation: 2008

68. Steven, A.A. *TOX (total organic halogen) is the non-specific parameter of the future?* EPA/600/D-84/169, US Environmental Protection Agency: Cincinnati, OH, 1984.

69. Korshin, G.V., Li, C-W and Benjamin, M.M. The decrease of UV absorbance as an indicator of TOX formation. *Water Res.* **1997**, 31(4), 946-949.

70. Bove, F.; Shim, Y.; Zeitz, P. Drinking water contaminants and adverse pregnancy outcomes: a review. *Environ. Health Perspect*. **2002**, 110, 61-74.

71. Waller, K.; Swan, S.H.; DeLorenze, G.; Hopkins, B. Trihalomethanes in drinking water and spontaneous abortion *Epidemiology*, **1998**, 9(2),134-140.

72. Savitz, D.A.; Andrews, K.W.; Pastore, L.M. Drinking water and pregnancy outcome in Central North Carolina: Source, amount, and trihalomethane. *Environ. Health Perspect*. **1995**, 103(6), 592-596.

73. King, W.D.; Dodds. L.; Allen, A.C. Relation between stillbirth and specific chlorination by-products in public water supplies. *Environ. Health Perspect.* **2000**,108, 883-886.

74. Savitz, D.A.; Singer, P.C.; Herring, A.H.; Hartmann, K.E.; Weinberg, H.S.; Makarushka, C. Exposure to drinking water disinfection by-products and pregnancy loss. *Am. J. Epidemiol.* **2006**, 164(11), 1043-1051.

75. Porter, C.K.; Putnam, S.D.; Hunting, K.L.; Riddle, M.R. The effect of trihalomethane and haloacetic acid exposure on fetal growth in a Maryland County. *Am. J. Epidemiol*. **2005**, 162(4), 334–344.

76. IARC (International *Agency for Research on Cancer). Monographs on the evaluation of carcinogenic risks to humans*, Volume 52, Lyon, France: 1999.

77. Villanueva, C.M.; Fernández, F.; Malats, N.; Grimalt, J.O.; Kogevinas, M. Meta-analysis of studies on individual consumption of chlorinated drinking water and bladder cancer. *J. Epidemiol. Community Health*. **2003**, 57, 166-173.

78. Richardson, S.D.; Simmons, J.E.; Rice, G. Disinfection byproducts: The next generation. *Environ. Sci. Technol.* **2002**, 36(9), 198A – 205A.

79. Iriarte, U.; Alvarez-Uriarte, J.I.; Lopez-Fonseca,·R.; Gonzalez-Velasco, J.R. Trihalomethane formation in ozonated and chlorinated surface water. *Environ. Chem. Lett,* **2003**, 1, 57- 61.

80..Commonwealth Australia. *Australian Greenhouse office Annual report 2003-2004,* Government of Australia: 2004

81. Gallard, H.; von Gunten, U. Chlorination of natural organic matter: kinetics of chlorination and of THM formation. *Water Res.* **2002**, 36, 65-74

82. Gang, D.D.; Segar, R.J.; Jr., Clevenger, T.E.; Banerji, S.K. Using chlorine demand to predict THM and HAA9 formation. *J. Am. Water Works Assoc*. **2002**, *94*(10), 76–85.

83. Gallard, H.; von Gunten, U. Chlorination of phenols: Kinetics and formation of chloroform. *Environ. Sci. Technol.* **2002**, 36, 884-890.

84. Norwood, D.L.; Johnson, J.D; Chrisman, R.F.; Hass, J.R.; Bobenrieth, M.J. Reactions of chlorine with selected aromatic models of aquatic humic material. *Environ. Sci. Technol.* **1980**, *14*(2), 187-189

85. Boyce, S.; Hornig, J. Reaction pathways of trihalomethane formation from the halogenation of dihydroxyaromatic model compounds for humic acid. *Environ. Sci. Technol.* **1983**, *17*, 202-211.

86. McClellan, J.N., Reckhow, D.A., Tobiason, J.E., Edzwald, J.K., Darrell, B. Smith, D.B. A *Comprehensive kinetic model for chlorine decay and chlorination byproduct formation*. In: Natural organic matter and disinfection byproducts, Barrett, S.E.; Stuart, W.; Krasner, S.W.; Amy G.L. (eds), ACS symposium series, Vol. 761: 2000, pp 223-246,

87. Reckhow, D.A.; Singer, P.C.; Malcom, R.L. Chlorination of humic materials: Byproduct formation and chemical interpretation. *Environ. Sci. Technol.* **1990**, *24*, 1655-1664.

88. WQRA (Water Quality Research Australia). Development of tools for improved disinfection control within distribution systems. Research report No 71: 2010

89. Consonery, J.J.; Luardi, P.J.; Kopansky, R.; Manning, R.L. *Total organic carbon: A reliable indicator of TTHM and HAA5 formation?* American Water Works Association, WQTC conference: 2004

90. Lekkas, T.D.; Nikolaou, A.D. Development of predictive models for the formation of trihalomethanes and haloacetic acids during chlorination of bromide-rich water. *Wat. Qual. Res. J. Can*. **2004**, 39, 149-159.

91. Obolensky, A.; Singer, P.C. Development and interpretation of disinfection byproduct formation models using the information collection rule database. *Environ. Sci. Technol.* **2008**, 42, 5654-5660.

92. Rodriguez, M.J.; Milot, J.; Srodes, J.B. Predicting trihalomethane formation in chlorinated waters using multivariate regression and neural networks. *J. Water Supply Res. T*. **2003**, 52(3), 199-215.

93. Shimazu, H.; Kouch, M.; Yonekura, Y.; Kumano, H.; Hashiwata, K.; Hirota, T.; Ozaki, N.; Fukushima, H. Developing a model for disinfection by-products based on multiple regression analysis in a water distribution system. *J. Water Supply Res. T*. **2005**, *54*(4), 225-237

94. Golfinopoulos, S.; Arhonditsis, G. Multiple regression models: A methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. *Chemosphere.* **2002**, 47, 1007-1018.

95. Baxter, C.W.; Smith, D.W.; Stanley, S.J. A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. *J. Environ. Eng. Sci*. **2004**, *3*, S45–S58.

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

97. Perdue, E.M.; Ritchie, J.D. *Dissolved organic matter in freshwater*. In: Surface and groundwater, weathering, and soils, volume 5; Drever, J.I. (ed.), Elsevier Inc.: San Diego, CA: 2005, pp 273-318.

98. ECB (European Chemical Bureau). *ECB News Letter*, Institute of Health and Consumer Protection: 2002, issue # 2

99. Liang, C.; Gallagher, D.A. QSPR prediction of vapour pressure from solely theorectically-derived descriptors*. J. Chem. Inf. Comput. Sci.* **1997**, 55(4), 321-324

100. Cronin, M.T.D. The current status and future applicability of quantitative structure-activity relationships (QSARs) in predicting toxicity. *Altern.Lab. Anim*. **2002**, 30, 81-84,

101. Liu, D.; Thomson, K.; Kaiser, K.L.E. Quantitative structure-toxicity relationship of halogenated phenols on bacteria. *Bull. Environ. Contam. Toxicol.* **1982**, 29, 130-136.

102. Toropov, A.A.; Pablo Duchowicz, P.; Castro, E.A. Structure-toxicity relationships for aliphatic compounds based on correlation weighting of local graph invariants. *Int. J. Mol. Sci.* **2003**, 4, 272-283.

103. Siraki, A.G.; Chan, T.S.; O'Brien, P.J.O. Application of quantitative structure-toxicity relationships for the comparison of p-benzoquinone cytotoxicity in primary cultured rat hepatocytes versus PC12 cells. *Toxicol. Sci*. **2004**, 81(1), 148-159.

104. Hammett, L.P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc*. **1937**, 59 (1), 96–103

105. Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **1969**, 2 (8), 232–239.

106. Carrasco-Velar, R., Padron, J.A.; Galvez, J. Definition of a novel atomic index for QSAR: the refractotopological state. *J. Pharm. Pharm. Sci.* **2004**, 7(1), 19-26

107. WHO (World Health Organization). *Guidelines for drinking- water quality,* 2nd ed., Volume I-Recommendations. WHO, Geneva: 1993.

108. WHO (World Health Organization). *Guidelines for drinking water quality,* 2nd ed., Volume 2-Health Criteria and other supporting information. WHO, Geneva: 1996.

109. Puzyn,T.; Falandysz, J. QSPR modeling of partition coefficients and Henry's law constants for 75 chloronaphthalene congeners by means of six chemometric approaches-A comparative study. *J. Phys. Chem*. **2007**, 36**,** 203-214

110. Tropsha, A.; Gramtica, P.; Gomba, V.K. The Importance of being earnest: Validation is absolute essential for successful application and interpretation of QSPR model. *QSAR Comb. Sci*. **2003**, *23*(1): 69-77.

111. Mon, J., Flury, M.; Harsh, J.B. A quantitative structure activity relationship (QSAR) analysis of triarylmethane dye tracers. *J. Hydrol.* **2006**, 336, 84-97

112. Schwarzenbach, R.P.; Gschwend, P.M.; Imboden, D.M. *Environmental Organic Chemistry*, 2nd, edn, John Wiley & Sons Inc., New Jersey: 2003.

113. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSPR/QSAR. *Chem. Rev*. **1996**, *96*(3), 1027-1044.

114. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev*. **1995**, 279-287

115. Xu, J.; Liang, H.; Chen, B.; Xu, W.; Shen, X.; Liu, H. Linear and nonlinear QSPR to predict refractive indices of polymers from cyclic dimmer structures. *Chemometr. Intell. Lab. Syst*. **2008**, 92, 152-156

116. Wang, Z.W.; Li, G.Z.; Mu, J.H.; Zhang, X.Y.; Lou, A.J. Quantitative structure-property relationship on prediction of surface tension of nonionic surfactants. *Chin. Chem. Lett.* **2002**, 13(4), 363-366

117. Kahn, I.; Fara, D.; Karelson, M.; Maran, U.; Andersson, P.L. QSPR treatment of the soil sorption coefficients of organic pollutants. *J. Chem. Inf. Model.* **2005**, *45* (1), 94-105

118. Ghasemi, J.; Saaidpour, S. QSPR prediction of aqueous solubility of drug-like organic compounds. *Chem. Pharm. Bull.* **2007**, 55(4), 669-674.

119. Katritzky, A.R.; Karelson, M.; Lobanov, V.S. QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure Appl. Chem.* **1997**, 69(2), 245-248

120. Katritzky, A.R.; Slavov, S.H.; Dobchev, D.D.; Karelson, M. QSPR modeling of UV absorption intensities. *J. Comput.-Aided Mol. Des.* **2007**, 21(7), 371-377,

121. Damborsky, J.; Shultz, T.W. Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. *Chemosphere.* **1997**, 34(2), 429-446

122. Croni, M T.; Bowers, G S.; Sinks, G D.; Schultz, T W**.** Structure-toxicity relationships for aliphatic compounds encompassing a variety of mechanisms of toxic action to *Vibrio fischeri*. *SAR QSAR Environ Res*. **2000**; 11(3-4): 301-12

123. Eroglu, E.; Palaz, S.; Oltulu, O.; Turkmen, H.; Ozaydın, C. Comparative QSTR study using semi-empirical and first principle methods based descriptors for acute toxicity of diverse organic compounds to the Fathead Minnow (2008). *Int. J. Mol. Sci.* **2007**, *8*, 1265-1283

124. Ferydoun, A.; Ali, R.A.; Najmeh, M. Study on QSPR method for theoretical calculation of heat of formation for some organic compounds. *Afr. J. Pure Appl. Chem.* **2008**, 2(1), 6-9

125. Vatani, A.; Mehrpooya, M.; Gharagheizi , F. Prediction of standard enthalpy of formation by a QSPR model. *Int. J. Mol. Sci.* **2007**, *8*, 407-432

126. Nikolič, S.; Mililcvič, A.; Trinajstie, N. QSPR study of polarographic half-wave reduction potentials of benzenoid hydrocarbons. *Croat. Chem. Acta.* **2006**, 79(1), 155-159.

127. Gobraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graph. Model*. **2002**, *20*(4), 269-276

128. Jorgensen, W.J. QSAR/QSPR and proprietary data. *J. Chem. Inf. Model.* **2006**, 46(3), 937

129. .Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry*. **2005**, 76, 319-347

130. Cabaniss, S.E., Madey, G., Leff, L., Maurice, P.A., Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part II. Molecular property distributions. *Biogeochemistry*. **2007**, 86, 269-286.

CHAPTER 2

STATISTICAL METHODOLOGY

2.1. Data Sources

The data for chlorine demand ($HOCl_{dem}$) and disinfection byproducts (DBPs) for model compounds were obtained from research papers published between 1978 and 2010 (Table 2.1). The data are comprised of small aromatic and aliphatic compounds with carboxyl, amine, alcohol, phenol, ether and other functional groups and some representative compounds in Figure 2.1.

Table 2.1. Data sources for model compounds and reaction conditions

Sources	Chlorine source	Chlorine dose	pH	Time, h	Temp
Boyce and Hornig (*1*)	Cl_2	10-10^2 mol/mol	7	24	10 ^oC
Norwood et al. (*2*)	HOCl	1.5-2.0 mol/mol-C	7	0.3-4	25 ^oC
de Laat et al. (*3*)	Cl_2	2 mol/mol	7	15	20 ^oC
Boyce and Hornig (*4*)	Cl_2	10-10^2 mol/mol	7	24	10 ^oC
Hureiki et al. (*5*)	Cl_2	8-20 mol/mol	8	72	20 ^oC
Gallard and von Gunten (*6*)	NaOCl	90 mol/2-6 mol	8	20	23 ^oC
Bull et al. (*7*)	Cl_2	20 mg/L	7	48	20 ^oC
Dickenson et al. (*8*)	Cl_2	4-7 mg/mL	8	24	22 ^oC
Bond et al. (*9*)	Cl_2	35 mol/mol	7	24	20 ^oC
Hong et al. (*10*)	NaOCl	10 mg /mg-C	7	96	20 ^oC
Larson and Rockwell (*11*)	NaOCl	10 mol/mol	7-8	0.33-24	25 ^oC

The choice of using a mixture of structurally diverse molecules is contrary to the norm in quantitative structure-property relationship (QSPR) or quantitative structure-activity relationship (QSAR), which is using molecules with similar

structures. The water samples usually contain a mixture of simple aliphatic and aromatic molecules derived from degradation of large molecules from dead animal and plants or derived from anthropogenic sources (*12*). Simple phenolic molecules have been detected in surface water, ground water and tap water (*13-17*) and occurrences of amino acids and organic nitrogenous compound have been reported in lake and sea waters (*18-20*).

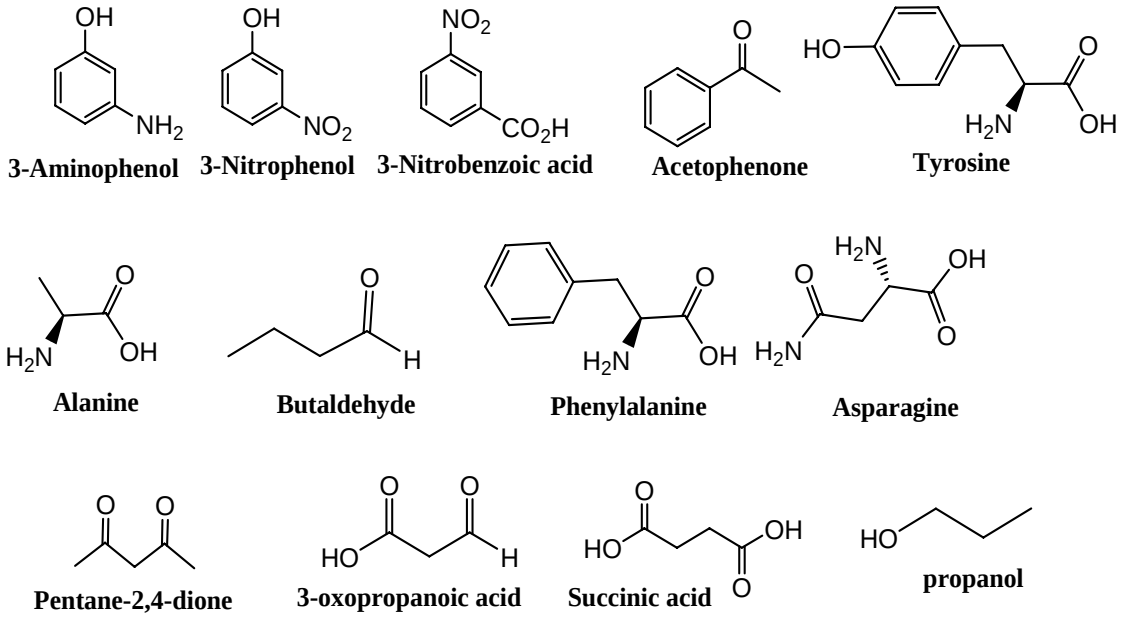


Figure 2.1. Examples of model precursor compounds used in this work

2.2. Descriptors generation

There are different molecular descriptors that can be used in QSAR/QSPR modeling which include constitutional descriptors (e.g., MW, atom counts), electrostatic descriptors (e.g., partial charges, polarity indices), geometric descriptors (e.g., molecular volume, solvent accessible surface area), quantum chemical descriptors (e.g., highest occupied molecular orbital, dipole moments) and topological descriptors (e.g., Wiener index, randic index) (*21,22*). The

advantages of constitutional descriptors are that they are easy to compute and do not require expensive software packages whereas disadvantages are that they cannot discriminate position isomers (e.g. catechol and resorcinol) and it becomes extremely laborious when you have a very large set of compounds and cannot explain effects of electronic, steric or geometric on reactivity of a molecule. In this work constitutional descriptors were used and were limited to those compatible with the AlphaStep model for NOM reactivity (*23*), including atom counts (the number of atoms of each element), functional group counts (the number of each functional group, including the number of aromatic rings), and variables which can be calculated from those (for example, H:C ratio, O:C ratio and number of phenol groups per ring). Two types of composite descriptors require explanation: the ring activation index (RAI) and the carbonyl index (CI).

The presence of electron donating substituents such as hydroxyl (OH) or amino group (NH_2) will activate the ring because they are strong electron donating groups. If any of the strong donors is present with a weak electron donating group (e.g., methoxy or ethoxy), strong donors will be directing the electrophilic substation (*24*). But if the weak electron donors are present with electron with drawing groups (e.g., cyano, carboxylic), the weak donor will direct the substitution (*24*). However, the increase in the number of strong electron donors in aromatic ring may not always increase electrophilic substitution because the relative position of the groups matters most (*24*). Thus, the RAI descriptor is motivated by the observation that aromatic rings with only electron withdrawing substituents showed much less $HOCl_{dem}$ and DBPs formation than

compounds with a single strong electron-donating substituent (*3*). Those aromatic rings with multiple electron donating group (not meta to each other) show intermediate chlorine demand. For example, benzoic acid and nitrobenzene consume < 0.5 mole HOCl per mole of substrate, phenol consumes 9.8 mole/mole, but 1,2-hydroxy benzenes and 1,4-dihydroxy benzenes consume $\leq$ 6 mole/mole (*3,6*). In addition, reactivity of aromatic molecules is influenced not only by the number of strong electron donating groups (OH and NH_2) in the molecule but also their relative position from each other. A 1,3-disubstitution (e.g., 1,3-dihydroxybenze) is more reactive than 1,2-disubstitution (e.g. 2-aminophenol) or 1,4-disubstitution (e.g., 1,4-dihydroxybenzene) due to cooperative effects in the former and antagonistic effect in the latter (*6,24*). However, it was also observed from the chlorine demand data that 2-hydroxy (or 2-amino) benzoic acid consumed less chlorine than corresponding meta and para isomers. The difference in reactivity could be attributed to higher stability of 2-hydroxy(2-amino) benzoic acid over the meta and para isomers due to hydrogen bonding effects (*25,26,27*). Thus, RAI descriptors were devised based upon the ratio of strong electron donating groups (-OH and -NH_2) to the number of aromatic rings (ED:AR) in a given molecule. RAI values range between 0 and 1 for each molecule (for molecules with several rings, e.g. tannic acid, a single average RAI is used). If ED:AR was $\leq$ 1, the RAI value was set equal to that ED:AR (except for ortho-carboxylic acids, which were assigned RAI = 0.75). If ED:AR was 2 or 3, the RAI was set equal to a lower value; the values which gave the lowest residuals were RAI = 0.6 for ED:AR = 2 and RAI = 0.5 for ED:AR = 3.

If -OH and -NH_2 groups were ortho or para to each other, as in 4-aminophenol or 2-aminophenol, the RAI index was 0.3 since two strong activators in such positions have antagonistic effects. The RAI was assigned 0.1 if the molecule had only alkoxy groups as ring activators to the ring (ED:AR = 0). If alkoxy groups were present along with OH or NH_2, as in 3,5-dimethoxyphenol, only the effects of strong activators affects the RAI. The RAI for molecules without -OH, -NH_2 or alkoxy groups on the ring was assigned a value of zero.

The carbonyl index, CI, is motivated by the observations that carbonyl compounds undergo substitution reaction through keto-enol tautomerization (*24,28,29*), and that β-dicarbonyl compounds consume more HOCl than other dicarbonyls (*8*). Carbonyl index (CI) relates to the lower pKa of hydrogen in a C-H bond located between two carbonyls (e.g., β-diketones) relative to other C-H bonds adjacent to a single carbonyl (*24,30*). The hydrogen can easily be abstracted by a base in solution to form keto-enol (or keto-enolate) tautomers and the enol form is the one that contributes to higher halogen substitution reaction (*30*). This concept was extended to carbons located between two hidden carbonyls (phenols) as in resorcinol. The index adds contributions from C atoms around carbonyl groups. The α-carbon between two keto groups was assigned a value of 2, the α-carbon between a keto-group and either an ester or acid group was assigned a value of 1.5. A carbon adjacent to an α-dicarbonyl is assigned a value of 1. The two carbons between γ-dicarbonyls (e.g. 4-oxoheptanedioic acid) are assigned a total value of 0.5, as is the α-carbon in ketones (e.g., acetophenone, acetone). Contributions from all carbonyl groups in a molecule

are used to calculate the CI. The list of constitutional and other derived descriptors is given in Table S2.1 and Table 2.2 gives the list of significant descriptors selected by descriptor selection process in this work..

Table 2.2. A list of significant descriptors and abbreviations

Descriptors	Abbreviation	Descriptors	Abbreviation
# Alkoxy groups attached to the aromatic ring without NH_2 and OH	ArORact	# Alkoxy groups attached to the aromatic ring without NH_2 and OH	ArORnoact
# Oxygen to carbon ratio	O:C	Ring activation index	RAI
Square root of # heteroatoms	sqrtHeA	# One three activated aromatic carbon	OTactC
Square root of #phenols	sqrtArOH	Difference of ArED:C and CORH:C	EDCORH
# sum of ArOR per carbon	ArOR:C	# Aliphatic C bonded reduced nitrogen (NR_2)	ACN
# Phenols	ArOH	# ArED per carbon	ArED:C
# Aliphatic sulfur	AS	Carbonyl index	CI
Log of hydrogen to carbon ratio	logH:C	# Phenol per carbon	ArOH:C
# Hydrogen to carbon ratio	H:C	Square root of ring index	sqrtRAI

2.3. Data Splitting and descriptor selection

The data collected for chlorine demand or DBPs formation was divided into training and external validation data sets using random data splitting or stratified random data splitting approaches (*31*) or pseudo-stratified random data splitting where necessary. The external data set is the list of compounds with similar functional groups but different chemical structure than members of the training data set. These data are not involved in model calibration whatsoever and therefore are used to evaluate the performance of the model. The training data is the list of compounds with diverse structure and functional groups used to calibrate the model and for internal validation of the model. In case of limited data size the whole training data set may be used to calibrate the model but

where there is enough data, the training data can be split into calibration and cross validation data sets.

Random data splitting allows each individual compound an equal chance of being selected into training and external validation data sets. However, the selection process that is based on random numbers does not take into account the structural attributes of the molecules. As a result you may end up with data sets that are not heterogeneous in terms of structural diversity which is important in QSPR modeling.

Alternatively, stratified random data splitting may be used, particularly when there are duplicate data points. In this case, each individual point is given an equal chance of being in training and external data sets but the process is repeated such that no compound is present in both training and external sets. Stratified data splitting is also used in splitting training data into calibration data sets and cross validation using the Leave-Many-Out approach. Splitting of training data may be repeated 2 to 10 times, each time leaving out 20-30% of data consistently in order to ensure that any data point should appear at least once in cross validation data. Pseudo stratified randomization was used where the entire data set was too small to be split into calibration, Leave-Many-Out (LMO) cross validation and external validation data sets and no duplicate measurements were available. This process ensures that the data splitting produces calibration and external data sets that are as heterogeneous as possible. In this research work stratified random data splitting was used to split the training data into calibration and cross validation five folds for $HOCl_{dem}$ and

TCM formation. Due to limited data size for TOX and TCAA formation, pseudo stratified random data splitting was used instead.

The simplest quantitative structure-property relationship (QSPR), both conceptually and statistically, expresses $HOCl_{dem}$ or DBP formation for a given molecule as the weighted sum of a set of M descriptor variables x_j (Equation 2.1).

$$HOCl_{dem}\,(or\ DBP) = \sum_{j=1}^{M} \beta_j x_j + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(Eq. 2.1)}$$

where β_j is the linear coefficient for the j^{th} descriptor variable, x_j and ε is the standard error of regression of the model where each descriptor represents a sub-structure of the molecule. The intercept is set to zero because pure water is expected to have zero chlorine demand or DBPs formation.

Selection of significant descriptors was done by multiple linear regression using Minitab® statistical software, StatGuide™ version 15 (*32*). Regressions began with a complete list of all potential descriptors and training model compounds (minimum 5 non-zero values for each descriptor). Successive elimination of the least significant descriptor was based on the p-value criterion. If the p-value for a given descriptor coefficient, β_j, exceeds 0.05, the descriptor was eliminated and multiple linear regression analysis was repeated on the remaining descriptors until all coefficients of descriptors were significant (that is, $p < 0.05$). Final significant descriptors obtained were tested for multi-colinearity by correlation analysis. If pair-wise correlations of descriptors had $r < 0.7$ ($r^2 = 0.49$), this indicates that they were sufficiently independent to be used as a group (*33*). These descriptors were used in all subsequent model calibration and validation.

2.4. Model calibration and validation

Once the significant descriptors were identified, the training data set was divided into calibration and cross validation data sets by Leave-Many-Out (LMO) approach in case there was sufficient data. In case of limited data, training data was used for model calibration and Leave-One-Out (LOO) was used for cross validation. Model calibration was performed using the Analysis ToolPak in MS Excel for Windows XP. Note that most commercial software and MS Excel versions on the market have errors in the default equations for calculation of R^2 and F-statistic for regression through origin (*34*). Preliminary multiple linear regression of the chlorine demand on a set of descriptors through origin using commercial statistical software packages (Minitab 15 and SPSS 15) and MS excels (MS Excel XP, MS Excel 2003 and MS Excel 2007) showed that only MS Excel for Windows XP gave correct calculations of R^2 and F-statistic.

In Leave-Many-Out cross validation, the training data were split into a calibration data set (70-80%) and a cross validation data set (20-30%) five times using stratified data splitting so that each compound is used at least once in cross validation (*31*). This resulted in five subsets of calibration and cross validation data. Multiple linear regression was then performed on each of the five calibration subsets using the selected significant descriptors to obtain five QSPR equations. Each of the five QSPR equations was used to predict chlorine demand or DBP formation of the respective cross validation subsets.

The final QSPR model was obtained from the five QSPR equations by averaging coefficients of descriptors and standard error of descriptors in the five equations. This equation was then used to predict $HOCl_{dem}$ or DBP formation of each compound in the five cross validation subsets from which averaged statistics of LMO cross validation (LMO_{CV}) were obtained. The average statistics of cross validation should be comparable to those obtained from individual cross validation sets using each QSPR equation above if there is little bias in data splitting. The equation was also used to predict $HOCl_{dem}$ or DBPs formation for the entire training data set and its statistics of cross validation were compared to the average statistics of cross validation obtained previously. The bias in data splitting using the LMO approach was cross checked by comparing the regression fits and cross validation statistics to that of LOO cross validation (LOO_{CV}).

In Leave-One-Out cross validation one compound out of 'N' compounds is set aside for validation and 'N-1' compounds are used for model calibration (*31*). The N-1 compounds are used to produce a multiple linear regression equation. The equation is then used to predict HOCl demand/DBP formation of the omitted compound. The process is repeated until each compound is used for validation. Therefore there will be 'N' equations and 'N' validated compounds in total (*31*). Then, average of statistics of fit, coefficients of descriptors and standard error of descriptors are calculated. This was followed by internal validation using Leave-One-Out cross validation (LMO_{CV}) where there were enough data and Leave-

One-Out cross validation (LOO_{CV}) was used where there were not enough training data to split into two.

External validation evaluates model stability on independent data that are not used in the model calibration process (*31*). Compounds used in external validation should have similar structures to compounds used in model calibration. The model has predictive power if statistics of external validation meet criteria of predictive power and are comparable to those of LMO_{CV}.

2.5. Model predictive power evaluation

The predictive ability of the QSPR model was assessed using the coefficient of determination (R_c^2) of regression, coefficient determination of cross validation (q^2), mean bias deviation (MBD) and root mean square error (RMSE) for cross validation and external validation. The R_c^2 can easily be calculated using the Equation 2.2.

$$R_c^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad \text{..... (Eq. 2.2)}$$

Where, y_i and $\hat{y}_i$ are experimental and predicted values respectively and $\bar{y}$ is the mean of experimental values.

q^2 is the square of correlation coefficient of predicted against experimental data for validation and was computed for both cross validation (q^2_{cv}) and external validation (q^2_{ext}). The q^2_{cv} and q^2_{ext} were calculated using Equations 2.3 and 2.4 respectively.

$$q^2_{cv} = 1 - \frac{\sum_{i=1}^{training}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{training}(y_i - \bar{y})^2} \quad \text{..... (Eq. 2.3)}$$

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{tr})^2} \quad \text{..... (Eq. 2.4)}$$

Where y_i and $\hat{y}_i$ are experimental and predicted HOCl demands or DBP formations (from external or cross validation data set) respectively; $\bar{y}_i$ is average of entire data in calibration data set, $\bar{y}_{tr}$ is the average of the experimental $HOCl_{dem}$ or DBP in the entire training data (*35,36*).

Since q^2 has some limitations an indicator of predictive power, additional tests of QSPR model predictive power recommended in the literature (*33,35,36*) were performed. The slope (k_i) and R_i^2 were obtained from regression line in a plot of predicted HOCl/DBP versus observed HOCl/DBP whereas slope (k_o) and R_o^2 were obtained by forcing the regression line through the origin (Figure 2.2).

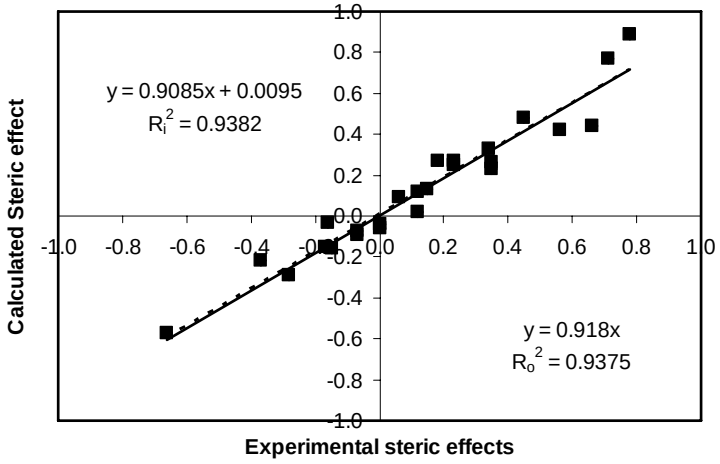


Figure 2.2. Normal regression of substituent steric effects in benzoic acid with q^2 = 0.99 and R_t = 0.001. Data source: Karelson (*37*)

Similarly one may obtain k_i', $R_i^{2'}$ and k_o', $R_o^{2'}$ from inverse regression lines with y-intercept and through origin, that is, a plot of observed HOCl/DBP versus predicted HOCl/DBP respectively (Figure 2.3).

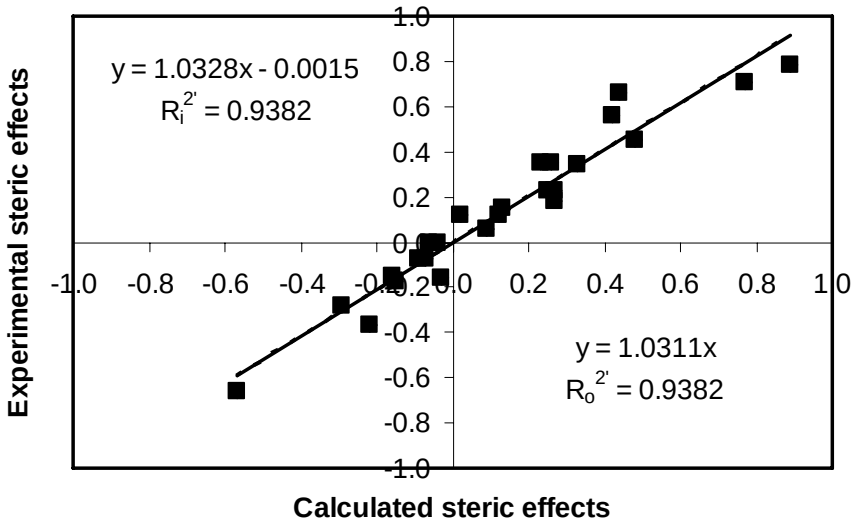


Figure 2.3. Inverse regression of substituent steric effects in benzoic acid with q^2 = 0.99 and R_t = 0.00. Data source: Karelson (*37*)

The model is said to have high predictive power and to be robust if it meets the following criteria for QSPR/QSAR predictive power: $R_c^2 > 0.6$ and $q^2 > 0.5$. The ratio $(R_i^2-R_o^2)/R_i^2$ or $(R_i^{2'}/R_o^{2'})/R_i^{2'}$ which, in this work, will be denoted by R_t and R_t' respectively should be less than 0.1. Thus, $R_t < 0.1$ and $0.85 \leq k \leq 1.15$ ($k = k_i$ & k_o) or $R_t' < 0.1$ and $0.85 \leq k' \leq 1.15$ ($k' = k_i'$ & k_o') or both. A QSPR that is close to an ideal model should have slopes k or k' close to 1 for both normal and inverse regression of predicted and observed data.

The root mean square error (RMSE) and model bias deviation (MBD) were computed using Equations 2.5 and 2.6.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad \text{..... (Eq. 2.5)}$$

$$MBD = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{\sum_{i=1}^{N} y_i} \times 100\% \quad \text{(Eq. 2.6)}$$

where y_i and $\hat{y}_i$ are experimental and predicted $HOCl_{dem}$ or DBP formation and N is the total number of observations. The RMSE of internal and external validation is expected to be lower and comparable to the residual standard deviation of model calibration with high predictive power. The MBD is issued is qualitative diagnostic predictive power. A model bias deviation of zero indicates that model has no net prediction bias whereas a negative MBD indicates that the model predicts lower that experimental value and positive MBD indicates that model predicts higher than experimental data. The magnitude of MBD does not necessarily indicate extent of bias because one compound that is either over-predicted or under-predicted may drive total residuals up or down respectively (Eq. 2.6).

The predictive power of the model was also evaluated using y-permutation, in which the chlorine demand or DBP formation values were randomly permuted while the descriptor values were fixed (*33,38*). This analysis is used to determine if the relationship between dependent variable and independent variables was by chance. The MLR r^2 and q^2 values obtained for 60 repetitions were compared with those using non-permuted $HOCl_{dem}$ or DBP formation. A model is robust if the average permuted $R^2 < 0.3$ and $q^2 < 0.05$ (*33*).

The model predictive power can also be assessed visually using the plot of predicted data versus observed data and vice versa. A model with high

predictive power is expected to have data points scattered closely around the 1:1 ideal model line (Figure 2.4).

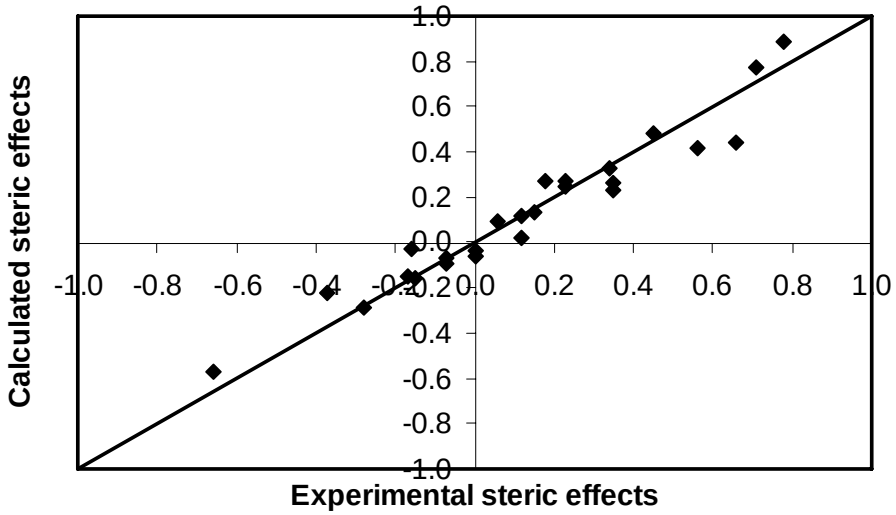


Figure 2.4. Deviation of the calculated substituents steric effects in benzoic acid from ideal model. Data source: Karelson (*37*)

The graphs may also have either confidence interval or predictive interval marginal lines or both. The plot of standardized residuals versus predicted data (Figure 2.5) is useful for testing constant variance of residuals or independence of variables and it may also show data points that are out of the ordinary. For a model derived at 95% confidence level, a data point with standardized residuals greater than +2.5 or less than -2.5 may be an outlier (*33,39,41*). A data point in training data becomes a suspected outlier if the standardized residual is outside or near the ±2.5 boundary and leverage (h) is greater than the cutoff leverage (h*), h>h*. The compound may be considered influential if the observed value have impacts on magnitude of model fits and coefficients (*42,43*).

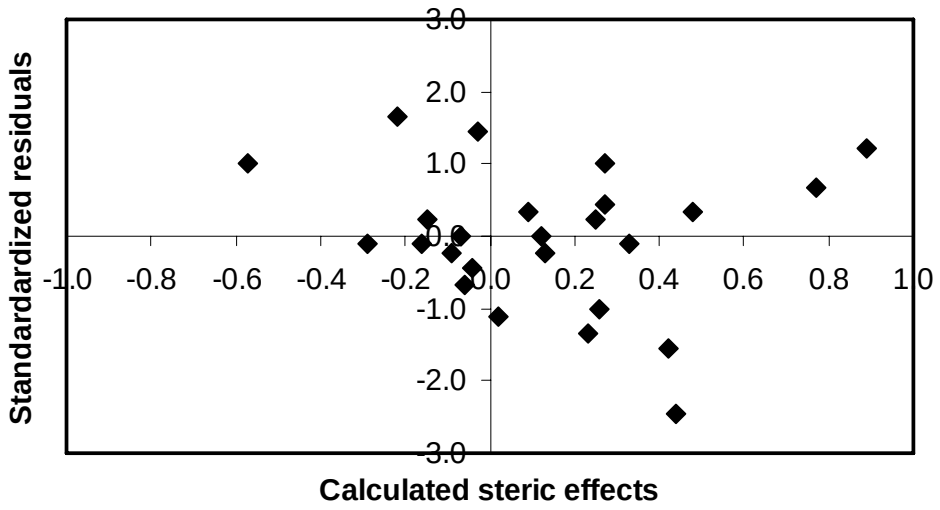


Figure 2.5. Residual plot of substituent steric effects in benzoic acid. Data source: Karelson (*37*)

The model with high predictive power will have the data points along the zero SDR line without a distinct pattern. Figure 2.5 shows there is some linear pattern in data points which is an indication that errors are not random and there is no constancy in variance though standardized residuals had an average around 1 (Mean = 0.97). The linear pattern of data points in Figure 2.5 suggests that the model had inadequate predictive power.

2.6. Model Applicability domain

The applicability domain is defined as the physico-chemical and structural space on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds (*33,36*). Model applicability domain is evaluated by calculating standardized residuals of cross validation and leverage for training and external data sets. Leverage is the potential of an observation to affect the model fit and influence refers to the actual effect the observation has on model fit due to having extreme descriptor

values (*30*). Standardized residuals were calculated by taking ratio of residual (predicted value-experimental value) to root mean square error (Equation 2.7). Leverage, h, was calculated using the Equation 2.8 and its magnitude is independent of the measured parameter (*31,43*).

$$SDR = \frac{(\hat{y}_i - y_i)}{RMSE} \quad \text{(Eq. 2.7)}$$

where, y_i and $\hat{y}_i$ are experimental and calculated values respectively

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, 2, \ldots, n) \quad \text{(Eq. 2.8)}$$

where,

X = N*k matrix of k model descriptor values of N training dataset

X^T = transpose of matrix vector X

x_i = row vector of compound x_i

x_i^T = transpose of row vector of compound x_i

h* = is a warning leverage (Fixed at 3k'/N)

The applicability domain can provide three important pieces of information. First it can show compounds that are outliers in terms of model fit if standardized residuals of cross validation are out of the cutoff value of ±2.5 (*40*). However, a data point in training data set that is an outlier may not necessarily be influential if its h < h* (*42*). Secondly, a data point from the training data set may be influential in determination of model parameters if its h > h* and it will be considered most influential if the data point is an outlier falling outside the applicability domain and its h is far from h* (*42*). Thirdly, predicted values of the external data set that fall within the applicability domain are considered reliable. Those predicted values

falling outside applicability domain are predicted due to over-extrapolation of the model (i.e., h> h*) and the results may be not be reliable. The prediction results of external data points with h close to h* may be accepted with cautions. Further information on use of Williams plot to determine model calibration and prediction outliers and influential data points can be found in the literature (*33,39,40,44,45*).

2.7. Other statistics for detection of outliers and influential data

There were other statistics that were used to detect data points that were influential or both outlier and influential in terms of model calibration include Cook's distance (D_i), DiFference in FiT Standardized (DFFITS) and Difference in BETA Standardized (DFBETAS) (*42,43*).

Cook's distance (D_i) of the i^{th} observation in a training data set is an overall measure of impact on regression model coefficients upon deletion of i^{th} observation in a training data set (*42*). Cook's distance (D_i) can be calculated manually using Equations 2.9 or 2.10. Nonetheless, most statistical software can calculate Cook's distance automatically if that option is selected when performing regression analysis. Data points with large residuals and/or high leverage may also have high Cook's distance and therefore may distort the outcome and accuracy of a regression. A simple operational guideline is that an i^{th} observation with $D_i > 1$ is considered an outlier. Such points require further examination using DFFITS and DFBETAS statistics in order to determine the impacts of the data points on model fits and coefficients respectively (*42*).

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\,MSE} \quad \text{..... (Eq. 2.9)}$$

$$D_i = \frac{e_i^{\,2}}{p\,MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \quad \text{..... (Eq. 2.10)}$$

where,

$\hat{y}_i$ is the prediction from the full regression model for observation j;

$\hat{y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation *i* has been omitted;

h_{ii} is the i[th] diagonal element of the hat matrix;

e_i is the crude residual (i.e., the difference between the observed value and the value fitted by the proposed model);

MSE is the mean square error of the regression model;

p is the number of fitted parameters in the model

DiFference in FiTs, Standardized (DFFITS) of an i[th] observation is a measure of a mean response variable obtained using model derived from training data set with the i[th] observation omitted (*42,43*). DFFITS can be calculated manually using Equation 2.11.

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{(ji)}}{s_{(i)}\sqrt{h_{ii}}} \quad \text{..... (Eq. 2.11)}$$

where h_{ii}, $\hat{y}_i$ and $\hat{y}_{j(i)}$ have the same meaning as described above, $s_{(i)}$ is the standard error estimated without the i[th] observation, and h_{ii} is the leverage for the point. Points with $DFFITS > 2(\frac{k'}{n})^{0.5}$ are potentially influential. Here k' is the number of predictors plus 1 in the model and n is total number of observations. The DFFITS of 2 is usually used as an operational cutoff value whereas absolute value for size adjusted DFFITS should not exceed 0.89 (*42*). Since the difference

in fit for an i^{th} observation may be due to influence of one or more independent variables, it important to evaluate the change in coefficients of each descriptor.

Difference in BETA, standardized (DFBETAS) is a measure of how much the coefficient, in standard deviation units, of an independent variable changes in regression model when an i^{th} observation is deleted from training data set (*42,43*). DFBETAS can be calculated manually using Equation 2.12 or 2.13 and DFBETAS exceeding $2n^{-0.5}$ is considered large (*43*). For such points, examine DFBETAS for those points with high DFFITS in order to determine the influential observations in the training data set. The operational cutoff of DFBETAS is 2 and absolute value for size adjusted DFBETAS should not exceed 0.63 (*42*).

$$DFBETAS = \frac{\sqrt{n} \times e_i}{s_{(i)}(n-1)} \quad \text{..... (Eq. 2.12)}$$

$$DFBETAS_{ij} = \frac{\beta_j - \beta_{j(i)}}{SE\beta_{j(i)}} \quad \text{..... (Eq. 2.13)}$$

where, n is data size, $s_{(i)}$ is standard deviation when i^{th} row has been deleted and $e_{(i)}$ is residual vector when i^{th} row has been deleted. β_j is the beta when the i^{th} observation is included and $\beta_{j(i)}$ is the beta when the i^{th} observation is excluded, $SE\beta_{j(i)}$ is standard error of beta when the i^{th} observation is excluded (*46*). Cook's distance, DFFITS and DFBETAS can be calculated using most of the commercial statistical software such as Minitab and Statistical Package for Social Science.

2.8. References

1. Boyce, S.D.; Hornig, J.F. Formation of chloroform from chlorination diketones and polyhydroxybenzenes in dilute aqueous solutions. In: Water chlorination: Environmental impacts and health, Jolly, J.L. (ed.), Ann Arbor Science Publishers, Ann Arbor, MI: 1979 pp131-140

2. Norwood, D.L.; Johnson, J.D; Chrisman, R.F.; Hass, J.R.; Bobenrieth, M.J. Reactions of chlorine with selected aromatic models of aquatic humic material. *Environ. Sci. Technol.* **1980**, 14(2), 187-189.

3. de Laat, J.; Merlet, N.; Doré, M. Chlorination of organic compounds: chlorine demand and reactivity in relationship to the trihalomethane formation. *Water Res.* **1982**, 16, 1437-1450

4. Boyce, S.; Hornig, J. Reaction pathways of trihalomethane formation from the halogenation of dihydroxyaromatic model compounds for humic acid. *Environ. Sci. Technol*. **1983**, 17, 202-211.

5. Hureiki, L.; Croué, J-P.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res.* **1994**, 28, 2521-2531.

6. Gallard, H.; von Gunten, U. Chlorination of phenols: Kinetics and formation of chloroform. *Environ. Sci. Technol*. **2002**, 36, 884-890.

7. Bull, R.J.; Reckhow, D.A.; Rotello, V.; Bull, O.M.; Kim, J. Use of toxicological and chemical models to prioritize DBP research; AWWA Research Foundation: 2006.

8. Dickenson, E.V.; Summers, S.; Croué, J-P.; Gallard, A. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol*. **2008**, 42, 3226-3233.

9. Bond T.; Henriet O.; Goslan E.H.; Parson S.A.; Jefferson B.. Disinfection byproducts and fractionation behavior of natural organic matter surrogates. *Environ. Sci. Technol.* **2009**, 43, 5982-5989

10. Hong, H.C.; Wong, M.H.; Liang, Y. Amino acids Precursors of trihalomethane and haloacetic acid formation during chlorination. *Arch. Environ. Toxicol*. **2009**, 56, 638-645.

11. Larson, R.A.; Rockwell, A.L. Chloroform and chlorophenol production by decarboxylation of natural acids during aqueous chlorination. *Environ. Sci. Technol*. **1979**, 13(3), 325-329

12. Moore, J.W.; Ramamoorthy, S. Organic chemicals in natural waters: Applied monitoring and impact assessment. Springer-verlag, New York: 1984.

13. Galceran, M.T.; Jairegui, O. Determination of phenols in sea water by liquid chromatography with electrochemical detection after enrichment by using solid-phase extraction cartridges and disks. *Anal. Chim. Acta.* **1995**, 304, 75- 84

14. Jairegui, O.; Galceran, M.T. Determination of phenols in water by on-line solid-phase disk extraction and liquid chromatography with electrochemical detection. *Anal. Chem. Acta.* **1997**, 340, 191-199.

15. Davì, M.L., Gnudi, F. Phenolic compounds in surface water. *Water Res.* **1999**, 33(14), 3213-3219

16. González-Toledo, E.; Prat, M.D.; Alpendurada, M.F. Solid-phase microextraction coupled to liquid chromatography for the analysis of phenolic compounds in water. *J. Chromat. A.* **2001**, 923(1-2), 45-52.

17. Ou J.; Hu, L.; Hu, L.; Li, X.; Zou, H. Determination of phenolic compounds in river water with on-line coupling bisphenol A imprinted monolithic precolumn with high performance liquid chromatography. *Talanta*. **2006**, 69, 1001-1006.

18. Daumas, R.A. Variations of particulate proteins and dissolved amino acids in coastal seawater, *Mar. Chem.* **1976**, 4(3), 225-242

19. Chinn, R.; Barrett, S.E. *Occurrence of amino acids in two drinking water sources.* In: Natural organic matter and disinfection byproducts: Characterization and control in drinking water, Barrett, S.E.; Krasner, S.W, Amy, G.L (eds), ACS symposium series, volume 17,: 2000, pp 96-108.

20. Dotson, A.; Westerhoff, P. Occurrence and removal of amino acids during drinking water treatment. *J. Am. Water Works Assoc.* **2009**, 101(9), 101-115.

21. Katritzky, A.R.; Lobanov, V.S. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev*. **1995**, 24, 279-287.

22. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSPR/QSAR. *Chem. Rev*. **1996**, 96(3), 1027-1044.

23. Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; and Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry*. **2005**, 76, 319-347.

24. Reusch, W. *Virtual textbook of organic chemistry*, 1999, (a 2008 revision). http://www.cem.msu.edu/~reusch/VirtualText/intro1.htm

25. Pinto, S.S.; Diogo, H.P.; Guedes, R.C.; Costa Cabral, B. J.; Minas da Piedade, M.E.; Martinho Simoes, J.A. Energetics of hydroxybenzoic acids

and of the corresponding carboxyphenoxyl radicals. Intramolecular hydrogen bonding in 2-hydroxybenzoic acid. *J. Phys. Chem. A.* **2005**, 109(42), 9700-9708.

26. Aarset, K.; Page, E. M.; Rice, D. Molecular structures of benzoic acid and 2-hydroxybenzoic acid, obtained by gas-phase electron diffraction and theoretical calculations. *J. Phys. Chem. A.* **2006**, 110(28), 9014-9019.

27. Stalin, T.; Rajendiran, N. Intramolecular charge transfer associated with hydrogen bonding effects on 2-aminobenzoic acid. J*. Photochem. Photobiol. A: Chem.* **2006**, 182(2), 137-150.

28. *Formation of halogenated organics by chlorination of water supplies*. USEPA (1975). National Service Center for Environmental Publication (NEPIS). http://nepis.epa.gov/EPA/html/pubs/pubtitleORD.htm

29. Larson, R.A.; Weber, E.J. Reaction *mechanisms in environmental organic chemistry*; Lewis Publishers: New York: 1994.

30. McMurry, J. *Organic chemistry*, 5th edn. Thomson Brooks/Cole, Pacific Grove, CA: 2000.

31. Herrell Jr., F.E. Regression modeling strategies with application to linear models, Logistic regression, and survival analysis; Springer-verlag: New York: 2001.

32. Minitab Inc. "Graphical data," Meet Minitab 15, p. 2-1 to 2-13: 2007. www.minitab.com

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

34. Eisenhauer, J.G. Regression through the origin. *Teaching Statistics*. **2003**, 25(3), 76-80.

35. Gobraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graph. Model*. **2002**, 20(4), 269-276

36. Tropsha, A.; Gramtica, P.; Gomba, V.K. The Importance of being earnest: Validation is absolute essential for successful application and interpretation of QSPR model. *QSAR Comb. Sci.* **2003**, 23(1), 69-77.

37. Karelson, M. *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience: New York: 2000.

38. Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR, *J. Chem. Inf. Model.* **2007**, 47, 2345-2357.

39. Rousseeuw, P.J.; van Zomeren, B.C. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* **1990**, 85(411), 633-639.

40. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, 26(5), 694-701.

41. Field, A. P. *Discovering statistics using SPSS*, 3rd edn, Sage Publications Ltd: 2009.

42. Freund, J.R.; Wilson, W.J; Sa, P. *Regression analysis: statistical modeling of a response variable*, 2nd edn, Academic Press: 2006.

43. Belsley, D.A.; Kuh, H.; Welsch, R.E. *Regression diagnostics. Identifying influential data and sources of colinearity*. Wiley Interscience, Hoboken, New Jersey: 1980.

44. Papa, E.; Villa, F.; Gramatica, P. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow). *J. Chem. Inf. Model*. **2005**, 45, 1256-1266.

45. Pan, Y.; Jiang J.; Wang, R.; Cao, H.; Cui, Y. A novel QSPR for prediction of lower flammability limits of organic compounds based on vector machine. *J. Hazard. Mater*. **2009**, 168, 962-969.

46. McLain, A. *Quantitative methods behavioral data I and II: Detecting influential data and collinearity*. Department of Statistics, Univesity of South Carolina .www.stat.sc.edu/~mclaina/psyc/4th%20lab%20notes%20710.pdf

CHAPTER 3

QSPR FOR PREDICTING CHLORINE DEMAND

Abstract

Conventional methods for predicting chlorine demand ($HOCl_{dem}$) due to dissolved organic matter (DOM) are based on bulk water quality parameters and ignore structural features of individual molecules that may better indicate reactivity towards the disinfectant. The Quantitative Structure-Property Relationship (QSPR) modeling approach can account for structural properties of individual molecules. Here we report a QSPR for $HOCl_{dem}$ based on eight constitutional descriptors. Model compounds with $HOCl_{dem}$ ranging from 0.1 to 13.4 mole chlorine (HOCl) per mole compound (Cp) were divided into a calibration and cross-validation data set (N = 159) and an external validation set (N = 42). The QSPR was calibrated using multiple linear regression in a 5-way Leave-Many-Out approach and has average R_c^2 = 0.86 and standard error of regression (SDE) = 1.24 mol-HOCl/mol-Cp and $p < 0.05$. LMO Cross validation has average q^2_{LMO} = 0.85 and the external validation has $q^2_{(Ext)}$ = 0.88, indicating a robust model. The leverage of 7 of 42 compounds in the external validation dataset exceeded the critical value, suggesting that these compounds may be over-extrapolated. However, root mean square error (RMSE) of prediction in the external validation was 1.17 mol-HOCl/mol-Cp, and all compounds were predicted with ±2.5 standardized residuals (SDR). Application of the QSPR to

model structures of NOM predicts $HOCl_{dem}$ comparable to reported measurements from natural water treatment.

3.1. Introduction

Freshwater contamination from microbial pathogens poses serious public health risks worldwide (*1,2,3*). Chlorination has been the most commonly used technology to eliminate microbes in drinking water (*1,4*). However, chlorine residue reacts with traces of dissolved organic matter (DOM) in water to produce disinfection byproducts (DBPs) mostly through electrophilic substitution reactions (*5*). Disinfection byproducts include trihalomethanes (THMs), haloacetic acids (HAAs), haloacetonitriles (HACN), halonitromethames (HNMs) and haloketones (6). Most DBPs are carcinogenic and tumorgenic to test animals (*7*) and are being regulated in most countries (*8,9,10*). The regulation is based on precautionary principle (*11*) for there is little evidence to link DBPs directly to reproductive problems and carcinogenicity in humans (*12,13*).

Since the first detection of DBPs in 1970's, the minimization of DBP production without compromising water quality has been a major challenge. A closely related problem is prediction of the amount of HOCl consumed, or chlorine demand ($HOCl_{dem}$). Quantitative prediction of $HOCl_{dem}$ for different water supplies has been undertaken in order to help optimize chlorine dosages while maintaining disinfection and minimizing DBP production.

Empirical models relate $HOCl_{dem}$ to bulk water quality parameters like pH, ultraviolet absorption and dissolved organic carbon (*7,8,14,16,19*). However, this approach treats DOM as a single**,** average entity while it is actually a mixture of

organic molecules with different chemical structures (*20*). The types, number and arrangement of functional groups in each molecule are expected to strongly influence reactivity towards chlorine during water treatment (*21,22,23*).

An alternative approach develops predictive models based on molecular structure and applies them to postulated DOM molecules. This approach is not a substitute for traditional experimental measurements, but can enable predictions of HOCl demand due to potential changes in the DOM-for example, changing pre-treatment methods or land use within the watershed. A predictive model has been developed which provides composition data for thousands of molecules in a DOM mixture (*24,25*). What is lacking is a rapid and quantitative method to predict HOCl demand from this molecular information.

Quantitative structure-property relationships (QSPRs) have been successfully used to predict physical properties of organic pollutants (*26*) and activities of pharmaceuticals (*27*). Common variables used in QSPR modeling include electrostatic (e.g., partial charge), geometric (e.g., molecular volume), quantum-chemical (e.g., dipole moment) and typological descriptors (e.g., Weiner index) modeling (*28*). Constitutional descriptors reflect only the chemical composition without any reference to geometry or electronic structure, and are attractive for work with large numbers of molecules because of their conceptual and computational simplicity (*29*).

Despite numerous applications of QSPRs in environmental studies, we are unaware of any attempts to predict $HOCl_{dem}$ using this approach. Here we use experimental data from the literature on HOCl consumption by small molecules to

develop, calibrate and validate a QSPR that predicts $HOCl_{dem}$ based solely on constitutional descriptors. This QSPR is used to predict $HOCl_{dem}$ by tannic acid and model DOM structures.

3.2. Methodology

3.2.1. Data collection

$HOCl_{dem}$ data for model compounds were obtained various sources (*19,20-23,30-34*) and the complete list of compounds and $HOCl_{dem}$ values is given in Tables S3.1, S3.2 and S3.3. There were 201 compounds with chlorine demand data that included both aromatic and aliphatic compounds with carboxyl, amine, alcohol, phenol, ether and other functional groups. The data were not acquired under consistent chlorination conditions; in particular, reaction times varied from 4 to 96 hours (Table 2.1). Since reaction with HOCl can require several days (*35*), the $HOCl_{dem}$ values in the shorter-time studies were adjusted by comparing the chlorine demand of compounds included in both longer- and shorter-time studies. Two studies could only be compared if one or more 'common' compounds were used in both. The ratio of $HOCl_{dem}$ at a shorter reaction time (s_i) to $HOCl_{dem}$ at longer reaction time (l_i) was calculated for each 'common' compound. If the average ratio s_i/l_i for the two studies was less than 0.85, only the shorter-time $HOCl_{dem}$ was adjusted using Equation 3.1. The adjusted chlorine demands for the 'common' compounds given in Tables S3.1 and S3.2 were used in calibration and validation along with those compounds which were not adjusted.

$$AdjHOCl_{dem} = HOCl_{dem} \frac{1}{N} \sum_{i=1}^{N} \frac{l_i}{s_i} \quad \text{..... (Eq. 3.1)}$$

3.2.2. Significant descriptor selection

The entire chlorine demand data comprised of 201 compounds. $HOCl_{dem}$ ranged from 0.1 mole/mole for non-reactive compounds like ethanol up to 13.40 mol-HOCl/mol-Cp for tyrosine, with a mean of 5.57 mol-HOCl/mol-Cp and standard deviation of 3.31 mol-HOCl/mol-Cp. The frequency distribution of $HOCl_{dem}$ for the 201 compounds is given Figure 3.1. The 201 compounds were split into training data set (N = 159) and external validation data set (N = 42). Multiple linear regression was algorithm of choice for selection of significant descriptors in training data set using Minitab® statistical software, StatGuide™ version 15 (*38*). $HOCl_{dem}$ was used as the dependent variable and descriptors as independent variables.

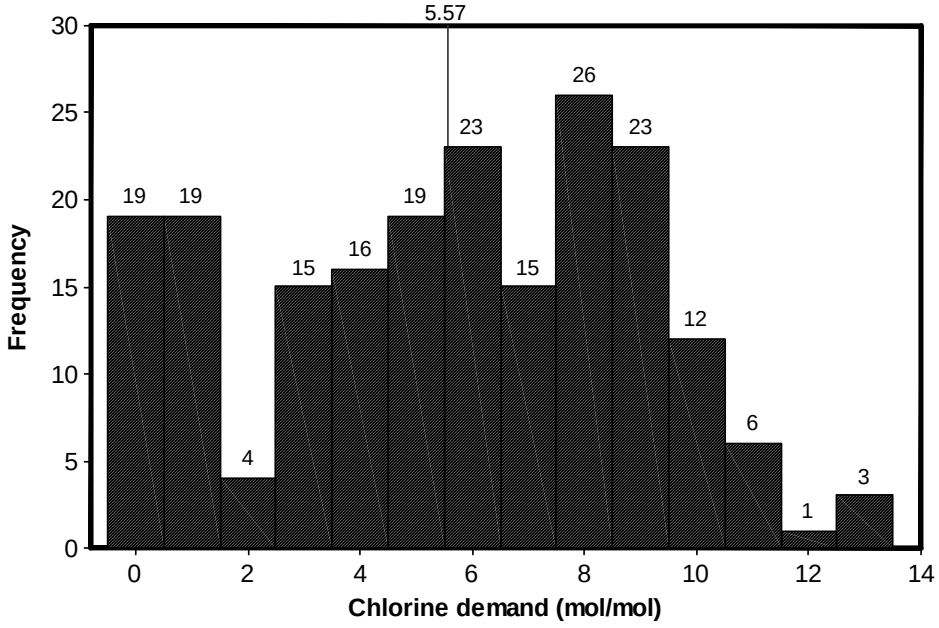


Figure 3.1. Distribution of the chlorine demand for 201 compounds

Descriptor selection process started with a complete list of all potential descriptors listed in Table 2.2 and 159 model compounds. Successive elimination of the insignificant descriptors ($p < 0.05$) left eight significant descriptors listed in Table 3.1. Correlation analysis of the eight descriptors indicated that the descriptors were independent because they had $r < 0.7$ (Table 3.1). These descriptors were used in model calibration and validation.

Table 3.1. Correlation matrix of the eight descriptors

	RAI	ArOH	ACN	CI	OC	AS	ArORact	ArORnact
RAI	1.00							
ArOH	0.63	1.00						
ACN	-0.44	-0.39	1.00					
CI	-0.17	-0.16	-0.13	1.00				
OC	-0.49	-0.33	0.12	0.13	1.00			
AS	-0.16	-0.17	0.25	-0.05	-0.09	1.00		
ArORact	-0.21	-0.22	-0.11	-0.05	-0.05	-0.06	1.00	
ArORnact	0.37	0.19	-0.18	-0.06	-0.04	-0.07	-0.09	1.00

3.2.3. Calibration and validation

The training data (N = 159) were used for model calibration using MS Excel for Windows XP and cross validations- Leave-Many-Out (LMO) and Leave-One-Out (LOO) cross-validations. In Leave-Many-Out cross validation, the training data were split into a calibration data set and a cross validation data set five times using stratified data splitting so that each compound is used at least once in cross validation (*39*). Each of the 5 calibration data sets had 109 compounds and each cross validation data set had 50 compounds (example, Tables S3.1 and S3.2) resulting in five subsets of calibration and cross validation data. Analysis Toolpak in MS Excel for Windows XP was used to perform multiple linear regression on each of the five calibration subsets using the eight

descriptors. Each of the five QSPR equations obtained was used to predict chlorine demand of the respective cross validation subsets and statistics of predictive power were obtained.

The final QSPR was obtained from the five QSPR equations by averaging coefficients of descriptors and standard error of descriptors in the five equations. This equation was then used to predict $HOCl_{dem}$ of each compound in the five cross validation subsets from which averaged statistics of LMO cross validation were obtained. The equation was also used to predict $HOCl_{dem}$ of the all 159 compounds in training data set and its statistics of cross validation were compared to the average statistics of cross validation obtained previously. The bias in data splitting in LMO was cross checked using Leave-One-Out (LOO) cross validation.

Leave-One-Out cross validation used the entire 159 compounds. Each time 1 compound was left out and 158 compounds were used to calibrate the model to obtain R_c^2 and SDE, coefficient of each descriptor and its standard error. The QSPR obtained was used to predicting the chlorine demand of the single compound left out. The compound left out was put back in training data and another compound was taken out and the process was repeated 159 times to obtain 159 QSPR equations. The observed and predicted chlorine demands for each compound were then used to determine statistics of predictive power. QSPR was tested for predictive power using 42 compounds in listed in Table S3.1. The equations for calculating model predictive power and threshold values for each statistic are given in Chapter 2 (Section 2.3.2).

3.3. Results and Discussion

The QSPR was calibration using the Leave-Many-Out (LMO) approach from which five QSPR equations for each of the five calibration data sets (Tables S3.4 and S3.5). The average QSPR model was obtained from the five equations. The coefficients, β, and standard error for each descriptor, based on Equation 2.1, are given in Table 3.2. The average model obtained had regression R_c^2 = 0.86 and SDE = 1.24 mol-HOCl/mol-Cp which is similar to R_c^2 = 0.87 and SDE = 1.21 mol-HOCl/mol-Cp obtained by regression of HOCl demand for all 201 compounds (Table S3.4). This indicates that stratified data splitting that was used in LMO had little bias and the effects of spread in the data using five-fold data splitting was minimized.

Table 3.2. The average coefficients and standard errors for the eight descriptors obtained using LMO approach (Tables S3.4 and S3.5)

Descriptor, x_j	Coefficient, β_j	StdE, ε_i	$(\varepsilon_j / \beta_j)*100$
RAI	7.61	0.34	4.5 %
ArOH	1.16	0.26	22.4 %
ACN	3.00	0.20	6.7 %
CI	1.23	0.23	18.7 %
OC	1.01	0.28	27.7 %
AS	2.37	0.54	22.8 %
ArORact	0.49	0.17	34.7 %
ArORnact	-0.72	0.28	38.9 %

3.3.1. LMO and LOO cross validations

In Leave-Many-Out cross validation (LMO_{CV}), five validation datasets (N = 50) were used (Table S3.2). Each QSPR equation (Tables S3.4 and S3.5) was used to predict chlorine demand for each of the five validation subsets and average statistics of cross validation were q^2_{LMO} = 0.83, $RMSE_{LMO}$ = 1.32 mol-HOCl/mol-Cp and MBD = 1.01 % (Table S3.6). The average statistics obtained

from five plots of predicted HOCl demand against observed HOCl demand with y-intercept were R_i^2 = 0.84, k_i = 0.86. When the intercept was set to zero the average values were R_o^2 = 0.81, k_o = 0.97 and R_t = 0.03 (Table S3.6). These statistics indicate that the five QSPR equations have high predictive power (*41,42*). On the other hand, Leave-One-Out cross validation (LOO_{CV}) cross validation had q^2_{LOO} = 0.85, $RMSE_{LOO}$ = 1.28 mol-HOCl/mol-Cp and MBD = -0.55%, R_i^2 = 0.85, k_i = 0.88, R_o^2 = 0.84, k_o = 0.97 and R_t = 0.01 (Figure 3.2 & Table S3.6,). These results are comparable to those obtained from LMO cross validation.

Prediction using the averaged model (Equation 2.1 and Table 3.2) on each of the five cross validation sets gave q^2_{LMO} = 0.85, $RMSE_{LMO}$ = 1.22 mol-HOCl/mol-C and MBD = 0.99 % (Table S3.7). The average statistics obtained from five plots of predicted HOCl demand against experimental HOCl demand with y-intercept were R_i^2 = 0.86, k_i = 0.87. When the intercept was set to zero the average values were R_o^2 = 0.84, k_o = 0.97 and R_t = 0.02 (Table S3.7). Finally the average QSPR predictive power was tested using the entire training data set (N = 159) and obtained q^2_{LMO} = 0.86, $RMSE_{LMO}$ = 1.21 mol-HOCl/mol-C and MBD = -0.28% (Table S3.7). The plot of predicted $HOCl_{dem}$ versus experimental $HOCl_{dem}$ with and without y-intercept had R_i^2 = 0.86 and k_i = 0.88 and R_o^2 = 0.85, k_o = 0.97 respectively (Figure 3.3) and R_t = 0.02. These statistics of predictive power are comparable and meet the criteria for checking predictive power of QSPR/QSAR (*41,42*). Therefore the average QSPR model obtained is robust

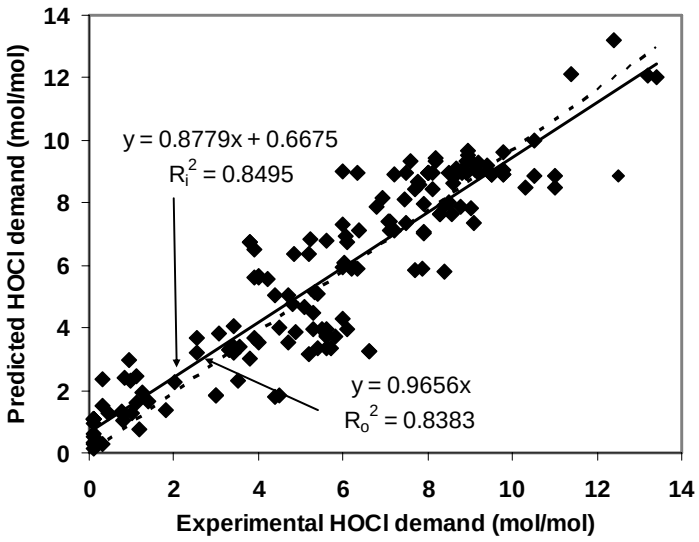


Figure 3.2. Regression of predicted $HOCl_{dem}$ on observed $HOCl_{dem}$ with y-intercept (R_i^2) and through origin (R_o^2) for LOO_{CV}

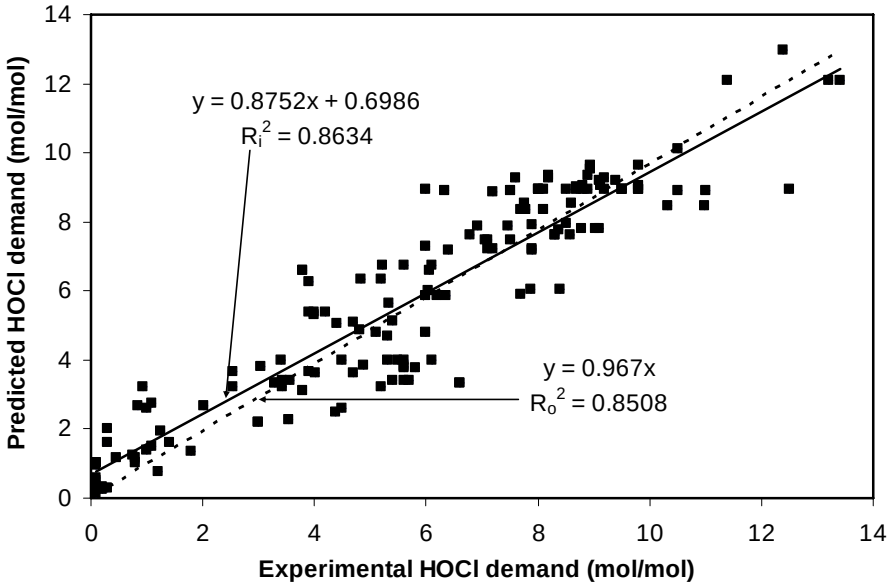


Figure 3.3. Regression of predicted $HOCl_{dem}$ on observed $HOCl_{dem}$ with y-intercept (R_i^2) and through origin (R_o^2) for LMO_{CV}

Plotting predicted $HOCl_{dem}$ against observed $HOCl_{dem}$ showed that the two were linearly related and most of the points along the line for the ideal QSPR there were only a few points fell outside the standard deviation margins (Figure 3.4).

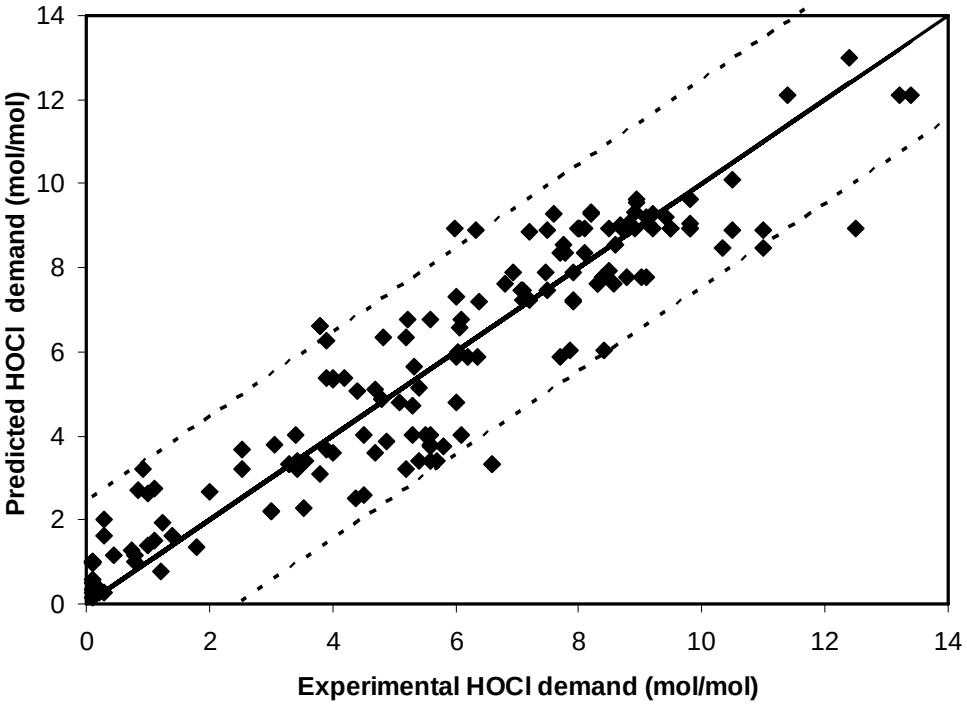


Figure 3.4. Deviation of predicted $HOCl_{dem}$ from ideal QSPR (N = 159) with ±2SDE margins

The plot of standardized residuals against predicted $HOCl_{dem}$ also supported the argument that the model has predictive power as the data points scattered with no distinct pattern with only a few points outside ±2.5 standardized residuals (Figure 3.5). The results from Leave-One-Out cross validation (N = 159) using statistical and graphical approaches were comparable to those obtained by LMO cross validation (Tables S3.6 and S3.7). The consistency of the two results implies that there was no serious bias in 5-way LMO data splitting employed to calibrate the QSPR model.

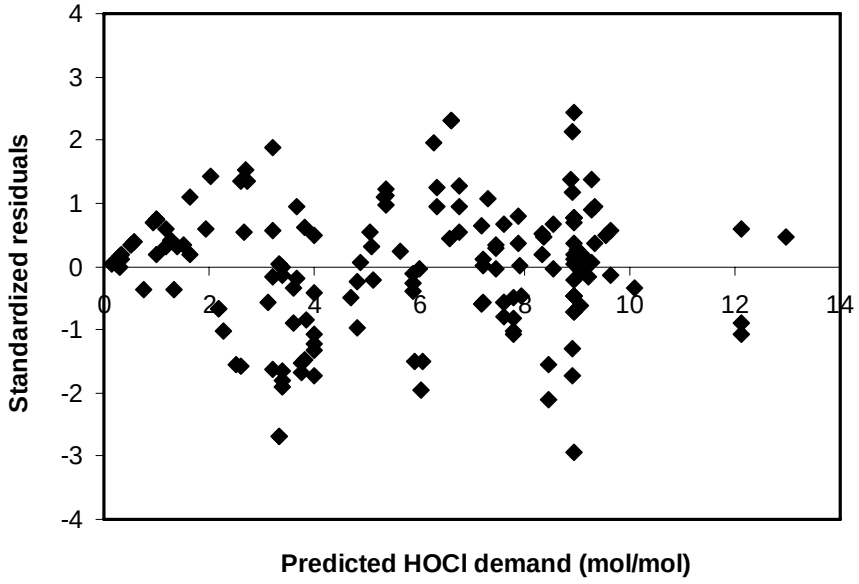


Figure 3.5. Standardized residuals for prediction of $HOCl_{dem}$ by LMO

3.3.2. External validation

The external validation data set (N = 42) has $HOCl_{dem}$ ranging from 0.1 to 11 mole chlorine per mole compound with a mean of 4.98 mol-HOCl/mol-Cp and standard deviation of 3.35 mol-HOCl/mol-Cp. The $HOCl_{dem}$ for each test compound was computed (Equation 2.1 and Table 3.2) from which we obtained q^2_{Ext} = 0.88, $RMSE_{Ext}$ = 1.17 mol-HOCl/mol-C and MBD = 11.42% (Table S3.8). The plot of predicted $HOCl_{dem}$ versus observed $HOCl_{dem}$ with y-intercept gave R_i^2 = 0.91, k_i = 0.90 and the regression through origin gave R_o^2 = 0.87 and k_o = 1.05 (Figure 3.6) and the ratio R_t = 0.04. These statistics of power were comparable to average values obtained when each of the five QSPR equations was used to predict the test compounds (Table S3.8).

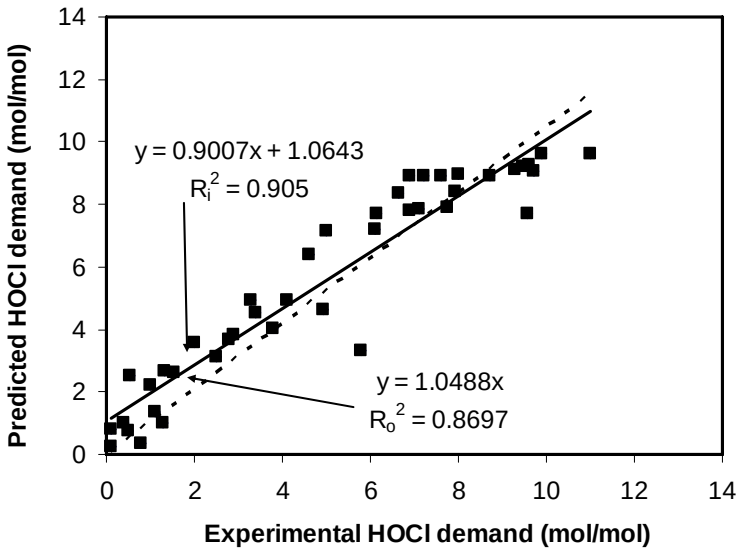


Figure 3.6. Plot of predicted $HOCl_{dem}$ against observed $HOCl_{dem}$ with y-intercept (R_i^2) and through origin (R_o^2) for external validation

y-Permutation test was performed 60 times and the average permuted R^2 was 0.073 and q^2 was -0.186 which are less than 0.3 and 0.05 respectively, implying that the model is robust. These external validation statistics are consistent with those obtained from LMO_{CV} and LOO_{CV}, satisfying the criteria for QSPR/QSAR predictive power (*41,42*). This confirms that the QSPR is robust.

The plot of predicted $HOCl_{dem}$ versus observed $HOCl_{dem}$ in the external validation dataset showed a good linear relationship (Figure 3.7). However, most of the points lie above the line bisecting the two axes, indicating a bias toward over-prediction for compounds with $HOCl_{dem}$ < 9 mol-HOCl/mol-Cp (Figure 3.7) consistent with MBD of 11%. Nonetheless, all compounds were predicted within ±2.5 standardized residuals (Figure 3.8). This indicates that there were no extreme outliers in validation data set.

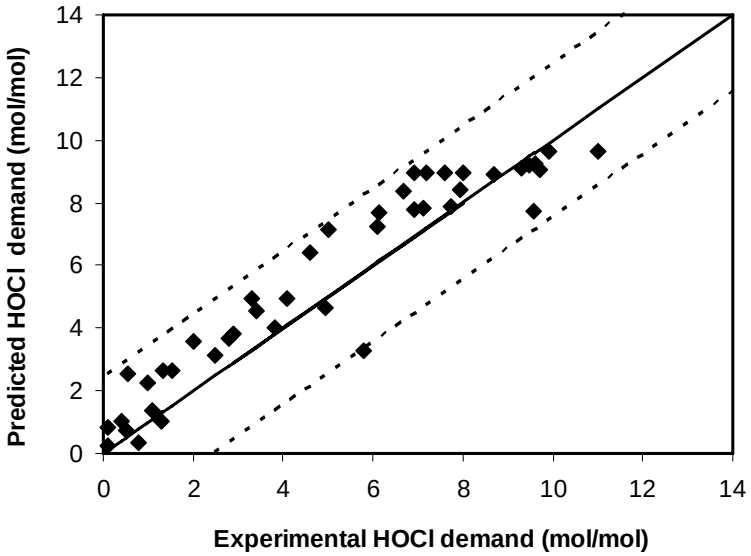


Figure 3.7. Deviation of predicted $HOCl_{dem}$ from ideal QSPR for external validation data with ±2SDE margins

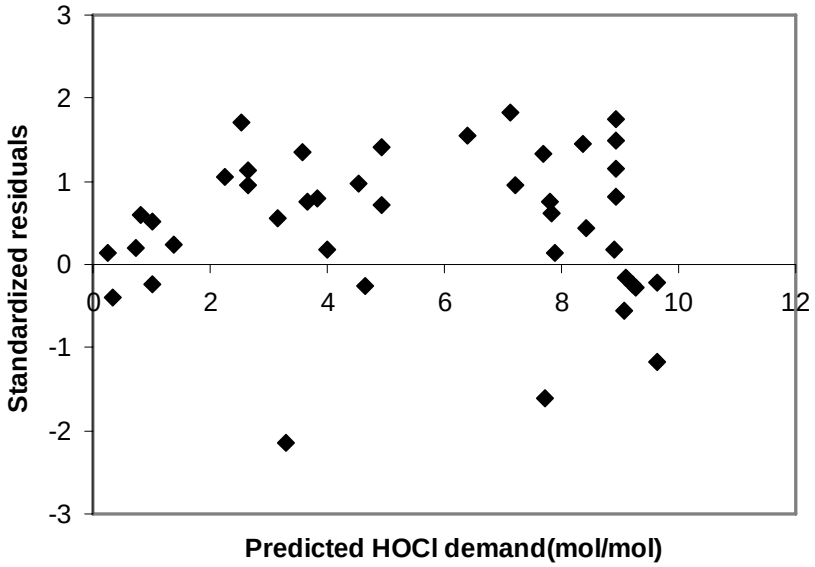


Figure 3.8. Standardized residuals for $HOCl_{dem}$.for external data

3.4. QSPR Applicability Domain

Results from the applicability domain evaluation indicate that 3 of the 159 compounds had cross validation standardized residuals beyond ± 2.5 (Figure 3.8). 4-Iodophenol (SDR = -2.94), leucine (SDR = -2.70) and isoleucine (SDR = -2.70) are therefore outliers in terms of fits in training data set. However they are not considered influential as they fall within model applicability domain and were therefore retained in the data set. One compound, acetylacetone acid (h = 0.30) had h > 0.25 (Figure 3.9).

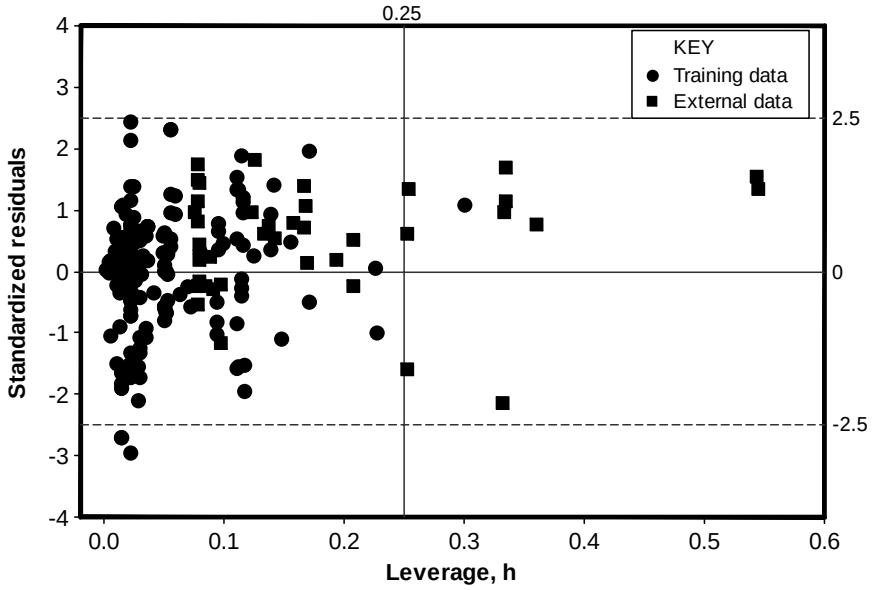


Figure 3.9. Williams plot for detection of outliers and influential observations in training and external validation data sets

The leverage of this compound is closer to the cutoff point of 0.25 and therefore, acetylacetone is not considered structurally most influential in determination of the model descriptors (*39*). Seven compounds in the external validation data had leverage, h > h* and far from acetylacetone. These compounds were 3-oxohexanedioic acid (h = 0.33), 3,4,5-Trimethoxybenzyl alcohol (h = 0.33), 3-(3,4,5-trimethoxyphenyl) propionic acid (h = 0.33), 3,4,5-

Triethoxybenzyl alcohol (h = 0.33), 1,2,3-trihydroxybenzene (h = 0.36) and 3-ethylaceto acetate (h = 0.54) and ornithinechlorohydrate (h = 54). The predicted values for these compounds were obtained due to over-extrapolation.

3.5. Mechanistic Implications of the Descriptors in the QSPR

The final model had 8 descriptors- RAI, ACN, CI, ArOH, AS, O:C, ArORnact and ArORact that differed in their contribution to $HOCl_{dem}$. The ring activation index (RAI), which accounts for the ratio of OH and NH_2 to the number of rings, is the most significant and represents the degree of ring activation that favors electrophilic substitution reactions. Anilinic and phenolic compounds consume more chlorine than non-activated aromatics; however, more than one electron donor on the ring slows the reaction due to steric effects and antagonism among the substituents particularly when they are ortho or para to each other. Aliphatic carbon bonded to reduced nitrogen, NH_2 (ACN), the second most significant descriptor, represents increased $HOCl_{dem}$ due to chlorine substitution on amine which accounts for most of substitution reaction in amino acids (*31*).

The remaining descriptors represent smaller but mechanistically distinct effect. β-dicarbonyl compounds have acidic carbons that allow HOCl to abstract the proton easily to form keto-enolates. However, the acidity of carbon decreases if sandwiched between two carboxylic acids or between carboxylic acid and keto group or aldehyde group. Sulfur increases reactivity of the molecules because sulfur has lone pairs of electron that could be donated and thiols tend to be more reactive than alcohols; e.g., benzothiomide had higher chlorine demand than

benzamide as was the case of phenylthiourea and acetanilide. The more oxidized molecules, (i.e., high O:C) are less reactive towards HOCl due to having fewer reduced carbons. This agrees with the observation that oxalic acid, fumaric and maleic acid had $HOCl_{dem}$ of less than 1 (*21*). Although NO_2 and COOH in molecules like nitrobenzene and benzoic acid deactivate the ring, the $HOCl_{dem}$ will always be positive (*21*) as account for by the O:C descriptor. 1-Nitroalkanes, such as nitromethane, has been reported to undergo chlorine substitution at α-carbon at pH 6-8 (*44*) and the O:C can account for prediction of its $HOCl_{dem}$ in this case as well. Alkoxy groups are weak ring activators relative to OH and NH_2. When alkoxy groups are present with strong ring activator (ArOH or $ArNH_2$) its contribution is less significant or negative due antagonistic effects. Thus $\beta_{ArORnonact}$ is negative, and the more the number of ArORnonact in the ring the lower the $HOCl_{dem}$ relative to aniline or phenol (*23*). However, when alkoxy groups are the only activating groups in the ring or they are present with deactivating groups, they will activate the ring and $\beta_{ArORact}$ is positive as observed in the case of 3,5-dimethoxybenzoic acid and 3,4,5-trimethoxybenzoic acid (*23,30*).

Constitutional descriptors may fail to explain reliably differences in $HOCl_{dem}$ of isomeric molecules because they cannot completely represent steric and electronic effects on reactivity towards HOCl. For instance, the aromatic isomeric pairs 1,2,4-trihydroxybenzene ($HOCl_{dem}$ = 3.9 mol-HOCl/mol-Cp) and 1,2,3-trihydroxybenzene ($HOCl_{dem}$ = 6.9 mol-HOCl/mol-Cp), 2-nitrophenol ($HOCl_{dem}$ = 9.60 mol-HOCl/mol-Cp) and 4-nitrophenol ($HOCl_{dem}$ = 7.60 mol-

HOCl/mol-Cp), and 2-methoxy phenol ($HOCl_{dem}$ = 7.7 mol-HOCl/mol-Cp) and 4-methoxyphenol ($HOCl_{dem}$ = 3.5 mol-HOCl/mol-Cp) show remarkable differences. Aliphatic isomers can show similar differences- For instance, 3-oxopentanedioic acid ($HOCl_{dem}$ = 5.3 mol-HOCl/mol-Cp) and 2-oxopentanedioic acid ($HOCl_{dem}$ = 1.4 mol-HOCl/mol-Cp).

3.6. Prediction of Chlorine Demand of Large DOM Surrogates

The QSPR model is derived using model compounds of low molecular weight. The predictive power of the QSPR model (Equation 2.1 and Table 3.3) on large molecules was tested using tannic acid ($C_{76}H_{52}O_{46}$, MW = 1701) which has an experimental chlorine demand of 32.4 mol-HOCl/mol-Cp (~35.5 mmol-HOCl/g-C) at pH 7 (*34*). The predicted chlorine demand with standard error of prediction was 33.42 ± 6.50 mol-HOCl/mol-Cp which is equivalent to 36.61 mmol-HOCl/g-C. This prediction is slightly higher than experimental, consistent with external validation results. Experimental results for typical fulvic acids are 27-33 mmol-HOCl/g-C (*35*), consistent with this prediction.

The predictive power of the QSPR model was also tested using two proposed model structures of fulvic acid (*45*). The first fulvic acid model structure (FA-1) and the second fulvic acid structure (FA-2) had molecular formulae of $C_{43}H_{44}O_{25}$ (MW = 960) and $C_{42}H_{44}O_{25}$ (MW = 948) respectively. The constitutional descriptors were calculated and used to predict $HOCl_{dem}$ for FA-1 of 8.37 mol-HOCl/mol-Cp (~16.21 mmol-HOCl/g-C) and 7.96 mol-HOCl/mol-Cp (~15.78 mmol-HOCl/g-C) for FA-2. These two model structures have two phenyl rings substituted with at most two hydroxyl groups and ketones which are

expected to contribute most to predicted $HOCl_{dem}$ at pH 7 and chain of aliphatic carboxylic acid and esters in the molecules would not. Despite an unavailability of experimental data, these $HOCl_{dem}$ predictions are within the expect range of $HOCl_{dem}$ of phenolic model compounds reported in literature (*21,22,23,32*).

3.7. Conclusions

In this study a robust QSPR for predicting chlorine was developed using chlorine demand data for model compounds. The QSPR had eight descriptors that explained over 84% of variance in chlorine demand. The model was validated by LMO cross validation and external validation data which indicated that it met criteria for predictive power. However, the model showed that it over-predicted chlorine demand of most compound in external data and prediction of 4 out 27 compounds in the data set were not reliable. One of the reasons is that the QSPR used constitutional descriptors which cannot explain steric and position isomerism effect on chlorine demand. Thus, combination of constitutional descriptors with quantum chemical, geometrical, electrostatic and typological descriptors may be important for both the prediction, determine reaction rates and extrapolation chlorine demand of molecules with diverse structures.

3.8. References

1. Lee S.H.; Levy D.A.; Craun, G.F.; Beach M.J.; Calderon R.L. Surveillance for waterborne-disease outbreaks-United States, 1999-2000. *Morbid. Mortal. Wkly Rep.,* 2002, 51(SS-8), 1-28.

2. Acosta, C.J.; Galindo, C.M.; Kimaro, J.; Senkoro, K.; Urasa, H.; Casalo, C.; Corachan, M.; Esko, N.; Tanner, M.; Mshindo, H.; Lwilla, F.; Vila, J.; Alonso, P.L. Cholera outbreak in Southern Tanzania: Risk factors and patterns of transmission. *Emerg. Infect. Dis.* **2001**, *7*(3), 583–587.

3. Anderson, Y.; Bohan, P. *Disease surveillance and waterborne disease outbreaks*. In: Water quality guidelines, standards and health: Risk assessment and management for water related infectious diseases; Fewtrell, L.; Batram, J. (eds); WHO and IWA Publishers: London: 2000, pp. 115-133. http://www.who.int/water_sanitation_health/dwq/iwachap6.pdf

4. Christman, K. History of chlorine. Chlorine Chemistry Council, Waterworld, September 1998. http://www.waterandhealth.org/drinkingwater/history.html

5. Larson, R.A.; Weber, E.J. *Reaction mechanisms in environmental organic Chemistry*; Lewis Publishers: New York: 1994.

6. Crittenden, J. C.; Trussell, R. R.; Hand, D. W.; Howe, K. J.; Tchobanoglous, G. *Water Treatment: Principles and design*, 2nd edn, John Wiley & Sons Inc., New York: 2005.

7. IPCS (International Program for Chemical Safety). *Disinfectants and disinfectant byproducts*: *Environmental Health Criteria 216*; WHO: Geneva: 2000. http://whqlibdoc.who.int/ehc/WHO_EHC_216.pdf

8. Iriarte, U.; Álvarez-Uriarte, J.I.; Lopéz-Fonseca,·R.; Gonzalez-Velasco, J.R. Trihalomethane formation in ozonated and chlorinated surface water. *Environ Chem Lett*. **2003**, ***1***, 57–61.

9. USEPA (US Environmental Protection Agency). *National primary drinking water regulations: disinfectants and disinfection by-products*; Final rule, Federal Registry, 63:241:69390; US Environmental Protection Agency: 1998. http://www.epa.gov/ogwdw/mdbp/dbpfr.html

10. EU (European Union). *Council directive 98/83/EC of November 3, 1998 on the quality of water intended for human consumption*; European Union/Commission Legislative Documents: European Union, 1998. http://www.doeni.gov.uk/1998_drinking_water_directive.pdf

11. Richardson, S.D. Simmons, J.E. and Rice, G. Disinfection byproducts: The next generation. *Environ. Sci. Technol.* **2002**, *36*(9), 198A–205A.

12. Porter, C. K., Putnam, S. D., Hunting, K. L. and Riddle, M. R. The effect of trihalomethane and haloacetic acid exposure on fetal growth in a Maryland County. *Am. J. Epidemiol*. **2005**, *162*, 334–344.

13. IARC *(*International Agency for Research*). Monographs on the evaluation of carcinogenic risks to humans, Volume 52,* IARC, IARC Press, Lyon, France: 1999. http://monographs.iarc.fr/ENG/Monographs/vol52/volume52.pdf

14. Gang, D.D.; Segar Jr., R.J.; Clevenger, T.E.; Banerji, S.K. Using chlorine demand to predict THM and HAA9 formation. *J. Am. Water Works Assoc*. **2002**, *94*(10), 76-85.

15. Liang, L.; Singer, P.C. Factors Influencing the formation and relative distribution of haloacetic acids and trihalomethanes in drinking water. *Environ. Sci. Technol*. **2003**, 37, 2920-2928.

16. Baxter, C.W.; Smith, D.W.; Stanley, S.J. A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. *J. Environ. Eng. and Sci*. **2004**, 3, S45-S58.

17. Shimazu, H.; Kouchi, M.; Yonekura, Y.; Kumano, H.; Hashiwata, K.; Hirota, T.; Ozaki, N.; Fukushima, H. Developing a model for disinfection by-products based on multiple regression analysis in a water distribution system. *J. Water Supply Res. T*. **2005**, *54*(4): 225-237.

18. Fitzgerald, F.; Chow, C.W.K.; Holmes, M. Disinfectant demand prediction using surrogate parameters-a tool to improve disinfection control. *J. Water Supply Res. T*. **2006**, *55*, 391-400.

19. Hong, H. C.; Wong, M. H.; Liang, Y. Amino acids Precursors of trihalomethane and haloacetic acid formation during chlorination. *Arch. Environ. Toxicol.* **2009**, 56, 638-645.

20. Perdue, E.M.; Ritchie, J.D. Dissolved organic matter in freshwater. In *Surface and Groundwater, Weathering, Erosion and Soils*, volume 5; Drever, J. ed.; Elsevier Inc.: San Diego, CA: 2005 pp 273-318.

21. de Laat, J.; Merlet, N.; Doré, M. Chlorination of organic compounds: chlorine demand and reactivity in relationship to the trihalomethane formation. *Water Res.* **1982**, *16*, 1437–1441

22. Boyce, S.; Hornig, J. Reaction pathways of trihalomethane formation from the halogenation of dihydroxyaromatic model compounds for humic acid. *Environ. Sci. Technol.* **1983**, *17*, 202–211.

23. Bull, R.J.; Reckhow, D.A.; Rotello, V.; Bull, O.M.; Kim, J. *Use of toxicological and chemical models to prioritize DBP research;* AWWA Research Foundation: 2006.

24. Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; and Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry*. **2005**, *76*, 319-347.

25. Cabaniss, S.E., Madey, G., Leff, L., Maurice, P.A., and Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part II. Molecular property distributions. *Biogeochemistry.* **2007**, 86, 269-286.

26. Schwarzenbach, R.P.; Gschwend, P.M.; Imboden, D.M. *Environmental organic chemistry*, 2nd edn, John Wiley & Sons Inc., Hoboken, New Jersey: 2003.

27. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSPR/QSAR. *Chem. Rev*. **1996**, *96*(3), 1027-1044.

28. Katritzky, A.R.; Lobonov, V.S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev*. **1995**, *24*, 279-287.

29. Karelson, M. *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience: New York: 2000.

30. Norwood, D.L.; Johnson, J.D; Chrisman, R.F.; Hass, J.R.; Bobenrieth, M.J. Reactions of chlorine with selected aromatic models of aquatic humic material. *Environ. Sci. Technol.* **1980**, *14*(2), 187–189.

31. Hureiki, L.; Croué, J-P.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res*. **1994**, *28*, 2521–2531.

32. Gallard, H.; von Gunten, U. Chlorination of phenols: Kinetics and formation of chloroform. *Environ. Sci. Technol.* **2002**, 36, 884-890.

33. Dickenson, E.V.; Summers, S.; Croué, J-P.; Gallard, A. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol*. **2008**, *42*, 3226–3233.

34. Bond T.; Henriet O.; Goslan E.H.; Parsons S.A.; Jefferson B. Disinfection byproducts and fractionation behavior of natural organic matter surrogates. *Environ. Sci. Technol.* **2009**, 43, 5982-5989.

35. Reckhow, D.A.; Singer, P.C.; Malcolm, R.L. Chlorination of humic materials: Byproduct formation and chemical interpretation. *Environ. Sci. Technol.* **1990**, 24, 1655-1664.

36. *USEPA (US Environmental Protection Agency). Formation of halogenated organics by chlorination of water supplies.* US-Environmental Protection Agency (1975). National Service Center for Environmental Publication (NEPIS). http://nepis.epa.gov/EPA/html/pubs/pubtitleORD.htm

37. Reusch, W. *Virtual textbook of organic chemistry*: 1999 (2008 revision). http://www.cem.msu.edu/~reusch/VirtualText/intro1.htm

38. Minitab Inc. (2007). "Graphical data," *Meet Minitab 15*, p. 2-1 to 2-13. www.minitab.com(accessed October 8, 2009)

39. Herrell Jr, F.E. *Regression modeling strategies with application to linear models, logistic regression, and survival analysis*; Springer-verlag, New York: 2001.

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

41. Tropsha, A.; Gramatica, P.; Gomba, V.K. The Importance of being earnest: Validation is absolute essential for successful application and interpretation of QSPR model. *QSAR Comb. Sci*. **2003**, *23*(1), 69-77.

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

43. Rücker, C.; Rücker, G.; Meringer, M. y-randomization and its variants in QSPR/QSAR, *J. Chem. Inf. Model.* **2007**, 47, 2345-2357.

44. Orvik, J.A. Kinetics and mechanism of nitromethane chlorination. A new rate expression. *J. Am. Chem. Soc*. **1980**, *102*(2), 740-743.

45. Leenheer, J.A.; Brown, G.K.; MacCarthy, P.; Cabaniss, S.E. Models of metal binding structures in fulvic acid from the Suwannee River, Georgia**.** *Environ. Sci. Technol.* **1998**, *32*, 2410-2416.

CHAPTER 4

QSPR FOR PREDICTING TOX FORMATION

Abstract

Chlorination of drinking water treatment produces many disinfection byproducts, the majority of them unknown. Total organic halide (TOX) is used as a surrogate for formation of toxic disinfection byproducts though it is currently not being regulated. In this study 49 compounds were used to train a quantitative structure-property relationship (QSPR) for predicting TOX formation in moles of chlorine (Cl) per mole of a compound (Cp) (mol-Cl/mol-Cp). The 4 descriptor QSPR had R_c^2 of 0.72 and standard deviation of estimation (SDE) of 0.43 mol-Cl/mol-Cp. The Leave-One-Out validation of the QSPR (q^2_{LOO} = 0.60, RMSE = 0.5 mol-Cl/mol-Cp, N = 49) and external validation (q^2_{Ext} = 0.67, RMSE = 0.48 mol-Cl/mol-Cp, N = 12). In addition, statistical power analysis showed the QSPR had high predictive power on external validation data because $R_t < 0.1$ and $0.85 \leq k' \leq 1.15$ indicating that the model is robust. The prediction of TOX formation of tannic acid and two model fulvic acids gave TOX formation between 4.05-4.98 mmol Cl/g-C which is consistent with 2.14-4.11 mmol-Cl/g-C range for TOX formation from dissolved organic matter reported for some river waters in the United States and Canada.

4.1. Introduction

Drinking water disinfection is imperative to protect the public from waterborne diseases. Chlorine is the most commonly used chemical disinfectant in most developed countries and is the disinfectant of choice in developing countries (*1,2*). The advantages of chlorine over the alternative disinfectants are that chlorine is the cheapest in terms of capital investment, operating and maintenance costs (*3,4*). The decline in waterborne diseases outbreaks in many countries is associated with chlorination of potable waters (*5,6*). However, the reaction of chlorine with traces of dissolved organic matter (DOM) produces disinfection byproducts which are potentially toxic. Since the discovery of trihalomethanes (THMs) in the early 1970's (*7,8*) there have been rapid advances in technology leading to identification of more classes of disinfection byproducts (DBPs) (*9,10,11,12*).

The chlorination of drinking water produces several chlorinated disinfection byproducts and only a fraction of them have been identified (*13,14*). However, toxicological studies have been based on individual trihalomethanes or haloacetic acids at doses which sometimes higher than normally present drinking water (*9*). The interactive toxic behavior of the mixture of DBPs may be more toxic than individual DBP. Thus, use of total organic halides (TOX) as a surrogate for toxic DBPs in drinking water becomes attractive. Studies have shown that TOX varies linearly with water quality parameters such as UV absorbance and dissolved organic carbon (*15,16*) and so do DBPs **(***17,18***).** Since both TOX and DBPs vary with DOC and UV absorbance, they are also expected to have some

linear relationship between themselves (*19*). Thus, total organic halide (TOX) is a potential surrogate for toxic DBPs in drinking water (*14,20*) though it is not currently being regulated (*14*).

The US Environmental Protection Agency (USEPA) released Stage 2 Disinfection Byproduct Rule (DBPR) that mandates monitoring of bulk water parameters (TOC and SUVA) as indicators of DBPs formation potential, for all drinking water utilities (*21*). The rule was released though USEPA was aware of the study by Weishaar et al. (*17*) that showed that the relationship between DOC and UV with DBPFP is not always linear and later studies supported results of their work (*22,23,24,25*). Thus, prediction of TOX formation in drinking water poses new challenge to water researchers.

It is important at this point to have a short discussion on total organic halides measurement protocol and interpretation of TOX data with respect to chlorine demand. Review of data for chlorine demand and TOX formation obtained from the same study showed that there is a lack of mass balance between the two (*26,27,28*). TOX analyzers have six key parts: adsorption module, boat inlet, pyrolysis furnace, titration cell, microcoulometer detector and strip chart recorder (*29*). Poor sample preparation and not adhering to sample analysis protocol at these six stages may lead to lack of mass balance. Higher TOX formation than expected may suggest contamination from reagents, glassware, gases or activated carbon used (*29*). Low TOX formation may suggest that there may loss of chlorine from chlorinated organic intermediates in the form of inorganic chlorine compounds as suggested by some proposed

reaction mechanisms (*26,30*) or there may be low recovery of TOX due to irreversible adsorption of some chlorinated byproducts (*31,32,33*) and other factors are detailed in Method 9020B (*29*),

The TOX formation models available in literature are mainly based on bulk water quality despite many studies showing water quality parameters such as DOC and UV fluctuate with time and space (*22,23,25*). The models derived based on these parameters may not be robust. Therefore, predictive models based on structural or functional group information, important in chemical reactivity, may be more stable than those based on bulk water parameters. Apparently no TOX formation model has been reported based on structural properties of individual molecules. Quantitative structure-property relationship (QSPR) using constitutional descriptors have been used in previous works to model degradation of natural organic matter and metal complexation (*34,35*). Constitutional descriptors are more attractive than quantum molecular descriptors (*36,37*) because there is no need to optimize the molecules, which is useful when dealing with mixtures of thousands or tens of thousands of structures (like DOM).

The objective of this paper is to develop a new and robust QSPR that predicts TOX formation from constitutional descriptors and to define its applicability domain. This QSPR would be integrated with an AlphaStep model for DOM in order to predict TOX formation from disinfection of surface waters.

4.2. Methodology

4.2.1. Data sources

This research work is based on data mining and utilized water chlorination experimental data for model compounds reported in three publications (*26,27,28*). TOX formation ranged from 0.004 to 3.00 mol-Cl per mol-Cp for 1,3-dihydroxybenzene (resorcinol) and (β-diketones. The average TOX formation for the entire data set was 0.84 mol-Cl/mol-Cp with standard deviation (Stdev) of 0.80 mol-Cl/mol-Cp (TOX formation data in Tables S4.1 and S4.2). The Stdev was high because TOX formation data was skewed to the low end (Figure 4.1).

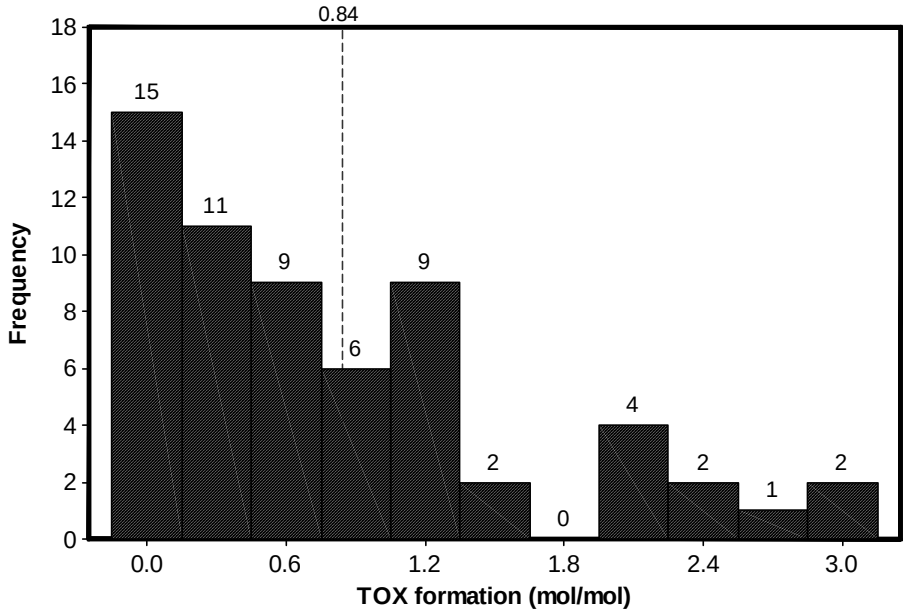


Figure 4.1. Frequency distribution of TOX formation data (N = 61)

4.2.2. Generation of descriptors

The data for 61 compounds were divided into model training data (N = 49) and external validation data (N = 12) by pseudo random data splitting. That is random splitting was done repeatedly until the external validation data contained

the various classes of compounds in the training data. The training data was not split further into calibration data set into Leave-Many-Out cross validation data set because the size of the training data was too small to have a robust average QSPR. The constitutional descriptors listed in Table 2.1 were computed from the structures of the compounds. Compounds lacking certain atom or group were given a value of zero for that descriptor. A descriptor was included in descriptor selection and model calibration only if it had non-zero values for at least five compounds. The data was split randomly into training and external validation data. The external validation data was set aside to validate the model because it was not use in model training.

4.2.3. Descriptor selection

The descriptors were selected by multiple linear regression of TOX using 49 compounds in training data set with constitutional descriptors listed in Table 2.1 using Minitab 15 Software (*38*). MLR through the origin was carried out using Equation 2.1 at 95% confidence interval. The descriptor selection process was repeated after dropping those with $p > 0.05$, until all remaining descriptors were significant ($p < 0.05$) and correlation coefficient of each pair of descriptors should be < 0.7 ($r^2 = 0.49$) as criteria for independence of descriptors (*39*).

4.2.4. Model calibration

The QSPR was calibrated by performing MLR of TOX formation (N = 49) on the four descriptors through the origin using Analysis ToolPak for MS Excel™ (Windows XP). The coefficient of determination for QSPR calibration (R_c^2) was obtained by using Equation 2.2. The MLR output provided coefficients of

descriptors and standard errors for the QSPR. The statistics of predictive power, standardized residuals (SDR), leverage, Cooks distance (D), difference in fit standardized (DFFITS) and difference in beta standardized (DFBETAS) were calculated using formulae given in Chapter 2.

4.2.5. Internal and external validations

The QSPR was validated using internal cross-validation and external validation from which statistics of predictive power of the QSPR were calculated. The statistical indicators evaluated include q^2, R_c^2, $R_t = (R_i^2-R_o^2)/R_i^2$ and k (*40,41*). The q^2 is defined as the coefficient of determination for validation (i.e., R^2) which can be calculated using Equations 2.3 and 2.4. The terms q^2_{LOO} and q^2_{ext} are used to denote q^2 for LOO_{CV} and external validation respectively.

4.3. Results and Discussion

4.3.1. QSPR calibration

The descriptor selection process gave four significant constitutional descriptors ($p < 0.05$) namely carbonyl index (CI), number phenols per carbon (ArOH:C), square root of number of heteroatoms (sqrtHeA), and the log of hydrogen to carbon ratio (logH:C). Correlation analysis showed that all pairs of descriptors had absolute correlation coefficients <0.7 (Table 4.1) indicating that they were mutually independent (*41*).

Table 4.1. Correlation matrix of the descriptors

	CI	ArOH:C	sqrtHeA	logH:C
CI	1.00			
ArOH:C	-0.18	1.00		
sqrtHeA	-0.18	-0.27	1.00	
logH:C	-0.06	-0.48	0.09	1.00

The calibration data set listed in Table S4.1 was used to derive the QSPR and for internal validation. The model descriptors listed in Table 4.2 show that order of importance of the descriptors was: Cl (14.9%) > ArOH:C (19.7%) ≈ sqrtHeA (19.5%) > logH:C (40.0%) The statistics of fit indicated that the QSPR could explain 72% (R_c^2 = 0.72) of the variance in TOX formation with a standard deviation of regression, SDE of 0.43 mol-Cl/mol-Cp. This suggests the model would have good predictive power particularly for those molecules that produce TOX above 0.43 mol-Cl/mol-Cp.

Table 4.2. QSPR for TOX formation (N = 49)

Descriptor, x_j	Coefficients, β_j	StdE, ε_j	$(\varepsilon_j / \beta_j)*100$
Cl	0.54	0.08	14.0%
ArOH:C	5.33	1.05	19.7%
sqrtHeA	0.33	0.06	19.5%
logH:C	-1.36	0.54	40.0%

4.3.2. QSPR validations

Internal validation

Leave-One-Out cross validation was used to validate the QSPR using calibration data set (Table S4.1) in order to obtain additional statistics of predictive power, q^2, k and R_t. Results showed that q^2_{LOO} = 0.60, $R_t^{'}$ = 0.001, k_i' = 0.95 and k_o' = 0.99 (Table 4.3) and all these parameters met criteria for predictive power (*40,41*). The RMSE of LOO_{CV} was 0.50 mol-Cl/mol-Cp which is similar to model standard error of 0.43 mol-Cl/mol-Cp.

Table 4.3. QSPR predictive power using internal and external validation (Unit of RMSE is in mol-Cl/mol-Cp and MBD is in %)

	q^2	MBD	RMSE	$R_i^{2'}$	k_i'	B	R_o^2	k_o'	R_t'
LOO_{CV}	0.60	-0.88	0.50	0.60	0.95	0.05	0.60	0.99	0.001
External data	0.67	-0.70	0.48	0.65	1.14	0.13	0.65	1.04	0.004

The performance of the QSPR was evaluated by visualizing the distribution of the points along the ideal model line in the graph of predicted TOX formation versus experimental TOX formation (Figure 4.2) and standardized residual plot (Figure 4.3).

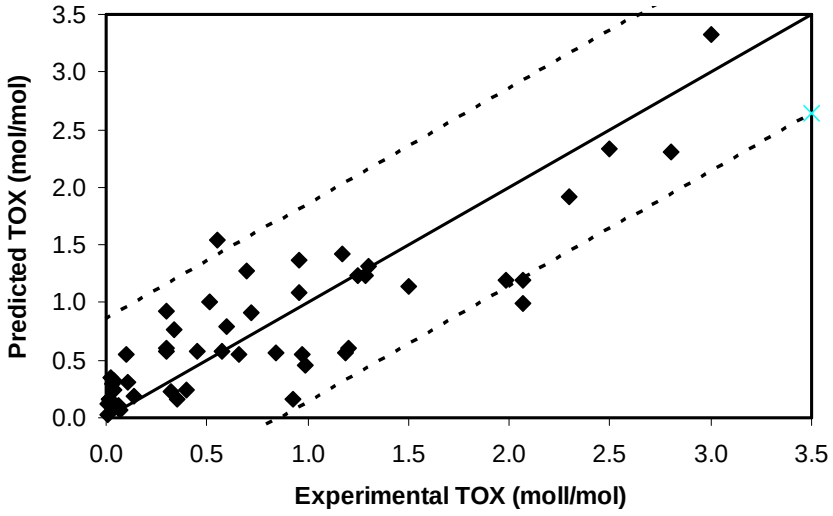


Figure 4.2. Deviation of predicted TOX formation to ideal QSPR (N = 49). The dotted lines represent the ±2SDE margins

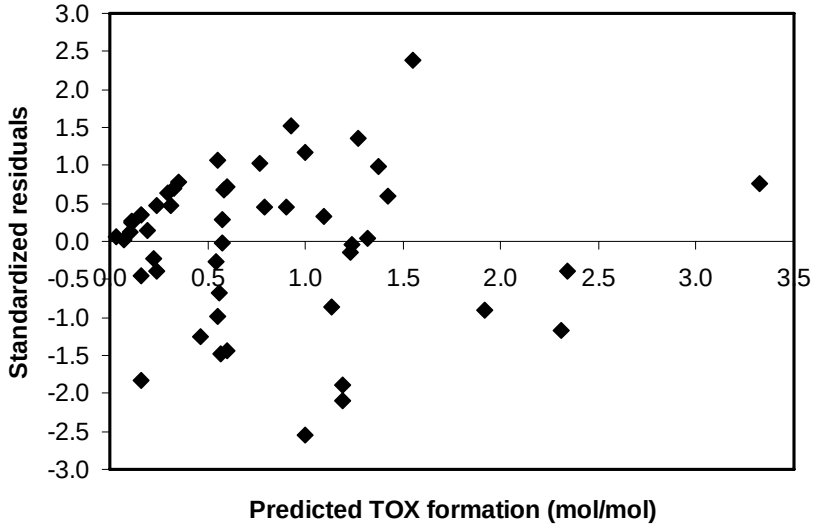


Figure 4.3. Predictive power diagnosis using SDR plot (N = 49)

Figures 4.2 and 4.3 revealed that 2-oxobutanedioic acid and coniferyl alcohol were slightly outside the ±2SDE margins. Thus, the QSPR underpredicted TOX formation for coniferyl alcohol (SDR = -2.56) but it was not considered an outlier because it is right at the margin.

The total TOX formation in calibration data is 38.98 mol-Cl/mol-Cp and the 49 compounds were divide into three groups based on TOX formation trends. The first group includes 8 compounds with higher TOX formation (1.50 ≤ TOX ≤ 3.00 mol-Cl/mol-Cp) which accounts for a total of 18.23 mol-Cl/mol-Cp (46.75%) of the total TOX formation in the training data set. Although the QSPR predicted slightly low for 7 out 8 compounds in this group, they were within ±2.5 standardized residuals. Coniferyl alcohol, ferulic acid, tyrosine and syringaldehyde showed slightly lower TOX formation than the rest of the compounds in the group. They lack β-diketone or hidden carbonyl (as in 1,3-dihydroxybenzene) functionality which are responsible for high chlorine consumption and DBPs formation of TOX (*27,42*). Nonetheless coniferyl alcohol and ferulic acid have aliphatic C=C in their structure and tyrosine has amine group in aliphatic substituent (alanine moiety) which are good sites for chlorine attack. These two functionalities set these compounds apart from phenolic compounds in the other groups in terms of TOX formation. This group contains the most reactive compounds and accounts for 16.3% of the total number of compounds in the training data set.

The second group has 20 compounds with intermediate TOX formation (0.45 ≤ TOX ≤ 1.30 mol-Cl/mol-Cp) and their total TOX is 17.81 mol-Cl/mol-Cp.

The TOX formation accounts for 45.69% of the total TOX formation in training data set. Nineteen compounds in this group were derivatives of phenol, aniline, 3,4,5-trimethoxy benzenes and 3,4,5-triethoxybenzenes. The QSPR predicted slightly higher than expected the majority of the compounds in this group. These compounds share one structural feature in common, i.e., they have NH_2, OH, OCH_3 or OC_2H_5 as electron donating groups in the aromatic ring. Though chlorine demand data for these compounds showed that anilines and phenols had higher chlorine demand than 3,4,5-trimethoxy benzenes or 3,4,5-triethoxy benzenes (*26*), there is no similar trend in TOX formation. This may attributed to differences in adsorption affinity of adsorbates to adsorbent (activated carbon) in TOX analyzers.

Adsorption of phenolics on activated carbon may occur irreversibly or reversibly as determined relative amounts of adsorbate recovered by desorption process (*31,32,33,43,44*). The factors that influence adsorption include: nature of adsorbent (e.g., pore size, functional groups, ash content), nature of adsorbates (e.g., concentration, functional groups, size, solubility) and conditions (e.g., pH, ionic strength) of liquid medium (*45*), Garcia-Araya et al. (*44*) studied reversible adsorption of benzoic acids on activated carbon and found that syringic acid was more adsorbed on activated carbon than gallic acid and para-hydroxybenzoic acid. Syringic acid is less soluble in water than other two phenolics and tends to spend more time on solid phase than liquid phase. Reversible adsorption of small chlorinated organic molecules on activated carbon has also been reported to be affected by solubility and hydrophobicity: Trichloroethylene > 1,2-

dichloroethylene > 1,1-dichloethane > carbon tetrachloride > 1,1,1-trichloroethane > chloroform (*43*). Based on adsorption studies, chlorinated phenolics or some of the chlorinated organic byproducts formed from ring opening might be strongly adsorbed. That may account for low recovery.

The third group of compounds had 21 compounds with lower TOX formation than the first two groups (0.004 ≤ TOX ≤ 0.40 mol-Cl/mol-Cp). These compounds contributed 2.94 mol-Cl/mol-Cp to the total TOX formation in training data set. This was only 7.6% of the total TOX formation but 39% of the compounds in the training data. It was found that over 60% of the compounds in this group were non-aromatic amino acids which had surprisingly very low TOX formation despite having chlorine demands between 2 to 8 mol-Cl_2/mol-Cp (*28*). The potential site for chlorine substitution for most aliphatic amino acids is NH_2 on the α-C. Nonetheless, the acidity of α-H is generally very weak because it is adjacent to C=O for carboxylic acid. Low TOX formation from chlorination may be attributed to two possibilities. Although amino acids consume high chlorine, most of the chlorine consumed might be converted into inorganic chlorine. The reaction mechanism proposed by Chu et al. (*30*) showed chlorine added to amino acids may be lost as HCl in later steps of chlorination reaction. Elimination of CO_2 and HCl from may lead to formation of aldehyde and cyanoalkane. Substitution of the two hydrogens on NH_2 may increase electron deficiency on the α-C. The nucleophilic attack on α-C (sp^3) may lead to a loss of NH_2 as monochloroamine ($ClNH_2$) or dichloroamine ($HNCl_2$). HCl and chloroamine, if they are formed, may not be detected by instruments used to measure TOX. The

study by Chu et al. (*30*) did not try to detect the possible inorganic byproducts from chlorination of amino acid. The lack of mass balance between chlorine consumed and TOX formation requires detailed investigations in order rule out or justify proposed mechanisms.

QSPR external validation

The QSPR was validated using an external validation data set of 12 compounds that represented most of the functional groups in the training data set. The descriptors in the QSPR were calculated from the structures and substituted in QSPR (Table 4.2) to estimated TOX formation. The statistics of predictive power for external validation were: q^2_{ext} = 0.67, R_t' = 0.004, k_i'= 1.14 and k_o' = 1.04 (Table 4.3). These statistics meet minimum requirements for QSAR/QSPR predictive power (*40,41*) and therefore the model is robust. In addition, external validation RMSE was 0.48 mol-Cl/mol-Cp which was comparable to SDE (SDE = 0.43 mol-Cl/mol-Cp). Qualitative evaluation of model bias showed that external validation had MBD of -0.70% (Table 4.3) suggesting that the QSPR performance was good. The plot of predicted versus experimental TOX formation showed some deviation of the ideal model fitting line but the TOX was predicted within ±2SDE (Figure 4.4). The standardized residual plot confirmed that all compounds were predicted within ±2.5 standardized residuals (Figure 4.5) which also indicates that the model has high predictive power and is robust.

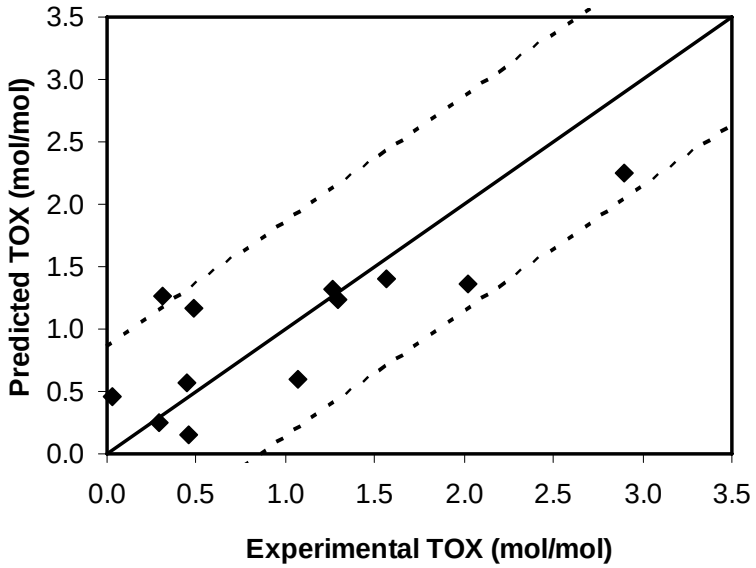


Figure 4.4. Deviation of predicted TOX from ideal QSPR for external validation data. The dotted lines represent the ±2SDE

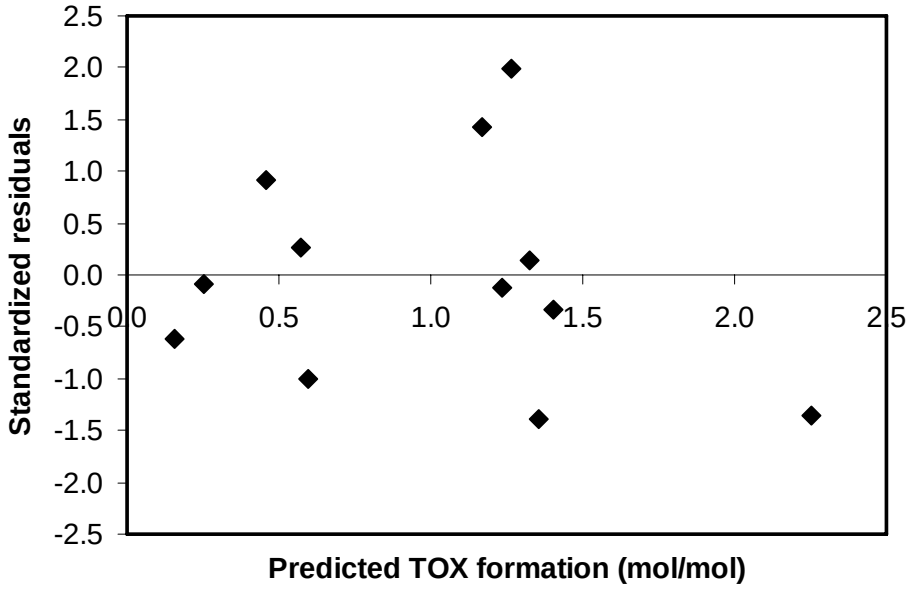


Figure 4.5. Model predictive power on external validation data using SDR plot

4.3.2. Applicability domain of the QSPR

The model applicability domain (AD) was used to assess presence of compounds in the training data set that were extreme to influence coefficients of the QSPR. It was also used to find out compounds in the external validation data that were prediction outliers (predicted due to extrapolation beyond the calibration data). The four descriptors QSPR calibrated by 49 compounds has warning leverage h* of 0.31 (Figure 4.6). The compounds 4-(3,4,5-Trimethoxybenzoyl) butyric acid (SDR = -0.38, h = 0.34), 5,7-dioxooctanoic acid (SDR = -1.17, h = 0.34) and 1,3-dihydroxybenzene (SDR = 0.77, h = 0.57) were out of the applicability domain but within the ±2.5 SDR boundary. Of these, 1.3-dihydrozybenzene is a potentially influential compound because its h is far from h*. The Cook's distance for 1.3-dihydrozybenzene was 0.45 and 0.07 which was less than the cutoff value of 1 (*46*) and therefore was not considered an outlier which agrees with results from residual plot analysis. The DFFITS and DFBETAS statistics (Chapter 2, Eqs 2.11 & 2.12) were used to evaluate their impacts on regression fit and coefficients of descriptors in QSPR upon omission of 1,3-dihydroxybenzene from the training data set. It was found that DFFITS was <2 and DFBETAS for each descriptor was less than <2. This means that the compounds had little influence on the coefficients of descriptors in the QSPR and model fit (*46*). Therefore 1,3-dihydroxybenzene was retained in the training data set. There were three compounds in the external validation data set with leverage far greater than 0.31, 2-aminophenol (SDR = 1.99, h = 0.44), cysteine (SDR = -0.62, h = 0.50) and 3-oxopentanedioic acid (SDR = -1.35, h = 0.96).

Therefore, these three compounds were predicted due to extrapolation of the QSPR.

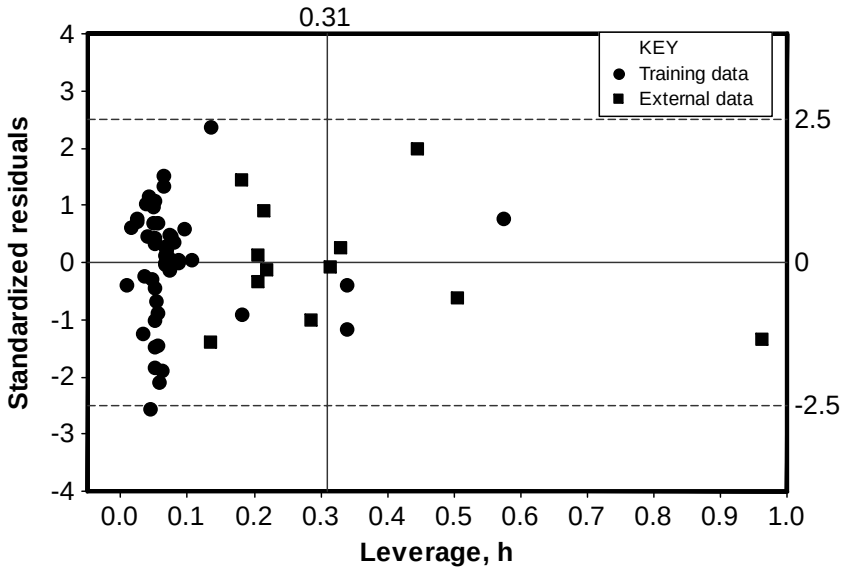


Figure 4.6. Williams plot indicating outliers and influential data points

4.3.3. Prediction of other model compounds

The QSPR was tested for its predictive performance using tannic acid (*47*) and two proposed structures of fulvic acid (*48*). The Cl, ArOH:C, sqrtHeA and logH:C were calculated from the three structures. The predicted TOX formation for tannic acid (MW = 1701) was 3.87 mol-Cl/mol-Cp (~150.63 µg-Cl/mg-C). The first fulvic acid model structure, FA-1 (MW = 960) had predicted TOX formation of 2.09 mol-Cl/mol-Cp (~143.88 µg-Cl/mg-C). The second fulvic acid, FA-2 (MW = 948) has TOX formation of 2.51 mol-Cl/mol-Cp (~176.89 µg-Cl/mg-C). The proposed fulvic model structures and tannic acid have phenols, ethylene and ketone functional groups which play role in chlorine demand and DBPs formation. The predicted TOX formations from the three structures were comparable to TOX formation for coniferyl alcohol, syringaldehyde, 5,7-

dioxooctanoic acid and 1,3-dihydroxybenzene (*26,27*). However there was no experimental formation for tannic acid, FA-1 or FA-2 to compare with. Some other studies showed raw water formation from Massachusetts, Virginia, Texas, Quebec and Manitoba had TOX formation (µg/L) per DOC (mg/L) ranging from 77.80 µg-Cl/mg-C to 192.50 µg-Cl/mg-C (*14*). The model predictions of TOX formation fall within this range.

The QSPR was used to predict TOX formation for tannic acid and two fulvic acid model structures. The structures of the molecules were phenolic in nature but differed in molecular weight and chlorine demand. Chlorine demand for tannic acid was 32.4 mol-Cl_2/mol-Cp (*47*) and calculated demands for FA-1 and FA-2 were 8.37 mol-Cl/mol-Cp and 7.96 mol-Cl/mol-Cp (Chapter 3). However, the QSPR TOX formation predictions ranged from 2.09 to 3.87 mol-Cl/mol-Cp. It would be expected that tannic acid should have had far higher TOX formation than fulvic acids based on chlorine demands. The low TOX formation, as predicted by QSPR, may suggest that the descriptors failed to explain the expected formation in tannic acid. The QSPR could also have given close prediction and low TOX formation may then be associated with irreversible adsorption of organic compounds on the activated carbon in the TOX analyzer. The unexpected low TOX formation relative to chlorine demands was also noted when experimental TOX formation data for phenolic compounds and 3,4,5-triethoxybenzennes or 3,4,5-triethoxybenzenes were compared. They did not show clear differences despite having quite different chlorine demands (Chapter 3). Thus, it is important to bear in mind that TOX recovery from sample is not

100% efficient (*37,49*) and that may affect the predicted TOX formation for some classes of phenolic compounds.

4.4. Conclusions

A robust QSPR for predicting TOX formation from chlorination was developed using 49 model compounds and four constitutional descriptors. The model showed high predictive power on external validation data. However, predictions of TOX formation for three compounds in the external validation data may not be reliable because they were out of the applicability domain (predicted due to extrapolation of the QSPR). The prediction of TOX formation of tannic acid fell within the range of TOX formation reported in raw waters in United States and Canada.

4.5. References

1. Chaidou, C.; Georgakilas, V.I.; Stalikas, C.; Saraçi, M.; Lahaniatis, E.S. Formation of chloroform by aqueous chlorination of organic compounds. *Chemosphere.* **1999**, 39(4), 587-594.

2. Galal-Gorchev, H. Chlorine in water disinfection. *Pure Appl. Chem.* **1996**, 68(9), 1731-1735.

3. ACC (American Chemistry Council). *The benefits of chlorine chemistry in water treatment.* 2008: 1-13.

4. PNL (Pacific Northwest Laboratories). *Disinfection technologies for potable water and wastewater treatment: Alternatives to chlorine.* Pacific Northwest Laboratories: 1998.

5. Armstrong, G.L.; Conn, L.A.; Pinner, R.W. Trends in infectious disease mortality in the United States during the 20th century. *J. Am. Med. Assoc.* **1999**, 281(1), 61-66.

6. CDC (Center for Disease Control and Prevention). Achievements in Public Health, 1900-1999: Control of infectious diseases. *Morbid. Mortal. Wkly Rep.* **1999**, 48(29), 621-629.

7. Bellar, T.L.; Lichtenberg, J.J.; Kroner, R.C. The Occurrence of organohalides in chlorinated drinking waters. *J. Am. Water Works Assoc.* **1974**, 66(12), 703-706.

8. Rook, J.J. Formation of haloforms during chlorination of natural waters . *Wat. Treat. Exam.* **1974**, 23(2), 234-243.

9. IPCS (International Program for Chemical Safety). *Environmental Health Criteria 216 Disinfectants and disinfectant byproducts*. WHO, Geneva: 2000.

10. Onstad, G.D.; Weinberg, H.S.; Stuart, W.K. Occurrence of halogenated furanones in U.S. drinking waters. *Environ. Sci. Technol.* **2008**, 42 (9), 3341-3348

11. Crittenden J.C.; Trussell R.R.; Hand, D.W.; Howe, K.J.; Tchobanoglous, G. *Water treatment: Principles and design,* 2nd edn. John Wiley & Sons Inc., New York: 2005

12. Krasner, S.K.; Weinberg, H.S.; Richardson, S.D.; Pastor, S.J.; Chinn, R.; Sclimenti, M.J.; Onstad, G.D.; Thruston Jr, A.D. Occurrence of a new generation of disinfection byproducts. *Environ. Sci. Technol.* **2006**, 40 (23), 7175-7185.

13. Stevens, A.A.; Dressman, R. C.; Sorrell, R. K.; Brass, H. J. T*OX (total organic halogen) is the non-specific parameter of the future?* Municipal

Environmental Research Laboratory, US Environmental Protection Agency, Cincinnati, OH: 1984

14. Reckhow, D.; Hua, G.; Kim, J.; Hatcher, P.; Caccamise, S.; Sachdeva, R. *Characterization of total organic halogen produced during disinfection processes*. AWWA Research Foundation Report 91176: 2008.

15. Korshin, G.V., Li, C-W; Benjamin, M.M. The decrease of UV absorbance as an indicator of TOX formation. *Water Res.* 1997, 31(4), 946-949.

16. Li, C.; Benjamin, M.M.; Korshin, G.V. The relationship between TOX formation and spectral changes accompanying chlorination of pre-concentrated or fractionated NOM. *Water Res.* **2002**, 36, 3265-3272

17. Weishaar J.L.; Aiken, G.R.; Bergamaschi, B.A.; Fram, M.S.; Fujii, R.; Kenneth M. Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. *Environ. Sci. Technol.* **2003**, 37(20), 4702-4708.

18. Jung, C-W.; Son, H-J. The relationship between disinfection by-products formation and characteristics of natural organic matter in raw water. *Korean J. Chem. Eng.* **2008**, 25(4), 714-720.

19. Yang, X.; Shang, C.; Lee, W.; Westerhoff, P.; Fan, C. Correlations between organic matter properties and DBP formation during chloramination. *Wat. Res.* **2008**, 42, 2329-2339.

20. Clesceri, L.S., Greenberg, A.E. and Eaton, A.D. (eds). *Standard methods for the examination of water and wastewater*, 20th edn, APHA, Washington DC: 1998.

21. USEPA (US Environmental Protection Agency). *National primary drinking water regulations: Stage 2 Disinfection and disinfection byproducts:* Final Rule: 2006.

22. Ates, N.; Kitis, M.; Yetis, U. Formation of chlorination by-products in waters with low SUVA--correlations with SUVA and differential UV spectroscopy. *Water Res.* **2007**, 41(18), 4139-48.

23. Reckhow, D.A.; Rees, P.L.; Nusslein, K.; Makdissy, G.; Devine, G. *Long-term variability of BDOM and NOM as precursors in watershed sources.* American Water Works Association: 2007.

24. Chow, A.T.; Dahlgren, R.A.; Zhang, Q.; Wong, P.K. Relationships between specific ultraviolet absorbance and trihalomethane precursors of different carbon sources. *J. Water Supply: Res. T.* **2008**, 57(7), 471-480.

25. Wei, Q.; Feng, C.; Wang, D.; Shi, B.Y.; Zhang, L.T.; Wei, Q.; Tang, H.X. Seasonal variations of chemical and physical characteristics of dissolved

organic matter and trihalomethane precursors in a reservoir: a case study. *J. Hazard. Mater*. **2008**, 150(2), 257-264.

26. Bull, R.; Reckhow, D.; Rotello, V.; Bull, O.M.; Kim, J. *Use of toxicological and chemical models to prioritize DBP research.* American Water Works Association: 2006

27. Dickenson, E.R.; Summers, R.S.; Croué, J.; Gallard, H. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol.* **2008**, 42(9), 3226-3233.

28. Hureiki, L.; Croue, J.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res.* 1994, 28(12), 2521-2531.

29. USEPA (US Environmental Protection Agency). Method for determination of total organic halides (TOX) in water sample, Method 9020B, Revision 2. US Environmental protection Agency, 1994. http://www.caslab.com/EPA-Methods/PDF/EPA-Method-9020B.pdf

30. Chu, W-H.; Gao, N-Y.; Deng, Y.; Dong, B-Z. Formation of chloroform during chlorination of alanine in drinking water. *Chemosphere*. **2009**, 77, 1346-1351.

31. Snoeyink, V.L.; Weber Jr., W.J.; Mark Jr., H.B. Sorption of phenol and nitrophenol by active carbon. *Envlron. Sci. Technol.* **1989**, 3(10), 918-926

32. Yonge, D.R.; Kelnath, T.M.; Poznanska, K.; Jiang, Z.P. Single-solute irreversible adsorption on granular activated carbon. *Envlron. Sci. Technol.* **1985**, 79, 690-894

33. Grant, T.M.; King, C.J. Mechanism of irreversible adsorption of phenolic compounds by activated carbons. *Ind. Eng. Chem. Res.* **1990**, 29, 264-271

34. Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry*. **2005**, 76(2), 319-347.

35. Cabaniss, S.E. Forward modeling of metal complexation by NOM: I. A priori prediction of conditional constants and speciation. *Environ. Sci. Technol.* **2009**, 43(8), 2838-2844.

36. Katritzky, A.R; Lobanov, V.S. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, 24, 279-287.

37. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSPR/QSAR. *Chem. Rev.* **1996**, 96(3), 1027-1044

38. Minitab Inc. Graphical data: Meet Minitab 15, p. 2-1 to 2-13. 2007. http://www.minitab.com.

39. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environ. Health Perspect.* **2003**, 111(10), 1361-1375.

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

41. Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Combin. Sci.* **2003**, 22(1), 69-77.

42. Arnold, W.A.; Bolotinm, J.; von Gunten, U.; Hofstetter, T.B. Evaluation of functional groups responsible for chloroform formation during water chlorination using compound specific isotope analysis. *Environ. Sci. Technol.* **2008**, 42(21), 7778-7785

43. Urano, K.; Yamamoto, E.; Tonegawa, M.; Fujie, K. Adsorption of chlorinated organic compounds on activated carbon from water, *Water Res.* **1991**, 25(12), 1459-1464

44. Garcia-Araya, J.F.; Beltran, F.J.; Alvarez, P.; Masa, F.J. Activated carbon adsorption of some phenolic compounds present in agroindustrial wastewater. *Adsorption.* **2003**, 9, 107-115,

45. Dąbrowski, A.; Podkościelny, P.; Hubicki, Z.; Barczak, M. Adsorption of phenolic compounds by activated carbon-a critical review. *Chemosphere*. **2005**, 58(8), 1049-1070

46. Freund**,** J.R.; Wilson**,** W.J.; Sa**,** P. Regression analysis: statistical modeling of a response variable, 2nd edition. Academic Press, 2006

47. Bond, T.; Henriet, O.; Goslan, E.H.; Parsons, S.; Jefferson, B. Disinfection byproduct formation and fractionation behavior of natural organic matter surrogates. *Environ. Sci. Technol.* **2009**, 43(15), 5982-9.

48. Leenheer, J.A.; Brown, G.K.; MacCarthy, P.; Cabaniss, S.E. Models of metal binding structures in fulvic acid from the Suwannee River, Georgia. *Environ. Sci. Technol.* **1998**, 32(16), 2410-2416.

49. Douglas, J.G.; Meznarich, H.K.; Olsen, J.R.; Ross, G.A.; Stauffer, M. *Investigation of the total organic halogen analytical method at the waste sampling and characterization facility.* Prepared for the U.S. Department of Energy Assistant Secretary for Environmental Management, EDC #: HNF-EDC-08-39196: 2008

CHAPTER 5

QSPR FOR PREDICTING TCM FORMATION

Abstract

Chlorination is the most widely used technique of water disinfection but leads to formation of chloroform (trichloromethane or TCM) and other disinfection by-products. This work reports the first quantitative structure-property relationship (QSPR) for predicting TCM formation in chlorinated drinking water. Model compounds (N = 117) drawn from ten literature sources were divided into training data (N = 90, analyzed by 5-way Leave-Many-Out internal cross-validation) and external validation data (N = 27). The QSPR internal cross validation had q^2 = 0.94 and root mean square error (RMSE) of 0.09 moles TCM per mole compound (Cp), consistent with external validation q^2 of 0.94 and RMSE of 0.08 mol-TCM/mol-Cp, and met criteria for high predictive power and robustness. In contrast, the QSPR using a logTCM performed poorly and did not meet criteria of predictive power. The QSPR predictions are consistent with experimental values for TCM formation from tannic acid and for model fulvic acid structures. The descriptors used are consistent with a relatively small number of important TCM precursor structures based upon 1,3-dicarbonyls or 1,3-dihydroxybenzenes.

5.1. Introduction

Most water treatment plants in the US and European countries use chlorine for chemical disinfection of water (*1*). Chlorine is often chosen because it is the cheapest of all known disinfectants in terms of investment and operating costs (*2*).While chlorine has contributed to the decline in water borne disease outbreaks (*3.4*), there has been growing public health concern over the products from the reaction of chlorine with dissolved organic matter (DOM) during drinking water treatment. Trihalomethanes (THMs) were the first class of compounds discovered (*5,6*) and since then more classes have been identified including trihaloacetic acids (THAAs) and trihaloacetonitriles (THANs) (*7*), these are collectively referred to as disinfection byproducts (DBPs). The extent of the threat posed to human health by DBPs remains controversial. Laboratory tests have shown that DBPs, particularly THMs and THAAs, are carcinogenic or tumorgenic to test animals (*8*). Epidemiological studies have found potential risks of miscarriage and stillbirth at highest sextile of THMs and THAAs concentrations (*9,10,11*). However, inconsistencies in various epidemiological studies led the International Agency for Research on Cancer (IARC) to conclude that there is no proven link between drinking water DBPs and cancer incidence (*12,13*). The IARC conclusions were supported by a later study which took into account methodological problems in the previous studies (*14*).

Current modeling practices for DBP formation use bulk water quality parameters like dissolved organic carbon (DOC) concentration, pH and ultraviolet light absorbance at 254 nm (UV_{254}) (*8,15,16,17*). The US Environmental Protection Agency (USEPA) Stage 2 Disinfection Byproduct Rule mandates

monitoring of bulk parameters as indicators of DBP formation potential for all drinking water utilities (*18*). Empirical models based on bulk water parameters have been summarized in the literature (*19,20*), and UV_{254} and DOC have been reported to have linear relationships with DBP formation (*19,21,22,23*). Nonetheless, these linear relationships between DOC or UV_{254} and DBP formation do not always hold (*24,25,26,27*) and no structural or functional group information on individual molecules can be derived from the bulk parameters. No model has been reported based on structural properties of individual molecules.

Combining an agent-based model of DOM chemistry with quantitative structure-property relationships (QSPRs) using constitutional descriptors enables a structure-based approach to this heterogeneous mixture (*28,29,30*). The agent-based model provides structural information for hypothetical DOM mixtures based on ecosystem and environmental factors (*28*). QSPRs using constitutional descriptors are rapidly and easily computed for each molecule in the mixture (DOM may include tens of thousands of structures).

The objective of this work is to develop a new and robust QSPR that predicts chloroform formation from constitutional descriptors and to define its applicability domain.

5.2. Methodology

5.2.1. Data source

This research work is based on data mining and utilized experimental data reported in ten publications (*31,32,33,34,35,36,37,38,39,40*). TCM formation ranged from <0.001 mol-CM/mol-Cp for aliphatic amino acids to >0.80 mol

TCM/mol compound for aromatic compounds with at least one 1,3-dihydroxy substituted aromatic ring and aliphatic β-diketones. The average TCM formation for the entire data set was 0.177 mol/mol with standard deviation (Stdev) of 0.318 mol-TCM/mol-Cp (TCM formation data in Tables S5.1, S5.2, S5.3). The high Stdev is attributed to a wide range in TCM formation from different compounds and the distribution of TCM formation data (Figure 5.1) were skewed to the low end. Ninety three compounds (79.5%) had low TCM formation of 0.0001-0.1400 TCM/mol-Cp, 20 compounds (17.1%) had high TCM formation of 0.680-1.140 TCM/mol-Cp and 4 compounds (3.4%) had TCM formation of 0.280-0.320 TCM/mol-Cp.

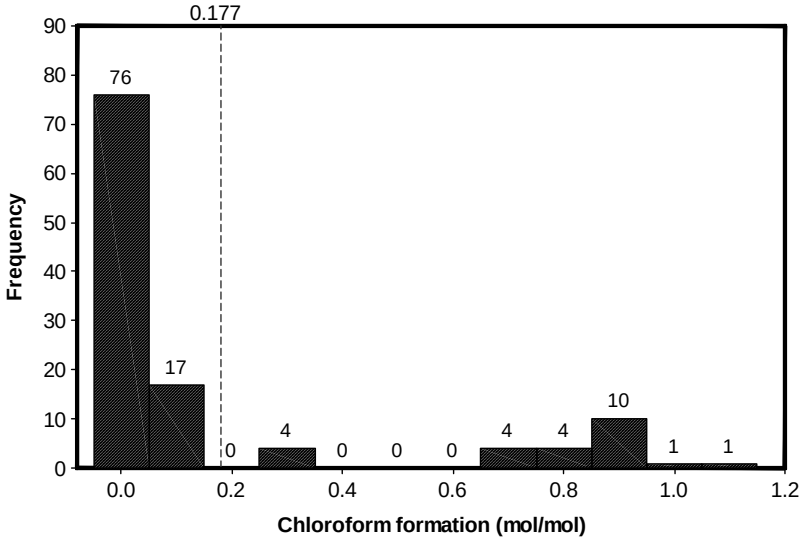


Figure 5.1. Frequency distribution of TCM formation (N = 117)

5.2.2. Generation of descriptors

The 117 data collected were divided into model training data (N = 90) and external validation data (N = 27). Constitutional descriptors were calculated by counting the number of atoms and functional groups for each of the 90 model

compounds. Compounds lacking a certain atom or group were given a value of zero for that descriptor. A descriptor was considered only if it had non-zero values for at least five compounds. Three were calculated differently from the rest. These were carbonyl index (CI), ring activation index (RAI) and both them have been used in previous work (*30*) and EDCORH and also were discussed in Chapter 2.

Carbonyl index (CI) relates to the lower pKa of hydrogen in a C-H bond located between two carbonyls (e.g., β-diketones) relative to other C-H bonds adjacent to a single carbonyl (*41,42*). The hydrogen can easily be abstracted by a base in solution to form keto-enol (or keto-enolate) tautomers and the enol form is the one that contributes to higher halogen substitution reaction (*41*). This concept was extended to carbons located between two hidden carbonyls (phenols) as in resorcinol. A compound that had at least one unsubstituted carbon in between 1,3-hidden dicarbonyls in aromatic molecules (e.g., 1,3-dihydroxybenzene) or 1,3-diketone in aliphatic molecules (e.g., 5,7-dioxooctanoic acid) was assigned a CI value of 2. A carbon between an aliphatic ester or acid carbonyl and keto group in 1,3 positions was assigned a CI value of 1.5 whereas a terminal carbon adjacent to a keto group in 1,3-dicarbonyls (e.g., 4,6-dioxoheptanedioic acid) is given a CI value of 1. A non-terminal carbon adjacent to a keto group in 1,3-dicarbonyl (3-oxohexanedioic acid) was given a value 0.5 and the same value was assigned for acetophenone and acetone or in molecules with carbonyls separated by more than two carbonyls (e.g., 4-oxoheptanedioic acid).

Ring activation index (RAI) reflects the fact that reactivity of aromatic molecules is influenced not only by the number of strong electron donating groups (OH and NH_2) in the molecule but also their relative position from each other. A 1,3-disubstitution is more reactive than 1,2 or 1,4-disubstitution due to cooperative effects in the former and antagonistic effect in the latter (*42*). However, it was also observed from the chlorine demand data that 2-hydroxy (or amino) benzoic acid consumed less chlorine than corresponding meta and para isomers. The difference in reactivity could be attributed to higher stability of 2-hydroxy(amino)benzoic acid over the meta and para isomers due to hydrogen bonding effects (*43,44,45*). Thus, RAI was calculated by taking the ratio of number of OH or NH_2 to the number of aromatic rings: if this ratio is 1, RAI was equal to 1; when the ratio was 2, RAI was set 0.6; when ratio was 3, RAI was set 0.5 (*30*). For a ratio of 1, RAI was reduced to 0.75 for 1,2 substitution with intramolecular H-bonding (e.g., 2-hydroxy benzoic acid). For ratio of 2 or 3, the RAI was reduced to 0.3 or 0.25 for ortho or para substitution, respectively. Molecules with only alkoxy groups, which are weak ring activators, had RAI set equal to 0.1. Aromatic molecules with only deactivating groups (e.g., acetophenone, 2-nitrobenzoic acid) and all aliphatic compounds were assigned RAI equal to zero (*30*) and in Chapter 2.

EDCORH is the difference between the sum of strong electron donating groups per carbon (ArED:C) and the sum of carbonyls (ketone and aldehydes) per carbon (CORH:C) in each molecule.

The number of one-three activated carbons (OTactC) in aromatic compounds was motivated by the observation that molecules with 1,3-disubstituted aromatic molecules (e.g., 1,3-dihydroxybenzene or 1,3-dihydroxy naphthalene) had higher experimental TCM formation than those with 1,2 or 1,4-disubstituted aromatic molecules (*31,32,33,39,42,46*). If a molecule had at least one 1,3 activated carbon, it was given a value of 1 and if it did not, it was given a value of zero. Molecules like 3-chlorophenol, 2,4,5-trichlorophenol and 2,3,4,6-tetrachlorophenol with at least one chlorine at 1,3-disubstution with OH were given a value of 1 as well.

5.2.3. Descriptor selection

Descriptor selection was carried out using Minitab 15 Software (*47*) by performing multiple regression (MLR) of TCM formation (as dependent variable, N = 90) on descriptors listed in Table 2.2 (as independent variables) generated using Equation 2.1. The y-intercept was set to zero and confidence interval was set at 95%. Any descriptor with $p<0.05$ was dropped and the process was repeated until all remaining descriptors had $p<0.05$. Three descriptors were finally obtained from the selection process. This was followed by correlation analysis in order to determine if indeed the descriptors were mutually independent since $r < 0.7$ (*48*).

5.2.4. Model calibration and validation

The training data (N = 90) were split in a 5-way Leave-Many-Out (LMO) design with 5 calibration data sets (N = 60) and cross validation data sets (N = 30). QSPR was calibrated by performing multiple linear regression of TCM

formation against the three descriptors (with intercept set to zero) for each of the five calibration data sets using Analysis ToolPak for MS Excel™ (Windows XP). The coefficient of determination for model calibration (R_c^2) is obtained by using Equation 2.2. Average statistics of fit, descriptor coefficients and standard errors were obtained for each of the five QSPRs.

Each of the five LMO QSPRs was validated using the cross validation data and the average QSPR was validated internally and externally. The predictive power of the model was evaluated using statistical and graphical methods. There are four statistical indicators of predictive power which are q^2, R_c^2, R_t and k (*49,50*) and the equations for calculating these statistics are described in Chapter 2 (Section 2.3.3).

Applicability domain analysis was performed to establish the range of applicability of the model and determine influential compounds. Leverage (h) and standardized residual (SDR) are the two parameters used to evaluate applicability domain (*51*). Leverage, a measure of how far an observation is from the neighboring means, is calculated by Equation 2.8.

5.3. Results and discussions

5.3.1. Model calibration

The process of descriptor selection gave three significant descriptors: the carbonyl index (CI), number of 1,3 activated aromatic carbons (OTactC), and the electron donor variable (EDCORH). Correlation analysis of the three descriptors showed that each pair of descriptors had $r < 0.7$ (Table 5.1) indicating that they were mutually independent (*48*). The five LMO QSPR equations are given in

Table S5.4**,** while the average QSPR is given in Table 5.2. CI is the best determined descriptor with a relative error of 5.71%. Average statistics of fit for the QSPR were: R_c^2 = 0.94, $AdjR_c^2$ = 0.92 and SDE = 0.08 mol-TCM/mol-Cp. The high R_c^2 and a low SDE of regression indicate good predictive power.

Table 5.1. Correlation matrix of the three descriptors

	CI	OTactC	EDCORH
CI	1.000		
OTactC	0.617	1.000	
EDCORH	0.495	0.008	1.000

Table 5.2. Average QSPR for TCM formation

Descriptor, x_j	Coefficient, β_j	Standard error, ε_j	$(\varepsilon_j/\beta_j)*100$
CI	0.275	0.016	5.7%
OTactC	0.266	0.036	13.6%
EDCORH	0.352	0.095	27.1%

5.3.2. QSPR LMO cross validation

The five LMO cross validations gave average statistics of q^2_{LMO} = 0.91, R_t = 0.008, k = 0.90 and k' = 0.94 (Table S5.5). All these parameters exceed threshold values for QSPR predictive power (*49,50*). The average RMSE of cross validation of 0.09 mol-TCM/mol-Cp is similar to model calibration error of 0.08 mol-TCM/mol-Cp, whereas MBD was 6.2% indicating the model predicts high for some compounds. TCM formation for all 90 compounds in the training data set was predicted using the average QSPR and results were: q^2_{cv} = 0.92, R_t = 0.007, k = 0.91 and k' = 0.95 (Table S5.5) with RMSE and MBD of 0.09 TCM/mol-Cp and 6.2 % respectively. These statistics were comparable to those obtained by the five LMO cross validations (Table S5.5). Overall, the QSPR had high predictive power based on LMO cross validation.

Graphical analysis of the plot of predicted versus observed TCM (Figure 5.2) and standardized residuals (Figure 5.3) shows that three data points were >2SDE (2 x 0.08 or 0.16 TCM/mol-Cp in this case) from the QSPR prediction. Two data points for 2-oxobutanedioic acid and one for 4-oxoheptanedioic acid were outliers.

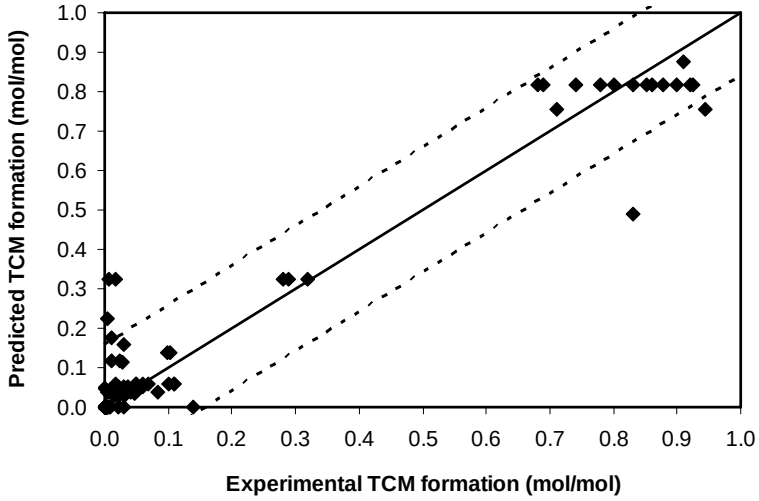


Figure 5.2. Predicted versus observed TCM formation for external validation data (N = 90). Dashed lines are ±2SDE and the solid line represents 1:1 prediction.

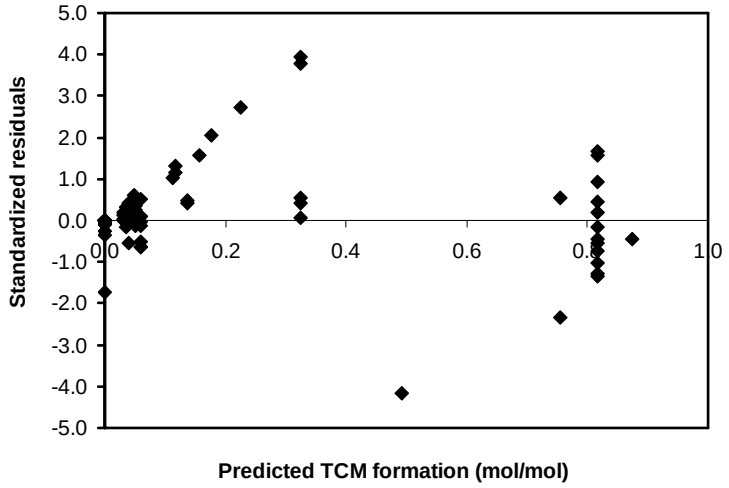


Figure 5.3. Standardized residuals of cross validation using average QSPR (N = 90)

5.3.3. QSPR external validation

The external validation set, never used in significant descriptor selection or model calibration, contained 27 compounds. TCM formations for these compounds were predicted using each individual QSPR and the average QSPR (Table S5.6), with comparable statistics of predictive power: q^2_{ext} = 0.94, R_t = 0.0001, k = 0.85 and k' = 0.85 (Table S5.6). The external calibration RMSE of 0.08 mol-TCM/mol-Cp was comparable to the 0.09 mol-TCM/mol-Cp obtained from LMO cross validation (Table S5.5) and SDE of 0.08 mol-TCM/mol-Cp obtained from model calibration (Table S5.4).

The MBD on external TCM data was -13.40% which indicates a bias toward underprediction, contrary to the overprediction bias in the internal cross-validation. The external validation bias was strongly influenced by a single compound, 2,4,6-trihydroxybenzoic acid, which had an absolute residual of -0.32 mol-TCM/mol-Cp (~SDR = -3.90) and was the only compound not predicted within ±2.5 standardized residuals (Figures 5.4 and 5.5).

It is not clear whether the underprediction of 2,4,6-trihydroxybenzoic acid is due to experimental error. 2,4,6-trihydroxybenzoic acid and phloroglucinol differ by a single carboxylic acid group, but the former has TCM formation potential of 1.14 mol-Cp/mol-Cp, much higher than 0.86 mol-TCM/mol-Cp reported for phloroglucinol in the same study (*32*) and 0.92 mol-TCM/mol-Cp in other work (*54,55*).

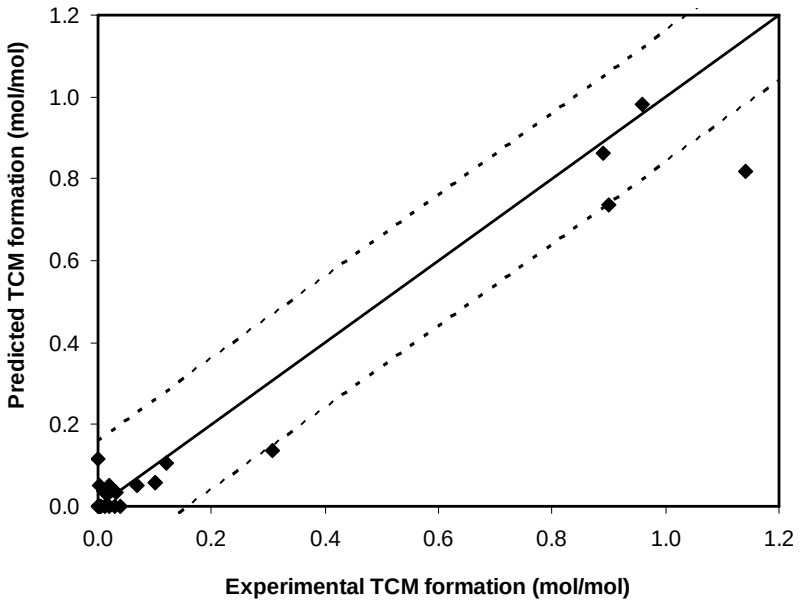


Figure 5.4. Predicted versus observed TCM formation for external validation data (N = 27). Dashed lines are ±2SDE and the solid line represents 1:1 prediction.

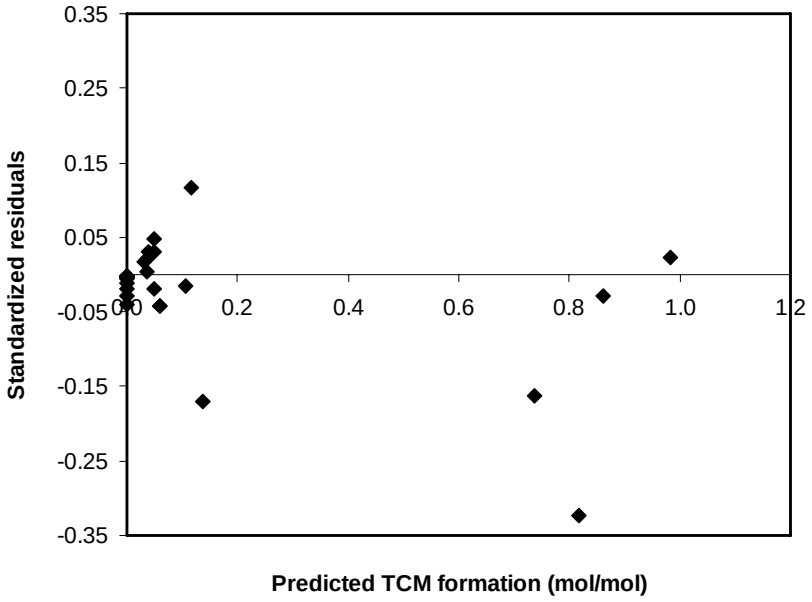


Figure 5.5. Standardized residuals of external validation using average QSPR (N = 27).

The reported TCM yield for 2,4,6-trihydroxybenzoic acid was also higher than for related resorcinols, which had TCM yields of about 0.87 mol-TCM/mol-Cp for 2,4-dihydroxybenzoic acids (*33,38,55*); about 0.83 mol-TCM/mol-Cp for 2,6-dihydroxybenzoic acids (*38,55*); about 0.75 mol-TCM/mol-Cp for 3,5-dihydrobenzoic acid (*33,38,55*); about 0.80 mol-TCM/mol-Cp for 3,5-hydoxytolune (*33,38,55*); about 0.85mol-TCM/mol-Cp for 1,3-dihydroxybenzene (*31,33,39,55*). Despite slight differences in experimental conditions the TCM formation potentials for resorcinols were closer to that of phloroglucinol because of these two groups of compounds had at least one pair of 1,3 hidden dicarbonyls and one 1,3-activated carbon. The presence of carboxylic acid may not likely increase TCM in phloroglucinol and thus the data point may have some errors.

5.3.4. QSPR model applicability domain

The model applicability domain was determined in order to assess if any of the training data were outliers (unduly influential in model calibration) and if any external data were prediction outliers (predicted due to extrapolation beyond the calibration data). The warning leverage h* for a 3-descriptor model and 90 training data was 0.13 (Figure 5.6). The training compounds 3-oxohexanedioic acid, 2-oxobutanedioic acid and 4-oxoheptanedioic acid fell outside the ±2.5 SDR boundary, indicating that they were training outliers. Deletion of these three compounds from training data (N = 87) did not cause a significant change in descriptor coefficients (relative to N = 90) (Table S5.7). The q^2 and RMSE for internal validation changed slightly from 0.92 and 0.09 mol-TCM/mol-Cp (N = 90)

to 0.96 and 0.06 TCM/mol-Cp (N = 87) respectively but there were no changes in q^2 and RMSE for external validation (Table S5.7). The analysis of outliers and influential data points showed that Cook's distances less was than 1, DFFITS<2 and DFBETAS<2 for the three data points. Therefore these data points were retained in the training data set because they were not considered outliers and also not influential (h < 0.13).

There were also five compounds in training data set with h>0.13 (Figure 5.6): 5,7-dioxooctanoic acid (SDR = -0.44, h = 0.24); 3-oxopentanedioic acid (SDR = -2.34, h = 0.17; SDR = -0.55, h = 0.17); 1,2,3-trihydroxybenzene (SDR = 2.05, h = 0.24) and 1,4-dihydroxybenzene (SDR = 1.34, h = 0.17), These compounds were not considered problematic, as their leverage values were closer to h* and their SDR were all within ±2.5; all were retained in the training data set.

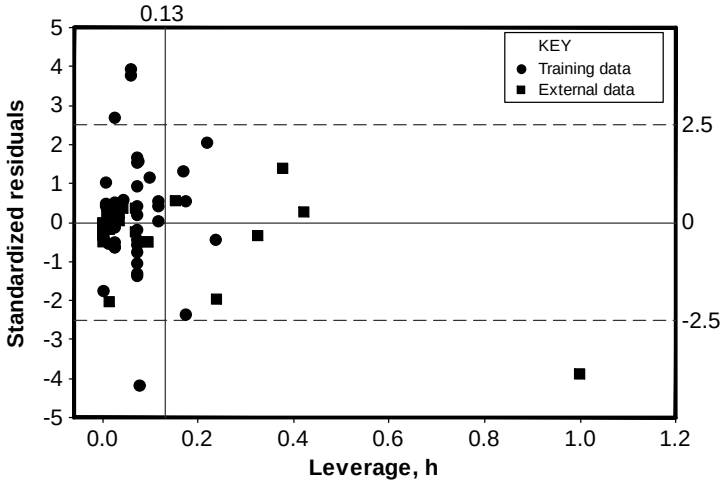


Figure 5.6. Williams plot indicating outliers and high leverage compounds

The Williams plot also revealed that five external compounds had h>0.13: 2,4,6-trihydroxybenzoic acid (SDR = -3.90, h = 1), 4,6-dioxoheptanedioic acid (SDR = -0.34, h = 0.32), 2,4-dioxopentane (SDR = 0.27, h = 0.42) and 1,4-

phenyldiamine (SDR = 1.40, h = 0.38) and 1,3-cyclohexanedioic acid (SDR = -1.96, h = 0.24). These compounds were extrapolated from the model and constitute 18.5% of the external data. Of these, only 2,4,6-trihydroxybenzoic acid (SDR = -3.90, h = 1) was not only over extrapolated but also a prediction outlier. As noted above, the experimental TCM for this compound seems questionable, but it was retained in the validation data set.

Several published models predict log-transformed formation potential from bulk water quality parameters (*59,60*), but we found that log-transformed TCM formation was less useful than the untransformed TCM potential in the QSPR modeling. The average QSPR from 5-way LMO had four descriptors (Hydrogen to carbon ratio, H:C; carbonyl index, CI; number of activating alkoxy groups, ArORact; and number of heteroatoms per carbon, HeA:C). The statistics of fit for the average QSPR were: R_c^2 = 0.75 and SDE = 0.55 log units (Table S5.8). Although R^2>0.6, the relative standard error of regression was exceedingly high (0.55 log units, greater than three-fold). Statistics of predictive power obtained from LMO cross validation using each QSPR were below the threshold values for QSPR, while those obtained using averaged QSPR for entire calibration data set were above the threshold values except k < 0.85 (Table S5.9). The q^2 on external validation fell short of other criteria of predictive power (Table S5.10) and external RMSE was very high (RMSE = 0.9 log units). Thus, log-transformation of TCM does not improve predictive performance of the QSPR.

Model deviations in predicting TCM formation may result from reactive intermediates and limitations of constitutional descriptors. For example,

chlorination of dicarbonyl compounds occurs through a nucleophilic keto-enol tautomer which forms several intermediate products (*41,56*). The rate of keto-enol tautomerization depends on both pKa of H and steric hindrance at α-C to keto (C=O) (*41*). In the absence of steric factors, the stability of the enol tautomer increases with substitutions at the C=C bond (E/Z-geometric isomers) and with the possibility of intramolecular hydrogen bonding via the O-H group (or resonance of enolate in case of (β-diketones) (41). Enol stabilization should increase rates of chlorination and formation of chloroform, which may be found in intermediates.

The pKa plugin in Marvin Sketch 4.1.1 software (57) was used to track changes in pKa during reaction sequences. For example, hydrogen atoms at carbon-2 and carbon-4 in 3-oxohexanedioic acid have pKa 19.2 and pKa 29.9 respectively. During stepwise chlorination, the pKa of the H on carbon-4 decreased from 29.9 to 17.2 after two chlorine atoms have been substituted onto carbon-2. This increases the probability of additional chlorination onto carbon-4, and thus increases the likelihood of TCM formation. On the other hand, substitution of chlorine at the carbon-3 position of 4-oxoheptanedioic acid followed by hydrolysis may lead to 2,2-dichlorobutane-1,4-dicarboxylic acid and 3,3,3-trichloropropanoic acid. Since neither compound is likely to undergo further chlorination, 4-oxoheptanedioic acid has low TCM formation in spite of the presence of multiple carbonyl groups. Full reaction-path simulation may lead to improved predictive capability for compounds like these.

5.3.6. Prediction of higher-MW compounds

Three model structures representing dissolved organic matter were examined to test the plausibility of the QSPR predictions for higher molecular-weight compounds of environmental relevance. Prediction of TCM formation from tannic acid was compared with experimental values (31), and TCM formation based on two proposed structures of fulvic acid (FA-1 and FA-2) (58) were compared with field results for DOM. Predicted TCM formation for tannic acid (MW = 1701) was 0.13 ± 0.03 mol-TCM/mol-Cp (~0.14 mmol-TCM/g-C), compared to an experimental yield of 0.05 mol-TCM/mol-Cp (~0.05 mmol-TCM/g-C) (*31*). While the relative error of this prediction is large, the error is small in absolute terms, comparable to the RMSE of QSPR calibration and validation (0.09 and 0.08 mol-TCM/mol-Cp, respectively). For FA-1 (MW = 960), predicted TCM was 0.13 ± 0.01 TCM/mol-Cp (~0.25 mmol-TCM/g-C) and for FA-2 (MW = 948) it was 0.27 ± 0.02 TCM/mol-Cp (~0.54 mmol-TCM/g-C). In comparison, trihalomethane production from field sampled DOM typically varies from ~0.15 to 0.35 mmol-TCM/g-C, depending on source (25). The field values overlap with but are somewhat lower than the QSPR/FA-based predictions; the higher estimates may be caused by high aromaticity of the FA-1 and FA-2 structures relative to bulk DOM samples.

Structural implications Figures 5.2 and 5.3 show three classes of compounds in the data set: 1) aliphatic compounds (amino acids, carboxylic acids, etc) and phenols without the β-diketone or ‘activated carbon’ structure showed lowest TCM formation potential (TCM<~0.2 mol-TCM/mol-Cp), 2)

chlorinated phenols showed intermediate TCM formation potential (0.20 <TCM< 0.32 mol-TCM/mol-Cp) and 3) 'activated' carbon (1,3-diphenols, β-diketones and their derivatives) showed high TCM formation potential (TCM > 0.50 mol-TCM/mol-Cp). Of the 90 compounds used in the model training, dihydroxy substituted aromatic compounds with hidden dicarbonyls and β-diketones contributed 83% of the total TCM formation. This is consistent with previous work emphasizing the importance of the activated carbon structures (*35,39,52*).

One consequence of this structural pattern is the relative simplicity of the QSPR for TCM production. Chlorine demand and TCM formation represent two different aspects of disinfection by-product formation: the reactant consumed and product created. Compared to a previously derived QSPR for chlorine demand30, the present QSPR for TCM formation has fewer descriptors (3 versus 8) and higher validation q^2 (0.94 versus 0.88). This improved performance with a simpler model arises from the small number of TCM producing structures relative to the larger number of structures which react with chlorine to form other byproducts.

A second consequence is that water treatment operators should be more concerned when source water has high proportion of the resorcinols (i.e., resorcinol and its derivatives) and β-diketones. Currently, DOC and specific ultraviolet absorbance (SUVA) are used as indicators of NOM reactivity (*21,52*), but cannot distinguish SUVA from different classes of compounds due to their spectral overlap. It has also been demonstrated that water with high SUVA may have low disinfection byproducts formation potential and vice versa (*25,26,27*).

Efficient water treatment requires a more specific method of evaluating and predicting TCM formation which can account variation in relative composition of more and less reactive sub-structures.

Apparent kinetic isotope effect is a promising alternative technique that has been reported for screening of functional groups responsible for TCM formation (*53*). Although the preliminary study used too few compounds to represent natural organic matter complexity, it presents an experimental avenue to improving mechanistic understanding and prediction of TCM formation due to resorcinols, β-diketones and phenols.

5.4. Conclusions

A robust QSPR for predicting chloroform formation from chlorination of model compounds (R_c^2 = 0.94, q^2_{ext} = 0.94, external validation RMSE = 0.08 mol-TCM/mol-Cp) uses only 3 constitutional descriptors. Model stability is limited by the quality of experimental data and by the number of different classes of model compounds represented in the training data. In terms of relative error, the model predicts β-diketones and resorcinols (higher TCM-forming compounds) more reliably than the lower TCM-forming compounds (amino acids, phenols, ketones, etc.). Arnold et al. (*53*) showed that β-diketones and resorcinols are not only very reactive toward chlorine substitution at room temperature but also give high TCM formation even at the WHO benchmark contact time of 30 minutes (*61*). However there is apparently no experimental technique that can quickly screen and discriminate presence of resorcinols and β-diketones from other compounds in drinking water sample.

5.6. References

1. Chaidou, C.; Georgakilas, V.I.; Stalikas, C.; Saraçi, M.; Lahaniatis, E.S. Formation of chloroform by aqueous chlorination of organic compounds. *Chemosphere*. **1999**, 39, 587-594.

2. ACC (American Chemistry Council). *The Benefits of chlorine chemistry in water treatment:* 2008, pp. 1-13. http://www.americanchemistry.com/s_chlorine/doc.asp?CID=1132&DID=8865&CTYPEID=10

3. Armstrong, G.L.; Conn, L.A.; Pinner, R.W. Trends in infectious disease mortality in the United States during the 20th century. *J. Am. Med. Assoc*. **1999**, 281(1), 61-66.

4. CDC (Center for Disease and Control and Prevention). Achievements in public health, 1900-1999: Control of infectious diseases. *Morbid. Mortal. Wkly Rep.* **1999**, 48(29), 621-629.

5. Bellar, T.L.; Lichtenberg, J.J.; Kroner, R.C.. The Occurrence of organohalides in chlorinated drinking waters. *J. Am. Water Works Assoc.* **1974**, 66, 703-706.

6. Rook, J.J. Formation of haloforms during chlorination of natural waters. *Wat. Treat. Exam.* **1974**, 23, 234-243.

7. Crittenden, J.C.; Trussell, R.R.; Hand, D.W.; Howe, K.J.; Tchobanoglous, C. *Water Treatment: Principles and Design,* 2nd Edn.; John Wiley & Sons Inc.,, New York: 2005.

8. IPCS (International Program on Chemical Safety). *Disinfectants and disinfectant byproducts*; *Environmental Health Criteria 216,* WHO: 2000. http://www.inchem.org/documents/ehc/ehc/ehc216.htm

9. Dodds, L.; King, W.; Allen, A.C.; Armson, B.A.; Fell, D.B.; Nimrod, C. Trihalomethanes in public water supplies and risk of stillbirth. *Epidemiology,* **2004**, 15, 179-186.

10. Savitz, D.A.; Andrews, K.W.; Pastore, L.M. Drinking water and pregnancy outcome in central North Carolina: Source, amount, and trihalomethane levels. *Environ. Health Perspect.* **1993**,103, 592-596.

11. Waller, K.; Swan, S.H.; DeLorenze, G.; Hopkins, B.. Trihalomethanes in drinking water and spontaneous abortion. *Epidemiology.* **1998**, 9, 134-140.

12. IARC (International Agency for Research on Cancer). *IARC Monographs on the evaluation of carcinogenic risks to humans,* volume 73, WHO: 1999.

13. IARC (International Agency for Research on Cancer). *IARC Monographs on the evaluation of carcinogenic risks to humans,* volume 52, WHO: 1991.

14. Savitz, D.A.; Singer, P.C.; Herring, A.H.; Hartmann, K.E.; Weinberg, H.S.; Makarushka, C. Exposure to drinking water disinfection by-products and pregnancy loss. *Am. J. Epidemiol.* **2006**,164, 1043-1051.

15. Golfinopoulos, S.; Arhonditsis, G. Multiple regression models: A methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. *Chemosphere.* **2002**, 47, 1007-1018.

16. Baxter, C.W.; Smith, D.W.; Stanley, S.J. A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. *J. Environ. Eng. Sci.* **2004**, 3, S45-S58.

17. Rodriguez, M.J.; Sérodes, J. Application of back-propagation neural network modeling for free residual chlorine, total trihalomethanes and trihalomethanes speciation. *J. Environ. Eng. Sci.* **2004**, 3, S25-S34.

18. USEPA (US Environmental Protection Agency). *National primary drinking water regulations: Stage 2 Disinfection and disinfection byproducts*: Final Rule. 2006.

19. Chowdhury, S.; Champagne, P. An investigation on parameters for modeling THMs formation. *Global NEST J.* **2008**, 10, 80-91.

20. Chowdhury, S.; Champagne, P.; McLellan, P.J. Models for predicting disinfection byproduct (DBP) formation in drinking waters: A chronological review. *Sci. Total Environ.* **2009**, 407, 4189-4206.

21. Chow, C.; Dexter, R.; Sutherland-Stacey, L.; Fitzgerald, F.; Fabris, R.; Drikas, M.; Holmes, H.; Kaeding, U. UV spectrometry in drinking water quality management. *Water.* **2007**, 40-43.

22. Fitzgerald, F.; Chow, C.; Holmes, M.. Disinfectant demand prediction using surrogate parameters - a tool to improve disinfection control. *J. Water Supply: Res. T.* **2006**, 55, 391-400.

23. Kitis, M.; Karanfil, T.; Kilduff, J.E.. The Reactivity of dissolved organic matter for disinfection byproduct formation. *Turkish J. Eng. Env. Sci.* **2004**, 28, 167-179.

24. Ates, N.; Kitis, M.; Yetis, U. Formation of chlorination by-products in waters with low SUVA-correlations with SUVA and differential UV spectroscopy. *Wat. Res.* **2007**, 41, 4139-4148.

25. Reckhow, D.A.; Rees, P.L.; Nusslein, K.; Makdissy, G.; Devine, G.. *Long-term variability of BDOM and NOM as precursors in watershed sources.* American Water Works Association: 2007.

26. Wei, Q.; Feng, C.; Wang, D.; Shi, B.Y.; Zang, L.T.; Wei, Q.; Tang, H.X. Seasonal variations of chemical and physical characteristics of dissolved

organic matter and trihalomethane precursors in a reservoir: a case study. *J. Hazard. Mater.* **2008**,150, 257-264.

27. Weishaar, J.L.; Aiken, G.R.; Bergamaschi, B.A.; Fram, M.S.; Fujii, R.; Mopper, K. Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. *Environ. Sci. Technol.* **2003**, 37, 4702-4708.

28. Cabaniss, S.E.; Madey, G.; Leff, P L.; Maurice, A.; Wetzel, R.A. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry*. **2005**, 76, 319-347.

29. Cabaniss, S.E. Forward modeling of metal complexation by NOM: I. A priori prediction of conditional constants and speciation. *Environ. Sci. Technol.* **2009**, 43, 2838-2844.

30. Luilo, G.B.; Cabaniss, S.E.. Quantitative structure-property relationship for predicting chlorine demand by organic molecules. *Environ. Sci. Technol.* **2010**, 44, 2503-2508.

31. Bond, T.; Henriet, O.; Goslan, E.H.; Parsons, S.; Jefferson, B. Disinfection byproduct formation and fractionation behavior of natural organic matter surrogates. *Environ. Sci. Technol.* **2009**, 43, 5982-5989.

32. Boyce, S.D.; Hornig, J.F. *Formation of chloroform from chlorination diketones and polyhydroxybenzenes in dilute aqueous solutions*. In: Water chlorination: Environmental impacts and health, Jolly, J.L.(ed.), Ann Arbor Science Publishers, Ann Arbor, Michigan: 1979, pp 131-140.

33. Boyce, S.D.; Hornig, J.F. Reaction pathways of trihalomethane formation from the halogenation of dihydroxyaromatic model compounds for humic acid. *Environ. Sci. Technol.* **1983**,17, 202-211.

34. Bull, R.; Reckhow, D.; Rotello, V.; Bull, O.M.; Kim, J. *Use of toxicological and chemical models to prioritize DBP research*. American Water Works Association: 2006.

35. Dickenson, E.R.; Summers, R.S.; Croué, J.; Gallard, H. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol.* **2008**, 42, 3226-3233.

36. Gallard, H.; von Gunten, U. Chlorination of phenols: Kinetics and formation of chloroform. *Environ. Sci. Technol*. **2002**, 36, 884–890.

37. Hureiki, L.; Croue, J.; Legube, B. Chlorination studies of free and combined amino acids. *Wat. Res.* **1994**, 28, 2521-2531.

38. Larson, R.A.; Rockwell, A.L. Chloroform and chlorophenol production by decarboxylation of natural acids during aqueous chlorination. *Environ. Sci. Technol.*. **1979**, 13(3), 325-329.

39. Norwood, D.L.; Johnson, J.D.; Christman, R.F.; Hass, J.R.; Bobenrieth, M.J.. Reactions of chlorine with selected aromatic models of aquatic humic material. *Environ. Sci. Technol.* **1980**,14, 187-190.

40. Hong, H.C.; Wong, M.H.; Liang, Y. Amino acids as precursors of trihalomethane and haloacetic acid formation during chlorination. *Arch. Environ. Contam. Toxicol.* **2009**, 56, 638-645.

41. McMurry, J. *Organic chemistry*. 5th edn. Thomson Brooks/Cole, Pacific Grove, California: 1999.

42. Reusch, W. Virtual Textbook of Organic Chemistry, 1999 (2008 revision). http://www.cem.msu.edu/~reusch/VirtualText/intro1.htm.

43. Aarset, K.; Page, E.M.; Rice, D. Molecular structures of benzoic acid and 2-hydroxybenzoic acid, obtained by gas-phase electron diffraction and theoretical calculations. *J. Phys. Chem.. A.* **2006**,110, 9014-9019.

44. Pinto, S.S.; Diogo, H.P.; Guedes, R.C.; Costa Cabral, B.J.; Minas da Piedade, M.E.; Martinho Simôes, J.A.. Energetics of hydroxybenzoic acids and of the corresponding carboxyphenoxyl radicals. Intramolecular hydrogen bonding in 2-hydroxybenzoic acid. *J. Phys. Chem.. A.* **2005**, 109, 9700-9708.

45. Stalin, T.; Rajendiran, N.. Intramolecular charge transfer associated with hydrogen bonding effects on 2-aminobenzoic acid. *J. Photochem. Photobiol. A: Chem.* **2006**,182, 137-150.

46. de Laat, J.; Merlet, N.; Dore. M. Chlorination of organic compounds: chlorine demand and reactivity in relationship to the trihalomethane formation. Incidence of ammoniacal nitrogen. *Wat. Res.* **1982**, 16, 1437-1450.

47. Minitab Inc. Graphical data: Meet Minitab 15, p. 2-1 to 2-13. 2007. http://www.minitab.com.

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

49. Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, 20, 269-276.

50. Tropsha, A.; Gramatica, P.; Gombar, V.. The Importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, 22, 69-77.

51. Herrell Jr, F.E. R*egression modeling strategies with application to linear models, Logistic regression, and survival analysis*. 1st edn, Springer-verlag, New York: 2001.

52. Reckhow D.A.; Singer, P.C.; Malcolm, R.L. Chlorination of humic materials: byproduct formation and chemical interpretations. *Environ. Sci. Technol.* **1990**, 24, 1655-1664.

53. Arnold, W.A.; Bolotin, J.; von Gunten, U.; Hofstetter, T.B.. Evaluation of functional groups responsible for chloroform formation during water chlorination using compound specific isotope analysis. *Environ. Sci. Technol*. **2008**, 42, 7778-7785.

54. Bove, F.; Shim, Y.; Zeitz, P. Drinking water contaminants and adverse pregnancy outcomes: a review. *Environ. Health Perspect*. **2002**, 110, 61-74.

55. de Leer, E.W.B.; Erkelens, C.. *Pathways for production of organochlorine compounds in the chlorination of humic materials*. In: Biohazards of water treatment Larson, R.A. (ed.), Lewis Publishers Inc., Chelsea, Michigan: 1989, pp 97-106.

56. Larson, R.A.; Weber, E.J.. *Reaction mechanisms in environmental organic chemistry*. Lewis Publishers, Boca Raton, FL: 1994.

57. ChemAxon Ltd. Marvin Sketch 4.1.1. 2006. http://www.chemaxon.com/marvin/sketch/index.php.

58. Leenheer, J.A.; Brown, G.K.; MacCarthy, P.; Cabaniss, S.E.. Models of metal binding structures in fulvic acid from the Suwannee River, Georgia. *Environ. Sci. Technol.* **1998**, 32, 2410-2416.

59. Lekkas, T.D.; Nikolaou, A.D. Development of predictive models for the formation of trihalomethanes and haloacetic acids during chlorination of bromide-rich water. *Wat. Qual. Res. J. Can*. **2004**, 39, 149-159.

60. Obolensky, A.; Singer, P.C. Development and interpretation of disinfection byproduct formation models using the information collection rule database. *Environ. Sci. Technol.* **2008**, 42, 5654-5660.

61. WHO (World Health Organization). *How to measure chlorine residual in water*. Technical note no 11. WHO; Geneva: 2006.

CHAPTER 6

QSPR FOR PREDICTING TCAA FORMATION

Abstract

In this work attempts were made to derive a quantitative structure-property relationship (QSPR) for predicting trichloroacetic acid (TCAA) formation using model compounds. A number of constitutional descriptors were used to derive QSPR for TCAA formation. None of the attempts were fruitful because the QSPRs did not meet all the requirements for QSPR predictive power. In the text five QSPRs are reported as examples of the failed QSPRs. All the QSPRs had high $q^2 > 0.5$ and $R_t < 0.1$ but ki and ko were below the accepted range of slope, k ($0.85 \leq k \leq 1.15$) and there was no significant linear relationship between predicted and experimental TCAA formation. There was linear relationship with $R_c^2 < 0.6$, for compounds with formation less than 0.2 mole TCAA per mole compound. From these results it was concluded that TCAA had weak linear relationship with constitutional descriptors. Trichlororoacetic acid formation may have a linear relationship with other descriptors that may include quantum-chemical, geometrical and electrostatic descriptors. Alternatively, the constitutional descriptors may be used to predict TCAA formation by QSPRs derived using non-linear algorithms. Alternatives to modeling TCAA formation from the molecular descriptors were beyond the scope of this study.

6.1. Introduction

Trichloroacetic acid is one the haloacetic acids formed when water is chlorinated to inactivate microbial pathogens (*1,2*). Over the years attempts have been made to develop models that would predict haloacetic acid and trihalomethane formation in drinking water. The models are generally based on bulk water quality and most recent models are summarized in two reviews (*3,4*). Most of the models reported were not tested for predictive power using independent data or surface waters. Since there is high variability in water quality parameters (UV_{254}, DOC, pH, turbidity, etc) from one water system to another and one season to another, water quality parameters may not always have linear relationship with disinfection byproducts formation (*5,6,7,8*). Use of molecular descriptors which may not change much with time and space would be useful for predicting disinfection byproducts formation. Of the known molecular descriptors reported in the literature (*9,10*), constitutional descriptors are easy to compute and fast because no molecular optimization and expensive software are required. There is no QSPR reported for TCAA formation so far and therefore the objective of this work is to develop a robust QSPR that predicts TCAA formation using constitutional descriptors.

6.2. Methodology

This work used 62 compounds obtained from three publications (*11,12,13*). TCAA formation ranged between 0.0001 mole of TCAA per mole compound (Cp) for amino acids and aliphatic compounds to 0.4760 mol-TCAA/mol-Cp for aromatic aldehydes and phenols (Table S6.1 and S6.2) and

Figure 6.1 shows the frequency distribution of the data. The constitutional descriptors listed in Table 2.1 were calculated and the whole data set was split by pseudo randomization into calibration data (N = 47) and external validation data (N = 15) given in Tables S6.1 and S6.2. Multiple linear regression (MLR) was used to select significant descriptors from the entire set of calibration data using Minitab 15 Software (*14*).

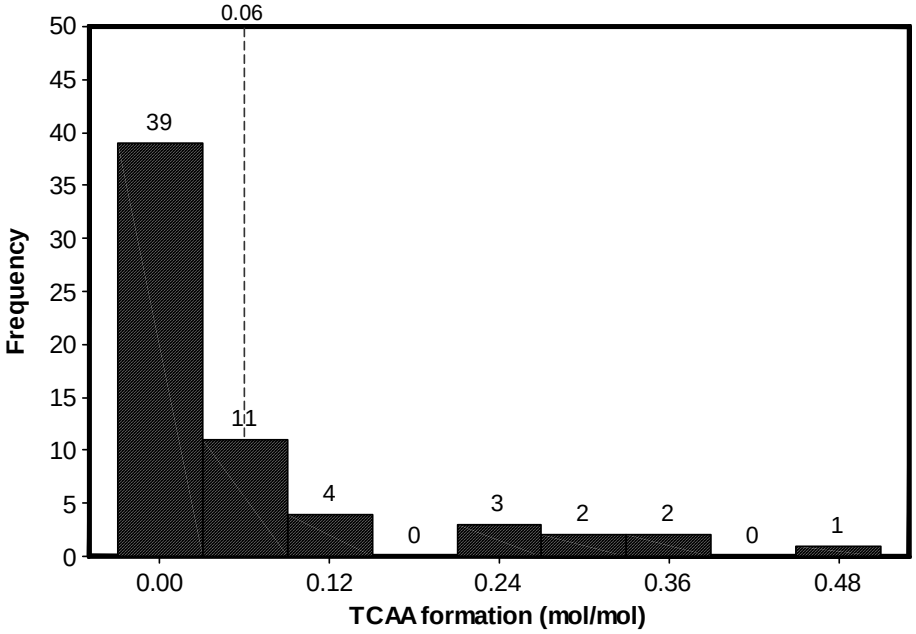


Figure 6.1. Frequency distribution of TCAA formation (N = 62)

A QSPR was derived using log-transformed TCAA formation and average for training data was -2.11 log units with standard deviation of 1.30 log units. The MLR of TCAA and four descriptors through origin was carried out using Equation 2.1 at 95% confidence interval using Analysis ToolPak for MS Excel^{TM} (Windows XP).

6.3. Results and Discussions

Equation 6.1 is the MLR-QSPR for log TCAA formation using hydrogen to carbon ratio (H:C), number of alkoxy groups on aromatic ring per carbon (ArOR:C) and number of strong electron donating groups on aromatic ring per carbon (ArED:C). The numbers in brackets are standard errors of coefficients.

$$\log TCAA_f = -1.72 * H:C(\pm 0.08) + 2.88 * ArOR:C(\pm 0.91) + 3.34 * ArED:C(\pm 0.94) \text{ .. (Eq. 6.1)}$$

The QSPR had R_c^2 = 0.74 and SDE = 0.68 log units. Although the QSPR had high R_c^2, the SDE indicates that the model may have poor performance because the SDE was a factor of 4X or 5X, much higher than experimental uncertainty.

The internal validation of the QSPR produced q^2= 0.74, k_i' = 1.00 and k_o' = 1.01, $R_t^{'}$ = 0.0001 and RMSE = 0.66 log units. Although these first four parameters met criteria for predictive power (*15,16*), the RMSE was about 4X. The close similarity of q^2 and RMSE to R_c^2 and SDE was expected because internal validation uses the same data set that used for QSPR calibration. The MBD 0.03% which implies the data points are more or less equally distributed around the ideal line in that the plot of predicted against experimental TCAA (Figure 6.2) whereas the standardized residual plot (Figure 6.3) showed that 3-hydroxy butyric acid (SDR = -2.67) was slightly outside the ±2.5 SDR margins.

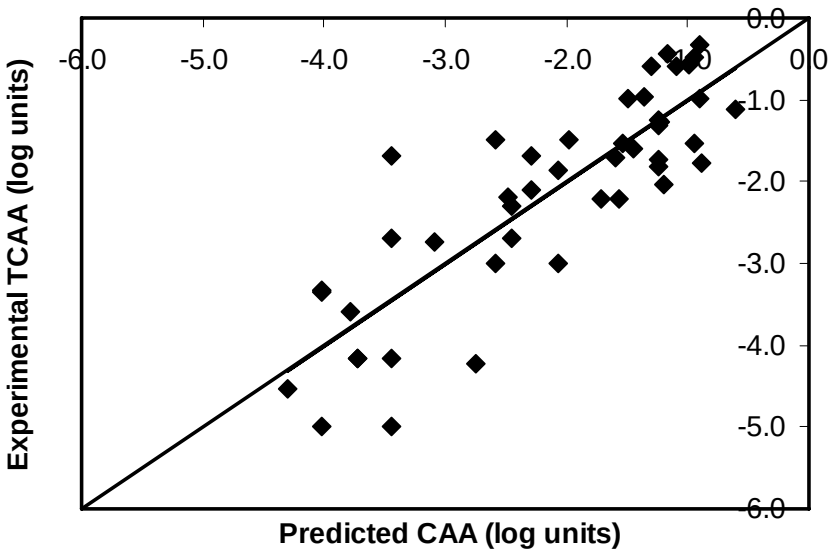


Figure 6.2. Deviation of predicted TCAA formation from ideal QSPR (N = 47)

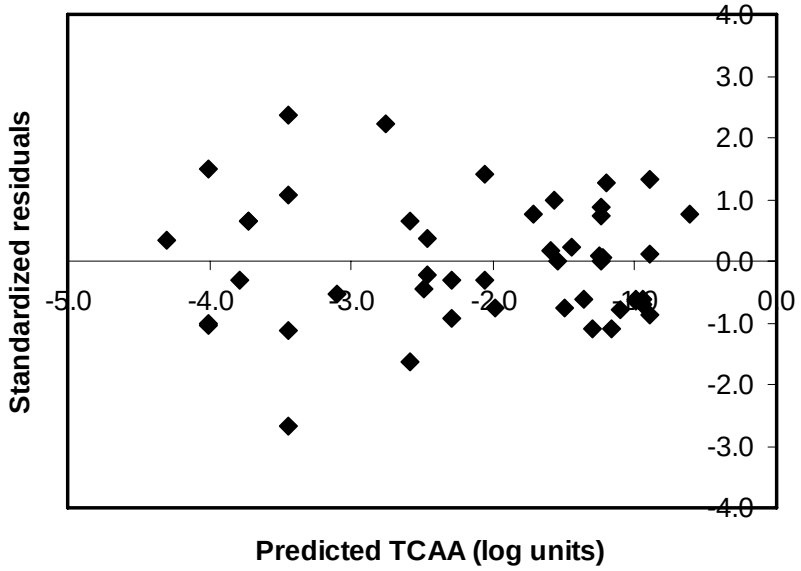


Figure 6.3. Standardized residual plot for logTCAA formation (N = 47).

The external validation data (Table S6.2) was used to test the performance of the QSPR. The statistics of predictive power showed that: q^2_{Ext} = 0.63, k_i = 0.97 and k_o = 1.00 and R_t = 0.003. These statistics meet minimum requirements for QSAR/QSPR predictive power but the RMSE was 0.51 log units

which is about 3X. The errors of 3 and 4 orders of magnitude are higher than the experimental TCAA formation of most compounds in the external validation data and calibration data respectively. The plot of predicted TCAA formation against experimental TCAA formation showed the relationship between the two was linear (Figure 6.4) and distribution of data points showed some pattern across the zero SDR (Figure 6.5) which suggests the QSPR lacked predictive power.

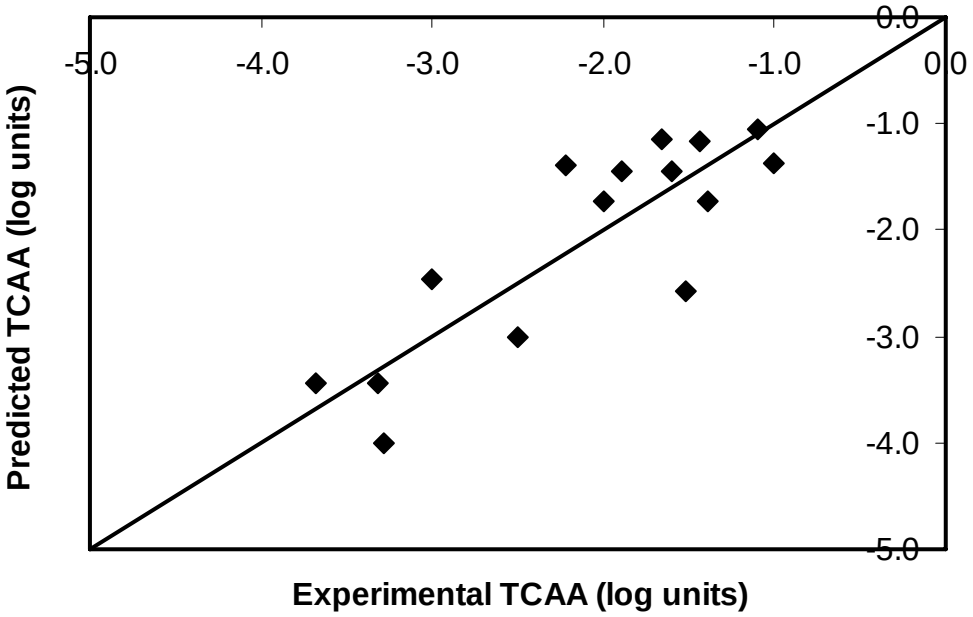


Figure 6.4. Deviation of the predicted logTCAA from ideal QSPR for external validation data

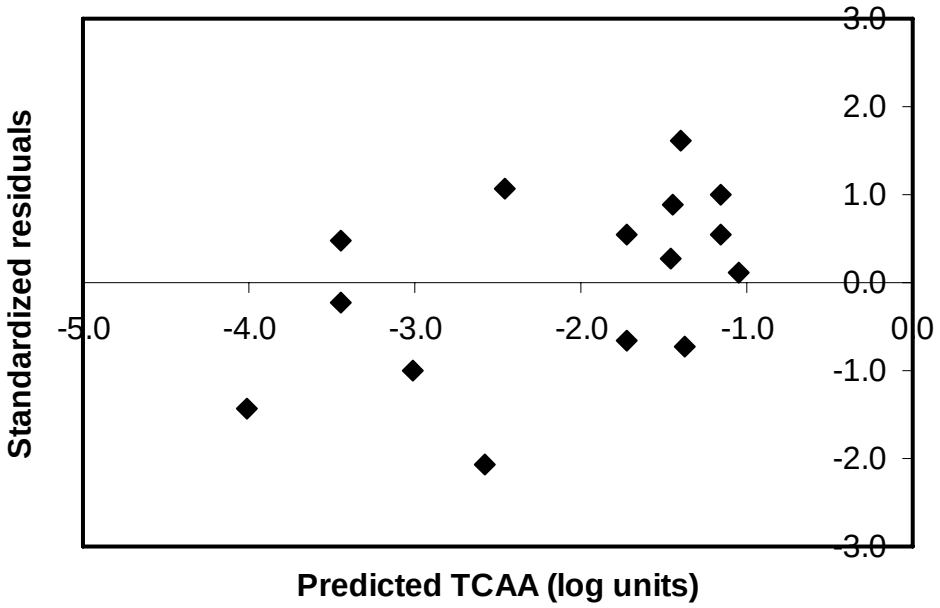


Figure 6.5. Assessment of model predictive power on external validation data using SDR plot

6.4. QSPR Applicability Domain

The model applicability domain of the QAPR was evaluated using William's plot to determine prediction outliers in external validation data and influential points in calibration data. Figure 6.6 shows that all compounds in calibration data were within applicability domain (SDR = ±2.5 and h ≤ 2.6) except 3-hydroxybutyric acid (SDR = -2.67, h = 0.05). This data point was not considered an outlier because its SDR was very close to the lower limit of SDR cutoff of -2.5. There were three compounds in external validation data set with h > 0.26, namely 2-oxopentanedioic acid (SDR = 1.64, h = 0.34), 3,4,5-trimethoxyacetophenone (SDR = -0.73, h = 0.41) and 3,4,5-trimethoxy-benzamide (SDR = 0.28, h = 0.39). These three data points were prediction outliers because h > h* and their predictions were due to extrapolation of the QSPR.

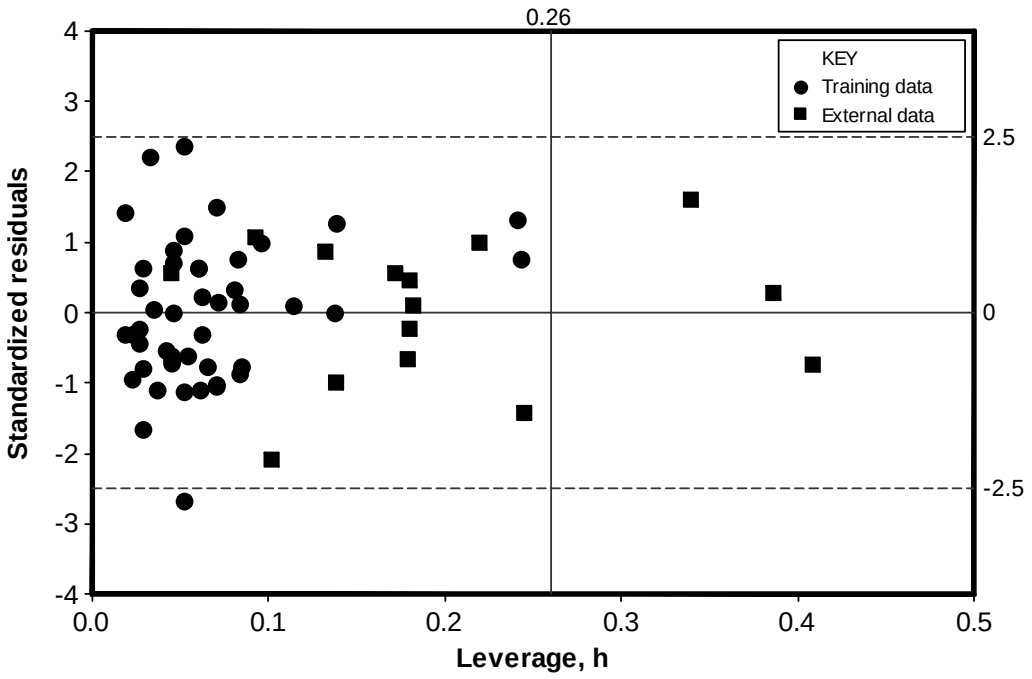


Figure 6.6. Williams plot for assessing outliers and influential compounds

6.5. Remarks on QSPR using logTCAA

The log-TCAA QSPR statistics of internal validation and external validation met criteria predictive power and Figures 6.1 and 6.3 were linear. These results suggest that model had high predictive power, which is not the case because the SDE and RMSEs were about 0.5-0.6 log units (factor of 3X to 4X). The SDE and RMSE error were higher than TCAA formation for most of the compounds in the data set. Predictions of compounds with TCAA formation less than RMSE or SDE are not reliable. The QSPR was not useful

Since TCAA formation are not measured in log units, the antilogs of predicted TCAA and observed TCAA were obtained and correlated in order to see if linear relationship would still hold. it was found that internal validation had q^2 = 0.22 and RMSE = 0.10 mol-TCAA/mol-Cp and external validation data had

q^2 = 0.44 and RMSE = 0.03 mol-TCAA/mol-Cp. The q^2 < 0.5 and the RMSE higher than experimental uncertainty in TCAA formation for most compounds in the training data confirms that the QSPR has low predictive power. This is a reminder that interpretation of results from the model that use log-transformed observation should be checked by taking antilog observed values in order to avoid misinterpretation of predictive power of the model.

6.6. Other QSPRs Attempted

Several further attempts were made to develop QSPRs using different combinations of constitutional descriptors and are reported here as examples. The first attempt used square root of TCAA (sqrtTCAA) as the dependent variable and number of phenols per carbon (ArOH:C), square root of number of hetero atoms (sqrtHeA) and square root of ring activation index (sqrtRAI) as independent variables (Eq 6.2). The numbers in brackets are standard errors.

$$sqrtTCAA_f = 0.83 * ArOH:C(\pm 0.29) + 0.03 * sqrtHeA(\pm 0.01) + 0.26 * sqrtRAI(\pm 0.05) \quad \text{(Eq. 6.2)}$$

Results showed that the QSPR had R_c^2 = 0.71 and SDE = 0.11 (mol-TCAA/mol-Cp)$^{1/2}$. Internal validation of square of predicted TCAA had q^2 = 0.54 and RMSE = 0.08 (mol-TCAA/mol-Cp)$^{1/2}$ (Table 6.1).

Table 6.1 Statistics of predictive power for the four QSPRs

	q^2	RMSE	R_i^2	k_i	b	R_o^2	k_o	R_t
Eq. 6.2	0.54	0.08	0.57	0.47	0.02	0.49	0.56	0.05
Eq. 6.3	0.61	0.07	0.63	0.54	0.05	0.58	0.63	0.03
Eq. 6.4	0.59	0.07	0.60	0.62	0.03	0.53	0.73	0.04
Eq. 6.5	0.61	0.07	0.61	0.51	0.03	0.54	0.70	0.04

RMSE was relatively higher than most of the compounds in the training data and the plot of experimental TCAA versus predicted TCAA was not linear (Figure 6. 6). The k_i and k_o statistics were lower than acceptable range of slope, k ($0.85 \leq k \leq 1.15$) and the final conclusion was that this QSPR had low predictive power.

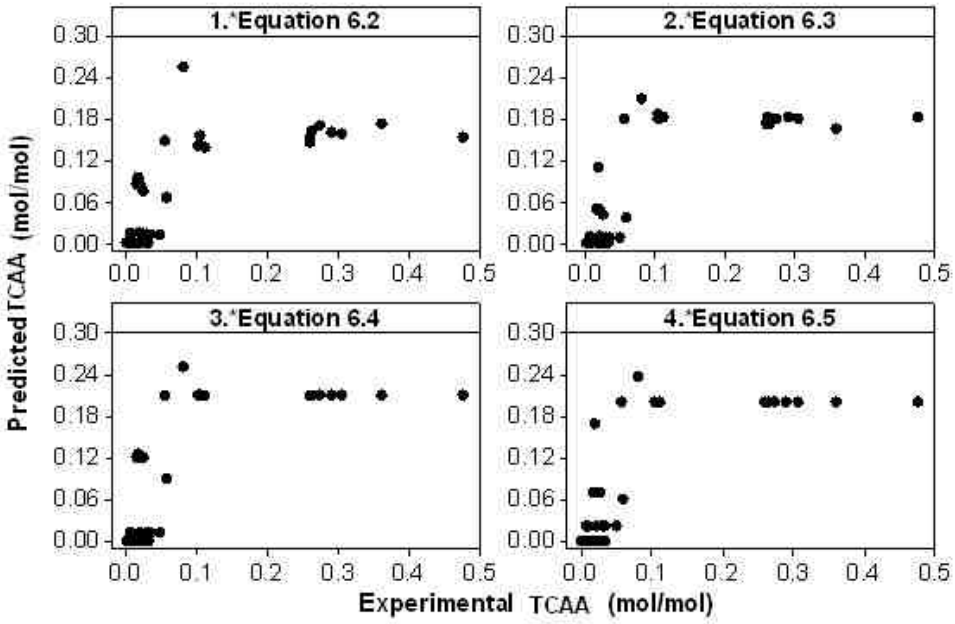


Figure 6.6. Scatter plot of predicted TCAA vs. experimental TCAA

The second attempt used square root of number of phenols (sqrtArOH), sqrtHeA, and sqrtRAI as independent variables and sqrtTCAA as dependent variable, and the QSPR obtained is given by Equation 6.3. The QSPR had R_c^2 = 0.75 and SDE = 0.10 $(\text{mol-TCAA/mol-Cp})^{1/2}$. The predicted sqrtTCAA from internal validation of the QSPR were transformed back to TCAA and data were used to determine predictive power of QSPR. Results showed that k_i and k_o were below acceptable values for slope ($85 \leq k \leq 1.15$). There was very weak linear relationship between predicted TCAA and experimental TCAA (Figure 6.6). Therefore the QSPR's was not useful.

$$sqrtTCAA_f = 0.20 * sqrtArOH(\pm 0.05) + 0.02 * sqrtHeA(\pm 0.01) + 0.19 * sqrtRAI(\pm 0.05)$$.(Eq. 6.3)

The third QSPR was developed using multiple linear regression of TCAA on RAI and ArOH to obtain Equation 6.4. The R_c^2 and SDE for the QSPR were 0.59 and 0.08 mol-TCAA/mol-Cp respectively and internal validation showed that k_i and k_o below the accept k values (Table 6.1). In addition the experimental and predicted TCAA formation had very poor linear relationship (Figure 6.6). The overall analysis was that the QSPR had low predictive power.

$$TCAA_f = 0.09 * ArOH(\pm 0.03) + 0.12 * RAI(\pm 0.03)$$ …… (Eq. 6.4)

Equation 6.5 represents the fourth QSPR relating TCAA formation to sqrtArOH and sqrt(RAI) and the model had R_c^2 = 0.61 and SDE = 0.07 mol-TCAA/mol-Cp. From analysis of internal validation it was found that the slopes k_i and k_o were below acceptable value of k (Table 6.1) and it was also shown that the relationship between predicted and experimental TCAA was poor (Figure 6.6). These results indicate that the QSPR had low predictive power.

$$TCAA_f = 0.13 * sqrtArOH(\pm 0.04) + 0.07 * sqrtRAI(\pm 0.03)$$ …… (Eq. 6.5)

6.7. Conclusions

Constitutional descriptors were used to develop QSPRs for predicting TCAA formation. None of QSPRs presented above met all the minimum requirements for QSPR predictive power. However, there were some linear relationships between predicted TCAA and experimental TCAA for compounds with TCAA formation less than 0.2 mol-TCAA/mol-Cp. The interpretations of these results are of two fold: (i) the descriptors used in this work were not

sufficient to explain the variation in TCAA formation; (ii) TCAA formation has statistically insignificant linear relationship with constitutional descriptors and therefore TCAA formation may be explained better using more sophisticated molecular descriptors (quantum-mechanical descriptors, typological descriptors, geometrical descriptors) or by using QSPR derived using non-linear algorithms. The application of descriptors other than constitutional descriptors and non-linear algorithms to derive QSPRs was beyond the scope of this study.

6.8. References

1. IPCS (International Program on Chemical Safety). *Disinfectants and disinfectant byproducts, Environmental Health Criteria 216*. WHO, Geneva: 2000.

2. Crittenden, J.C.; Trussell, R.R.; Hand, D.W.; Howe, K.J.; Tchobanoglous, G. *Water treatment: Principles and design,* 2nd edn, John Wiley and Sons Inc., New York: 2005.

3. Chowdhury, S.; Champagne P. An investigation on parameters for modeling THM formation. *Global NEST. J*. **2008**, 10(1), 80-91.

4. Chowdhury, S.; Champagne, P.; McLellan, P.J. Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. *Sci. Total Environ.* **2009**, 407(14), 4189-206.

5. Ates, N.; Kitis, M.; Yetis, U. Formation of chlorination by-products in waters with low SUVA--correlations with SUVA and differential UV spectroscopy. *Water Res.* **2007**, 41(18), 4139-48.

6. Reckhow, D.A.; Rees, P.L.; Nusslein, K.; Makdissy, G.; Devine, G. *Long-term variability of BDOM and NOM as precursors in watershed sources*. American Water Works Association: 2007

7. Wei, Q.; Fengm C.; Wang, D, et al. Seasonal variations of chemical and physical characteristics of dissolved organic matter and trihalomethane precursors in a reservoir: a case study. *J. Hazard. mater.* 2008, 150(2), 257-64.

8. Weishaar, J.L.; Aiken, G.R.; Bergamaschi, B.A.; Fram, M.S.; Fujii, R.; Kenneth, M. Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. *Environ. Sci. Technol.* **2003**, 37(20), 4702-4708.

9. Katritzky, A.R.; Lobanov, V.S. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev*. **1995,** 24, 279-287.

10. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSPR/QSAR. *Chem. Rev*. **1996**, 96(3), 1027-1044

11. Bull, R.; Reckhow, D.; Rotello, V.; Bull, O.M.; Kim, J. *Use of toxicological and chemical models to prioritize DBP research.* American Water Works Association: 2006.

12. Dickenson, E.R.; Summers, R.;S, Croué, J.; Gallard, H. Haloacetic acid and trihalomethane formation from the chlorination and bromination of Aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol.* **2008**, 42(9), 3226-3233.

13. Hong, H.C.; Wong, M.H.; Liang, Y. Amino acids as precursors of trihalomethane and haloacetic acid formation during chlorination. Arch. *Environ. Contam. Toxicol.* **2009**, 56(4), 638-45.

14. Minitab Inc. Graphical data: Meet Minitab 15, p. 2-1 to 2-13. 2007. http://www.minitab.com.

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

16. Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, 22(1), 69-77.

CHAPTER 7

RESEARCH SUMMARY AND RECOMMENDATIONS

Abstract

This chapter provides a summary of results from quantitative structure-property relationship (QSPR) for chlorine (HOCl) demand and QSPRs for, total organic halide (TOX), trichloromethane (TCM) formation and trichloroacetic acid (TCAA) formation. In also presents results from integration of QSPRs (chlorine demand and TOX formation) with AlphaStep model of NOM. All QSPRs, except for TCAA, showed high predictive power using external data. None of the QSPRs for predicting TCAA formation had met the minimum criteria for predictive power. The QSPRs for chlorine demand and TOX formation were integrated with AlphaStep NOM. Results from simulation of HOCl demand and TOX formation potential from chlorination of NOM in surface water were 27.55 µmol-HOCl/mg-C and 183.6 µg-Cl/mg-C respectively. These results are consistent with reported values for humic acid and fulvic acids and also whole DOM.

7.1. Summary

Supply of good quality potable water to public that meets regulatory requirements is the goal of water supply authorities. Use of chemical disinfectants, particularly chlorine, to treat drinking water has saved millions of lives from waterborne diseases (*1,2*). The detection of THMs in drinking water (*3,4*) showed that chemical disinfection also yield toxic products. Early studies on

animals showed that THMs and HAAs were toxic and carcinogenic (*5*) and regulatory authorities responded by issuing standards that limit the amounts of disinfection byproducts. The response of water researchers to DBPs has focused on designing water treatment processes and technologies that minimize formation of disinfection byproducts without compromising disinfection efficiency. Mathematical modeling is one of the tool that is used to improve water treatment conditions and most of empirical models are based bulk water quality parameters (*6,7*)**.** Application of empirical models based on water quality parameters have some drawbacks which have been discussed in Chapter 1.

This research work was designed to address these challenges by developing models for predicting chlorine demand and disinfection byproducts formation using molecular structures of model compounds. Quantitative structure-property relationship (QSPR), which relate structure of molecules and physical chemical property, were used for the first time predict chlorine demand and disinfection byproducts formation. The research work produced multiple linear regressions (MLR) QSPRs for predicting: chlorine demand, total organic halide (chlorine) formation, chloroform formation and trichloacetic acid (and HAAs) formation.

7.1.1. QSPR for chlorine demand

The QSPR for predicting chloride demand had eight constitutional descriptors: ring activation index (RAI), number of phenols (ArOH), carbonyl (CI), number of aliphatic carbons attached to reduced nitrogen, $-NH_2$ (ACN), atomic oxygen to carbon ratio (O:C), number of aliphatic sulfur (AS), number of alkoxy

groups on an aromatic ring without NH_2 an OH (ArORact) and number of alkoxy groups on aromatic ring with NH_2 and OH (ArORnact). The QSPR (Eq. 7.1) had R_c^2 and SDE of estimate of 1.24 mole HOCl per mole compound (Cp).

$$HOCl_{dem} = 7.61 * RAI + 1.16 * ArOH + 3.00 * ACN + 1.23 * CI + 2.37 * AS + 1.01 * O:C + 0.49 * ArORact - 0.72 * ArORnact \quad \text{..... (Eq. 7.1)}$$

The predictive power of the QSPR was evaluated using cross validation -Leave-One-Out cross validation (LOO_{CV}) and Leave-Many-Out cross validation (LMO_{CV}) and external data. The LMO_{CV} and LOO_{CV} showed that the QSPR had high predictive power as q^2_{LMO} = 0.86, $RMSE_{LMO}$ = 1.21 mol-HOCl/mol-Cp, k_i = 0.88, k_o = 0.97 and R_t = 0.015 and q^2_{LOO} = 0.85, $RMSE_{LOO}$ = 1.28 mol-HOCl/mol-Cp, k_i = 0.88, k_o = 0.97 and R_t = 0.013 met the criteria for predictive power (*8,9*). The QSPR also performed well on external data as indicated by statistics of predictive power: q^2_{Ext} = 0.88, $RMSE_{Ext}$ = 1.17 mol- HOCl/mol-Cp, k_i = 0.90, k_o = 1.05 and R_t = 0.039. The q^2 and RMSE for external validation were also comparable to R_c^2 and SDE those from LMO_{CV} and QSPR calibration respectively. However, the applicability domain revealed that chlorine demand predictions for 7 compounds in the external data (N = 42) were due to extrapolation of the QSPR. Of the eight constitutional descriptors, RAI, CI and ACN were the most important descriptors. RAI represents the ring activation from OH. NH_2 groups and CI is important in aliphatic compounds with β-diketone and ACN is important descriptor for amino acids and other molecules with amine. Nitrogen is more electronegative than carbon and therefore induces electron deficiency on α-C in of most amino acids. The hydrogen on nitrogen and on α-C can be substituted for chlorine. Although

QSPR had high predictive power, it may not do well with position isomers because constitutional descriptors do not explain steric and electronic properties of the molecules. The model may not work well with aliphatic with C=C bonds which are good site for chlorine addition because the QSPR does not have descriptor to explain this molecular structure.

7.1.2. QSPR for TOX formation

Chlorination of drinking water produces several disinfection byproducts and toxic nature of the mixture is of health concern. This makes total organic halide (TOX) as an appropriate surrogate of presence of toxic disinfection byproducts (*10*). In this work the QSPR for predicting TOX formation was developed is given by Equation 7.2. The QSPR has four constitutional descriptors which were carbonyl index (CI), ratio of phenol to carbon (ArOH:C), square root of number hetero atoms (sqrtHeA) and log of atomic hydrogen to carbon ratio (logH:C).

$$TOX_f = 0.54 * CI + 5.33 * ArOH:C + 0.33 * sqrtHeA - 1.36 * \log H:C \quad \text{..... (Eq. 7.2)}$$

The four descriptors explained about 72% of variation in TOX formation (R_c^2 = 0.72) with SDE of 0.43 mol-Cl/mol-Cp. The model predictive power was validated by LOO_{CV} (N = 49) and had q^2_{LOO} = 0.60, RMSE = 0.50 mol-Cl/mol-Cp , k_i' = 0.95, k_o'= 0.99 and R_t' = 0.001 whereas external validation (N = 12) had q^2_{Ext} = 0.67, RMSE = 0.48 mol-Cl/mol-Cp. These statistics indicate that the QSPR has high predictive power and robust. The predictions of TOX for three compounds in external data were due to extrapolation of the QSPR and therefore may not be

reliable. The predictions of TOX formation from tannic acid, fulvic acid (FA-1) and fulvic acid (FA-2) were 150.63 µg-Cl/mg-C, 143.89 µg-Cl/mg-C and 176.88 µg-Cl/mg-C respectively and these fell in 77.80-192.50 µg-Cl/mg-C range for finished waters reported in literature (*10*). This QSPR will over-predict most of the compounds with very low TOX formation which were mostly amino acids. Since the contribution of the amino acids to the total TOX is very small the QSPR is still useful for prediction because they do not affect the overall performance of the model.

The ArOH:C represents that number of phenols per carbon and is the most important descriptor of all would in TOX formation. But the descriptor may not be able to catch the difference in TOX formation of 1,2-dihydroxybenzene from 1,3-dihydroxybenzne and 1,4-dihydroxybenzene. The descriptor, sqrtHeA, is the square root of the sum of O, N and S and of these N and S are neutral nucleophiles which can attack HOCl. Oxygen is very electronegative atom (present in functional groups such as -OH, -C=O) and can cause electron deficiency to a carbon to which it is bonded. This in turn lower the pKa of protons attached to the **α**-carbon which promotes chlorine substitution.

When chlorine demand and TOX formation data were compared it was found that the amount of HOCl consumed is generally larger than the amount of TOX detected for model compounds and water samples. The lack of mass balance may arise from irreversible adsorption chlorinated organics to activated carbon or volatile organics during TOX analysis or loss of chlorine as inorganic chlorine. The QSPR may under-predict TOX formation potential some

compounds because the model was derived using amount of TOX obtained after passing sample on activated carbon in TOX analyzer. Apparently there is no study that has evaluated how much of chlorine consumed during chlorination of water is lost as inorganic chlorine or irreversibly adsorbed on activated carbon. However, the average amount of TOX to chlorine consumed by model compounds used in this study was 22.38% which was within 18-30% range found for whole NOM or fulvic acid and humic acid fractions.

7.1.3. QSPR for TCM formation

Trichloromethane (TCM) is one of the trihalomethane that was discovered early in 1970s and it is the only trihalomethane that produced in water free of bromine. The QSPR for predicting chloroform formation was derived (Eq. 7.3) and had R_c^2 = 0.94 and SDE = 0.08 mol-TCM/mol-Cp. The model had three descriptors namely carbonyl index (CI), one-three activated carbon in aromatic ring (OTactC) and EDCORH is difference between sum of number of strong electron donors (OH and NH_2) on aromatic ring and number of aldehyde and ketone groups (CORH). These descriptors explained over 90% of variance in chloroform formation.

$$TCM_f = 0.28 * CI + 0.27 * OTactC + 0.35 * EDCORH \quad \text{…….} \quad \text{(Eq. 7.3)}$$

The OTactC descriptor for molecules with resorcinol like structures (3-chlorophenols, 3-aminophenols and 1,3-dihydroxybenzenes) represents the ortho (2) carbon between the two groups which becomes highly activated making it more nucleophilic. The nucleophilic carbon can attack the electron

deficient chlorine in HOCl which eventually leads to chlorine substitution. The QSPR showed high predictive power using LMO_{CV} (q^2_{LMO} = 0.92, RMSE = 0.09 mol-TCM/mol-Cp, k_i = 0.91, k_o= 0.95 and R_t = 0.007) and external validation (q^2_{Ext} = 0.94, $RMSE_{Ext}$ = 0.08 mol-TCM/mol-Cp, k_i' = 0.85, k_o'= 0.85 and R_t' = 0.0003). These statistics indicate that the model had high predictive power. But applicability domain analysis showed that 5 out 27 compounds in external data were predicted due to extrapolation of the QSPR. The high predictive of the QSPR was tested on tannic acid for which predicted TCM formation (0.33 mol-TCM/mol-Cp) which was higher than experimental value (0.13 mol-TCM/mol-Cp). The model over predicted molecules with low chloroform formation and did better with high chloroform producing molecules. Most of the compounds with low chloroform production were mostly amino acids and aliphatic compounds (excluding β-diketones). The QSPR may not predict reliably molecules with aliphatic C=C, and compound with high number of ketone and aldehyde relative to number of OH and NH_2 (which ortho-para to each other) in aromatic ring.

7.1.4. QSPR for TCAA formation

Chlorination of drinking water also produces haloacetic acids which are among the DBPs of health concern. In this work attempts were made to derive QSPR for predicting trichloroacetic acid (TCAA) using various combinations of constitutional descriptors listed in Table 2.1. All QSPR reported had high R_c^2 and SDE but the models failed to meet the criteria of QSPR predictive power on validation data. The results also showed that there was insignificant linear relationship between experimental TCAA and predicted TCAA. It was concluded

the constitutional descriptors which were used in the study could not explain the variation in TCAA formation and had very poor linear relationship between constitutional descriptors and TCAA formation. The relationship between the two may be explained by non-linear QSPR which could be derived using non-linear algorithms such as artificial neural network (ANN).

7.2. Integration of QSPRs with AlphaStep model of NOM

7.2.1. Background

Natural organic matter is a complex mixture of organic molecules with diverse structure, size and functional groups (*11*). The individual molecules interact with water (hydrophobic and hydrophilic interactions), each other (hydrogen bonding or acid-base interaction), metals (complexation, charge transfer), with bacteria (biodegradation), light, oxygen and chemical oxidants (redox and radical drive reactions), just to mention a few. The reaction of chlorine with components of natural organic matter is one of example of interaction of low level components which leads to change in properties of parent molecules or formation of new molecules. Thus, changes driven by individual molecules may ultimately impact the behavior of macrosystem (high level interaction). Agent-based models (ABMs) are said to be appropriate for studying natural organic matter interactions because with various components in water matrix may act as agents (*12*).

Alphastep model is an agent based modeling (ABM) software and is defined as “a Windows-based program which simulates transformation of environmental biological materials into NOM and eventual consumption or

destruction" (*13*). The transformation and consumption (or destruction) reaction probabilities are influenced by various biological, chemical and physical factors which all together constitute environmental parameters (*13*). It is useful for estimating not only reaction probabilities for transformation of macromolecules based on structural properties but also examining properties of different molecules (end products) after simulation (*13*).

Previous studies shown that it is possible to combine AlphaStep with QSAR or QSPR derived using constitutional descriptors in order to predict Cu(II) binding to NOM (*14*), conditional metal-ligand binding constants of metals to NOM (*15*) and pKa of NOM (*16*). The results from each study were consistent experimental values reported in literature. In this work, the QSPRs for HOCl demand and TOX formation were derived using chlorine demand and TOX formation data based on bench scale experiments using dissolved organic precursors with well defined structure and functional groups. The reaction behavior of the molecules used in this study may not necessarily reflect the behavior of natural organic matter in a natural system. In order to evaluate the implication of the descriptors in natural systems QSPRs were combined with AlphaStep model in order to estimate chlorine demand and TOX formation in surface waters.

7.2,2. AlphaStep algorithms

The general algorithm for the AlphaStep is reported in literature (*11*) and reaction conditions used for simulation of transformation of natural organic matter precursors from aquatic and soil NOM matrices are summarized. Soil NOM had

an input of 2000 molecules each of abietic acid, flavonol and gallic acid with pH = 5.0, dissolved oxygen = 0.1 mM, enzyme (protease and oxidase) activity = 0.01, carboxylase and bacterial abundance = 0.010, simulation time = 5000 h reaction time (~90s computing time) and humid soil incubated in darkness. Simulation of aquatic NOM precursors was performed using an input of 400 molecules each of lignin and protein, pH = 7.0, dissolved oxygen = 0.3 mM, temperature = 25^{o}C, enzyme activity (protease, oxidase, decarboxylase) = 0.010, simulation time = 5000 h reaction time (~90s computing time) and incubated under constant light intensity of 2.0 x 10^{-8} mol.cm^{2}.h^{-1}. A brief description of simulation process is given below (*11*).

Each molecule is treated as an agent with defined chemical structure and reaction probabilities with other molecules or with environmental variables such as light, oxygen, bacteria. The molecules are therefore allowed to undergo chemical transformation through a series of steps in the course of simulation time. For each step, molecules are tested for possible chemical reactions that may lead to their transformation into new molecular products. The reaction probabilities of products in each step are allowed to react and their reaction probabilities are calculated and the process is repeated until the simulation time has elapsed. Since the program is capable of reporting reaction frequencies (counts) and also calculating aggregate properties of end products (NOM matrix), it is possible to obtain molecular descriptors that can be substituted in the QSPRs in order to estimate reactivity properties of NOM. In this case, the

molecular properties in the output were used to calculate the constitutional descriptors in order to estimate the chlorine demand and TOX formation of NOM.

7.2.3. Results and discussion

Simulation results for chlorine demand of natural organic matter in water using the eight constitutional descriptors (Eq. 7.1) was 27.55 µmol-HOCl/mg-C. This value is comparable to typical chlorine demand of fulvic acids in natural water of 27-33 µmol-HOCl/mg-C (*17*). However, the predicted chlorine demand for NOM from soil was 18.59 µmol-HOCl/mg-C which is lower than that for aquatic NOM. The soil NOM is derived primarily from chemical and biodegradation of terrestrial plants and animals as well as from anthropogenic activities. On the other hand aquatic NOM is a mixture of NOM terrestrial sources that reach water through surface runoffs and NOM derived from planktons. The differences in chlorine demand may reflect the differences in composition of NOM fractions and aqueous medium provides viable environment for faster biodegradation of NOM into smaller molecules than in terrestrial environment.

Simulations of total organic halides showed that TOX formation from aquatic NOM was higher than TOX formation from soil NOM. The predicted TOX formation from aquatic NOM and Soil NOM were 183.6 µg-Cl/mg-C and 136.4 µg-Cl/mg-C respectively. Previous studies showed that TOX formation from chlorination of humic acids (HA) and fulvic acids (FA) were 136-232 µg-Cl/mg-C for FA and 230-288 µg-Cl/mg-C for HA and for the whole dissolved organic matter was 170-298 µg-Cl/mg-C (*18*). The predictions TOX formation from simulation are consistent with Singer's study and are also in agreement with

Reckhow et al. (*10*) study which showed that chlorination of surface waters in the US gave 77.8-192.5 µg-Cl/mg-C.

7.2.4. Conclusions

The simulation of chlorine demand and TOX formation from NOM suing AlphaStep and descriptors from QSPRs gave results consistent with empirical data from chlorination of NOM in water. These results compliment to conclusions reached in earlier chapters that QSPRs derived using constitutional had high predictive power and robust and that the constitutional descriptors have mechanistic implications. The differences in between TOX formation and chlorine demand have also been revealed in simulation where HOCl demand was higher than TOX formation. This agrees with the hypothesis that there is lack of mass balance between the two which implies loss of chlorine in inorganic form, as VOCs or irreversible activated carbon in TOX analyzers.

7.3. Implications of the descriptors to chlorination reaction

This research has reported three multiple linear regression QSPRs that showed high predictive power. However, each QSPR had different numbers of descriptors used to explain the variation in dependent variable though there were few descriptors that were common to all QSPRs. The QSPR for chlorine demand had more descriptors than either the QSPRs for TOX and TCM formation QSPR (Eqs 7.1, 7.2 & 7.3). The QSPR for chlorine demand had more descriptors because there are different classes of molecules in water that vary in structural properties (or functional groups) that influence their chemical reaction with chlorine. The chlorinated organic molecules may retain parent structure or may

undergo further transformation to known and unknown chlorinated compounds (DBPs) and unchlorinated organic products. The RAI, ArOH, CI, ACN and AS were more important descriptors than O:C, ArORact and ArORnact. The RAi and ArOH were key descriptors for aromatic organic molecules with strong electron donating substituents. The ArOH and RAI (the relative number of strong electron donor (NH_2 and OH) to number of aromatic rings) are explain for the degree to which the ring is activated. These functional groups make the ring highly electron rich though the number and the relative position of the groups to each in the ring may increase or decrease ring electron richness due cooperative or antagonistic effects respectively. In general, the more electron rich the ring is the higher the electrophilic substitution.

Carbonyl index was an important additive descriptor to explain reactivity of carbonyl (aldehydes and ketones) α, β-and γ-dicarbonyl compounds that may present in water sample matrix. The pKa of hydrogen on alpha-carbon in between two carbonyl groups is relative lower than in alpha or gamma carbonyls. The lower the pKa the easier the abstraction of proton by a base in water to generate a keto-enol intermediate that eventually leads to chlorine substitution. The ACN and AS are two descriptors related to number of aliphatic nitrogen (amines) and sulfur (thiols, thiomide, thioethers, thioketone). Nitrogen and sulfur have a pair of electrons p-orbital which can act as neural bases. The lone pair can attack the Cl atom in HOCl which may eventually lead to addition chlorine to sulfur or nitrogen. The addition of chlorine to sulfur or nitrogen may induce electron deficiency on the carbon to which they are bonded and that may lead to

nucleophilic or electrophilic substitution with a release of inorganic chlorine compounds. This descriptor is additive and therefore chlorine consumed is expect increase with increasing number of aliphatic sulfur and amine in a molecule. The ACN is more important for amino acids (including cysteine and methionine) which moderate chlorine consumption and AS is for any molecule with aliphatic sulfur.

Formation of disinfection byproducts from a precursor is a multistep processes with several intermediate products. Disinfection byproducts may be formed from any of intermediates by addition reactions, substitution reaction, elimination reactions, etc. Chlorination of aromatic rings (in aromatic compounds) or aliphatic compounds may occur at different rates and not every chlorinated molecule may undergo ring opening to form THMs and HAAs. Some of the molecules may retain its parent structure and some might be broken down into compounds with unknown identity whereas breakdown of intermediates for some compounds may favor formation of THMs over HAAs and vice versa. The mechanisms by which THMs and HAAs are formed from intermediates may not be highly dependent on the structures of parent molecules. Thus, it is not surprising to find that QSPRs for DBPs formation have been explained by fewer descriptors than chlorine demand and just a few of the descriptors in chlorine demand model also appear TOX and TCM formation models. Thus, CI, ArOH or sum NH_2 and OH in aromatic ring (ED) emerged as important descriptors for TOX and TCM formation.

Chlorination of aromatic and aliphatic compounds eventually leads to formation of disinfection byproducts which may imply that chlorine consumed should be equal to chlorine incorporated in organic molecules of known and unknown identities (TOX). This is generally not the case because some of the chlorine may be lost from intermediate products as inorganic chlorine which cannot be detected by TOX analyzers. Further analysis of the chlorine demand and TOX formation from individual model compounds (*19,20,21*) showed only 22.38% of the chlorine consumed is detected as TOX for majority of them (Tables S7.1). Similarly amount TOX detected relative to chlorine form fulvic and humic fractions as well as raw water (*18*) was about 18-33% (Tables S7.2 & S7.3). These two results agree to each other and in this case it showed that about 70-80% of chorine consumed is lost probably as inorganic chlorine. Consequently, the linear relationship between chlorine demand and TOX formation may not always hold. This is one of the reasons chlorine demand and TOX formation QSPRs do not have the same number of descriptors.

The application of the QSPRs developed in this study is limited by the conditions the data for chlorine demand and DBPs were collected. The prediction of chlorine demand, TOX and TCM may be reliable for samples chlorinated at pH 7-8; water should not have high levels of copper (II), iron (II) and manganese (II) which can increase chlorine demand because these ions can be oxidized by chlorine**.** These metal ions may also coordinate to carbonyls or amines in molecules like β-diketone. The consequence of this phenomenon is that pKa of hydrogen on α-carbon will be lowered and study has shown that chloroform

formation from carbonyl compounds or 2-hydroxybenzoic acids and catechols was higher in presence of copper (II) than when it was absent (*22,23*). Presence of high levels of bromide may also affect predictions of DBPs because bromide reaction with HOCl to produce HOBr. The presence of these two oxidants will increase formation of DBPs significantly while the QSPRs were derived using data that were collected in samples without bromide. The nitrogenous aromatic compounds may have exocyclic directly bonded to the carbon in the ring or endocyclic amine when nitrogen part of the atoms forming the ring. The QSPRs for chlorine demand, TOX and TCM formation may not have high predictive in nitrogenous compounds with endocyclic nitrogen or oxygen or sulfur because there is descriptor in the QSPRs that account for endocyclic ring activation.

7.4. Recommendations

This work used constitutional descriptors to derived QSPRs. It was found that constitutional descriptors that were used in this study failed to explain variation in TCAA formation. This suggests that there is either no linear relationship between TCAA and constitutional descriptors or the relationship is non-linear. It is recommended TCAA formation should be predicted using non-linear QSPR which could be derived using non-linear algorithms such artificial neural network.

Constitutional descriptors cannot explain the differences in DBPs formation or chlorine demand positional isomers or constitutional isomers. Alternatively, use of descriptors such as geometrical descriptors, quantum-

mechanical descriptors, typological descriptors is highly recommended in order to improve the predictions.

Natural water may have traces of bromide from point and non-point sources and chlorination of such waters may affective predictive power of models derived using bromine free water sample matrices. There is an opportunity of QSPR studies on chlorination of water (with bromide at various concentrations) so that we may derive correction factors to account for presence of bromide

The $HOCl_{dem}$ and DBPs formation experiments were performed at contact times between 4-96 h and in excess chlorine dose which do not reflect the routine contact time and chlorine doses used in typical drinking water treatment plants. More work is needed to build up data base for $HOCl_{dem}$ and DBPs formation under typical water treatment conditions. Monitoring of chlorine demand and DBPs formation from chlorination of most reactive model compounds with time will shed light on which model structures will react at the water treatment plant and which ones will react in distribution systems. Such data will also be useful for evaluating the predictive power of the QSPRs reported in this work and designing strategies to control the most reactive precursors.

7.6. References

1. Christman, K. *History of chlorine. Chlorine Chemistry* Council, Waterworld: 1998. http://c3.oreg/chlorine_knowledge_center/hioistroy.htm.

2. Lee, S.H.; Levy, D.A.; Craun, G.F.; Beach M.J.; Calderon R.L. Surveillance for waterborne-disease outbreaks-United States, 1999–2000. *Morbid. Mortal. Wkly Rep.,* 2002, 51(SS-8), 1-28.

3. Bellar, T.L.; Lichtenberg, J.J.; Kroner, R.C. The Occurrence of organohalides in chlorinated drinking waters. *J. Am. Water Works Assoc.* **1974**, 66(12), 703-706.

4. Rook, J.J. Formation of haloforms during chlorination of natural waters. *Wat. Treat. Exam.* **1974**, 23(2), 234-243.

5. IPCS (International Program on Chemical Safety). *Disinfectants and disinfectant byproducts, Environmental Health Criteria 216*, WHO: 2000.

6. Chowdhury, S.; Champagne, P. An investigation on parameters for modeling THMs formation. *Global NEST J.* **2008**,10(1), 80-91.

7. Chowdhury, S.; Champagne, P.; McLellan, P.J. Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. *Sci. Total Environ.* **2009**, 407(14), 4189-206.

8. Golbraikh, A.; Tropsha, A.; Beware of q2! *J. Mol. Graphics Model.* **2002**, 20(4), 269-276.

9. Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, 22(1), 69-77.

10. Reckhow, D.; Hua, G.; Kim, J.; Hatcher, P.; Caccamise, S.; Sachdeva, R. Characterization of total organic halogen produced during disinfection processes. AWWARF Report 91176: 2008.

11. Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics. *Biogeochemistry.* 2005, 76(2), 319-347.

12. Devillers, J.; Devillers, H.; Decourtye, A.; Aupinel, P. Internet resources for agent-based modeling. *SAR QSAR Environ. Res.* 2010, 21(3-4), 337-350

13. AlphaStep User Guide, Version 0.3, October, 2003

14. Cabaniss, S.E.; Maurice, P.A.; Madey, G. A stochastic model for the synthesis and degradation of natural organic matter. Part III: Modeling Cu(II) complexation. *Appl. Geochem.* 2007, 22,1646-1658

15. Cabaniss, S.E. Forward modeling of metal complexation by NOM: I. A priori prediction of conditional constants and speciation. *Environ. Sci. Technol.* 2009, 43(8), 2838-2844.

16. Cabaniss, S.E.; Madey, G.; Leff, L.; Maurice, P.A.; Wetzel, R. A stochastic model for the synthesis and degradation of natural organic matter part II: molecular property distributions. *Biogeochemistry.* 2007, 86, 269-286

17. Reckhow, D.A.; Singer, P.C.; Malcom, R.L. Chlorination of humic materials: Byproduct formation and chemical interpretation. *Environ. Sci. Technol.* 1990, 24, 1655-1664

18. Singer, P.C. Humic substances as precursors for potentially harmful disinfection byproducts. *Wat. Sci. Tech.* 1999, 40(9), 25-30

19. Bull, R.J.; Reckhow, D.A.; Rotello, V.; Bull, O.M.; Kim, J. Use of toxicological and chemical models to prioritize DBP research; AWWA Research Foundation: 2006.

20. Dickenson, E.V.; Summers, S.; Croué, J-P.; Gallard, A. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol.* 2008, 42, 3226-3233.

21. Hureiki, L.; Croué, J-P.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res.* 1994, 28, 2521-2531.

22. Blatchley III, E.R.; Margetas, D.; Duggirala, R.; Copper catalysis in chloroform formation during water chlorination. *Water Res.* 2003, 37, 4385-4394.

23. Fu, J.; Qu, J.; Liu, R.; Qiang, Z.; Liu, H.; Zhao, X. Cu(II)-catalyzed THM formation during water chlorination and monochloramination: A comparison study. *J. Hazard. Mater.* 2009, 170, 58-65.

APPENDICES

Table S2.1. A list of constitutional descriptors and abbreviations.

Descriptors	Abbreviation	Descriptors	Abbreviation
# Alkoxy groups attached to the aromatic ring without NH_2 and OH	ArORact	# Alkoxy groups attached to the aromatic ring without NH_2 and OH	ArORnoact
# Oxygen to carbon ratio	O:C	Ring activation index	RAI
Square root of number of heteroatoms	sqrtHeA	# One three activated aromatic carbon	OTactC
Square root of number of phenols	sqrtArOH	Difference of ArED:C and CORH:C	EDCORH
Sum of weak of strong electron donors (alkoxys) bonded to aromatic ring per carbon	ArOR:C	# Aliphatic C bonded reduced nitrogen (NH_2)	ACN
# Phenols	ArOH	Sum ArED per carbon	ArED:C
# Aliphatic sulfur	AS	Carbonyl index	CI
Log of hydrogen to carbon ratio	logH:C	# Phenol per carbon	ArOH:C
# Hydrogen to carbon ratio	H:C	Square root of ring index	sqrtRAI
# Anilines	$ArNH_2$	Log of number of heteroatoms	logHeA
# Nitrogen to carbon ratio	N:C	Log of number of carbons	logC
# Carboxylic acids and acid amides	$COOH/NH_2$	# Carboxylic acids and acid amides per carbon	COOH/NH2:C
#Aldehydes and ketones	CORH	# Aldehydes and ketones per carbon	CORH:C
# C-C pi-bonds	piB	# C-C pibonds per carbon	piB:C
# Hydroxyl, amino and alkoxy groups bonded to aromatic ring	EDOR	# Hydroxyl, amino and alkoxy groups bonded to aromatic ring per carbon	EDOR:C
# Heteroatoms	HeA	Log of number of heteroatoms per carbon	logHeA:C
# Aniline per carbon	$ArNH_2$:C	# Alcohols and thiols	RO(S)H
# Carboxylic acid and acid amines per carbon	$COOH/NH_2$:C	# Aromatic rings	Ar
# Aldehyes and ketones per carbon	CORH:C	#Aliphatic amine	RNH_2
# Hydroxyl, amino and alkoxy groups bonded to aromatic ring per carbon	EDOR:C	# Aromatic aldehyde and ketones	ArCORH
# C-C pibonds per carbon	piB:C	# Aliphatic aldehyde and ketones	RCORH
# Oxygen atoms	O	# Aromatic carboxylic acid and acid amines per carbon	$ArCOOH/NH_2$: C
# Hydrogen atoms	H	# Aromatic aldehyde and ketones per carbon	ArCORH:C
# Carbon atoms	C	# Aliphatic amine per carbon	RNH_2:C
# Activated aromatic carbon	ActC	Sum of strong electron donors (OH and NH_2) bonded to aromatic ring	ArED
# Halogens bonded to aromatic	ArX	Sum of weak of strong electron donors (alkoxys) bonded to aromatic ring	ArOR
Square root of number of carbon	sqrtC	# Aliphatic carbons	R-C
Square root of number of oxygen atoms	sqrtO	Square root of hydrogen atoms	sqrtH
# Aromatic and aliphatic esters	COOR	# Aromatic carbons	Ar-C

Table S3.1. Leave-Many-Out calibration data set (N = 109) for chlorine demand (mol-HOCl/mol-Cp). AdjHOCl = adjusted HOCl demand

Compound name	Source	HOCl	AdjHOCl	RAI	ArOH	ACN	Cl	OC	AS	ArORact	ArORnact
2,4-Dihydroxybenzoic acid	(*1*)	7.50		0.60	2.00	0.00	0.00	0.57	0.00	0.00	0.00
3,5-Dihydroxybenzoic acid	(*1*)	7.10		0.60	2.00	0.00	0.00	0.57	0.00	0.00	0.00
1,4-Phenyldiamine	(*2*)	3.53		0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2-Aminobenzoic acid	(*2*)	6.04		0.75	0.00	0.00	0.00	0.29	0.00	0.00	0.00
2-Aminophenol	(*2*)	4.70		0.30	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3,4,5-Triethoxybenzoic acid	(*2*)	0.99		0.10	0.00	0.00	0.00	0.38	0.00	3.00	0.00
3,4,5-Trimethoxyacetophenone	(*2*)	0.93		0.10	0.00	0.00	0.50	0.36	0.00	3.00	0.00
3,4,5-Trimethoxybenzamide	(*2*)	5.33		0.10	0.00	1.00	0.00	0.40	0.00	3.00	0.00
3,4,5-Trimethoxyphenyl acetonitrile	(*2*)	4.38		0.10	0.00	0.00	0.00	0.27	0.00	3.00	0.00
3,4,5-Trimethoxyphenylacetic acid	(*2*)	0.84		0.10	0.00	0.00	0.00	0.45	0.00	3.00	0.00
4-(3,4,5-trimethoxybenzoyl) butyric acid	(*2*)	2.01		0.10	0.00	0.00	0.00	0.43	0.00	3.00	0.00
4-Allyl-2,6-dimethoxyphenol	(*2*)	6.80		1.00	1.00	0.00	0.00	0.27	0.00	0.00	2.00
4-Aminobenzoic acid	(*2*)	7.90		1.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00
4-Hydroxybenzophenone	(*2*)	8.94		1.00	1.00	0.00	0.00	0.15	0.00	0.00	0.00
4-Methyl-2,6-dimethoxyphenol	(*2*)	4.87		0.50	1.00	0.00	0.00	0.33	0.00	0.00	2.00
Acetosyringone	(*2*)	7.79		1.00	1.00	0.00	0.50	0.40	0.00	0.00	2.00
Aniline	(*2*)	8.58		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ethyl-(3,4,5-trimethoxybenzyl) acetate	(*2*)	4.50		0.10	0.00	0.00	0.00	0.36	0.00	3.00	0.00
Ferulic acid	(*2*)	10.32		1.00	1.00	0.00	0.00	0.40	0.00	0.00	1.00
Syringaldehyde	(*2*)	8.38		1.00	1.00	0.00	0.00	0.44	0.00	0.00	2.00
Syringic acid	(*2*)	6.93		1.00	1.00	0.00	0.00	0.56	0.00	0.00	2.00

1,3-Dihydroxybenzene	(*3*)	7.90		0.60	2.00	0.00	0.00	0.33	0.00	0.00	0.00
3-Oxobutanedioic acid	(*3*)	3.80		0.00	0.00	0.00	1.50	1.25	0.00	0.00	0.00
4,6-dioxoheptanoic acid	(*3*)	4.80		0.00	0.00	0.00	3.50	0.57	0.00	0.00	0.00
4-Oxoheptanedioic acid	(*3*)	1.25		0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.00
5,7-Dioxooctanoic acid	(*3*)	6.00		0.00	0.00	0.00	3.50	0.50	0.00	0.00	0.00
Phenol	(*3*)	9.50		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2,4,6-Trichlorophenol	(*4*)	8.00		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2,4-Dichlorophenol	(*4*)	8.00		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3-Chlorophenol	(*4*)	9.50		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Cyanophenol	(*4*)	11.0		1.00	1.00	0.00	0.00	0.14	0.00	0.00	0.00
Phenol	(*4*)	9.50		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
Alanine	(*5*)	3.90		0.00	0.00	1.00	0.00	0.67	0.00	0.00	0.00
Arginine	(*5*)	8.90		0.00	0.00	3.00	0.00	0.33	0.00	0.00	0.00
Asparagine	(*5*)	6.10		0.00	0.00	2.00	0.00	0.75	0.00	0.00	0.00
Cysteine	(*5*)	8.40		0.00	0.00	1.00	0.00	0.67	1.00	0.00	0.00
Glutamine	(*5*)	3.80		0.00	0.00	2.00	0.00	0.60	0.00	0.00	0.00
Glycine	(*5*)	3.40		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Leucine	(*5*)	6.60		0.00	0.00	1.00	0.00	0.33	0.00	0.00	0.00
Lysine	(*5*)	5.20		0.00	0.00	2.00	0.00	0.33	0.00	0.00	0.00
Methionine	(*5*)	6.20		0.00	0.00	1.00	0.00	0.50	1.00	0.00	0.00
Phenylalanine	(*5*)	5.20	3.82	0.00	0.00	1.00	0.00	0.22	0.00	0.00	0.00
Serine	(*5*)	4.50		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Threonine	(*5*)	5.80		0.00	0.00	1.00	0.00	0.75	0.00	0.00	0.00

Valine	(*5*)	5.70		0.00	0.00	1.00	0.00	0.40	0.00	0.00	0.00
4-Iodophenol	(*6*)	12.5		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
Alanine	(*6*)	2.81	3.56	0.00	0.00	1.00	0.00	0.67	0.00	0.00	0.00
Asparagine	(*6*)	4.1	5.21	0.00	0.00	2.00	0.00	0.75	0.00	0.00	0.00
Aspartic acid	(*6*)	5.50		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Glutamic acid	(*6*)	2.40	3.05	0.00	0.00	1.00	0.00	0.80	0.00	0.00	0.00
Glycine	(*6*)	5.60		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Isoleucine	(*6*)	2.60	3.30	0.00	0.00	1.00	0.00	0.33	0.00	0.00	0.00
Phenylalanine	(*6*)	2.70	3.43	0.00	0.00	1.00	0.00	0.22	0.00	0.00	0.00
Proline	(*6*)	5.40		0.00	0.00	1.00	0.00	0.40	0.00	0.00	0.00
Tyrosine	(*6*)	13.4		1.00	1.00	1.00	0.00	0.33	0.00	0.00	0.00
1,2,4-Trihydroxybenzene	(*7*)	3.90		0.30	3.00	0.00	0.00	0.50	0.00	0.00	0.00
1,3-Dihydroxybenzene	(*7*)	7.20		0.60	2.00	0.00	0.00	0.33	0.00	0.00	0.00
1-Naphthol	(*7*)	7.20		1.00	1.00	0.00	0.00	0.10	0.00	0.00	0.00
2-Chlorophenol	(*7*)	8.90		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2-Hydroxybenzoic acid	(*7*)	6.00		0.75	1.00	0.00	0.00	0.43	0.00	0.00	0.00
2-Hydroxytoluene	(*7*)	7.50		1.00	1.00	0.00	0.00	0.14	0.00	0.00	0.00
2-Methoxylphenol	(*7*)	7.70		1.00	1.00	0.00	0.00	0.29	0.00	0.00	1.00
2-Oxoacetic acid	(*7*)	1.10		0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00
3,4,5-Trichlorophenol	(*7*)	5.2		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3-Aminophenol	(*7*)	7.70		0.60	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3-Hydroxybenzaldehyde	(*7*)	9.80		1.00	1.00	0.00	0.00	0.29	0.00	0.00	0.00
3-Hydroxybenzoic acid	(*7*)	9.10		1.00	1.00	0.00	0.00	0.43	0.00	0.00	0.00

3-Methoxylphenol	(7)	8.10		1.00	1.00	0.00	0.00	0.29	0.00	0.00	1.00
3-Nitroaniline	(7)	8.50		1.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
3-Nitrobenzoic acid	(7)	0.10		0.00	0.00	0.00	0.00	0.57	0.00	0.00	0.00
3-Nitrophenol	(7)	9.20		1.00	1.00	0.00	0.00	0.50	0.00	0.00	0.00
4,4-Dihydroxy biphenyl	(7)	10.5		1.00	2.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Aminophenol	(7)	5.40		0.50	1.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Chloro-3,5-dimethylphenol	(7)	4.70		0.50	1.00	0.00	0.00	0.13	0.00	0.00	0.00
4-Chlorobenzoic acid	(7)	0.10		0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00
4-Chlorophenol	(7)	8.70		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Hydroxyacetophenone	(7)	9.80		1.00	1.00	0.00	0.50	0.25	0.00	0.00	0.00
4-Hydroxybenzaldehyde	(7)	8.80		1.00	1.00	0.00	0.00	0.29	0.00	0.00	0.00
4-Hydroxybenzoic acid	(7)	9.40		1.00	1.00	0.00	0.00	0.43	0.00	0.00	0.00
4-Hydroxytoluene	(7)	5.50	6.33	1.00	1.00	0.00	0.00	0.14	0.00	0.00	0.00
4-Nitrophenol	(7)	7.60		1.00	1.00	0.00	0.00	0.50	0.00	0.00	0.00
Acetathiomide	(7)	4.20		0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Acetic acid	(7)	0.10		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Acetone	(7)	0.10		0.00	0.00	0.00	0.50	0.33	0.00	0.00	0.00
Acetylacetone	(7)	4.00		0.00	0.00	0.00	2.50	0.40	0.00	0.00	0.00
Anisole	(7)	1.00		0.10	0.00	0.00	0.00	0.14	0.00	1.00	0.00
Benzaldehyde	(7)	0.10		0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00
Benzoic acid	(7)	0.30		0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00
Benzothiomide	(7)	4.00		0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Citric acid	(7)	0.80		0.00	0.00	0.00	0.00	1.17	0.00	0.00	0.00

Ethanol	(*7*)	0.10		0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00
Fumaric acid	(*7*)	0.10		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Maleic acid	(*7*)	0.10		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Malonic acid	(*7*)	1.80		0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.00
N,N-Diethylbenzenamine	(*7*)	8.30		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Nitrobenzene	(*7*)	0.10		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
Oxalic acid	(*7*)	0.30		0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00
Phenoxyacetic acid	(*7*)	0.30		0.10	0.00	0.00	0.00	0.38	0.00	1.00	0.00
Phenylthiourea	(*7*)	12.4		1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Propanal	(*7*)	0.20		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
Succinic acid	(*7*)	0.10		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Thiourea	(*7*)	3.90		0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Tyrosine	(*7*)	11.4		1.00	1.00	1.00	0.00	0.33	0.00	0.00	0.00
3-(4-Hydroxy-3,5-dimethoxyphenyl)propanoic acid	(*8*)	6.06		1.00	0.00	0.00	0.00	0.42	0.00	0.00	2.00
3,5-Dihydroxytoluene	(*8*)	6.26	6.39	0.60	2.00	0.00	0.00	0.29	0.00	0.00	0.00
3,5-Dimethoxybenzoic acid	(*8*)	3.00		0.10	0.00	0.00	0.00	0.44	0.00	2.00	0.00
3,5-Dimethoxybenzoic acid	(*8*)	3.00		0.10	0.00	0.00	0.00	0.44	0.00	2.00	0.00
Sinapic acid	(*8*)	6.09	8.77	1.00	1.00	0.00	0.00	0.45	0.00	0.00	2.00
Vanillic acid	(*8*)	5.38	7.75	1.00	1.00	0.00	0.00	0.50	0.00	0.00	1.00

Table S3.2. Leave-Many-Out cross validation data set (N = 50) for chlorine demand (mol-HOCl/mol-Cp)

Compound name	Source	HOCl	AdjHOCl	RAI	ArOH	ACN	Cl	OC	AS	ArORact	ArORnact
1,3-Dihydroxybenzene	(*1*)	7.10		0.60	2.00	0.00	0.00	0.33	0.00	0.00	0.00
1,3-Dihydroxynaphthalene	(*1*)	5.10		0.30	2.00	0.00	0.00	0.20	0.00	0.00	0.00
3,5-Dihydroxytoluene	(*1*)	7.90		0.60	2.00	0.00	0.00	0.29	0.00	0.00	0.00
3,4,5-Trimethoxybenzoic acid	(*2*)	1.10		0.10	0.00	0.00	0.00	0.50	0.00	3.00	0.00
4-Hydroxyacetophenone	(*2*)	8.94		1.00	1.00	0.00	0.50	0.25	0.00	0.00	0.00
4-Hydroxybenzaldehyde	(*2*)	9.12		1.00	1.00	0.00	0.00	0.29	0.00	0.00	0.00
Acetovanillione	(*2*)	8.68		1.00	1.00	0.00	0.50	0.33	0.00	0.00	1.00
Vanillic acid	(*2*)	8.60		1.00	1.00	0.00	0.00	0.50	0.00	0.00	1.00
2-Oxopentanedioic acid	(*3*)	1.40		0.00	0.00	0.00	0.50	1.00	0.00	0.00	0.00
3-Chlorophenol	(*3*)	8.80		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3-hydroxybutyric acid	(*3*)	1.20		0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00
3-Oxopentanedioic acid	(*3*)	5.30		0.00	0.00	0.00	3.00	1.00	0.00	0.00	0.00
Citric acid	(*3*)	0.46		0.00	0.00	0.00	0.00	1.17	0.00	0.00	0.00
Methylacetic acid	(*3*)	0.79		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
2,4-Dichlorophenol	(*4*)	8.10		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3,4,5-Trichlorophenol	(*4*)	8.5		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Chlorophenol	(*4*)	9.80		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
4-Hydroxytoluene	(*4*)	10.5		1.00	1.00	0.00	0.00	0.14	0.00	0.00	0.00
4-Nitrophenol	(*4*)	8.20		1.00	1.00	1.00	0.00	0.50	0.00	0.00	0.00
Aspartic acid	(*5*)	6.10		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Glutamic acid	(*5*)	5.6		0.00	0.00	1.00	0.00	0.80	0.00	0.00	0.00

Glutamine	(*5*)	3.80		0.00	0.00	2.00	0.00	0.60	0.00	0.00	0.00
Isoleucine	(*5*)	6.60		0.00	0.00	1.00	0.00	0.33	0.00	0.00	0.00
Proline	(*5*)	5.60		0.00	0.00	1.00	0.00	0.40	0.00	0.00	0.00
Tyrosine	(*5*)	13.20		1.00	1.00	1.00	0.00	0.33	0.00	0.00	0.00
Valine	(*5*)	2.70	3.43	0.00	0.00	1.00	0.00	0.40	0.00	0.00	0.00
Arginine	(*6*)	8.20		0.00	0.00	3.00	0.00	0.33	0.00	0.00	0.00
Cysteine	(*6*)	6.20	7.87	0.00	0.00	1.00	0.00	0.67	1.00	0.00	0.00
Leucine	(*6*)	2.60	3.30	0.00	0.00	1.00	0.00	0.33	0.00	0.00	0.00
Lysine	(*6*)	3.80	4.83	0.00	0.00	2.00	0.00	0.33	0.00	0.00	0.00
Methionine	(*6*)	6.00		0.00	0.00	1.00	0.00	0.50	1.00	0.00	0.00
Serine	(*6*)	5.30		0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
Threonine	(*6*)	5.60		0.00	0.00	1.00	0.00	0.75	0.00	0.00	0.00
1,3,5-Trihydroxybenzene	(*7*)	9.10		0.50	3.00	0.00	0.00	0.50	0.00	0.00	0.00
2,4,6-Trichlorophenol	(*7*)	6.8	8.02	1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2-Aminophenol	(*7*)	3.9	4.02	0.30	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2-Chlorophenol	(*7*)	9.20		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2-Naphthol	(*7*)	4.40		0.50	1.00	0.00	0.00	0.10	0.00	0.00	0.00
Alanine	(*7*)	2.0	2.54	0.00	0.00	1.00	0.00	0.67	0.00	0.00	0.00
Aniline	(*7*)	8.30		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Asparagine	(*7*)	5.60		0.00	0.00	2.00	0.00	0.75	0.00	0.00	0.00
Butanal	(*7*)	0.20		0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00
Malic acid	(*7*)	0.75		0.00	0.00	0.00	0.00	1.25	0.00	0.00	0.00
Methionine	(*7*)	5.00	6.35	0.00	0.00	1.00	0.00	0.50	1.00	0.00	0.00

Phenol	(*7*)	9.80		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
Phenylalanine	(*7*)	2.0	2.54	0.00	0.00	1.00	0.00	0.22	0.00	0.00	0.00
3,5-dihydroxybenzoic acid	(*8*)	7.06		0.60	2.00	0.00	0.00	0.57	0.00	0.00	0.00
Ferullic acid	(*8*)	7.63	10.99	1.00	1.00	0.00	0.00	0.40	0.00	0.00	1.00
Sinapic acid	(*8*)	9.03		1.00	1.00	0.00	0.00	0.45	0.00	0.00	2.00
Syringic acid	(*8*)	5.10	7.46	1.00	1.00	0.00	0.00	0.56	0.00	0.00	2.00

Table S3.3. External validation data set (N = 42) for chlorine demand (mol-HOCl/mol-Cp)

Compound name	Source	HOCl	AdjHOCl	RAI	ArOH	ACN	Cl	OC	AS	ArORact	ArORnact
Manose	(*9*)	1.30		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
5-Methylfurfural	(*9*)	0.80		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
Arabinose	(*9*)	0.4		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
3,4,5-Triethoxybenzyl alcohol	(*2*)	0.55		0.10	0.00	0.00	0.00	0.31	0.00	3.00	0.00
3,4,5-Trimethoxybenzyl alcohol	(*2*)	1.52		0.10	0.00	0.00	0.00	0.40	0.00	3.00	0.00
3-Aminobenzoic acid	(*2*)	7.74		1.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00
4-Hydroxybenzoic acid	(*2*)	9.48		1.00	1.00	0.00	0.00	0.43	0.00	0.00	0.00
Coniferyl alcohol	(*2*)	6.66		1.00	1.00	0.00	0.00	0.30	0.00	0.00	1.00
Methylsyringate	(*2*)	7.11		1.00	1.00	0.00	0.00	0.50	0.00	0.00	2.00
p-Cummaric acid	(*2*)	9.29		1.00	1.00	0.00	0.00	0.33	0.00	0.00	0.00
Sinapyl alcohol	(*2*)	6.14		1.00	1.00	0.00	0.00	0.36	0.00	0.00	2.00
Trans-3,5-dimethoxy-4-hydroxycinnamate	(*2*)	9.59		1.00	1.00	0.00	0.00	0.38	0.00	0.00	2.00
Vanillin	(*2*)	7.92		1.00	1.00	0.00	0.00	0.38	0.00	0.00	1.00
2-Oxobutyric acid	(*3*)	1.10		0.00	0.00	0.00	0.50	0.75	0.00	0.00	0.00
3-Oxohexanedioic acid	(*3*)	5.80		0.00	0.00	0.00	2.00	0.83	0.00	0.00	0.00
N-Acetylneuraminic acid	(*3*)	2.9		0.00	0.00	1.00	0.00	0.82	0.00	0.00	0.00
Citraconic acid	(*3*)	0.10		0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00
2,3,4,6-Tetrachlorophenol	(*4*)	7.20		1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
Ornithinechlorohydrate	(*6*)	4.60		0.00	0.00	2.00	0.00	0.40	0.00	0.00	0.00
β-Alanine	(*6*)	2.80		0.00	0.00	1.00	0.00	0.67	0.00	0.00	0.00
1,2,3-Trihydroxybenzene	(*7*)	6.90		0.50	3.00	0.00	0.00	0.50	0.00	0.00	0.00

1,2-Dihydroxybenzene	(*7*)	4.10	0.30	2.00	0.00	0.00	0.33	0.00	0.00	0.00
1,4-Dihydroxybenzene	(*7*)	3.30	0.30	2.00	0.00	0.00	0.33	0.00	0.00	0.00
2,3,6-Trichlorophenol	(*7*)	6.90	1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2,3-Dichlorophenol	(*7*)	8.00	1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
2-hydroxyacetophenone	(*7*)	9.90	1.00	1.00	0.00	0.50	0.25	0.00	0.00	0.00
2-Hydroxybenzaldehyde	(*7*)	9.70	1.00	1.00	0.00	0.00	0.29	0.00	0.00	0.00
2-Nitrophenol	(*7*)	9.60	1.00	1.00	1.00	0.00	0.50	0.00	0.00	0.00
3,5-Dichlorophenol	(*7*)	7.60	1.00	1.00	0.00	0.00	0.17	0.00	0.00	0.00
3-hydroxyacetophenone	(*7*)	11.00	1.00	1.00	0.00	0.50	0.25	0.00	0.00	0.00
3-hydroxytoluene	(*7*)	8.70	1.00	1.00	0.00	0.00	0.14	0.00	0.00	0.00
4,6-Dichloro-1,3-dihydroxybenzene	(*7*)	5.00	0.60	2.00	0.00	0.00	0.25	0.00	0.00	0.00
4-Chloro-1,3-dihydroxybenzene	(*7*)	6.10	0.60	2.00	0.00	0.00	0.33	0.00	0.00	0.00
4-Methoxyphenol	(*7*)	3.40	0.50	1.00	0.00	0.00	0.29	0.00	0.00	1.00
Acetophenone	(*7*)	0.50	0.00	0.00	0.00	0.50	0.13	0.00	0.00	0.00
Benzamide	(*7*)	2.50	0.00	0.00	1.00	0.00	0.14	0.00	0.00	0.00
Ethylaceto acetate	(*7*)	2.00	0.00	0.00	0.00	2.50	0.50	0.00	0.00	0.00
Phenylacetic acid	(*7*)	0.10	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00
Pyruvic acid	(*7*)	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
Urea	(*7*)	3.80	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00
3-(3,4,5-trimethoxyphenyl) propionic acid	(*8*)	1.32	0.10	0.00	0.00	0.00	0.42	0.00	3.00	0.00
3-(-4-hydroxy-3-methoxyphenyl)propanoic acid	(*8*)	4.94	0.50	1.00	0.00	0.00	0.40	0.00	0.00	1.00

Table S3.4. Calibration statistics and coefficient of the 5-fold LMO QSPRs for chlorine demand.

	Regression statistics				Descriptors							
QSPRs	**R**	**R^2**	**AdjR^2**	**SDE**	**RAI**	**ArOH**	**ACN**	**Cl**	**OC**	**AS**	**ArORact**	**ArORnact**
Set-1	0.93	0.87	0.85	1.26	7.81	0.92	3.08	1.19	0.85	1.96	0.40	-0.68
Set-2	0.95	0.90	0.88	1.12	7.60	1.28	2.97	1.38	0.88	2.04	0.28	-0.67
Set-3	0.91	0.82	0.80	1.35	7.62	1.04	2.87	0.87	1.36	2.85	0.63	-0.98
Set-4	0.92	0.85	0.83	1.29	7.47	1.30	3.02	1.38	0.90	2.52	0.73	-0.60
Set-5	0.94	0.88	0.86	1.20	7.52	1.25	3.05	1.31	1.06	2.50	0.40	-0.68
Total	3.72	3.45	3.38	4.95	38.03	5.79	15.00	6.13	5.06	11.87	2.43	-3.62
Mean	**0.93**	**0.86**	**0.84**	**1.24**	**7.61**	**1.16**	**3.00**	**1.23**	**1.01**	**2.37**	**0.49**	**-0.72**
Stdev	0.02	0.04	0.04	0.10	0.13	0.17	0.08	0.21	0.21	0.37	0.19	0.15
All (N = 159)	0.93	0.86	0.85	1.24	7.68	1.08	2.97	1.19	1.04	2.44	0.34	-0.64
All (N = 201)	0.93	0.87	0.86	1.21	7.75	0.93	2.91	1.21	1.03	2.50	0.25	-0.68

Table S3.5. Standard errors of descriptors for the five-fold LMO calibration for chlorine demand

	Descriptors								
QSPRs	**RAI**	**ArOH**	**ACN**	**Cl**	**OC**	**AS**	**ArORact**	**ArORnact**	**Mean**
Set-1	0.34	0.25	0.22	0.19	0.28	0.55	0.15	0.25	5.51
Set-2	0.30	0.22	0.19	0.19	0.25	0.49	0.14	0.22	5.60
Set-3	0.33	0.23	0.20	0.31	0.31	0.58	0.22	0.42	6.04
Set-4	0.41	0.31	0.19	0.23	0.28	0.51	0.22	0.28	5.58
Set-5	0.34	0.27	0.21	0.21	0.29	0.57	0.15	0.24	5.62
Total	1.72	1.28	1.00	1.13	1.41	2.70	0.87	1.41	28.35
Mean	**0.34**	**0.26**	**0.20**	**0.23**	**0.28**	**0.54**	**0.17**	**0.28**	**5.67**
Stdev	0.04	0.04	0.01	0.05	0.02	0.04	0.04	0.08	0.21
Training (N = 159)	0.27	0.19	0.16	0.17	0.23	0.44	0.14	0.22	5.73
All data (N = 201)	0.23	0.16	0.15	0.16	0.21	0.42	0.12	0.18	5.57

Table S3.6. Predictive power for 5-fold LMO_{CV} and LOO_{CV} for chlorine demand

Data sets (N = 50)	q^2	MBD	RMSE	R_i^2	k	R_o^2	k'	R_t
Set-1	0.83	-3.86	1.29	0.84	0.95	0.84	0.96	0.000
Set-2	0.75	-2.23	1.50	0.77	0.89	0.76	0.96	0.008
Set-3	0.90	3.41	1.19	0.91	0.80	0.86	0.95	0.057
Set-4	0.86	4.03	1.25	0.87	0.85	0.83	1.00	0.038
Set-5	0.81	3.69	1.35	0.82	0.81	0.77	0.99	0.060
Total	4.17	5.03	6.58	4.21	4.30	4.07	4.85	0.163
Mean	**0.83**	**1.01**	**1.32**	**0.84**	**0.86**	**0.81**	**0.97**	**0.033**
Stdev	0.06	3.75	0.12	0.05	0.06	0.04	0.02	0.027
LOO_{cv} (N = 159)	**0.85**	**-0.55**	**1.28**	**0.85**	**0.88**	**0.84**	**0.97**	**0.013**

Table S3.7. Predictive power for LMO_{CV} using average QSPR for chlorine demand.

Data sets	q^2	MBD	RMSE	R_i^2	K	R_o^2	k'	R_t
Set-1	0.84	-1.69	1.26	0.85	0.95	0.85	0.98	0.001
Set-2	0.78	-1.76	1.43	0.78	0.87	0.77	0.96	0.014
Set-3	0.94	3.63	0.94	0.95	0.85	0.92	0.97	0.030
Set-4	0.89	2.43	1.12	0.89	0.87	0.87	0.99	0.024
Set-5	0.82	2.34	1.33	0.82	0.81	0.78	0.98	0.054
Total	4.27	4.96	6.08	4.29	4.36	4.18	4.87	0.122
Mean	**0.85**	**0.99**	**1.22**	**0.86**	**0.87**	**0.84**	**0.97**	**0.024**
Stdev	0.06	2.53	0.19	0.06	0.05	0.06	0.01	0.020
Training (N = 159)	**0.86**	**-0.28**	**1.21**	**0.86**	**0.88**	**0.85**	**0.97**	**0.015**
All data (N = 201)	0.87	1.90	1.21	0.87	0.88	0.85	0.98	0.020

Table S3.8. Predictive power for the external data using QSPRs for chlorine demand

Each QSPR	q^2	MBD	RMSE	R_i^2	k	R_o^2	k'	R_t
Set-1	0.90	8.56	1.09	0.91	0.91	0.89	1.03	0.026
Set-2	0.88	11.64	1.16	0.91	0.93	0.88	1.06	0.027
Set-3	0.88	10.60	1.22	0.89	0.86	0.84	1.03	0.054
Set-4	0.86	13.22	1.27	0.89	0.89	0.85	1.06	0.053
Set-5	0.88	12.45	1.19	0.91	0.91	0.87	1.06	0.041
Total	4.40	56.47	5.92	4.51	4.50	4.33	5.24	0.201
Mean	**0.88**	**11.29**	**1.18**	**0.90**	**0.90**	**0.87**	**1.05**	**0.040**
Stdev	0.01	1.81	0.07	0.01	0.03	0.02	0.01	0.014
Average QSPR	**0.88**	**11.42**	**1.17**	**0.91**	**0.90**	**0.87**	**1.05**	**0.039**

Table S4.1. TOX formation (mol-Cl/mol-Cp) for compounds in training data set (N = 49

Compound name	Source	TOX	CI	ArOH:C	√HeA	logH:C
1,3-Dihydroxybenzene	(*3*)	3.00	2.00	0.33	1.41	0.00
1,4-Phenyldiamine	(*2*)	0.03	0.00	0.00	*1.41*	0.12
2-Aminobenzoic acid	(*2*)	0.45	0.00	0.00	1.73	0.00
2-Oxobutanedioic acid	(*3*)	0.55	1.50	0.00	2.24	0.00
2-Oxobutyric acid	(*3*)	0.30	0.50	0.00	1.73	0.18
2-Oxopentanedioic acid	(*3*)	0.34	0.50	0.00	2.24	0.18
3-(3,4,5-trimethoxyphenyl) propionic acid	(*2*)	1.19	0.00	0.00	2.24	0.12
3,4,5-Triethoxybenzoic acid	(*2*)	0.66	0.00	0.00	2.24	0.14
3,4,5-Trimethoxyacetophenone	(*2*)	0.60	0.50	0.00	2.00	0.10
3,4,5-Trimethoxybenzamide	(*2*)	0.30	0.00	0.00	2.24	0.11
3,4,5-Trimethoxybenzoic acid	(*2*)	0.97	0.00	0.00	2.00	0.08
3,4,5-Trimethoxybenzyl alcohol	(*2*)	0.99	0.00	0.00	2.00	0.15
3,4,5-Trimethoxyphenyl acetonitrile	(*2*)	0.85	0.00	0.00	2.00	0.07
3-Hydroxybutyric acid	(*3*)	0.93	0.00	0.00	1.73	0.30
3-Oxohexanedioic acid	(*3*)	2.30	2.50	0.00	2.24	0.12
4-(3,4,5-trimethoxybenzoyl) butyric acid	(*2*)	0.30	0.50	0.00	2.45	0.11
4,6-dioxoheptanoic acid	(*3*)	2.50	3.50	0.00	2.00	0.15
4-Allyl-2,6-dimethoxyphenol	(*2*)	0.72	0.00	0.10	1.73	0.15
4-Aminobenzoic acid	(*2*)	0.58	0.00	0.00	1.73	0.00
4-Hydroxybenzaldehyde	(*2*)	1.30	0.00	0.14	1.41	-0.07
4-Hydroxybenzoic acid	(*2*)	1.17	0.00	0.14	1.73	-0.07
5,7-Dioxooctanoic acid	(*2*)	2.80	3.50	0.00	2.00	0.18
Acetovanillione	(*2*)	0.96	0.50	0.11	1.73	0.05
Alanine	(*6*)	0.01	0.00	0.00	1.73	0.30
Aniline	(*2*)	0.40	0.00	0.00	1.00	0.07
Arginine	(*6*)	0.11	0.00	0.00	2.45	0.37
Citraconic acid	(*3*)	0.10	0.00	0.00	2.00	0.08
Coniferyl alcohol	(*2*)	2.07	0.00	0.10	1.73	0.08
Ethyl-(3,4,5-trimethoxybenzyl) acetate	(*2*)	1.20	0.00	0.00	2.45	0.15
Ferulic acid	(*2*)	1.99	0.00	0.10	2.00	0.00
Glutamine	(*6*)	0.04	0.00	0.00	2.24	0.30
Isoleucine	(*6*)	0.02	0.00	0.00	1.73	0.34

Leucine	(*6*)	0.00	0.00	0.00	1.73	0.34
Lysine	(*6*)	0.04	0.00	0.00	2.24	0.37
Methionine	(*6*)	0.14	0.00	0.00	2.00	0.34
Methoxyacetic acid	(*3*)	0.35	0.00	0.00	1.73	0.30
Ornithine	(*6*)	0.06	0.00	0.00	1.73	0.34
p-Cummaric acid	(*2*)	1.29	0.00	0.11	1.73	-0.05
Phenylalanine	(*6*)	0.02	0.00	0.00	1.73	0.16
Proline	(*6*)	0.32	0.00	0.00	1.73	0.26
Serine	(*6*)	0.01	0.00	0.00	2.00	0.37
Sinapyl alcohol	(*2*)	0.51	0.00	0.09	2.00	0.10
Syringaldehyde	(*2*)	2.07	0.00	0.11	2.00	0.05
Syringic acid	(*2*)	0.70	0.00	0.11	2.24	0.05
Trans-3,5-dimethoxy-4-hydroxycinnamate	(*2*)	0.95	0.00	0.08	2.24	0.07
Tyrosine	(*6*)	1.50	0.00	0.11	2.00	0.09
Valine	(*6*)	0.01	0.00	0.00	1.73	0.40
Vanillin	(*2*)	1.25	0.00	0.13	1.73	0.00
β-Alanine	(*6*)	0.07	0.00	0.00	*1.73*	0.37

Table S4.2. TOX formation (mol-Cl/mol-Cp) for external validation data set (N = 12)

Compound name	**Source**	**TOX**	**Cl**	**ArOH:C**	**√HeA**	**logH:C**
2-Aminophenol	(*2*)	0.31	0.00	0.17	1.41	0.07
3-Aminobenzoic acid	(*2*)	0.45	0.00	0.00	1.73	0.00
3-Oxopentanedioic acid	(*3*)	2.90	3.00	0.00	2.24	0.08
4-Hydroxyacetophenone	(*2*)	1.56	0.50	0.13	1.41	0.00
Acetosyringone	(*2*)	2.02	0.50	0.10	2.00	0.08
Cysteine	(*6*)	0.46	0.00	0.00	2.00	0.37
Glutamic acid	(*6*)	0.03	0.00	0.00	2.24	0.20
p-Cummaric acid	(*2*)	1.29	0.00	0.11	1.73	-0.05
Sinapic acid	(*2*)	0.48	0.00	0.09	2.24	0.04
Threonine	(*6*)	0.29	0.00	0.00	2.00	0.30
Vanillic acid	(*2*)	1.26	0.00	0.13	2.00	0.00
3,4,5-trimethoxyphenylacetic acid	(*2*)	1.07	0.00	0.00	2.24	0.10

Table S5.1. LMO calibration data set (N = 60) for TCM formation (mol-TCM/mol-Cp)

Compound name	Source	TCM	CI	OTactC	EDCORH
3-Oxopentanedioic acid	(*10*)	0.9440	3.00	0.00	-0.20
1,3-dihydroxybenzene	(*9*)	0.9250	2.00	1.00	0.00
2,6-Dihydroxybenzoic Acid	(*10*)	0.9000	2.00	1.00	0.00
1,3-Dihydroxybenzene	(*1*)	0.8300	2.00	1.00	0.00
3-Oxohexanedioic Acid	(*3*)	0.8300	2.00	0.00	-0.17
2,4-Dihydroxybenzoic Acid	(*1*)	0.8000	2.00	1.00	0.00
3,5-Dihydroxybenzoic Acid	(*10*)	0.7800	2.00	1.00	0.00
3,5-Dihydroxytoluene	(*1*)	0.6900	2.00	1.00	0.00
1,3-Dihydroxynaphthalene	(*1*)	0.6800	2.00	1.00	0.00
3-Chlorophenol	(*4*)	0.3200	0.00	1.00	0.17
2.3,4,6-Tetrachlorophenol	(*4*)	0.2900	0.00	1.00	0.17
2,4,5-Trichlorophenol	(*4*)	0.2800	0.00	1.00	0.17
Maleic acid	(*10*)	0.1400	0.00	0.00	0.00
4-Chlorophenol	(*4*)	0.1100	0.00	0.00	0.17
Phenol	(*4*)	0.1100	0.00	0.00	0.17
Acetovanillione	(*2*)	0.0980	0.50	0.00	0.00
Tyrosine	(*9*)	0.0830	0.00	0.00	0.11
Aniline	(*2*)	0.0690	0.00	0.00	0.17
Phenol	(*3*)	0.0600	0.00	0.00	0.17
2-Aminobenzoic acid	(*4*)	0.0570	0.00	0.00	0.14
4-Iodophenol	(*4*)	0.0500	0.00	0.00	0.17
Ferulic acid	(*9*)	0.0470	0.00	0.00	0.10
Vanillic acid	(*2*)	0.0400	0.00	0.00	0.13
3-Aminobenzoic acid	(*2*)	0.0370	0.00	0.00	0.14
3-(4-Hydroxy-3,5-Dimethoxyphenyl) Propanoic Acid	(*8*)	0.0330	0.00	0.00	0.09
1,4-Cyclohexanedione	(*11*)	0.0300	1.00	0.00	-0.25
4-Hydroxybenzoic acid	(*2*)	0.0300	0.00	0.00	0.14
Sinapic Acid	(*9*)	0.0300	0.00	0.00	0.09
4-(3,4,5-Trimethoxybenzoyl) Butyric Acid	(*2*)	0.0280	0.50	0.00	-0.07
2-Aminophenol	(*2*)	0.0240	0.00	0.00	0.33
2-Hydroxybenzoic Acid	(*10*)	0.0210	0.00	0.00	0.14

Syringaldehyde	(*2*)	0.0210	0.00	0.00	0.00
3-Oxobutanedioic acid	(*10*)	0.0170	1.50	0.00	-0.25
Trans-3,5-dimethoxy-4-hydroxycinnamate	(*2*)	0.0170	0.00	0.00	0.17
Sinapic acid	(*2*)	0.0160	0.00	0.00	0.09
1,2,3-Trihydroxybenzene	(*11*)	0.0100	0.00	0.00	0.50
Aspartic acid	(*6*)	0.0080	0.00	0.00	0.00
Threonine	(*6*)	0.0070	0.00	0.00	0.00
Syringic acid	(*8*)	0.0050	0.00	0.00	0.11
4-Oxoheptanedioic acid	(*3*)	0.0050	1.00	0.00	-0.14
Asparagine	(*6*)	0.0050	0.00	0.00	0.00
Lysine	(*6*)	0.0050	0.00	0.00	0.00
Citraconic acid	(*3*)	0.0040	0.00	0.00	0.00
3,4,5-Triethoxybenzyl alcohol	(*2*)	0.0030	0.00	0.00	0.00
Arginine	(*6*)	0.0030	0.00	0.00	0.00
Isoleucine	(*5*)	0.0023	0.00	0.00	0.00
Phenylalanine	(*5*)	0.0023	0.00	0.00	0.00
Leucine	(*5*)	0.0021	0.00	0.00	0.00
3,4,5-Trimethoxybenzoic acid	(*2*)	0.0020	0.00	0.00	0.00
3,4,5-Trimethoxyphenylacetic acid	(*2*)	0.0020	0.00	0.00	0.00
Serine	(*6*)	0.0020	0.00	0.00	0.00
Valine	(*5*)	0.0016	0.00	0.00	0.00
Lysine	(*5*)	0.0011	0.00	0.00	0.00
Alanine	(*6*)	0.0010	0.00	0.00	0.00
Glutamic acid	(*6*)	0.0010	0.00	0.00	0.00
Serine	(*5*)	0.0006	0.00	0.00	0.00
Cysteine	(*5*)	0.0004	0.00	0.00	0.00
Glutamine	(*5*)	0.0003	0.00	0.00	0.00
Proline	(*5*)	0.0003	0.00	0.00	0.00
Methionine	(*5*)	0.0001	0.00	0.00	0.00

Table S5.2. LMO-cross validation data set (N = 30) for TCM formation (mol-TCM/mol-Cp)

Compound name	Source	TCM	CI	OTactC	EDCORH
2,4-dihydroxybenzoic acid	(*10*)	0.9200	2.00	1.00	0.00
5,7-dioxooctanoic acid	(*3*)	0.9100	3.50	0.00	-0.25
1,3-dihydroxybenzene	(*8*)	0.8770	2.00	1.00	0.00
1,3,5-trihydroxybenzene	(*11*)	0.8600	2.00	3.00	0.00
3,5-Dihydroxytolune	(*8*)	0.8520	2.00	1.00	0.00
3,5 dihydroxybenzoic acid	(*11*)	0.7400	2.00	1.00	0.00
3-Oxopentanedioic acid	(*3*)	0.7100	3.00	0.00	-0.20
4-Hydroxyacetophenone	(*2*)	0.1030	0.50	0.00	0.00
2,4-Dichlorophenol	(*4*)	0.1000	0.00	0.00	0.17
4-Aminobenzoic acid	(*2*)	0.0600	0.00	0.00	0.14
4-Nitrophenol	(*4*)	0.0500	0.00	0.00	0.17
Tyrosine	(*5*)	0.0414	0.00	0.00	0.11
Vanilic acid	(*8*)	0.0378	0.00	0.00	0.13
3-Hydroxybutyric acid	(*3*)	0.0290	0.00	0.00	0.00
Ferulic acid	(*2*)	0.0290	0.00	0.00	0.10
Sinapic acid	(*8*)	0.0228	0.00	0.00	0.09
1,4-dihydroxybenzene	(*11*)	0.0100	0.00	0.00	0.33
3-(-4-hydroxy-3-methoxyphenyl) Propanoic acid	(*8*)	0.0098	0.00	0.00	0.10
3-Oxobutanedioic acid	(*3*)	0.0060	1.50	0.00	-0.25
Syringic acid	(*2*)	0.0060	0.00	0.00	0.11
3,4,5-Trimethoxybenzyl alcohol	(*2*)	0.0030	0.00	0.00	0.00
Aspartic acid	(*5*)	0.0017	0.00	0.00	0.00
Asparagine	(*5*)	0.0010	0.00	0.00	0.00
2-Oxobutyric acid	(*3*)	0.0010	0.50	0.00	-0.25
Isoleucine	(*6*)	0.0010	0.00	0.00	0.00
Leucine	(*6*)	0.0010	0.00	0.00	0.00
Glutamic acid	(*5*)	0.0010	0.00	0.00	0.00
Arginine	(*5*)	0.0009	0.00	0.00	0.00
Alanine	(*5*)	0.0006	0.00	0.00	0.00
Threonine	(*5*)	0.0002	0.00	0.00	0.00

Table S5.3. External validation data set (N = 27) for TCM formation (mol-TCM/mol-Cp)

Compound name	Source	TCM	CI	OTactC	EDCORH
2,4,6-Trihydroxybenzoic acid	(*11*)	1.1400	2.00	1.00	0.00
2,4-Pentanedione	(*11*)	0.9600	4.00	0.00	-0.33
1,3-Cyclohexanedione	(*11*)	0.9000	3.00	0.00	-0.25
4,6-Dioxoheptanoic acid	(*3*)	0.8900	3.50	0.00	-0.29
Acetosyringone	(*2*)	0.3070	0.50	0.00	0.00
3,4,5-Trimethoxyacetophenone	(*2*)	0.1210	0.50	0.00	-0.09
2,4,6-Trichlorophenol	(*4*)	0.1000	0.00	0.00	0.17
2-Chlorophenol	(*4*)	0.1000	0.00	0.00	0.17
4-Cyanophenol	(*4*)	0.0700	0.00	0.00	0.14
Vanillin	(*2*)	0.0400	0.00	0.00	0.00
4-Allyl-2,6-dimethoxyphenol	(*2*)	0.0310	0.00	0.00	0.10
4-Hydroxybenzaldehyde	(*2*)	0.0290	0.00	0.00	0.00
3-(3,4,5-trimethoxyphenyl) Propionic acid	(*2*)	0.0200	0.00	0.00	0.00
4-Hydroxytoluene	(*4*)	0.0200	0.00	0.00	0.14
Coniferyl alcohol	(*2*)	0.0140	0.00	0.00	0.10
Sinapyl alcohol	(*2*)	0.0140	0.00	0.00	0.09
Citric acid	(*3*)	0.0120	0.00	0.00	0.00
p-Cummaric acid	(*2*)	0.0090	0.00	0.00	0.11
3,4,5-Trimethoxyphenyl acetonitrile	(*2*)	0.0060	0.00	0.00	0.00
3,4,5-Trimethoxybenzamide	(*2*)	0.0040	0.00	0.00	0.00
Ethyl-(3,4,5-trimethoxybenzyl) acetate	(*2*)	0.0030	0.00	0.00	0.00
Ornithine	(*6*)	0.0030	0.00	0.00	0.00
2-Oxopentanedioic acid	(*3*)	0.0020	0.50	0.00	-0.25
3,4,5-Triethoxybenzoic acid	(*2*)	0.0020	0.00	0.00	0.00
B-Alanine	(*6*)	0.0020	0.00	0.00	0.00
1,4-Phenyldiamine	(*2*)	0.0010	0.00	0.00	0.33
Methoxyacetic acid	(*3*)	0.0010	0.00	0.00	0.00

Table S5.4. QSPR calibration by Leave-Many-Out and entire data for TCM formation

	Statistics of model fit			Coefficients of descriptors			Standard errors		
	R^2	$AdjR^2$	SDE	CI	OTactC	EDCORH	CI	OTactC	EDCORH
Model-1	0.91	0.89	0.09	0.29	0.23	0.33	0.02	0.04	0.10
Model-2	0.94	0.92	0.07	0.27	0.30	0.34	0.01	0.03	0.09
Model-3	0.94	0.92	0.08	0.27	0.27	0.33	0.02	0.04	0.09
Model-4	0.93	0.91	0.08	0.27	0.27	0.40	0.02	0.04	0.10
Model-5	0.95	0.94	0.07	0.28	0.26	0.37	0.01	0.03	0.09
Sum	4.68	4.58	0.39	1.38	1.33	1.76	0.08	0.18	0.48
Mean	**0.94**	**0.92**	**0.08**	**0.28**	**0.27**	**0.35**	**0.02**	**0.04**	**0.10**
Stdev	0.02	0.02	0.01	0.01	0.03	0.03	0.00	0.00	0.01
All **(N = 90)**	**0.92**	**0.90**	**0.09**	**0.26**	**0.29**	**0.24**	**0.01**	**0.03**	**0.08**

Table S5.5. Cross validation of each QSPR and average QSPR for TCM formation

	Q^2	MBD	RMSE	R_i^2	k_i	b	R_o^2	k_o'	R_t
Model-1	0.95	6.25	0.08	0.95	0.94	0.03	0.95	0.98	0.004
Model-2	0.90	11.25	0.11	0.90	0.94	0.04	0.89	0.99	0.010
Model-3	0.91	7.58	0.09	0.91	0.92	0.03	0.91	0.96	0.007
Model-4	0.93	0.53	0.09	0.93	0.87	0.02	0.93	0.90	0.006
Model-5	0.88	5.29	0.11	0.88	0.84	0.03	0.87	0.89	0.014
Sum	4.57	30.91	0.46	4.58	4.52	0.15	4.54	4.72	0.041
Mean	**0.91**	**6.18**	**0.09**	**0.92**	**0.90**	**0.03**	**0.91**	**0.94**	**0.008**
Stdev	0.03	3.89	0.01	0.03	0.04	0.01	0.03	0.04	0.004
Avg model									
LMO_{cv} (N = 90)	**0.92**	**6.90**	**0.09**	**0.92**	**0.91**	**0.03**	**0.91**	**0.95**	**0.007**
LOO_{cv} (N = 90)	**0.92**	**5.25**	**0.08**	**0.92**	**0.91**	**0.03**	**0.92**	**0.94**	**0.006**

Table S5.6. External validation using each QSPR and average QSPR for TCM formation

	Q^2	MBD	RMSE	R_i^2	k_i	b	R_o^2	k_o	R_t
Model-1	0.94	-10.22	0.08	0.95	0.88	0.00	0.95	0.89	0.0001
Model-2	0.94	-13.97	0.08	0.96	0.85	0.00	0.96	0.85	-0.0012
Model-3	0.94	-13.37	0.08	0.96	0.85	0.00	0.96	0.86	0.0000
Model-4	0.93	-14.78	0.09	0.96	0.82	0.01	0.96	0.83	0.0004
Model-5	0.94	-13.00	0.08	0.96	0.84	0.00	0.96	0.85	0.0001
Sum	4.70	-65.35	0.42	4.78	4.25	0.02	4.78	4.27	-0.0006
Mean	**0.94**	**-13.07**	**0.08**	**0.96**	**0.85**	**0.00**	**0.96**	**0.85**	**-0.0001**
Stdev	0.00	1.73	0.00	0.01	0.02	0.00	0.01	0.02	0.0006
Avg model									
N = 27	0.94	-13.40	0.08	0.96	0.85	0.00	0.96	0.85	0.0003

Table S5.7. Change in statistics of regression and predictive power upon deletion of compounds with SDR > 2.5 or SDR < -2.5.

	Regression statistics				Coefficients of descriptors			Standard errors		
	R	**R^2**	**AdjR_c^2**	**SDE**	**CI**	**OTactC**	**EDCORH**	**CI**	**OTactC**	**EDCORH**
Mean (90)	0.97	0.95	0.93	0.07	0.281	0.223	0.357	0.014	0.027	0.085
Mean (89)	0.97	0.94	0.92	0.07	0.265	0.281	0.350	0.015	0.034	0.089
Mean (87)	0.97	0.95	0.93	0.07	0.281	0.223	0.357	0.014	0.027	0.085
Int validation										
Each QSPR	**Q^2**	**MBD**	**RMSE**	**R_i^2**	**k**	**b**	**R_o^2**	**k'**	**R-ratio**	
Mean (90)	0.91	6.18	0.09	0.92	0.90	0.03	0.91	0.94	0.01	
Mean (89)	0.93	12.37	0.08	0.92	0.90	0.03	0.91	0.94	0.01	
Mean (87)	0.95	6.38	0.06	0.96	0.96	0.02	0.96	0.98	0.00	
Avg QSPR	**q^2**	**MBD**	**RMSE**	**R_i^2**	**k**	**b**	**R_o^2**	**k'**	**R-ratio**	
N = 90	0.92	6.90	0.09	0.92	0.91	0.03	0.91	0.95	0.01	
N = 89	0.94	6.47	0.07	0.94	0.93	0.02	0.94	0.96	0.00	
N = 87	0.96	0.89	0.06	0.94	0.93	0.02	0.94	0.96	0.00	
Ext Validation										
Avg QSPR	**q^2**	**MBD**	**RMSE**	**R_i^2**	**k**	**b**	**R_o^2**	**k'**	**Ratio**	
N = 90	0.93	-12.31	0.09	0.95	0.85	0.00	0.95	0.86	0.00	
N = 89	0.93	-15.79	0.09	0.96	0.82	0.00	0.96	0.83	0.00	
N = 87	0.93	-12.31	0.09	0.95	0.85	0.00	0.95	0.86	0.00	

Table S5.8. QSPR calibration using logTCM LMO data splitting and entire data

	Statistics of model fit				Coefficients of descriptors				Standard errors			
	R	R^2	AdjR^2	StdE	HC	CI	ArORact	ΣHeA/C	HC	CI	ArORact	ΣHeA/C
Model-1	0.84	0.70	0.66	0.59	-1.06	0.55	-0.16	-0.58	0.14	0.09	0.08	0.29
Model-2	0.87	0.75	0.72	0.55	-1.08	0.56	-0.16	-0.68	0.12	0.08	0.09	0.26
Model-3	0.87	0.75	0.72	0.56	-0.90	0.59	-0.18	-1.06	0.15	0.08	0.08	0.30
Model-4	0.86	0.75	0.72	0.56	-1.00	0.55	-0.14	-0.84	0.13	0.08	0.08	0.26
Model-5	0.89	0.79	0.76	0.52	-1.02	0.57	-0.22	-0.82	0.12	0.07	0.08	0.25
Sum	4.33	3.74	3.59	2.77	-5.06	2.81	-0.86	-3.97	0.65	0.40	0.41	1.35
Avg	0.87	0.75	0.72	0.55	-1.01	0.56	-0.17	-0.79	0.13	0.08	0.08	0.27
Stdev	0.02	0.03	0.04	0.03	0.07	0.02	0.03	0.18	0.01	0.01	0.00	0.02
All (90)	0.85	0.73	0.71	0.57	-0.99	0.57	-0.17	-0.85	0.11	0.07	0.07	0.22

Table S5.9. LogTCM Cross validation of each QSPR and average QSPR

	q^2	MBD	RMSE	R_i^2	K	B	R_o^2	k'	R-ratio
Model-1	0.21	-13.80	1.02	0.33	0.49	-0.64	0.19	0.75	0.41
Model-2	0.38	-1.07	0.87	0.44	0.64	-0.53	0.35	0.88	0.20
Model-3	0.26	-1.25	0.89	0.38	0.57	-0.67	0.27	0.89	0.30
Model-4	0.17	-2.75	0.97	0.30	0.51	-0.82	0.11	0.85	0.63
Model-5	0.29	3.73	0.85	0.41	0.64	-0.64	0.30	0.93	0.27
Sum	1.30	-15.14	4.60	1.86	2.85	-3.30	1.22	4.28	1.81
Avg	0.26	-3.03	0.92	0.37	0.57	-0.66	0.24	0.86	0.36
Stdev	0.08	6.50	0.07	0.06	0.07	0.10	0.09	0.07	0.17
Avg QSPR									
All (90)	0.73	-2.13	0.56	0.74	0.79	0.32	0.71	0.93	0.04

Table S5.10. LogTCM External validation using each QSPR and average QSPR

	q^2	MBD	RMSE	R_i^2	K	B	R_o^2	K'	R-ratio
Model-1	0.80	-10.99	0.90	0.25	0.45	-0.71	0.08	0.78	0.69
Model-2	0.77	-1.56	0.96	0.24	0.46	-0.76	0.05	0.81	0.77
Model-3	0.78	-8.26	0.95	0.22	0.46	-0.74	0.06	0.80	0.74
Model-4	0.80	-7.75	0.91	0.25	0.46	-0.75	0.06	0.81	0.76
Model-5	0.79	-4.88	0.94	0.22	0.46	-0.80	0.03	0.83	0.89
Sum	3.94	-33.44	4.67	1.19	2.29	-3.77	0.28	4.04	3.85
Avg	0.79	-6.69	0.93	0.24	0.46	-0.75	0.06	0.81	0.77
Stdev	0.01	3.59	0.03	0.01	0.00	0.03	0.02	0.02	0.07
Avg QSPR									
n = 27	0.80	-10.99	0.90	0.25	0.45	-0.71	0.08	0.78	0.69

Table S6.1. TCAA formation (mol-TCAA/mol-Cp) for compounds in training data set (N = 47)

Compound name	Source	TCAA	logTCAA	√TCAA	H:C	ArOR:C	ArED:C	ArOH:C	√HeA	√RAI	√ArOH
Syringaldehyde	(*2*)	0.47600	-0.32	0.690	1.11	0.22	0.11	0.11	2.00	1.00	1.00
Phenol	(*3*)	0.36000	-0.44	0.600	1.00	0.00	0.17	0.17	1.00	1.00	1.00
Vanillin	(*2*)	0.30500	-0.52	0.552	1.00	0.13	0.13	0.13	1.73	1.00	1.00
Vanillic acid	(*2*)	0.29000	-0.54	0.539	1.00	0.13	0.13	0.13	2.00	1.00	1.00
4-Hydroxybenzoic acid	(*2*)	0.27300	-0.56	0.522	0.86	0.00	0.14	0.14	1.73	1.00	1.00
4-Hydroxybenzaldehyde	(*2*)	0.26300	-0.58	0.513	0.86	0.00	0.14	0.14	1.41	1.00	1.00
Ferulic acid	(*2*)	0.26000	-0.59	0.510	1.00	0.10	0.10	0.10	2.00	1.00	1.00
4-Hydroxyacetophenone	(*2*)	0.25900	-0.59	0.509	1.00	0.00	0.13	0.13	1.41	1.00	1.00
Sinapyl alcohol	(*2*)	0.11100	-0.95	0.333	1.27	0.18	0.09	0.09	2.00	1.00	1.00
Syringic acid	(*2*)	0.10400	-0.98	0.322	1.11	0.22	0.11	0.11	2.24	1.00	1.00
4-Allyl-2,6-dimethoxyphenol	(*2*)	0.10200	-0.99	0.319	1.40	0.20	0.10	0.10	1.73	1.00	1.00
1,3-Dihydroxybenzene	(*3*)	0.07700	-1.11	0.277	1.00	0.00	0.33	0.33	1.41	0.77	1.41
2-Aminobenzoic acid	(*2*)	0.05700	-1.24	0.239	1.00	0.00	0.14	0.00	1.73	0.87	0.00
Acetovanillione	(*2*)	0.05500	-1.26	0.235	1.11	0.11	0.11	0.11	1.73	1.00	1.00
3,4,5-Trimethoxyphenyl acetonitrile	(*2*)	0.04900	-1.31	0.221	1.18	0.27	0.00	0.00	2.00	0.32	0.00
3,4,5-Triethoxybenzyl alcohol	(*2*)	0.03300	-1.48	0.182	1.54	0.23	0.00	0.00	2.00	0.32	0.00
2-Oxobutyric acid	(*3*)	0.03200	-1.49	0.179	1.50	0.00	0.00	0.00	1.73	0.00	0.00
3,4,5-Trimethoxybenzyl alcohol	(*2*)	0.02900	-1.54	0.170	1.40	0.30	0.00	0.00	2.00	0.32	0.00
3-Hydroxybutyric acid	(*3*)	0.02900	-1.54	0.170	2.00	0.00	0.00	0.00	1.73	0.00	0.00
Aniline	(*2*)	0.02500	-1.60	0.158	1.17	0.00	0.17	0.00	1.00	1.00	0.00
Citric acid	(*3*)	0.02100	-1.68	0.145	1.33	0.00	0.00	0.00	2.83	0.00	0.00
4-(3,4,5-trimethoxybenzoyl) butyric acid	(*2*)	0.02000	-1.70	0.141	1.29	0.21	0.00	0.00	2.45	0.32	0.00
4-Aminobenzoic acid	(*2*)	0.01900	-1.72	0.138	1.00	0.00	0.14	0.00	1.73	1.00	0.00

2-Aminophenol	(*2*)	0.01700	-1.77	0.130	1.17	0.00	0.33	0.17	1.41	0.55	1.00
3-Aminobenzoic acid	(*2*)	0.01500	-1.82	0.122	1.00	0.00	0.14	0.00	1.73	1.00	0.00
3-Oxopentanedioic acid	(*3*)	0.01400	-1.85	0.118	1.20	0.00	0.00	0.00	2.24	0.00	0.00
3,4,5-Trimethoxybenzoic acid	(*2*)	0.00900	-2.05	0.095	1.20	0.30	0.00	0.00	2.00	0.32	0.00
3-Oxohexanedioic acid	(*3*)	0.00800	-2.10	0.089	1.33	0.00	0.00	0.00	2.24	0.00	0.00
Phenylalanine	(*5*)	0.00639	-2.19	0.080	1.44	0.00	0.00	0.00	1.73	0.00	0.00
3-(3,4,5-Trimethoxyphenyl) propionic acid	(*2*)	0.00600	-2.22	0.077	1.33	0.25	0.00	0.00	2.24	0.32	0.00
3,4,5-Triethoxybenzoic acid	(*2*)	0.00600	-2.22	0.077	1.38	0.23	0.00	0.00	2.24	0.32	0.00
Ethyl-(3,4,5-trimethoxybenzyl) acetate	(*2*)	0.00500	-2.30	0.071	1.43	0.00	0.00	0.00	2.45	0.32	0.00
4,6-Dioxoheptanoic acid	(*3*)	0.00200	-2.70	0.045	1.43	0.00	0.00	0.00	2.00	0.00	0.00
Methoxyacetic acid	(*3*)	0.00200	-2.70	0.045	2.00	0.00	0.00	0.00	1.73	0.00	0.00
Proline	(*5*)	0.00181	-2.74	0.043	1.80	0.00	0.00	0.00	1.73	0.00	0.00
5,7-Dioxooctanoic acid	(*3*)	0.00100	-3.00	0.032	1.50	0.00	0.00	0.00	2.00	0.00	0.00
Citraconic acid	(*3*)	0.00100	-3.00	0.032	1.50	0.00	0.00	0.00	2.00	0.00	0.00
Arginine	(*5*)	0.00048	-3.32	0.022	2.33	0.00	0.00	0.00	2.45	0.00	0.00
Cysteine	(*5*)	0.00045	-3.35	0.021	2.33	0.00	0.00	0.00	2.00	0.00	0.00
Methionine	(*5*)	0.00026	-3.59	0.016	2.20	0.00	0.00	0.00	2.00	0.00	0.00
Glutamine	(*5*)	0.00007	-4.15	0.008	2.00	0.00	0.00	0.00	2.24	0.00	0.00
Isoleucine	(*5*)	0.00007	-4.15	0.008	2.17	0.00	0.00	0.00	1.73	0.00	0.00
Leucine	(*5*)	0.00007	-4.15	0.008	2.17	0.00	0.00	0.00	1.73	0.00	0.00
Glutamic acid	(*5*)	0.00006	-4.22	0.008	1.80	0.00	0.00	0.00	2.24	0.00	0.00
Valine	(*5*)	0.00003	-4.52	0.005	2.50	0.00	0.00	0.00	1.73	0.00	0.00
Serine	(*5*)	0.00001	-5.00	0.003	2.33	0.00	0.00	0.00	2.00	0.00	0.00
Threonine	(*5*)	0.00001	-5.00	0.003	2.00	0.00	0.00	0.00	2.00	0.00	0.00

Table S6.2. TCAA formation (mol-TCAA/mol-Cp) for external validation data set (N = 15)

Compound name	Source	TCAA	logTCAA	√TCAA	H:C	ArOR:C	ArED:C	ArOH:C	√HeA	√RAI	√ArOH
2-Oxobutanedioic acid	(*3*)	0.03	-1.52	0.21	1.00	0.00	0.00	0.00	2.24	0.00	0.00
2-Oxopentanedioic acid	(*3*)	0.01	-2.22	0.17	1.20	0.00	0.00	0.00	2.24	0.00	0.00
3,4,5-Trimethoxyacetophenone	(*2*)	0.10	-1.00	0.08	1.27	0.27	0.00	0.00	2.00	0.32	0.00
3,4,5-Trimethoxybenzamide	(*2*)	0.03	-1.60	0.31	1.30	0.30	0.00	0.00	2.24	0.32	0.00
3,4,5-Trimethoxyphenylacetic acid	(*2*)	0.01	-2.00	0.07	1.27	0.27	0.00	0.00	2.24	1.00	0.00
4-Oxoheptanedioic acid	(*3*)	0.00	-3.00	0.03	1.43	0.00	0.00	0.00	2.24	0.00	0.00
Acetosyringone	(*2*)	0.02	-1.66	0.15	1.20	0.20	0.10	0.10	2.00	1.00	1.00
Alanine	(*5*)	0.00	-3.68	0.01	2.33	0.00	0.00	0.00	1.73	0.00	0.00
Asparagine	(*5*)	0.00	-3.32	0.07	2.00	0.00	0.00	0.00	2.24	0.00	0.00
Aspartic acid	(*5*)	0.00	-2.50	0.06	1.75	0.00	0.00	0.00	2.24	0.00	0.00
Coniferyl alcohol	(*2*)	0.01	-1.89	0.11	1.20	0.10	0.10	0.10	2.00	1.00	1.00
Lysine	(*5*)	0.00	-3.28	0.02	2.33	0.00	0.00	0.00	2.00	0.00	0.00
p-Cummaric acid	(*2*)	0.04	-1.44	0.19	0.89	0.00	0.11	0.11	1.73	1.00	1.00
Sinapic acid	(*2*)	0.08	-1.10	0.28	1.09	0.18	0.09	0.09	2.24	1.00	1.00
Tyrosine	(*5*)	0.04	-1.39	0.20	1.22	0.00	0.11	0.11	2.00	1.00	1.00

Table S.7.1. Relative amounts of TOX formation to chlorine demand for model compounds

Compound Name	Source	Chlorinedemand mol/mol	Chlorinedemand mol-Cl_2/mol-C	TOX formation mol/mol	TOX formation Mol-Cl/mol-C	%TOX
3-(3,4,5-trimethoxyphenyl)Propionic acid	(*2*)	1.32	1.32	1.19	1.19	90.00
3,4,5-Trimethoxybenzoic acid	(*2*)	1.10	1.32	0.97	1.16	88.18
3,4,5-Triethoxybenzoic acid	(*2*)	0.99	0.91	0.66	0.61	66.87
3,4,5-Trimethoxybenzyl alcohol	(*2*)	1.52	1.82	0.99	1.18	64.80
3,4,5-Trimethoxyacetophenone	(*2*)	0.93	1.01	0.60	0.65	64.19
1,4-Phenyldiamine	(*2*)	*0.06*	0.12	0.03	0.07	54.10
Coniferyl alcohol	(*2*)	6.66	7.99	2.07	2.48	31.08
Ethyl-(3,4,5-trimethoxybenzyl) acetate	(*2*)	4.50	3.86	1.20	1.03	26.70
Acetosyringone	(*2*)	7.79	9.35	2.02	2.43	25.98
Syringaldehyde	(*2*)	8.38	11.17	2.07	2.76	24.71
3,4,5-Trimethoxyphenyl acetonitrile	(*2*)	4.38	4.77	0.85	0.92	19.31
Ferulic acid	(*2*)	10.32	12.39	1.99	2.38	19.23
4-Hydroxyacetophenone	(*2*)	8.94	13.41	1.56	2.34	17.49
Vanillin	(*2*)	7.92	11.88	1.25	1.88	15.81
Vanillic acid	(*2*)	8.60	12.90	1.26	1.89	14.68
4-(3,4,5-trimethoxybenzoyl) butyric acid	(*2*)	2.01	1.73	0.30	0.25	14.65
4-Hydroxybenzaldehyde	(*2*)	9.12	15.64	1.30	2.23	14.26
p-Cummaric acid	(*2*)	9.29	12.39	1.29	1.72	13.89
p-Cummaric acid	(*2*)	9.29	12.39	1.29	1.72	13.89
4-Hydroxybenzoic acid	(*2*)	9.48	16.25	1.17	2.01	12.35
Acetovanillione	(*2*)	8.68	11.58	0.96	1.27	11.01
4-Allyl-2,6-dimethoxyphenol	(*2*)	6.80	8.15	0.72	0.86	10.58
Syringic acid	(*2*)	6.93	9.24	0.70	0.93	10.10
Trans-3,5-dimethoxy-4-hydroxycinnamate	(*2*)	9.59	9.59	0.95	0.95	9.95
Sinapyl alcohol	(*2*)	6.14	6.70	0.51	0.56	8.37
2-Aminobenzoic acid	(*2*)	6.04	10.36	0.45	0.77	7.43
4-Aminobenzoic acid	(*2*)	7.90	13.54	0.58	0.99	7.29
2-Aminophenol	(*2*)	4.70	9.39	0.31	0.62	6.58

3-Aminobenzoic acid	(*2*)	7.74	13.27	0.45	0.77	5.78
3,4,5-Trimethoxybenzamide	(*2*)	5.33	6.40	0.30	0.35	5.53
Sinapic acid	(*2*)	9.03	9.85	0.48	0.53	5.35
Aniline	(*2*)	8.58	17.16	0.40	0.80	4.65
Citraconic acid	(*3*)	0.10	0.24	0.10	0.24	100.00
3-hydroxybutyric acid	(*3*)	1.20	3.60	0.93	2.79	77.50
3-Oxopentanedioic acid	(*3*)	5.30	12.72	2.90	6.96	54.72
4,6-dioxoheptanoic acid	(*3*)	4.80	8.23	2.50	4.29	52.08
5,7-dioxooctanoic acid	(*3*)	6.00	9.00	2.80	4.20	46.67
Methoxyacetic acid	(*3*)	0.79	3.16	0.35	1.40	44.30
3-Oxohexanedioic acid	(*3*)	5.80	11.60	2.30	4.60	39.66
1,3-dihydroxybenzene	(*3*)	7.90	15.80	3.00	6.00	37.97
2-Oxobutyric acid	(*3*)	1.10	3.30	0.30	0.90	27.27
2-Oxopentanedioic acid	(*3*)	1.40	4.20	0.34	1.02	24.29
2-Oxobutanedioic acid	(*3*)	3.80	11.40	0.55	1.65	14.47
Tyrosine	(*6*)	13.40	17.87	1.50	2.00	11.19
Proline	(*6*)	5.40	12.96	0.32	0.76	5.89
Cysteine	(*6*)	8.43	33.72	0.46	1.84	5.44
Threonine	(*6*)	5.60	16.80	0.29	0.87	5.20
β-Alanine	(*6*)	2.80	11.20	0.07	0.26	2.36
Methionine	(*6*)	6.00	14.40	0.14	0.33	2.27
Arginine	(*6*)	8.20	16.40	0.11	0.21	1.30
Ornithine	(*6*)	4.60	11.04	0.06	0.14	1.26
Glutamine	(*6*)	3.80	9.12	0.04	0.08	0.92
Glutamic acid	(*6*)	3.26	7.82	0.03	0.06	0.77
Lysine	(*6*)	5.17	10.34	0.04	0.08	0.75
Phenylalanine	(*6*)	3.67	4.89	0.02	0.03	0.65
Isoleucine	(*6*)	3.54	7.08	0.02	0.03	0.42
Alanine	(*6*)	3.80	15.20	0.01	0.05	0.32
Serine	(*6*)	5.30	21.20	0.01	0.05	0.23
Valine	(*6*)	3.67	11.01	0.01	0.02	0.14
Leucine	(*6*)	3.54	7.08	0.00	0.01	0.11

Table S.7.2. Relative amounts of TOX formation to chlorine demand fulvic acid (FA) and humic acid (HA) fractions (*12*).

		Chlorine demand		TOX		
Water source	**Fraction**	**mg-Cl_2/mg-TOC**	**mol-Cl_2/mol-C**	**mg-Cl/mg-TOC**	**Mol-Cl/mol-C**	**%TOX**
Black Lake	FA	1.28	0.22	0.21	0.07	32.50
	HA	2.28	0.39	0.29	0.10	25.26
Coal Creek	FA	1.64	0.28	0.23	0.08	28.29
	HA	2.02	0.34	0.27	0.09	26.53
Ogeechee River	FA	1.52	0.26	0.22	0.07	28.42
	HA	2.12	0.36	0.26	0.09	24.72
Ohio River	FA	1.24	0.21	0.16	0.05	25.97
	HA	2.14	0.36	0.23	0.08	21.68
Missouri River	FA	1.1	0.19	0.14	0.05	24.73
	HA	2.14	0.36	0.23	0.08	21.50

Table S.7.3. Relative amounts of TOX formation to chlorine demand in raw water from different water systems (*12*).

	Chlorine demand		TOX		
Water system	**Mg-Cl_2/mg-TOC**	**mol-Cl/mol-C**	**mg-Cl/mg-TOC**	**mol-Cl/mol-Cl**	**%TOX**
1	2.01	0.34	0.30	0.10	29.75
1	1.66	0.28	0.20	0.07	24.10
2	1.77	0.30	0.24	0.08	27.57
3	1.87	0.32	0.17	0.06	18.40
4	1.77	0.30	0.17	0.06	19.21
5	2.01	0.34	0.22	0.07	21.49
6	2.14	0.36	0.30	0.10	27.85
7	1.58	0.27	0.19	0.07	24.43

References

1. Boyce, S.; Hornig, J. Reaction pathways of trihalomethane formation from the halogenation of dihydroxyaromatic model compounds for humic acid. *Environ. Sci. Technol*. **1983**, 17, 202-211.

2. Bull, R.J.; Reckhow, D.A.; Rotello, V.; Bull, O.M.; Kim, J. Use of toxicological and chemical models to prioritize DBP research; AWWA Research Foundation: 2006.

3. Dickenson, E.V.; Summers, S.; Croué, J-P.; Gallard, A. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. *Environ. Sci. Technol.* **2008**, 42, 3226-3233.

4. Gallard, H.; von Gunten, U. Chlorination of phenols: Kinetics and formation of chloroform. *Environ. Sci. Technol*. **2002**, 36, 884–890.

5. Hong, H.C.; Wong, M.H.; Liang, Y. Amino acids Precursors of trihalomethane and haloacetic acid formation during chlorination. *Arch. Environ. Toxicol*. **2009**, 56, 638-645.

6. Hureiki, L.; Croué, J-P.; Legube, B. Chlorination studies of free and combined amino acids. *Water Res.* **1994**, 28, 2521-2531.

7. de Laat, J.; Merlet, N.; Doré, M. Chlorination of organic compounds: chlorine demand and reactivity in relationship to the trihalomethane formation. *Water Res.* **1982**, 16, 1437-1450.

8. Norwood, D.L.; Johnson, J.D; Chrisman, R.F.; Hass, J.R.; Bobenrieth, M.J. Reactions of chlorine with selected aromatic models of aquatic humic material. *Environ. Sci. Technol.* **1980**, 14(2), 187-189.

9. Bond T.; Henriet O.; Goslan E.H.; Parson S.A.; Jefferson B.. Disinfection byproducts and fractionation behavior of natural organic matter surrogates. *Environ. Sci. Technol.* **2009**, 43, 5982-5989.

10. Larson, R.A.; Rockwell, A.L. Chloroform and chlorophenol production by decarboxylation of natural acids during aqueous chlorination. *Environ. Sci. Technol.* **1979**, 13(3), 325-329.

11. Boyce, S.D.; Hornig, J.F. Formation of chloroform from chlorination diketones and polyhydroxybenzenes in dilute aqueous solutions. In: Jolly et al Water chlorination: Environmental impacts and health. Ann Arbor Science Publishers, Ann Arbor, Michigan: 1979, pp131-140.

12. Singer, P.C. Humic substances as precursors for potentially harmful disinfection byproducts. *Wat. Sci. Tech.* **1999**, 40(9), 25-30