THE PREDICTION OF CHLORINE CONSUMPTION OF ORGANIC MOLECULES IN DRINKING WATER TREATMENT

by

Birgül Karataş Süzergöz

B.Sc. in Biology, İstanbul University, 2008

M.Sc. in Biology Education, İstanbul University, 2009

Submitted to the Institute of Environmental Sciences

in partial fulfillment of the requirements for the degree of

Master of Science

in

Environmental Sciences

Boğaziçi University

2018

THE PREDICTION OF CHLORINE CONSUMPTION OF ORGANIC MOLECULES IN DRINKING WATER TREATMENT

APPROVED BY:

Prof. Dr. Melek Türker Saçan Thesis Advisor	. . . . .
Prof. Dr. Ferhan Çeçen	. . . . .
Prof. Dr. Emine Ubay Çokgör	. . . . .

DATE OF APPROVAL: 04/06/2018

To my daughter

Gamze Zümra Süzergöz

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my beloved thesis advisor Prof. Dr. Melek Türker Saçan for her patience, support and guidance throughout the study. I feel very grateful to her for giving me the chance to study with her. The compilation of this study could not have been possible without her expertise. I would also like to thank to the jury members; Prof. Dr. Ferhan Çeçen and Prof. Dr. Emine Ubay Çokgör for spending their valuable time to evaluate this thesis.

I am also thankful to all the other faculty and staff members of Institute of Environmental Sciences for their kind co-operation and help.

I sincerely thank to my co-workers Serdar Demir and Tuğba Karamanlı for answering my continuous questions in the field of pharmacy. Their contribution and companionship mean a lot to me.

I would like to express my deepest gratitude to my family for their endless love, support and understanding. I have no valuable words to express my gratitude to my daughter Gamze Zümra Süzergöz for her presence and unconditional love, without her support this study could not become a reality.

Finally, I would like to thank everyone who has encouraged me throughout this study. The completion of this study could not have been possible without the participation and assistance of so many people whose names may not all be mentioned. Their contributions are sincerely appreciated and acknowledged.

ABSTRACT

THE PREDICTION OF CHLORINE CONSUMPTION OF ORGANIC MOLECULES IN DRINKING WATER TREATMENT

In the present study, the chlorine demand of a diverse set of organic chemicals present in water bodies was investigated by a quantitative structure-property relationship (QSPR) model. The descriptors required for the model development were obtained by SPARTAN (v.10), DRAGON (v.6.0), and ADMET (v.8.0) software packages. The selection of descriptors was carried out via the tools implemented in QSARINS (v.2.2.1) software. Numerous division trials were performed on the data set as training and test sets which comprise the 80% and 20% of the whole data set, respectively. The generated models were validated internally and externally in line with the Organization of Economic Co-operation and Development (OECD) principles. Six descriptors from DRAGON (v.6.0) and one descriptor from ADMET (v.8.0) constitute the final model. These descriptors stem from various blocks including GETAWAY, WHIM, information indices, molecular properties, simple constitutional, 2D autocorrelation and 2D atom pairs. The predictive ability of the final model was tested using an external data set consisting of various pharmaceuticals and personal care products (PPCPs) with no experimental chlorine demand data. The proposed QSPR model covers structurally 91% of the external chemicals. The AD of the generated model was also strictly defined by the range of descriptors' approach. The predictive ability of the generated model was found to be reliable for most of the tested PPCPs. Antibiotics are the highlighted pharmaceuticals among the tested PPCPs with the highest chlorine demand.

ÖZET

SU ARITIMINDA BULUNAN ORGANIK MOLEKÜLLERİN KLOR TÜKETİMLERİNİN TAHMİNİ

Bu çalışmada, suda bulunan çeşitli organik bileşiklerin klor ihtiyaçları kantitatif yapı-özellik ilişkileri modeli kullanılarak incelenmiştir. Model geliştirmek için gerekli olan tanımlayıcılar SPARTAN (v.10), DRAGON (v.6.0), and ADMET (v.8.0) yazılımları kullanılarak elde edilmiştir. Tanımlayıcıların seçimi QSARINS (v.2.2.1) yazılımında bulunan araçlarla yapılmıştır. Veri setinin sırasıyla %80 i eğitim seti ve %20 si test seti olacak şekilde çok sayıda ayrımı denenmiştir. Oluşturulan modeller Ekonomik İş birliği ve Kalkınma Örgütü'nün belirlediği kriterlere uygun olur ve içsel ve dışsal olarak doğrulanmıştır. DRAGON (v.6.0) yazılımdan gelen 6 adet tanımlayıcı ve ADMET (8.0) yazılımdan gelen bir adet tanımlayıcı en son modeli oluşturmaktadır. Bu tanımlayıcılar geometri-topoloji-atom ağırlıkları birleşiminden oluşan tanımlayıcılar, ağırlıklı bütünsel yapıya bağlı değişebilen tanımlayıcılar, bilgi indeksi tanımlayıcıları, moleküler özelliklere bağlı tanımlayıcılar, basit yapısal tanımlayıcılar, 2 boyutlu öz ilinti ve 2 boyutlu atom çiftlerinden oluşmaktadır. Elde edilen son modelin tahmin kapasitesi klor tüketim verisi bulunmayan çok çeşitli ilaç ve kişisel bakım malzemesini içeren bir dışsal veri set kullanılarak test edilmiştir. Önerilen kantitatif yapı-özellik ilişkileri modeli harici kimyasalları yapısal olarak %91 ini kapsamaktadır. Oluşturulan modelin uygulanabilirlik alanı tanımlayıcı aralığı yaklaşımı ile de kesin olarak tanımlanmıştır. Oluşturulan modelin tahmin kapasitesi test edilen çok sayıda ilaç ve kişisel bakım malzemesi için güvenilir bulunmuştur. Test edilen ilaç ve kişisel bakım ürünleri içerisinde en yüksek klor ihtiyacıyla öne çıkan grup antibiyotiklerdir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....iii
ABSTRACT.....iv
ÖZET.....v
TABLE OF CONTENTS.....vi
LIST OF FIGURES.....viii
LIST OF TABLES.....ix
LIST OF SYMBOLS/ABBREVIATIONS.....x
1. INTRODUCTION.....1
1.1. Purpose of the Study.....4
2. THEORETICAL BACKROUND.....5
2.1. Disinfection.....5
2.1.1. Chlorine.....5
2.1.2. Sodium Hypochlorite.....6
2.1.3. Calcium Hypochlorite.....6
2.1.4. Chlorine Dioxide.....6
2.1.5. Chloramines.....7
2.2. Literature Survey on Chlorine Demand.....7
2.3. Quantitative Structure- Activity/Property Relationship (QSA/PR).....12
2.4. Multiple Linear Regressin.....16
2.5. Pharmaceuticals and Personal Care Products (PPCPs) and Their Impact on the Environment.....17
4. MATERIALS AND METHODS.....19
3.1. Data Sets.....19
3.2. QSA/PR Model Development.....19
3.3. Structure Optimization and Calculation of Descriptors.....20
3.4. Training and Test Set Divisions.....22
3.5. Descriptor Selection.....23
3.6. Validation of a QSPR Model.....25
3.6.1. Internal Validation.....27
3.6.1.1. R^2 (Coefficient of determination).....27
3.6.1.2. Leave-one-out (LOO) cross-validation (Q^2_{LOO}).....27
3.6.1.3. Y-scrambling.....28

3.6.2. External Validation.....28
3.6.2.1. Predictive Squared Correlation Coefficients (Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}).....29
3.6.2.2. Concordance Correlation Coefficient (*CCC*).....29
3.6.2.3. The r^2_m, $\bar{r}^2_m$, Δr^2_m.....30
3.6.2.4. The Golbraikh and Tropsha method.....31
3.6.2.5. $RMSE_{Ext}$ (Root mean square errors of prediction).....32
3.6.2.6. Mean Absolute Error (*MAE*) based criteria…….....32
3.6.2.7. External Validation Based on the Mean Absolute Error.....33
3.7. Selection of the Best QSPR Model.....34
3.8. Applicability Domain.....34
3.9. Predictive Performance of the Generated Models.....36
4. RESULTS AND DISCUSSION.....38
4.1. QSAR Modelling on the Chlorine Demand of Diverse Organic Chemicals.....38
4.2. Analysis of the Applicability Domain.....50
4.3. Prediction of HOCldem values of External Set Chemicals with no data.....59
4.4. Comparison with the Proposed QSPR Model with the Literature Model.....63
5. CONCLUSION.....66
REFERENCES.....68
APPENDIX A: INFORMATION ON THE TR-TS SET DIVISIONS OF THE MODELS……...80
APPENDIX B: INFORMATION ON THE PREDICTED CHLORINE DEMAND VALUES OF EXTERNAL SET CHEMICALS.....88

LIST OF FIGURES

Figure 3.1. Flowchart for the generation of a QSA/(P)R model…….…….....20

Figure 3.2. Pre-filtration step in QSARINS 2.2.1.....24

Figure 3.3. Error types of Xternal validation plus tool.....34

Figure 4.1. The change in R^2 and Q^2 along with the increase in the number of variables…….....39

Figure 4.2. Plot of calculated/predicted vs. observed values of HOCldem for the training/test set compounds by model equation.....……...44

Figure 4.3. Williams plot for the Eq.4.1.…….…….....50

Figure 4.4. Training and Test Set compounds vs hat value graph…….51

Figure 4.5. Graphical representation of Eq. 4.1. for the prediction of external set……...……61

LIST OF TABLES

Table 3.1. Descriptor blocks and types in Dragon 6.0 software……..22

Table 4.1. The developed models, their fit and internal validation parameters……...40

Table 4.2. External Validation Parameters of the generated QSAR models…..……..40

Table 4.3. The results of Xternal Validation Plus tool for the developed model…….....42

Table 4.4. List of Dragon 6.0 and ADMET Predictor 8.0 descriptors appeared in Eq. 4.1…….....45

Table 4.5. Chemicals, their experimental and predicted HOCldem values, descriptor values and residuals.…….....….....….....….....…….....52

Table 4.6. Concentrations of Active Pharmaceutical Ingredients found in finished drinking water worldwide differences method….....59

Table 4.7. The Compounds that fall outside the AD and their Anatomical Therapeutic Chemical Classification codes and predicted HOCldem values.....62

Table 4.8. Range of descriptors appeared in Eq. 4.1.....…….....63

Table 4.9. Average Model Coefficients and Standard Errors for descriptors used in the published model.....64

LIST OF SYMBOLS/ABBREVIATIONS

Symbol	Explanation	Unit
$Å^2$	CPK Area	
$Å^3$	CPK Volume	
B02[C-N]	Presence/absence of C-N at topological distance 2	
E	Gas-phase Energy	eV
E1e	1^{st} component accessibility directional WHIM index/ weighed by Sanderson electronegativity	
E_{aq}	Aqueous-phase Energy	eV
E_{HOMO}	Energy of the Highest Occupied Molecular Orbital	eV
E_{LUMO}	Energy of the Lowest Unoccupied Molecular Orbital	eV
f	Function	
*h**	Critical Hat Value	
MATS3s	Moran autocorrelation of lag 3 weighed by I-state	
NaOCl	Sodium Hypochlorite	
nArOH	Number of aromatic -OH groups	
NH_3	Ammonia	
O:C	Oxygen carbon atomic ratio	
R^2	Coefficient of Determination	
R^2_{adj}	Adjusted (for degrees of freedom) squared correlation coefficient	
R8v+	R maximal autocorrelation of lag 8/weighted by van der Waals volume	
RAI	Ring activation index	
V_x	McGowan volume	
WHIM	Weighted Holistic Invariant Molecular Descriptors	

Symbol	Explanation	Unit
ϕ	Biological System	
μ	Dipole Moment	
ω	Electrophilicity Index	
σ	Hammet Sigma	
η	Hardness	
σ	Softness	

Abbreviation	Explanation	Unit
ACN	Aliphatic C-bonded NH_2	
AD	Applicability Domain	
ANN	Artificial Neural Networks	
API	Active Pharmaceutical Ingredient	
ArHdrxl_-OH	Number of aromatic hydroxyl groups	
ArOH	Number of Phenols	
ArORnonact	Alkoxy groups attached to the aromatic ring with NH2 and OH	
ArORact	Alkoxy groups attached to the aromatic ring without NH_2 and OH	
AS	Sulfur in the Carbon Chain	
BE	Backward Elimination	
C	Chemical Constitution	
CAS	Chemical Abstracts Service	
CI	Carbonyl Index	
CCC	Concordance Correlation Coefficient	
DBP	Disinfection by-products	
DOM	Dissolved Organic Matter	
DXAA	Dichloro- or Dibromo Acetic Acid	
ECHA	European Chemical Agency	
ECVAM	European Centre for the Validation of Alternative Methods	
EPA	Environmental Protection Agency	
FS	Forward Selection	
GA	Genetic Algorithm	
GABA	Gamma Amino Butyric Acid	
GETAWAY	GEometry, Topology, and Atom-Weights AssemblY	
HAA	Halo Acetic Acid	
HM	Heuristic Method	
HOCldem	Chlorine Demand	
HPI	Hydrophilic	
HPOA	Hydrophobic Acids	
HPON	Hydrophobic Neutrals	

Abbreviation	Explanation	Unit
MAE	Mean Absolute Error	
MBD	Mean Bias Deviation	
MCDM	Multi-Criteria Decision Making	
MLR	Multiple Linear Regression	
NOM	Natural Organic Matter	
NTU	Nephelometric Turbidity Unit	
OECD	Organization for Economic Co-operation and Development	
OLS	Ordinary Least Squares	
PBT	Persistent, Bio accumulative, and Toxic	
PDI	Packing Density Index	
PPCP	Pharmaceuticals and Personal Care Products	
QSAR	Quantitative Structure-Activity Relationships	
QSPR	Quantitative Structure-Property Relationships	
RM	Replacement Method	
RMSE	Root Mean Squared Error	
SA	Simulated Annealing	
SCI	Structural Content Index	
SD	Standard Deviation	
SPA	Successive Projections Algorithm	
SS	Stepwise Selection	
TCM	Trichloro methane	
THMs	Trihalomethanes	
TOX	Total Organic Halides	
UV	Ultraviolet	

1. INTRODUCTION

Freshwater contamination from microbial pathogens remains a serious threat to public health throughout the world (Luilo and Cabaniss, 2010). In the middle of 1880s, the germ theory suggested by Pasteur shows the epidemiological relation between disease and water. Waterborne diseases were well understood in 1854 that the well had become contaminated by a pathogen called *Vibrio cholera*. It was one of the first diseases to be recognized as capable of being waterborne (EPA, 1999). Another well-known cholera epidemic that took place in Haiti in 2010 result in more than 530.000 reported cases and over 7000 deaths (Yang and Wang, 2018).

World Health Organization (WHO) reports that common waterborne pathogens including cholera, dysentery, typhoid fever, *E. coli*, giardia and cryptosporidium, and many others account for 3.7 % of the total global burden of diseases and conduce to about 1.7 million human deaths each year (WHO, 2002; Yang and Wang, 2018). In addition to waterborne pathogens, schistosomiasis, trachoma, ascariasis, trichuriasis and hookworm disease were also attributed to hygienically unsafe water (WHO, 2002).

Water disinfection has been a prerequisite since the early 20^{th} century as it has produced water of a quality safe for human consumption, protected freshwater contamination from the microbes or disease-causing organisms, stopped outbreaks of waterborne infections and parasitic diseases and protected public health. The main operations employed to produce hygienically safe water are sedimentation, filtration, and disinfection (Kumar et al., 2012).

The disinfection is a chemical or physical process that kills or inactivates microorganisms in water. In addition to removing pathogens from drinking water, they are also used for removing taste and color; oxidizing iron and manganese; improving coagulation and filtration efficiency; preventing algal growth in sedimentation basins and filters and preventing biological regrowth in the water distribution system (EPA, 1999). Chemical processes consist of treatment of water with halogens, ozone, hypohalous salts, enzymes, and silver cations, while the physical processes consist of thermal treatment, application of ultrasound and electromagnetic radiation such as ultraviolet rays, X-rays, and γ- radiation, filtration through filters capable of retaining bacteria, and reverse osmosis (Kumar et al., 2012). Adoption of these processes depends on the quality and quantity of water to be treated, location of the water supply, desired quality of treated water, safety, and economics (Kumar et al., 2012).

Despite the development of alternative disinfectants instead of chlorine, chlorinated compounds have been increasingly used for providing hygienically safe drinking water. In 1999, Environmental Protection Agency (EPA) reported that, more than 60 percent of the treatment systems used chlorine as disinfectant or oxidant in surface and groundwater disinfection treatment systems. European Chemical Agency (ECHA) also reported that, 32 thousand tons of chlorine was used as a disinfectant for drinking purposes in 2003.

Chlorination reduces the risk of pathogenic infection by inactivating microorganisms but can pose a chemical threat to human health regarding the formation of disinfection residues and their byproducts if the organic and inorganic precursors are present in the water (Sadiq et al., 2004; Gopal et al., 2007; Hu et al., 2018). There are numerous organic materials such as amino acids, proteins and/or pharmaceuticals and personal care products (PPCPs) such as antibiotics, analgesics, contraceptive drugs, fragrances in water distribution systems that can react with free chlorine to yield halogenated compounds (Chen et al., 2015; Hu et al., 2018). During chlorination of water, chlorine reacts with the traces of organic matter to form disinfection by-products (DBPs) both in the presence or absence of bromide and iodide. DBPs and halogenated DBPs that are known to express toxicity toward humans with harmful long-term effects (Kulkarni et al., 2010; Zhang et al., 2015; Hu et al., 2018). Also, many of DBP precursors are characterized by low molecular weight and thus they may not be completely removed by conventional water treatment plants using traditional coagulation-sedimentation- filtration techniques.

Richardson (2008) reported that, all methods used for disinfection produce their own suite of DBPs and bio-reactive compounds in drinking water. Some of these are also reactive in iodine and bromine containing DBPs formation, which in turn may be even more toxic than their chlorinated counterparts (Krasner et al., 2006; Singer et al., 2006; Ateş et al., 2007; Richardson et al., 2008). The most commonly found DBPs that result from chlorination are trihalomethanes (THMs), haloacetic acids, haloacetonitriles, chloral hydrate and chlorinated phenols. Most of the studies focus on the behavioral profiles of DBPs but the mixture of DBPs have been shown to more cytotoxic, genotoxic, and carcinogenic (Richardson et al., 2008).

However, the formation of DBPs in drinking water depends on several other factors such as temperature, pH, dose, contact time, inorganic compounds and natural organic matter present in water; the complex effects of different factors in water systems remains unclear (Gopal et al., 2007).

After the discovery of trihalomethane (THM) formation from the chlorination of natural organic matter in Rotterdam water supply in 1974, more than 700 DBPs have been identified but only a small number has been assessed for their toxic effects or quantitative occurrence due to the inherent heterogeneity of NOM, the complex background chemistry of municipal water supplies and large variations in water quality of surface water supplies with season and location in terms of NOM concentrations, origin, and characteristics (Krasner et al., 2006; Richardson et al., 2008; Chen et al., 2015).

Considering the increase in the formation of DBPs and the limited available experimental data on the behavioral profiles of DBPs; an improved understanding is important for the development of effective strategies to regulate their formation regarding chlorine consumptions (Abdullah et al., 2009). Empirical methods for predicting chlorine demand due to NOM are based on bulk water quality parameters and do not consider structural properties of molecules that can help to understand reactivity toward chlorine (Luilo and Cabaniss, 2010). Basically, NOM is composed of a mixture of organic molecules with different chemical structures and the amount and/or characteristics of NOM changes with climate, geology and topography (Fabris et al., 2008; Wei et al., 2008; Luilo and Cabaniss, 2010). Dissolved organic matter (DOM), found in aquatic compartments, is also a heterogeneous mixture of various organic materials including humic, fluvic and non-humic compounds, plays a vital role as it undergoes various biological and photochemical transformations (Hu et al., 2016). The composition of DOM also influences the quality of drinking water as it is another major precursor of DBPs. Hence, tremendous effort has been made to give insight into the characteristics of DOM (Hu et al., 2016).

Quantitative structure-activity relationships (QSAR) method is an alternative *in silico* method that can be used to make predictions of chlorine demand due to potential changes in NOM. QSAR method also used to optimize chlorine dosages while maintaining disinfection and minimizing DBP production. Chen et al. (2015) also stated that, the diversity and quantity of DBPs poses daunting challenges to researchers, regulators, and water suppliers to completely understand their analysis, occurrence, toxicity, property, formation, degradation, and removal. Therefore, it is beneficial to find effective ways to prioritize future DBP studies, which puts emphasis on quantitative structure–activity relationship (QSAR) models. However, it is not a substitute for experimental measurements, this approach can be used to enable predictions of chlorine demand (HOCldem) and disinfection by-products(DBPs) due to potential changes in the DOM such as changing pre-treatment methods or land use within the watershed.

The major goal of quantitative structure – activity/property relationship (QSA/(P)R) studies is to find a mathematical relationship between the activity or property under investigation and molecular structure (Katritzky et al., 1995). Despite numerous applications of QSA/(P)R in environmental studies, only Luilo and Cabaniss (2010) attempted to predict HOCldem using QSAR method. In their work, experimental data of disinfection by-product formation from small molecules (HOCldem) collated from literature were used to develop multiple linear regression-based QSPR model using limited number of descriptors. The main drawback of this model is that once they identified the significant descriptors than they divided the training data set into calibration and cross validation data sets by Leave-Many-Out (LMO) approach. Additionally, their model was not tested with up-to-date internal and external validation criteria which provide information on model stability and to what extent it can be used to predict the property of interest using new data. In this study, we plan to carry out a QSPR study for the prediction of chlorine demand of a wide variety of chemicals including pharmaceuticals and personal care products(PPCPs).

1.1. Purpose of the Study

The aims of this study were to redevelop valid QSPR models on the prediction of chlorine consumption of organic molecules present in drinking water using their chlorine demand data and structures, to generate robust QSPR models which complies with the Organization for Economic Co-operation and Development (OECD) principles, to indicate the reliability of predicted chlorine demand (HOCldem) properties of test set chemicals regarding the applicability domain of the models, to predict HOCldem properties of 110 diverse group of chemicals including pharmaceuticals with no chlorine demand data, to compare the published model in terms of up-to-date validation criteria.

A satisfactory result is the creation of a QSPR model that has to be both descriptive (pinpointing the key descriptors) and predictive (able to predict both the chlorine demand of compounds which are not included in the QSPR determinations and the chlorine demand of external chemicals with no data), and superior to the only QSPR model reported by Lulio and Cabaniss (2010).

2. THEORETICAL BACKGROUND

2.1. Disinfection

In literature, there are many kinds of disinfection methods used to eliminate pathogenic microorganisms. Considering various types of disinfectants each with their own characteristic, one should consider the quality and quantity of water to be treated, the location of the water supply, the desired quality of treated water, and also safety and economics. Water disinfection can be achieved chemically such as treatment with halogens, ozone, and silver cations, or physically such as thermal treatment, application of ultrasound and electromagnetic radiation, reverse osmosis and/or a combination of above processes (hybrid methods) such as O_3 and H_2O_2, O_3 and UV, and H_2O_2 and UV (Galal- Gorchev, 1996; Kumar et al., 2012). Some of these methods will be explained briefly as for their advantages and limitations in terms of cost, efficacy, stability, ease of application and the nature of disinfection by-products (DBPs).

As stated by Kumar et al. (2012) and Galal-Gorchev (1996) chlorinated compounds are the most commonly used disinfectants. In developing countries, the use of chlorine is often the only affordable method for both drinking water and wastewater treatment. In this study, we will focus on the chlorine demand of organic compounds in natural water treatment system in which chlorine containing disinfectants are used.

2.1.1. Chlorine

Chlorine used as a disinfectant more than a century. In addition, chlorine is very effective in removing odor to improve the quality of drinking water and in removing pathogens. Chlorine-based disinfectants also provide long time residual protection to control the re-growth of microorganisms throughout the distribution systems, as well as to destroy biofilms during back flushing of biological activated carbon and reverse osmosis systems (Du et al., 2017). Chlorine is typically applied in one of three forms; elemental chlorine gas, sodium hypochlorite solution and calcium hypochlorite solid. All of them should be applied and stored carefully due to their toxicity or corrosiveness.

2.1.2. Sodium Hypochlorite (Bleach)

Ordinary household bleach contains around 5 to 6% sodium hypochlorite (NaOCl) and can be used to purify water if it contains no other active ingredients, scents, or colorings. Bleach is far from an ideal source due to its bulkiness (only 5% active ingredient), and instability over time of the chlorine content in bleach. Chlorine loss is further increased by agitation or exposure to air (Kumar et al., 2012). Use of sodium hypochlorite solution is easier than gas form but it is more expensive than chlorine gas. It should not be stored more than one month in a cool, dark and dry place and care should be taken due to its corrosive effect.

2.1.3. Calcium Hypochlorite (Bleaching Powder or Chlorinated Lime)

Calcium hypochlorite is another chlorinating chemical used primarily in smaller applications. It was patented in 1799 and called bleaching powder. It is a white, dry solid containing approximately 65% chlorine, and is commercially available in granular and tablet forms. Use of calcium hypochlorite solid as a disinfectant is easier than other forms. It is stable and can be stored almost one year but it should be kept away from organic materials due to reactions between calcium hypochlorite and organic material. The calcium takes a long time to dissolve completely and it can also cause calcium scaling and deposits on surfaces and in water circulating equipment (Kumar et al, 2012). As stated by Kumar et al. (2012) this is the most effective method of chlorine treatment in the field.

2.1.4. Chlorine Dioxide

The characteristics of chlorine dioxide (ClO_2) are quite different from chlorine. It is an effective bactericide and viricide under the ranges of pH from 7.0 to 10.0, temperature from 25 to 37°C, and turbidity from 0.5 to 10 nephelometric turbidity unit (NTU) (depending on the source of raw water) that are expected in the treatment of potable water (Kumar et al., 2012). It is found as a dissolved gas in solution that makes it largely unaffected by pH, but it is volatile and can be relatively easily stripped from the solution. Chlorine dioxide is also a strong disinfectant and is selective in its attack on organic materials. But the use of chlorine dioxide is limited to 1 mg/L due to its unknown long-term health effects (Symons et al., 1977; Kumar et al., 2012).

2.1.5. Chloramines

The disinfectant potential of chlorine-ammonia compounds or chloramines was identified in the early 1900s (EPA, 1999). If the chlorinated water contains ammonia (NH_3), it combines with aqueous chlorine to form chloramines for the treatment of drinking water (Kumar et al., 2012). Initially, chloramines were used for taste and odor control, but they were found more stable than free chlorine in the distribution system. Consequently, chloramines were found to be effective for controlling bacterial regrowth. As a result, chloramines were used regularly during the 1930s and 1940s for disinfection (EPA, 1999).

Chloramines produce less DBPs than other chlorinated compounds. However, during the treatment process, the reaction between chlorine and ammonia result in unpleasant taste and odor in finished water (Symons et al., 1977; EPA, 1999; Kumar et al., 2012). To prevent unpleasant taste and odor, both pre-ammoniation or post-ammoniation can be applied. Pre-ammoniation can prevent the formation of taste and odor that are caused by the reaction of chlorine with phenols and other substances (EPA, 1999; Kumar et al., 2012). However, post-ammoniation is generally preferred in the ammonia-chlorine water treatment process as reported by White (references cited in Kumar et al., 2012).

2.2. Literature Survey on Chlorine Demand

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 (Reckhow et al., 1990). The reaction of DOM with chlorine in water is mostly through electrophilic substitution reactions for aromatic compounds (Baxter et al., 2004) and electrophilic addition or elimination in aliphatic compounds and haloform reaction in the case of simple ketones or β –diketones (Arnold et al., 2008).

The formation of THMs which were the first class of DBPs to be detected in potable drinking water in the US during disinfection with chlorine. The formation of THMs is emerging as a persistent problem associated with the maintenance of potable water quality (Boyce et al., 1983). Soon after the discovery of THM formation during chlorination of water containing NOM, numerous studies have been conducted in literature (Rook, 1974; Bond et al., 2009, Richardson et al., 2007, Du et al., 2017).

In 1980s, it is known that the chlorination of natural water containing humic substances produces chloroform, but little is known about the chemical structural properties of aquatic humic material and less is known regarding their reactions with aqueous chlorine (Norwood et al., 1980). Norwood et al. (1980) attempted to measure the chloroform yields and chlorine demands for structures representative of humic degradation products and to identify other chlorinated end products with using a series of compounds. They found two types of pattern. The first pattern was typified by the orcinol data that reflects a generally rapid and simultaneous exertion of both chlorine demand and chloroform production and the second pattern was typified by the 3,5-dimethoxy-4-hydroxycinnamic acid data that reflects an initial chlorine demand in excess of chloroform production. The compounds exhibiting first pattern chlorine demand values were reported as mol of Cl_2/mol of compound as 6.26 for orcinol, 6.6 for resorcinol, 7.6 for 3,5-dihydroxy benzoic acid, 7.63 for 3-methoxy -4-hydroxy-cinnamic acid, 3 for 3,5-dimethoxy-benzoic acid and the compounds exhibiting second pattern chlorine demand values were reported as 6.09 for 3,5-dimethoxy-4-hydroxy-cinnamic acid, 6.06 for β-(3,5-dimethoxy-4-hydroxyphenyl) propionic acid, 5.38 for vanillic acid, 5.18 for syringic acid, 4.94 for 3-methoxy-4-hydroxyhydro-cinnamic acid (Norwood et al., 1980).

De Laat et al. (1982) studied the formation conditions of volatile and non-volatile organochlorides during chlorination of surface water. Chlorination of several organic compounds in diluted aqueous solutions and in a neutral medium (pH 7.0) were evaluated according to chlorine demand and chloroform production. The results obtained from this study show that aromatic compounds such as phenol and aniline derivatives have higher chlorine demand compared to a number of aliphatics such as acids, aldehydes, alcohols. As far as chloroform production is concerned in neutral medium (molar yields < 5%). Many of model compounds produce low quantities of chloroform except a few specific structures such as metapolyhydroxybenzenes and metachlorophenols. The results obtained during this study showed that the chlorine found in the chloroform constitutes only a small proportion of the chlorine demand. The liberation of chloroform in highly reactive precursors such as resorcinol accompanied by the formation of trichloroacetic acid and of monochloromaleic acid. The interactions between chlorine and ammonia nitrogen with the organic compounds can lead to the fast production of notable amounts of chloroform in highly reactive precursors (metapolyhydroxybenzenes and meta chlorophenols).

Boyce and Hornig (1983) conducted another study on the reaction pathways of THM formation from the halogenation of dihydroxy aromatic model compounds for humic substances. The halogenation reactions of 1,3-dihydroxyaromatic compounds and simple methyl ketones were studied in order to explain the mechanism of analogous processes involving naturally occurring humic

material. Model compounds were halogenated at pH 8-10. The results showed that the resorcinol derivatives were found highly reactive towards the formation of THM precursors in dilute aqueous solution at neutral and weakly alkaline pH. The pH dependence and reaction stoichiometry of chloroform and bromoform production were similar. Other halogenated by-products varied as a function of pH and the relative concentrations of halogen and substrate. They reported that the chlorine demand (Cl_2 /substrate) for 1,3-dihydroxynaphthalene as 5.1, 1,3-dihydroxybenzene as 7.1, 3,5-dihydroxy benzoic acid as 7.1, 2,4-dihydroxy benzoic acid as 7.5 and 3,5-dihydroxy toluene as 7.9 at pH 7 (Boyce et al., 1983).

Nitrogenous compounds in raw and treated water are of great concern with their high chlorine demand. The free and combined amino acids are also considered as vital components of natural waters due to the algal activities. Of the amino acids tyrosine and tryptophan have been reported to consume chlorine as 11.4 and 16 mol/mol, respectively. The chlorination of 22 free amino acids, some polypeptides and proteins are conducted by Hureiki et al. (1994). Compounds were halogenated at pH 8.0 and 20℃ for 72 h. Results have shown that the reactivity towards chlorine is somehow related to chemical structures of compounds. Both free and combined amino acids show high chlorine reactivity towards chlorine (references cited in Hureiki et al., 1994). In addition to higher chlorine demand of nitrogenous organic compounds, they produce halogenated DBPs including THMs, halo acids, halo nitriles and haloketones.

In 2002, von Gunten and Gallard conducted another study on the kinetics of chlorination and of THM formation. They combined the kinetics of chlorine consumption with THM formation to further elucidate the nature of the THM precursors in NOM with different types of natural waters. They also investigated the effects of water composition such as type of NOM, seasonal variation and of pretreatment such as UV/visible irradiation and pretreatment with ozone and chlorine dioxide. As a result, THM precursors were divided into a fast and a slowly reacting fraction. They found that resorcinol type structures may be responsible for the fast reacting THM precursors whereas other phenolic compounds might be responsible for the slowly reacting THM precursors. UV/visible irradiation before chlorination of natural water did not affect the THM formation kinetics but the chlorine demand values were increased. In this study, chlorine demand values of ten different phenolic compounds were reported (Gallard and von Gunten, 2002).

Four years later, chlorination of 55 model compounds was conducted by Bull et al. (2006). The studied compounds consist of purines and pyrimidines, simple aromatic amines, pyridine derivatives, structural amino sugars (sialic acids), lignin monomers and derivatives. Many of them contain an

acetylamino group. The sialic acids have been identified in marine and fresh water systems which recently recognized as major components of natural organic matter (NOM) and have sometimes been classified as part of the “hydrophilic neutral” and “colloidal” fractions (Bull et al., 2006). Samples were chlorinated for 48 h at 20℃ and pH 7 in the laboratory. These conditions were selected to mimic the chemical environment that is typical of treatment and distribution systems. Results of this study showed that all compounds were quite reactive to chlorine except for the N-acetylmuramic acid and a few of the pyridine derivatives. The amount of halogenated by-product varied greatly from one compound to another. Between 16% and 100% of the total organic halide(TOX) was observed to be in the form of non-regulated or “unknown” compounds. The chlorine demand was found to be very high for ferulic acid and trans-3,5-dimethoxy-4-hydroxycinnam, 10.32 and 9.588 ($mMCl_2$ /mM), respectively. The THM formation was found significant at neutral pHs for the aminobenzoic acid and aniline.

Bond et al. (2009) conducted another study on the DBP formation and fractionation behavior of NOM surrogates in order to identify significant precursors of regulated and nonregulated DBPs from diverse NOM surrogates. NOM surrogates were fractionated with XAD resins. Their trihalomethane (THM), haloacetic acid (HAA), halo acetaldehyde, haloacetonitriles, and haloketone formations after chlorination (with and without bromide) were recorded. Model compounds were chlorinated both with and without bromide for 24 h at 20±2℃ and pH 7.0. Analysis of fractionation behavior brought out three main groups; hydrophobic acids (HPOA), hydrophobic neutrals (HPON), and hydrophilies (HPI). According to the combination of high/low levels for both chlorine demand and DBP substitution, four divisions were found. There was no significant relationship found between physical properties and formation of any DBP groups (Bond et al., 2009).

It was known that chlorine reacts with NOM and produces DBPs during the water treatment process. Algal-derived organic matter containing lower aromatic carbon contribute greatly to the DBP precursor pool. Therefore, their presence in raw and treated water and the difficulty of removing them are of great concern due to higher chlorine demand and higher DBP formation (Hureiki et al., 1994, Hong et al., 2009). As far as laboratory studies are concerned, algal cells and extracellular organic matter (EOM) may also be the potential DBP precursors for THM and HAA formation due to the higher amino acid content. For these reasons, chlorine demand, TOX and THM formation potentials of 22 free amino acids, 3 proteins, and four polypeptides (containing 4-8 amino acids) were studied by Hureiki et al. (1994) as stated above. Another research on the chlorination of twenty amino acids as precursors of THM and HAA formation conducted by Hong et al. (2009). The chlorine consumption, the HAA and THM formation potential and aromaticity of each amino acid were

determined in this study. Excess chlorine dose applied in two studies and incubated at 20℃ in darkness. The first study was conducted at pH 8.0 whereas the second study was conducted at pH 7.0, and the contact times for the studies were 72 h and 96h, respectively. The results of study conducted by Hong et al. (2009) showed that chloroform, dichloroacetic acid and trichloroacetic acid were the predominant DBP species while the amount of other DBPs such as brominated species were negligible. The chlorine demand of the 20 amino acids ranged from 3.4 to 10 mg Cl_2/mg compound, relatively higher than those of humic substances containing aromatic carbon (1.1- 2.3 mg Cl_2/mg compound) (Reckhow et al., 1990) even they were chlorinated under similar conditions (pH 7.0; temperature 20℃; Cl_2 /DOC, 4-10 mg Cl_2 / mg C; contact time, 3 days) (Hong et al., 2009). The more electron-donating functional groups (-OH, -S, and $–NH_2$) and double bonds in the structure lead to higher chlorine demand (Hureiki et al., 1994; Hong et al., 2009). Specific UV absorbance at 254 nm (SUV_{254}) (meaning aromaticity of the amino acids) showed that amino acids with higher SUV_{254} values, generally produced more THMs (2.57- 147 μg/ mg C) than non-aromatic ones (<4.19 μg/ mg C). Complementary to these results, HAA formation potential shows two types of pattern. Amino acids with chain structures generally exhibit a slow increase in HAA formation as the chlorine demand increased. It indicates that electron donating functional groups in the chain structures act as chlorine-reactive sites but not significant HAA precursors. The second pattern including amino acids with ring structures and two amino acids with a chain structure (aspartic acid and asparagine) showed fast increase in both HAA formation and chlorine demand (Hong et al., 2009).

Among other chlorination by-product precursors, limited studies have been carried out on aliphatic type structures such as *β*-diketone-acid and *β*-keto-acid as THM and HAA precursors. Dickenson et al. (2008) studied the formation of THM, HAA, and TOX from both the chlorination and bromination of a number of aliphatic acid compounds in order to determine the important THM and HAA precursors that can be found in natural waters. In addition to the effect of bromination to the types of THM and HAA formation, the effect of pH was investigated by using two pH values, 8.0 and 5.5, which were applied in water treatment plants. Thus, model compounds were chlorinated and brominated separately at 22℃ and pH 8.0 or 5.5. The results of this study showed that resorcinol exhibit a high 24 h chloroform molar yield (67 % mol/mol) with fast formation kinetics. The resorcinol type structures such as 4,6-dioxoheptanoic acid and 5,7-dioxooctanoic acids which contain *β*-diketone functional group with a distant carboxylic acid group also formed significant amounts of chloroform and a small amount of dichloro acetic acid after 24 h at pH 8.0. This result was also consistent with the study conducted by Gallard and von Gunten (2002). The DXAA (dichloro- or dibromoacetic acid) formation was similar between pH 5.5-8.0 and relatively higher at the lower pH. The possible reaction mechanisms for the formation of DXAA such as decarboxylation, enolization,

and halogenation for model compounds were proposed in this study. The bromoform and DBAA formation were faster compared to chloroform and dichloro acetic acid at both pH 8.0 and 5.5. This is due to bromine being a better electrophile than chlorine. Dickenson et al. (2008) suggested that for aliphatic acid compounds a β-dicarbonyl group consisting of either two ketone groups or one ketone and one carboxyl group is necessary for the formation of chloroform or bromoform and it is important for one of the carbonyl groups to be a ketone and the other carbonyl group to be within a carboxylic acid or carboxylic acid ester functional group for the formation of DXAA. In addition to resorcinol and phenol which are defined as fast and slow reacting THM precursors, respectively, other significant fast- and slow- reacting precursors were defined. For effective control of THM and HAA formation in chlorinated drinking waters, the presence of aliphatic β-dicarbonyl structures or structures that can be readily oxidized to β-dicarbonyl structures should also be considered (Dickenson et al., 2008).

2.3. Quantitative Structure – Activity/Property Relationship (QSA/PR)

QSA/PR analysis is a computational tool based on the assumption that the biological activity (or property, reactivity etc.) of a new designed, untested, and even non-synthesized chemical have the correlation with molecular structure, or properties of molecular structures of similar compounds whose activities/properties have already been assessed. In other words, it provides an option to construct a mathematical equation for a set of chemicals for their specific activity/ property/ toxicity behavior using information encoded in the chemical structure (Roy et al., 2015). At the beginning of the QSAR history, the developed models depend on sets of congeneric compounds, however; more general QSAR models suitable for diverse molecules belonging to different chemical classes were developed in progress of time.

Many different algorithms and computer software are available for QSAR model development. Most of them are based on linear or multiple linear regression (MLR) with variable selection, partial least squares (PLS), as well as non-linear (artificial neural networks) methods. In all approaches, descriptors derived from molecular structure serve as independent variables, and chemico-biological activity or physico-chemical property serve as dependent variables.

The main steps of QSAR study include;

1. data preparation,
2. data processing,
3. data prediction and validation,
4. data interpretation (Roy et al., 2015).

The gradual evolution of QSAR ideology was supposed to be born in the field of toxicology once Cros (1863) observed a relationship between toxicity of primary aliphatic alcohols and aqueous solubility. However, the fundamentals of the molecular structures revealed by Kekule, Couper, and Crum-Brown. In 1841, Blake had noted similar action revealed by salts of isomorphous bases. Ten years later, a relationship between the taste of compounds and their structures was observed by Horsford and Baird. Few years later, Pelikan (1854) and Borodin (1858) discovered that the chemical composition has an impact on the toxicological behavior of the compound (references cited in Roy et al., 2015).

The early suggestion for the existence of a mathematical relationship between chemical structure and activity was done by Crum-Brown and Fraser in 1868 by suggesting the existed linkage between biological activity of different alkaloids and their molecular constitution (Todeschini et al., 2009; Roy et al., 2015). Especially, they defined the physiological action of a compound in certain biological system (ϕ) as a function (f) of its chemical constitution (C):

$$\phi = f(C) \tag{2.1}$$

Thus, the differentiation in chemical structure, ΔC, would be reflected by an effect on biological activity, $\Delta \phi$. This equation can be considered as the first general formulation of QSAR.

In 1869, Richardson noticed that the molecular weight of primary alcohols influences their narcotic behavior. In 1874, Körner supported a hypothesis on the existence of correlations between molecular structure and physico-chemical properties, which represent the synthesis of disubstituted benzenes. The different colors of disubstituted benzenes were thought to be an indicator of different molecular structure and thus, the indicator variables for *ortho*, *meta*, and *para* substitution can be considered as the first three molecular descriptors (Körner, 1869; Todeschini et al., 2009). In 1877, Reynolds reported an effect of chemical constitution on physiological activity of chemicals. A few years later, Richet (1893) observed that the toxicity of alcohol, ether, and ketones is contrarily

influenced by their solubility. In 1901, Overton observed a systemic increase in narcotic potency in tadpoles once the chain length in groups was increased and concluded the partitioning of the compounds into lipid cells to be responsible for such activity. While conducting experiments with morphine Overton also noticed varying toxic effects to human and tadpoles (references cited in Roy et al., 2015).

At the end of the 19th century, Richet (1893), Meyer (1899), and Overton (1901) separately observed a linear correlation between lipophilicity and biological effects (Walker et al., 2003). Specifically, the first QSAR studies arise from the search for relationship between the potency of local anesthetics and the oil/water partition coefficient.

Further estimations in the realm of structure-activity relationships were added by Traube (1904) who observed narcosis to be linearly with the surface tension of the chemicals. Seidell (1912) noted both the partition coefficient and solubility measure to establish such correlation. In 1939, Ferguson assumed that the relative saturation of the substance in the applied phase is associated with narcotic activity. In the middle of the 20th century, the discovery of the effect of ionization of bases and weak acids toward their bacteriostatic activity contributed the pioneering effect of structure-activity relationships (references cited in Roy et al., 2015).

In 1935, the study conducted by Hammett led to the development of the well-known electronic substituent constant Hammett sigma (σ). The goal of the study was to find a linear free energy related (LFER) models (Hammett, 1935, 1937; Nantasenamat et al., 2009; Todeschini et al., 2007; Roy et al., 2015).

The first theoretical molecular descriptors were proposed in 1947 based on the graph theory. The Wiener index and the Platt number were used to model the boiling point of hydrocarbons (Wiener, 1947; Platt, 1947). After the proposal of Wiener Index and Platt number, several other studies were conducted based on the graph theory. On the other hand, the use of quantum-chemical descriptors in the field of QSAR modeling made progress concurrently with quantum chemistry (references cited in Roy et al., 2015).

In the middle of the 20th century, Taft (1956) proposed an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds complementary to the Hammett equation. The contributions from Hammett and Taft set forth the mechanistic basis for QSAR development by Corwin Hansch. Fujita integrated hydrophobic parameters with Hammett's

electronic constants. Therefore, the use of regression analysis to derive QSARs was first proposed by Corwin Hansch and collaborators at Pomona College in the middle of the 20th century (references cited in Roy et al., 2015).

At the same time, Free and Wilson (1964), developed a model of additive substituent contributions to biological activities. Their model called "de novo approach" depends on a biological response via the presence/absence of substituent groups on a common molecular skeleton (Todeschini et al., 2007). In the beginning of the early 1970s, the QSAR modelling was compromised of both substituent effects and the molecular structure-based indices.

The studies of Cros (1863), Richet (1893), Meyer (1899), Overton (1901) and others contributed to the basic understanding of structure-activity relationships (SARs); however, the application of regression analysis and other statistical methods to derive QSARs by Hansch and Fujita (Walker, 2003).

Woo et al. (2002) revealed carcinogenic DBPs found in drinking water by mechanism-based structure-activity relationships analysis. In this study, 209 DBPs were analyzed; some of them were of concern with a moderate or high- moderate rating. Of these, four were structural analogs of 3-chloro-4-(dichloromethyl)-5-hydroxy-2(5*H*)-furanone and five were haloalkanes that presumably will be controlled by existing and future THM regulations and the other eleven DBPs, which included halo nitriles, haloketones, halo aldehyde, halo nitroalkane, and dialdehyde are suitable priority candidates for future carcinogenicity testing and/or mechanistic studies.

Luilo and Cabaniss (2010) used experimental data (HOCldem) of disinfection by-product formation from small molecules to develop multiple linear regression-based QSPR model from limited number of descriptors and studied the prediction of chloroform (trichloro methane- TCM) formation in drinking water disinfection. The 117 experimental data points were collated from 9 literature studies from 1980 onwards. 20 constitutional descriptors were calculated by counting the number of atoms and functional groups for each of the 90 model compounds (training set compounds). They found a robust QSPR model with only three constitutional descriptors for predicting TCM formation from chlorination of model compounds.

Thus, the QSAR philosophy began with the idea of serving an essential role for the interdisciplinary exploration of knowledge on the behavioral manifestation of chemicals. As the time

progressed, in the realm of QSAR it is not surprising to face with different applications such as drug design and molecular design also modification of existing methods (Roy et al., 2015).

2.4. Multiple Linear Regression

Linear methods assume the existence of a linear relationship between independent variables and response variable. Once descriptors were generated, a forward stepwise regression method was used to develop a linear relationship between descriptors and the property of interest. Multiple linear regression attempts to model the relationship between two or more independent variables and a single dependent variable by finding a linear equation (Eq. 2.2) as shown below;

$$\hat{Y}=b_0 + b_1*X_1 +b_2*X_2……..+b_n*X_n \,, \qquad (2.2)$$

$\hat{Y}$ is the dependent variable, in this case, that is chlorine demand; where b_o is the intercept, X-X_n are theoretical molecular descriptors and b_1-b_n are regression coefficients.

Considering the large number of descriptors calculated the use of all numerous combinations of the available ones for model calculation by means of the MLR would be impossible (Gramatica, 2014). Also, the greatest technological impact on modern QSAR has been the unbridled generation of molecular descriptors, thus, to mitigate the redundancy of intercorrelated descriptors, many methods have been developed such as Forward Selection (FS), Backward Elimination (BE), Combined selection/elimination (stepwise selection; SS), Heuristic Method (HM), Genetic Algorithm (GA), Simulated Annealing (SA), Replacement Method (RM), leaps and bounds method, Successive projections algorithm (SPA) and Artificial Neural Networks (ANN), etc. (Doweyko, 2008; Yousefinejad et al., 2015). GA is the most preferable method among these methods because of its superior performance in variable selection.

In the present study, all the possible combinations of the selected descriptors are explored *via* the “Allsubset” method in QSARINS. The best linearly correlated combinations are listed by the software in terms of leave–one–out cross–validated R^2 (Q^2_{LOO}). GA is adaptive heuristic search algorithm based on evolutionary ideas of natural selection and genetics. It combines survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. GA is defined as a search approach which uses random choice as a tool to guide a highly exploitative search through a coding

of a parameter space (Goldberg, 1989). The mechanistic of a simple genetic algorithm is involving copying strings and swapping partial strings based on three operators:

- Reproduction
- Crossover
- Mutation.

QSARINS employs Tournament Selection method to select best representative descriptors *via* GA.

2.5. Pharmaceuticals and Personal Care Products (PPCPs) and Their Impact on the Environment

Any chemical encountering the environment can entail risks to the ecosystem by exerting undesired hazardous outcomes. In the modern era, along with DBPs, pharmaceuticals and personal care products (PPCPs) have been recognized as contaminants of emerging concern due to their persistent and continuous transformation to the environmental compartments. The term PPCPs simply refers to any product regarding healthcare or used for medical purposes for humans and/or animals (Yang et al., 2017). PPCPs can also be divide into multiple groups according to their properties and purposes. Pharmaceuticals include antibiotics, hormones, analgesics, anti-inflammatory drugs, blood-lipid regulators, and personal care products include preservatives, bactericides, fragrances (Yang et al., 2017).

Since the discovery of penicillin by bacteriologist Alexander Fleming, pharmaceuticals have become an indispensable part of modern life with undeniable benefits for diagnosis and treatment of human and/or animal diseases (Klatte et al., 2017). The annual production of pharmacologically active compounds exceeds 100 000 tons globally, a proportion of which enters aquatic environments through their universal consumption and/or production, low human and/or animal metabolic capability, and improper medicine disposal via toilets and sinks has been gaining more and more attention (Gramatica and Sangion, 2016; Hird et al., 2016; Yang et al., 2017).

PPCPs or unwanted residues of Active Pharmaceutical Ingredients (APIs) have been one of the major groups of emerging contaminants that are commonly found and targeted in aquatic and terrestrial environments such as drinking water sources, sewage treatment plants, water treatment plants, and soil since 1970 (Gramatica and Sangion, 2016).

More than 600 APIs mainly antibiotics, analgesics, antidepressants, anti-inflammatory drugs, blood lipid regulators, hormones or their metabolites or transformation products have been found in aquatic and terrestrial environmental compartments at concentrations ranging from ng l^{-1} up to µg l^{-1} in 71 countries worldwide (Jones et al., 2005; Gramatica and Sangion, 2016; Yang et al., 2017). In Germany, approximately half of the 2300 APIs approved for human medicine are considered to be potentially relevant for the environment since they are persistent, bio-accumulative, and toxic (PBT) (Klatte et al., 2017). However, it is difficult to identify the concentrations, origin, metabolism and transformation pathways, short-term and/or long-term effects of pharmaceuticals entering the environment, as well as their impact on non-target organisms. A well-known unexpected effect on non-target organisms reported in several studies. The first one reveals that the synthetic estrogen 17α-ethinylestradiol affect sexual characteristics of male fish (Ankley et al., 2007). The second one shows that the psychotropic oxazepam affect the behavior of perch (Klatte et al., 2017). Another well-known pharmaceutical was fluoxetine (Prozac) belongs to antidepressants lead to changes in weight, metabolism, and behavior of a marine invertebrate *Hediste diversicolor* once it was taken up and metabolized (Ankley et al., 2007). Thus, the findings of this study can be used in assessment of the potential of a chemical to change the behavior of non-target organisms.

3. MATERIALS AND METHODS

3.1. Data Set

The data set used in this study was taken from Luilo and Cabaniss (2010). The data set were compiled from 9 literature studies conducted by Norwood et al., (1980), de Laat et al., (1982), Boyce et al., (1983), Hureiki et al., (1994), Gallard and von Gunten, (2002), Bull et al., (2006), Dickenson et al., (2008), Bond et al., (2009), and Hong et al., (2009). The data set was composed of 150 diverse organic chemicals with chlorine demand values expressed as mol of HOCl / mol of compound ranging from 0.1 to 12.67. However, the compiled data were not acquired under consistent conditions, since the reaction with HOCl may take several hours to several days (35 days), the chlorine doses, pH, temperature and time vary in different studies. Luilo and Cabaniss (2010) used a formulation that the HOCldem values in the shorter time studies were adjusted upward by comparing the chlorine demand of compounds included in both longer- and shorter-time studies by using "common" compounds (Eq.3.1).

$$\mathrm{AdjHOCl_{dem}} = \mathrm{HOCl_{dem}} \; 1/N \; \sum_{i=1}^{N} l_i / s_i \qquad (3.1)$$

The ratio of HOCldem at a shorter reaction time s_i to HOCldem at a 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 HOCldem was adjusted using equation (Eq.3.1).

3.2. QSA/PR Model Development

Flowchart of the model development procedure was given in Figure 3.1. Modelling was done by following data set compilation, data set curation, geometry optimization and descriptor calculation, data splitting (training and test set), descriptor selection, model selection, testing internal and external validation of the model steps. Finally, the best model was tested for its predictive ability by using a number of external set compounds.

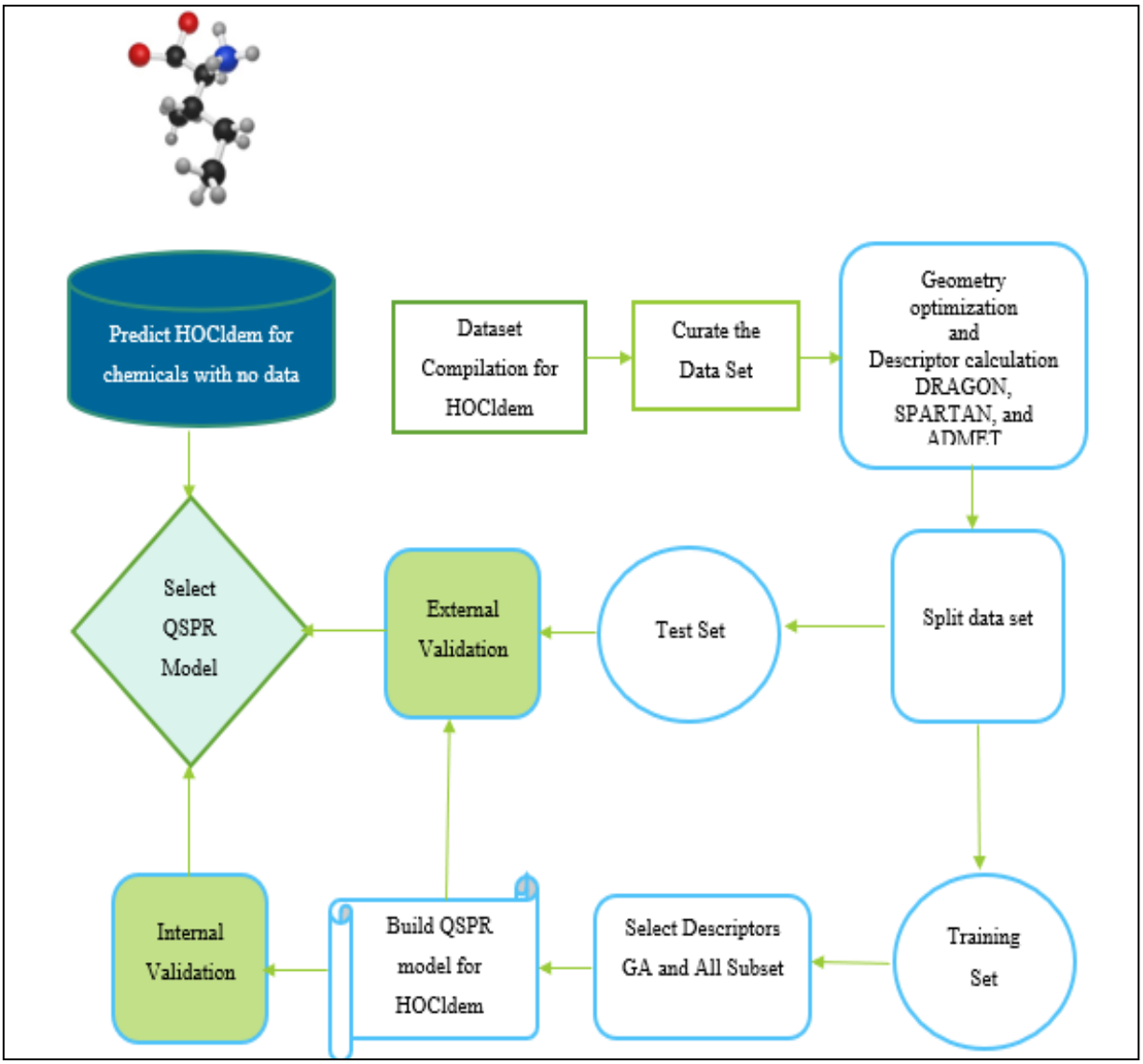


Figure 3.1. Flowchart for the generation of a QSA/(P)R model.

3.3. Structure Optimization and Calculation of Descriptors

Drawing structures is crucial for the calculation of molecular descriptors that capture distinct compositional, electronic, and steric properties (Tropsha et al., 2003). Compounds were represented by theoretical molecular descriptors that cover the information encoded in the chemical structures (Karelson, 2000).

A great number of molecular descriptors were calculated using Dragon 6.0 (Talete Inc., 2014), Spartan 10 (Wavefunction, 2010), and Admet Predictor 8.0 (Simulation Plus Inc., 2015) software packages. In the present study, all chemical structures were drawn and geometrically optimized with Spartan 10 (Wavefunction, 2010) using the semi-empirical PM6 method. Aqueous-phase energy (E_{aq})

values were calculated for each compound and the optimized geometry of the lowest aqueous-phase energy conformer was selected.

The lowest energy conformers of the compounds were used for further descriptor calculations. Aqueous-phase energy (E_{aq}), gas-phase energy (E), the highest occupied molecular orbital energy (E_{HOMO}), the lowest unoccupied molecular orbital energy (E_{LUMO}), molecular weight, dipole moment(μ), CPK volume and CPK area ($Å^2$) were obtained from Spartan 10 (Wavefunction, 2010). Additional descriptors such as E_{LUMO} - E_{HOMO} gap, hardness (η), softness (σ), chemical potential and electrophilicity index (ω) were calculated from the energies obtained from Spartan 10 (Wavefunction, 2010) according to the equations defined by LoPachin et al., (2007).

The first group of the theoretical descriptors were calculated using Dragon 6.0 software (Talete Inc., 2014). Dragon 6.0 (Talete Inc.,2014) software can calculate a total 4885 theoretical molecular descriptors belonging to 29 different types of groups. The descriptor groups are provided in Table 3.1. The descriptor values derived from Spartan 10 (Wavefunction, 2010) were saved as excel file to be imported in to Dragon 6.0 (Talete Inc., 2014). Spartan files were saved as .mol2 file, to be submitted to Dragon 6.0 (Talete, 2014).

Additionally, Admet Predictor 8.0 (Simulations Plus Inc, 2015) descriptors were calculated. Molecular structures were also uploaded to the GaussView (v.3.09, Gaussian Inc., Pittsburgh, USA) to be converted to the MDL mol file format for the calculation of ADMET Predictor 8.0 (Simulation Plus Inc., 2015) descriptors. Finally, 2839 Dragon 6.0 (Talete Inc., 2014) descriptors, 15 Spartan 10 (Wavefunction, 2010) descriptors and 202 Admet Predictor 8.0 (Simulation Plus Inc., 2015) descriptors were calculated.

Finally, the calculated/derived descriptors from Spartan 10 (Wavefunction, 2010) and the descriptors from ADMET Predictor 8.0 (Simulation Plus Inc., 2015) were saved as text file to be imported to QSARINS (v.2.2.1) (Gramatica et al., 2012).

Table 3.1. Descriptor blocks and types in Dragon 6.0 software.

ID Block	Block Description	Number of Descriptors
1	Constitutional descriptors	43
2	Ring descriptors	32
3	Topological Indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	48
7	2D matrix-based descriptors	550
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P_VSA like descriptors	45
11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	90
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GATEWAY descriptors	273
20	Randic molecular profiles	41
21	Functional group counts	154
22	Atom-centred fragments	115
23	Atom-type E-state indices	170
24	CATS 2D	150
25	2D Atom Pairs	1596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug like indices	27

3.4. Training and Test Set Divisions

In order to obtain a validated and predictive QSAR/QSTR model, the data set should be divided into the training set of compounds that are used to build the model, and test sets of compounds that are not used in the model development, but the model validation. Although there is not a definite ratio about the number of chemicals to be divided between the sets, at least 20% of the data set is recommended to be used for testing purposes (Gramatica, 2007).

There are numerous division methods in the literature, such as periodical division, Kohonen networks, and cluster analysis (Papa et al., 2005). Additionally, the division of the data set can be done randomly (random division set by QSARINS v.2.2.1), it can also be done manually or performing hierarchical cluster analysis in SPSS (v.17.0, SPSS Inc., 2008). For random division, the

response values (dependent variable) (in this study the chlorine demand values of compounds) were ordered in ascending order, or order can be done according to the molecular features of the compounds. This splitting should guarantee that the test set covers the entire range of the experimental responses leaving the minimum and maximum values of chlorine demand in the training set. Test set chemicals were selected randomly and were not considered in the calculation of model but used for the external validation: they were in turn put aside and used to check the predictive ability of the training set. In this case, the procedure of training and test set selection and external validation were repeated several times to identify the QSAR model with training set that affords the best prediction power for the test set. To guarantee its ability in providing reliable predictions on new chemicals, it must have good performances in predicting the external dataset called external validation and the statistical parameters should be as high as possible in the best model (Gramatica, 2007).

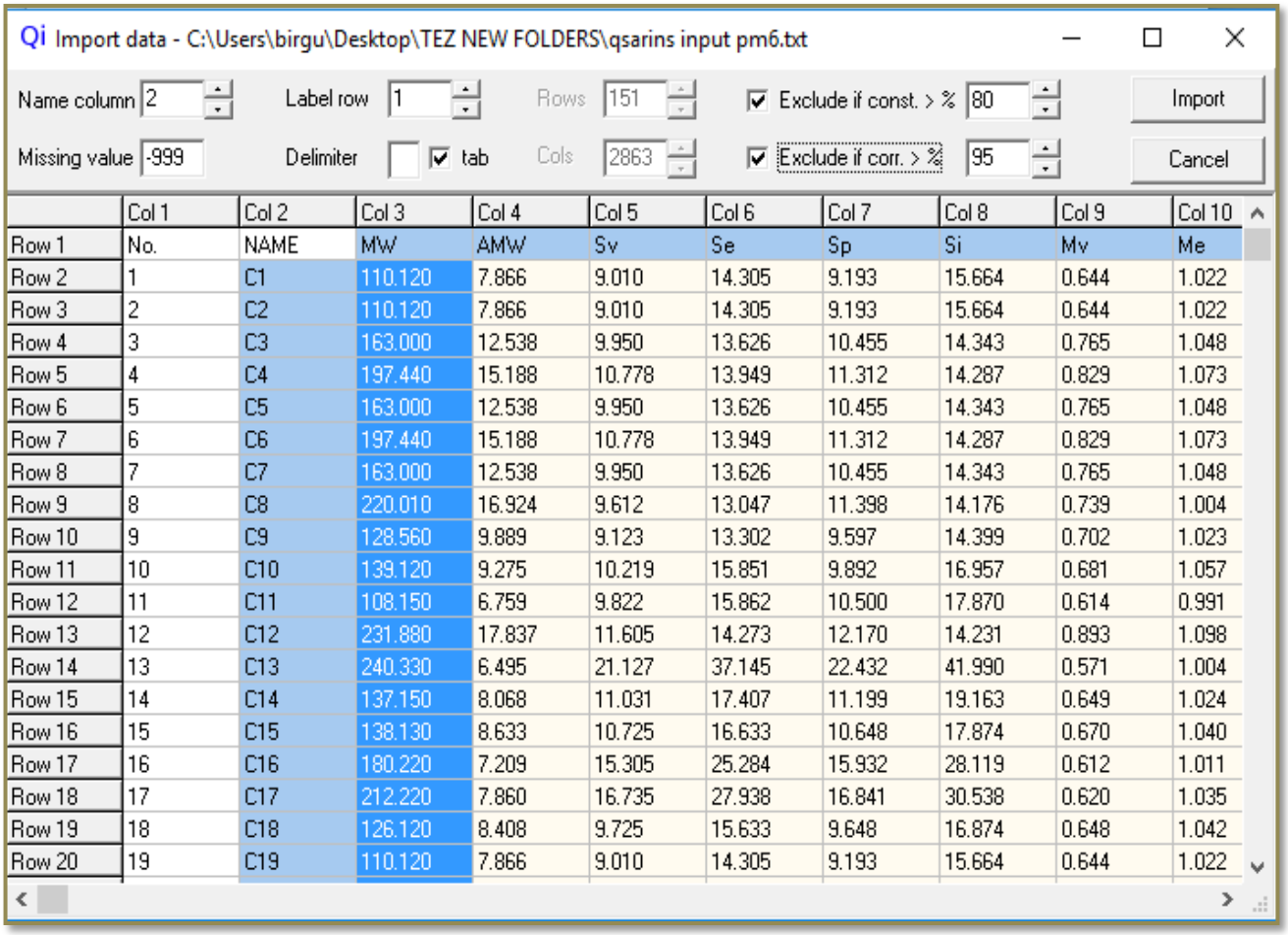
In this study, the approach of Katritzky et al. (1995) will be used for the training and test set division. After the division of the data set, the training set was also checked for normality via Kolmogorov-Smirnov test (SPSS Inc., 2008). Training sets that are not normally distributed will not be modeled. If necessary, another division method such as Kohonen networks will be utilized. In this study, four training/test set divisions were created. Divisions were made by using the response and structure splitting setups in QSARINS (v.2.2.1).

3.5. Descriptor Selection

Models were developed using Genetic Algorithm, all subsets and by holding model and adding descriptors one by one. All subsets can be employed to calculate models with small dimensions, as the combinations grow exponentially when higher dimensions are chosen, and that process requires too much time when the number of descriptors is increasing. If the selected method does not give the desired results, genetic algorithm can be used. This method is a search technique categorized as global search heuristics. Same as algorithms in nature, it is used to select the best descriptors that shows the higher correlation with the data set and discard the others. GA method generally select descriptors randomly. The selected descriptors are used for further calculations to build model. GA tool in QSARINS 2.2.1 allows modification in population size, the mutation rate, and the number of generations for genetic algorithm.

In addition, QSARINS 2.2.1 software allows to pre-filtration based on an objective selection through,

- tests of identical values (constant variables);
- pair-wise correlations (according to a user defined cut off value (suggestion: 95%);
- descriptors can also be deleted if the percentage of the compounds sharing the same value is too high (suggestion: 80%) (Figure 3.2).



	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10
Row 1	No.	NAME	MW	AMW	Sv	Se	Sp	Si	Mv	Me
Row 2	1	C1	110.120	7.866	9.010	14.305	9.193	15.664	0.644	1.022
Row 3	2	C2	110.120	7.866	9.010	14.305	9.193	15.664	0.644	1.022
Row 4	3	C3	163.000	12.538	9.950	13.626	10.455	14.343	0.765	1.048
Row 5	4	C4	197.440	15.188	10.778	13.949	11.312	14.287	0.829	1.073
Row 6	5	C5	163.000	12.538	9.950	13.626	10.455	14.343	0.765	1.048
Row 7	6	C6	197.440	15.188	10.778	13.949	11.312	14.287	0.829	1.073
Row 8	7	C7	163.000	12.538	9.950	13.626	10.455	14.343	0.765	1.048
Row 9	8	C8	220.010	16.924	9.612	13.047	11.398	14.176	0.739	1.004
Row 10	9	C9	128.560	9.889	9.123	13.302	9.597	14.399	0.702	1.023
Row 11	10	C10	139.120	9.275	10.219	15.851	9.892	16.957	0.681	1.057
Row 12	11	C11	108.150	6.759	9.822	15.862	10.500	17.870	0.614	0.991
Row 13	12	C12	231.880	17.837	11.605	14.273	12.170	14.231	0.893	1.098
Row 14	13	C13	240.330	6.495	21.127	37.145	22.432	41.990	0.571	1.004
Row 15	14	C14	137.150	8.068	11.031	17.407	11.199	19.163	0.649	1.024
Row 16	15	C15	138.130	8.633	10.725	16.633	10.648	17.874	0.670	1.040
Row 17	16	C16	180.220	7.209	15.305	25.284	15.932	28.119	0.612	1.011
Row 18	17	C17	212.220	7.860	16.735	27.938	16.841	30.538	0.620	1.035
Row 19	18	C18	126.120	8.408	9.725	15.633	9.648	16.874	0.648	1.042
Row 20	19	C19	110.120	7.866	9.010	14.305	9.193	15.664	0.644	1.022

Figure 3.2. Pre-filtration step in QSARINS v.2.2.1.

During the model development, it is possible to set up the software holding best descriptors from the statistically acceptable model and continue to the addition of new descriptors during the calculation of new models. If the increase in the squared coefficient (R^2) should be no less than 0.02 for models, it is considered as a good result. The addition of descriptors continues to find the best model. This procedure is repeating many times with different combinations, the result at the end of the process will be the population of models.

Usually, the number of calculated descriptors is abundant (hundreds, sometimes even thousands) to have the possibility to represent different features of the chemical structure in different ways, but the problem is that most of them could be intercorrelated and redundant giving very similar structural information (Collinearity).

The use of collinear descriptors adds nothing to mechanistic interpretation of a QSAR and may influence statistical analysis adversely by causing instability in the regression coefficients (Dearden et al., 2009). During the model development; QUIK rule parameter (Delta K) was set to 0.05 to prevent the use of collinear descriptors and eliminate the generation of unbridled molecular descriptors. QUIK rule is a simple criterion based on the K multivariate index that allows the rejection of models with high predictive collinearity that can lead to chance correlation (Todeschini et al., 2004). This rule is derived from the evident assumption that the total correlation in the set given by the model predictors X plus the response Y (K_{XY}) should always be greater than that measured only in the set of predictors (K_X) (Todeschini et al., 2004). Therefore, the QUIK rule is: only models with the K_{XY} correlation among the [X + Y] variables greater than the K_X correlation among the [X] variables can be accepted, or if $[K_{XY}] - [K_X] < \delta K$ reject the model (Todeschini et al., 2004). δK (Delta K) is a limit defined by the user. The QUIK rule has been demonstrated to be very effective in avoiding models with multicollinearity without prediction power (Todeschini et al., 2004).

3.6. Validation of a QSPR Model

Validation is an indispensable part to verify reliability of models. The first condition for model validity deals with the ratio of the number of molecules over the number of selected descriptors (Topliss and Costello, 1972; Bultinck et al., 2007). As the number of descriptors increase in a model, it gets harder to interpret mechanistically every descriptor and its contribution to the overall equation. The recommended Topliss ratio should have a value of at least 5. The quality of a QSAR model should always be examined by validation techniques, which allows to detect over fitting due to variable multicollinearity, noise, sample specificity, and unjustified model complexity. The second condition for model validity deals with the statistical performance of the model and the regression coefficients (Bultinck et al., 2007).

A certain number of validation techniques that allow the evaluation of the effective prediction ability of model has been developed. In this study, QSAR models were developed using MLR (Multiple Linear Regression) and evaluated considering the Organization for Economic Co-operation and Development (OECD) principles (Setubal principles) that were agreed by OECD member

countries at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology (2004).

The QSPR development strictly followed the internationally recognized OECD principles for regulatory purposes and model validation and it should also be associated with the following criteria;

- a defined endpoint
- an unambiguous algorithm
- a defined domain of applicability
- appropriate measures of goodness-of–fit, robustness and predictivity
- a mechanistic interpretation, if possible

According to OECD principle-2, model development with an unambiguous algorithm should be clearly defined, the simplest and most easily reproducible, and thus continuously applicable in a way that the calculations for the prediction of the endpoint can be reproduced by a wide number of users (Gramatica, 2006; 2014).

Applicability domain (AD) of the models were defined in accordance with OECD validation principle-3, meaning that the developed model cannot reliably predict the modeled property for all chemicals, so only the chemicals that fall in the model's domain can be considered reliable. The AD of the model should be defined as a theoretical region that reveals outliers in terms of structural and/or response. It should be also noted that each QSAR model has its own specific AD based on the training set chemicals.

The developed QSAR models should have appropriate measures of goodness-of-fit, robustness and predictivity associated with the OECD validation principle 4. While internal performance of the model is characterized by using a training set, the external predictivity of the model characterized by using an appropriate test set (OECD, 2007).

The statistical performances of the developed model were evaluated by the coefficient of determination (R^2), cross validation leave one out (Q^2_{LOO}), standard error of the estimate (SE), Fisher criterion (F), the root mean squared error of training set ($RMSE_{\mathrm{Tr}}$), the concordance correlation coefficient for training set (CCC_{Tr}) and the predictive abilities were evaluated by the determination coefficient of test set (R^2_{Test}), the root mean squared error of test set ($RMSE_{\mathrm{Test}}$), the concordance correlation coefficient for test set (CCC_{Test}), r^2_{m}, Δr^2_{m} metrics (Ojha et al., 2011) the predictive

squared correlation coefficients, namely Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and the Golbraikh and Tropsha criteria (Golbraikh and Tropsha, 2003).

3.6.1. Internal Validation

The statistical quality and internal predictivity of the MLR models has been judged by the parameters like the correlation coefficient of determination (R^2), the adjusted squared determination coefficient (R^2_{adj}), the leave-one-out cross-validation (Q^2_{LOO}), and Y-scrambling. The root mean squared of error (*RMSE*) for the training ($RMSE_{Tr}$) set that summarize the overall error of the model was calculated as an additional measure of the accuracy.

3.6.1.1. R^2 (Coefficient of determination). R^2 is the most widely used parameter to assess the ability of a QSAR model to reproduce the data in the training set (goodness of fit) and is a measure of the quality of the fit between model-predicted and experimental values (Gramatica,2007). The value of R^2 can be predicted from the following equation (Eq. 3.2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} \qquad (3.2)$$

Where y, is the observed dependent variable (the experimental response), $\bar{y}$ its mean and y_i is the corresponding calculated(predicted) value. *RSS* is the residuals sum of squares, and *TSS* is the total sum of squares for n elements of the modeled data set.

3.6.1.2. Leave-one-out (LOO) cross-validation (Q^2_{LOO}). The cross-validation leave-one-out is the most known technique and used as a fitness function during the model development step. In this step, certain number of compounds are involved in the model calculation (training set), while the others (test set) are in turn put aside and used to assess the model's prediction ability. It assesses the ability of the model to predict new chemicals in the data set one by one, putting them iteratively in the test set. The value greater than 0.5 is generally regarded as good (Eriksson et al., 2003).

The formula of the calculated parameter Q^2_{LOO} value is as follows;

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_{tr}})^2} = 1 - \frac{PRESS}{TSS} \qquad (3.3)$$

Where y_i, is the observed dependent variable (the experimental response), $\hat{y}_{i/i}$ is the predicted value of the response calculated excluding the i^{th} element from the model computation, $\bar{y}$ is the mean value of the dependent variable, *PRESS* is the predictive error sum of squares, and the *TSS* is the total sum of squares for *n* elements of the complete data set.

This process continuous till the increase in the number of descriptors does not efficiently improve the Q^2_{LOO} value. The increase in the Q^2_{LOO} value goes on in consonance with the increase in the fitting measure R^2 value, after a certain number of descriptors is reached an overfitted, but not predictive, model can result (Chirico and Gramatica, 2011).

3.6.1.3. Y-scrambling. Response randomization (Y-scrambling) is an attempt to observe the action of chance in fitting given data. This is done by intentionally destroying the connection between target variable *y* and independent variables *x* (molecular descriptors in QSAR) by randomly permuting the response values (*y*), leaving all *x* data untouched, and performing the whole model building procedure as it would be done for real *y* data (Rücker et al., 2015). This technique must be used in accordance with cross-validation (CV) and must always be applied to check the significance of the developed QSAR model obtained by chance correlation (Gramatica, 2007).

3.6.2. External Validation

External validation (test set validation) is regarded as the most conclusive process to verify predictive performance of any developed QSAR model, since a new data set that has not been used for model development is employed for prediction to check the reliability of the developed model (Gramatica, 2011, 2007). The basic difference of internal and external validation is that, during the internal validation process all molecules are included in the model development, since iterations depend on the excluding one or more compounds at a time (if a compound is in the test set in at least one validation step, no chemicals remains new at the end of the process) (Chirico and Gramatica, 2011).

The quality of the developed models is assessed based on different external parameters such as Q^2_{Fn} (predictive squared correlation coefficients- Q^2_{F1} (Shi et al., 2001), Q^2_{F2} (Schüürmann et al., 2008) Q^2_{F3} (Consonni et al., 2009), r^2_m (Ojha et al., 2011), Δr_m^2 (Ojha et al., 2011), $RMSE_{Test}$, *MAE* (Mean absolute error), CCC_{Test} (Lin, 1989) and Golbraikh and Tropsha method (2003) and only good models stable and internal predictive are subjected to external validation.

3.6.2.1. Predictive Squared Correlation Coefficients (Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}). The form is similar to Q^2_{LOO}, but here the sums are over the external prediction set elements and, in lieu of PRESS (which is calculated using the training set in cross-validation), the sum of the squared differences between the prediction set experimental values (y_i) and those calculated by the model ($\hat{y}_i$) is used. Q^2_{F1} shows the degree of correlation between the experimental and predicted activity of the data set (Shi et al., 2001).

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{TR})^2} \quad (3.4)$$

In the denominator, the sum of the squared differences of the prediction set experimental values (y_i) and the average of the training set ($\bar{y}_{TR}$) is used, instead of the plain total sum of squares-TSS (which is calculated using the training set values). The use of the average of the training set values, instead of the average of the prediction set values, is a way of keeping track of the "distance" between the two sets (Gramatica et al., 2011).

An alternative criterion was proposed by Schüürmann et al. (2008). It differs from Q^2_{F1} because the average value at the denominator is calculated using the prediction data set instead of the training set.

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y}_{TEST})^2} \quad (3.5)$$

Consonni et al. (2009) proposed another validation metric called Q^2_{F3}, the formulation differs from both Q^2_{F1} and Q^2_{F2} as the denominator is calculated on the training set, and both numerator and denominator are divided by the number of the corresponding elements.

$$Q^2_{F3} = 1 - \frac{[\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2]/n_{TEST}}{[\sum_{i=1}^{n_{tr}} (y_{i\,TR} - \bar{y}_{TR})^2]/n_{TR}} \quad (3.6)$$

3.6.2.2. Concordance Correlation Coefficient (CCC). Concordance correlation coefficient is the concordance between a new test or measurement (y) and a gold standard test or measurement (x) proposed by Lin (1989, 1992). It is apposite to measure the agreement between experimental and predicted data, which should be the main objective of any developed QSAR models. For CCC_{Test}, Gramatica et al. (2012) proposed a cutoff value of 0.85 defined by verifying scatter plots on external

data. *CCC* can be described the most optimistic, modified form of the correlation coefficient and can be calculated with the equation (Eq.3.7) (Lin, 1989). This metric can be computed both training and external sets as CCC_{Tr} and CCC_{Test}, respectively.

$$CCC = \frac{2\sum_{i=1}^{n}(x_{obs(test)}-\overline{x_{obs(test)}})(y_{pred(test)}-\bar{y}_{pred(test)})}{\sum_{i=1}^{n}(x_{obs(test)}-\overline{x_{obs(test)}})^2+\sum_{i=1}^{n}(y_{pred(test)}-\overline{y_{pred(test)}})^2+n(\overline{x_{obs(test)}}-\overline{y_{pred(test)}})^2} \quad (3.7)$$

In the above equation (Eq.3.7.), $x_{obs(test)}$ and $y_{pred(test)}$ correspond to the observed and predicted values of the test compounds, *n* is the number of chemicals, and $\overline{x_{obs(test)}}$ and $\overline{y_{pred(test)}}$ correspond to the averages of the observed and predicted values, respectively, for the test compounds (Gramatica, 2011; Roy et al., 2016).

3.6.2.3. The r^2_m, $\bar{r}^2_m$, Δr^2_m. In the same period, two different versions of r_m^2 metric were introduced by Ojha et al. (2011). The r_m^2 predicts the relationship between the order of the experimental activity and the predicted activity. If the experimental and predicted values fit each other, the difference between these values is expected to be 0. The r_m^2 can be calculated with the following equation (Eq.3.8):

$$r_m^2 = r^2 \left(1 - \sqrt{r^2 - r_o^2}\right) \quad (3.8)$$

Where r^2 and r_0^2 correspond to R^2 and R_0^2 in the Golbraikh and Tropsha method, are the determination coefficients of linear relations between the observed and predicted values of the test set compounds with and without intercept, respectively. The value of r^2 is always greater than the value of r_0^2.

This formula can be applied in both internal and external validation. The main point of this formula is the difference between r^2 and r_0^2. While calculating r_m^2 the slopes of the regression lines were not considered by Roy et al. (2011).

In a robust model, the following conditions should be provided:

- $r_m^2 > 0.5$
- $\bar{r}^2_m = (r_m^2 + r_{m'}^2) / 2 > 0.5$ (3.9)
- $\Delta r_m^2 = | r_m^2 - r_{m'}^2 | < 0.2$

The r_m^2 metric does not count in the differences between individual responses and the training set mean, by the way it prevents overestimation of the quality of prediction when the data sets with wide response range (Ojha et al., 2011). This parameter is used to evaluate the quality of prediction only after model development and prediction steps are over.

3.6.2.4. The Golbraikh and Tropsha method. Golbraikh and Tropsha (2003) set some criteria for external prediction. The models were considered acceptable if the following conditions are satisfied:

I. $Q^2_{Tr} > 0.5$,
II. $R^2_{Test} > 0.6$,
III. R_0^2 and $R_0'^2$ are close to R^2 (3.10)

$\frac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \leq k \leq 1.15$ or

$\frac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \leq k' \leq 1.15$

IV. $|r_0^2 - r_0'^2| < 0.3$

Where correlation coefficient R between the predicted and observed activities, coefficients of determination (r_0^2 is predicted vs. observed activities, $r_0'^2$ is observed vs. predicted activities), slopes k and k' of regression lines (predicted vs. observed activities and observed vs. predicted activities) through the origin (Golbraikh and Tropsha, 2003). This concept is based on the simple idea that the more the predicted values match the experimental ones, the better the model performance in prediction. It is necessary to be sure that; the slopes of the regression lines (those related to R^2 and R^2_0) should not too different from 1 and that R^2 and R^2_0 should be close enough to each other (Golbraikh and Tropsha, 2003).

3.6.2.5. $RMSE_{Ext}$ (Root mean square errors of prediction). Root mean squared error in prediction is widely used error based external validation metric which formulated as:

$$RMSE_{ext} = \sqrt{\frac{1}{n_{ext}} x \sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2} \quad (3.11)$$

That measures the discrepancies among the experimental values versus the ones predicted by the model. It can be calculated for both training and prediction sets. As stated by Gramatica (2011), the stability of RMSE for training and prediction sets can be considered a measure of the model's generalizability. According to Willmott and Matsuura, *RMSE* is composed of magnitude of the prediction errors in the squared form, the magnitude of the average error, and the squared root of the number of samples while *MAE* is simpler metric (references cited in Roy et al., 2016).

3.6.2.6. Mean Absolute Error (MAE) based criteria. In lieu of *PRESS* and *RMSE*, some authors prefer to use mean absolute error (*MAE*) as an error-based metric considering its simplicity. The formulae of *RMSE* squaring the higher prediction error values will have more weight than the lower errors while *MAE* provides an equal weight to all the errors (Roy et al., 2016).

MAE can be calculated with the following equation:

$$MAE = \frac{\sum_{i=1}^{n_{ext}} |y_i - \hat{y}_i|}{n_{ext}} \quad (3.12)$$

I. Good predictions:

From a general notation, an error of 10% of the training set range should be acceptable while an error value more than 20% of the training set should be a very high error. Thus, the criteria for good predictions should be the following:

MAE ≤ 0.1 x training set range and *MAE*+ 3σ ≤ 0.2 x training set range.

Where, the σ value refers to the standard deviation of the absolute error values for the test set data. Considering a normal distribution pattern, mean ± 3σ covers 99.7% of the data points.

II. Bad predictions:

A value of *MAE* more than 15% of the training set range should be high while an error more than 25% of the training set is considered very high. Hence, the predictions could be considered very high. Hence, the predictions could be considered when:

MAE > 0.15 x training set range or *MAE* + 3σ > 0.25 x training set range.

The predictions which do not fall under either of the above two conditions may be considered as of moderate quality. The mentioned criteria should be used in cases where there are more than 10 data points in the test set (Roy et al., 2016).

3.6.2.7. External Validation Based on the Mean Absolute Error. Xternal Validation Plus is another tool developed by Roy et al. (2015) that checks the presence of systematic errors in the model and further computes all the required external validation parameters, while judges the performance of actual prediction quality of a QSAR model based on the mean absolute error.

(i) NPE/NNE > 5* or NNE/NPE >5*

Here, **NPE**: Number of positive errors AND **NNE**: Number of negative errors

(ii) ABS(MPE/MNE) > 2* or ABS(MNE/MPE) > 2*

Here, **MPE**: Mean of positive error; **MNE**: Mean of negative errors; **ABS**: Absolute value

(iii) AAE - ABS(AE) < 0.5*×AAE

Here, **AAE**: Average of absolute prediction errors; **AE**: Average of predictions errors; **ABS**: Absolute value

(iv) R^2 (i^{th} vs $(i-1)^{th}$ residuals) > 0.5* for residuals sorted on experimental response values (Yobs)

(v) R^2(Yobs and residuals) > 0.5*

Figure 3.3. Error types of Xternal validation plus tool.

If any one or more conditions stated in Figure 3.3. were met, Systematic error occurs. If the systematic error present in the output file, the model should be discarded.

3.7. Selection of the Best QSPR Model

In accordance with the above-mentioned criteria, Multi-Criteria Decision Making (MCDM) implemented in QSARINS 2.2.1 was used for the ranking of generated models. It consists of the summary of performances of a certain number of criteria associated with the internal and external validation concurrently. Each validation criteria value ranges from 0 to 1 and the geometric average of all the values obtained from the desirability functions creates the MCDM value (QSARINS 2.2.1). The MCDM of fitting (maximizing R^2, $R^2_{adj,}$ and CCC_{Tr} while minimizing $R^2 - R^2_{adj}$) cross-validation (maximizing Q^2_{LOO}, Q^2_{LMO} and CCC_{CV}, while minimizing $R^2_{\text{Y-SCR}}$) and external validation (maximizing Q^2_{F1}, $Q^2_{\mathrm{F2,}}$ $Q^2_{\mathrm{F3,}}$ and CCC_{Test}) parameters are automatically calculated using all the corresponding criteria. The model with the best MCDM compromise among the selected validation criteria will be sorted as the best.

3.8. Applicability Domain

Applicability domain (AD) of the models was defined to be consistent with OECD principle-3, the need to define an applicability domain (AD) is crucial for further predictions since in the absence of AD each model can predict the activity of any compound. AD expresses the fact that QSARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions (OECD, 2007). The basic meaning of AD was defined by Netzeva et al. (2005) in the 52th workshop of the European Centre for the Validation of Alternative Methods (ECVAM, 2005). “The applicability domain of a QSAR model is the response and the chemical structure space in which the model makes predictions with a given reliability”. It can be described as a theoretical region in chemical space confined to both the model descriptors and modeled response in this study chlorine demand that allows one to estimate the uncertainty in the prediction of a compound based on how similar it is to the training compounds employed in the model development (Roy et al., 2015).

The domain of applicability of molecules plays a crucial role in determining the uncertainty in the prediction of specific molecules and it is impossible to predict a whole universe of chemicals employing a single QSAR model (Gramatica, 2007). By means of AD there are various approaches have been developed by researchers and each method has its own pros and cons. The existing methods for determining AD are the ranges in the descriptor space, the geometrical methods, the distance-based methods, the probability density distribution, and the range of the response variable (Roy et al., 2015).

In this study, the AD of the model was defined via leverage approach called Williams Plot, Standardization Approach Roy et al., (2015), and ranges in the descriptor space. Williams plot reveal that the compounds included in the AD perimeter were considered not outlier in terms of their structural features (structural domain), their descriptor values (physico-chemical domain) or their response values (response domain). To visualise the outliers in the model, a plot of standardised residuals (R) versus leverages (or hat values, *h*) detects the outliers for the response (Y-outliers) and those for the structure (X-outliers). It consists of plotting the standardized residuals on the y-axis and the leverage values from the *hat* matrix diagonal on the x-axis. Leverage values represent the degree of influence that the structure of every single chemical has on the model. A compound with high leverage in a QSAR model is the driving force for the variable selection if this compound is in the training set (good leverage).

The leverage (h_i) values calculated with descriptor matrix of the training set was used to construct the Williams plot (Eq. 3.13);

$$h_i = X_i^T (X^T X) X_i \quad (3.13)$$

where X_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values while X^T is its transpose matrix.

The critical hat value (h^*) was generally fixed at *3*p´/n;*

$$h^* = 3 p' / n \quad (3.14)$$

where *n* is the number of training chemicals in the model, and p' the number of variables plus one.

Query compounds with leverage higher than this defined threshold of *3*p´/n* are considered to be unreliably predicted, conversely if the leverage values of any compound are lower than the critical value, they are considered as reliably predicted (Gramatica, 2007).

Another approach developed by Roy et al. (2015) called "Applicability domain using standardization approach" is used to define the AD of generated models. This program is available at http://dtclab.webs.com/software-tools and/or http://teqip.jdvu.ac.in/QSAR_Tools/. The algorithm and methodology of the proposed approach was explained in detail in their paper (Roy et al., 2015).

The idea is similar to leverage approach. The developed model is applied for prediction of test set compounds which are expected to be structurally similar to the training set compounds. If a small fraction of the training set contains features very dissimilar to the rest, those features are not properly included in the training step and these compounds called as X-outliers. If test set compounds show similarity to this small fraction of training set compounds, then their predictions are expected not to be good, as the model has not captured the features of those training set compounds which have a small manifestation and are different from majority of the compounds. So those test set compounds are expected to be outside of the AD of the model. Based on this method, all the descriptors of the training set compounds should follow a normal distribution pattern. According to this distribution, 99.7% of the population will remain within the range mean ±3 standard deviation (*SD*). Thus, mean ±3*SD* represents the zone where most of the training set compounds belong to. Any compound outside this zone is dissimilar to the rest and majority of the compounds. Thus, after a descriptor column is standardized based on the corresponding mean and standard deviation for the training set compounds only, if the corresponding standardized value for descriptor *i* of compound *k* (S_{ki}) is more than 3, then the compound should be an X-outlier (if in the training set) or outside AD (if in the test set) based on descriptor *i*. If the maximum S_i value of a compound *k* is lower than 3, then the compound is quite similar to a good number of compounds in the training set with respect to all descriptors. If the minimum S_i value of a compound *k* is higher than 3, than the compound is quite dissimilar to most of the training set compounds with respect to other descriptors (Roy et al., 2015). The algorithm and methodology of the proposed approach was explained in detail at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. (Roy et al., 2015). Additionally, a compound was identified as being out of the AD in a simple way; at least one descriptor is out of the span of the ranges approach.

3.9. Predictive Performance of the Generated Models

One of the major applications of QSAR models is predictive ability of properties of untested chemicals in helping for further synthesis and testing that result in significant gain in resources in terms of material, manpower, money, and time (Roy et al., 2011). However, some researchers consider cross-validation is adequate if properly done. Gramatica et al. (2011) support the idea to test the real predictivity of a QSAR model by using completely new chemicals (called prediction set or external set).

There are discordant opinions on the use of completely new chemicals to verify the predictive ability of the developed model. The prediction set can be confused with the test set, but it differs from

the test set of the internal validation. In contrast to the test set of internal validation, external validation does not use excluded molecules (prediction/blind set) during model development that leads to an accurate evaluation of the model's prediction ability.

After the development of the QSAR model, the final model was used to predict chlorine demand of 110 compounds with no chlorine demand data using Insubria Graph proposed by Gramatica et al., (2013).

4. RESULTS AND DISCUSSION

4.1. QSAR Modelling on the Chlorine Demand of Diverse Organic Chemicals

The experimental chlorine demand data compiled from different resources were divided into two categories (training and test sets). The training set was composed of 80% of the whole data with 120 compounds and the test set was composed of 20% of the whole data with 30 compounds. The normality of the chlorine demand data was evaluated using the Kolmogorov-Simirnov test in SPSS (v.17., SPSS Inc.,2008). The difference between minimum and maximum values in a data set should be at least 2 logarithmic unit in order to be used for QSAR modelling (Cronin et al., 2009). The range of logarithmic chlorine demand is between (-1) and (1.1). Thus, HOCldem values were ordered and compounds with minimum and maximum HOCldem values were left in the training set. Further splitting was made using the tool in QSARINS (v.2.2.1) software. The QSPR models were created by different combinations of compounds using the response and structure splitting setups. The test set chemicals that resulted in the most robust model were given in Appendix A, Table A.1.

Molecular descriptors obtained from three software packages (DRAGON v.6.0, SPARTAN v.10 and ADMET Predictor v.8.0) combined in QSARINS (v.2.2.1) software. Numerous QSPR models based on ordinary least square (OLS) method were generated using different training and test set divisions together with "All subset" and "Genetic Algorithm"-based iterative tool of QSARINS (v.2.2.1) software. QUIK Rule was set to 0.05 before starting to scan for models in order to eliminate the models with intercorrelated descriptors. We also attained Topliss Ratio (1972) and tested the R^2 and Q^2 increase with respect to number of variables (Figure 4.1). The increase in the R^2 and Q^2 values is greater than 0.02 in each step indicating that the addition of a new variable to the model is not redundant. Models with descriptor numbers varying from 6 to 7 were generated and ranked using MCDM as well as the internal and external validation criteria. The best model from each division was selected with MCDM analysis. Furthermore, models were tested for their external prediction capacity by regarding the highest number of compounds within their applicability domain. All models were listed together with their fit and internal validation parameters, and external validation parameters in Table 4.1. and 4.2, respectively. The complete list of chemicals in the data set was given in Table 4.5. together with the CAS number and chlorine demand (HOCldem) values and the test set chemicals for the generated models are also given in Appendix A Table A.1.

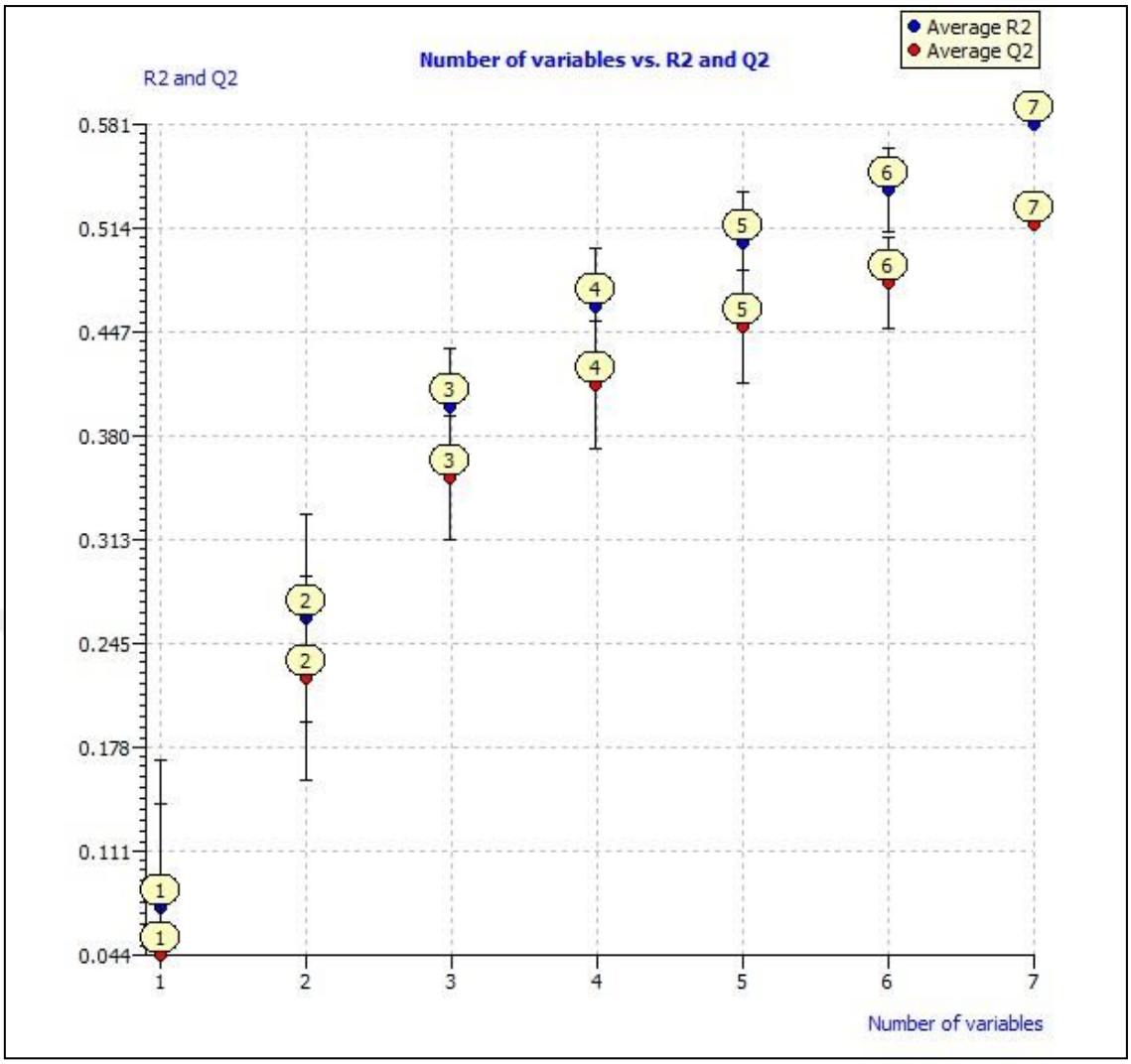


Figure 4.1. The change in R^2 and Q^2 along with the increase in the number of variables.

The models selected as for the higher R^2 and Q^2 values, higher prediction performances of the prediction set and the minimum number of response and structural outliers.

The developed models were verified for the internal predictive power which is judged by the parameters like R^2 (the coefficient of determination) and Q^2_{LOO} (the leave-one-out cross-validation). For the 6-7-descriptor based models, R^2 and Q^2_{LOO} values were very similar to each other which reveal the stability of all models (R^2 range is between 0.58 and 0.63, and Q^2_{LOO} range is between 0.52 and 0.58).

Table 4.1. The developed models for HOCldem property, their fit and internal validation parameters.

Model No	**Nu. of descriptors**	**Descriptors**	R^2	R^2_{adj}	$RMSE_{Tr}$	CCC_{TR}	F	Q^2_{LOO}	$RMSE_{CV}$	CCC_{CV}
M1*	**7**	**SIC2 MATS3s E1e B02[C-N] PDI ArHdrxl_-OH R8v+**	**0.60**	**0.56**	**2.25**	**0.74**	**23.18**	**0.52**	**2.41**	**0.70**
M2	7	CIC2 GATS6i E1e E2e P_VSA_m_2 B0[O-O] ArHdrxl_-OH	0.61	0.58	2.17	0.76	24.64	0.55	2.32	0.72
M3	6	CIC2 SpMAD_G/D nArOH GATS6s P_VSA_m_2 CATS2D_03_AA	0.58	0.56	2.29	0.73	25.71	0.52	2.43	0.70
M4	6	CIC2 SpPosA_B(p) TDB07s E1e F01[C-N] ArHdrxl_-OH	0.63	0.61	2.15	0.77	32.08	0.58	2.29	0.75

*Selected Model.

Table 4.2. External Validation Parameters of the generated QSPR models for HOCldem.

Model No	*Training/ Test Set*	r^2_{ext}	$RMSE_{Ext}$	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC_{Ext}	r^2_m *av.*	Δr^2_m	*k'*	*k*	*Prediction %*
M1*	***120/30***	***0.76***	***1.39***	***0.79***	***0.79***	***0.84***	***0.87***	***0.67***	***0.17***	***0.86***	***0.98***	***91***
M2	*120/30*	*0.77*	*1.51*	*0.76*	*0.76*	*0.81*	*0.87*	*0.67*	*0.17*	*0.99*	*0.94*	*90*
M3	*120/30*	*0.83*	*1.18*	*0.83*	*0.82*	*0.87*	*0.90*	*0.70*	*0.15*	*0.97*	*0.99*	*75*
M4	*120/30*	*0.78*	*1.35*	*0.75*	*0.75*	*0.86*	*0.88*	*0.69*	*0.08*	*0.96*	*0.98*	*45*

*Selected Model.

Having high the R^2_{Test} values (0.76 to 0.83) and low $RMSE_{Test}$ proved to have a good predictive power for all models.

All QSPR models were considered in terms of the following Golbraikh and Tropsha (2003) criteria;

I. $q^2 > 0.5$ (0.52)
II. $R^2 > 0.6$ (0.60),
III. R_0^2 and $R_0'^2$ are close to R^2
*1-(R_0^2 / R^2) (0.007) < 0.1 and $0.85 \leq k$ *(1.07)* ≤ 1.15 or
*1-($R_0'^2$ / R^2) (0.090) < 0.1 and $0.85 \leq k'$ (0.88) ≤ 1.15
IV. $|R_0^2 - R_0'^2|$ (0.07) < 0.3

With the abovementioned results, all models satisfy the Golbraikh and Tropsha (2003) criteria. *CCC* is an additional criterion for the external validation suggested by Lin (1989) which verifies the agreement of experimental and predicted data and the suggested cutoff value for the *CCC* was 0.80. However, Chirico and Gramatica (2012) suggested a higher threshold value for this parameter.

For the developed models, all of the external validation criteria were in agreement with the threshold values suggested in the literature. Threshold values attained were given below for CCC_{Test}, Q^2_{Fn}, $\bar{r}^2_m$ and Δr_m^2. All These results indicate that the developed models are robust, validated and predictive.

I. $CCC = 0.85$
II. $Q^2_{Fn} = 0.70$
III. $\bar{r}^2_m = 0.65$
IV. $\Delta r^2_m = 0.2$

Of the generated models, the 7-descriptor MLR model labelled as M1 was highlighted in Tables 4.1. and 4.2., since it has no response and structural outliers and its structural coverage is 91% for the external set chemicals (110 chemicals from different class) with no HOCl dem data.

The highlighted model in Table 4.1. gives the following equation, Eq. 4.1. This equation is together with the descriptors involved and their regression coefficients and the numbers in parenthesis indicate the standard deviation of the coefficient of descriptors.

$HOCl_{dem}$ = -16.704 (±8.690) + 0.5879 (±0.627) ArHdrxl_-OH + 0.4066 (±6.573) PDI + 0.3537 (±5.159) SIC2 + 0.2119(±1.077) B02[C-N] – 0.1862 (±1.907) MATS3s – 0.1729 (±9.515) E1e + 0.1458 (±87.766) $R8v_+$ (4.1)

n_{Tr} = 120, $r^2_{adj=}$ 0.56, $RMSE_{Tr}$ = 2.25, F = 22.18, CCC_{Tr} = 0.74,
n_{Test}= 30, r^2_{Test}= 0.81, $RMSE_{Test}$ = 1.39, Q^2_{F1} = 0.78, Q^2_{F2} = 0.78, Q^2_{F3} = 0.84, CCC_{Test} = 0.87
s = 2.33

where n_{Tr} and n_{Test} refer to the number of compounds in the training and test sets, respectively. r^2_{adj} refers to the adjusted coefficient of determination and F indicates Fischer statistics.

Besides employed rigorous external validation, the prediction quality of the developed model tested via MAE-based approach (Roy et al., 2015) and regarded as good. This indicates that there was not any systematic error occurring in the model.

Table 4.3. The results of Xternal Validation Plus tool for the developed model.

MAE-based Metrics	**Parameters**	**MAE-based parameters'value**
Model biasness test	**Systematic Error Result**	**Absent**
	nPE / nNE	1.1429
	nNE / nPE	0.8750
	|MPE / MNE|	0.7511
	|MNE / MPE|	1.3314
	AAE - |AE|	0.1402
	R^2 (Residuals; serial correlation)	0.0457
	R^2 (Residuals and Yobs values)	0.3130
	R^2_{Test} (100% data)	0.6067
	$R_0{}^2_{Test}$ (100% data)	0.6056
	$R_0{}'^2_{Test}$ (100% data)	0.4626
Classical Metrics	**Q^2_{F1} (100% data)**	0.9973
(for 100% data)	**Q^2_{F2} (100% data)**	0.6029
	Scaled Avg. r_m^2(100% data)	0.4682
	Scaled Δr_m^2 (100% data)	0.2250
	***CCC* (100% data)**	0.7654

Table 4.3. Continued.

MAE-based Metrics	Parameters	MAE-based parameters'value
	R^2_{Test} (95% data)	0.7753
	$R_0{}^2_{Test}$ (95% data)	0.7459
Classical Metric	$R_0{}'^2_{Test}$ (95% data)	0.3989
(after removing	Q^2_{F1} (95% data)	0.9984
5% data with	Q^2_{F2} (95% data)	0.7450
high residuals)	Scaled Avg r_m^2(95% data)	0.5065
	Scaled Δr_m^2(95% data)	0.2678
	CCC (95% data)	0.8306
	$RMSE_P$ (100% data)	0.2325
Error-based metrics	*SD* (100% data)	0.1792
(for 100% data)	*SE* (100% data)	0.0327
	MAE (100% data)	0.1517
	$RMSE_P$ (95% data)	0.1765
Error-based metric	*SD* (95% data)	0.1329
(after removing 5% data	*SE* (95% data)	0.0251
with high residuals)	*MAE* (95% data)	**0.1188**
	MAE+3**SD* (95% data)	**0.5175**
BASIC DATA STRUCTURE INFORMATION		
	NCompTest	30
Number of test set compounds.	Train range	12.6600
Range and Mean (train and test)	TrainYMean	5.0788
	Test range	1.6198
	TestYMean	0.5951
Distribution of observed	%Y(+/-0.5)TestMean	86.6667
response values of Test set	%Y(+/-1.0)TestMean	96.6667
around Test mean (in %)	%Y(+/-1.5)TestMean	100.0000
	%Y(+/-2.0)TestMean	100.0000
Distribution of observed	%Y(+/-0.5)TrainMean	0.0000
response values of Test set	%Y(+/-1.0)TrainMean	0.0000
around Train mean (in %)	%Y(+/-1.5)TrainMean	0.0000
	%Y(+/-2.0)TrainMean	0.0000
	%NComp>(0.1*TR)	0.0000
Distribution of prediction	%NComp>(0.15*TR)	0.0000
errors (in %)	%NComp>(0.2*TR)	0.0000
	%NComp>(0.25*TR)	0.0000
	(0.1*TrainingSetRange)	1.2660
Threshold values utilized	(0.15*TrainingSetRange)	1.8990
to judge the model predictions	(0.2*TrainingSetRange)	2.5320
	(0.25*TrainingSetRange)	3.1650
RESULT (*MAE*-based criteria applied on 95% data)	Prediction Quality	GOOD

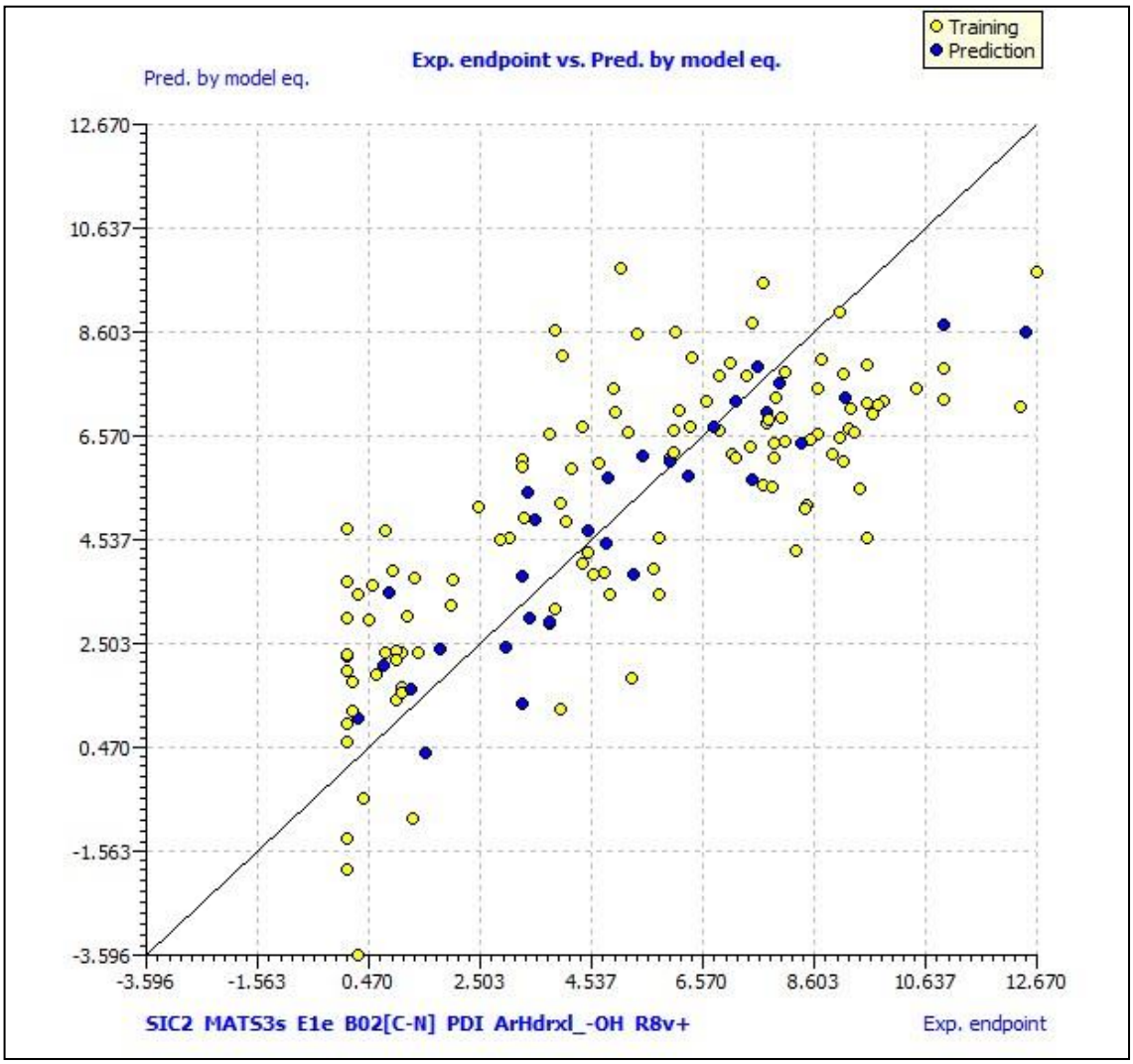


Figure 4.2. Plot of calculated/predicted vs. observed values of HOCldem for the training/test set compounds by model equation (Eq.4.1), with yellow labeled training set compounds and blue labeled test set compounds.

The proposed model includes 7 descriptors from various blocks of descriptors using 2 software packages. Six descriptors from Dragon 6.0 (Talete Inc., 2014) and one descriptor from ADMET Predictor 8.0 (Simulation Plus Inc., 2015) manifest several aspects of the molecular structure and classified as in Table 4.4.

Table 4.4. List of Dragon 6.0 and ADMET Predictor 8.0 descriptors appeared in the proposed model presents the abbreviations and meaning of models' descriptors.

Abbreviation of Descriptor	**Description**	**Descriptor Block**
ADMET Pred. 8.0		
ArHdrxl_-OH	Number of aromatic hydroxyl groups	Simple Constitutional Descriptors
DRAGON 6.0		
PDI	Packing Density Index	Molecular Properties
SIC2	Structural information content index (neighborhood symmetry of 2-order)	Information Indices
B02[C-N]	Presence/absence of C-N at topological distance 2	2D Atom Pairs
MATS3s	Moran autocorrelation of lag 3 weighed by I-state	2D Autocorrelation
R8v+	R maximal autocorrelation of lag 8/weighted by van der Waals volume	GETAWAY Descriptors
E1e	1st component accessibility directional WHIM index/ weighed by Sanderson electronegativity	WHIM Descriptors

The Dragon 6.0 (Talete Inc., 2014) based descriptors such as Packing Density Index (PDI), structural information content index(neighborhood symmetry of 2 order) (SIC2), presence/absence of C-N at topological distance 2 (B02[C-N]), Moran autocorrelation of lag 3 weighed by I-state (MATS3s), 1st component accessibility directional WHIM index/ weighed by atomic Sanderson electronegativity (E1e), R maximal autocorrelation of lag 8/weighted by van der Waals volume (R8v+) and only one ADMET Predictor 8.0 (Simulation Plus Inc., 2015) based descriptor the number of aromatic hydroxyl groups (ArHdrxl_-OH) occur in the model.

7 descriptors were used in the QSPR model proposed for HOCldem. Regardless of their sign, their importance could be written and explained as below based on the magnitude of standardized coefficients:

ArHdrxl_-OH > PDI > SIC2 > B02[C-N] > MATS3s > E1e > R8v+

The number of aromatic hydroxyl groups (ArHdrxl_-OH) is the first and the only ADMET Predictor (Simulation Plus Inc., 2015) descriptor appearing in the developed QSPR model belongs to simple constitutional descriptor group. They are normally relevant descriptors since the number of a certain functional group will obviously affect the properties of the molecule. ArHdrxl_-OH is also similar to the nArOH descriptor appearing in the priori model. The priori descriptor was calculated for each compound as the sum of aromatic hydroxyl groups. These types of descriptors are the most

basic and commonly used descriptors, reflecting the molecular composition of a compound without any additional information about its molecular geometry and topology, and also insensitive to any conformational change (Consonni and Todeschini, 2009; Luilo and Cabaniss, 2010). The number of atoms (atom numbers), number of bonds (bond number), absolute and relative numbers of specific atom-types (count descriptors), absolute and relative numbers of single, double, triple, and aromatic bonds, number of rings (cyclomatic number), number of rings divided by the number of atoms or bonds, number of benzene rings, number of benzene rings divided by the number of atoms, molecular weight and average molecular weight, atomic composition indices, and information index on size can be considered the most common constitutional descriptors (Consonni and Todeschini, 2009).

The packing density index (PDI) is a molecular property defined as the ratio between the McGowan volume (V_x) and the total surface area from P_VSA-like descriptors. PDI also simply defined as the fraction of the space filled by the shapes that creates the packing (Talete Inc., 2014). In literature, PDI was used for the utilization of structural descriptors to predict metabolic constants of xenobiotics in mammals (Pirovano et al.,2015). They found the positive correlation coefficient of PDI shows that binding increases with substrate size for cytochrome P450 (CYP) enzymes. In addition to this, it was revealed that a larger molecular size increases the possibility of interactions with the binding site and the hydrophobic nature of the molecules. In this study, PDI also shows positive correlation coefficient meaning that a larger molecular structure increases the possibility of interactions with chlorine.

The structural information content index of order 2 (SIC2) is another important molecular descriptor appearing in the Eq.4.1 which belongs to information indices as a neighborhood symmetry of order 2. The 2nd order SIC2 is defined in a normalized form of the information content to delete the influence of graph size. This topological descriptor is calculated as follows;

$$SIC_r = \frac{ICm}{\log 2\ A} \tag{4.2}$$

where A is the number of graph vertices. SIC2 considers a graph based on neighbor degree and edge multiplicity. Thus, it represents a measure for the structural complexity.

Atom pairs are substructure descriptors that collecting counts of occurrences of predefined structural properties (functional groups, augmented atoms, pharmacophore point pairs, atom pairs and triangles, surface triangles, etc.) in molecules or binary variables specifying their

presence/absence (Crowe et al., 1970; Lynch et al., 1971). Substructure descriptors can be divided into two categories, 2D substructure descriptors that are based on topological representation of molecules and 3D substructure descriptors that encode spatial relationships. The presence absence of C-N at topological distance 2 (B02[C-N]) belongs to 2D-Atom pairs. These types of molecular descriptors are also mentioned as topological atom pairs and sensitive to long-range correlations between the atoms in molecules (Consonni and Todeschini, 2009). The two-dimensional (also called as topological) representation of a molecule considers how the atoms are connected, that is, it defines the connectivity of atoms in the molecule in terms of the presence/absence and nature of chemical bonds.

Molecular descriptors based on the autocorrelation function AC_k, defined as;

$$AC_k = \int_a^b f(x)f(x+k)dx \tag{4.3}$$

where $f(x)$ is any function of the variable x and k is the lag representing an interval of x, and a and b define the total studied interval of the function. The autocorrelation function AC_k is the integration of the products of the function values calculated at x and $x+k$. This function expresses how numerical values of the function at intervals equal to the lag are correlated (Consonnni and Todeschini, 2009).

Moran autocorrelation of lag 3 weighed by I state (MATS3s) belongs to 2D autocorrelation descriptors and based on the topological distance matrix weighted on the electro topological I state. These descriptors A(d), in general, are calculated using the following equation (Eq. 4.4):

$$A(d) = \sum_{j=1}^{a} \sum_{i=1}^{a} \sigma(d_{ij} - d)\, p_i p_j$$

$$\sigma = \begin{cases} 1 (d_{ij} = d \\ 0 (d_{ij} \neq d \end{cases} \tag{4.4}$$

Where d refers to a topological distance which can take a number between 1 and the maximum distance in a given molecule, σ is a function of d_{ij}, which is the topological distance between atoms i and j, a refers to the number of atoms in the given molecule and p_i p_j are the properties of atoms i and j, respectively (Consonni and Todeschini, 2009).

Autocorrelation descriptors of chemical compounds can be calculated by using various molecular properties that can be represented at the atomic level or molecular surface level or else. The most widely used spatial autocorrelation molecular descriptors are obtained by taking the molecule atoms as the set of discrete points in space and an atomic property as the function evaluated at those points. In addition to the common weighting schemes mentioned above, the weighting schemes for atoms can be based on local vertex invariants such as the topological (vertex degrees), Kier-Hall (intrinsic states) or E-state indices (normalized distance complexity index) and related indices, alternatively (Consonnni and Todeschini, 2009).

It was shown that to get the best surface autocorrelation vectors for QSAR modeling, the van der Waals surface area is better than other molecular surfaces.

E1e is another molecular descriptor appeared in the model and negatively contributes the HOCldem activity of the molecules. 1st component accessibility directional WHIM index/ weighted by Sanderson electronegativity called E1e, belongs to WHIM descriptors (Weighted Holistic Invariant Molecular descriptors). WHIM descriptors based on statistical indices calculated on the projections of the atoms along principal axes (Todeschini and Gramatica, 1997; Consonni and Todeschini, 2009). The aim is to manifest 3D molecular information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. Basically, to calculate them weighting schemes were divided into six different categories for the atoms: the unweighted case (w_i=1 for all the atoms), atomic mass (m), van der Waals volume (v), Sanderson atomic electronegativity (e), atomic polarizability (p), and electro topological state indices of Kier and Hall (S).

$$s_{jk} = \frac{\sum_{i=1}^{A} w_i (qij - \bar{q}_j)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^{A} w_i} \qquad (4.5)$$

Where s_{jk} is the weighted covariance between the j^{th} and k^{th} atomic coordinates, A is the number of atoms, w_i the weight of of the i^{th} atom, q_{ij} and q_{ik} represent the jth and kth coordinate (j, k =x, y, z) of the i^{th} atom respectively, and $\bar{q}$ the corresponding average value.

Based on the type of weighting scheme, different covariance matrices and different principal axes (principal components) are obtained. Thereby, WHIM method can be classified as a general search for principal axes regarding a defined atomic property (weighting scheme). A set of statistical

indices is calculated on the atoms projected onto the principal component for each weighting scheme (Consonni and Todeschini, 2009).

The last descriptor appearing in the model (R8v+) R maximal autocorrelation of lag 8 weighed by van der Waals volume belongs to the GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors and defined in analogy to H-indices, (Consonni and Todeschini, 2009):

$$H= M \times (M^T \times M)^{-1} \times M^T \quad (4.6)$$

Where M is the molecular matrix composed of the centered Cartesian coordinates *x, y, z* of the molecule atoms (hydrogens included) in a chosen conformation. These indices encode similar information from the 3D-geometries of compounds. Especially, R8v provides information on the chance of a given atom *i* to interact with those atoms *j* at a topological distance. $d_{ij} = 8$, under the influence of the atomic van der Waals volumes (Cruz-Monteagudo et al., 2007). The last character of the descriptor "*v*" is the manifestation of the van der Waals volume.

GATEWAY descriptors were used in some literature study; for instance, R8v+ was used to model and check the anticancer activities of 1,4-naphthoquinone derivatives (Prachayasittikul et al., 2014). In the study conducted by Prachayasittikul et al. (2014), the most influential descriptor was R8v+ as deduced from its highest regression coefficient value. This study reveals that the shape and size of compounds are likely to play predominant roles in the cytotoxic activity against cancer cell lines. Another study including the computational modeling tools for the design of potent antimalarial bisbenzamidines reveal that elevated high van der Waals volumes could play a positive role in the molecular interactions responsible for the desired drug conformation that is required for the optimal binding to the macromolecular target (Cruz-Monteagudo et al., 2007).

In the study of Gramatica et al. (2008), E1e and MATS2e contemporaneously characterize molecular size and polarizability. The main conclusion of these two studies reveals the importance of electrostatic interaction for the HOCldem reactions. However, R8v+, E1e, and MATS3s have a positive/negative/negative weight and relation to the HOCldem activity, respectively; it should be considered that these descriptors are not simply and specifically indicators of electronegativity and are related to some other structural and geometrical properties (Consonni and Todeschini, 2009; Yousefinejad et al., 2017). Overall, the developed model had 2 indicator variables. For instance, B02[C-N] in model equation (4.1) descriptor has a range between 0-1 and ArHdroxl_-OH has a range between 0-2.

4.2. Analysis of the Applicability Domain

The AD of the Eq.4.1 exploited by Williams plot is shown in Figure 4.3.

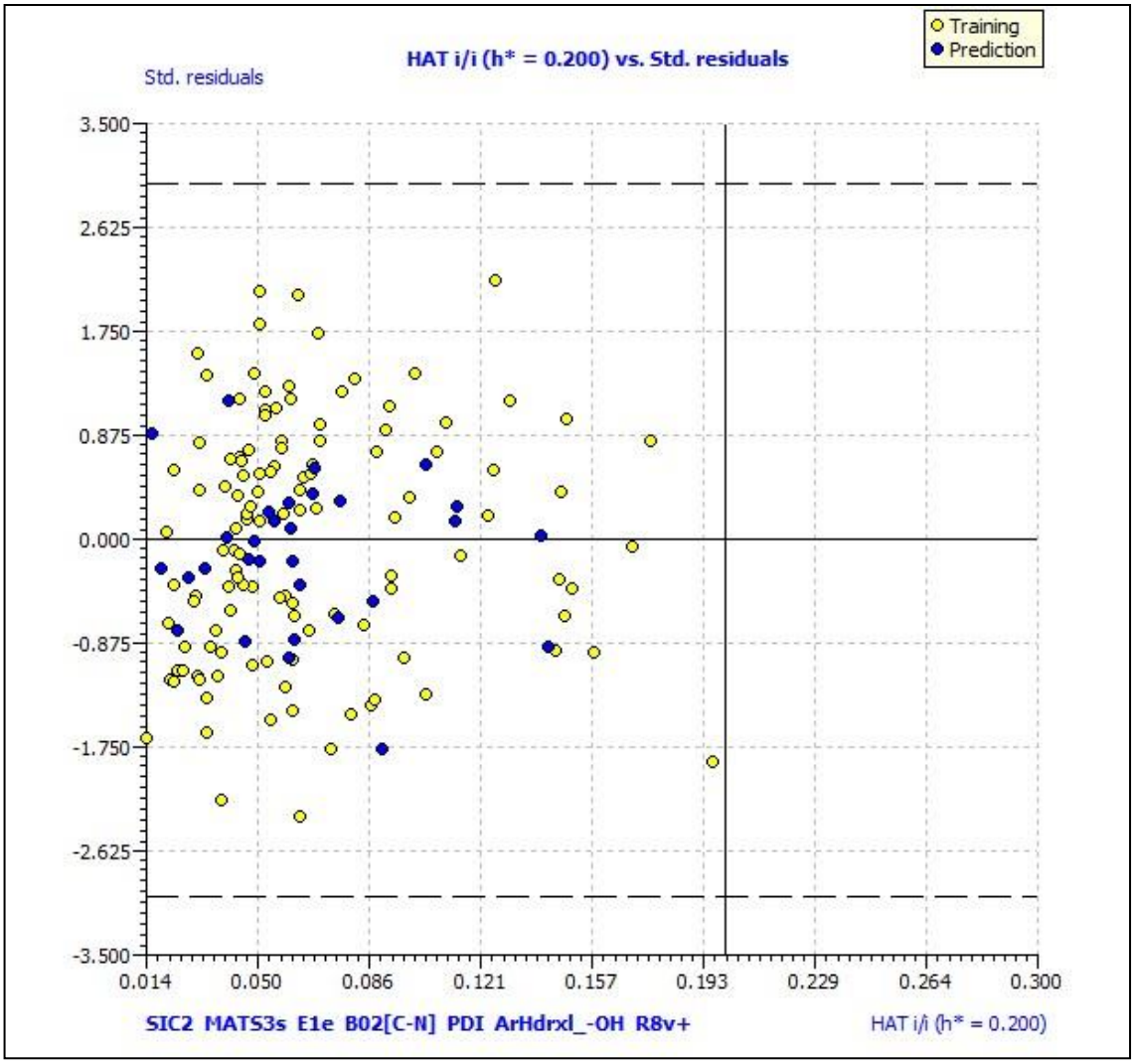


Figure 4.3. Williams plot for the Eq.4.1. with the yellow-labeled training set chemicals and the blue-labeled prediction (test set) chemicals. The dotted lines are the 3σ limit and the warning value of hat (h^* = 0.200), respectively.

It is clear that hat values of all the chemicals were lower than the critical hat value (h^*=0.200). For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units, meaning that the data set has no response values higher than the response outlier limit (±3σ). It is also important to note that there is no structural outlier for the developed model. In other words, all the HOCldem values of data set compounds are well predicted by the model equation (Eq.4.1.). Experimental and predicted HOCldem from Eq. 4.1, and descriptor values of training and test set chemicals in data set are given in Table 4.5.

The standardization method proposed by Roy et al., (2015) was also applied to our data set to check if there is outlier or not in the proposed model (Figure 4.4). The outcome of this method is consistent with our findings from leverage approach indicated by Williams Plot. Thus, there was not any outlier appearing in the model regarding the standardization approach.

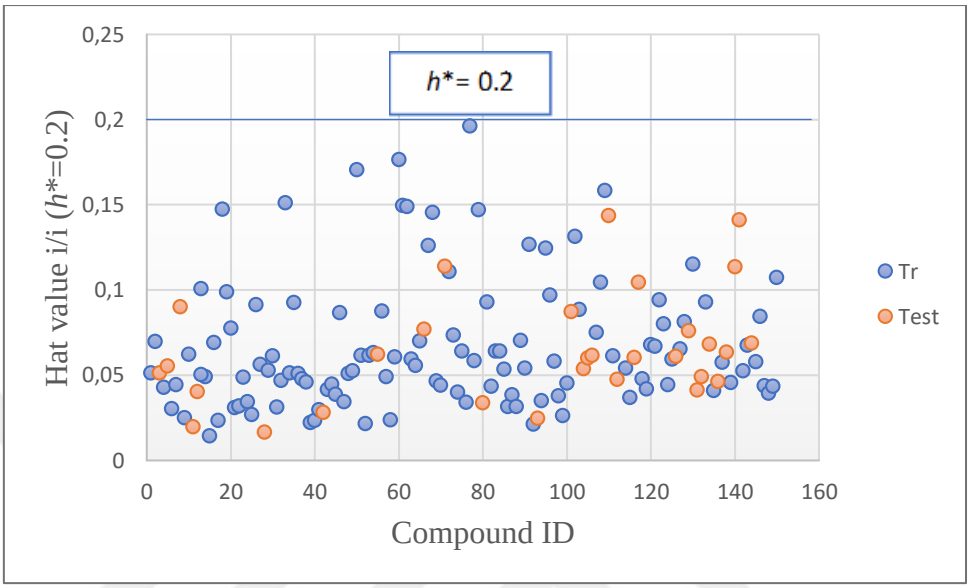


Figure 4.4. Training and Test Set compounds vs hat value graph.

Table 4.5. Chemicals that are used to model HOCldem, their experimental and predicted HOCldem values, descriptor values and residuals.

Compound	**Status**	**SIC2**	**MATS3s**	**E1e**	**B02[C-N]**	**PDI**	**R8v+**	**ArHdrxl_-OH**	**Pred. by model eq.**	**Exp. endpoint**	**Residual**
1,3-Dihydroxybenzene	Training	0.700	-0.246	0.506	0	0.813	0.000	2	7.7813	7.4	0.3813
5-Methylfurfural	Training	0.873	0.004	0.561	0	0.891	0.000	0	4.7273	0.8	3.9273
2,3-Dichlorophenol	Prediction	0.860	0.119	0.504	0	0.884	0.000	1	7.6091	8	-0.3909
2,3,6-Trichlorophenol	Training	0.818	0.176	0.528	0	0.890	0.000	1	6.6962	6.9	-0.2038
3,5-Dichlorophenol	Prediction	0.776	-0.486	0.515	0	0.884	0.000	1	7.9599	7.6	0.3599
2,4,6-Trichlorophenol	Training	0.776	-0.033	0.510	0	0.890	0.000	1	6.9375	8.02	-1.0825
2,4-Dichlorophenol	Training	0.860	-0.059	0.522	0	0.884	0.000	1	7.842	8.1	-0.258
4-Iodophenol	Prediction	0.751	-0.131	0.427	0	0.922	0.000	1	8.6043	12.5	-3.8957
4-Chlorophenol	Training	0.751	-0.196	0.512	0	0.875	0.000	1	6.7207	9.25	-2.5293
4-Nitrophenol	Training	0.761	-0.043	0.591	1	0.844	0.000	1	6.4487	7.9	-1.4513
4-Hydroxytoluene	Prediction	0.738	-0.121	0.545	0	0.869	0.000	1	5.7893	6.33	-0.5407
2,3,4,6-Tetrachlorophenol	Prediction	0.834	0.030	0.540	0	0.896	0.000	1	7.2533	7.2	0.0533
3,4,5-Triethoxybenzyl alcohol	Training	0.656	0.078	0.441	0	0.853	0.011	0	3.6544	0.55	3.1044
3-Aminobenzoic acid	Training	0.856	-0.062	0.522	1	0.890	0.000	0	6.848	7.74	-0.892
4-Hydroxybenzoic acid	Training	0.781	-0.092	0.548	0	0.824	0.002	1	5.5693	9.44	-3.8707
Coniferyl alcohol	Training	0.901	0.062	0.597	0	0.836	0.011	1	7.2541	6.66	0.5941
Methylsyringate	Training	0.795	0.037	0.496	0	0.825	0.003	1	6.2308	7.11	-0.8792
1,2,3-Trihydroxybenzene	Training	0.688	0.216	0.548	0	0.773	0.000	3	7.7811	6.9	0.8811
1,2-Dihydroxybenzene	Training	0.662	0.181	0.599	0	0.813	0.000	2	4.8939	4.1	0.7939
1,4-Dihydroxybenzene	Training	0.587	-0.195	0.509	0	0.813	0.000	2	6.1078	3.3	2.8078
2-Hydroxyacetophenone	Training	0.803	-0.177	0.508	0	0.868	0.000	1	7.2762	9.9	-2.6238
2-Hydroxybenzaldehyde	Training	0.829	-0.006	0.513	0	0.864	0.000	1	7.0202	9.7	-2.6798
2-Nitrophenol	Training	0.795	-0.022	0.563	1	0.844	0.000	1	7.2186	9.6	-2.3814

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
3-Hydroxyacetophenone	Training	0.830	-0.220	0.542	0	0.868	0.000	1	7.2928	11	-3.7072
3-Hydroxytoluene	Training	0.801	-0.172	0.554	0	0.869	0.000	1	6.6394	8.7	-2.0606
4,6-Dichloro-1,3-dihydroxybenzene	Training	0.775	-0.093	0.645	0	0.841	0.000	2	7.0512	5	2.0512
4-Chloro-1,3-dihydroxybenzene	Training	0.836	-0.132	0.577	0	0.828	0.000	2	8.6211	6.1	2.5211
4-Methoxyphenol	Prediction	0.701	-0.140	0.541	0	0.852	0.004	1	5.4834	3.4	2.0834
Acetophenone	Training	0.667	-0.138	0.551	0	0.916	0.000	0	2.9976	0.5	2.4976
Benzamide	Training	0.682	-0.062	0.544	1	0.939	0.000	0	5.2037	2.5	2.7037
Phenylacetic acid	Training	0.729	-0.206	0.523	0	0.872	0.002	0	3.7228	0.1	3.6228
Pyruvic acid	Training	0.857	0.009	0.545	0	0.721	0.000	0	1.3981	1	0.3981
Urea	Training	0.583	-1.074	0.507	0	0.811	0.000	0	2.9176	3.8	-0.8824
Ethylaceto acetate	Training	0.833	-0.379	0.602	0	0.803	0.003	0	3.2592	2	1.2592
3-(3,4,5-Trimethoxyphenyl) propanoic acid	Training	0.687	-0.141	0.546	0	0.851	0.017	0	3.8203	1.32	2.5003
1,3-Dihydroxynaphthalene	Training	0.728	-0.137	0.515	0	0.917	0.001	2	9.8744	5.1	4.7744
3,5-Dihydroxytoluene	Training	0.748	-0.283	0.551	0	0.824	0.000	2	8.1275	6.39	1.7375
3,4,5-Trimethoxybenzoic acid	Training	0.657	-0.005	0.503	0	0.843	0.006	0	2.3474	1.1	1.2474
4-Hydroxyacetophenone	Training	0.777	-0.198	0.564	0	0.868	0.004	1	6.6491	9.37	-2.7209
4-Hydroxybenzaldehyde	Training	0.795	-0.054	0.549	0	0.864	0.000	1	6.215	8.96	-2.745
Acetovanillone	Training	0.855	-0.094	0.541	0	0.857	0.004	1	7.4993	8.68	-1.1807
Vanillic acid	Prediction	0.867	-0.026	0.521	0	0.821	0.004	1	7.0426	7.75	-0.7074
3-Chlorophenol	Training	0.834	-0.258	0.525	0	0.875	0.000	1	7.8115	9.15	-1.3385

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
Butanal	Training	0.818	-0.048	0.590	0	0.790	0.000	0	1.7768	0.2	1.5768
Phenol	Training	0.618	-0.100	0.505	0	0.865	0.000	1	4.6011	9.6	-4.9989
Ferulic acid	Training	0.877	-0.077	0.586	0	0.838	0.015	1	7.9048	10.99	-3.0852
2-Chlorophenol	Training	0.792	0.152	0.530	0	0.875	0.000	1	6.0959	9.15	-3.0541
2-Aminophenol	Training	0.795	0.155	0.525	1	0.890	0.000	1	8.1544	4.02	4.1344
2-Napthol	Training	0.648	-0.099	0.529	0	0.967	0.001	1	6.7684	4.4	2.3684
1,3,5-Trihydroxybenzene	Training	0.594	-0.401	0.487	0	0.773	0.000	3	8.9963	9.1	-0.1037
L-Aspartic acid	Training	0.844	-0.233	0.486	1	0.735	0.000	0	4.6014	5.8	-1.1986
3-Methoxyphenol	Training	0.788	-0.201	0.533	0	0.852	0.000	1	6.494	8.1	-1.606
3-Nitroaniline	Training	0.781	-0.007	0.593	1	0.915	0.000	0	5.2439	8.5	-3.2561
3-Nitrobenzoic acid	Training	0.856	0.004	0.607	1	0.850	0.000	0	4.7521	0.1	4.6521
3-Nitrophenol	Prediction	0.829	-0.060	0.596	1	0.844	0.000	1	7.3281	9.2	-1.8719
4,4'-Dihydroxy biphenyl	Training	0.564	-0.095	0.565	0	0.915	0.007	2	7.4982	10.5	-3.0018
4-Aminophenol	Training	0.761	-0.169	0.523	1	0.890	0.000	1	8.5936	5.4	3.1936
4-Chloro-3,5-dimethylphenol	Training	0.709	-0.130	0.515	0	0.881	0.000	1	6.0653	4.7	1.3653
4-Chlorobenzoic acid	Training	0.795	-0.056	0.603	0	0.877	0.000	0	3.0173	0.1	2.9173
Acetothioamide	Training	0.763	-1.025	0.593	1	0.825	0.000	0	5.96	4.2	1.76
Acetic acid	Training	0.802	-1.021	0.575	0	0.684	0.000	0	2.2883	0.1	2.1883
Acetone	Training	0.473	-0.738	0.579	0	0.766	0.000	0	-1.2784	0.1	-1.3784
Acetyl acetone	Training	0.599	-0.559	0.534	0	0.802	0.000	0	1.2204	4	-2.7796
Anisole	Training	0.639	-0.058	0.532	0	0.902	0.000	0	2.3949	1	1.3949
Benzaldehyde	Training	0.655	0.066	0.556	0	0.919	0.000	0	2.2868	0.1	2.1868
Benzoic acid	Prediction	0.687	0.019	0.616	0	0.869	0.000	0	1.0559	0.3	0.7559

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
Citric acid	Training	0.715	-0.320	0.397	0	0.682	0.003	0	1.9241	0.63	1.2941
Ethanol	Training	0.763	-0.184	0.672	0	0.676	0.000	0	-1.9141	0.1	-2.0141
Fumaric acid	Training	0.721	-0.449	0.556	0	0.718	0.000	0	0.6097	0.1	0.5097
Maleic acid	Training	0.721	-0.449	0.531	0	0.718	0.000	0	0.9448	0.1	0.8448
Malonic acid	Prediction	0.737	-0.838	0.475	0	0.691	0.000	0	2.4164	1.8	0.6164
Nitrobenzene	Training	0.618	0.081	0.604	1	0.894	0.000	0	2.2921	0.1	2.1921
N, N-Diethylaniline	Training	0.625	0.022	0.509	1	0.911	0.002	0	4.3532	8.3	-3.9468
Propanal	Training	0.797	-0.058	0.578	0	0.766	0.000	0	1.2175	0.2	1.0175
Phenylthiourea	Training	0.729	-0.296	0.534	1	0.952	0.003	0	7.1466	12.4	-5.2534
Phenoxyacetic acid	Training	0.748	-0.076	0.558	0	0.857	0.008	0	3.4846	0.3	3.1846
Oxalic acid	Training	0.667	0.398	0.552	0	0.652	0.000	0	-3.5965	0.3	-3.8965
Succinic acid	Training	0.662	-0.420	0.494	0	0.719	0.000	0	0.6011	0.1	0.5011
Thiourea	Training	0.583	-1.061	0.541	0	0.850	0.000	0	3.1885	3.9	-0.7115
3,5-Dimethoxybenzoic acid	Prediction	0.728	-0.111	0.580	0	0.849	0.004	0	2.4497	3	-0.5503
Sinapic acid	Training	0.781	-0.023	0.460	0	0.834	0.015	1	8.0995	8.77	-0.6705
Vanillin	Training	0.882	0.038	0.516	0	0.853	0.000	1	7.3505	7.92	-0.5695
Sinapyl alcohol	Training	0.805	0.146	0.516	0	0.832	0.014	1	7.0748	6.14	0.9348
2-Oxobutyric acid	Training	0.860	0.241	0.521	0	0.748	0.000	0	1.6687	1.1	0.5687
3-Oxohexanedioic acid	Training	0.802	-0.416	0.545	0	0.755	0.010	0	3.4951	5.8	-2.3049
Citraconic acid	Training	0.816	-0.250	0.542	0	0.741	0.000	0	1.9752	0.1	1.8752
P-Coumaric acid	Training	0.815	-0.141	0.565	0	0.843	0.010	1	7.1176	9.29	-2.1724
2-Oxopentanedioic acid	Training	0.844	-0.100	0.530	0	0.738	0.003	0	2.3581	1.4	0.9581
3-Hydroxybutyric acid	Training	0.885	-0.355	0.478	0	0.692	0.000	0	3.0706	1.2	1.8706

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
3-Oxopentanedioic acid	Training	0.719	-0.633	0.542	0	0.738	0.002	0	1.8573	5.3	-3.4427
1,2,4-Trihydroxybenzene	Training	0.688	-0.056	0.537	0	0.773	0.000	3	8.653	3.9	4.753
2-Hydroxybenzoic acid	Training	0.813	-0.084	0.518	0	0.824	0.000	1	6.1681	6	0.1681
2-Hydroxytoluene	Prediction	0.769	0.020	0.551	0	0.869	0.000	1	5.7441	7.5	-1.7559
2-Methoxyphenol	Training	0.759	0.156	0.498	0	0.852	0.000	1	5.6279	7.7	-2.0721
2-Oxoacetic acid	Training	1.000	-0.016	0.622	0	0.682	0.000	0	1.5665	1.1	0.4665
d-Mannose	Training	0.742	0.105	0.501	0	0.648	0.003	0	-0.9076	1.3	-2.2076
3-Aminophenol	Training	0.829	-0.226	0.527	1	0.890	0.000	1	9.593	7.7	1.893
3-Hydroxybenzaldehyde	Training	0.863	-0.079	0.548	0	0.864	0.000	1	7.1961	9.8	-2.6039
3-Hydroxybenzoic acid	Training	0.844	-0.110	0.524	0	0.824	0.000	1	6.5678	9.1	-2.5322
L-Glutamic acid	Training	0.851	-0.111	0.520	1	0.753	0.003	0	4.5739	3.05	1.5239
3,4,5-Trimethoxybenzyl alcohol	Prediction	0.658	0.127	0.614	0	0.841	0.005	0	0.3793	1.52	-1.1407
3-(4-Hydroxy-3-methoxyphenyl) propionic acid	Training	0.840	-0.117	0.614	0	0.835	0.019	1	7.4993	4.94	2.5593
L-Alanine	Training	0.860	0.032	0.434	1	0.744	0.000	0	4.9803	3.33	1.6503
L-Isoleucine	Prediction	0.801	0.123	0.499	1	0.782	0.000	0	3.8265	3.3	0.5265
L-Proline	Prediction	0.856	0.026	0.462	1	0.828	0.000	0	6.2076	5.5	0.7076
L-Valine	Prediction	0.758	0.094	0.501	1	0.768	0.000	0	3.0339	3.43	-0.3961
L-Serine	Training	0.925	0.023	0.557	1	0.707	0.000	0	3.4948	4.9	-1.4052
L-Tyrosine	Training	0.836	-0.008	0.521	1	0.843	0.015	1	9.8134	12.67	-2.8566
L-Arginine	Training	0.885	-0.008	0.628	1	0.834	0.019	0	6.532	8.55	-2.018
L-Leucine	Prediction	0.781	0.084	0.671	1	0.782	0.000	0	1.36	3.3	-1.94

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
L-Cysteine	Training	0.925	0.040	0.490	1	0.771	0.000	0	5.5968	7.87	-2.2732
L-Lysine	Prediction	0.768	0.107	0.515	1	0.815	0.006	0	4.4797	4.83	-0.3503
3,5-Dihydroxybenzoic acid	Training	0.777	-0.248	0.530	0	0.788	0.000	2	7.9974	7.08	0.9174
2,4-Dihydroxybenzoic acid	Training	0.834	-0.210	0.534	0	0.788	0.002	2	8.804	7.5	1.304
4-Aminobenzoic acid	Training	0.799	-0.048	0.536	1	0.890	0.003	0	6.1767	7.9	-1.7233
4-Cyanophenol	Prediction	0.775	-0.087	0.472	1	0.865	0.000	1	8.7563	11	-2.2437
1.4-Phenyldiamine	Prediction	0.563	-0.186	0.522	1	0.974	0.000	0	4.935	3.53	1.405
2-Aminobenzoic acid	Training	0.827	-0.050	0.503	1	0.890	0.000	0	6.6863	6.04	0.6463
Methoxyacetic acid	Training	0.822	-0.230	0.491	0	0.724	0.000	0	2.3532	0.79	1.5632
L-Asparagine	Training	0.914	-0.230	0.482	1	0.790	0.000	0	6.6482	5.21	1.4382
L-Threonine	Training	0.903	-0.032	0.512	1	0.709	0.000	0	3.9919	5.7	-1.7081
L-methionine	Training	0.876	0.085	0.420	1	0.801	0.004	0	6.7635	6.35	0.4135
Aniline	Training	0.618	-0.135	0.523	1	0.956	0.000	0	5.1633	8.44	-3.2767
L-Phenylalanine	Training	0.785	0.059	0.533	1	0.876	0.008	0	5.9883	3.28	2.7083
1-Napthol	Training	0.639	0.010	0.537	0	0.967	0.000	1	6.1485	7.2	-1.0515
L-Glycine	Prediction	0.880	-0.087	0.486	1	0.737	0.000	0	4.7286	4.5	0.2286
3,4,5-Triethoxybenzoic acid	Training	0.654	0.021	0.561	0	0.854	0.011	0	2.1909	0.99	1.2009
3,4,5-Trimethoxyacetophenone	Training	0.654	-0.081	0.436	0	0.871	0.006	0	3.9547	0.93	3.0247
3,4,5-Trimethoxybenzamide	Prediction	0.658	-0.044	0.588	1	0.883	0.007	0	3.876	5.33	-1.454
3,4,5-Trimethoxyphenylacetonitrile	Training	0.658	-0.058	0.605	1	0.872	0.013	0	4.0889	4.38	-0.2911
3,4,5-Trimethoxyphenylacetic acid	Prediction	0.687	-0.188	0.511	0	0.847	0.009	0	3.5123	0.84	2.6723

Table 4.5. Continued.

Compound	Status	SIC2	MATS3s	E1e	B02[C-N]	PDI	R8v+	ArHdrxl_-OH	Pred. by model eq.	Exp. endpoint	Residual
4-Allyl-2,6-dimethoxyphenol	Prediction	0.781	0.252	0.482	0	0.861	0.007	1	6.7751	6.8	-0.0249
L-Arabinose	Training	0.838	0.124	0.521	0	0.636	0.000	0	-0.4974	0.4	-0.8974
4-methyl-2,6-dimethoxyphenol	Prediction	0.725	0.252	0.446	0	0.849	0.002	1	5.7661	4.87	0.8961
Acetosyringone	Training	0.752	-0.012	0.458	0	0.850	0.004	1	6.8918	7.79	-0.8982
Syringaldehyde	Prediction	0.755	0.112	0.456	0	0.845	0.003	1	6.4275	8.38	-1.9525
Syringic acid	Training	0.750	0.031	0.434	0	0.819	0.003	1	6.3643	7.46	-1.0957
3-Oxobutanedioic acid	Prediction	0.917	-0.371	0.557	0	0.717	0.000	0	2.9664	3.8	-0.8336
4,6-Dioxoheptanoic acid	Training	0.818	-0.403	0.566	0	0.793	0.008	0	3.9267	4.8	-0.8733
4-Oxoheptanedioic acid	Prediction	0.633	-0.308	0.563	0	0.769	0.016	0	1.618	1.25	0.368
5,7-Dioxooctanoic acid	Prediction	0.830	-0.305	0.503	0	0.802	0.020	0	6.0811	6	0.0811
L-Glutamine	Training	0.907	-0.118	0.495	1	0.802	0.003	0	6.6262	3.8	2.8262
Benzothioamide	Training	0.682	0.017	0.541	1	0.951	0.000	0	5.2677	4	1.2677
L-Malic acid	Prediction	0.863	-0.242	0.474	0	0.670	0.000	0	2.1023	0.75	1.3523
4-(3,4,5-trimethoxybenzoyl) Butyric acid	Training	0.729	-0.126	0.563	0	0.854	0.013	0	3.7555	2.01	1.7455
Trans-3,5-dimethoxy-4-hydroxycinnam aldehyde	Training	0.786	0.084	0.495	0	0.858	0.016	1	7.9832	9.59	-1.6068
L-Ornithine	Training	0.805	0.091	0.555	1	0.806	0.002	0	3.8887	4.6	-0.7113
Ethyl-(3,4,5-trimethoxybenzyl) acetate	Training	0.735	-0.141	0.515	0	0.865	0.009	0	4.321	4.5	-0.179
3-(4-hydroxy-3,5 dimethoxyphenyl) Propanic acid	Training	0.759	-0.067	0.556	0	0.831	0.012	1	6.2708	6.06	0.2108
N-Acetylneuraminic acid	Training	0.844	0.012	0.489	1	0.704	0.012	0	4.5398	2.9	1.6398

4.3. Prediction of HOCldem Values of External Set Chemicals with No Data

When PPCPs are not completely degraded by the conventional treatment methods, they are transferred into sewage treatment plants (STPs) which is originally designed for the removal of organic matter and suspended solids to meet the minimum discharge standard (Yang et al., 2017). Thereby, the excreted antibiotic metabolites and resistant bacteria from urban and hospital wastewater get into the aquatic environment with sewage sludge and/or sewage effluent and can be further modified in receiving water bodies (Klatte et al., 2017). Even though several advanced treatment systems, including membrane filtration, granular activated carbon, and advanced oxidation processes have been used for the effective removal of individual PPCPs, they might be insufficient when it comes to partially metabolized PPCP removal and antibiotic resistance (mentioned in section 2.5) (Klatte et al., 2017; Yang et al., 2017). It is also known that PPCP residues are found in drinking water (Table 4.6). Besides the variety of the released chemicals, chlorination may add new chlorinated derivatives of these chemicals. Although it is advisable to find alternative options on the basis of scientific knowledge to minimize the emission pathways in environmental fragments, considering the large number of PCPPs influencing our lives, it is also important to search for their chlorine demand.

Table 4.6. Concentrations of active pharmaceuticals ingredients (APIs) found in finished drinking water worldwide (Jones et al., 2005).

Compound	**Therapeutic Group**	**Max. con. Detected (ng l^{-1})**	**Country**	**Reference**
Bezafibrate	Lipid regulator	27	Germany	(Stumpf et al.,1996)
Bleomycin	Anti-neoplastic	13	UK	(Aherne et al., 1990)
Clofibric acid	Lipid regulator	(+) identification 270 5.3	UK Germany Italy	(Fielding et al., 1981) (Heberer et al.,1997) (Zuccato et al.,2000)
Carbamazepine	Anti-epileptic	24 258	Canada USA	(Asana et al., 2004) (Stackelberg et al., 2004)
Diazepam	Psychiatric drug	10 23.5	UK Italy	(Waggot et al., 1981) (Zuccato et al.,2000)
Diclofenac	Analgesic and Anti-pyretic	6	Germany	(Stumpf et al.,1996)
Gemfibrozil	Lipid regulator	70	Canada	(Tauber et al., 2003)

Table 4.6. Continued.

Compound	Therapeutic Group	Max. con. Detected (ng l^{-1})	Country	Reference
Ibuprofen	Analgesic and Anti-pyretic	3	Germany	(Stumpf et al.,1996)
Phenazone	Analgesic and Anti-pyretic	250 400	Germany Germany	(Zhülke et al., 2004) (Reddersen et al., 2002)
Propylphenazone	Analgesic and Anti-pyretic	80 120	Germany Germany	(Zhülke et al., 2004) (Reddersen et al., 2002)
Tylosin	Macrolide	1.7	Italy	(Zuccato et al.,2000)

Therefore, in the present study, a new valid QSPR model is proposed to be used for the prediction of chlorine demand of various class of pharmaceuticals with no experimental HOCldem data. The proposed model (Eq. 4.1.) was also employed for the prediction of HOCldem values of 110 chemicals with no chlorine demand data using Insubria Graph (Figure 4.5). This external set of chemicals are mainly pharmaceuticals from different class, such as antibiotics, antidepressants, anti-inflammatory, anti-psychotic, anti-viral, sedative and hypnotic drugs. Surprisingly, among 110 chemicals, only limited number of chemicals were out of the structural applicability domain.

Analysis of the predicted HOCldem data of external set chemicals with regard to applicability domain via leverage approach and descriptor range presents a useful overall estimate for the chlorine demand profiles of chemicals. The chemicals that are mainly used in industry including aldehydes (Capron aldehyde), volatile organics (Ethyl bromide), and glycol ethers (1-Methoxy-2-propanol) have the lowest chlorine demand (expressed as 0.78, 0.54, 1.5 mol HOCl/mol compound, respectively) whilst the pharmaceuticals including antibiotics (tetracycline, pazufloxacin, cinoxacin) and an anti-viral(famciclovir) have the highest chlorine demand (expressed as 10.87, 10.55, 10.07, 9.65 mol HOCl/mol compound, respectively).

The chemical groups of the compounds, the predicted HOCldem values from equation 4.1. and the descriptor values of external set chemicals with no chlorine demand data are given in Appendix B.

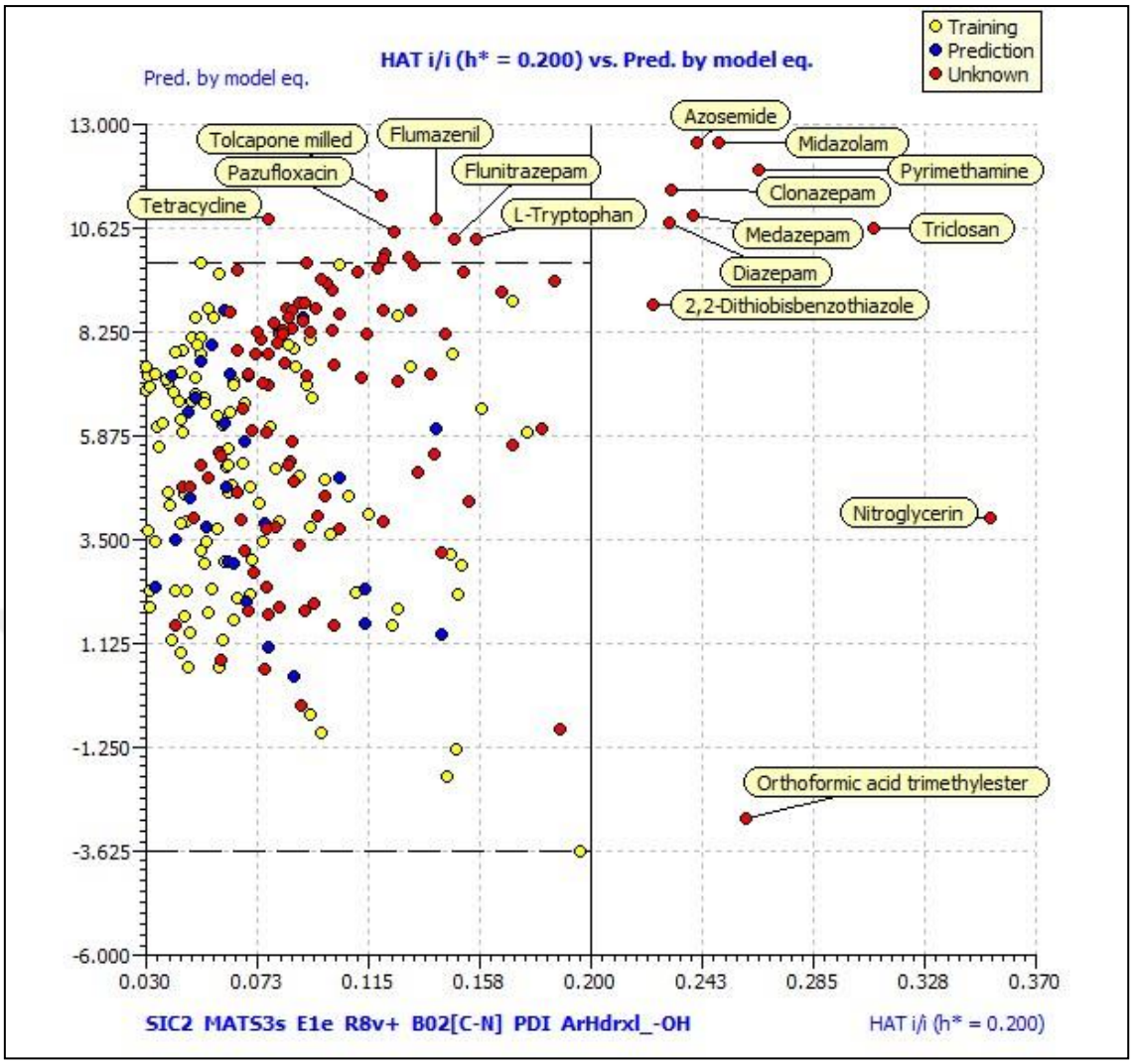


Figure 4.5. Insubria graph indicating the predicted HOCldem values from Eq.4.1 for training, test and external set chemicals; training set in yellow, test set in blue, and external set in red.

The developed model has 91% structural coverage with these diverse group of medicinal chemicals. PCPPs are considered major groups of emerging contaminants that are found in several environmental compartments with different kinds of by-products and different concentration. The external set has an impact on the environment due to consisting of various medicinal groups. Most of antibiotics, anti-inflammatory drugs, hormones, beta-blockers, and industrial chemicals well predicted with equation 4.1. with no experimental chlorine demand data.

The compounds that fall outside the AD mainly belong to psychotropics (benzodiazepines), loop diuretics (Azosemide), anti-protozoans (Pyrimethamine), and antibiotics (Triclosan). The reason behind this should be the heterogenicity of the external data set, there were many different chemical structures. However, 100 compounds belonging to different medicinal classes were well predicted by

the model equation (Eq.4.1). This model was considered reliable for predicting chlorine demand values of PCPP groups.

The compounds that fall outside the AD are given in Table 4.7 together with their Anatomical Therapeutic Chemical (ATC) Classification codes. The main chemical group that fall outside the AD are benzodiazepines. They exert their effect by binding to the benzodiazepine receptor at the Gamma-aminobutyric acid (GABA) receptors. Some of these compounds were used to treat epilepsy by increasing the release of GABA in the brain, to affect neurotransmitters in the brain that may be unbalanced in people with anxiety. This procedure helps to calm the excessive electrical nerve activity of the central nervous system and brain. They also used in the field of anesthesia. The most common forms of benzodiazepines are Xanax, Diazem, and Tranko-buskas. Currently, the sale and supply of psychotropic drugs is strictly regulated under the Turkish Ministry of Health. Azosemide is the second pharmaceutical compound with higher chlorine demand that belong to loop diuretics. The third group including triclosan is widely used as anti-bacterial and anti-fungal agent in personal care products such as soaps, skin moisturizers, deodorants and toothpaste. The fourth category including 2,2'-dithiobisbenzathiazole belongs to allergens that exert its effect by means of increased histamine release and trimethyl orthoformate is defined as a reagent in organic synthesis and a number of pharmaceutical intermediates made from this compound.

Table 4.7. The Compounds that fall outside the AD and their ATC (Anatomical Therapeutic Chemical) Classification codes, predicted HOCldem values (mol HOCl/mol Compound).

Chemical Group	**Compound Name**	**ATC Code**	**Brand Name**	**Pred. HOCldem**	**HAT values (*h**=0.2)**
Sedative-hypnotic	Midazolam	N05CD08		12.62	0.26
Anti-epileptic (Benzodiazepine)	Clonazepam	N03AE01	Rivotril	11.5	0.24
	Medazepam	N05BA03	Tranko-buskas	10.96	0.24
	Diazepam	N05BA01	Diazem	10.77	0.23
Loop diuretics	Azosemide	SLC12A1		12.6	0.25
Anti-protozoal	Pyrimethamine	P01BD01		11.96	0.26
Anti-bacterial Anti-fungal agent	Triclosan	D09AA06	Kursept	10.64	0.35

Table 4.7. Continued.

Chemical Group	Compound Name	ATC Code	Brand Name	Pred. HOCldem	HAT values (h^*=0.2)
Industrial	2,2'-Dithiobisbenzothiazole			8.90	0.23
Orthoester	Trimethyl orthoformate			-2.84	0.26
Vasodilator Vasoconstractor	Nitroglycerin			4.00	0.35

The AD of proposed model was also tested via the ranges of descriptor space for the external set chemicals. The range of each descriptor appearing in the proposed model is given in Table 4.8. Of the descriptors, the PDI value of 17 compounds fall outside of the descriptor range. 7 fluoroquinolone antibiotics, 4 industrial chemicals, 2 antidepressants, 2 sedative and hypnotics, a diuretic, a gastroprotective and an anti-psychotic belonging to this group stay outside the PDI descriptor range. The second descriptor R8v+ with two compounds which fall outside of the descriptor range mainly belonging to muscle relaxants and herbicides. The third descriptor E1e with one compound which belongs to amino sugars. Even though abovementioned chemicals fall in the applicability of the QSPR model, their prediction may be unreliable. Therefore, with the application of the two methods on the prediction of HOCldem for external set chemicals, final coverage of the proposed model was found as 73%.

Table 4.8. Range of descriptors appeared in the proposed model.

Descriptor	Minimum Value	Maximum Value	% External Set within the descriptor range
R8v+	0	0.019	91%
PDI	0.636	0.967	78%
E1e	0.397	0.672	99%
SIC2	0.473	1	99%
ArHdrxl_-OH	0	3	100%
MATS3s	-1.074	0.398	100%
B02[C-N]	0	1	100%

4.4. Comparison with the Proposed QSPR Model with the Literature Model

It is difficult to make a strict comparison between our model and the model developed by Luilo and Cabaniss (2010) since the quality of prediction depends on various parameters. But even in this case, general considerations can be made. The studied physico-chemical property, chlorine demand has already been modeled by using initial 26 constitutional descriptors. These variables 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 that can be calculated from those (for instance; H:C ratio, O:C ratio and the number of phenol groups per ring) for each compound make up the descriptor pool (Table 4.9). Whereas, in the present study 3056 descriptors that covers such as quantum chemical, geometrical, electrostatic, topological etc. were calculated by using different software packages.

Table 4.9. Average Model Coefficients and Standard Errors for descriptors used in the published model.

Descriptor (x_j)	**Coefficient (β_j)**	**StdE**	**(StdE/β_j) *100**
RAI	7.61	0.34	4.5 %
ArOH	1.16	0.26	22.4 %
ACN	3	0.2	6.7 %
CI	1.23	0.23	18.7 %
O:C	1.01	0.28	27.7 %
AS	2.37	0.54	22.8 %
ArORact	0.49	0.17	34.7 %
ArORnonact	-0.72	0.28	38.9 %

It is not surprising to find common descriptors for both models. The descriptor "nArOH" appearing in the published model that emphasizes the contribution of the number of aromatic -OH groups to reactivity toward chlorine consistent with the ArHdrxl_-OH in the developed model. ArHdrxl_-OH is the only Admet Predictor 8.0 (Simulation Plus Inc., 2015) descriptor that belongs to simple constitutional descriptors appearing in the developed model. It is clear that the number of OH atoms effective in reactivity toward chlorine.

It is known that the QSAR model validation is essential to ensure the reliability of predicted data. In the published model, the data set split into three categories, a calibration, a cross validation and an external validation set. Multiple Linear Regression (MLR) was used to calibrate the published QSPR model. The statistical parameters like the determination coefficient (R^2), the standard error of regression ($StdE_{reg}$), mean bias deviation (*MBD*), and root-mean- squared error (*RMSE*) for cross-validation and external validation were used to validate the model. In this study, along with the above-mentioned metrics, different validation metrics such as *CCC*, r^2_m, Golbraikh and Tropsha criteria were used.

In other words, Luilo and Cabaniss (2010) reported limited statistical metrics for their QSPR model and their model has not been tested with the up-to-date internal and external validation metrics. However, our model with an external set including many emerging contaminants such as antibiotics, anti-inflammatories, hormones, sedative and psychotropics, has 73% predictive ability. Thus, in the present study, the proposed model has a potential to notify the environmental scientists about the underdetermined fate of chemicals in the aquatic environments.

5. CONCLUSION

In the present study, chlorine demand of organic chemicals was determined. Numerous QSPR models were generated for different training and test set divisions. All models were validated internally and externally using recent metrics reported in the literature and the OECD principles. Their applicability domains were defined via Williams plot, standardization approach and ranges of descriptor space. The best model from each division was selected using MCDM approach as implemented in QSRAINS. The predictive performance of final models was compared with an external set of 110 chemicals mainly comprised of pharmaceuticals. Of the final four models, the one with the highest structural coverage for the external set chemicals was proposed as the best model.

The chlorine demand of the compounds was found to be related to the six theoretical descriptors from DRAGON 6.0 (Talete Inc., 2014) and one descriptor from ADMET Predictor 8.0 (Simulation Plus Inc., 2015). Descriptors appearing in the model were based on 2D and 3D geometries of the molecule. The atomic van der Waals volumes, the packing shape of molecule related to the 3D geometry, the weighted Sanderson atomic electronegativity, neighborhood symmetry of 2-order, the presence/absence of C-N at a topological distance 2, and the number of aromatic hydroxide are found to be effective in reactivity toward chlorine.

The structural coverage of the final proposed model was 91 % for the external set chemicals with no chlorine demand data when leverage approach was used. Additionally, the structural coverage of the proposed model was evaluated by the descriptor range. The predictive ability of the proposed model is 73% when the range of descriptors space was used to define the AD. Within this context, the predicted chlorine demand of compounds that fall in the applicability domain of the model was evaluated as unreliable since some of their descriptors were out of the descriptor range. A significant amount of chemicals' chlorine demand (80 compounds) was reported for the first time in the literature. Also, the proposed model has a potential to be used to fulfill the data gaps of similar emerging contaminants that fall in the AD of the model since the data set mainly composed of a diverse set of pharmaceuticals. In this regard, the proposed model was found to be superior to the previously published literature model.

The unexpected findings of this study are the higher chlorine consumption rates and the lower removal efficiency of pharmaceuticals. Comparing external set chemicals with regard to chlorine consumption, pharmaceuticals mainly composed of antibiotics, and anti-virals consume chlorine far

more than industrial chemicals including aldehydes, and disinfectants. Recent technologies such as ozone and/or granular activated carbon may be used in water treatment instead of chlorination. They are capable of removing a wide range of organic substances that might be present in source water but they are not effective for the compounds with higher water solubility and/or poor degradability. Although the pharmaceutical concentrations detected in tap water is far below the regulatory limits, considering the potential health risks in long-term exposure and/or short-term exposure little is known. Another side of the same coin is the synergistic power of mixtures. Considering the diversity and quantity of chemicals that released to the environment, it is crucial to completely understand their analysis, occurrence, toxicity, property, formation, degradation, and removal for the researchers, regulators, and water suppliers. An important feature of this work was to emphasize the significance of toxicity tests, regarding the variety of chemicals that consume chlorine to form or trigger the formation of various disinfection byproducts. Additionally, the proposed model can be used to screen the chemicals with high chlorine demand, and to designate the priority of these chemicals for further action.

REFERENCES

Abdullah, M.P., Yee, L.F., Ata, S., Abdullah, A., Ishak, B., Abidin, K.N.Z., 2009. The study of interrelationship between raw water quality parameters, chlorine demand and the formation of disinfection by-products. Physics and Chemistry of the Earth, 34, 806–811.

Adamson, G.W., Lynch, M.F., and Town, W.G., 1971. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part II. Atom-centered fragments. Journal of the Chemical Society: C, 3702-3706.

Ankley, G.T., Brooks, B.W., Huggett, D.B., Sumpter, J.P., 2007. Repeating History: Pharmaceuticals in the environment. Environmental Science and Technology, 41, 8211-8217.

Aptula, O.A., Jeliazkova, N.G., Schultz, T., Cronin, M.T.D., 2005. The better predictive model: q^2 for the training set or low root mean square error of prediction for the test set. QSAR Combinatorial Science, 24, 385-396.

Arnold, W.A., Bolotin, J., von Gunten, U., Hofstetter, T.B., 2008. Evaluation of functional groups responsible for chloroform formation during water chlorination using compound specific isotope analysis. Environmental Science and Technology, 42 (21), 7778-7785.

Ates, N., Kitis, M., Yetis, U., 2007. Formation of chlorination by-products in waters with low SUVA – correlations with SUVA and differential UV spectroscopy. Water Research, 41, 4139-4148.

Aydın, E., Talinli, I., 2013. Analysis, occurrence and fate of commonly used pharmaceuticals and hormones in the Buyukcekmece Watershed, Turkey. Chemosphere, 90, 2004-2012.

Baxter, C.W., Smith, D.W., Stanley, S.J., 2004. A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. Journal of Environmental Engineering and Science, 3, 45-58.

Bhhatarai, B., Garg, R., Gramatica, P., 2010. Are mechanistic and statistical QSAR approaches really different? MLR studies on 158 cycloalkyl-pyranones. Molecular Informatics, 29, 511-522.

Bond, T., Henriet, O., Goslan, E.H., Parsons, S.A., Jefferson, B., 2009. Disinfection byproducts and fractionation behavior of natural organic matter surrogates. Environmental Science and Technology, 43, 5982-5989.

Boyce, S., Hornig, J., 1983. Reaction pathways of trihalomethane formation from the halogenation of dihydroxy aromatic model compounds for humic acid. Environmental Science and Technology, 17, 202-211.

Bull, R.J., Reckhow, D.A., Rotello, V., Bull, O.M., Kim, J., 2006. Use of toxicological and chemical models to prioritize DBP research. AWWA Research Foundation, Journal of American Water Works Association, 35-41, U.S.A.

Bultinck, P., Van Damme, S., 2007. A new computer program for QSAR-analysis: ARTE-QSAR. Journal of Computational Chemistry, 28, 1924-1928.

Chattaraj, P.K., Sarkar, U., and Roy, D.R., 2006. Electrophilicity index. Chemical Reviews, 106, 2065-2091.

Chen, B., Zhang, T., Bond, T., Gan, Y., 2015. Development of quantitative structure activity relationship (QSAR) model for disinfection by-product (DBP) research: A review of methods and resources. Journal of Hazardous Materials, 299, 260-279.

Chiang, P.C., Chang, E.E., Chuang, C.C., Liang, C.H., Huang, C.P., 2010. Evaluating and elucidating the formation of nitrogen-contained disinfection by-products during pre-ozonation and chlorination. Chemosphere, 80(3), 327-333.

Chirico, N., Gramatica, P., 2011. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. Journal of Chemical Information and Modeling, 51(9), 2320-2335.

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q^2 parameter for QSAR validation. Journal of Chemical Information and Modeling, 49(7), 1669-1678.

Crowe, J.E., Lynch, M.F., Town, W.G., 1970. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part I. Non-cyclic fragments. Journal of Chemometrics, 18, 990-997.

Cruz-Monteagudo, M., Borges, F., Perez Gonzalez, M., Soeiro Cordeiro, N.D., 2007. Computational modeling tools for the design of potent antimalarial bis-benzamidines: Overcoming the antimalarial potential of pentamidine. Bioorganic and Medicinal Chemistry, 15(15), 5322-5339.

Dearden, J.C., Cronin, M.T.D., Kaiser, K.L.E., 2009. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR and QSAR in Environmental Research, 20, 241-266.

de Laat, J., Merlet, N., Dore´, M., 1982. Chlorination of organic compounds: Chlorine demand and reactivity in relationship to the trihalomethane formation. Water Research, 16, 1437-1450.

Dickenson, E. V., Summers, S., Croue´, J. P., Gallard, A., 2008. Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic-dicarbonyl acid model compounds. Environmental Science and Technology, 42(9), 3226-3233.

World Health Organization Geneva, 2000. Disinfectants and disinfectant by-products, http://whqlibdoc.who.int/ehc/WHO_EHC_216.pdf. Date accessed December 2014.

DRAGON for Windows 6.0.38, Talete Inc., Milan, 2014. http://www.talete.mi.it/help/dproperties_help/index.html?molecular_properties.htm. Date accessed June 2017.

Du, Y., Lv, X.T., Wu, Q.Y., Zhang, D.Y., Zhou, Y.T., Peng, L., Hu, H.Y., 2017. Formation and control of disinfection by-products and toxicity during reclaimed water chlorination: a review. Journal of Environmental Sciences, 58, 51-63.

Environment Directorate Joint Meeting of the Chemicals Committee and the Working Party on Chemicals. Pesticides and Biotechnology. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships Models. OECD Principles, 2007. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2. Date accessed January 2015.

Environmental Protection Agency Home Page, Alternative Disinfectants and Oxidants Guidance Manual, 1999. http://www.epa.gov/ogwdw/mdbp/alternative_disinfectants_guidance.pdf. Date accessed March 2015.

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

Fabris, R., Chow, C.W.K., Drikas, M., Eikebrokk, B., 2008. Comparison of NOM character in selected Australian and Norwegian drinking waters. Water Research, 42, 4188-4196.

Galal-Gorchev, H., 1996. Chlorine in water disinfection. Pure and Applied Chemistry, 68, 1731-1735.

Gallard, H., Von Gunten, U., 2002. Chlorination of natural organic matter: kinetics of chlorination and of THM formation. Water Research, 36, 65-74.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Publishing Company Inc., U.S.A.

Golfinopoulos, S., Arhonditsis, G., 2002. Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. Chemosphere, 47(9), 1007-1018.

Gopal, K., Tripathy, S.S., Bersillon, J. L., Dubey, S.P., 2007. Chlorination by-products, their toxicodynamics and removal from drinking water. Journal of Hazardous Materials, 140, 1-6.

Gramatica, P., Todeschini, R., 1997. Sd-modelling and prediction by whim descriptors. Part 5. Theory development and chemical meaning of whim descriptors. Journal of Quantitative Structure-Activity Relationships, 16, 113-119

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR and Combinatorial Science, 26(5), 694-701.

Gramatica, P., Kovarich, S., Papa, E., 2009. Development, validation and inspection of the applicability domain of QSPR models for physicochemical properties of polybrominated diphenyl ethers. QSAR and Combinatorial Science, 28(8), 790-796.

Gramatica, P., Chirico, N., 2012. Real external predictivity of QSAR Models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. Journal of Chemical Information and Modeling, 52(8), 2044-2058.

Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. Journal of Computational Chemistry, 34, 2121-2132.

Gramatica, P., Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'min, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha, A., 2014. QSAR modeling: where have you been? Where are you going to?. Journal of Medicinal Chemistry, 57(12), 4977-5010.

Gramatica, P., Shao, Y., Liu, J., Wang, M., Shi, L., Yao, X., 2014. Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modelling methods. Atmospheric Environment, 88, 212-218.

Gramatica, P., 2014. External evaluation of QSAR models, in addition to cross-validation: verification of predictive capability on totally new chemicals. Molecular Informatics, 33(4), 311-314.

Gramatica, P., Sangion, A., 2016. Hazard of pharmaceuticals for aquatic environment: prioritization by structural approaches and prediction of ecotoxicity. Environment International, 95, 131-143.

Hird, C.M., Urbina, M.A., Lewis, C.N., Snape, J.R., Galloway, T.S., 2016. Fluoxetine exhibits pharmacological effects and trait-based sensitivity in a marine worm. Environmental Science and Technology, 50(15), 8344-8352

Hong, H.C., Wong, M.H., Liang, Y., 2009. Amino acids as precursors of trihalomethane and haloacetic acid formation during chlorination. Archives of Environmental Contamination and Toxicology, 56(4), 638-645.

Hu, J., Chu, W., Sui, M., Xu, B., Gao, N., Ding, S., 2018. Comparison of drinking water treatment processes combinations for the minimization of subsequent disinfection by-products formation during chlorination and chloramination. Chemical Engineering Journal, 335, 352-361.

Hureiki, L., Croue, J.P., Legube, B., 1994. Chlorination studies of free and combined amino acids. Water Research, 28, 2521–2531.

Jones, O.A., Lester, J.N., Voulvoulis, N., 2005. Pharmaceuticals: a threat to drinking water? Trends in Biotechnology, 23(4), 163-167.

Karelson, M., 2000. Molecular Descriptors in QSAR/QSPR. John Wiley & Sons, Inc., U.S.A.

Katritzky, A., Lobanov, S.V., Karelson, M., 1995. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews, 4, 279-287.

Klatte, S., Schaefer, H.C., Hempel, M., 2017. Pharmaceuticals in the environment- a short review on options to minimize the exposure of humans, animals and ecosystems. Sustainable Chemistry and Pharmacology, 5, 61-66.

Krasner, S.W., Weinberg, H.S., Richardson, S.D., Pastor, S.J., Chinn, R., Sclimenti, M.J., Onstad, G.D., Thruston Jr., A.D., 2006. Occurrence of a new generation of disinfection by-products. Environmental Science and Technology, 40(23), 7175-7185.

Kulkarni, P., Chellam, S., 2010. Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with treatment. Science of the Total Environment, 408, 4202-4210.

Kumar, J.K., Pandit, A.B., 2012. Drinking water disinfection techniques, 1-9, CRC Press Inc., U.S.A.

Lekkas, T.D., Nikolaou, A.D., 2004. Development of predictive models for the formation of trihalomethanes and haloacetic acids during chlorination of bromide-rich water. Water Quality Research Journal of Canada, 39, 149-159.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate the reproducibility. Biometrics, 45(1), 255-268.

Lin, L.I., 1992. Assay validation using the concordance correlation coefficient. Biometrics, 48, 599-604.

Lipnick, L., R., 1985. A perspective on quantitative structure-activity relationships in ecotoxicology. Environmental Toxicology and Chemistry, 4, 255-257.

Liu, H., Papa, E., Gramatica, P., 2008. Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. Chemosphere, 70, 1889-1897.

LoPachin, R.M., Gavin, T., Geohagen, B.C., Das, S., 2007. Neurotoxic mechanisms of electrophilic type-2 alkenes: soft–soft interactions described by quantum mechanical parameters. Toxicological Sciences, 98(2), 561-570.

Luilo, G.B., Cabaniss, S.E., 2010. QSPR for predicting chlorine demand by organic molecules. Environmental Science and Technology, 44(7), 2503-2508.

Luilo, G.B., 2011. Quantitative structure- property relationships for predicting chlorine demand and disinfection by-products formation in drinking water, Ph.D. Thesis, The University of New Mexico, U.S.A.

Luilo, G.B., Cabaniss, S.E., 2011. QSPR for predicting chloroform formation in drinking water disinfection. SAR and QSAR in Environmental Research, 22, 489-504.

Madjarova, G., Tadjer, A., Cholakova, T. P., Dobrev, A.A., and Mineva, T., 2005. Selectivity descriptors for the Michael addition reaction as obtained from density functional based approaches. Journal of Physical Chemistry A, 109(2), 387-393.

Mallakin, A., Mezey, P.G., Zimpel, Z., Berenhaut, K.S., Greenberg, B.M., Dixon, D. G., 2005. Use of quantitative structure-activity relationship to model the photoinduced toxicity of anthracene and oxygenated anthracenes. QSAR and Combinatorial Science, 24, 844-852.

Maykel Cruz-Monteagudo, M., Borges, F., Gonza´lez, M.P., Nata´lia Dias Soeiro Cordeiro, M. 2007. Computational modeling tools for the design of potent antimalarial bisbenzamidines: overcoming the antimalarial potential of pentamidine. Bioorganic and Medicinal Chemistry, 15, 5322-5339.

Mitra, I., Roy, P.P., Kar, S., Ojha, P.K., Roy, K., 2010. On further application of r_m^2 as a metric for validation of QSAR models. Journal of Chemometrics, 24, 22-33.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., Prachayasittikul, V., 2009. A practical overview of QSAR. EXCLI Journal, 8, 74-88.

Environmental Protection Agency Home Page. http://water.epa.gov/drink/contaminants/basicinformation/disinfectionbyproducts.cfm. Date accessed December 2014.

Netzeva, T.I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D., Schultz, T., Stanton, D.W., van de Sandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Alternatives to Laboratory Animals, 33, 155-173.

Norwood, D.L., Johnson, J.D., Christman, R.F., Hass, J.R., Bobenrieth, M.J., 1980. Reactions of chlorine with selected aromatic models of aquatic humic material. Environmental Science and Technology, 14, 187-189.

Obolensky, A., Singer, P.C., 2008. Development and interpretation of disinfection by-product formation models using the information collection rule database. Environmental Science and Technology, 42, 5654-5660.

Oğuz, M., Mihçiokur, H., 2014. Environmental risk assessment of selected pharmaceuticals in Turkey. Environmental Toxicology and Pharmacology, 38, 79-83.

Ojha, P.K., Mitra, L., Das, R.N., Roy, K., 2011. Further exploring r^2_m metrics for validation of QSPR models. Chemometrics and Intelligent Laboratory Systems, 107 (1), 194-205.

Organization for Economic Co-operation and Development, 2004. Guidance document on the validation of (quantitative) structure–activity relationships [(Q)SAR] models, Final Report, Paris, France.

Papa, E., Villa, F., Gramatica, P., 2005. Statistically validated QSARs, based on theoretical descriptors, for modeling aquaitic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). Journal of Chemical Information and Modeling, 45(5), 1256-1266.

Pearson, R.G., 1990. Hard and soft acids and bases - the evolution of a chemical concept. Coordination Chemistry Reviews, 100, 403-425.

Pirovano, A., Brandmaier, S., Huijbregtsa, M.A.J., Ragas, A.M.J., Veltmana, K., Hendriks, A.J. 2015. The utilization of structural descriptors to predict metabolic constants of xenobiotics in mammals. Environmental Toxicology and Pharmacology, 39(1), 247-258.

Prachayasittikul, V., Pingaew, R., Worachartcheewan, A., Nantasenamat, C., Prachayasittikul, S., Ruchirawat, S., Prachayasittikul, V., 2014. Synthesis, anticancer activity and QSAR study of 1,4-naphthoquinone derivatives. European Journal of Medicinal Chemistry, 84, 247-263.

Reckhow, D.A., Singer, P.C., Malcolm, R.L., 1990. Chlorination of humic materials: by-product formation and chemical interpretations. Environmental Science and Technology, 24, 1655-1664.

Richardson, S.D., 2002. The role of GC-MS and LC-MS in the discovery of drinking water disinfection by-products. Journal of Environmental Monitoring, 4, 1-9.

Richardson, S.D., Fasano, F., Ellington, J.J., Crumley, F.G., Buettner, K.M., Evans. J.J., Blount, B.C., Silva, L.K., Waite, T.J., Luther, G.W., McKague, A.B., Miltner, R.J., Wagner, E.D., Plewa, M.J., 2008. Occurrence and mammalian cell toxicity of iodinated disinfection byproducts in drinking water. Environmental Science and Technology, 42, 8330-8338.

Roy, K., Roy P., Leonard, J. 2008. On some aspects of validation of predictive QSAR models. Chemistry Central Journal, 2, 11-13.

Roy, K., 2015. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems, 145, 22-29.

Roy, K., Kar, S., Das, R.N., 2015. A primer on QSAR/QSPR modeling, in statistical methods in QSAR/QSPR, New York, Springer, 35-62.

Roy, K., Das, R.D., Ambure, P., Aher. R.B., 2016. Be aware of error measures: further studies on validation of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems, 152, 18-33.

Sadiq, R., Rodriquez, M.J., 2004. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. Science of the Total Environment, 321, 21-46.

Schüürmann, G., Ebert, R., Chen, J., Wang, B., Kühne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient - test set activity mean vs training set activity mean. Journal of Chemical Information and Modeling, 48, 2140-2145.

Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C. L., Sheehan, D.M., 2001. QSAR models using a large diverse set of estrogens. Journal of Chemical Information and Computer Sciences, 41(1), 186-195.

Shimazu, H., Kouch, M., Yonekura, Y., Kumano, H., Hashiwata, K., Hirota, T., Ozaki, N., Fukushima, H., 2005. Developing a model for disinfection byproducts based on multiple regression analysis in a water distribution system. Journal of Water Supply Research and Technology, 54(4), 225-237.

Schultz, T.W., Cronin, M.T.D., Walker, J.D., Aptula. A.O., 2003. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. Journal of Molecular Structure, 622, 1-22.

Singer, P.C., Archer, A.D., 2006. An evaluation of the relationship between SUVA and NOM coagulation using ICR database. Journal of American Water Works Association, 98 (7), 110-123.

SPSS Statistics 17.0 for Windows, 2008. SPSS Inc., USA.

SPARTAN v. 10, Wavefunction, Inc., 2010, Irvine, USA, http://wavefun.com.

Todeschini, R., Consonni, V., Mauri, A., Pavan, M., 2004. Detecting “bad” regression models: multicriteria fitness functions in regression analysis. Analytica Chimica Acta, 515(1), 199-208.

Todeschini, R., Consonni, V., 2009. Molecular descriptors for cheminformatics, Second Ed., WILEY-VCH, Berlin, Germany.

Topliss, J. G., Costello, R. J., 1972. Chance correlations in structure–activity studies using multiple regression analysis. Journal of Medicinal Chemistry, 15, 1066-1068.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Combinatorial Science, 22, 69-77.

Türkdoğan, F. I., Yetilmezsoy, K., 2009. Appraisal of potential environmental risks associated with human antibiotic consumption in Turkey. Journal of Hazardous Materials, 166, 297-308.

Rücker. C., Rücker. G., Meringer. M., Y-randomization – A useful tool in QSAR validation or Folklore?, http://www.mathe2.uni-bayreuth.de/markus/pdf/pub/YRandQsar.pdf. Date accessed January 2015.

Yang, Y., Ok, Y.S., Kim, K.H., Kwon, E.E., Tsang, Y.F., 2017. Occurrences and removal of pharmaceuticals and personal care products (PPCPs) in drinking water and water/sewage treatment plants: a review. Science of the Total Environment, 596(597), 303-320.

Yang, C., Wang, J., 2018. On the intrinsic dynamics of bacteria in waterborne infections. Mathematical Biosciences, 296, 71-81.

Yousefinejad, S., Hemmateenejad, B., 2015. Chemometrics tools in QSAR/QSPR studies: a historical perspective. Chemometrics and Intelligent Laboratory Systems, 149, 177-204.

Yousefinejad, S., Eftekhari, R.B., Honarasa, F., Zamanian, Z., Sedaghati, F., 2017. Comparison between the gas-liquid solubility of methanol and ethanol in different organic phases using structural properties of solvents. Journal of Molecular Liquids, 241, 861-869.

Walker, J.D., 2003. QSARs promote more efficient use of chemical testing resources. Environmental Toxicology and Chemistry, 22, 1651-1652.

Wang, F., Ruan, M., Lin, H., Zhang, Y., Hong, H., Zhou, X., 2014. Effects of ozone pretreatment on the formation of disinfection by-products and its associated bromine substitution factors upon chlorination/chloramination of Tai Lake water. Science of The Total Environment, 475, 23-28.

Wei, Q., Feng, C., Wang, D., Shi, B., Zhang, L., Tang, H., 2008. Seasonal variations of chemical and physical characteristics of dissolved organic matter and trihalomethane precursors in a reservoir: a case study. Journal of Hazardous Materials, 150, 257–264.

Woo, Y.T., Lai, D., McLain, J.L., Manibusan, M.K., Dellarco, V., 2002. Use of mechanism-based structure–activity relationships analysis in carcinogenic potential ranking for drinking water disinfection by-products. Environmental Health Perspectives, 110, 75-87.

World Health Organization Home Page. http://www.who.int/whr/2002/chapter4/en/index7.html. Date accessed December 2017.

World Health Organization Home Page. http://www.who.int/mediacentre/factsheets/fs194/en/. Date accessed November 2017.

Zhang, X., Li, W., Blatchley, E.R., Wang, X., Ren, P., 2015. UV/chlorine process for ammonia removal and disinfection by-product reduction: comparison with chlorination. Water Research, 68, 804-811.

APPENDIX A: INFORMATION ON THE TRAINING-TEST SET DIVISIONS OF THE MODELS

Table A.1. The test set divisions of the developed models with their chlorine demand (molHOCl/mol compound) values.

No	CAS	Compound	Exp. Endpoint HOCldem (mol HOCl/mol compound)	M1	M2	M3	M4
1	108-46-3	1,3-Dihydroxybenzene	7.4	7.4	7.4	7.4	7.4
2	620-02-0	5-Methylfurfural	0.8	0.8	0.8	0.8	0.8
3	576-24-9	2,3-Dichlorophenol	8	**8**	8	8	**8**
4	933-75-5	2,3,6-Trichlorophenol	6.9	6.9	**6.9**	**6.9**	6.9
5	591-35-5	3,5-Dichlorophenol	7.6	**7.6**	7.6	7.6	**7.6**
6	88-06-2	2,4,6-Trichlorophenol	8.02	8.02	8.02	8.02	8.02
7	120-83-2	2,4-Dichlorophenol	8.1	8.1	8.1	8.1	8.1
8	540-38-5	4-Iodophenol	12.5	**12.5**	12.5	12.5	12.5
9	106-48-9	4-Chlorophenol	9.25	9.25	9.25	9.25	9.25
10	100-02-7	4-Nitrophenol	7.9	7.9	7.9	7.9	7.9
11	106-44-5	4-Hydroxytoluene	6.33	**6.33**	6.33	**6.33**	**6.33**
12	58-90-2	2,3,4,6-Tetrachlorophenol	7.2	**7.2**	7.2	7.2	**7.2**
13	39727-75-8	3,4,5-Triethoxybenzyl alcohol	0.55	0.55	0.55	0.55	0.55
14	99-05-8	3-Aminobenzoic acid	7.74	7.74	7.74	**7.74**	7.74
15	99-96-7	4-Hydroxybenzoic acid	9.44	9.44	9.44	9.44	9.44
16	458-35-5	Coniferyl alcohol	6.66	6.66	6.66	6.66	6.66
17	884-35-5	Methylsyringate	7.11	7.11	**7.11**	**7.11**	7.11
18	87-66-1	1,2,3-Trihydroxybenzene	6.9	6.9	6.9	**6.9**	6.9

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
19	120-80-9	1,2-Dihydroxybenzene	4.1	4.1	4.1	4.1	4.1
20	123-31-9	1,4-Dihydroxybenzene	3.3	3.3	3.3	3.3	3.3
21	118-93-4	2-Hydroxyacetophenone	9.9	9.9	9.9	9.9	9.9
22	90-02-8	2-Hydroxybenzaldehyde	9.7	9.7	9.7	9.7	9.7
23	88-75-5	2-Nitrophenol	9.6	9.6	9.6	9.6	9.6
24	121-71-1	3-Hydroxyacetophenone	11	11	11	**11**	11
25	108-39-4	3-Hydroxytoluene	8.7	8.7	8.7	8.7	8.7
26	137-19-9	4,6-Dichloro-1,3-dihydroxybenzene	5	5	5	**5**	5
27	95-88-5	4-Chloro-1,3-dihydroxybenzene	6.1	6.1	6.1	**6.1**	6.1
28	150-76-5	4-Methoxyphenol	3.4	**3.4**	**3.4**	3.4	**3.4**
29	98-86-2	Acetophenone	0.5	0.5	0.5	0.5	0.5
30	55-21-0	Benzamide	2.5	2.5	2.5	**2.5**	2.5
31	103-82-2	Phenylacetic acid	0.1	0.1	0.1	0.1	0.1
32	127-17-3	Pyruvic acid	1	1	1	1	1
33	57-13-6	Urea	3.8	3.8	3.8	3.8	3.8
34	141-97-9	Ethylaceto acetate	2	2	**2**	**2**	2
35	90-50-6	3-(3,4,5-Trimethoxyphenyl) propanoic acid	1.32	1.32	1.32	1.32	1.32
36	132-86-5	1,3-Dihydroxynaphthalene	5.1	5.1	5.1	5.1	5.1
37	504-15-4	3,5-Dihydroxytoluene	6.39	6.39	6.39	6.39	6.39
38	118-41-2	3,4,5-Trimethoxybenzoic acid	1.1	1.1	1.1	1.1	1.1
39	99-93-4	4-Hydroxyacetophenone	9.37	9.37	9.37	**9.37**	9.37
40	123-08-0	4-Hydroxybenzaldehyde	8.96	8.96	8.96	8.96	8.96

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
41	498-02-2	Acetovanillone	8.68	8.68	8.68	8.68	8.68
42	121-34-6	Vanillic acid	7.75	**7.75**	**7.75**	**7.75**	**7.75**
43	108-43-0	3-chlorophenol	9.15	9.15	9.15	**9.15**	9.15
44	123-72-8	Butanal	0.2	0.2	0.2	0.2	0.2
45	108-95-2	Phenol	9.6	9.6	9.6	**9.6**	9.6
46	1135-24-6	Ferulic acid	10.99	10.99	10.99	10.99	10.99
47	95-57-8	2-Chlorophenol	9.15	9.15	9.15	9.15	9.15
48	95-55-6	2-Aminophenol	4.02	4.02	4.02	4.02	4.02
49	135-19-3	2-Napthol	4.4	4.4	4.4	4.4	**4.4**
50	108-73-6	1,3,5-Trihydroxybenzene	9.1	9.1	9.1	**9.1**	9.1
51	56-84-8	L-Aspartic acid	5.8	5.8	5.8	5.8	5.8
52	150-19-6	3-Methoxyphenol	8.1	8.1	**8.1**	8.1	8.1
53	99-09-2	3-Nitroaniline	8.5	8.5	8.5	**8.5**	8.5
54	121-92-6	3-Nitrobenzoic acid	0.1	0.1	0.1	0.1	0.1
55	554-84-7	3-Nitrophenol	9.2	**9.2**	9.2	9.2	**9.2**
56	92-88-6	4,4'-Dihydroxy biphenyl	10.5	10.5	10.5	10.5	10.5
57	123-30-8	4-Aminophenol	5.4	5.4	5.4	5.4	5.4
58	88-04-0	4-Chloro-3,5-dimethylphenol	4.7	4.7	4.7	4.7	4.7
59	74-11-3	4-Chlorobenzoic acid	0.1	0.1	0.1	**0.1**	0.1
60	62-55-5	Acetothioamide	4.2	4.2	**4.2**	**4.2**	4.2
61	64-19-7	Acetic acid	0.1	0.1	0.1	0.1	0.1
62	67-64-1	Acetone	0.1	0.1	**0.1**	0.1	0.1

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
63	123-54-6	Acetyl acetone	4	4	4	4	4
64	100-66-3	Anisole	1	1	1	**1**	1
65	100-52-7	Benzaldehyde	0.1	0.1	**0.1**	**0.1**	0.1
66	65-85-0	Benzoic acid	0.3	**0.3**	0.3	0.3	**0.3**
67	77-92-9	Citric acid	0.63	0.63	**0.63**	0.63	0.63
68	64-17-5	Ethanol	0.1	0.1	0.1	0.1	0.1
69	110-17-8	Fumaric acid	0.1	0.1	0.1	0.1	0.1
70	110-16-7	Maleic acid	0.1	0.1	**0.1**	**0.1**	0.1
71	141-82-2	Malonic acid	1.8	**1.8**	**1.8**	1.8	**1.8**
72	98-95-3	Nitrobenzene	0.1	0.1	0.1	0.1	0.1
73	91-66-7	N, N-DIETHYLANILINE	8.3	8.3	8.3	8.3	8.3
74	123-38-6	Propanal	0.2	0.2	0.2	0.2	0.2
75	103-85-5	Phenylthiourea	12.4	12.4	12.4	12.4	12.4
76	122-59-8	Phenoxyacetic acid	0.3	0.3	0.3	**0.3**	0.3
77	144-62-7	Oxalic acid	0.3	0.3	0.3	0.3	0.3
78	110-15-6	Succinic acid	0.1	0.1	0.1	**0.1**	0.1
79	62-56-6	Thiourea	3.9	3.9	3.9	3.9	3.9
80	1132-21-4	3,5-Dimethoxybenzoic acid	3	**3**	**3**	3	**3**
81	530-59-6	Sinapic acid	8.77	8.77	8.77	8.77	8.77
82	121-33-5	Vanillin	7.92	7.92	7.92	7.92	7.92
83	537-33-7	Sinapyl alcohol	6.14	6.14	6.14	6.14	6.14
84	600-18-0	2-Oxobutyric acid	1.1	1.1	1.1	1.1	1.1

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
85	689-31-6	3-Oxohexanedioic acid	5.8	5.8	5.8	**5.8**	5.8
86	498-23-7	Citraconic acid	0.1	0.1	0.1	**0.1**	0.1
87	501-98-4	P-Coumaric acid	9.29	9.29	**9.29**	9.29	9.29
88	328-50-7	2-Oxopentanedioic acid	1.4	1.4	1.4	1.4	1.4
89	300-85-6	3-Hydroxybutyric acid	1.2	1.2	1.2	**1.2**	1.2
90	542-05-2	3-Oxopentanedioic acid	5.3	5.3	5.3	5.3	5.3
91	533-73-3	1,2,4-Trihydroxybenzene	3.9	3.9	3.9	3.9	3.9
92	69-72-7	2-Hydroxybenzoic acid	6	6	**6**	6	6
93	95-48-7	2-Hydroxytoluene	7.5	**7.5**	7.5	7.5	**7.5**
94	90-05-1	2-Methoxyphenol	7.7	7.7	**7.7**	7.7	7.7
95	298-12-4	2-Oxoacetic acid	1.1	1.1	1.1	1.1	1.1
96	3458-28-4	d-Mannose	1.3	1.3	1.3	1.3	1.3
97	591-27-5	3-Aminophenol	7.7	7.7	7.7	7.7	7.7
98	100-83-4	3-Hydroxybenzaldehyde	9.8	9.8	9.8	**9.8**	9.8
99	99-06-9	3-Hydroxybenzoic acid	9.1	9.1	9.1	9.1	9.1
100	56-86-0	L-Glutamic acid	3.05	3.05	3.05	3.05	3.05
101	3840-31-1	3,4,5-Trimethoxybenzyl alcohol	1.52	**1.52**	**1.52**	1.52	**1.52**
102	1135-23-5	3-(4-Hydroxy-3-methoxyphenyl) propionic acid	4.94	4.94	4.94	4.94	4.94
103	56-41-7	L-Alanine	3.33	3.33	3.33	3.33	3.33
104	73-32-5	L-Isoleucine	3.3	**3.3**	**3.3**	3.3	**3.3**
105	147-85-3	L-Proline	5.5	**5.5**	5.5	5.5	**5.5**
106	72-18-4	L-Valine	3.43	3.43	3.43	3.43	**3.43**

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
107	56-45-1	L-Serine	4.9	4.9	4.9	4.9	4.9
108	60-18-4	L-Tyrosine	12.67	12.67	12.67	**12.67**	12.67
109	74-79-3	L-Arginine	8.55	8.55	8.55	8.55	8.55
110	61-90-5	L-Leucine	3.3	**3.3**	3.3	3.3	**3.3**
111	52-90-4	L-Cysteine	7.87	7.87	7.87	7.87	7.87
112	56-87-1	L-Lysine	4.83	**4.83**	4.83	4.83	**4.83**
113	99-10-5	3,5-Dihydroxybenzoic acid	7.08	7.08	**7.08**	7.08	7.08
114	89-86-1	2,4-Dihydroxybenzoic acid	7.5	7.5	**7.5**	7.5	7.5
115	150-13-0	4-Aminobenzoic acid	7.9	7.9	**7.9**	7.9	7.9
116	767-00-0	4-Cyanophenol	11	**11**	**11**	11	**11**
117	106-50-3	1,4-Phenyldiamine	3.53	**3.53**	3.53	3.53	**3.53**
118	118-92-3	2-Aminobenzoic acid	6.04	6.04	6.04	6.04	6.04
119	625-45-6	Methoxyacetic acid	0.79	0.79	**0.79**	0.79	0.79
120	70-47-3	L-Asparagine	5.21	5.21	**5.21**	5.21	5.21
121	72-19-5	L-Threonine	5.7	5.7	**5.7**	5.7	5.7
122	63-68-3	L-methionine	6.35	6.35	6.35	6.35	6.35
123	62-53-3	Aniline	8.44	8.44	8.44	8.44	8.44
124	63-91-2	L-Phenylalanine	3.28	3.28	3.28	3.28	3.28
125	90-15-3	1-Napthol	7.2	7.2	**7.2**	7.2	7.2
126	56-40-6	L-Glycine	4.5	**4.5**	4.5	4.5	**4.5**
127	6970-19-0	3,4,5-Triethoxybenzoic acid	0.99	0.99	0.99	0.99	0.99
128	1136-86-3	3,4,5-Trimethoxyacetophenone	0.93	0.93	0.93	0.93	0.93

Table A.1. Continued.

No	CAS	Compound	Exp. Endpoint HOCldem (molHOCl/mol compound)	M1	M2	M3	M4
129	3086-62-2	3,4,5-Trimethoxybenzamide	5.33	**5.33**	5.33	5.33	**5.33**
130	13338-63-1	3,4,5-Trimethoxyphenylacetonitrile	4.38	4.38	**4.38**	4.38	4.38
131	951-82-6	3,4,5-Trimethoxyphenylacetic acid	0.84	**0.84**	**0.84**	0.84	**0.84**
132	6627-88-9	4-Allyl-2,6-dimethoxyphenol	6.8	**6.8**	6.8	6.8	**6.8**
133	5328-37-0	L-Arabinose	0.4	0.4	0.4	0.4	0.4
134	7.05.6638	4-methyl-2,6-dimethoxyphenol	4.87	**4.87**	4.87	4.87	**4.87**
135	2478-38-8	Acetosyringone	7.79	7.79	**7.79**	7.79	7.79
136	134-96-3	Syringaldehyde	8.38	**8.38**	8.38	8.38	**8.38**
137	530-57-4	Syringic acid	7.46	7.46	7.46	7.46	7.46
138	328-42-7	3-Oxobutanedioic acid	3.8	**3.8**	3.8	3.8	**3.8**
139	51568-18-4	4,6-Dioxoheptanoic acid	4.8	4.8	**4.8**	4.8	4.8
140	502-50-1	4-Oxoheptanedioic acid	1.25	**1.25**	1.25	1.25	**1.25**
141	51568-19-5	5,7-Dioxooctanoic acid	6	**6**	6	6	6
142	56-85-9	L-Glutamine	3.8	3.8	3.8	3.8	3.8
143	2227-79-4	Benzothioamide	4	4	4	4	4
144	97-67-6	L-Malic acid	0.75	**0.75**	0.75	0.75	**0.75**
145	34759-04-1	4-(3,4,5-trimethoxybenzoyl) Butyric acid	2.01	2.01	2.01	2.01	2.01
146	4206-58-0	Trans-3,5-dimethoxy-4-hydroxycinnam aldehyde	9.59	9.59	9.59	9.59	9.59
147	70-26-8	L-Ornithine	4.6	4.6	4.6	4.6	4.6
148	3044-56-2	Ethyl-(3,4,5-trimethoxybenzyl) acetate	4.5	4.5	4.5	4.5	4.5
149	14897-78-0	3-(4-hydroxy-3,5-dimethoxyphenyl)Propanoic acid	6.06	6.06	6.06	6.06	**6.06**

150	131-48-6	N-Acetylneuraminic acid	2.9	2.9	2.9	2.9	2.9

*bold numbers indicate test set chemicals in each division.

APPENDIX B: INFORMATION ON THE PREDICTED CHLORINE DEMAND VALUES OF EXTERNAL SET CHEMICALS

Table B.1. Predicted chlorine demand values of external set chemicals from model Eq.4.1. along with hat and descriptor values.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Antibiotics										
28657-80-9	Cinoxacin	0.895	0.040	0.456	0.011	1	0.939	0	10.0674	0.1217
85721-33-1	***Ciprofloxacin***	0.823	0.029	0.529	0.010	1	**0.984**	0	8.9395	0.089
112398-08-0	***Danofloxacin***	0.854	0.031	0.554	0.011	1	**0.989**	0	9.2105	0.1016
98106-17-3	***Difloxacin***	0.798	-0.014	0.504	0.008	1	**0.983**	0	8.8322	0.0842
74011-58-8	Enoxacin	0.803	0.042	0.523	0.012	1	0.951	0	8.2821	0.0814
93106-60-6	***Enrofloxacin***	0.799	0.037	0.542	0.011	1	**0.968**	0	8.2166	0.0811
79660-72-3	Fleroxacin	0.809	-0.016	0.538	0.013	1	0.947	0	8.34	0.0858
100986-85-4	Levofloxacin	0.845	0.010	0.501	0.012	1	0.961	0	9.414	0.0992
98079-51-7	Lomefloxacine	0.827	0.011	0.547	0.011	1	0.943	0	8.1019	0.074
70458-96-7	Norfloxacin	0.800	0.025	0.529	0.012	1	0.956	0	8.3048	0.0824
82419-36-1	Ofloxacin	0.845	0.010	0.501	0.012	1	0.961	0	9.414	0.0992
127045-41-4	Pazufloxacin	0.906	-0.055	0.471	0.009	1	0.964	0	10.5471	0.1252
51940-44-4	Pipemidic Acid	0.796	0.041	0.540	0.011	1	0.946	0	7.7636	0.072
151096-09-2	***Moxifloxacin***	0.841	0.047	0.543	0.010	1	**0.975**	0	8.7668	0.086
98105-99-8	***Sarafloxacin***	0.800	-0.011	0.485	0.011	1	**0.994**	0	9.6292	0.1115
111542-93-9	Sparfloxacin	0.822	0.046	0.536	0.009	1	0.962	0	8.2547	0.0728

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Antibiotics										
64544-07-6	Cefuroxime axetil	0.933	-0.180	0.520	0.010	1	0.903	0	9.4934	0.0971
56-75-7	Chloramphenicol	0.875	0.073	0.543	0.019	1	0.824	0	7.1279	0.1267
79660-72-3	Fleroxacin	0.809	-0.016	0.510	0.012	1	0.947	0	8.6123	0.0845
54-85-3	Isoniazid	0.856	0.130	0.565	0.000	1	0.928	0	6.502	0.0672
98079-51-7	Lomefloxacin	0.827	0.011	0.561	0.012	1	0.943	0	8.0173	0.0805
67-20-9	Nitrofurantoin	0.942	-0.144	0.648	0.011	1	0.920	0	8.2361	0.1446
59-87-0	Nitrofurazone	0.931	-0.082	0.601	0.014	1	0.888	0	8.2395	0.1149
60-54-8	Tetracycline	0.881	-0.100	0.495	0.008	1	0.873	1	10.8677	0.0771
Anticoagulant										
5543-58-8	R(+)- Warfarin	0.763	-0.131	0.467	0.011	0	0.948	1	9.8679	0.0916
Antidepressant										
50-48-6	***Amitriptyline***	0.679	-0.083	0.408	0.013	1	**0.993**	0	9.436	0.1867
54910-89-3	Fluoxetine	0.733	0.061	0.495	0.011	1	0.950	0	7.5566	0.0831
61869-08-7	***Paroxetine***	0.818	-0.007	0.460	0.010	1	**0.986**	0	9.933	0.1207
76-75-5	Thiopental	0.734	-0.004	0.429	0.015	1	0.861	0	7.3025	0.1389
Anti-epileptic										
50-06-6	Phenobarbital	0.765	-0.105	0.466	0.015	1	0.926	0	8.7551	0.121
1622-61-3	**Clonazepam**	0.827	-0.027	0.449	**0.021**	1	**0.992**	0	11.5033	0.2308
Anti-hyperlipidemic										
637-07-0	Clofibrate	0.710	0.013	0.446	0.018	0	0.868	0	5.4901	0.1405
882-09-7	***Clofibric acid***	0.746	-0.011	0.512	**0.021**	0	0.848	0	5.0651	0.1339

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Anti-inflammatory										
50-78-2	Acetylsalicylic acid	0.840	-0.226	0.453	0.003	0	0.857	0	5.9955	0.0761
15687-27-1	S(+)- Ibuprofen	0.708	-0.031	0.602	0.012	0	0.863	0	2.7741	0.0716
22204-53-1	S(+)- Naproxen	0.811	-0.073	0.562	0.008	0	0.917	0	5.4291	0.0592
103-90-2	Paracetamol	0.806	-0.303	0.539	0.006	1	0.876	1	9.6773	0.065
Anti-psychotic										
110-89-4	Piperidine	0.604	-0.108	0.534	0.000	1	0.906	0	3.7824	0.0762
106266-06-2	***Risperidone***	0.791	-0.006	0.508	0.007	1	**1.003**	0	8.9519	0.0911
Anti-viral										
59277-89-3	Acyclovir	0.907	-0.045	0.573	0.014	1	0.914	0	8.7057	0.1041
104227-87-4	Famciclovir	0.826	-0.192	0.478	0.018	1	0.911	0	9.6507	0.1515
Beta blocker										
72956-09-3	Carvedilol	0.744	0.076	0.424	0.010	1	0.964	0	8.7843	0.1311
525-66-6	S(+)-Propranolol	0.762	0.029	0.508	0.013	1	0.900	0	7.082	0.0769
22664-55-7	Metipranolol	0.741	-0.082	0.594	0.009	1	0.852	0	4.5977	0.065
Chemotherapeutic agents										
154361-50-9	Capecitabine	0.840	0.064	0.595	0.006	1	0.839	0	4.9445	0.0541
59-05-2	Methotrexate	0.884	-0.116	0.645	0.008	1	0.926	0	7.2411	0.1125
Corticosteroid										
51333-22-3	Budesonide	0.805	0.039	0.530	0.008	0	0.878	0	4.7189	0.0445
Glucocorticoid										
5251-34-3	Cloprednol	0.852	0.047	0.536	0.009	0	0.846	0	4.7184	0.047

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Hormones										
57-63-6	Ethinylestradiol	0.808	0.060	0.485	0.011	0	0.897	1	8.7188	0.0627
57-85-2	Testosterone propionate	0.718	-0.020	0.554	0.007	0	0.915	0	4.0206	0.0484
10161-34-9	Trenbolone acetate	0.837	-0.084	0.489	0.008	0	0.954	0	7.5036	0.1021
Industrial										
107-98-2	1-Methoxy-2-propanol	0.820	-0.102	0.507	0.000	0	0.713	0	1.5566	0.0413
120-78-5	**2,2-Dithiobisbenzothiazole**	0.643	0.105	0.553	**0.020**	1	**1.078**	0	8.895	0.2243
39263-32-6	2-Amino-5-bromobenzonitrile	0.885	-0.033	0.536	0.000	1	0.955	0	8.2361	0.0825
696-23-1	2-Methyl-4(5)-nitroimidazole	0.873	0.038	0.627	0.000	1	0.886	0	5.3215	0.0853
371-40-4	4-Fluoroaniline	0.737	-0.161	0.584	0.000	1	0.957	0	6.0114	0.0709
107-02-8	Acroleine	0.917	0.068	0.625	0.000	0	0.769	0	1.9008	0.0911
5329-14-6	Amidosulfonic acid	0.833	-1.043	0.505	0.000	0	0.786	0	5.6869	0.1707
98-88-4	Benzoyl-chloride	0.655	0.104	0.576	0.000	0	0.923	0	1.9956	0.0815
501-53-1	Benzyl-chloroformiate	0.729	0.049	0.592	0.013	0	0.901	0	3.8182	0.08
100-46-9	Benzylamine	0.678	0.097	0.605	0.000	1	0.947	0	4.0657	0.0962
66-25-1	Capronaldehyde	0.724	-0.016	0.607	0.000	0	0.819	0	0.7841	0.0591
367-21-5	Chlorfluoroaniline	0.873	-0.029	0.598	0.000	1	0.956	0	7.255	0.0919
46755-94-6	Diphenylpropanediol(S)	0.607	0.159	0.440	0.014	0	0.860	0	3.2484	0.1433
1141-88-4	***Dithiodianiline***	0.614	0.035	0.619	0.016	1	**1.009**	0	6.0536	0.1815
106-89-8	Epichlorohydrine	0.880	0.143	0.552	0.000	0	0.861	0	3.9849	0.0666
58-08-2	Caffeine	0.723	-0.012	0.551	0.000	1	0.940	0	5.5395	0.0579

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Industrial										
74-96-4	Ethyl bromide	0.719	-0.192	0.640	0.000	0	0.809	0	0.5491	0.0756
108-31-6	Maleic anhydride	0.720	-0.488	0.629	0.000	0	0.858	0	2.4546	0.0761
105-53-3	Malonic acid diethylester	0.659	-0.255	0.586	0.015	0	0.803	0	2.0738	0.0943
2211-28-1	***Monobenzoate***	0.690	0.020	0.504	0.014	0	**0.974**	1	8.819	0.0949
110-91-8	Morpholine	0.659	-0.081	0.579	0.000	1	0.878	0	3.2897	0.0679
121-69-7	N. N'-dimethylaniline	0.585	-0.148	0.563	0.000	1	0.914	0	3.4046	0.0887
54-11-5	Nicotine (S izomer)	0.847	-0.015	0.480	0.002	1	0.947	0	8.4851	0.0792
57849-23-7	***Octabase H***	0.758	0.007	0.479	0.009	1	**0.975**	0	8.5281	0.0906
149-73-5	**Orthoformic acid trimethylester**	**0.453**	0.322	0.455	0.000	0	0.759	0	-2.8413	0.2597
3282-30-2	Pivaloyl chloride	0.480	0.032	0.421	0.000	0	0.782	0	-0.8065	0.1882
119-64-2	***Tetralin***	0.653	-0.034	0.517	0.000	0	**0.994**	0	4.5133	0.0984
935-92-2	Trimethylquinone	0.596	-0.029	0.521	0.000	0	0.901	0	1.8757	0.0696
Sedatives and Hypnotics										
57-43-2	Amobarbital	0.688	-0.048	0.528	0.013	1	0.859	0	5.2381	0.0849
1622-62-4	***Flunitrazepam***	0.819	-0.015	0.453	0.014	1	**0.981**	0	10.3758	0.148
Sedatives and Hypnotics										
2898-12-6	**Medazepam**	0.772	-0.019	0.471	**0.023**	1	**1.007**	0	10.9571	0.2393
59467-70-8	**Midazolam (base)**	0.847	-0.064	0.459	0.019	1	**1.048**	0	12.6199	0.2489
76-74-4	Pentobarbital	0.688	-0.040	0.463	0.010	1	0.859	0	5.7789	0.086
76-73-3	R(+)-Secobarbital	0.802	-0.057	0.472	0.010	1	0.865	0	7.3314	0.0691
439-14-5	**Diazepam**	0.775	-0.012	0.481	**0.023**	1	**1.003**	0	10.7661	0.2303

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Sedatives and Hypnotics										
78755-81-4	***Flumazenil***	0.902	-0.027	0.483	0.013	1	**0.974**	0	10.866	0.1412
111-42-2	Diethanolamine	0.680	0.023	0.542	0.006	1	0.732	0	1.5552	0.1022
Diuretic										
27589-33-9	**Azosemide**	0.919	-0.137	0.512	**0.020**	1	**1.019**	0	12.5951	0.2409
396-01-0	***Triamterene***	0.756	0.010	0.494	0.009	1	**1.053**	0	9.8151	0.1329
Disinfectant/Preservative										
111-30-8	Glutaraldehyde	0.693	-0.081	0.642	0.000	0	0.802	0	-0.2545	0.0897
3380-34-5	**Triclosan**	0.761	-0.102	0.544	**0.031**	0	0.940	1	10.6364	0.3083
1330-20-7	Xylene (mixed isomers)	0.580	-0.146	0.564	0.000	0	0.923	0	1.8284	0.0769
Gastroprotective										
76824-35-6	Famotidine	0.847	-0.299	0.537	0.013	1	0.954	0	9.7474	0.1189
103577-45-3	***Lansoprazole***	0.868	-0.042	0.558	0.014	1	**0.996**	0	9.9826	0.1305
Other										
7535-00-4	D-Galactosamine[a]	0.838	0.082	0.420	0.002	1	0.693	0	3.9537	0.1208
3416-24-8	***D-Glucosamine[a]***	0.838	0.082	**0.387**	0.002	1	0.693	0	4.3961	0.1536
14307-02-9	D-Mannosamine[a]	0.838	0.082	0.442	0.003	1	0.693	0	3.7619	0.1045
71-00-1	L-Histidine[b]	0.954	0.005	0.527	0.004	1	0.867	0	7.8643	0.0652
73-22-3	L-Tryptophan[b]	0.891	0.066	0.491	0.018	1	0.950	0	10.4119	0.1562
91-53-2	Ethoxyquin[c]	0.766	-0.333	0.543	0.010	1	0.923	0	7.7699	0.0768
67306-00-7	Fenpropidin[d]	0.652	-0.134	0.587	0.010	1	0.909	0	4.8662	0.0871
98319-26-7	Finasteride[e]	0.687	-0.060	0.534	0.006	1	0.898	0	5.2165	0.0513
1665-48-1	***Metaxalone[f]***	0.803	-0.170	0.536	**0.021**	1	0.930	0	9.1898	0.1664

Table B.1. Continued.

CAS	Compound	SIC2	MATS3s	E1e	R8v+	B02[C-N]	PDI	ArHdrxl_-OH	Pred. by eq.	HAT i/i (h*=0.2000)
Other										
63675-72-9	Nisoldipine[g]	0.792	-0.013	0.448	0.012	1	0.900	0	8.2926	0.1015
55-63-0	**Nitroglycerin[h]**	0.627	0.120	0.435	**0.025**	1	0.733	0	4.0056	0.3531
51-52-5	Propylthiouracil[i]	0.883	-0.124	0.598	0.002	1	0.917	0	7.0853	0.0747
58-14-0	**Pyrimethamine[j]**	0.818	-0.117	0.409	**0.020**	1	**0.987**	0	11.9592	0.2642
58-55-9	Theophylline[k]	0.810	-0.006	0.459	0.000	1	0.958	0	8.2609	0.0934
134308-13-7	Tolcapone [l]	0.790	0.045	0.585	0.010	1	0.893	2	11.3988	0.1199

[a] Aminosugar; [b] Amino acid; [c] Antioxidant; [d] Pesticide; [e] Synthetic analog of testosterone; [f] Muscle relaxant; [g] Calcium channel blocker; [h] Vasodilator/Vasoconstrictor; [I] Anti-hyperthyroidism; [j]Antiparasitic drug; [k] Anti-asthma; [l] to treat Parkinson's disease.

*The chemicals that were out of the descriptor range written in bold and the chemicals that falls in the AD of the model but falls out of the descriptor range written in bold and italic