ASSESSMENT AND IN SILICO MODELLING OF THE TOXICITY OF SELECTED EMERGING POLLUTANTS TO *CHLORELLA VULGARIS*

by

Gülçin Tuğcu

BS. in Math., Hacettepe University, 1995

MS. in Appl.Math., Georgia Southern University, 2004

MS. in E.Sc., Boğaziçi University, 2011

Submitted to the Institute of Environmental Sciences in partial fulfillment of

the requirements for the degree of

Doctor

of

Philosophy

Boğaziçi University

2017

ACKNOWLEDGMENTS

Foremost, I would like to thank my thesis advisor Prof. Melek Türker Saçan for her support, guidance, and ideas during the development of this thesis. I would also like to express my gratitude to Prof. Meral Birbir, Prof. Ferhan Çeçen, Prof. Nilsun İnce, and Prof. Safiye Sağ Erdem for their insightful and valuable comments and also for their participation in my thesis committee.

I offer my deepest thanks to my colleagues Dr. Doğa Ertürk, Filiz Ayılmaz, Gülhan Özkösemen, and Dr. Ayşe Tomruk for sharing their expertise. I also would like to thank Nagihan Elif Kahraman and Defne Şahin for technical support. I would like to thank Prof. P. Gramatica for providing QSARINS program.

I am indebted to my close friends Serli, Hüma, Leyla, Neda, and Serap for their friendship, accompany during my study.

Finally, I would like to thank my family for their love, patience, and support in every sense. Without them, it would be literally impossible to finish this thesis.

This study was carried out in Ecotoxicology and Chemometrics Laboratory of Institute of Environmental Sciences, Bogazici University, Istanbul, Turkey. The financial supports of Boğaziçi University Scientific Research Funds (Projects 6052, 6730, and 8502) and TÜBİTAK (Project 214Z225) are appreciated.

ASSESSMENT AND IN SILICO MODELLING OF THE TOXICITY OF SELECTED EMERGING POLLUTANTS TO *CHLORELLA VULGARIS*

Release of emerging pollutants such as pesticides, phthalates, and substituted phenols and anilines is detrimental threat for the aquatic environment. Registration, Evaluation, Authorization and Restriction of CHemicals (REACH) regulation requires algal toxicity data for regulatory risk assessment purposes. Quantitative Structure–Toxicity Relationships (QSTRs) are well accepted tools for data gap-filling. Therefore, studying the toxic effects of chemicals on algae via experimental and *in silico* methods would provide invaluable information for the chemicals with no toxicity data; and the knowledge gained through this study forms a scientific basis towards the protection of aquatic ecosystems. In the present study, the 96-h algal toxicity tests were performed with nitro-, chloro-, methoxy-, and methyl- substituted phenols and anilines to *Chlorella vulgaris*. Merging these data with the previously reported toxicity data of our laboratory enabled a high quality single source algal toxicity data for toxicity modeling. Consequently, models for the prediction of acute toxicity and low-toxic-effect concentrations were developed and verified based on the principles of OECD. Interspecies models were also developed using algae-algae and algae ciliate toxicity data. Developed models displayed decent predictivity and have a high potential to assess the toxicity of untested phenols and anilines on *C. vulgaris* within the applicability domain of models.

YENİ ORTAYA ÇIKAN BAZI KİRLETİCİLERİN *CHLORELLA VULGARIS* TOKSİSİTELERİNİN BELİRLENMESİ VE MODELLENMESİ

Pestisitler, ilaçlar, fitalatlar ve fenol ve anilin türevleri gibi yeni ortaya çıkan kirleticilerin çevreye salınması sucul çevre açısından yıkıcı tehdit oluşturmaktadır. Registration, Evaluation, Authorization and Restriction of CHemicals (REACH) regülasyonu risk belirleme amaçları için alg toksisite verileri gerektirmektedir. Kantitatif Yapı-Toksisite İlişkileri (KYTİ) eksik verileri tamamlamada kabul edilir araçlardır. Bu nedenle, toksisite verisi olmayan kimyasalların alg üzerindeki toksik etkilerinin deneysel ve bilgisayarla modelleme yoluyla çalışılması çok değerli bilgi sağlayacaktır. Bu çalışma ile elde edilecek bilgi sucul ekosistemlerin korunmasına bilimsel bir taban oluşturacaktır. Bu çalışmada, nitro, kloro, metoksi ve metil eklenmiş fenol ve anilin türevlerinin *Chlorella vulgaris*'e olan etkileri 96 saatlik alg toksisite deneyleriyle belirlenmiştir. Toksisite modellemesi için bu verinin daha önceden laboratuvarımızdan raporlanmış veri ile birleştirilmesi tek kaynaktan yüksek kalitede alg toksisite verisini mümkün kılmıştır. Sonrasında, akut toksisite ve düşük toksik etki konsantrasyonlarının tahmini için modeller geliştirilmiş ve OECD ilkelerine dayanılarak doğrulanmıştır. Alg-alg ve alg-silli toksisite verileri kullanılarak türlerarası modeller de geliştirilmiştir. Geliştirilen modeller, iyi tahmin edilebilirlik gösterdi. Bu modeller, uygulanabilirlik alanı dahilinde, test edilmemiş fenollerin ve anilinlerin *C. vulgaris* üzerindeki toksisitesini değerlendirmek için yüksek bir potansiyele sahiptir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

Symbol/Abbreviation	Explanation	Units used
ACR	Acute to Chronic Ratio	
AD	Applicability Domain	
BPNN	Back Propagation Neural Networks	
CAS	Chemical Abstracts Service	
ChV	Chronic value	
CPANN	Counter Propagation Artificial Neural Networks	
C. vulgaris	*Chlorella vulgaris*	
E	Gas phase energy	eV
E_{aq}	Aqueous energy	eV
E_{HOMO}	Energy of the Highest Occupied molecular Orbital	eV
E_{LUMO}	Energy of the Lowest Unoccupied molecular Orbital	eV
EC_x/IC_x	Effective Concentration/ Inhibitory Concentration that reduces the observed endpoint by x%	
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals	
ECHA	European Chemicals Agency	
ECOSAR	Ecological Structure Activity Relationships	
ECOTOX	ECOTOXicology database	
F	Fischer statistics	
GA	Genetic Algorithm	
h^*	Critical hat value	
IC_{50}	Concentration that inhibits algal growth by 50%	mg L^{-1}
k-MCA	k-Means Cluster Analysis	
Log D	Distribution coefficient	
Log K_{ow}	Logarithm of *n*-octanol/water partition coefficient	
Log P	Partition coefficient	
LOEC	Lowest Observed Effect Concentration	mg L^{-1}
MAE	Mean Absolute Error	
MATC	Maximum Acceptable Toxicant Concentration	

Symbol/Abbreviation	Explanation	Units used
MLR	Multiple Linear Regression	
mM	Millimolar	
MOA	Mode of Action	
NOEC	No Observed Effect Concentration	mg L^{-1}
OECD	Organisation for Economic Co-operation and Development	
PBT	Persistent Bioaccumulative Toxic	
PEC	Predicted Environmental Concentration	
PLS	Partial Least Squares	
PNEC	Predicted No Effect Concentration	
pT	Negative logarithm of IC_{50} or EC_{50}	mM
Q^2_{Loo}	Leave-one-out cross validation parameter	
QSAR	Quantitative Structure-Activity Relationship	
QSTR	Quantitative Structure-Toxicity Relationship	
QTTR	Quantitative Toxicity-Toxicity Relationship	
R	Pearson correlation coefficient	
R^2	Coefficient of determination	
R^2_{adj}	Adjusted (for degrees of freedom) squared correlation coefficient	
REACH	Registration, Evaluation, Authorization, and Restriction of CHemical substances	
RMSE	Root Mean Squared Error	
SE	Standard Error of the estimate	
SOM	Self-organizing Maps	
SVM	Support Vector Machine	
SVR	Support Vector Regression	
US EPA	United States Environmental Protection Agency	

1. INTRODUCTION

In the last decades, pollutants originating from pesticides, personal care products, nanomaterials, and pharmaceuticals have been released into the environment and considered as emerging pollutants. The definition of emerging pollutants is; synthetic/manufactured or natural chemicals that have no regulatory standard and have been recently noticed in the environment. Growing evidence suggests that adverse effects could occur at environmentally relevant concentrations of these pollutants. However, their environmental releases are not included in routine monitoring programs for testing their presence in the environment. These chemicals are candidates for future regulation depending on their ecotoxicity, potential health effects, public perception, and frequency of occurrence in environmental compartments (Hoenicke et al., 2007; EPA, 2008). Some of these chemicals have been classified as Persistent Bioaccumulative and Toxic (PBT) chemicals such as phthalates, pesticides, polyaromatic hydrocarbons (PAHs), and substituted phenols and anilines by United States Environmental Protection Agency (US EPA) (http://www.epa.gov/pbt/). Therefore, to determine the toxicity of these contaminants to non-target species, such as algae, is beneficial to understand their impact to ecosystems.

Green algae play an important role in the equilibrium of aquatic ecosystems, being the first level of the trophic chain to produce nutrients and oxygen. The disturbance of sensitive algal communities has the potential leading to a biomagnified response by higher aquatic species living in the same aquatic ecosystems. *Chlorella vulgaris* is an environmentally significant green algae due to its widespread distribution in natural waters (Ventura et al., 2010). Therefore, studying the toxic effects of selected chemicals in the present study on freshwater algae, namely *C. vulgaris* would provide valuable information regarding their toxic potencies, and the knowledge gained through the toxicity tests forms the scientific basis towards the protection of aquatic ecosystems.

There are many studies which explore benefits of algae, since they have valuable cellular components such as pigments, fatty acids, vitamins, antioxidants, etc. (Priyadarshani and Rath, 2012). Algae are used as food supplement (Tokuşoglu and Ünal, 2003), animal feed (Vanthoor-Koopmans et al., 2014), pharmaceutical (Vo et al., 2012), dye (Gouveia et

al., 2007), and fuel feedstock (Mallick et al., 2011). They are also benefited in hydrogen production (Eroglu and Melis, 2016), carbon dioxide reduction (Raeesossadati et al., 2014), and water treatment (Lim et al., 2010).

Due to the ecological significance of algae, Registration, Evaluation, Authorization, and restriction of CHemicals (REACH) legislation requires ecotoxicity data, including algal growth inhibition test results, for chemicals manufactured in or imported into the European Union (EC, 2006). As a consequence, a significant amount of data is necessary to fulfill the requirements. Regarding a greater testing demand in the European Union (EU) and the REACH legislation, the use of valid and quality *in silico* methods like quantitative structure-(activity/toxicity) relationship (QS(A/T)R) models are encouraged to meet the regulatory testing needs. Understanding the relationship between the molecular structure and a particular effect in a biological system will lead to useful models. The formal development of these predictive models relating the molecular structure and a particular activity quantitatively is called QSAR (Cronin, 2010). Similarly, Quantitative Structure – Toxicity Relationship (QSTR) studies, relating the physicochemical properties of chemicals with their toxicity on the basis that similar compounds have similar toxicities, are expected to reduce the cost and the number of organisms used for toxicity testing (Sullivan et al., 2014). The generation of a proper QSTR model is based on the quality of the toxicity data used for modeling. However, the development of QSTR models using compiled data from the literature has the risk of yielding misleading results originating from the discrepancies between laboratories. Therefore, high quality experimental data generated in the same laboratory according to a REACH compatible endpoint is of paramount importance for risk assessment and necessary for the calculation of Predicted No Effect Concentrations (PNECs). Besides, these data can be used in interspecies correlations, read-across and provide a valuable basis to explore QSTR. While toxicity values are predicted for untested and designed chemicals within the applicability domain (AD), QSTRs can be used for screening as well as prioritization.

QSTR models are categorized into two broad groups, as linear and nonlinear. While linear models are known to be transparent and easily applicable, nonlinear models are preferred in explaining nonlinear relations between the modelled variable and the structure of the molecule.

Organisms in the environment are exposed continuously to low concentrations of a variety of compounds simultaneously and thus, chronic effects are likely to occur to aquatic and terrestrial organisms. Therefore, to determine the toxic level of a chemical on a certain species, not only acute toxicity but also chronic toxicity data are needed and have gained significant attention. Chronic toxicity values for algae are obtained via standardized bioassays. In a regular 72-h or 96-h batch algal toxicity test, NOEC and/or a low-toxic effect concentration is obtained. Then, in environmental risk assessment, no-effect concentration of the subject chemical is estimated using available toxicity data such as NOEC and low-toxic effect concentration (e.g. EC_{10}). Where these values are not available, acute-to-chronic extrapolation is used.

Besides QSTR, Quantitative Toxicity-Toxicity Relationship (QTTR) is becoming an important tool for determining the toxicity of a chemical using interspecies relations. QTTRs also have the potential to fill the gaps where toxicity data are scarce. There have been studies on interspecies toxicity prediction in the literature to fill these gaps and also to understand the toxic mechanism of chemicals. From the aquatic environment, bacteria, ciliate, algae, daphnia, and fish toxicities were found to be correlated (Kar and Roy, 2010; Zhang et al., 2010; Aruoja et al., 2011; Singh et al., 2014; Furuhama et al., 2015). Therefore, it is worth searching interspecies toxicity relationships to make toxicity predictions for algae using other species' toxicity data.

1.1. The Aim and Contribution of the Thesis

The aim of the present study is three-fold. Firstly, performing toxicity bioassays for environmentally significant chemicals were targeted. These experiments were carried out using freshwater algae, *C. vulgaris*, due to their importance in the aquatic environment and in the food chain. The second aim of the present study is to develop validated QSTRs for the toxicity prediction of untested chemicals. This part includes both acute and chronic toxicity estimation. Finally, the third aim of the present is to develop QTTR models using the generated toxicity data for aquatic species.

To achieve these goals, 96-h algal toxicity assays of 62 chemicals with different mode of actions (MOAs) for *C. vulgaris* were performed. The generated toxicity data were then

combined with the results of the toxicity values obtained previously in our laboratory. Consequently, the data set of 84 chemicals were used to develop validated QSTRs for *C. vulgaris*. QTTR models using the toxicity data for *C. vulgaris* and two other aquatic species, namely *Tetrahymena pyriformis* and *Pseudokirchneriella subcapitata*, were developed.

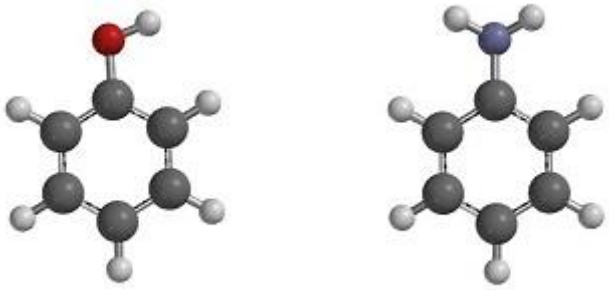
The contributions of this thesis are as follows:

- The new toxicity data for untested and hazardous chemicals were transmitted.
- The data are the outcome of 96-h batch algal assays run in the same laboratory, and have been extended the previous work in our laboratory.
- Validated linear (MLR) and nonlinear (SVM and BPNN) QSTRs models with wide ranges of applicability domains in line with OECD principles were developed.
- Validated QTTRs predicting algal toxicity values were developed using other two species' toxicity data.
- The data gaps on acute and chronic algal toxicity required for regulatory assessment of studied chemicals were filled.

2. THEORETICAL BACKGROUND

2.1. Environmental Significance of Phenols and Anilines

Environmental contamination due to organic chemicals including phenols, anilines and their derivatives is prevalent as a result of anthropogenic activities. The phenolic compounds in the aquatic environment arise from industrial activities, agricultural practices, and natural substance degradation (Dimou et al., 2006). Pharmaceuticals, food additives, and personal care products are also sources of substituted phenols (Selassie and Verma, 2015). Nitrophenols are used in dyes, solvents, plastics, and explosives production (Michalowicz and Duda, 2007). Aniline and its derivatives are introduced into the environment from many different fields of applications, such as the production of isocyanates, rubber processing chemicals, dyes, pesticides, and pharmaceuticals (Aruoja et al., 2011). Active pesticide ingredients are estimated to be used 1 to 2.5 million tons annually, mainly in agriculture. Production and use of these compounds potentially end in the aquatic ecosystems via industrial wastewater discharge, storm-water discharges, and return flows from irrigated fields (Mandaric et al., 2016). Aminophenols are intermediates for dyes (Morel and Christie, 2011) and pharmaceuticals (Sun et al., 2004). Production and use of these compounds ultimately end in aquatic ecosystems via either industrial wastewater discharge or surface runoff. The steady growth of chemical industry has led to vast number of chemicals in the environment. The environmental concentrations of these chemicals may be observed as high as 2000 μg/L for chlorophenols, 0.04-10 μg/L for nitrophenols, and 204 μg/L for methylphenols in river waters in Japan (Michalowicz and Duda, 2007). Due to their extensive usage of phenols and anilines (Figure 2.1 (a) and (b), respectively), especially as industrial, biocidal, and pharmaceutical chemicals draws attention to these chemicals as emerging and environmentally important substances. They are included in the category of PBT chemicals.



(a) (b)

Figure 2.1. (a) phenol and (b) aniline.

2.1.1. Mode of action of studied chemicals

Phenols and anilines comprise two large groups of environmentally important chemicals. The majority of the industrial chemicals exhibit a narcotic mode of action (Bradbury and Lipnick, 1990), and chemicals with different modes of action are also included in industrial chemicals. Based on the classification reported by Cronin et al. (2002), the phenols and anilines selected for toxicological assessment in the present study are expected to elicit toxicity through one of the following mechanisms: polar narcosis, respiratory uncoupling, pro- or soft electrophilic. While polar narcotic chemicals are well-correlated with a hydrophobicity parameter such as log *P* or log *D*, others exert excess toxicity.

2.2. Environmental Significance of *Chlorella vulgaris*

Determining the adverse effects of chemicals to algae is of paramount importance for risk assessment and environmental regulation. Algal assemblages are used to monitor the impacts of aquatic stressors and aquatic toxicity because of their sensitivity to pollutants and their short life cycle. Among the algal species, *C. vulgaris* is a preferred species in algal toxicity studies (Ma et al., 2002; Cronin et al., 2004; Cai, et al., 2009; Sevcik, et al., 2009; Murkovski and Skórska, 2010; Ertürk and Saçan, 2013) due to its widespread distribution (Ventura et al., 2010), natural presence in freshwater ecosystems, and fast growth (Murkovski and Skórska, 2010). However, only 111 chemicals (~10%) have *C. vulgaris* toxicity data regarding a compiled 72-h and 96-h algal toxicity data of 1081 chemicals for 26 species (Fu et al., 2015). Therefore, a significant amount of data is necessary to fill

toxicity data gap for many organic chemicals towards this algae. Owing to its versatile structure, *C. vulgaris* is also produced for different purposes such as food, fuel, and pharmaceuticals (Safi et al., 2014).

C. vulgaris is a unicellular spherical green alga with a diameter of 2-10 μm (Figure 2.2). Their cells have rigid cell walls and a single chloroplast. The cells are non-motile and reproduce asexually and rapidly (Rai et al., 2013). It was first discovered by Beijerinck in 1890 (Safi et al., 2014). The scientific classification of *C. vulgaris* is given in Table 2.1. *C. vulgaris* is found in both freshwater and marine environments naturally (Reynolds, 1984; Chaminda Lakmal et al., 2015). It is also a habitant in natural waters of Turkey (Tas and Gonulol, 2007; Baykal et al., 2011).

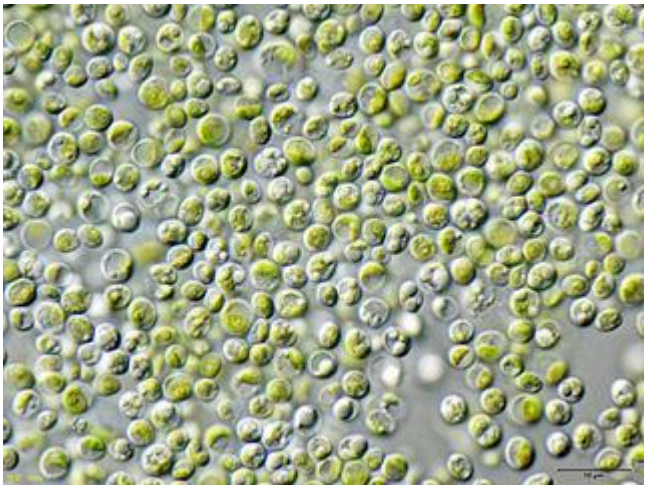


Figure 2.2. Microscopic view of *C. vulgaris* (Beijerinck NIES-2170) (https://commons.wikimedia.org/wiki/Category:Chlorella_vulgaris).

Table 2.1. Scientific classification of *C. vulgaris*.

Domain	Eukaryota
Kingdom	Viridiplantae
Division	Chlorophyta
Class	Trebouxiophyceae
Order	Chlorellales
Family	Chlorellaceae
Genus	Chlorella
Species	*Chlorella vulgaris*

2.3. Algal Toxicity Tests

Algal toxicity tests, dating back to 1910, are done according to standards and guidelines. However, standardized assays with freshwater algae were developed in the 1960s (Janssen and Heijerick, 2003). The standardized tests are described in several guidelines: Ecological Effects Test Guidelines OCSPP 850.4500 (EPA, 2012), OECD 201 (OECD, 2011), APHA, AWWA, and WEF (2012), Environmental Science and Technology Centre of Canada (Environment Canada, 2007), and books (Staveley and Smrchek, 2005; Stauber et al., 2005).

Toxic effects of chemicals are determined using several bases some of which are:

- Growth rate inhibition calculated according to
    - Average specific growth rate
    - Yield
    - Biomass
- The disappearance of fluorescein diacetate (FDA)
- Dissolved oxygen production
- Cell number

Test duration for acute algal bioassays is usually 72-96 h. Although Organisation for Economic Co-operation and Development (OECD) test design is 72-h (OECD, 2011), there are also 96-h batch assays (EPA, 2012; Staveley and Smrchek, 2005). Algal stocks are inoculated during growth phase (4-8 days old). Starting with a population of 100,000 cells, assays are run in exponential growth phase. Although *Pseudokirchneriella subcapitata* and *Desmodesmus subspicatus* are recommended species by OECD Guideline No: 201, any non-chain forming (Stauber et al., 2005) or non-attached (OECD, 2011) microalgae can be used as the test species as long as it is confirmed that their exponential growth can be maintained throughout the exposure phase. The important parts of the assays are: the *p*H should not change more than 1.5 units, the light intensity and temperature should be kept constant, and the test should be performed in growth phase. Reference toxicants are used to assess the reproducibility and reliability of the results. These results for a reference toxicant are compared with the test results obtained in the previous tests in the literature.

Since several generations are produced during the tests, low-toxic-effect concentration results are considered in chronic values (Ahlers et al., 2006; Chen et al., 2009; ECOSAR, 2012). European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) database considers algal tests as chronic studies for those longer than 12 hours (ECETOC, 2003). 72-h (or longer) no observed effective concentration (NOEC) and lowest observed effective concentration (LOEC) can be regarded as a long-term result according to the European Chemicals Agency (ECHA, 2008).

2.3.1. Response variable calculation

Quantifying the concentration-response curve from an algal toxicity test was done by selecting an empirical mathematical model that describes a sigmoid curve and by applying conventional curve-fitting techniques to the data. Feasible models include Probit, Logit, and Weibull methods (Christensen et al., 2009), and ICp (EPA, 2002; Norberg-King T., 1988). NOEC and LOEC values were estimated by using Dunett's test or Steel's many-one rank test.

2.4. Quantitative Structure – Toxicity Relationships (QSTRs)

QSTRs are well-defined models that correlate properties of a chemical with its toxicity. Several requirements and principles are defined for a QSTR model that can be used officially. According to the OECD principles, a QSTR model should have appropriate measures of goodness-of-fit, robustness, and predictivity for a reliable model associated with the following information:

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness-of-fit, robustness, and predictivity
5. a mechanistic interpretation, if possible (OECD, 2007).

Guidance on information requirements and chemical safety assessment of REACH framework (ECHA, 2008) states that the information generated by QSTRs may be used instead of experimental data, if they provide the following conditions.

- Predictions should be obtained from a QSTR model whose scientific validity has been established,
- Predicted value of the substance should fall within the applicability domain of the QSTR model used,
- The results should be adequate for the purpose of classification and labeling, and/or risk assessment, and
- Adequate and reliable documentation of the applied method should be provided.

The generation of a good QSTR model is based on the quality of toxicity data used for modeling. However, the development of QSTR models using compiled data from the literature has the risk of yielding misleading results originating from the discrepancies between laboratories. Therefore, high quality data generated in the same laboratory according to a REACH compatible endpoint provide a valuable basis to explore QSTR, which can be used to predict the toxicity of untested and designed compounds. While toxicity values are predicted for untested chemicals within the applicability domain (AD), QSTRs can also be used for screening and prioritization. In this respect, QSTR models have been studied extensively for phenol and aniline derivatives due to their widespread usage and hazard (Furuhama et al., 2015; Dieguez-Santana et al., 2016; Fan et al., 2016; Chen et al., 2017; Abbasitabar et al., 2017).

2.4.1. Molecular descriptors

A QSTR relates toxicity of a set of similar chemicals with their selected properties. These properties are quantitative parameters called descriptors. While empirical descriptors are obtained via experiments, theoretical descriptors are calculated using some mathematical algorithms implemented in software. A calculated molecular descriptor is the final result of a logic and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a number (Todeschini and Consonni, 2009). Descriptors are usually classified as physicochemical, structural, topological, electronic, and

geometric (Roy et al., 2015a). There are numerous software packages that calculate molecular descriptors some of which are CODESSA (www.semichem.com), DRAGON (www.talete.mi.it), HyperChem (www.hyper.com), Mopac (www.openmopac.net), ADMET (www.simulations-plus.com), SPARTAN (www.wavefun.com), and PaDEL-Descriptor (www.yapcwsoft.com/dd/padeldescriptor/). In the present study, DRAGON, SPARTAN, and ADMET descriptors were calculated for the studied chemicals.

2.4.2. Methods used for training/test set division

External validation is a way to establish the reliability of a QSTR model. To validate a model, the data sets are usually split into training and test sets. There are numerous division methodologies such as factorial designs, D-optimal designs, periodical division, self-organized maps etc., in the literature. To select a representative subset of samples from the whole dataset, factorial designs, and D-optimal designs (Eriksson and Johansson, 1996) are used. Factorial designs presume that different sample properties (such as substituent groups at certain positions) are divided into groups. Training set should include one representative for each combination of the properties. For a diverse dataset, this approach is impractical, and fractional factorial designs are used, in which only a part of all combinations is included into the training set. Training and test sets can be selected by using sphere exclusion algorithms (Golbraikh and Tropsha, 2002) and longest minimum distance (LMD) method (Ghasemi, et al., 2013) which are the other division methods. In periodical division, the data set is ordered with respect to the dependent variable. Then, starting from the second chemical, every third chemical, for example, is allocated into the test set, such that, the chemicals with the lowest and the highest toxicity values are left in the training set. For this division, the descriptor values are not used.

Kohonen neural networks, also known as self-organized maps (SOMs), are able to select a representative training set, and a test set similar to it (Devillers, 1996, Zupan and Gasteiger, 1999). “Self-organized” indicates that the learning does not need a dependent variable. Kohonen networks project multi-dimensional space into a 2D array of neurons. The projection, which is called learning of network, runs in two steps. In the first step, an object (represented by a vector) is presented to all neurons and the algorithm selects the most similar neuron, called the “winning neuron”. In the second step, the weights of the winning

neuron are modified to the vector values and at the same time the neighboring neurons are modified to become similar to it (Vracko, 2005). At the end of the learning session, all compounds locate in cells of the map such that; while the most similar compounds dwell in the same cell, neighboring cells have less similar ones. Blank neurons are also possible. A sample Kohonen map (Figure 2.3) was obtained using Kohonen and CPANN Toolbox for MATLAB (Ballabio et al., 2009, Ballabio and Vasighi, 2012) for 46 substituted phenols with 1800 descriptors calculated by using DRAGON (www.talete.mi.it) and SPARTAN (www.wavefun.com) software packages. The closeness of neurons is related to the similarity. For example, the chemicals in the upper right neuron, 32 (2,3-dimethylphenol), 35 (2,6-dimethylphenol), and 36 (3,4-dimethylphenol), are closely related. Similarly, the chemicals in the neuron just below the upper right neuron, 33 (2,4-dimethylphenol), 34 (2,5-dimethylphenol), and 37 (3,5-dimethylphenol) are also closely related. However, chemicals in different neurons are less similar considering their descriptor values. Complete list of the numbered chemicals in Figure 3.2 can be found in Table 2.2.

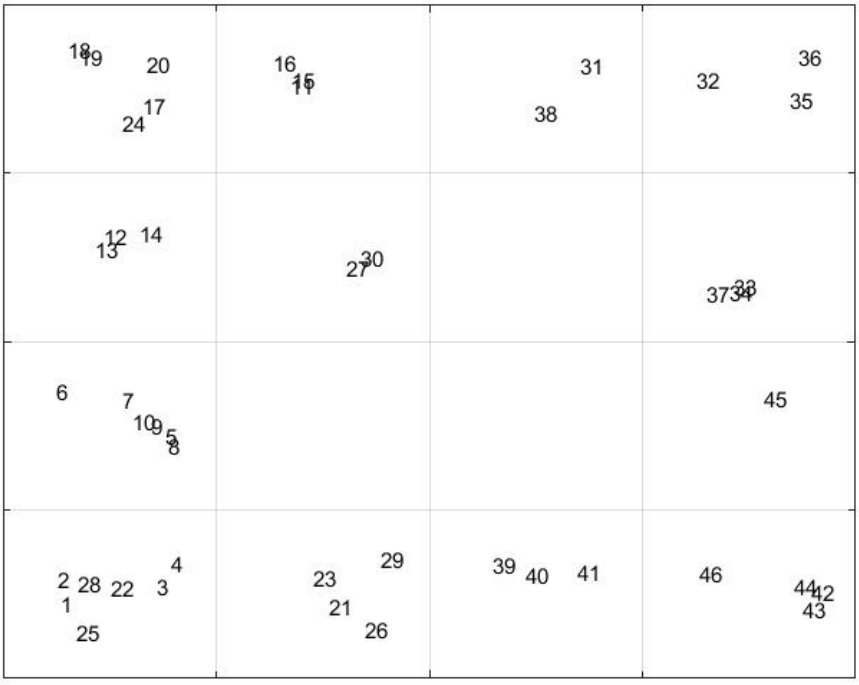


Figure 2.3. A sample Kohonen top-map for 46 chemicals listed in Table 2.2. spanning onto a 4x4 grid.

Table 2.2. The sample data set for Kohonen network grouping.

ID	Chemical	ID	Chemical
1	Phenol	24	Tetrachlorohydroquinone
2	2-chlorophenol	25	Catechol
3	3-chlorophenol	26	4-chlorocatechol
4	4-chlorophenol	27	3,5-dichlorocatechol
5	2,3-dichlorophenol	28	Resorcinol
6	2,4-dichlorophenol	29	4-chlororesorcinol
7	2,5-dichlorophenol	30	4,6-dichlororesorcinol
8	2,6-dichlorophenol	31	2-methylphenol
9	3,4-dichlorophenol	32	2,3-dimethylphenol
10	3,5-dichlorophenol	33	2,4-dimethylphenol
11	2,3,4-trichlorophenol	34	2,5-dimethylphenol
12	2,3,5-trichlorophenol	35	2,6-dimethylphenol
13	2,3,6-trichlorophenol	36	3,4-dimethylphenol
14	2,4,5-trichlorophenol	37	3,5-dimethylphenol
15	2,4,6-trichlorophenol	38	4-chloro-3-methylphenol
16	3,4,5-trichlorophenol	39	2-nitrophenol
17	2,3,4,5-tetrachlorophenol	40	3-nitrophenol
18	2,3,4,6-tetrachlorophenol	41	4-nitrophenol
19	2,3,5,6-tetrachlorophenol	42	2,4-dinitrophenol
20	Pentachlorophenol	43	2,5-dinitrophenol
21	1,2,3-trihydroxybenzene	44	3,4-dinitrophenol
22	Hydroquinone	45	5-methyl-2-nitrophenol
23	Chlorohydroquinone	46	2-methyl-4,6-dinitrophenol

The k-Means Cluster Analysis (k-MCA) is also used to divide the data into training and test sets, so that general characteristics appear in both sets. To ensure a statistically acceptable data, the data is partitioned into several clusters in terms of the response variable. Particular characteristics of all compounds are represented in each cluster derived from k-MCA. Selection is carried out by taking, in a random way, chemicals belonging to each cluster (Caballero and Fernandez, 2006).

The primary purpose of hierarchical clustering is to display the data in such a way as to emphasize its patterns. In hierarchical clustering, the descriptors are clustered into subgroups in a series of partitions. The basic process of hierarchical clustering starts by assigning each chemical as a cluster. Then the similarities between the clusters are determined by the distances (similarities) between the descriptors they contain. In the next step, the most similar (closest) pair of clusters are merge into a new cluster. The previous two steps are iterated until all chemicals are clustered into a single cluster containing all chemicals. The

results, which are of qualitative nature, are presented in the form of a dendrogram allowing one to visualize the chemicals in a 2D space (He and Jurs, 2005)

In the present study, Kohonen, k-means cluster analysis, hierarchical clustering, and periodical division methods were used to obtain different division sets.

2.4.3. Descriptor selection

With the help of vast number of software, descriptors that can be used in QSAR models exceed 5000. The selection method of significant descriptors to be used in the models is, therefore, important. QSAR models are expected to have the number of descriptors as low as possible. This idea relies on the following facts: a) the dependent variable is intended to be explained in the simplest way, which corresponds to the smallest model, b) redundant descriptors will add noise to the estimation, c) collinearity risk increases with the increase in the number of descriptors, d) time is saved by not measuring the redundant descriptors. The significant descriptors for model development are selected by different methods. Some of these are; All Subsets method in QSARINS (Gramatica et al., 2013; Gramatica et al., 2014), Stepwise Regression, and Genetic Algorithm (GA). All Subsets method in QSARINS explores all the possible combinations of the descriptor pool (model size). The best linearly correlated combinations are listed by the software in terms of Q^2_{LOO} (cross validation leave-one-out). This method guarantees that the best subset of variables is found. However, it is very time consuming when the number of descriptors is high (Cassotti et al., 2014). In stepwise multiple regression, the independent variables are entered to control the contribution of the other variables already in the model. Variables are added to/removed from the regression equation one at a time, which may lead to a suboptimal solution. The process of adding more variables stops, when it is not possible to make a statistically significant improvement in R^2 using any of the variables not yet included (Tamhane and Dunlop, 2000). GA imitates properties such as adaptability, and heredity of living beings as natural selection. The use of the heuristic organized operations of “reproduction”, “crossing”, and “mutation” from casual or user-selected starting “populations” generates the new “chromosomes”- or models / descriptor sets (Kuz’min et al., 2010). QSARINS employs Tournament Selection method to select the best representative descriptors via GA. Reshaped Sequential Replacement method, which is an augmented form of Sequential Replacement

method starts with a model consisting of randomly chosen set of descriptors. At each iteration, a descriptor is replaced to see if a better model can be obtained. The procedure continues until no better model can be found (Cassotti et al., 2014, Grisoni et al., 2014).

2.5. Modeling Techniques

QSAR models could be broadly categorized as linear and nonlinear models. While linear models are easy to interpret and easy to apply, nonlinear models are more successful in explaining nonlinear relationships between an activity and the structure of the related molecule. In a study by Timofei et al. (1997), the authors inferred that the Multiple Linear Regression (MLR) approach leads to a better interpretation of the contribution of individual terms, but neural networks can extract more 'information' from the data than statistical methods, especially where nonlinear relationships are involved. Caballero and Fernandez (2006) constructed models to predict antifungal activity using MLR and Bayesian-regularized neural networks. Although the nonlinear model performed better, they concluded that the same features play important role during the process of linear and nonlinear descriptor selection. Carlsson and his co-authors (2009) have showed that it is possible to interpret nonlinear machine-learning methods. The authors devised linear (Partial Least Square (PLS)) and nonlinear (Support Vector Machine (SVM)) models for a simulated Ames mutagenicity data. They found that PLS results are poor, explaining the fact that a linear method cannot accurately describe nonlinear data; and in terms of interpretability, linear methods are of less value when applied to nonlinear relationships. In a study by Saçan et al. (2010), toxicity of organic chemicals to freshwater algae were modeled via Counter-Propagation Neural Networks (CPANN) and MLR. Authors observed that CPANN models have higher correlation coefficients and slightly higher Root Mean Squared Errors (RMSEs) than MLR models. Raevsky et al. (2011) developed and compared linear and nonlinear QSAR models of acute intravenous toxicity of organic chemicals for mice. They eventually concluded that the linear and nonlinear QSAR relationships should be explored due to multifactor phenomenon like toxicity. They suggested that linear models may be used as local models for assessment of toxicity of specific functional groups, while nonlinear models may be used for heterogeneous sets of chemicals or chemicals containing several functional groups. In a study by Xu et al., (2012), MLR, PLS, and SVM methods were used to construct models for the prediction of human oral bioavailability (log B). They reported that models

had slight differences in their performances, and concluded that the linear and nonlinear methods they employed were appropriate for predicting log B. MLR and back-propagation neural networks (BPANNs) were used as feature mapping techniques for prediction of the dermal penetration rate of some volatile and nonvolatile chemicals. Those linear and nonlinear models resulted in similar outputs (Fatemi and Malekzadeh, 2012).

MLR is a frequently used method in QSAR models owing to its easy application and interpretability. The relation between the descriptors and the modeled activity is transparent. The generalized expression of an MLR model is the following equation:

$$Y=c+a_1x_1+a_2x_2+\ldots+a_nx_n \quad (2.1)$$

With an n-descriptor MLR equation, Y is the dependent variable for the modeled activity, c is the constant term, and a_i is the corresponding coefficient of the descriptor x_i (Montgomery et al., 2012). While a positive coefficient suggests a positive contribution, a negative coefficient suggests a negative contribution to the modeled activity. However, when descriptors are highly intercorrelated these interpretations might be inaccurate. Therefore, the descriptors in the equation should not be intercorrelated, *e.g.* a Pearson correlation coefficient less than 0.7 or variance inflation factor (VIF) less than 5. QUIK rule is another test to detect collinearity in linear models (Todeschini and Consonni, 2009). This method tests whether the total correlation among the block of descriptors (K_{XX}) is higher than the correlation among them and the responses (K_{XY}). If $K_{XY} - K_{XX} >$ DeltaK, then the descriptors are regarded as not collinear. Delta K is a threshold value defined by the user. In the present study, MLR models were developed setting Delta K as 0.05. When the number of descriptors is high with respect to the number of chemicals, the possibility of intercorrelation increases. Hence, the number of descriptors in the equation should follow the Topliss ratio, *i.e.* the number of chemicals is at least five times the number of descriptors. In addition to the conditions explained above, each coefficient should be significant at $p<0.05$ level, which is checked with a *t*-test (Yee and Wei, 2012; Roy et al., 2015a).

The CPANN models generally have two layers, the input (Kohonen) layer and the output layer. CPANNs are built up from two layers of neurons arranged in 2D rectangular matrices. The Kohonen layer receives the input variables. Afterwards, it converts 3D input

into 2D map such that similar chemicals (having similar descriptors) are located in the same neuron. The output layer, which has the same topological arrangement of neurons as the input layer, receives the target (toxicity) values during the learning process.

Kriging (also known as Gaussian process regression) was first implemented in geostatistics for gold mining and it has also received some attention in the QSAR literature. Kriging has been used to model ciliate toxicity (Burden, 2001), algal toxicity (Tugcu et al., 2014), absorption, distribution, metabolism, and excretion (ADME) properties of organic compounds (Obrezanova et al., 2007), and basicities of pyridine derivatives (Hawe et al., 2010). Kriging basically estimates the unobserved points via a weighted linear combination of observed values (dependent variables), where the weights are determined to minimize the variance of the error by using proximity between descriptors (Fang et al., 2004).

Support Vector Machine method was initially proposed for classification problems by Cortes and Vapnik (1995) and later extended to regression applications (Support Vector Regression (SVR)) (Drucker et al., 1996). The main advantages of SVM are: results are stable, reproducible, and largely independent of the optimization algorithm, solution is guaranteed to be optimum without getting stuck at local minima, a simple geometric interpretation is attainable, and few parameters have to be adjusted like the regularization parameter (Doucet and Panaye, 2010). The performance of SVR models depends on type (epsilon (ε) or nu (ν)), cost (C, the regularization parameter), and the kernel type in general. There are four types of kernel functions, namely, linear, polynomial, radial basis function, and sigmoid function. The radial basis function type is widely preferred in regression problems (Panaye et al., 2006). SVM method has been successfully used in QSAR models to predict mode of action of toxic chemicals (Michielan et al., 2010), structural class of protein (Fernandez et al., 2011), bioactivity of HIV-1 integrase ST inhibitors (Xuan et al., 2013), aqueous solubility of drug-like molecules (Liang et al., 2011), adsorption of dyes on activated carbon (Örücü et al., 2014), and phenol toxicity on *Photobacterium phosphoreum* (Asadollahi-Baboli, 2012; Zhou et al., 2015).

Backpropagation neural networks (BPNN) are popular neural network architectures used in QSAR models. BPNNs are multilayer feed-forward neural networks trained by backpropagation of errors. They constitute of neurons organized in layers (input, hidden, and

output) and connected through weights. During the learning process, the weights are modified so that the response to a given input is similar to the target (Mazzatorta et al., 2005). The learning speed of the network is affected by the learning rates and momentum parameters (Niculescu, 2003). While their high preciseness of prediction is appreciated, the interpretability of the models is questioned. In a study by Baskin et al. (2002), the influence of descriptors on activities in a BPNN model was proven to be interpretable. Therefore, the term "black box" used for ANN models could be reevaluated. BPNN models have been found to be appealing in QSAR studies. The maximum adsorption capacities of dyes was modeled by Örücü et al. (2014), and BPNN model was found to be superior to the other methods performed in their study. Subchronic inhalation toxicity values in rodents was modeled with BPNN technique (Dobchev et al., 2013). Further information on BPNN in QSAR studies could be found in the literature (Zupan and Gasteiger, 1999).

In the present study, as linear regression, MLR; as nonlinear, SVR and BPNN methods were used for QSTR modelling.

2.6. Model Validation

Validation is a crucial aspect to prove reliability of models. Statistically robust and predictive models are capable of making accurate and reliable predictions of the modeled endpoint of untested chemicals. The model validation corresponds to OECD principle no 4 (OECD, 2007). The validation of QSAR models has two main components: internal validation and external validation. While internal validation is performed on training set and model itself, external validation involves testing the model on a test set. There are various validation strategies adopted. The widely accepted parameters and limits are explained below. Formulas were given for all validation parameters in Appendix Table A.1.

The robustness of a model, and the predictivity to some extent is tested by cross-validation leave-one-out (Q^2_{LOO}). The reliability of MLR models are tested by response randomization (Y-scrambling) procedure. For model randomization, the dependent variables of the training set are shuffled and new coefficient of determinations are calculated. The process is repeated several times. The significantly low correlation coefficients of the new

models indicate that the originally proposed model was not obtained by chance correlation (Gramatica, 2013).

Robustness of the models is verified by the parameter Q^2, later called Q_{F1}^2, which is a parameter proving the success of prediction of the model on a test set. Schüürmann (2008) showed that any difference between training and test set means may yield an overestimation of the prediction capability and proposed a new Q^2 parameter (Q_{F2}^2). Afterwards, Consonni et al. (2009) formulated a novel external correlation coefficient (Q_{F3}^2) for a test set based on sum of squares (SS) referring to mean deviations of observed values from a training set mean over a training set, instead of an external evaluation set. They concluded that correlation coefficients using either training set activity mean or test set activity mean have drawbacks. Therefore, the external predictive ability of the models should have information about the whole data set. In addition to these parameters, r_m^2 average is calculated for the predictions (Roy and Roy, 2009). This parameter penalizes a model for large differences between observed and predicted values. Any possibility of systematic error (Roy et al., 2017) is explored on the test set applying normality test (e.g. Kolmogorov-Smirnov). A normal distribution of residuals of test set with a mean close to zero is an indication of absence of systematic error.

Another set of criteria were developed by Golbraikh et al. (2002) for model validation. Models are considered to have acceptable prediction power, if they satisfy all of the following conditions:

I. $R_{cv}^2 > 0.5$ (2.2)

II. $R^2 > 0.6$ (2.3)

III. R_0^2 or $R_0'^2$ close to R^2

i.e.: **(a)** $(R^2 - R_0^2)/R^2 < 0.1$ and $0.85 \leq k \leq 1.15$ (2.4)

or

(b) $(R^2 - R_0'^2)/R^2 < 0.1$ and $0.85 \leq k' \leq 1.15$ (2.5)

IV. $|R_0^2 - R_0'^2| < 0.3$ (2.6)

Where R^2 is the coefficient of determination, *k* and *k*' are slopes, R_0^2 (predicted vs. observed) and $R_0'^2$(observed vs. predicted) are coefficients of determination without intercept. The first condition investigates the robustness of the model, while the second one checks the fit. The third and the fourth conditions analyze the closeness of the regression line to y=x. Concordance correlation coefficient (CCC) has been proposed by Lin (1989; 1992) and later revaluated for QSAR validations by Chirico and Gramatica (2011). This parameter is used of fitting of predicted vs. observed points onto y=x and any divergence of the regression line from the concordance line (the slope 1 line passing through the origin) gives as a value of CCC smaller than 1. They noted that this coefficient measures both precision and accuracy, and evaluated the possibility of using this coefficient in lieu Golbraikh and Tropsha criteria.

2.7. Applicability Domain

The third principle of OECD validation criteria states the necessity of "a defined domain of applicability" for a model. The presence of outliers in any model can significantly change its predictive power. Moreover, they could highlight the drawbacks of the model and also could help explaining mechanisms of those chemicals. There is a variety of methods to define outliers for identifying both poorly predicted chemicals and the chemicals that are significantly distant from the other chemicals in the data set. Defining the applicability domain (AD) provides both the applicable space of a model and the detection of outliers. The most common approaches to define AD are ranges in the descriptor space, range of the response variable, geometrical methods, distance-based methods, and probability density distribution. Distance based methods include leverage, Mahalanobis distance (MD), City block, and Euclidean distance (ED) approaches (Roy et al., 2015b). Distance-based approaches calculate the distance from each point to a particular point in the data set. A summary of calculation of these distances were summarized in a study by Javorska et al. (2005). Mahalanobis and Euclidean distances have been preferred as distance measures for the detection of structural outliers in nonlinear models (Martincic et al., 2015). For MD, structurally distant molecules can be identified with $MD^2 > \chi^2$ (chi-squared) (Rousseeuw and Leroy, 1987). If the data are normally distributed with n dimensions, the values follow a χ^2 distribution with n degrees of freedom. Chemicals having MD^2 exceeding the particular quantile of the χ^2 distribution are called structural outliers. For Euclidean distance, the chemical is defined as a vector and its distance is calculated to the model's centroid. The

threshold value for ED is defined as the largest distance in the training set of chemicals (Minovski et al., 2013).

The chemicals that are structurally very influential in determining the model parameters, i.e. creating leverage effect, are demonstrated in the Williams graph. This graph is obtained by plotting the hat (leverage) values versus standardized residuals. The standardized residual formula is given in Appendix Table A.1. The leverag564373e of a chemical provides a measure of the distance of the chemical from the centroid of its training set. If the vector of observed values is denoted by y and the vector of estimated values by $\hat{y}$, then $\hat{y}=Hy$, where H is the hat matrix. Diagonal elements of the hat matrix (H) are the leverage (hat) values such that h_{ii} is the leverage value of the i^{th} chemical. The hat matrix is given as $H=X(X^T X)^{-1} X^T$ where X is the design matrix consist of descriptors (Egan and Morgan, 1998). In this approach, if the hat value of a test set chemical is greater than the critical hat value (h^*), then the chemical is identified as a structural outlier. Critical hat value is set at $3p/n$, where p is the number of descriptors plus one and n is the number of chemicals in the model (Papa, et al., 2007). While the high-leverage ($h > h^*$) chemicals of training set chemical is structurally influential in the model, chemicals in test set with high-leverage are assumed to be extrapolated, and identified as structural outliers.

In the present study, hat values for linear models and Euclidean distance (ED) for nonlinear models were used to indicate the AD.

2.8. Risk Assessment Perspective

While acute and chronic toxicity tests on species are the traditional risk assessment methods, non-testing methods such as QSAR have become more and more inevitable considering the number of chemicals in use. The use of QSARs takes place in various applications of the regulatory assessment of chemicals including to support priority settings, to supplement in weight-of-evidence approaches, and to substitute animal experiments (Worth, 2010).

General procedure for regulatory assessment of chemicals involves the assessment of predicted environmental concentrations (PECs) and predicted no effect concentrations

(PNEC). PNEC is a level for which lower concentrations are considered to cause no adverse effects to the aquatic organisms. PNECs are derived using two different methods. The first one is obtained by dividing acute or chronic toxicity values (whichever is available) by the assessment factors. The assessment factor ranges between 10 and 10,000 depending on the toxic data available. When the available data are limited due to short term test results of three species from three trophic levels, the concentration is divided by 10,000. On the contrary, if abundant data are available from long-term test results and more taxonomic groups, the assessment factor is smaller (TGD, 2003). For the second method, if substantial amount of data are available, the species distribution is used. PECs are assessed either via monitoring or making predictions correlated to production and release of the contaminants. PNECs are then compared with the PECs in the environmental risk assessment. Hazard quotient (HQ) calculated as PEC/PNEC is a parameter used for risk characterization. An HQ greater than 1.0 shows a potential risk for the environment, and consecutively, further steps are taken (TGD, 2003).

Low-toxic effect concentrations (NOEC, LOEC, EC_x (x% effective concentration e.g. EC_{10}, EC_{20})) are used for the risk assessment purposes. NOEC is defined as the chemical concentration that causes no significant difference compared to the controls, and LOEC is the lowest observed effective concentration. These metrics are calculated via hypothesis testing. On the other hand, EC_x is calculated via point estimation. Arguments on which one is to be used started around the publication of the OECD report in 1998 (OECD, 1998). While some researchers criticize the use of NOEC and LOEC (Newman, 2008; van Dam et al., 2012; Landis and Chapman, 2011; Jager, 2012), others explain the drawbacks of EC_x (Christensen et al., 2009). ECHA prefers low effect percentiles (5-20%) over NOEC because they are considered to be more comparable metrics than NOEC (ECHA, 2008). Shieh et al. (2001) showed that although NOEC has drawbacks, it is a better parameter than EC_{10} for the protection of algae using Cd and Ni as toxicants. Iwasaki et al. (2015) analyzed the choice of NOEC and EC_x (e.g. EC_{10}, EC_{20}) in the calculation of Hazardous Concentration for 5% of species (HC5), and concluded that there was no profound effect. In addition to the parameter to be used, the variability of toxicity data ranges in the literature is another challenge on risk assessment (Koller et al., 2000). There is a paucity in chronic toxicity QSARs (Cronin, 2017). Moreover, the evaluation of relationship between NOEC and EC_x based on empirical data is essential (Beasley et al., 2015).

Low-effect concentrations and the acute median effective concentration (EC_{50}) are used for the calculation of the acute-to-chronic ratio (ACR) in risk assessment. There are two basic types of ACR. One is calculated with acute toxicity and NOEC values, the other is calculated with acute toxicity and maximum acceptable toxicant concentration (MATC) values. Due to the critics on NOEC, some researchers also used EC_{10} or EC_{20} as the chronic value (Kumar et al., 2016). ECOSAR defines chronic value (ChV) as the geometric mean of NOEC and LOEC values of the chemicals, which is also known as MATC (ECOSAR, 2012). The U.S. Environmental Protection Agency (EPA) defines ACR based on the ratio of acute LC/EC_{50} to the chronic MATC value or regression derived EC_{20} value (Hoff et al., 2010). EU Technical Guidance Document (TGD) defines ACR as EC_{50}/NOEC (TGD, 2003).

QSARs estimating chronic toxicity to aquatic organisms are limited in the literature. There have been efforts to correlate low-toxic-effect concentrations with properties of the chemicals (Chen et al., 2009; Austin and Eadsforth, 2014; Fan et al., 2016), median effective concentrations, and other species endpoints (Chen et al., 2009). Additionally, correlations between ACRs of aquatic organisms and properties of chemicals have been investigated (May et al., 2016). Considering that the mode of action (MOA) of a chemical may differ among species, QSARs for intraspecies are more reliable than interspecies models. Additionally, higher organisms have different types of chronic values using survival and reproduction as endpoint. Therefore, correlating algal chronic values with daphnia or fish chronic values is complicated. For instance, in a study by May et al. (2016), the authors concluded that fish and daphnia ACRs are not related. However, Chen et al. (2009) found a correlation between algae and fish NOECs. As a widely used chronic value estimator, US EPA ECOSAR, has been criticized by several researchers (Claeys et al., 2013). It estimates chronic value, but not NOEC or EC_{10}. Hence, the predictions are not compliant with regulation requirements. Another important point for chronic toxicity QSARs is validation. Given the fact that the majority of the QSAR models do not consider the OECD validation principles (OECD, 2007), there is a need for validated models.

In the present study, validated models for low-toxic-effect concentrations, namely, NOEC and IC_{20} were generated using algal toxicity data reported in the present study.

2.9. Interspecies Toxicity Predictions

Besides QSTR, interspecies toxicity prediction models are becoming an important tool for determining the toxicity of a chemical using interspecies relations. An important difference between QSTR and QTTR is that while theoretical descriptors are used in QSTRs, other species toxicities are used as descriptors in QTTRs. QTTRs have also a potential to fill the gaps where toxicity data are scarce. There have been studies on interspecies toxicity prediction in the literature to fill these gaps and also to understand the toxic mechanism of chemicals. From the aquatic environment, bacteria, ciliate, algae, daphnia, and fish toxicities were found to be correlated (Kar and Roy, 2010; Zhang et al., 2010; Aruoja et al., 2011; Singh et al., 2014; Furuhama et al., 2015). In a recent study by Kar et al. (2016), interspecies models were exhaustively reviewed. Their importance on toxic mechanisms, species-specific toxicities, and reduction on animal usage were summarized. Therefore, it is worth searching interspecies toxicity relationships to make toxicity predictions for algae using other species.

In the present study, algae-ciliate and algae-algae interspecies toxicity models were developed using MLR.

3. MATERIALS AND METHODS

3.1. Test Chemicals

In the present study, the selected chemicals for batch 96-h algal toxicity bioassay include chloro-, methyl-, -ethyl, methoxy-, and nitro- substituted phenols and anilines, and their names, ID and CAS numbers are listed in Table 3.1. together with their hazard warnings. The chemicals have been selected according to the following criteria. The chemicals should be:

- Phenol and aniline derivatives with no algal toxicity for *C. vulgaris*,
- Environmentally significant,
- Soluble enough for aquatic toxicity testing,
- Commercially available with at least 97% purity,
- Having a vapor pressure < 0.3 mm Hg (Not highly volatile).

The chemicals were purchased from Sigma-Aldrich (Taufkirchen, Germany) and Dr. Ehrenstorfer (Wesel, Germany) with the purity of minimum 97% (Appendix Table B.1). No further purification was undertaken. Majority of the stock solutions were prepared freshly for each experiment by dissolving the chemical in deionized water. The chemical stock solutions were sterilized through 0.2 μm (PVDF membrane with glass fiber) sterile filters. The chemicals with low solubility were prepared in dimethyl sulfoxide (>99.9 purity). The proportion of the solvent did not exceed 0.1 % (v/v).

Nominal and actual concentrations of each chemical was measured via either Gas Chromatography (GC, Agilent, 6890N equipped with an automatic sampler, split/splitless injection port and flame ionization detector) or UV-vis spectrophotometer (Lasany CADAS 200; Dr Bruno Lange GmbH) at the beginning of the experiments. Additionally, a test concentration of the tested chemical was analyzed in a separate test vessel without algae at the beginning and at the end of the experiments to check if there was a significant chemical loss (more than 20%) due to volatilization, adsorption on the test vessel etc. throughout the

experiment. UV-Vis spectrum was taken by scanning absorption of each chemical between 230 and 680 nm during experiments. The chemicals analyzed by spectrophotometer were listed in Appendix Table B.2. For GC analyses, actual concentrations were prepared in methylene chloride and a calibration curve was prepared to evaluate the concentration of stock solutions. Gas chromatography analysis were carried out with an HP-5 ms capillary column, which was 30-m long, with 0.25-mm inner diameter and 0.25-mm film thickness. Helium was used as the carrier gas at a constant flow rate of 33.3 cm/s. GC oven was programmed for an initial temperature of 40 ^{o}C for 1 min, increased to 140 ^{o}C for 10 ^{o}C/min, and then to 260 ^{o}C for 20 ^{o}C/min. The injector and detector temperatures (250 ^{o}C and 300 ^{o}C, respectively) were held constant during the analysis. The chemicals analyzed via GC analysis were listed in Appendix Table B.3.

In order to identify and prioritize hazardous chemicals for the protection of human health and the environment, the classification, labelling, and packaging (CLP) of substances and mixtures regulation (EC No 1272/2008) was released. In this regulation, chemicals are labeled with the Globally Harmonised System of Classification and Labelling of Chemicals (GHS) pictograms, such that they are easily classified. The pictograms related to the subject chemicals are given in Table 3.2. Related warnings to each chemical are also given in Table 3.1.

The previous 96-h algal toxicity test results of our laboratory (Ertürk, 2013) were merged with the new results for the modelling part of this work. Chemicals with ID numbers in Table 3.1. 63-92 belong to the previous work.

Table 3.1. The tested chemicals, chemicals from the previous study, their ID and CAS numbers, and hazard classifications.

ID	CAS No	Name	Warning
1	95-48-7	2-methylphenol	C-T
2	526-75-0	2,3-dimethylphenol	C-T-H-E
3	105-67-9	2,4-dimethylphenol	C-T-W-E
4	95-87-4	2,5-dimethylphenol	C-T-E
5	576-26-1	2,6-dimethylphenol	C-T-E
6	95-65-8	3,4-dimethylphenol	C-T-E
7	108-68-9	3,5-dimethylphenol	C-T
8	123-07-9	4-ethylphenol	C-W
9	150-76-5	4-methoxyphenol	H
10	5150-42-5	2,3-dimethoxyphenol	W
11	91-10-1	2,6-dimethoxyphenol	W
12	2033-89-8	3,4-dimethoxyphenol	W
13	500-99-2	3,5-dimethoxyphenol	W
14	697-82-5	2,3,5-trimethylphenol	C-W-T-E
15	527-60-6	2,4,6-trimethylphenol	C-W-T-H-E
16	533-73-3	1,2,4-trihydroxybenzene	C-W
17	824-46-4	Methoxyhydroquinone	W
18	95-71-6	Methylhydroquinone	C-H-W-E
19	700-13-0	2,3,5-trimethylhydroquinone	C-W-E
20	824-69-1	2,5-dichlorohydroquinone	C-W
21	504-15-4	5-methylresorcinol	W
22	6640-27-3	2-chloro-4-methylphenol	W
23	615-74-7	2-chloro-5-methylphenol	C-W-E
24	1570-64-5	4-chloro-2-methylphenol	C-T-E
25	59-50-7	4-chloro-3-methylphenol	C-W-E
26	88-04-0	4-chloro-3,5-dimethylphenol	W
27	88-75-5	2-nitrophenol	W-E
28	554-84-7	3-nitrophenol	C-H-W-T
29	100-02-7	4-nitrophenol	T-H-W
30	51-28-5	2,4-dinitrophenol	T-H-E-F
31	329-71-5	2,5-dinitrophenol	T-H-E
32	577-71-9	3,4-dinitrophenol	T-H-E
33	4920-77-8	3-methyl-2-nitrophenol	W
34	2581-34-2	3-methyl-4-nitrophenol	W
35	119-33-5	4-methyl-2-nitrophenol	W
36	2042-14-0	4-methyl-3-nitrophenol	W-T
37	700-38-9	5-methyl-2-nitrophenol	W
38	534-52-1	2-methyl-4,6-dinitrophenol	W-C-T-H-E
39	2423-71-4	2,6-dimethyl-4-nitrophenol	W
40	619-08-9	2-chloro-4-nitrophenol	W
41	89-64-5	4-chloro-2-nitrophenol	W
42	610-78-6	4-chloro-3-nitrophenol	W

Table 3.1. continued.

ID	CAS No	Name	Warning
43	618-80-4	2,6-dichloro-4-nitrophenol	W
44	95-55-6	2-aminophenol	W-H
45	591-27-5	3-aminophenol	W-E
46	123-30-8	4-aminophenol	W-H-E
47	95-84-1	2-amino-4-methylphenol	W-E
48	1687-53-2	5-amino-2-methoxyphenol	W
49	95-85-2	2-amino-4-chlorophenol	W-H
50	99-57-0	2-amino-4-nitrophenol	W-T
51	88-74-4	2-nitroaniline	T-H
52	99-09-2	3-nitroaniline	T-H
53	100-01-6	4-nitroaniline	T-H-E
54	97-02-9	2,4-dinitroaniline	T-H-E
55	618-87-1	3,5-dinitroaniline	T-H
56	603-83-8	2-methyl-3-nitroaniline	T-H-E
57	119-32-4	4-methyl-3-nitroaniline	T-H-E
58	121-87-9	2-chloro-4-nitroaniline	W-E
59	89-63-4	4-chloro-2-nitroaniline	T-H-E
60	635-22-3	4-chloro-3-nitroaniline	T-H-E
61	3531-19-9	6-chloro-2,4-dinitroaniline	T-H-E
62	4421-08-3	4-hydroxy-3-methoxybenzonitrile	W
63	108-95-2	Phenol	C-T-H-E
64	95-57-8	2-chlorophenol	W-E
65	108-43-0	3-chlorophenol	W-E
66	106-48-9	4-chlorophenol	C-W-E-T
67	576-24-9	2,3-dichlorophenol	W-E
68	120-83-2	2,4-dichlorophenol	C-T-E
69	583-78-8	2,5-dichlorophenol	C-W
70	87-65-0	2,6-dichlorophenol	C-W-E
71	95-77-2	3,4-dichlorophenol	C-W-E
72	591-35-5	3,5-dichlorophenol	C-W-T-E
73	15950-66-0	2,3,4-trichlorophenol	C-W
74	933-78-8	2,3,5-trichlorophenol	W-E
75	933-75-5	2,3,6-trichlorophenol	W
76	95-95-4	2,4,5-trichlorophenol	W-E
77	88-06-2	2,4,6-trichlorophenol	W-H-E
78	609-19-8	3,4,5-trichlorophenol	C-W-E
79	4901-51-3	2,3,4,5-tetrachlorophenol	C-T-E
80	58-90-2	2,3,4,6-tetrachlorophenol	T-E
81	935-95-5	2,3,5,6-tetrachlorophenol	T-E
82	87-86-5	Pentachlorophenol	T-H-E
83	87-66-1	1,2,3-trihydroxybenzene	W-H
84	123-31-9	Hydroquinone	C-W-H-E
85	615-67-8	Chlorohydroquinone	C-W

Table 3.1. continued.

ID	CAS No	Name	Warning
86	87-87-6	Tetrachlorohydroquinone	C-W
87	120-80-9	Catechol	C-W-T-H
88	2138-22-9	4-chlorocatechol	C
89	13673-92-2	3,5-dichlorocatechol	W
90	108-46-3	Resorcinol	C-W-H-E
91	95-88-5	4-chlororesorcinol	W
92	137-19-9	4,6-dichlororesorcinol	W

Table 3.2. Hazard classification symbols (GHS pictograms) and their descriptions.

	Acute toxicity (oral, dermal, inhalation), categories 1,2,3 (T)		Health hazard (H)
	Acute toxicity (oral, dermal, inhalation), category 4 (W)		Flammable liquid (F)
	Hazardous to the aquatic environment (E)		Corrosive (C)

3.2. Growth Inhibition Tests with *Chlorella vulgaris*

Algal growth inhibition tests were performed in batch cultures according to the standard test procedures (OECD No: 201, 2011) using the freshwater algae *C. vulgaris*. The parent cultures of *C. vulgaris* strain (CCAP 211/11B) were purchased from Culture Collection of Algae and Protozoa – (CCAP, The Scottish Association for Marine Science, Scottish Marine Institute, Dunbeg, Argyll, UK).

The population density was quantified spectrophotometrically at 680 nm (Lasany UV-VIS Spectrophotometer Double Beam Variable Band Width LI-2804). A linear relationship was found via correlating the algal cell counts with the absorbance using 56 data points (Appendix Figure C.1). After the range finding tests, the definitive toxicity tests were carried out using three replicates of the five different concentrations of the chemicals. The test conditions and other relevant information for the algal growth inhibition assays are summarized in Table 3.3. Tests were carried out in a climate room (Figure 3.1) with a fixed temperature and lighting described in test conditions.

Table 3.3. Test conditions for incubation and toxicity tests.

Test type	Static non-renewal
Test organism	*C. vulgaris*
Starting inoculum	10^5 mL^{-1}
Temperature	24 ± 0.5 ºC
Light quality	Cool white fluorescent lighting
Light intensity	60 µmol photons m^{-2} s^{-1}
Photoperiod	Continuous illumination
Test chamber size	500 mL
Test solution volume	100 mL
Replicates	3
Agitation	Daily by hand
Test concentrations	Five and a control
Test duration	96 hours
Endpoint	Growth (optical density at 680 nm)

Figure 3.1. A view from the climate room where toxicity assays were conducted.

The validity of the bioassays was assessed based on the OECD criteria (2006). Although the test duration is recommended as 72 hours, this period could be modified to either 48 or 96 hours. At the end of the test period, the population in the controls should increase at least 16-fold. The increase in the *p*H of the test medium during the test should not be greater than 1.5 units, preferably be 0.5. Additionally, the coefficient of variation should be less than 10% among the controls at the end of the test. Algal toxicity of each chemical was determined by statistical analysis of the average specific growth rate as the

response variable. Percent inhibition relative to the control growth rate was fitted against the test substance concentration, and the inhibitory concentration that reduces the response variable by 50% (IC_{50}) was calculated using the ICp method as executed in ToxCalc software (v. 5.0.32, Tidepool Scientific, CA, USA). The growth medium used in the experiments for *C. vulgaris* is bold basal medium with 3-fold nitrogen and vitamins as given in by Culture Collection of Algae and Protozoa – (CCAP, The Scottish Association for Marine Science, Scottish Marine Institute, Dunbeg, Argyll, UK) (Table 3.4).

Table 3.4. Bold basal medium with 3-fold nitrogen and vitamins used in bioassays.

Chemical	**Concentration (mg L^{-1})**
$NaNO_3$	750
$CaCl_2.2H_2O$	25
$MgSO_4.7H_2O$	75
$K_2HPO_4.3H_2O$	75
KH_2PO_4	175
NaCl	25
$FeCl_3.6H_2O$	0.97
$MnCl_2.4H_2O$	0.41
$ZnCl_2$	0.05
$CoCl_2.6H_2O$	0.02
$Na_2MoO_4.2H_2O$	0.04
Na_2EDTA	7.5
Thiaminehydrochloride	12
Cyanocobalamin	10

3.3. Modeling Methods

3.3.1. Calculation of low-toxic-effect and median inhibitory concentrations

The dependent variable of the toxicity models was determined following the algal responses to the tested chemicals. After a 96-h exposure period, the inhibitory concentrations of the chemicals that result in a 50% reduction (IC_{50}) in the average specific growth rate of algal cultures was assessed. Specific growth rate μ (d^{-1}) was calculated as in Equation 3.1.

$$\mu = \frac{X_2 - X_1}{(t_2 - t_1)X} \tag{3.1}$$

Where X_1 and X_2 are the final and initial populations on days t_1 and t_2, respectively.

Flowchart for the statistical analysis of algal growth response data is given in Figure 3.2. The procedure to analyze the toxicity data was as follows: Once the raw data are entered as the mean cell count per replicate after 96 h, the readings taken from blank samples are subtracted and cell enumeration is calculated as cells mL^{-1}. Hypothesis testing is then used to detect statistically significant differences between treatments. This requires that the basic assumptions of hypothesis testing, i.e. that the observations within treatments are independent and normally distributed and that the variance is homogenous across all concentrations and controls, be validated. Normality was tested using the Shapiro-Wilks test and homogeneity of variances is tested with the Bartlett's test. If these two assumptions are violated, then the data must be transformed using a log transformation and the assumptions must be tested again. If the data pass the normality and homogeneity of variance tests, a parametric multiple comparisons test, e.g. Dunnett's test, was applied. If the data fail the normality and homogeneity of variance tests with original and transformed data, then a non-parametric method such as Steel's Many One Rank test or Wilcoxon Rank Sum test was performed. The 96-h IC_{50} values were calculated using the point estimate techniques while LOEC and NOEC values for growth inhibition were obtained using a hypothesis testing approach such as Dunnett's procedure or Steel's Many-one Rank Test (EPA, 2002). Low-toxic-effect concentrations (LOEC, NOEC, and IC_{20}) for the same duration were used in the calculation of acute to chronic toxicity ratio (ACR) (Section 3.4).

96-h IC_{20} and 96-h IC_{50} with associated 95% confidence intervals were calculated using linear interpolation combined with bootstrapping (ICp) method as executed in ToxCalc (v. 5.0.32, 2009, Tidepool Scientific, USA). The response variable, the average specific growth rate, was calculated for each chemical as recommended in the OECD guideline (2011). Decimal logarithm of the reciprocal IC_{50} values in mM ($\log (1/IC_{50})$) determined at the end of 96 h, denoted as pT, was used as the dependent variable to construct algal QSTRs.

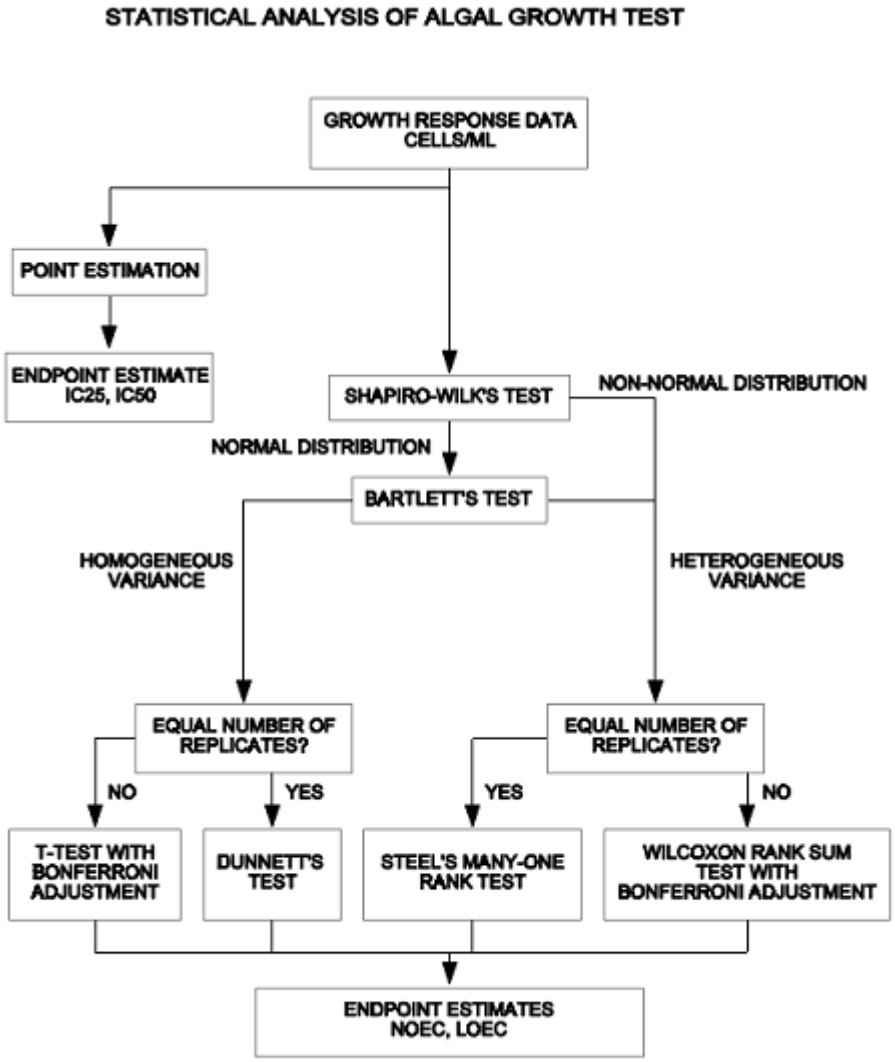


Figure 3.2. Flowchart for the statistical analysis of algal growth response data (EPA, 2002).

3.3.2. General procedure for QSTR modeling

Flowchart of QSTR modeling used in the present study is given in Figure 3.3. After toxicity values were calculated, molecular optimization was performed considering the lowest aqueous energy of the molecule. The descriptors for all chemicals were calculated with this energy optimized molecule. The data set was divided into training and test sets so that training set of chemicals were normally distributed and the test set was representative of the training set. Models were developed using the training set with the calculated descriptors. Descriptor selection was made with All Subsets method. The developed model was internally validated and tested using the test set. If the test set passed the external valida-

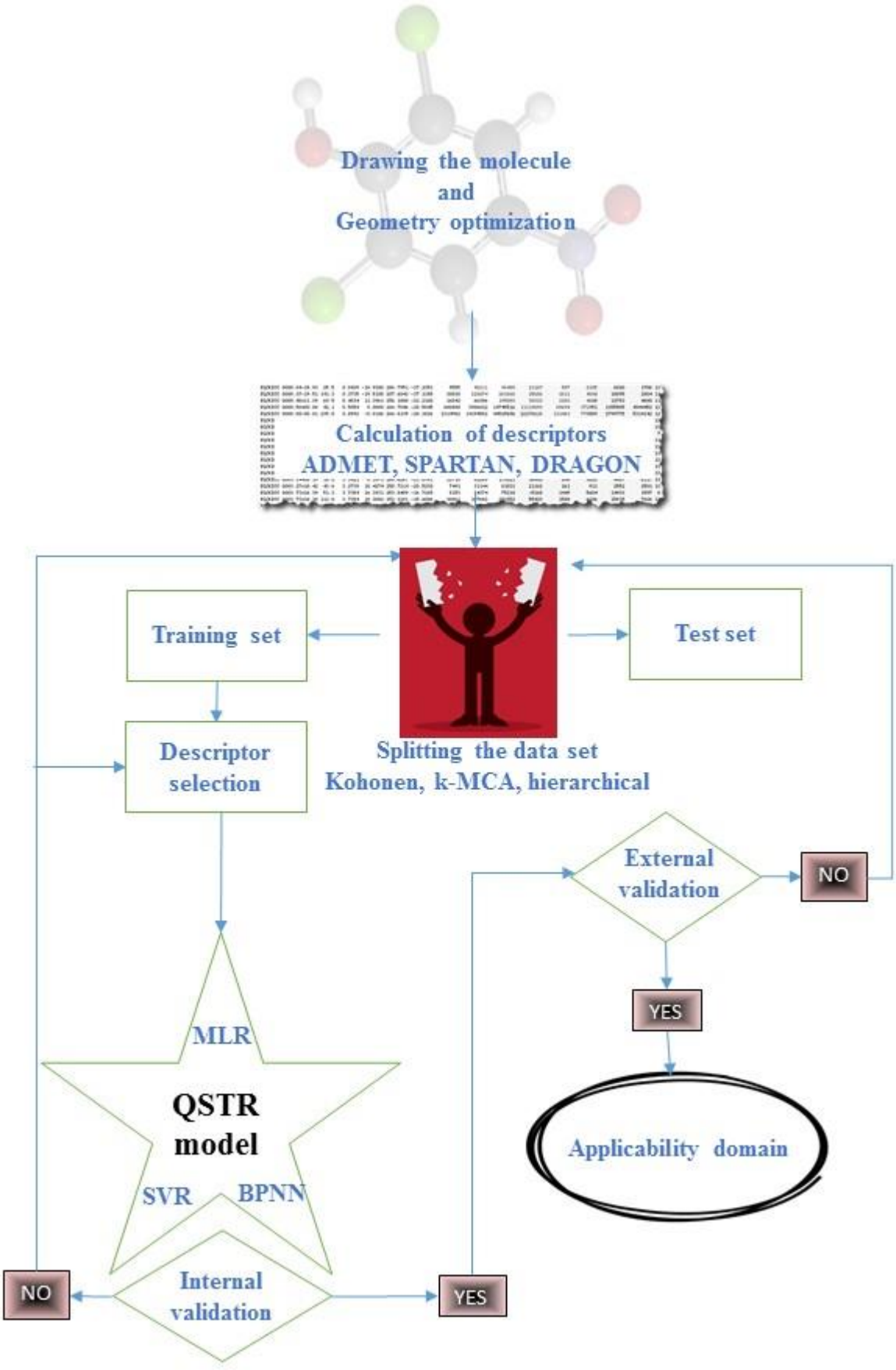


Figure 3.3. Flowchart of QSTR modeling.

tion criteria, the applicability domain was defined. The selected model was further tested on an external set. Additionally, predictivity of the model was explored on chemicals with no toxicity data.

3.3.3. Molecular descriptors

The logarithm of the 1-octanol/water partition coefficient (log K_{ow} or log P) of the chemicals, representing the hydrophobicity, was obtained from ECOSAR (v. 1.1, 2011, United States Environmental Protection Agency). The distribution coefficient (log D) of the chemicals were obtained from Danish (Q)SAR Database. For both log P and log D, the measured values were used whenever available.

The geometry optimization of the molecules prior to descriptor calculation was carried out using the semi-empirical PM6 method in SPARTAN software (v.10, 2011, Wavefunction, Inc., Irvine, California, USA). Semi-empirical molecular descriptors namely gaseous phase energy (E) (eV), aqueous phase energy (E_{aq}), highest occupied molecular orbital energy (E_{HOMO}), lowest unoccupied molecular orbital energy (E_{LUMO}), dipole moments (debyte), hardness, softness, electrophilicity, E_{LUMO} - E_{HOMO} gap, CPK volume ($Å^3$) and CPK area ($Å^2$) of the molecules were calculated (Tugcu et al., 2012). The theoretical molecular descriptors were calculated using the minimum aqueous energy conformation of the molecule. The first group of the theoretical descriptors were calculated using DRAGON software (v.6, 2010, Talete, Milano, Italy). The principal descriptor groups include: Constitutional indices, topological indices, connectivity indices, 2D matrix based descriptors, ETA indices, 3D descriptors, functional group counts, molecular properties, etc. Additionally, ADMET Predictor v.8.0.4.6 (Simulations Plus, Inc, USA) descriptors were calculated. The input files (MDL MOL format) were obtained via converting .mol2 files to .mol files using GaussView (v.3.0, Gaussian Inc., Pittsburgh, USA).

3.3.4. Training set/ test set division

The data set were divided into training and test sets using periodical division, Kohonen networks, hierarchical clustering, and k-MCA methods. In periodical division, only chemical toxicity values were considered. The toxicity values were sorted and in a periodical manner,

every third or fourth chemical was selected for the test set. The most and the least toxic chemicals were allocated into the training set. Division of the data set with Kohonen networks were performed via Kohonen and CPANN Toolbox for MATLAB (Ballabio et al., 2009, Ballabio and Vasighi, 2012). The test set was selected such that the test set contains chemicals coming from each neuron. Hierarchical clustering was performed selecting between-groups linkage cluster method with squared Euclidean distance using SPSS (SPSS Inc., 2008). For k-MCA method, the chemicals in the test set were selected such that the test set contains chemicals coming from each cluster. While neurons/clusters with more chemicals contribute more on the test set, less chemicals were selected from sparingly occupied neurons/clusters. Difference between Kohonen network and the k-MCA is that while clusters are independent from each other, neighboring neurons in a Kohonen network comprise similar chemicals. After the division of the data set, the training set was checked for normality via Kolmogorov-Smirnov test using SPSS software (SPSS Inc., 2008). Training sets those were not normally distributed were not modeled.

3.3.5. Selection of descriptors and modeling

The significant descriptors to be used in the model development were selected via All Subsets and GA. Highly inter-correlated (R >0.9) descriptors were not used in the model development. QUIK rule parameter (Delta K) was set at 0.05 to test collinearity and eliminate models with collinear descriptors.

QSTR models for *C. vulgaris* were developed using linear and nonlinear methods namely MLR, and BPNN, and SVM, respectively. The linear model was obtained using QSARINS program (Gramatica et al., 2013; Gramatica et al., 2014). The nonlinear models were developed using Molegro Data Modeller (MDM).

In order to obtain chemicals for external validation, the data set was divided into a training set and a test set using the methods described in Section 3.3. For the linear models, coefficient of determination (R^2), adjusted (for degrees of freedom) coefficient of determination (R^2_{adj}), Fischer statistics (F), and standard error of the model (SE) were calculated. Internal validation of all models was tested with the leave-one-out (LOO) procedure and cross validated leave one-out (Q^2_{LOO}) parameter for each model was

calculated. Collinearity between the independent variables in each of the generated models were checked with Kxx (<0.05). The reliability of MLR models was also tested by response randomization (Y-scrambling) procedure. For model randomization, the dependent variables of the training set were shuffled and new correlation coefficients were calculated (R_{ys}^2). The process was repeated at least 5000 times. The significantly low correlation coefficients of the new models indicated that the originally proposed model was not obtained by chance correlation. Y-scrambling procedure were run in QSARINS program with randomly generated 5000 models. For training and test sets of all models Root Mean Square Error ($RMSE_{tr}$, $RMSE_{test}$, respectively) were calculated.

The same divisions and the same descriptors used in linear models were employed in the nonlinear models (SVR and BPNN). R^2, Q_{LOO}^2, and $RMSE_{tr}$ parameters for training set were used for the evaluation of fit and robustness of nonlinear models. The selection of kernels and parameters is an important factor for the performance of SVR models. Radial basis function kernel was preferred due to its effectiveness and speed in training processes for regression studies (Panaye et al., 2006). Then, *epsilon* and *nu* models were experimented using the same parameters for comparison. After the selection of model type, other parameters such as number of support vectors and cost were optimized using Fine-tuning Optimization method via the grid-based search implemented in MDM software (2.6.0, Molegro ApS, 2011). This tool searches combinations of parameters to be fine-tuned. Similar to SVR models, the best BPNN models were searched via fine-tuning of learning rate, maximum number of training epochs, and number of neurons in the hidden layer. Relevance scores belong to BPNN models calculated to find the relevance of each descriptor in the model. The coefficients and coordinates of support vectors of SVR models were reported. These models were validated on the corresponding test set. After the fine-tuning, the best model were selected considering R^2, Q_{LOO}^2, and $RMSE_{tr}$, and validation results on the test set.

The statistical quality of all models were evaluated and compared using their goodness-of-fit (R^2), robustness (Q_{LOO}^2), interpretability (the meaning and relevance of descriptors), and predictivity (parameters on test set). To evaluate the predictivity of the models, the following cut-off values recommended by Chirico and Gramatica (2012) were used to reflect a successful model: Q_{F1}^2, Q_{F2}^2, and Q_{F3}^2 > 0.70; r_m^2 > 0.65, and concordance correlation

coefficient (*CCC*) > 0.85 for the test set. Golbraikh and Tropsha criteria (Golbraikh et al., 2003) were also considered for the acceptance of models. Any possibility of systematic error was explored on the test set applying normality test.

3.3.6. Applicability domain

A chemical was identified as a response outlier in MLR models if its predicted value was higher than three standardized residuals. Additionally, chemicals structurally very influential in determining model parameters, i.e., creating leverage effect, were demonstrated in the Williams plot. Chemicals in the test set that were predicted due to extrapolation of the model (i.e., fall outside the applicability domain) were detected when their leverage (hat) values were greater than the critical hat value (h^*) (Section 2.7). The ADs of the linear models were verified by using the descriptor space and the toxicity values, and also the leverage approach.

Applicability domain of nonlinear models were described via standardized residuals and Euclidean distance of molecules. Similar to Williams plot, but X-axis had Euclidean distances instead of hat values. X-axis cut-off value was defined at the largest distance for training set chemical of the model.

The coverage of each model was defined with the model's AD. For this purpose, chemicals that do not have experimental values, were predicted using the developed models. Then, predicted toxicity values against leverages/Euclidean distances were plotted for chemicals whose descriptors values were within the defined ranges. The chemicals whose predicted toxicities and leverages/Euclidean distances lie within the AD were accepted as successful predictions.

3.4. Calculation of Acute to Chronic Toxicity Ratio and Modeling of Low-Toxic Effect Concentrations

Of the studied 84 chemicals, 60 have NOEC values in total together with 16 chronic values retrieved from Ertürk (2013). NOEC, LOEC, and IC_{20} values of these 60 chemicals were calculated using ToxCalc software. MATC of chemicals was calculated as geometric

mean of NOEC and LOEC as described in ECOSAR methodology document (2012). Three types of acute to chronic ratio (ACR) were defined for the comparison with the literature using these endpoints and 96-h IC_{50}: (1) $ACR_{MATC}=IC_{50}/MATC$ (Hoff et al., 2010), (2) $ACR_{20}=IC_{50}/IC_{20}$ (Hoff et al., 2010), and (3) $ACR_{NOEC}=IC_{50}/NOEC$ (EC, 2003). The calculated acute and chronic toxicity values were compared with the studies in the literature, ECOTOX database, ECHA database, predictions of Danish (Q)SAR database and ECOSAR.

For the modelling part of the chronic toxicity, all types of ACRs, NOEC, and IC_{20} values were considered as dependent variable. SPARTAN (v.10), DRAGON (v.6.0), and ADMET Predictor (v.8) programs were used for the calculation of descriptors as described section 3.3.3. The chemicals were divided into training and test sets using periodical division, Kohonen networks, hierarchical clustering, and k-MCA methods. Two types of models were developed for each endpoint; one with the empirical descriptor (96-h IC_{50}) and the other with theoretical descriptors. The statistical quality of all models were evaluated and compared using their goodness-of-fit (R^2), robustness (Q^2_{LOO}), interpretability (the meaning and relevance of descriptors), and predictivity (parameters on test set). Additionally, *SE* and *F* calculated. To evaluate the predictivity of the models, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, r^2_m, and *CCC* parameters were calculated. Golbraikh and Tropsha criteria (Golbraikh et al., 2003) were also considered for the acceptance of models. AD of models defined using Williams plot. Further external validation was performed on chemicals those have unbounded NOEC values.

3.5. Compilation of Toxicity Data from Databases and the Literature

Toxicity assay results as well as predicted toxicity values from all models generated in the present study for studied chemicals were compared with those obtained from the literature. US EPA ECOTOX database (ECOTOX, US EPA) (Green algae, 24-96 h EC_{50} , IC_{50}, IC_{10}, IC_{20}, NOEC, and LOEC values), and ECHA Database (https://echa.europa.eu/information-on-chemicals), and ECETOC databases of EU (2003) were searched for algal toxicity values of studied chemicals. Toxicity data from Fu et al. (2015) were considered as a comprehensive source. Since *C. vulgaris* toxicity data for tested compounds lack the same conditions in the literature, algal toxicity predictions from this

study were also compared to US EPA ECOSAR 1.1 (v.1.1, US EPA, 2011) and Danish (Q)SAR database (http://qsar.food.dtu.dk) for tested compounds.

3.6. Interspecies Toxicity Relationships

To search for interspecies toxicity relationships, toxicity data were compiled from the literature for the studied chemicals on ciliate and algae, namely *Tetrahymena pyriformis* and *Pseudokirchneriella subcapitata*, respectively. Although there are various data sources, single source data were preferred for the sake of data quality. In this respect, the acute toxicity data producing a 50% growth inhibition ($pT_{T.pyriformis}$= -log IGC_{50}) on *T. pyriformis* (40-h) were retrieved from Cronin et al. (2002). 64 common chemicals of the *T. pyriformis* and the *C. vulgaris* sets were considered for the modeling purpose. $pT_{T.pyriformis}$ values ranged between -0.652 and 2.710 (as mM). *p*T range of *C. vulgaris* data was between -0.60 and 2.34 (as mM).

72-h algal *(P. subcapitata*) toxicity data ($pT_{P.subcapitata}$= -log EC_{50}) were taken from Aruoja et al. (2011). There are 23 common chemicals between *C. vulgaris* and *P. subcapitata* data sets. $pT_{P.subcapitata}$ values ranged between -0.321 and 1.941 (as mM). *p*T range of *C. vulgaris* data was between -0.60 and 1.86 (as mM).

Since *C. vulgaris* toxicity data are scarcer than the other species data in this part, *C. vulgaris* toxicity was considered as dependent variable. The models were developed without involving additional descriptors. The data sets were divided into training and test sets via periodical division. The most and the least toxic chemicals were allocated into the test set. Linear models were developed and validated according to OECD principles. The statistical quality of all models were evaluated using their goodness-of-fit (R^2), robustness (Q^2_{LOO}), and predictivity (parameters on test set). Additionally, *SE* and *F* calculated. To evaluate the predictivity of the models, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, r^2_m, and *CCC* parameters were calculated. Golbraikh and Tropsha criteria (Golbraikh et al., 2003) were also considered for the acceptance of models. AD of models was defined using Williams plot.

4. RESULTS AND DISCUSSION

4.1. Toxicity of the Selected Chemicals to *Chlorella vulgaris*

Standardized static 96-h algal growth inhibition tests for 62 chemicals were performed using freshwater green algae *C. vulgaris*. During the bioassay, the algal biomass increased by at least 16-fold and the coefficients of variation for the control growth rates were below 10% at the end of experiments. The *p*H of the medium was 6.35 (±0.1) at the beginning of the assay, and it did not rise more than 0.5 units. The toxic chemicals were tested in five different concentrations (Figure 4.1). The toxicity assay for a reference chemical, 3,5-dichlorophenol, to *C. vulgaris* was performed to compare with the ring test (ISO, 2004). The toxicity of 3,5-dichlorophenol was found to be 3.3 mg L^{-1} (±0.2), which coincides with the results of international standards.

Figure 4.1. Fading color of algal cultures with increasing chemical concentration.

Of the studied chemicals, 96-h IC_{50} and IC_{20} values of 54 chemicals were determined (Table 4.1) using the ICp method in the ToxCalc software. 96-h algal toxicity values of eight chemicals in the data set could not be determined due to the reasons explained below. The acute and low-toxic effect concentrations of the studied chemicals are given in Table 4.1., together with their expected mode of actions (MOA). In the present data set, phenols include

four different MOAs (polar narcotic, respiratory uncoupler, pro-electrophile, and soft electrophile), whereas anilines include three MOAs (polar narcotic, respiratory uncoupler, and pro-electrophile). NOEC and LOEC were estimated by using Dunett's test or Steel's many-one rank test. During the range finding tests of 4-aminophenol and 5-amino-2-methoxyphenol, the test medium color turned into black and precipitation occurred. Both the dark color of the medium and the insoluble form of the chemical did not allow attaining a toxicity value. Therefore, toxicity tests for these two chemicals could not be performed. The highest concentration tested for 2,6-dimethoxyphenol during range finding was 600 mg L^{-1}. Since the chemical was not toxic to *C. vulgaris* (IC_{50} > 600 mg L^{-1}), further experiment was not done for this chemical. 2,3,5-trimethylhydroquinone was not soluble enough for algal toxicity testing. Therefore, a solvent was used for toxicity assay. However, a significant difference was observed during range finding tests with and without solvent. Hence, toxicity test was not conducted. The magnitude of the inhibition was insufficient for calculation an IC_{50} value due to the solubility limits of 4-nitroaniline, 2,3-dimethoxyphenol, 2,6-dimethoxyphenol, 3,4-dimethoxyphenol, and 2-amino-4-nitrophenol.

The change in the nominal and final concentrations of each chemical during the 96-h assay was analyzed using either GC or spectrophotometer as described in Section 3.1. After testing the nominal and final concentrations of each chemical during the bioassay we observed that all chemicals, except methylhydroquinone, had concentrations within the 20% nominal concentration. This chemical was found to decrease to 72.5% of the initial concentration. Therefore, its actual concentration was recalculated. The concentration of other chemicals was considered as nominal concentrations.

Among the tested chemicals, the least toxic chemical is 3-aminophenol (*p*T: -0.68) and the most toxic one is 4-chloro-2-nitrophenol (*p*T: 1.85). The toxicity of studied chemicals was dependent on the type and the number of the substituents. In general, the toxicity was higher when the number of substituents increased. Chloronitrophenols were more toxic than chloromethylphenols. Chloronitroanilines were more toxic than nitro- and methyl-substituted anilines. Dinitroanilines were more toxic than chloronitroanilines. Phenols were more toxic than anilines with the same substituents in the same positions.

The evaluation of toxicities of studied chemicals to algae indicated that their mechanisms vary depending upon the on the intrinsic reactivity pattern of the compounds. The toxicity of nitrophenols can be attributed to the fact that they act as uncoupling agents in oxidative phosphorylation. From the chemical viewpoint, the nitro group is a strong π-electron acceptor, lowering the electron density of the aromatic ring. Inside the nitro group, excess electronic charge is mainly localized at the oxygen atoms, while the nitrogen atom is typically electron-deficient. As a consequence, nitroaromatic compounds show enhanced reactivity for the attack of nucleophiles at aromatic ring carbons as well as for reactions with reducing agents, and in phenol derivatives the nitro substituent leads to a pronounced enhancement of the acidity of the OH group. Whether or not efficient oxidation/reduction cycle of the nitroaromatic compound exists depends on the balance between the oxidative and reductive pathways and probably also on further aspects such as physiological properties of the organism. Clearly, the variety of metabolic routes and bioactive agents formed from nitroaromatic compounds makes it difficult to decide which of the different modes of action will be of primary importance for a given test organism and endpoint. Aminophenols could be partially auto-oxidized during the experiment, leading to the formation of more or less toxic oligomers. 2-aminophenol is an example in the data set. The dinitro compounds that are possibly exert their toxicity in their reduced form are likely to have an increase in toxicity.

4.1.1. Correlation of *C. vulgaris* toxicity with hydrophobicity

The ability of a chemical to reach to the active site of action and elicit its adverse effects on aquatic organisms are mainly controlled by the hydrophobicity of the chemical in question. In general, hydrophobicity of a chemical is described by the logarithm of the octanol–water partition coefficient, log *P*. However, in some cases, for the chemicals that partly or completely ionize at the test medium *p*H, using the *p*H corrected hydrophobicity (log *D*) might be a better idea to account for the ionization phenomena. For example, in a previous study, Ertürk and Saçan (2013) found that ionization of phenols in the *C. vulgaris* test system had a considerable impact on their toxicity to *C. vulgaris*. Accordingly, using log *D* instead of log *P* yielded better results in explaining the toxicity of polar narcotic phenols. Therefore, using log *D* instead of log *P* to describe the relationship between the algal toxicity of phenols and their hydrophobicity has been preferred.

Tested chemicals were examined according to their expected mode of actions (Table 4.1). Their toxicity values were plotted against their log *D* values in order to inspect any possible correlation (Figure 4.2). An obvious relationship between hydrophobicity and algal toxicity of polar narcotic phenols (blue marks) can be observed with the following equation (Eq.4.1).

$$pT = 0.7 \log D - 1.13 \quad (4.1)$$
$$n = 52, R = 0.86$$

As evidenced by the good agreement of many chemicals in the data set, hydrophobicity underpins this set of chemicals. However, the visual analysis also clearly points out that nitrophenols, in particular, dinitrophenols, and also polyphenols (except polar narcotic resorcinols) displayed toxicity in excess of that predicted by their hydrophobicity.

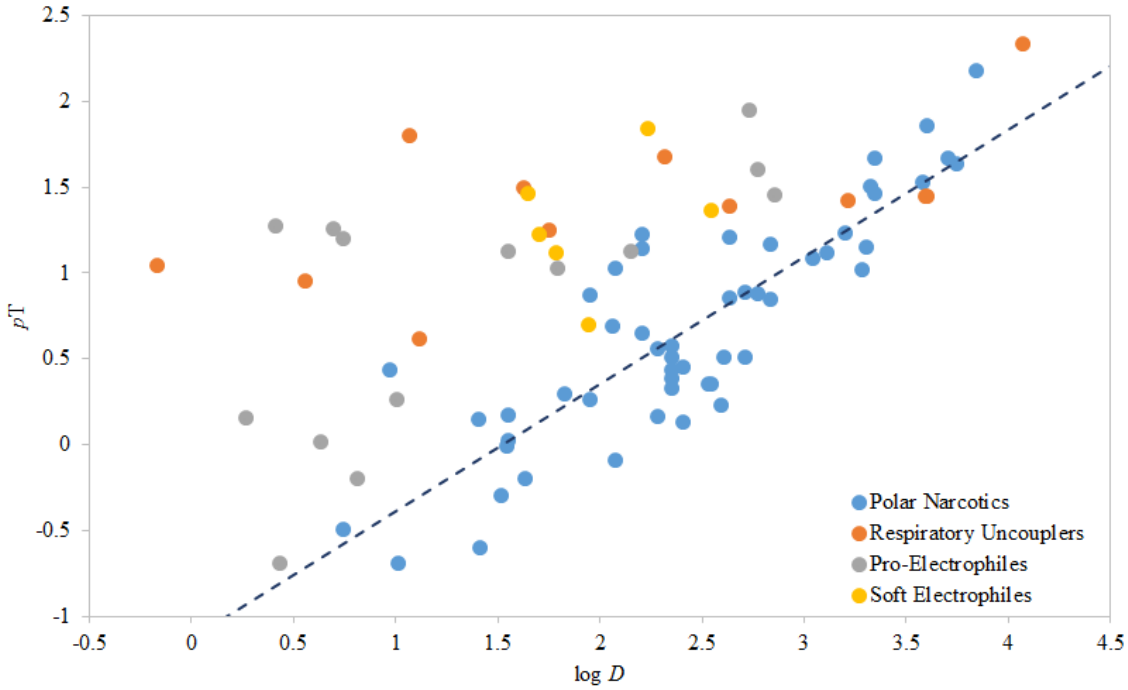


Figure 4.2. Relationships between log *D* and 96-h algal *p*T values of the tested chemicals.

According to this simple hydrophobicity model depicted by the line in Figure 4.2, the chemicals along the line can be assumed to elicit toxicity through polar narcosis. It is also possible to state that there should be other mechanisms than simple membrane perturbations

Table 4.1. Chemicals tested in the present study and previous study (Ertürk, 2013), their expected mode of actions (MOA), 96-h 50% and 20% inhibitory concentrations (IC_{50} and IC_{20}) with their 95% confidence intervals, NOEC and LOEC values (mg L^{-1}), and $pT=\log(1/IC_{50})$.

ID	MOA*	96-h IC_{50}	96-h IC_{20}	NOEC	LOEC	pT (mM)
1	1	131.4 (111.5-149.2)	100.6 (68.9-107.3)	25.0	50.0	-0.08
2	1	50.1 (44.1-58.1)	30.2 (19.1-36.8)	15.0	30.0	0.39
3	1	44.1 (39.3-52.5)	32.5 (30.5-35.7)	15.0	30.0	0.44
4	1	56.8 (48.8-62.1)	37.3 (29.1-62.7)	25.0	50.0	0.33
5	1	89.0 (74.1-101.2)	54.5 (43.6-64.7)	20.0	40.0	0.14
6	1	32.3 (29.5-35.5)	20.6 (12.5-24.6)	<20.0	20.0	0.58
7	1	37.4 (35.8-39.2)	25.8 (23.9-27.5)	<20.0	20.0	0.51
8	1	71.0 (68.3-73.8)	48.9 (46.2-51.1)	11.9	23.8	0.24
9	1	244.8 (232.6-258.9)	131.7 (81.0-171.4)	50.0	100.0	-0.29
13	1	238.4 (224.2-256.3)	129.9 (99.9-164.3)	50.0	100.0	-0.19
14	1	42.1 (39.2-43.7)	24.2 (20.6-26.3)	5.1	10.2	0.51
15	1	60.0 (57.9-64.4)	43.4 (41.1-45.1)	21.3	42.5	0.36
16	3	87.9 (80.2-91.2)	40.1 (27.6-46.3)	10.0	40.0	0.16
17	3	8.7 (8.3-9.2)	5.7 (5.2-6.1)	2.5	5.0	1.21
18	3	190.9 (159.3-214.1)	68.2 (42.79-89.7)	<14.5	14.5	-0.19
20	3	4.4 (4.0-5.3)	2.7 (2.2-3.1)	1.8	2.9	1.61
21	1	599.9 (545.6-648.7)	401.8 (280.5-500.2)	200.0	400.0	-0.68
22	1	44.1 (43.1-45.2)	30.9 (28.6-32.1)	3.8	7.6	0.51
23	1	18.2 (17.4-19.2)	8.3 (7.2-9.4)	2.5	5.0	0.89
24	1	19.9 (18.6-21.0)	12.4 (8.3-14.9)	6.0	12.0	0.86
25	1	9.6 (7.9-10.7)	6.1 (2.8-8.2)	3.0	6.0	1.17
26	1	11.0 (8.7-12.3)	5.6 (5.1-6.3)	2.0	4.0	1.15
27	4	10.5 (9.9-11.1)	3.1 (0.8-4.1)	2.5	5.0	1.12

Table 4.1. continued.

ID	MOA*	96-h IC_{50}	96-h IC_{20}	NOEC	LOEC	*p*T (mM)
28	4	27.6 (25.5-29.8)	15.2 (9.6-27.2)	6.0	12.0	0.70
29	4	8.2 (6.8-10.1)	5.4 (4.6-6.2)	2.5	5.0	1.23
30	2	16.6 (16.4-16.8)	13.1 (11.8-14.1)	2.5	5.0	1.05
31	2	2.9 (2.7-3.0)	1.9 (1.4-2.3)	1.0	2.0	1.81
32	2	43.7 (42.7-44.6)	26.7 (23.9-29.2)	<10.0	10.0	0.62
33	1	34.0 (32.1-35.8)	13.4 (11.4-15.5)	7.7	15.4	0.65
34	1	14.3 (13.2-15.5)	10.0 (7.3-11.6)	5.0	10.0	1.03
35	1	9.0 (8.0-9.9)	4.6 (3.9-5.2)	<1.5	1.5	1.23
36	1	41.5 (35.9-46.1)	30.2 (15.8-39.1)	14.9	29.9	0.57
37	1	10.9 (10.2-11.8)	7.4 (0.4-9.6)	2.0	4.0	1.15
38	2	7.9 (7.1-8.6)	4.0 (0-8.5)	<1.5	1.5	1.40
39	1	10.3 (9.8-10.7)	6.2 (5.1-7.1)	2.0	4.0	1.21
40	4	5.9 (5.9-6.0)	4.6 (4.4-4.7)	1.0	2.1	1.47
41	4	2.5 (2.2-2.8)	0.9 (0.7-1.1)	0.3	0.6	1.85
42	4	7.5 (6.6-8.0)	5.0 (4.6-5.4)	2.0	3.9	1.37
43	2	22.9 (22.4-23.3)	17.4 (16.6-18.0)	8.0	16.0	0.96
44	3	6.0 (5.5-6.5)	2.4 (1.9-2.9)	<1.3	1.3	1.26
45	3	527.7 (479.2-567.8)	215.3 (188.5-235.0)	75.0	150.0	-0.68
47	1	44.9 (38.9-49.8)	15.0 (4.4-27.7)	5.0	10.0	0.44
49	3	13.4 (12.4-14.4)	4.7 (4.4-5.1)	5.0	10.0	1.03
51	1	68.9 (63.7-72.6)	37.9 (32.9-41.2)	12.8	25.6	0.30
52	1	97.7 (91.9-103.8)	55.5 (48.6-62.5)	20.5	41.0	0.15
54	2	10.3 (9.8-11.1)	5.2 (3.8-7.3)	<1.8	1.8	1.25
55	2	5.7 (5.5-6.4)	3.7 (2.8-4.5)	2.1	4.1	1.50

Table 4.1. continued.

ID	MOA*	96-h IC_{50}	96-h IC_{20}	NOEC	LOEC	*p*T (mM)
56	1	141.7 (138.2-144.2)	106.4 (99.8-111.3)	49.3	98.5	0.03
57	1	101.5 (94.7-107.9)	29.1 (26.0-32.3)	<9.9	9.9	0.18
58	1	23.0 (21.5-25.1)	6.0 (5.7-6.3)	<2.5	2.5	0.87
59	1	23.7 (21.1-27.5)	11.0 (7.8-14.7)	3.0	6.1	0.86
60	1	34.9 (33.1-36.4)	18.3 (16.9-19.5)	3.3	6.6	0.69
61	2	4.5 (4.0-5.2)	2.6 (2.5-2.9)	1.0	2.0	1.68
62	1	149.5 (130.5-171.3)	66.4 (57.9-76.1)	17.6	35.3	0.00
63	1	374.9 (362.7-386.6)	276.3 (246.1-304.7)	120.0	240.0	-0.60
64	1	86.3 (79.6-93.0)	55.4 (26.6-70.1)	15.0	30.0	0.17
65	1	56.3 (54.1-58.3)	32.5 (29.7-35.1)	<5.0	5.0	0.36
66	1	44.9 (41.6-47.1)	24.9 (22.6-26.9)	5.0	10.0	0.46
67	1	13.3 (12.5-14.1)	5.5 (4.2-6.6)	<2.0	2.0	1.09
68	1	9.3 (8.3-10.3)	4.7 (3.7-5.4)	2.0	4.0	1.24
69	1	12.5 (11.6-13.3)	5.2 (2.7-6.5)	<2.0	2.0	1.12
70	1	21.5 (20.5-22.5)	7.9 (6.6-9.4)	<5.0	5.0	0.88
71	1	5.5 (4.9-6.0)	3.3 (3.0-3.6)	0.7	1.3	1.47
72	1	3.5 (3.3-3.6)	1.6 (1.9-1.7)	0.8	1.6	1.67
73	1	4.5 (4.2-4.8)	1.3 (1.0-1.7)	<0.5	0.5	1.64
74	1	2.7 (2.4-3.0)	1.3 (0.5-1.8)	<0.6	0.6	1.86
75	1	6.1 (5.8-6.3)	3.1 (2.8-3.3)	1.0	2.0	1.51
76	1	4.2 (3.9-4.4)	2.1 (1.6-2.4)	0.5	1.0	1.67
77	1	5.8 (5.2-6.4)	2.6 (2.1-3.0)	<1.0	1.0	1.53
78	1	1.3 (1.2-1.4)	0.5 (0.08-0.7)	<0.1	0.1	2.18
79	2	1.05 (1.0-1.1)	0.3 (0.2-0.4)	<0.4	0.3	2.34

Table 4.1. continued.

ID	MOA*	96-h IC_{50}	96-h IC_{20}	NOEC	LOEC	*p*T (mM)
80	2	8.3 (7.9-8.7)	5.5 (4.8-6.1)	2.5	5.0	1.45
81	2	8.6 (7.8-9.3)	4.7 (3.8-5.3)	2.0	4.0	1.43
82	2	9.4 (9.1-9.7)	4.8 (4.3-5.2)	1.0	2.0	1.45
83	3	6.6 (6.1-7.2)	3.0 (2.3-3.9)	<1.0	1.0	1.28
84	3	105.7 (99.0-112.2)	60.3 (42.7-83.5)	10.0	20.0	0.02
85	3	10.7 (9.6-11.7)	3.4 (2.4-5.1)	<1.0	1.0	1.13
86	3	8.5 (8.0-9.0)	3.8 (3.1-4.5)	<1.0	1.0	1.46
87	3	59.7 (55.6-63.4)	14.1 (11.5-16.6)	<10.0	10.0	0.27
88	3	10.6 (10.0-11.0)	5.0 (4.1-6.5)	1.5	3.0	1.13
89	3	2.0 (1.8-2.2)	0.9 (0.7-1.2)	<0.5	0.5	1.95
90	1	342.5 (308.9-375.3)	159.0 (101.7-228.9)	50.0	100.0	-0.49
91	1	78.5 (69.5-88.2)	46.3 (41.9-49.9)	20.0	40.0	0.27
92	1	16.9 (16.3-17.5)	11.5 (10.8-12.0)	5.0	10.0	1.02

*Expected MOA of chemicals are according to Cronin et al., 2002 and Schultz et al., 1998. 1: polar narcotic; 2: respiratory uncoupler; 3: pro-electrophile; 4: soft electrophile.

in action for the chemicals eliciting toxicity above the domain of polar narcosis. Based on the classification used by Cronin et al. (2002), the phenols selected for toxicological assessment in this study are expected to elicit toxicity either through polar narcosis or respiratory uncoupling or pro- or soft-electrophilic (Table 4.1). From a MOA perspective, Figure 4.2 reveals that classifying dinitrophenols, tetrachlorophenols, and pentachlorophenol in the same MOA (i.e., respiratory uncoupling) seems to be misleading as all of the chlorophenols fell within the domain of polar narcosis; whereas, dinitrophenols displayed excess toxicity than polar narcotics. As expressed previously by some researchers, although common functional groups imply the same mode of action, more often they do not share the same MOA (Netzeva et al., 2008). Similar trend of higher toxicity of dinitrophenols than mononitrophenols was also seen in mono- and dinitrobenzenes (Schmitt et al., 2000). Therefore, for future QSTR studies dealing with toxicity data obtained from *C. vulgaris* test system, a distinct MOA classification might be required. Based on our findings, classifying chlorophenols as polar narcotics might be more convenient from a QSTR perspective. Visual analysis also revealed that although 3-nitrophenol and tetrachlorohydroquinone were categorized as soft electrophile and pro-redox, respectively, these two chemicals were found to lie within the polar narcosis domain. The behavior of tetrachlorohydroquinone can be explained on the basis of its hydrophobicity because with increasing hydrophobicity, the toxicity of electrophiles and pro-electrophiles converges on a narcosis mechanism (Aptula et al., 2005). As for 3-nitrophenol, nitro group in the *meta* position is not very active (pK_a=8.36) as opposed to that in *ortho* (pK_a=7.23) or *para* (pK_a=7.15) position. Hence, this could be the reason why 3-nitrophenol also converged on a polar narcotic MOA.

Aruoja et al. (2011) stated that aniline toxicity for *Pseudokirchneriella subcapitata* does not depend on hydrophobicity. Anilines of our data set were analyzed to check if there was a correlation between polar anilines (8 chemicals) and hydrophobicity. Although polar narcotic aniline derivatives are not significantly correlated with log *P* (R=0.52), there was a relatively higher correlation with log *D* (R=0.66). In general, aromatic amines are classified as polar narcotics. However, they have excess toxicity for *Daphnia magna* (Urrestarazu Ramos et. al, 2002). The similar tendency was observed in polar narcotic anilines used in this study for *C. vulgaris*. Polar narcotic anilines in the present data set located above the polar narcosis line.

Consequently, the visual analyses revealed that majority of the phenols and anilines in the data set presented in this study are expected to elicit toxicity through polar narcosis. The toxicity of these chemicals can be predicted reliably with either log *P* or log *D*. On the other hand, for chemicals acting through more reactive mechanisms than polar narcosis, descriptors other than a hydrophobicity term or in addition to a hydrophobicity term are required to explain their toxicity to *C. vulgaris*. Therefore, it is worthwhile to search for descriptors that can be used to relate algal toxicity to chemical structures regarding the excess toxicity.

4.2. QSTR Models of the 96-h Algal Toxicity Data Set

QSTR models were developed to predict 96-h acute toxicities (*p*T= log (1/IC_{50}) mM). 84 chemicals were used in QSTR modelling step (Table 4.1). Log (1/IC_{50}) values spanned wide range (3.02 log unit) and were normally distributed (non-parametric Kolmogorov-Smirnov test, $p>0.05$) with a mean of 0.850 (Figure 4.3), and found to be appropriate for modelling. The log *P* values of these chemicals vary from 0.21 to 5.12, which allows developing hydrophobicity-based QSTRs.

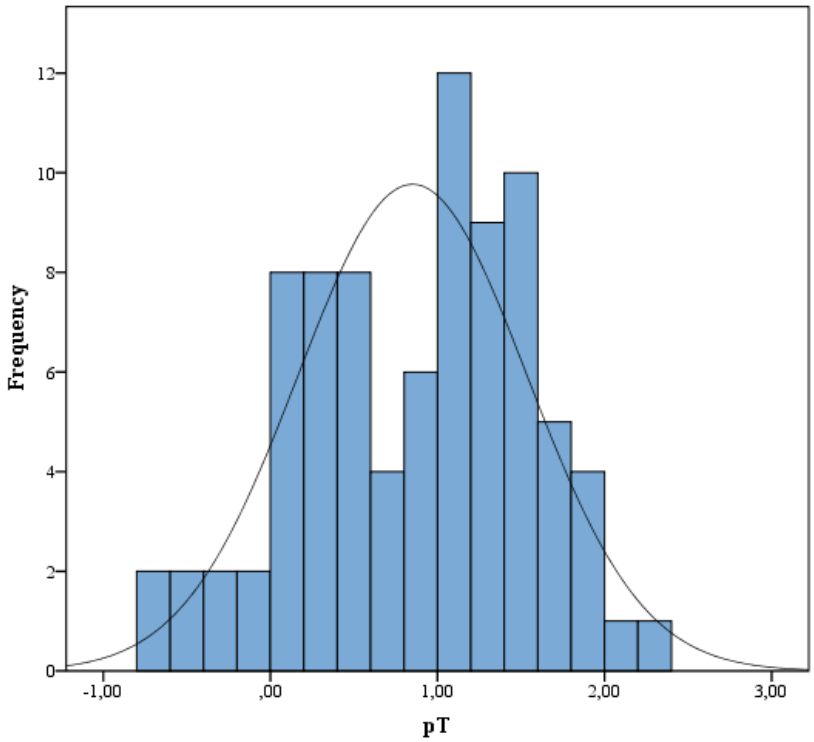


Figure 4.3. The histogram of *p*T values of studied chemicals.

The relationship between the toxicity and the hydrophobicity (Figure 4.2) suggested that an additional parameter is required to model the toxicity of all chemicals. The chemicals having excess toxicity (above the log *D* correlation line) need to have additional descriptors to log *D*. In order to develop a model to successfully predict algal toxicity of chemicals in the data set, the molecular structure of chemicals were optimized and the descriptors were calculated as described in Section 3.3.3. Then, the data set was split into training and test sets for the model development and validation purposes. To construct a representative test set to the training set, similar chemicals were grouped together using Kohonen networks, k-means clustering, and hierarchical clustering considering all 2818 descriptors.

Kohonen was used to find the most and the least similar molecules, independent of their activity, using Kohonen Toolbox (Ballabio et al., 2009; Ballabio and Vasighi, 2012). To this end, 3x3 and 4x4 normally bound maps with 100 epochs were used, respectively (Figure 4.4. a and b). The similar chemicals are located in the same neurons identified with their ID numbers.

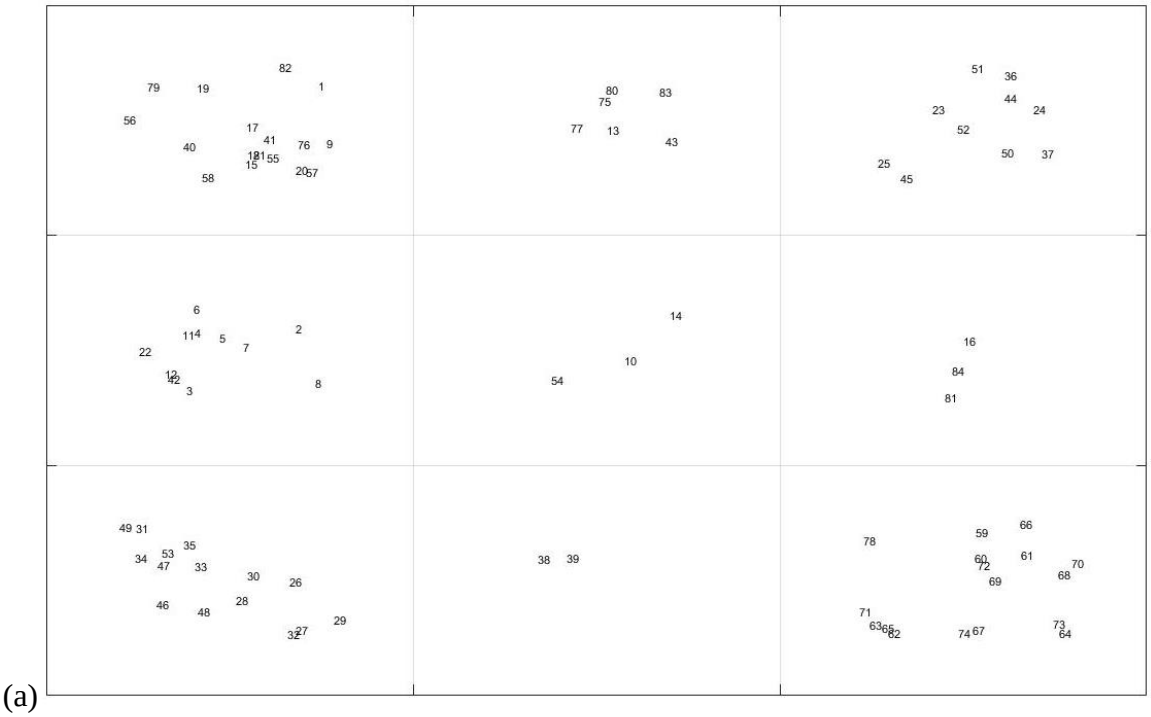
K-means clustering with maximum 10 iterations was applied to the descriptor set of chemicals. Chemicals were grouped into 5 clusters (Table 4.2). Hierarchical clustering was performed to group similar molecules. Considering their non-standardized descriptor values as variables, between-groups linkage cluster method with squared Euclidean distance has given the dendrogram in Figure 4.5. These groupings were performed in SPSS (v.17.0, SPSS Inc., 2008) software.

Table 4.2. K-means clustering lists.

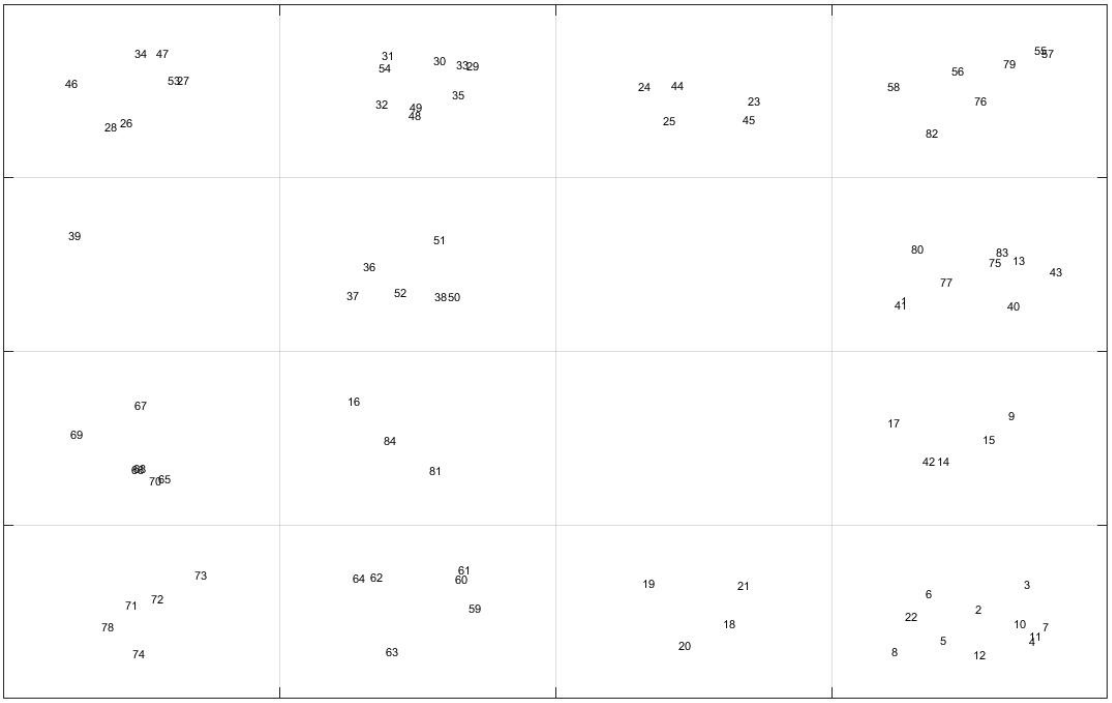
Cluster No	Chemicals
1	1,16,18,21-25,44,45,49,51,63-72,83-85,87,88,90,91
2	43,79-82,86
3	2-15,17,26,47
4	30-32,38,54,55,61
5	20,27-29,33-37,39-42,52,56-60,62,73-78,89,92

For periodical division, chemicals were sorted according to the toxicity values. The most and the least toxic chemicals were allocated into the training set and the remaining chemicals were selected, so that, every third or fourth chemical was selected for the test set. In this way, two sets were equivalent with respect to the molecular similarity (Kohonen map,

K-means clustering, and dendrogram) and response value equity. This assured that test set is a representative of the training set.



(a)



(b)

Figure 4.4. Kohonen top map of the chemicals. (a) 3x3 map and 100 epochs; (b) 4x4 map and 100 epochs.

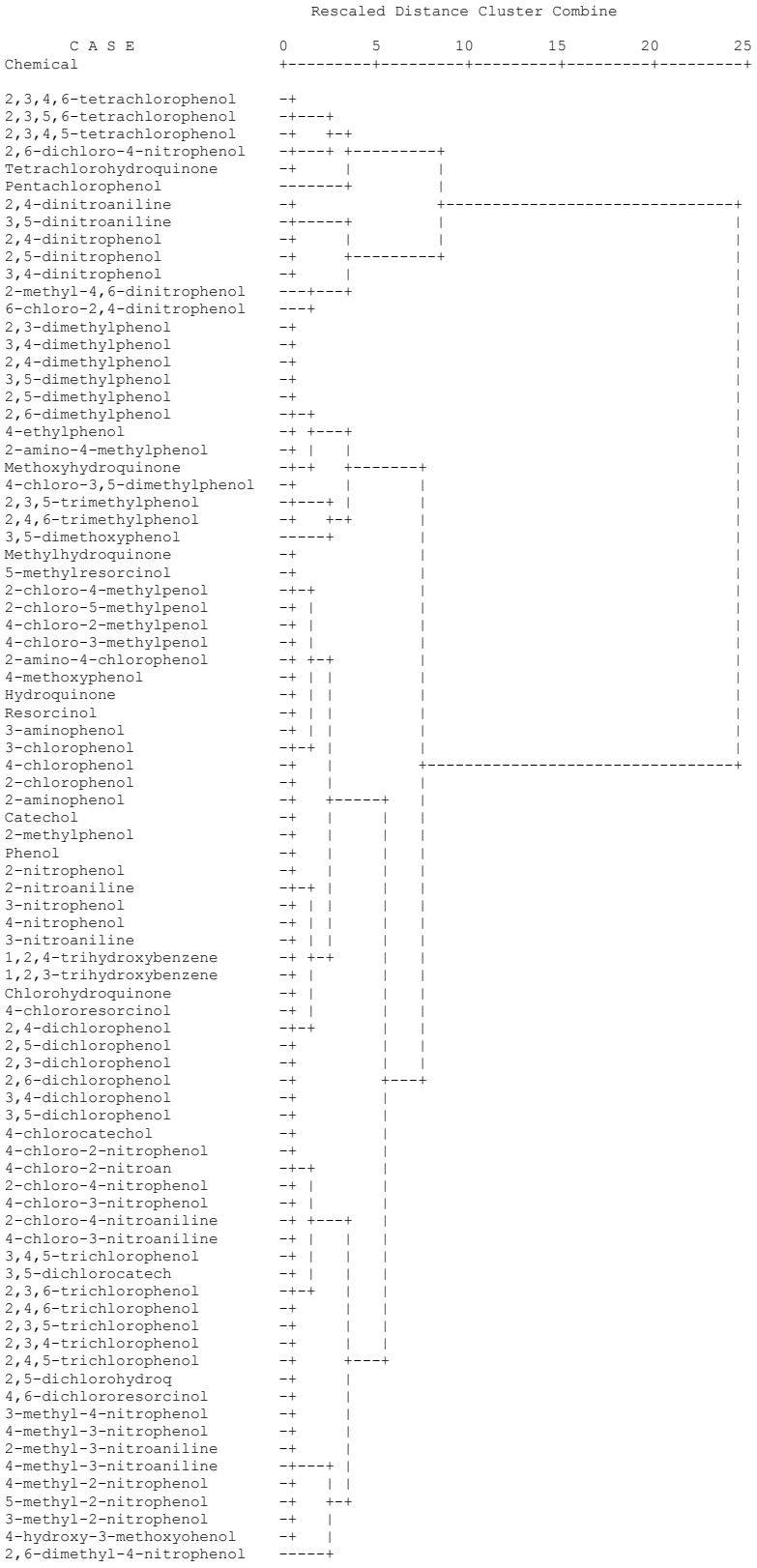


Figure 4.5. Dendrogram from hierarchical cluster analysis of the data set.

4.2.1. Linear models

Given the fact that, a model's goodness-of-fit increases with the increasing number of independent variables. There is a trade-off between the simplicity of a model and its goodness-of-fit. Therefore, during the model selection step, the principle of parsimony was taken into account; the model with the best statistical metrics for both training and test sets, but having as few parameters as possible was selected. The descriptors for model development were selected among 2818 descriptors from DRAGON 6, ADMET 8, and SPARTAN 10 software packages.

Previous findings showed that log *P* and log *D* are proper descriptors in algal toxicity prediction. Log *P* and log *D*-based empirical and theoretically calculated descriptors were given priority. These priority descriptors were obtained from ECOSAR (experimental log *P* was preferred), Danish (Q)SAR database, DRAGON, and ADMET Predictor programs.

The first linear model (MLR1) with three descriptors was obtained for 67 training set chemicals together with nonlinear models (Table 4.3) via All Subsets module in QSARINS program (Eq. 4.2).

$$pT = 2.010\ (\pm 0.818)\mathrm{ATSC3e} + 0.472\ (\pm 0.146)\mathrm{Admet\_MlogP} + 0.210\ (\pm 0.113)\mathrm{PEoDIa\_3D} - 0.768\ (\pm 0.320) \quad (4.2)$$
$$n_{tr} = 67,\ R^2 = 0.67$$

95% Confidence intervals of descriptor coefficients were given in parantheses. Standardized coefficients of the model was interpreted as the contribution of each descriptor. The contribution of each descriptor in the model, hence, lined up as Admet_MlogP (0.588), ATSC3e (0.442), and PEoDIa_3D (0.210). The internal and external validation parameters were given in Section 4.2.2. Figure 4.6 shows predicted vs. observed toxicities for the model, the dashed line stands for the unity line. Those chemicals located away from the unity line were subjected to outlier analysis in Section 4.3.1.

The descriptor ATSC3e appeared in MLR1 model is a 2D descriptor calculated as centered Broto-Moreau autocorrelation of lag 3 weighted by Sanderson electronegativity (Todeschini and Consonni, 2009). ATSC descriptors were shown to be representative in explaining toxicity in the literature. Gramatica et al. (2016) developed an MLR model with ATSC descriptor to predict personal care products toxicity for fish. PEoDIa_3D is an ADMET descriptor. It represents proximity effects of electron donors of type I (including atoms with hydrogens). All descriptors are positively correlated with algal toxicity.

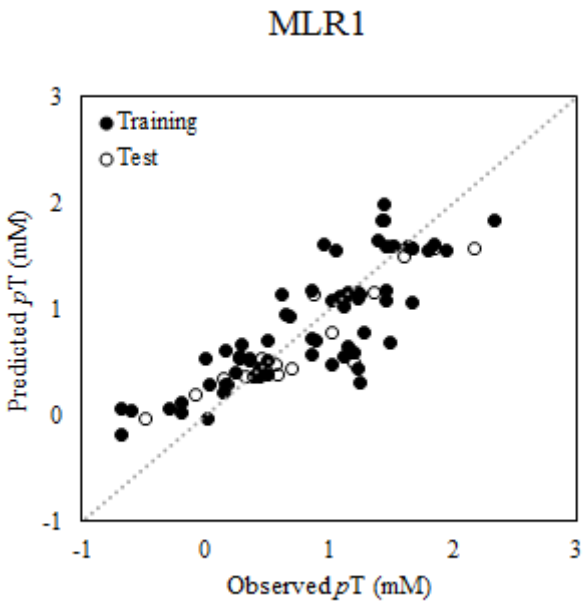


Figure 4.6. Predicted vs. observed *p*T values for the MLR1 model.

Table 4.3. Observed and predicted *p*T values, hat (leverage) values, Euclidean distances (ED), and standardized residuals for models of the second division.

		MLR1			SVR1			BPNN1		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St. Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St. Res.**
1*	-0.085	0.189	0.051	0.702	-0.035	0.959	0.146	0.039	0.959	0.345
2	0.387	0.361	0.053	-0.076	0.295	0.986	-0.306	0.358	0.986	-0.093
3	0.442	0.359	0.054	-0.212	0.295	0.986	-0.470	0.354	0.986	-0.248
4*	0.333	0.359	0.054	0.075	0.295	0.986	-0.114	0.354	0.986	0.069
5*	0.138	0.341	0.057	0.525	0.288	0.987	0.480	0.319	0.987	0.517
6*	0.578	0.379	0.050	-0.524	0.303	0.984	-0.896	0.393	0.984	-0.541
7	0.515	0.377	0.050	-0.347	0.302	0.985	-0.672	0.389	0.985	-0.350
8	0.236	0.403	0.046	0.424	0.315	0.983	0.244	0.438	0.983	0.573
9	-0.295	0.065	0.050	0.925	-0.286	1.088	0.015	-0.176	1.088	0.329
13	-0.189	0.113	0.071	0.799	-0.053	1.179	0.443	-0.070	1.179	0.345
14	0.510	0.523	0.062	0.033	0.646	1.073	0.440	0.631	1.073	0.348
15	0.356	0.506	0.066	0.385	0.639	1.074	0.904	0.603	1.074	0.701
16	0.157	0.298	0.131	0.375	0.503	1.256	1.111	0.317	1.256	0.454
17	1.208	0.582	0.034	-1.622	0.778	0.807	-1.397	0.830	0.807	-1.097
18	-0.187	0.025	0.052	0.560	-0.338	1.090	-0.480	-0.253	1.090	-0.181
20*	1.607	1.504	0.039	-0.275	1.500	1.079	-0.355	1.491	1.079	-0.344
21	-0.684	0.053	0.051	1.910	-0.302	1.089	1.222	-0.199	1.089	1.387
22	0.509	0.698	0.032	0.486	0.737	1.047	0.733	0.911	1.047	1.158
23	0.893	0.698	0.032	-0.495	0.737	1.047	-0.496	0.911	1.047	0.061
24	0.855	0.578	0.047	-0.734	0.634	1.052	-0.731	0.727	1.052	-0.385

Table 4.3. continued.

		MLR1			SVR1			BPNN1		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
25	1.172	0.586	0.045	-1.518	0.640	1.052	-1.716	0.740	1.052	-1.242
26	1.153	0.645	0.078	-1.335	0.943	1.177	-0.670	0.809	1.177	-0.985
27	1.122	1.095	0.073	-0.066	1.188	1.668	0.220	1.140	1.668	0.058
28*	0.703	0.439	0.054	-0.682	0.810	1.026	0.354	0.643	1.026	-0.163
29	1.232	0.439	0.054	-2.065	0.810	1.026	-1.360	0.643	1.026	-1.693
30	1.045	1.560	0.116	1.376	1.258	2.191	0.673	1.388	2.191	0.975
31	1.806	1.560	0.116	-0.676	1.258	2.191	-1.786	1.388	2.191	-1.219
32	0.624	1.143	0.121	1.416	0.903	1.873	0.916	1.259	1.873	1.844
33	0.653	0.945	0.040	0.765	0.855	1.068	0.664	1.106	1.068	1.317
34	1.029	0.487	0.027	-1.398	0.547	0.780	-1.563	0.688	0.780	-0.987
35	1.232	1.157	0.105	-0.196	1.023	1.638	-0.671	1.162	1.638	-0.196
36*	0.567	0.487	0.027	-0.215	0.547	0.780	-0.074	0.688	0.780	0.341
37	1.146	1.157	0.105	0.019	1.023	1.638	-0.412	1.162	1.638	0.035
38	1.397	1.642	0.116	0.653	1.543	2.151	0.463	1.415	2.151	0.042
39*	1.210	0.495	0.027	-1.839	0.324	0.521	-2.866	0.672	0.521	-1.554
40	1.466	1.181	0.035	-0.748	1.331	1.113	-0.450	1.325	1.113	-0.420
41	1.849	1.607	0.079	-0.642	1.600	1.638	-0.810	1.459	1.638	-1.130
42*	1.367	1.165	0.034	-0.531	1.323	1.112	-0.153	1.313	1.112	-0.166
43	0.958	1.617	0.050	1.711	1.489	1.083	1.713	1.549	1.083	1.702
44	1.261	0.309	0.041	-2.465	0.110	0.662	-3.722	0.383	0.662	-2.531
45	-0.684	-0.188	0.074	1.298	-0.473	1.238	0.670	-0.590	1.238	0.259
47	0.438	0.448	0.048	0.020	0.232	0.397	-0.673	0.580	0.397	0.405

Table 4.3. continued.

		MLR1			SVR1			BPNN1		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
49*	1.030	0.781	0.025	-0.640	0.662	0.301	-1.192	1.012	0.301	-0.051
51	0.302	0.668	0.047	0.958	0.596	1.137	0.957	0.842	1.137	1.564
52	0.150	0.216	0.044	0.170	0.345	1.028	0.632	0.261	1.028	0.321
54	1.251	1.135	0.063	-0.302	1.356	1.742	0.342	1.211	1.742	-0.114
55	1.505	0.676	0.089	-2.191	1.292	1.495	-0.673	0.941	1.495	-1.613
56	0.031	0.284	0.036	0.656	0.155	0.789	0.403	0.354	0.789	0.936
57*	0.176	0.290	0.036	0.283	0.164	0.789	-0.051	0.365	0.789	0.533
58	0.875	0.722	0.021	-0.381	0.806	0.564	-0.207	0.986	0.564	0.336
59	0.862	1.176	0.039	0.819	1.126	1.062	0.862	1.284	1.062	1.224
60	0.694	0.931	0.029	0.621	1.017	1.108	1.057	1.114	1.108	1.224
61	1.681	1.569	0.068	-0.293	1.472	1.653	-0.671	1.463	1.653	-0.627
62	-0.001	0.540	0.022	1.385	0.483	0.619	1.562	0.749	0.619	2.162
63	-0.600	0.047	0.051	1.685	-0.295	1.012	0.987	-0.222	1.012	1.091
64	0.173	0.605	0.024	1.118	0.487	0.960	1.024	0.795	0.960	1.805
65	0.359	0.533	0.027	0.445	0.392	0.962	0.103	0.677	0.962	0.914
66*	0.457	0.533	0.027	0.188	0.392	0.962	-0.220	0.677	0.962	0.625
67	1.088	1.122	0.033	0.083	1.297	1.145	0.670	1.358	1.145	0.774
68	1.244	1.098	0.032	-0.366	1.276	1.145	0.116	1.338	1.145	0.284
69	1.115	1.098	0.032	-0.057	1.276	1.145	0.505	1.338	1.145	0.630
70*	0.880	1.136	0.034	0.662	1.309	1.145	1.387	1.369	1.145	1.413
71	1.472	1.084	0.032	-0.996	1.263	1.145	-0.669	1.326	1.145	-0.414
72	1.668	1.058	0.031	-1.578	1.239	1.145	-1.394	1.304	1.145	-1.057

Table 4.3. continued.

		MLR1			SVR1			BPNN1		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
73	1.642	1.587	0.068	-0.139	1.674	1.413	0.109	1.610	1.413	-0.087
74*	1.864	1.579	0.067	-0.738	1.677	1.413	-0.593	1.606	1.413	-0.732
75*	1.510	1.595	0.069	0.224	1.670	1.413	0.519	1.613	1.413	0.298
76	1.672	1.579	0.067	-0.239	1.677	1.413	0.022	1.606	1.413	-0.184
77	1.532	1.587	0.068	0.151	1.674	1.413	0.465	1.610	1.413	0.230
78*	2.181	1.571	0.066	-1.599	1.680	1.412	-1.619	1.603	1.412	-1.666
79	2.344	1.845	0.098	-1.323	1.658	1.558	-2.207	1.695	1.558	-1.863
80	1.446	1.845	0.098	1.055	1.658	1.558	0.673	1.695	1.558	0.706
81	1.431	1.845	0.098	1.109	1.658	1.558	0.738	1.695	1.558	0.764
82	1.452	1.991	0.119	1.461	1.658	1.694	0.673	1.728	1.694	0.803
83	1.281	0.786	0.050	-1.288	1.248	1.329	-0.103	1.026	1.329	-0.732
84	0.018	-0.025	0.075	-0.119	-0.226	1.232	-0.796	-0.331	1.232	-1.014
85	1.131	0.561	0.060	-1.488	0.520	0.995	-1.974	0.767	0.995	-1.049
86	1.465	1.584	0.112	0.335	1.330	1.165	-0.421	1.629	1.165	0.488
87	0.266	0.526	0.022	0.657	0.477	0.643	0.670	0.732	0.643	1.333
88	1.135	1.034	0.017	-0.245	1.097	0.289	-0.108	1.267	0.289	0.395
89	1.952	1.551	0.043	-1.035	1.561	0.820	-1.258	1.566	0.820	-1.108
90*	-0.493	-0.025	0.075	1.227	-0.226	1.232	0.854	-0.331	1.232	0.458
91	0.265	0.561	0.060	0.763	0.520	0.995	0.808	0.767	0.995	1.434
92	1.025	1.084	0.075	0.170	1.116	0.964	0.309	1.361	0.964	0.983

* Test set chemical

In an attempt to develop a model using a *p*H corrected hydrophobicity descriptor (log *D*), S+logD was found to be significant in a linear model. Using another division, 18 chemicals allocated into the test set (Table 4.4). The second linear model (Eq. 4.3) (MLR2) was obtained with three descriptors and 66 chemicals.

$$pT = 3.253\ (\pm 0.641)\ \text{ATSC3e} + 0.457\ (\pm 0.132)\ \text{S+logD} + 0.557\ (\pm 0.236)\text{EEM\_Xfon} - 1.101(\pm 0.406) \quad (4.3)$$
$$n_{tr} = 66,\ R^2 = 0.689$$

Again a DRAGON descriptor, ATSC3e, and two ADMET descriptors, S+logD and EEM_Xfon, appeared in MLR2 model. Predicted vs. observed toxicity values were plotted on Figure 4.7, together with y=x line. Those chemicals located away from the unity line were subjected to outlier analysis in Section 4.3.1.

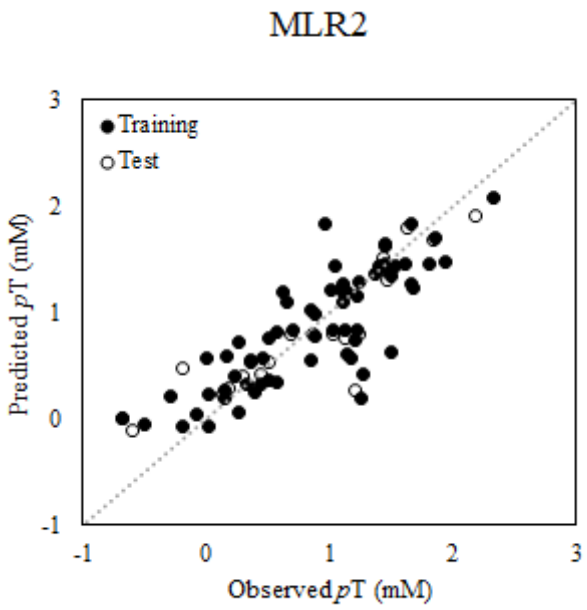


Figure 4.7. Predicted vs. observed *p*T values for MLR2 model.

In MLR2 model, standardized coefficients of descriptors revealed that the importance of descriptors in explaining algal toxicity was ordered as S+logD (0.759), ATSC3e (0.733), and EEM_Xfon (0.516). S+logD is a *p*H correlated hydrophobicity descriptor. EEM_XFon is the sigma Fukui index provides information on local reactivity of sigma electrons in the

Table 4.4. Observed and predicted *p*T values, hat (leverage) values, Euclidean distances (ED), and standardized residuals for models of the second division.

		MLR2			SVR2			BPNN2		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
1	-0.085	0.038	0.054	0.318	-0.042	0.616	0.122	-0.175	0.616	-0.286
2	0.387	0.254	0.048	-0.364	0.248	0.836	-0.451	0.092	0.836	-0.898
3	0.442	0.324	0.050	-0.312	0.367	0.934	-0.232	0.208	0.934	-0.700
4	0.333	0.317	0.050	-0.034	0.357	0.926	0.087	0.198	0.926	-0.398
5*	0.138	0.235	0.051	0.255	0.260	0.849	0.383	0.096	0.849	-0.132
6	0.578	0.341	0.047	-0.643	0.352	0.917	-0.726	0.200	0.917	-1.146
7	0.515	0.360	0.048	-0.404	0.390	0.948	-0.383	0.237	0.948	-0.824
8	0.236	0.410	0.045	0.455	0.407	0.952	0.530	0.267	0.952	0.082
9	-0.295	0.205	0.040	1.325	-0.062	0.535	0.725	-0.132	0.535	0.478
13*	-0.189	0.480	0.028	1.781	0.213	0.510	1.282	0.158	0.510	1.050
14*	0.510	0.536	0.059	0.071	0.684	1.175	0.555	0.557	1.175	0.143
15	0.356	0.527	0.062	0.453	0.701	1.191	1.087	0.572	1.191	0.639
16	0.157	0.190	0.144	0.086	0.658	1.328	1.586	0.285	1.328	0.377
17*	1.208	0.276	0.088	-2.563	0.385	1.038	-2.625	0.175	1.038	-3.122
18	-0.187	-0.069	0.066	0.327	-0.332	0.701	-0.451	-0.365	0.701	-0.528
20	1.607	1.466	0.050	-0.387	1.457	0.972	-0.486	1.343	0.972	-0.807
21	-0.684	0.004	0.059	1.849	-0.268	0.663	1.312	-0.311	0.663	1.113
22	0.509	0.760	0.035	0.666	0.758	1.125	0.788	0.654	1.125	0.433
23	0.893	0.785	0.037	-0.281	0.801	1.158	-0.284	0.701	1.158	-0.571
24	0.855	0.560	0.044	-0.805	0.650	1.125	-0.668	0.508	1.125	-1.063

Table 4.4. continued.

		MLR2			SVR2			BPNN2		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
25	1.172	0.577	0.044	-1.591	0.668	1.135	-1.597	0.527	1.135	-1.940
26	1.153	0.613	0.068	-1.459	0.923	1.348	-0.724	0.796	1.348	-1.067
27	1.122	1.094	0.055	-0.070	1.022	1.131	-0.313	0.959	1.131	-0.484
28	0.703	0.836	0.049	0.367	0.954	1.389	0.808	0.896	1.389	0.592
29	1.232	0.844	0.049	-1.038	0.956	1.381	-0.872	0.898	1.381	-1.002
30	1.045	1.447	0.082	1.085	1.279	1.492	0.729	1.298	1.492	0.749
31	1.806	1.452	0.082	-0.981	1.278	1.488	-1.694	1.298	1.488	-1.544
32	0.624	1.201	0.110	1.614	1.280	1.805	2.101	1.307	1.805	2.073
33	0.653	1.092	0.078	1.207	0.956	1.074	0.973	0.895	1.074	0.739
34	1.029	0.839	0.061	-0.517	0.816	1.268	-0.680	0.803	1.268	-0.685
35	1.232	1.150	0.084	-0.220	0.986	1.036	-0.777	0.915	1.036	-0.949
36	0.567	0.815	0.060	0.662	0.804	1.291	0.745	0.795	1.291	0.680
37	1.146	1.199	0.090	0.136	1.012	1.015	-0.439	0.933	1.015	-0.654
38	1.397	1.445	0.072	0.122	1.267	1.402	-0.424	1.246	1.402	-0.465
39	1.210	0.748	0.084	-1.264	0.686	1.184	-1.669	0.640	1.184	-1.719
40	1.466	1.406	0.044	-0.171	1.241	0.903	-0.728	1.181	0.903	-0.871
41*	1.849	1.686	0.069	-0.446	1.171	0.739	-2.161	1.225	0.739	-1.887
42	1.367	1.360	0.042	-0.028	1.237	0.936	-0.424	1.162	0.936	-0.628
43	0.958	1.828	0.067	2.356	1.191	0.598	0.737	1.358	0.598	1.200
44	1.261	0.184	0.054	-2.900	0.002	1.108	-4.005	-0.015	1.108	-3.847
45	-0.684	0.012	0.067	1.877	-0.143	1.226	1.708	-0.171	1.226	1.535
47*	0.438	0.411	0.061	-0.079	0.221	0.876	-0.697	0.138	0.876	-0.913

Table 4.4. continued.

		MLR2			SVR2			BPNN2		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
49*	1.030	0.791	0.028	-0.635	0.442	0.461	-1.873	0.523	0.461	-1.529
51*	0.302	0.410	0.058	0.297	0.451	1.416	0.481	0.472	1.416	0.518
52	0.150	0.264	0.066	0.309	0.378	1.562	0.727	0.402	1.562	0.759
54*	1.251	0.793	0.082	-1.250	1.287	1.771	0.118	1.094	1.771	-0.470
55	1.505	0.620	0.106	-2.439	1.271	1.928	-0.728	1.062	1.928	-1.321
56	0.031	0.235	0.074	0.560	0.258	1.492	0.726	0.283	1.492	0.765
57*	0.176	0.288	0.072	0.295	0.297	1.448	0.371	0.318	1.448	0.415
58*	0.875	0.788	0.030	-0.218	0.869	1.117	-0.004	0.794	1.117	-0.229
59	0.862	1.015	0.031	0.414	0.911	0.866	0.162	0.885	0.866	0.075
60*	0.694	0.802	0.031	0.299	0.874	1.110	0.586	0.802	1.110	0.337
61	1.681	1.225	0.065	-1.232	1.416	1.375	-0.839	1.256	1.375	-1.281
62	-0.001	0.567	0.020	1.501	0.224	0.315	0.713	0.280	0.315	0.845
63*	-0.600	-0.108	0.065	1.334	-0.264	0.594	1.070	-0.345	0.594	0.769
64	0.173	0.595	0.026	1.129	0.447	0.800	0.882	0.322	0.800	0.459
65	0.359	0.559	0.030	0.529	0.477	0.914	0.371	0.334	0.914	-0.077
66	0.457	0.565	0.030	0.278	0.485	0.920	0.080	0.344	0.920	-0.351
67	1.088	1.211	0.035	0.322	1.242	1.155	0.484	1.112	1.155	0.067
68	1.244	1.280	0.042	0.107	1.353	1.286	0.359	1.231	1.286	-0.027
69	1.115	1.266	0.041	0.390	1.334	1.269	0.680	1.211	1.269	0.273
70	0.880	0.984	0.028	0.277	0.945	0.838	0.207	0.818	0.838	-0.186
71*	1.472	1.310	0.047	-0.429	1.415	1.354	-0.175	1.297	1.354	-0.522
72	1.668	1.271	0.046	-1.071	1.383	1.358	-0.914	1.270	1.358	-1.207

Table 4.4. continued.

		MLR2			SVR2			BPNN2		
ID	***p*T**	**Pred. *p*T**	**Hat Val.**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**	**Pred. *p*T**	**ED**	**St.Res.**
73*	1.642	1.806	0.084	0.454	1.898	1.529	0.822	1.688	1.529	0.144
74	1.864	1.705	0.070	-0.423	1.791	1.429	-0.221	1.591	1.429	-0.811
75	1.510	1.353	0.045	-0.422	1.387	0.971	-0.393	1.248	0.971	-0.791
76	1.672	1.841	0.090	0.470	1.945	1.583	0.874	1.734	1.583	0.193
77	1.532	1.445	0.047	-0.228	1.484	1.099	-0.145	1.329	1.099	-0.606
78*	2.181	1.920	0.106	-0.721	2.044	1.687	-0.434	1.830	1.687	-1.055
79	2.344	2.083	0.117	-0.717	2.096	1.652	-0.776	1.839	1.652	-1.511
80	1.446	1.629	0.064	0.484	1.644	1.127	0.618	1.483	1.127	0.100
81*	1.431	1.522	0.060	0.248	1.545	1.000	0.367	1.411	1.000	-0.058
82	1.452	1.649	0.067	0.541	1.679	1.155	0.728	1.505	1.155	0.166
83	1.281	0.411	0.121	-2.429	0.829	1.192	-1.436	0.546	1.192	-2.216
84	0.018	-0.063	0.084	-0.228	-0.265	0.929	-0.907	-0.336	0.929	-1.074
85*	1.131	0.754	0.036	-1.004	0.685	0.596	-1.415	0.576	0.596	-1.671
86	1.465	1.457	0.094	-0.007	1.459	0.734	-0.004	1.472	0.734	0.036
87	0.266	0.063	0.084	-0.568	-0.046	0.977	-1.006	-0.180	0.977	-1.358
88	1.135	0.842	0.034	-0.768	0.783	0.621	-1.104	0.679	0.621	-1.359
89	1.952	1.479	0.052	-1.268	1.465	0.956	-1.545	1.356	0.956	-1.791
90	-0.493	-0.058	0.083	1.183	-0.262	0.922	0.725	-0.333	0.922	0.474
91	0.265	0.729	0.038	1.227	0.665	0.596	1.259	0.553	0.596	0.854
92	1.025	1.206	0.053	0.502	1.249	0.694	0.728	1.167	0.694	0.444

* Test set chemical

molecule, and can be further decomposed into nucleophilicity and electrophilicity of the molecule (Todeschini and Consonni, 2009). Generally, QSTR models are functions of a molecule's hydrophobic, electronic, and structural properties (Sanderson et al., 2004). In the proposed models, log *P* and log *D* descriptors stand for hydrophobicity, ATSC3e and EEM_Xfon stand for electronic properties. ATSC3e also bears traces of structural properties of the molecules as it considers defined on a molecular graph (Todeschini and Consonni, 2009). The descriptor values used in models were given in Appendix Table D.1.

When two MLR models are compared, while MLR2 had stronger regression fits (R^2_{tr}), MLR2 had better predictions (all external validation parameters) on test set. Although *k*' values are below the 0.85 limit, *k* values satisfy the condition for MLR models. The models' reliability and robustness were also checked by the Y-scrambling procedure. The average R^2 of shuffled models were significantly low (Table 4.5), proving that there is no chance correlation in the models. Possible outliers of model was inspected in Section 4.3.1.

4.2.2. Nonlinear models

In order to explore nonlinear relationship between the molecular properties and the toxicity, two nonlinear methods were employed in the prediction of algal toxicity of tested chemicals. The same training and test sets in linear models were employed in nonlinear modelling part in order to compare their performances.

For the selection of model type of SVR, an empirical approach was employed. All parameters were kept constant and the model types were compared. It was observed that *nu* was superior to *epsilon*, therefore *nu* type SVR was selected.

ATSC3e, Admet_MlogP, and PEoDIa_3D were used as input variable in SVR1 model and ATSC3e, S+logD, and EEM_Xfon were used in SVR2 model. Fine-tuning for optimum parameters were done via grid search on Cost and *nu*. Internal and external validation parameters of SVR1 model were given in Table 4.5. The parameters used in model development were given in Table 4.6. Models' support vector details were given in Appendix Table D.2 and Table D.3.

Table 4.5. Summary of statistical parameters used for internal and external validations of linear and nonlinear models.

Validation criteria	MLR1	MLR2	SVR1	SVR2	BPNN1	BPNN2
Internal						
R^2_{tr}	0.672	0.689	0.789	0.779	0.735	0.754
R^2_{adj}	0.657	0.674	n/a	n/a	n/a	n/a
Q^2_{LOO}	0.629	0.641	0.788	0.778	0.729	0.724
$RMSE_{tr}$	0.382	0.370	0.308	0.312	0.347	0.348
SE	0.394	0.381	n/a	n/a	n/a	n/a
F	43.063	45.815	n/a	n/a	n/a	n/a
R^2_{ys}	0.047	0.046	n/a	n/a	n/a	n/a
External						
R^2_{test}	0.846	0.775	0.821	0.766	0.874	0.773
$RMSE_{test}$	0.308	0.348	0.308	0.352	0.261	0.377
Q^2_{F1}	0.809	0.765	0.946	0.923	0.961	0.912
Q^2_{F2}	0.809	0.765	0.946	0.923	0.961	0.912
Q^2_{F3}	0.787	0.725	0.786	0.719	0.847	0.677*
CCC_{test}	0.879	0.857	0.894	0.862	0.918	0.839*
r^2_m	0.722	0.676	0.762	0.691	0.756	0.701
$(R^2 - R^2_0)/R^2$	0.006	0.001	0.008	0.033	0.043	0.013
k	1.135	1.069	0.862	0.860	0.896	0.781*
$(R^2 - R'^2_0)/R^2$	0.056	0.067	0.004	0.002	0.007	0.009
k'	0.821*	0.849*	1.071	1.052	1.054	1.158*
$\lvert R^2_0 - R'^2_0 \rvert$	0.042	0.052	0.003	0.024	0.032	0.003
Mean residual	-0.072	-0.046	-0.080	-0.013	-0.058	-0.151

* Value that did not fulfill the criteria outlined in Sections 2.6 and 3.3.5

Table 4.6. Architecture of SVR1 and SVR2 models.

Property	**Specification**
Model type	Nu-SVR
Kernel	RBF
Termination criterion tolerance	0.001
Cost	5
Gamma	0.333
Nu	0.5
Normalization	-1.0 – 1.0
Rho	-0.171

The predicted vs. observed toxicity values of both training and test sets are depicted in Figure 4.8. The dashed line stands for y=x line. Those chemicals located away from the unity line were subjected to outlier analysis in Section 4.3.2.

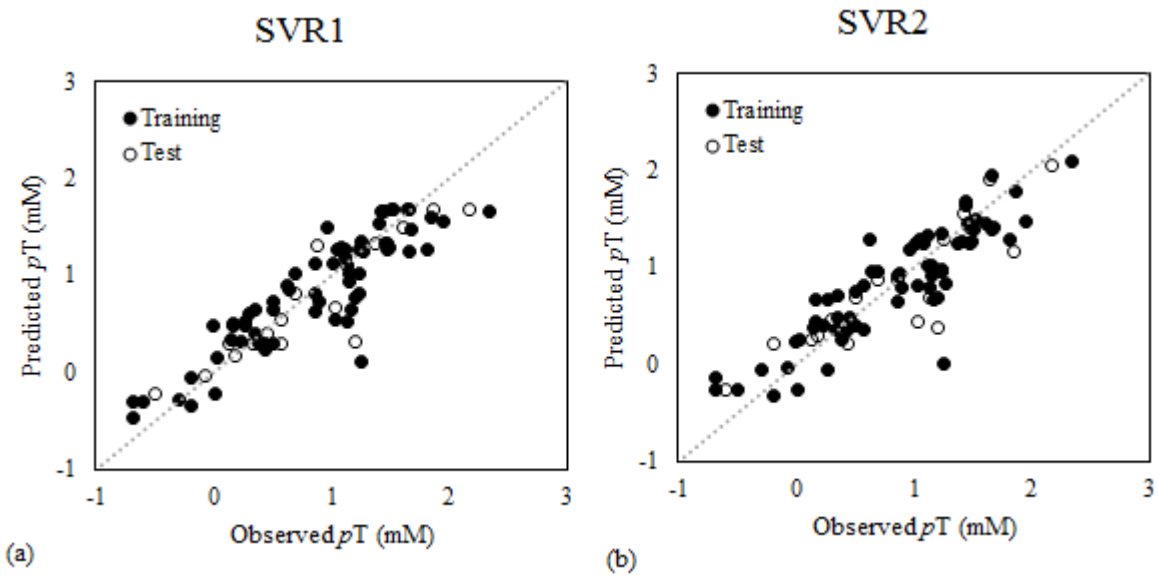


Figure 4.8. Predicted vs. observed *p*T values for SVR models. (a) SVR1 and (b) SVR2.

For the development of BPNN models, again the same training set with the same descriptors for the same training/test set division used in linear modeling part were employed. The data were normalized between 0.1 and 0.9 and sigmoidal transfer function was used in the neural networks. Fine-tuning for optimum parameters were done via grid search on maximum number of epochs and the number of neurons in the hidden layer. The parameters used in model development were given in Table 4.7. Model functions were given in Appendix Table D.4. Table D.5. The correlation of the predicted and observed *p*T values

of both BPNN models are given in Figure 4.9. Internal and external validation parameters of BPNN models were given in Table 4.5.

Table 4.7. Architecture of BPNN models.

	BPNN1	**BPNN2**
Property	**Parameter**	**Parameter**
Maximum training epochs	1500	2000
Learning rate	0.3	0.3
Output layer learning rate	0.3	0.3
Momentum	0.2	0.2
Data range normalization	0.1-0.9	0.1-0.9
Number of neurons in hidden layer	3	3
Initial weight range	±0.5	±0.5

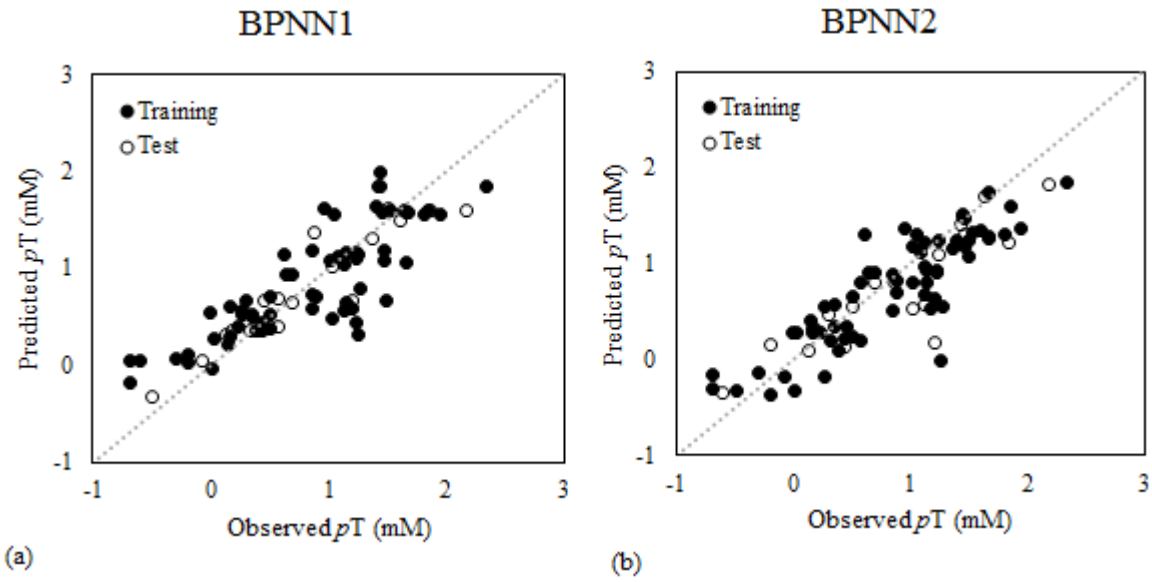


Figure 4.9. Predicted vs. observed *p*T values for BPNN models. (a) BPNN1 and (b) BPNN2.

The parameters used in model development are given in Table 4.8. Relevance scores of the descriptors used in BPNN1 ordered as Admet_MlogP, PEoEDIa_3D, and ATSC3e. An interesting outcome of BPNN2 model is that the relevance of descriptors in MLR2 and the relevance scores of BPNN2 models coincided.

Table 4.8. Relevance scores of the input variables for BPNN models.

Model	Index	Descriptor	Relevance score
	0	Admet_MlogP	100
BPNN1	1	PEoEDIa_3D	81
	2	ATSC3e	62
	0	S+logD	100
BPNN2	1	ATSC3e	30
	2	EEM_XFon	4

A comparison of the internal and external validation results of four nonlinear models was summarized in Table 4.5. together with the validation results of linear models. Regression fits of all models were in acceptable qualities. Cross-validation coefficient (Q^2_{LOO}) representing internal predictive capacities were above the acceptance limit (0.50). Additionally, the low difference between Q^2_{LOO} and R^2 indicated the robustness of the developed models. All models fulfilled the internal validation criteria. However, BPNN2 model failed some of the external validation criteria. Out of the limits for Q^2_{F3}, CCC_{test}, and the slopes (k and k') of regression lines, indicated that although the fit of the regression line was good, it was not close enough to the unity line. On the other hand, the limit for Q^2_{F3} is considered 0.5 by some researchers (Roy et al., 2015a), and its value may be accepted as above the limit. Considering the internal validation parameters, nonlinear models performed better than their linear counterparts in general. While models of the first division performed better in model parameters, models of the second division were better on test sets. All models were further analyzed for the presence of any systematic error. The residuals of the prediction set were normally distributed (Kolmogorov-Smirnov test) with means close to zero (Table 4.5), implying that there was no systematic error in the predictions.

4.3. Applicability Domains of All Models

4.3.1. Applicability domains of linear models

The ADs of linear models were defined by the boundaries of the descriptor and the toxicity range (Table 4.9). Williams plot (Figure 4.10) shows the structural distance of chemicals to each other and their prediction accuracy. While the distances are represented with leverage values, errors are represented by standardized residuals. The vertical reference

line is at the critical hat value (h^*=0.182 and 0.179, MLR1 and MLR2, respectively), and the horizontal reference lines are ±3σ, the cut-off values for the response outliers. Both models had neither response, nor structural outliers.

Table 4.9. Boundaries of *p*T and descriptors used in models.

Variable	Minimum value	Maximum value
*p*T	-0.68	2.34
ATSC3e (MLR1)	0.045	0.533
ATSC3e (MLR2)	0.040	0.533
Admet_MlogP	0.447	3.909
PEoEDIa_3D	0	4
S+logD	-1.008	3.571
EEM_XFon	0.138	1.759

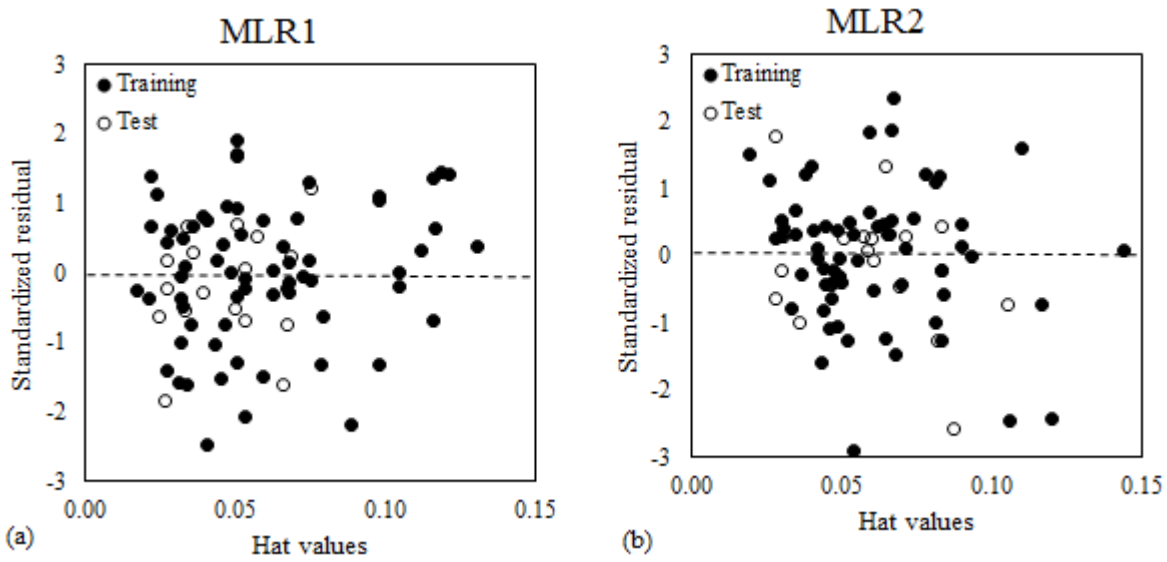


Figure 4.10. Williams plots of (a) MLR1 (b) MLR2, y-axis has standardized residuals and x-axis has hat values.

4.3.2. Applicability domains of nonlinear models

Applicability domains of nonlinear models were defined by the boundaries of the descriptors and the toxicity range (Table 4.9). Euclidean distances of chemicals were calculated and marked on the graphs to spot possible structurally distant chemicals (Figure 4.11). While standardized residuals are on the y-axis, distances are on the x-axis. The horizontal reference lines are the cut-off values (±3σ) for the response outliers. X-axis cut-off values are defined at the largest Euclidean distance for the training set chemicals. SVR1

and BPNN1 had the threshold 2.191, whereas SVR2 and BPNN2 had the threshold 1.928 for ED. 2-Aminophenol (44) appeared to be the common response outlier in three models (SVR1, SVR2, and BPNN2). It was underestimated by these models. BPNN2 had an additional outlier, methoxyhydroquinone (17). It was also underestimated by this model.

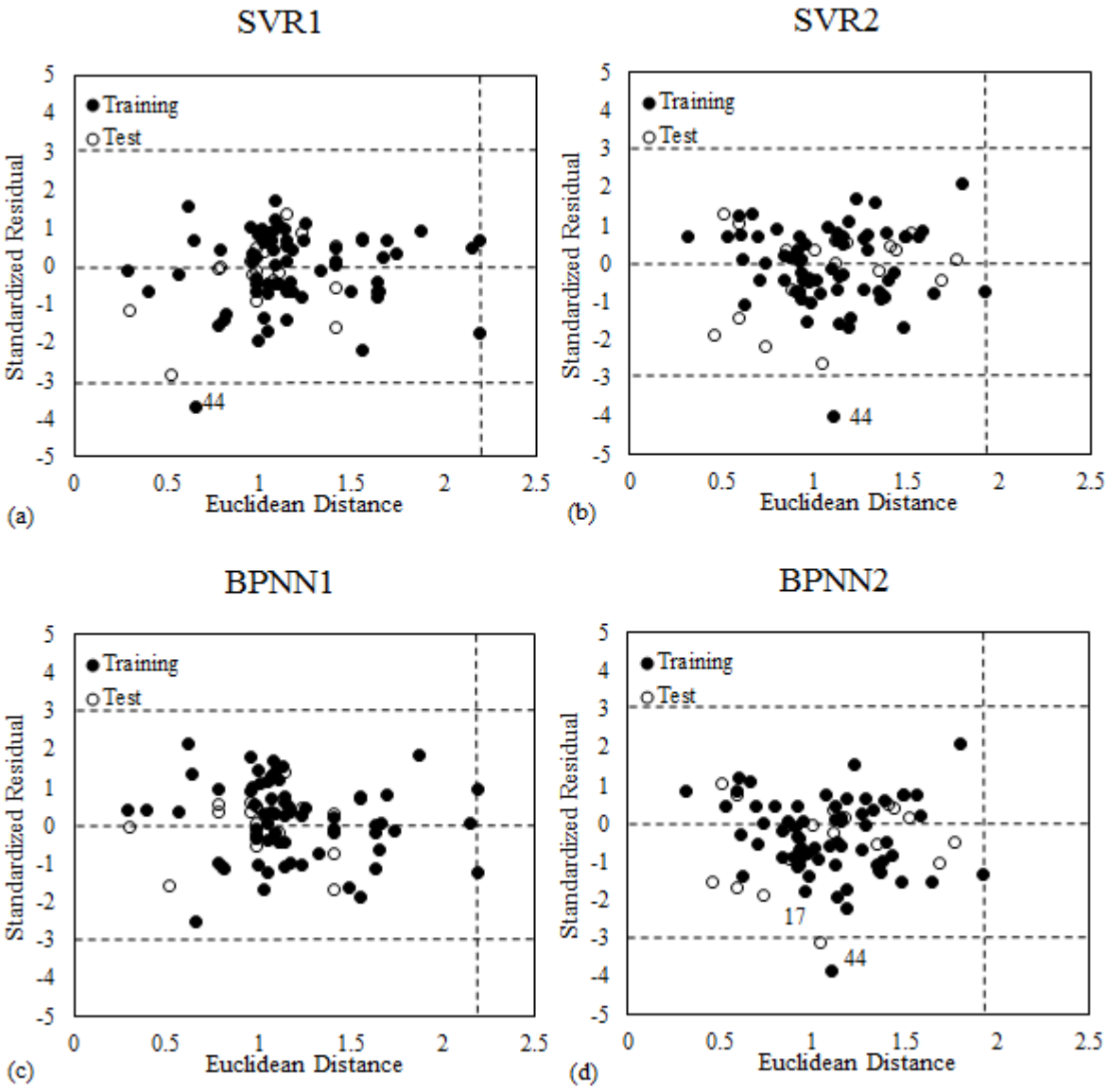


Figure 4.11. Applicability domain of nonlinear models. (a) and (b) SVR models, (c) and (d) BPNN models.

All models were tested on an external set of diverse chemicals with no algal toxicity data to estimate the predictive ability of models. In this respect, 152 chemicals were obtained from Fu et al. (2015) (Appendix Table D.6). This diverse data set consisted of phenol and

aniline derivatives, pesticides, pharmaceutical ingredients, and phthalates, i.e. a compilation of emerging pollutants. Majority of the data set consist of following classes of chemicals: 63 pesticides, 21 pharmaceuticals, 54 phenol, aniline, and other benzene derivatives. Regarding the predictions with developed models, while diethylamine and 2-propanol from the industrial chemicals group are the least toxic chemicals, anilophos and fenitrothion from the pesticide group are the most toxic chemicals. Toxicities of these chemicals to other species were reported in Appendix Table D.6.

The descriptors used in models were calculated for each chemical in the set (Appendix Table D.6). The chemicals that were out of the descriptor ranges were removed from the data set. Remaining chemicals were tested on all models (Figure 4.12). Response outlier limits were defined with observed toxicity values of the training sets (-0.68 – 2.34). Structural outlier limits for external sets were given in parentheses for each model: MLR1 (0.182), MLR2 (0.179), SVR1 and BPNN1 (2.191), SVR2 and BPNN2 (1.928).

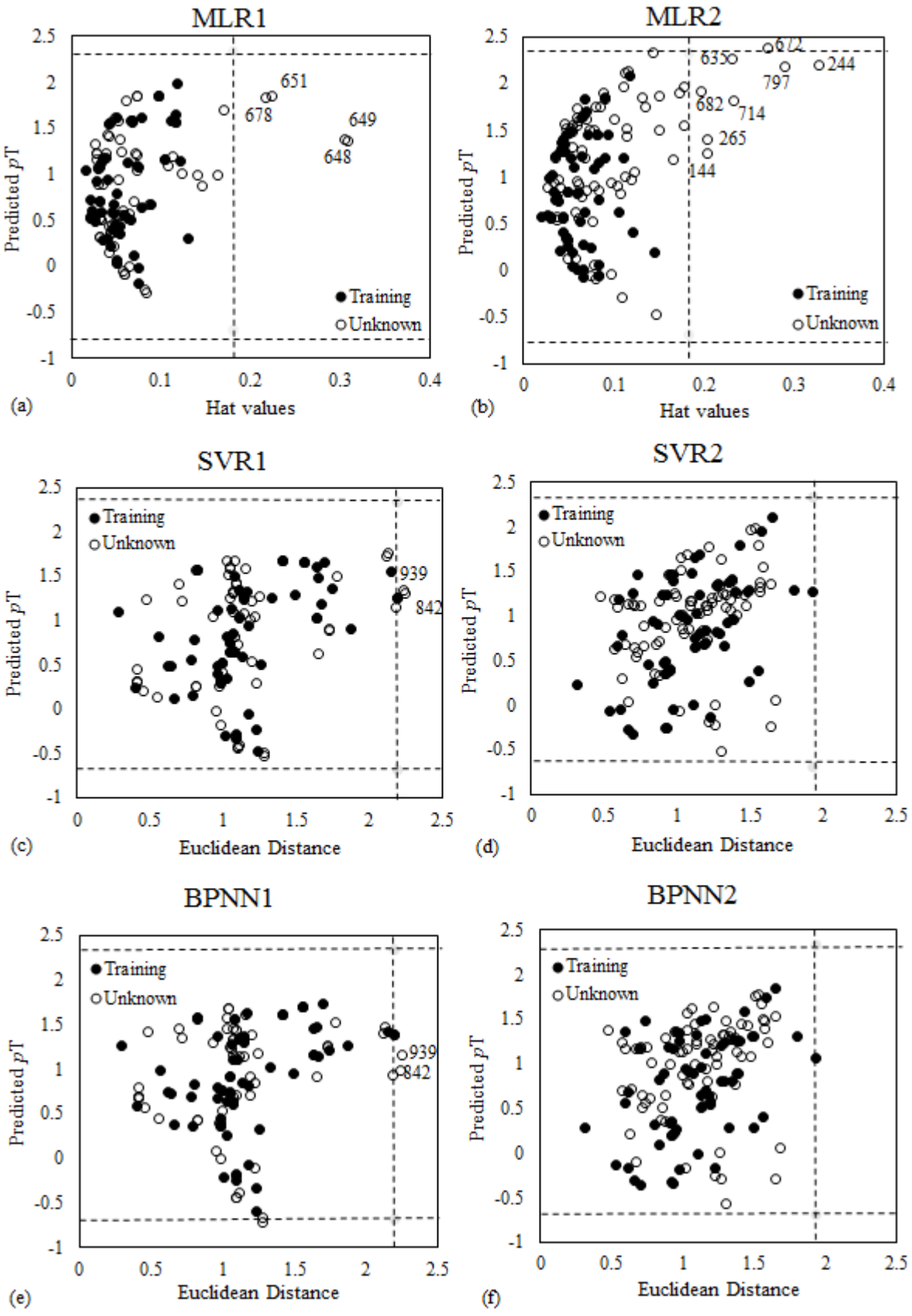


Figure 4.12. Structural coverage of all models for chemicals with no toxicity data. (a) MLR1, (b) MLR2, (c) SVR1, (d) SVR2, (e) BPNN1, (f) BPNN2.

The percent coverages of all models were given in Table 4.10. MLR1 model was found to have wider coverage on external set than MLR2. This superiority was also seen in its external validation statistics. Among the nonlinear models, SVR2 and BPNN2 models predicted all of the external set chemicals within the prediction and Euclidean distance ranges. Majority of structurally distant chemicals spotted in structural coverage of MLR1 and MLR2 were pesticides. SVR1 and BPNN1 models had two structural outliers that both are antibiotics, Sulfamethoxazole (842) and Sulfaquinoxaline (939).

Table 4.10. Coverage of all models in their ADs.

Model	**Number of chemicals in external set**	**Number of chemicals in AD**	**Structural Coverage (%)**
MLR1	57	53	93.0
MLR2	73	65	89.0
SVR1	57	55	96.5
SVR2	73	73	100
BPNN1	57	55	96.5
BPNN2	73	73	100

4.4. Comparison of Acute Toxicity Results with the Literature

Table 4.11. gives a summary of literature algal toxicity values and database predictions of studied chemicals. There were not any toxicological data with the same conditions available, i.e. for *C. vulgaris* 96-h growth inhibition test. Therefore, 24-h, 48-h, and 96-h toxicity values for any green algae were included in the comparison table. Danish (Q)SAR database predictions belong to 72-h toxicity to *Pseudokirchneriella subcapitata* specie and ECOSAR (2011) predictions belong to 96-h toxicity to green algae.

Most of the ECOSAR predictions were out of the predicted IC_{50} range of the present study. Only 10% of ECOSAR predictions were within the range. Majority of ECOSAR predictions (as mg L^{-1}) were lower than the predicted values. However, chloronitrophenols had higher predictions than the prediction range.

Almost half of the Danish (Q)SAR database predictions were within the prediction range. Approximately 15% of the predictions were higher than the prediction range. While

polyphenols predictions were significantly higher, methylnitrophenol predictions were slightly higher than the prediction ranges.

ECHA, ECOTOX, and literature toxicity values were found to vary extremely, as much as 470-fold and 78,000-fold (2-methylphenol and 2-methyl-4,6-dinitrophenol, respectively). Additionally, for 26 chemicals out of 62 tested chemicals have no algal toxicity data available in the literature (24 to 96-h EC_{50}/IC_{50}).

Linear and nonlinear QSTR models reported in the literature were compared to the best models of this study. The most-used statistical parameters of models developed on algal toxicity data from various studies were summarized in Table 4.12. The analysis of the results given in Table 4.12. showed that former models in the literature lack external validation metrics. Compared to latter studies, MLR1 model had satisfactory statistics with low number of descriptors. Having more chemicals in training and test sets, the data set also have diverse set of chemicals with four different MOAs. Ertürk and Saçan (2013) developed an MLR model with log *D* and $E_{(HOMO-LUMO)}$ for the prediction of chlorophenols. While this model had better statistics than our MLR model, its domain covers only chlorophenols. The model proposed in Aruoja et al. (2011) is developed for non-polar and polar narcosis phenol and anilines. The model is developed with a hydrophobicity parameter, log *P*. Although the model includes phenol and aniline classes, our model is better than that model in terms of diverse MOA. Lu et al. (2008) developed a model for phenol and aniline derivatives with log *P* and $E_{(HOMO-LUMO)}$. While this model has superior statistics, the number of chemicals in the model is very low. In our previous study (Tugcu et al., 2017), we developed two linear models. While the two-descriptor model is constructed with log *P* and hardness, the other one is developed with log *D* and two other structural descriptors. These models had only phenol derivatives in model development, having narrower applicability domain. In a study by Pramanik and Roy (2014), a diverse set of 74 organic chemicals (phenols, anilines, pesticides, PAHs, etc.) were modelled with MLR using five descriptors, one of which is log *P*, and the others are structural descriptors. While this model has comparable statistics with

Table 4.11. The comparison of the acute toxicity values for the tested chemicals. The concentrations are given as mg L^{-1}.

Chemical ID	Predicted 96-h IC_{50} range from selected models in the present study	ECOSAR predictions	ECHA[a]	Danish (Q)SAR database[b]	ECOTOX database range[c]	Literature[d]
1	69.92-99.08	23.91	100	59.93	75-100	0.27 - 127
2	53.23-68.01	12.54	50	52.30	-	48.1
3	53.48-57.94	12.54	-	45.44	-	9.7 - 19.3
4	53.48-58.85	12.54	32.5	44.82	-	29.3, 32.5
5	55.75-71.16	12.54	10 - 48	51.72	-	41.6 - 47.5
6	49.47-55.75	12.54	-	48.04	-	32.0
7	49.91-53.37	12.54	14 – 22	63.73	-	27.2
8	44.52-48.30	13.57	-	34.91	-	21.72-62.65
9	77.39-186.24	52.78	19 - 54.7	52.15	-	-
13	51.10-181.29	58.52	-	48.56	110-220	90[e]
14	31.88-40.90	6.49	-	43.58	-	13.62
15	33.98-42.44	6.49	5.59-8.64	42.67	170-300	17-30 [e]
16	60.73-81.37	8.15	-	742.61	-	-
17	20.73-74.16	7.15	-	190.91	-	-
18	117.22-222.12	3.61	-	197.12	-	-
20	5.61-6.12	2.92	-	28.41	-	-
21	109.88-196.48	3.61	-	268.86	-	-
22	17.50-28.56	12.77	-	23.97	-	-
23	17.50-28.56	12.77	-	24.98	-	-
24	26.75-39.31	12.77	14.81	17.27	-	-
25	25.96-37.81	12.77	10-15	18.34	150	15
26	24.33-38.18	6.51	-	14.03	-	-
27	10.08-11.20	38.06	-	35.21	1.08-6.82	1.1 - 43.0
28	20.27-50.67	38.06	-	31.42	6.7	6.7 - 98.5
29	19.92-50.67	38.06	10.4 – 32	35.09	4.19-32	0.25 - 65.1

Table 4.11. continued.

Chemical ID	Predicted 96-h IC_{50} range from selected models in the present study	ECOSAR predictions	ECHA[a]	Danish (Q)SAR database[b]	ECOTOX database range[c]	Literature[d]
30	5.08-7.54	65.04	10.9 – 78	8.48	0.9-40	0.9 - 12.7
31	5.08-7.54	65.04	-	23.71	-	-
32	10.15-13.25	65.04	-	21.79	-	-
33	11.99-17.37	19.45	-	31.03	-	-
34	22.19-49.94	19.45	-	32.43	-	-
35	10.55-10.85	19.45	-	22.79	12-32	-
36	23.46-49.94	19.45	-	27.27	-	-
37	9.68-10.67	19.45	-	25.38	-	-
38	4.52-7.63	32.49	-	-	0.0014-110	3.78-5.58
39	29.84-53.45	9.86	-	21.87	-	-
40	6.81-11.45	19.23	-	13.33	-	-
41	3.58-6.04	19.23	-	7.33	-	6.16
42	7.58-11.88	19.23	-	7.81	-	-
43	3.09-5.87	9.34	-	8.12	-	-
44	45.14-71.39	7.12	-	109.88	-	0.15-0.33
45	106.23-424.72	10.50	2.44-160	135.94	-	161.42
47	32.36-47.85	4.39	-	-	4.6-8	4.6-8 [e]
49	13.96-23.77	4.60	-	38.09	-	-
51	19.89-53.75	1.56	64.6	24.37	-	65
52	75.25-84.10	2.10	-	22.66	58	33.91-57.58
54	11.27-29.47	2.28	-	4.36	-	3
55	20.97-43.90	3.08	-	3.55	-	-
56	67.31-88.49	60.66	-	20.28	-	-
57	65.70-78.34	1.71	-	19.7	-	-
58	17.80-32.76	1.84	-	9.75	-	-

Table 4.11. continued.

Chemical ID	Predicted 96-h IC_{50} range from selected models in the present study	ECOSAR predictions	ECHA[a]	Danish (Q)SAR database[b]	ECOTOX database range[c]	Literature[d]
59	8.98-16.66	1.36	-	5.78	-	3.86-9.48
60	13.27-27.21	1.84	-	6.37	-	-
61	5.87-12.95	1.90	-	1.55	-	-
62	26.59-43.05	79.21	-	33.57	-	-

[a] https://echa.europa.eu/search-for-chemicals, 72 h and 96 h endpoint
[b] Battery predictions http://130.226.165.14/User_Manual_Danish_Database.pdf (ECB, 2005).
[c] US EPA ECOTOX database (Green algae, 1-4 days EC_{50} and IC_{50} values) Search was performed on May 2017.
[d] Fu et al., 2015 if not stated otherwise.
[e] Kuhn and Pattard, 1990

Table 4.12. Linear QSTR models from various studies developed on algal toxicity data.

Number of descriptors	Chemical class	Training set			Test set			Reference
		n	R^2	Q^2_{LOO}	*n*	R^2	*RMSE*	
Linear models								
2	Chlorophenols	24	0.82	0.73	6	0.64	0.40	Ertürk and Saçan (2013)
1	Phenols and anilines	58	0.60	-	-	-	-	Aruoja et al. (2011)
2	Phenols and anilines	14	0.95	-	6	0.94	-	Lu et al. (2008)
2	Phenols	35	0.61	0.51	11	0.79	0.28	Tugcu et al. (2017)
3	Phenols	35	0.86	0.81	11	0.94	0.16	Tugcu et al. (2017)
8	Organic chemicals	389	0.66	0.64	66	0.72	0.61	Önlü and Saçan (2016)
8	Nitrobenzenes	42	0.92	0.89	-	-	-	Bao et al. (2012)
2	Phenols and anilines	21	0.88	-	-	-	-	Wang et al. (2007)
2	Nitroaromatics	25	0.72	-	-	-	-	Yan et al. (2005)
5	Organic chemicals	74	0.77	0.71	31	0.80	0.565	Pramanik and Roy (2014)
3[a]	Phenols and anilines	67	0.67	0.63	17	0.85	0.31	**This study**
Nonlinear models								
3[b]	Chlorophenols	24	0.80	0.73	6	0.92	0.26	Ertürk et al. (2012)
2[c]	Organic chemicals	73	n.a	n.a	18	0.92	0.44	Tugcu et al. (2014)
3[d]	Phenols and anilines	67	0.74	0.73	17	0.87	0.26	**This study**

[a]: MLR1; [b]: CPANN; [c]: Kriging; [d]: BPNN1

MLR1, it is unfavorable with higher number of descriptors. Likewise, the MLR model developed by Önlü and Saçan (2016) has a diverse set of organic chemicals. Their global model has eight descriptors with a hydrophobicity parameter and structural property descriptors, providing comparable results with MLR1 using higher number of descriptors and diverse chemical classes. In a study by Ertürk et al. (2012) marine algal toxicity was predicted with a CPANN model with three DRAGON descriptors. While their data set consisted of only chlorophenols, our nonlinear model has a wider applicability domain and more robust in terms of R^2- Q^2_{LOO} difference. A Kriging model with a calculated log *P* and a DRAGON descriptor was developed for a set of organic chemicals in our previous study (Tugcu et al., 2014). While this model has a stronger fit for test set of chemicals, its RMSE

is higher than BPNN1. Additionally, the Kriging model was not developed on a standardized algal toxicity data. Since neither SVR nor BPNN model for prediction of algae is available with similar chemicals, a comparison could not be performed.

4.5. Low-Toxic-Effect Concentration Models

NOEC, LOEC, and IC_{20} values together with additional low-toxic-effect values (Ertürk, 2013) to extend the data set for modelling were reported in Table 4.1. 60 chemicals having NOEC values were used in modelling.

The NOEC, ChV, and ACR values reported in the literature and the databases are considered to be in general agreement with the present study. In a study by Urrestarazu Ramos et al. (1999), 11 narcotic chemicals consisting of phenol and aniline derivatives were tested with *Chlorella pyrenoidosa*. Their average of NOEC was 16.55, whereas the present study had an average NOEC as 15.99 including reactive chemicals. Kuhn and Pattard (1990) studied the toxicity of various organic and inorganic chemicals on *Scenedesmus subspicatus* and reported EC_{10} and EC_{50} for the inhibition of growth rate at 48-96h. The average EC_{10} was 15.93 and the average ACR for chemical groups including aromatic nitro compounds, halogenated aromatics, phenols, and halogenated aromatics was 4.25. The average ACR of organic chemicals for algae reported in ECOSAR Methodology Document (2012) is 4. This value is obtained using 72/96-h EC_{50} and ChV of their data set. The present data have an average ACR_{MATC} of 3.75. The average ACR reported in EU TGD for algae is 5.4. Likewise, Chen et al. (2009) reported an ACR_{NOEC} of 5.8 for closed system (n=50). The present data have an average ACR_{NOEC} of 5.34 (Table 4.13). Where ACR_{NOEC} for 11 non-polar narcosis is 4.5 (McGrath et al., 2004), the calculated ACR_{NOEC} for the present study is 5.45. This is due to the fact that non-polar ACRs are less than the polar ones (e.g. Chen et al., 2009; Roex et al., 2000).

Considering all chemicals, average ACR_{NOEC} for algae is less than factor of 10 suggesting that regulations are conservative for *C. vulgaris*. However, three chemicals have ACR_{NOEC} higher than 10, namely, hydroquinone, 4-chloro-3-nitroaniline, and 2-chloro-4-methylphenol. Regarding the MOA, pro-electrophiles have a higher ACRs than other MOA

groups (Table 4.13). Interestingly, two of the chemicals slightly exceeding the limit of 10 are polar narcotics (Table 4.13).

Table 4.13. Comparison of ACRs overall and with respect to MOAs.

class	***n***	**ACR_{MATC}**	**ACR_{EC20}**	**ACR_{NOEC}**
All	60	3.75	1.80	5.34
Polar narcotic	43	3.85	1.76	5.45
Respiratory uncoupler	11	3.24	1.58	4.59
Pro-electrophile	7	4.02	2.07	6.02
Soft electrophile	6	3.46	2.03	4.89

In line with the previous findings by Ahlers et al. (2006) and May et al. (2016), no significant relation was found between log *P* and ACR. Thereby, we opted to study on QSAR modeling of NOEC and IC_{20}. Strong correlations were observed between the acute median and the low-toxic-effect concentrations (Table 4.14). The promising results encouraged us to study further on these relations.

Table 4.14. Correlations between the low-toxic effects and the median inhibitory concentration.

Pearson correlation coefficient (significant at 0.01 level)	log(1/NOEC)	log(1/LOEC)	log(1/ IC_{20})	log(1/ IC_{50})
log(1/NOEC)	1			
log(1/LOEC)	0.998	1		
log(1/ IC_{20})	0.963	0.958	1	
log(1/ IC_{50})	0.961	0.937	0.985	1

QSAR models were developed to predict the NOEC (*p*NOEC=log(1/NOEC) mM) and IC_{20} (pIC_{20}=log(1/IC_{20}) mM) values. 60 chemicals with NOEC values were used in QSAR models (Table 4.15). Log(1/NOEC) and log(1/IC_{20}) values spanned wide ranges (2.943 and 2.772 log units, respectively) and both were normally distributed (non-parametric Kolmogorov-Smirnov test), appropriate for modelling. Seven chemicals with unbounded NOEC values were used as the external set data. Although a significant relationship between log *P* and NOEC was found for polar narcotics, this was not the case for the rest of the chemicals. Given the fact that all MOA groups had more or less close ACR values (Table 4.13), global models were aimed regardless of MOA. In the first attempt, theoretical

descriptors were used to estimate low-effect concentration. In the second attempt, median inhibitory concentration belong to the same experiment was used as the predictor variable.

4.5.1. Modeling of NOEC

The NOEC values of the tested 60 chemicals were sorted in increasing order. The most and the least toxic chemicals were allocated into the training set in order to define AD as wide as possible. Among 2830 descriptors in the descriptor pool a model with acceptable statistical performance was selected via the All Subsets tool in the software QSARINS. Equation 4.4 (model 1) was constructed with two DRAGON descriptors (Table 4.15). The numbers in parentheses are the 95% confidence intervals. The plot of observed versus predicted NOEC values from Equation 4.4 (model 1) is presented in Figure 4.13 (a). The Williams graph (Figure 4.13 (b)), having neither response nor structural outliers (h^*=0.188), supported the model.

In QTTR model 1, SM04_EA(bo) is a 2D descriptor defined as the spectral moment of order 4 from edge adjacency matrix weighted by bond order. This topological index describes the structural information of molecules. It is derived from the edge adjacency matrix, which represents the connections between adjacent pairs of atoms giving the information about branching (Estrada, 1997). The other descriptor in the model, E1m, is a 3D descriptor (1st component accessibility directional WHIM index/weighted by atomic masses) (Consonni and Todeschini, 2010). Apparently, molecular shape indices obtained by the number of paths and the length of the bonds within a graph for the atoms in a molecule had an effect on explaining NOEC. E1m has a slightly higher standardized coefficient than SM04_EA(bo), taking a more important role in the model.

Table 4.15. Internal and external validation parameters and equations for models developed for low toxic-effect concentrations.

	Model 1	**Model 2**	**Model 3**	**Model 4**
Parameters	**Eq. 4.4**	**Eq. 4.5**	**Eq. 4.6**	**Eq. 4.7**
	*p*NOEC= 1.220(±0.483) SM04_EA(bo) +1.210 (±0.387)E1m -6.905 (±2.953)	*p*NOEC= 0.984 (±0.090)pIC_{50} +0.711(±0.089)	pIC_{20}= 0.608 (±0.142)T_Grav3 +1.500 (±0.586)Mor09m -5.816(±1.729)	pIC_{20}= 1.004 (±0.044) pIC_{50} +0.253(±0.044)
Training set				
n_{tr}	48	48	48	48
R^2_{tr}	0.634	0.914	0.653	0.979
Q^2_{LOO}	0.586	0.906	0.614	0.977
SE	0.414	0.199	0.407	0.099
F	38.95	488.662	42.276	2130.077
Y-randomization R^2 aver.	0.044	0.021	0.041	0.021
Test set				
n_{test}	12	12	12	12
R^2_{test}	0.756	0.958	0.849	0.991
Q^2_{F1}	0.757	0.952	0.818	0.983
Q^2_{F2}	0.750	0.951	0.816	0.983
Q^2_{F3}	0.703	0.941	0.798	0.981
CCC_{test}	0.855	0.976	0.892	0.990
r^2_m aver.	0.665	0.926	0.794	0.926

*p*NOEC: log(1/NOEC); pIC_{20}: log(1/IC_{20})

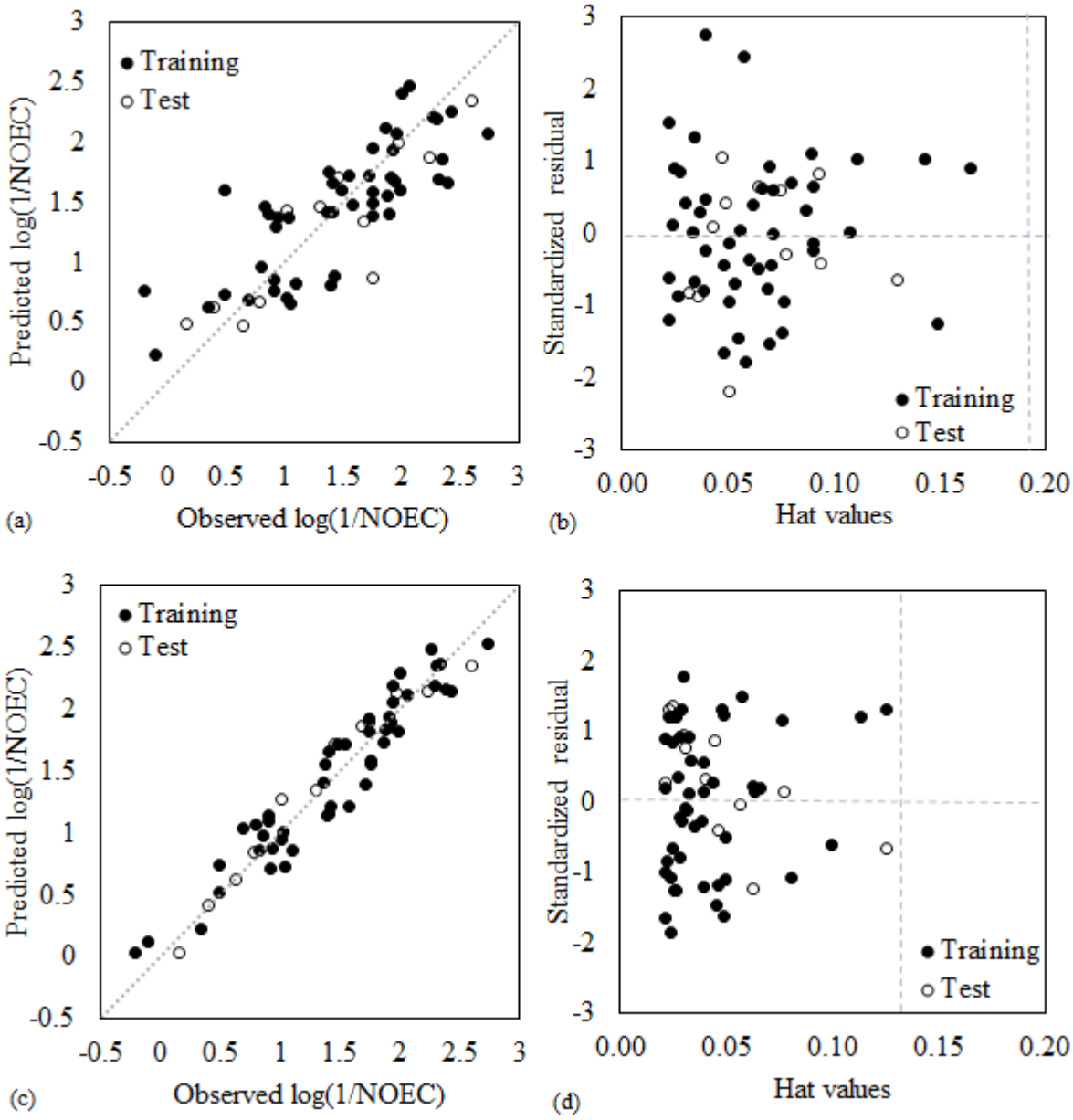


Figure 4.13. (a) Predicted from model 1 vs. observed NOEC (b) Williams plot for model 1 (c) Predicted from model 2 vs. observed NOEC (d) Williams plot for model 2.

In an attempt to estimate NOEC value using the IC_{50} as independent variable yielded Eq.4.5 (model 2) with exceptionally good statistics (Table 4.15). The plot of observed versus predicted values for the model is presented in Figure 4.13 (c). The Williams graph, did not have any response outlier and two chemicals were at the edge of the critical hat limit (h^*=0.125)(Figure 4.13 (d)).

Predicted NOEC and IC_{20} values are given and test set of chemicals are marked in Table 4.16. The descriptor values are given in Appendix Table E.1 and leverage and standardized residuals are given in Appendix Table E.2.

4.5.2. Modeling of IC_{20}

Using 2830 descriptors in the descriptor pool, Eq. 4.6 (model 3) was developed for IC_{20} with two theoretical descriptors (Table 4.15). This two-descriptor linear model fulfilled the external validation criteria outlined in Sections 2.6 and 3.3.5. The observed vs. predicted IC_{20} values from Eq. 4.6 are presented in Figure 4.14 (a). An ADMET and a DRAGON descriptor appeared in the model 3. Both descriptors are representative of both atomic masses and the geometry of the molecules. T_Grav3 is a geometrical descriptor, which is a modified form of the gravitational index. This descriptor characterizes the mass distribution of the molecule using atomic masses of considered atoms and interatomic distances (Todeschini and Consonni, 2009). Mor09m (signal 09 / weighted by mass) is a 3D-MoRSE DRAGON descriptor. 3D-MoRSE descriptors provide information from the three-dimensional structure of a molecule. The interatomic distances of atoms in the optimized molecule are calculated to be used in the radial basis function. Mor09m was also found significant in explaining acute algal toxicity of phenols in a previous study of the authors (Tugcu et al., 2017). Considering the standardized coefficients, T_Grav3 contributed more than Mor09m in explaining IC_{20}.

It was found that all the chemicals were within the AD of the model as the leverage value of the chemicals were lower than the critical hat value (h^*=0.188) of the model, with the exception of pentachlorophenol (Figure 4.14 (b)) for Eq. 4.6 with a standardized residual of -0.594. This high leverage chemical belongs to the training set and is influential in the model regression. After a meticulous examination, it was found that it has a unique structure among the chemicals. It is the only chemical with five substituents in the data set. Moreover, T_Grav3 value of pentachlorophenol is very high, originating from the exceptionally high molecular weight (266) of it.

Using the IC_{50} as independent variable, Eq. 4.7 (model 3) was developed with better statistical metrics than Equation 4.6 (model 4) (Table 4.15). This model had neither response nor structural outlier (h^*=0.125) (Figure 4.14 (c and d)).

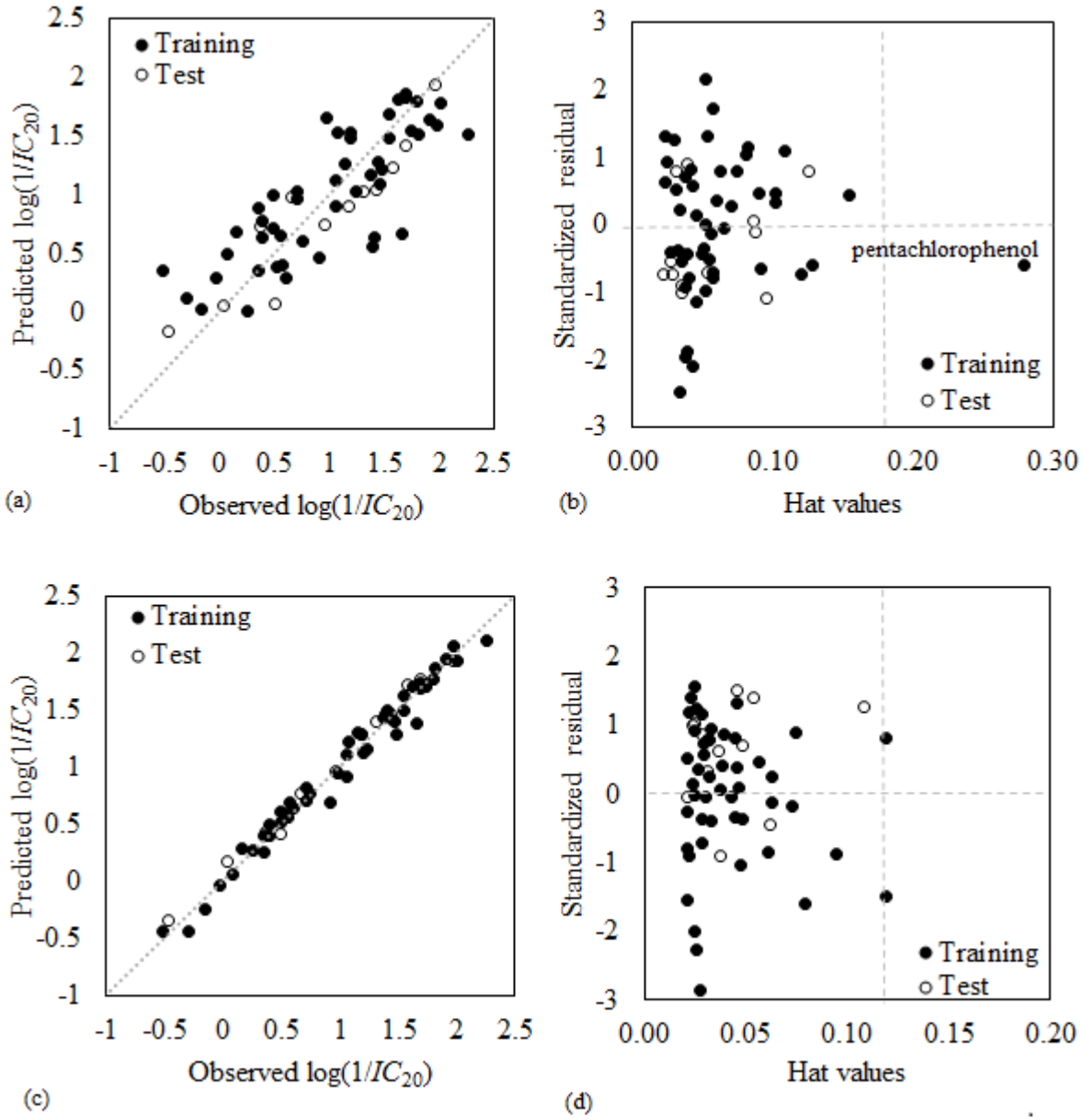


Figure 4.14. (a) Predicted from Eq. 4.6 vs observed IC_{20} (b) Williams plot for model 3 (c) Predicted from Eq. 4.7 vs observed IC_{20} (d) Williams plot for model 4.

Table 4.16. Predicted values belong to each low-toxic-effect model.

ID	Predicted *p*NOEC[a] from Equation 4.4	Predicted *p*NOEC[a] from Equation 4.5	Predicted pIC_{20}[b] from Equation 4.6	Predicted pIC_{20}[b] from Equation 4.7
1	0.471*	0.627*	0.059*	0.167*
2	0.853	1.092	0.294	0.641
3	0.763	1.146	0.400	0.696
4	0.686	1.038	0.376	0.587
5	0.669*	0.847*	0.349	0.391
8	0.697	0.943	0.626	0.490
9	0.628*	0.421*	0.284	-0.044
13	0.725	0.525	0.493	0.063
14	0.877	1.213	0.602	0.765
15	0.961	1.061	0.709	0.610
16	0.818	0.865	0.075*	0.410*
17	0.863*	1.899*	0.555	1.465
20	2.398	2.291	1.504	1.866
21	0.769	0.038	0.345	-0.434
22	1.476	1.212	0.985*	0.764*
23	1.576	1.589	1.022	1.149
24	1.745	1.552	1.112	1.111
25	1.340*	1.864*	1.172	1.429
26	1.400	1.845	1.275	1.410
27	1.388	1.814	0.667	1.379
28	1.412	1.402	0.750*	0.958*
29	1.493	1.923	0.628	1.490
30	2.116	1.739	1.254	1.302
31	2.204	2.487	1.593	2.066
33	1.467*	1.353*	0.903	0.908
34	1.603	1.723	0.893*	1.286*
36	1.436*	1.269*	0.965	0.822
37	1.558	1.838	1.022*	1.403*
39	1.935	1.901	1.034*	1.468*
40	1.874*	2.153*	1.220*	1.725*
41	2.065	2.529	1.508	2.109
42	1.673	2.055	1.681	1.625
43	1.662	1.653	1.525	1.214
45	0.490*	0.038*	0.109	-0.434
47	0.803	1.142	0.464	0.692
49	1.710*	1.724*	1.207	1.287
51	1.375	1.008	0.654	0.556
52	1.456	0.858	0.774	0.403
55	2.074	2.191	1.421*	1.764*
56	1.604	0.741	0.675	0.284

Table 4.16. continued.

ID	Predicted *p*NOEC[a] from Equation 4.4	Predicted *p*NOEC[a] from Equation 4.5	Predicted pIC_{20}[b] from Equation 4.6	Predicted pIC_{20}[b] from Equation 4.7
59	1.941	1.559	1.524	1.118
60	1.721	1.393	1.655	0.949
61	1.861	2.364	1.638	1.940
62	1.291	0.710	0.885	0.252
63	0.230	0.121	-0.166*	-0.350*
64	1.369	0.881	0.725*	0.426*
66	1.409	1.160	1.030	0.711
68	1.708	1.934	1.483	1.502
71	1.667	2.159	1.859	1.731
72	1.697	2.351	1.777	1.927
75	2.199	2.196	1.788	1.769
76	2.346*	2.355*	1.931*	1.931*
80	1.998*	2.133*	1.807	1.704
81	2.465	2.118	1.824	1.689
82	2.250	2.139	1.539	1.710
84	0.663	0.729	0.008	0.271
88	1.598	1.827	1.093	1.392
90	0.618	0.226	0.021	-0.242
91	1.395	0.972	0.997	0.519
92	1.728	1.719	1.473	1.282
External set				
6	0.849	1.279	0.417	0.835
7	0.825	1.217	0.520	0.765
32	1.804	1.325	0.810	0.875
38	2.232	2.085	1.306	1.658
54	1.993	1.941	1.280	1.508
57	1.456	0.884	1.017	0.433
70	1.777	1.576	1.301	1.136

[a] *p*NOEC: log(1/NOEC), [b] pIC_{20}: log(1/ IC_{20}), * Test set of chemical.

4.5.3. Testing the models using external set chemicals

The developed models were applied on the external set of seven chemicals with experimental NOEC data that were not used in modelling and testing steps. All of the chemicals were located in their ADs of models (Figure 4.15). Each chemical had a leverage value less than the model's critical hat and their predictions were within the experimental interval. For the NOEC models, model 1 estimated all seven chemicals correctly. Model 2 had six correct predictions out of seven chemicals. Model 3 and 4 had satisfactory fits (0.69 and 0.90 R^2, respectively) for this external set (Table 4.17).

Table 4.17. Estimation results for all models on the external set of chemicals.

ID	**Observed *p*NOEC**	**Predicted *p*NOEC (model 1)**	**Predicted *p*NOEC (model 2)**	**Observed pIC_{20}**	**Predicted pIC_{20} (model 3)**	**Predicted pIC_{20} (model 4)**
6	>0.786	0.849	1.279	0.772	0.417	0.835
7	>0.786	0.825	1.217	0.676	0.520	0.765
32	>1.265	1.804	1.325	0.838	0.810	0.875
38	>2.121	2.232	2.085	1.691	1.306	1.658
54	>1.998	1.993	1.941	1.547	1.280	1.508
57	>1.188	1.456	0.884*	0.719	1.017	0.433
70	>1.513	1.777	1.576	1.315	1.301	1.136

*p*NOEC: log(1/NOEC); pIC_{20}: log(1/IC_{20}), * Unsuccessful prediction

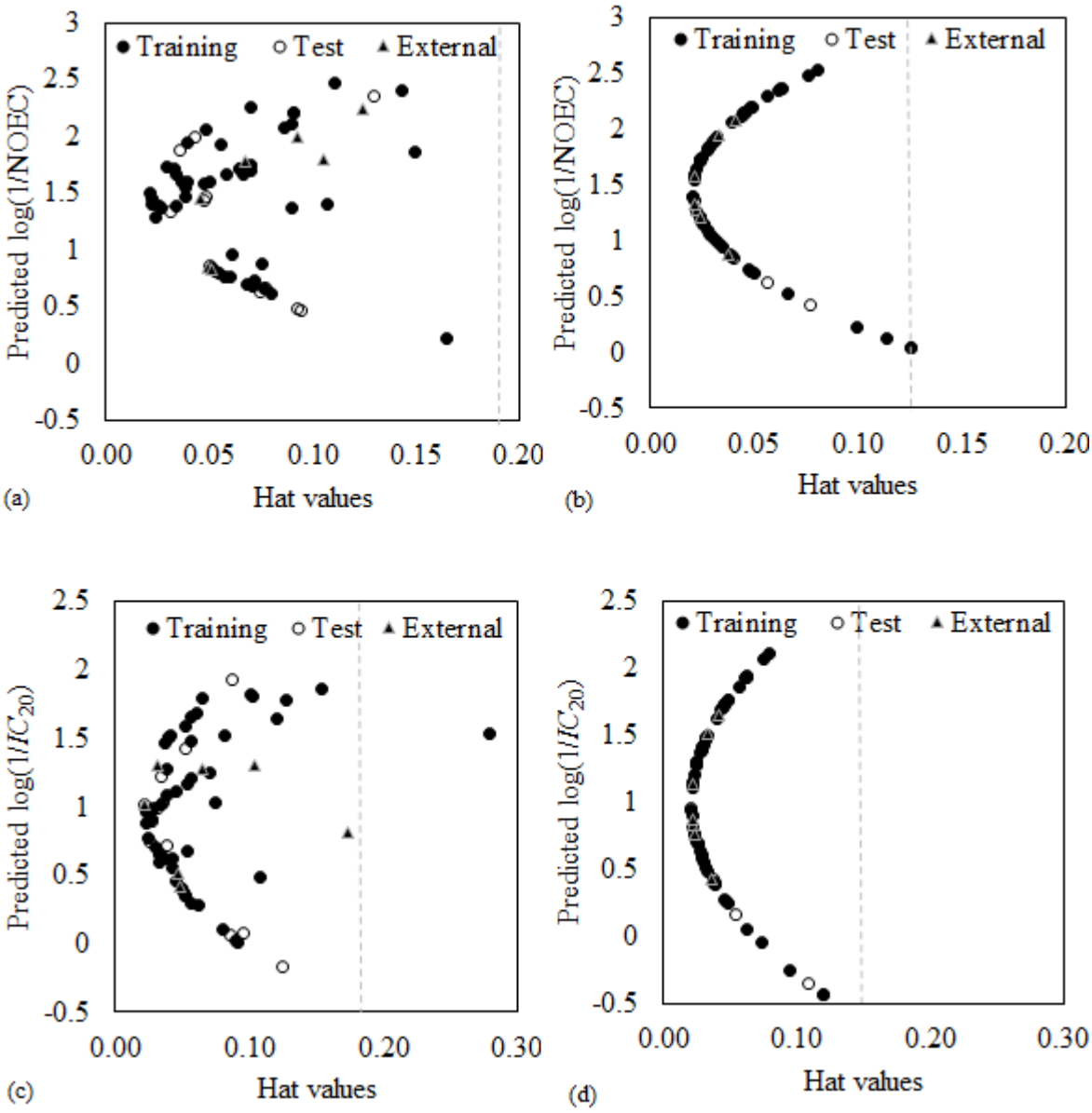


Figure 4.15. Predicted NOEC values vs. hat values for the training, test and external set of chemicals (a) model 1, (b) model 2; Predicted IC_{20} values vs. hat values for the training, test and external set of chemicals (c) model 3, (d) model 4.

Toxicity assay results for our study are listed together with those obtained in the literature in Table 4.18. Toxicity values of studied chemicals were searched in US EPA ECOTOX database (Green algae, 24-96 h EC_{10}, EC_{20}, EC_{25}, NOEC, and LOEC values), and ECHA Database. There is not much chronic algal toxicity values available in the literature. While no chronic value is present for 33 chemicals, most of the others have limited endpoints. The difference between the literature and the present study's NOEC and LOEC values largely depended on species difference, test duration, and concentrations tested in the bioassay. Therefore, it is unfounded to make a solid comparison based on these attributes. The information in Table 4.18 is provided for the sake of completeness.

4.6. Interspecies Toxicity Models

Our previous studies which showed a good relationship between *T. pyriformis* and *D. tertiolecta* (Ertürk and Saçan, 2012), and between *P. subcapitata* and *D. tertiolecta* (Ertürk et al., 2012) for chlorophenols led us to further probe these relationships also for the extended data set. Since *C. vulgaris* toxicity values are scarce, *C. vulgaris* toxicity was considered as dependent variable.

4.6.1. Ciliate-algae QTTR

Previous interspecies toxicity studies confirmed significant relationship between ciliate and algal species (Cronin et al., 2004; Huang et al., 2007). In order to develop a QTTR model the acute toxicity data producing a 50% growth inhibition ($pT_{T.pyriformis}$=-log IGC_{50}) on *T. pyriformis* (40-h) were retrieved from Cronin et al. (2002). 64 common chemicals of the *T. pyriformis* and the *C. vulgaris* sets (Table 4.20) were considered for modeling purpose. There is a significant correlation (R=0.88) between *T. pyriformis* and *C. vulgaris* toxicity values excluding methylhydroquinone. In order to develop a QTTR, 50 chemicals were selected for training set (Table 4.19).

Table 4.18. Literature low-toxic-effect concentrations (mg L^{-1}, unless otherwise noted) for the studied chemicals.

	ECETOC	ECHA	ECOTOX				Literature				The present study
ID	**NOEC**	**NOEC unless otherwise noted**	***EC*$_{10}$**	**NOEC**	**LOEC**	***EC*$_{25}$**	**NOEC**	**LOEC**	***EC*$_{10}$**	***IC*$_{20}$**	**NOEC/LOEC/*IC*$_{20}$**
1		6.8	2.2								25.0/50.0/100.6
2											15.0/30.0/30.2
3				50			<6.27[a]	6.27[a]	3.7[a]		15.0/30.0/32.5
4				50							25.0/50.0/37.3
5		2-4		50							20.0/40.0/54.5
6				50							<20.0/20.0/20.6
7		1.7		50							<20.0/20.0/25.8
8											11.9/23.8/48.9
9											50.0/100.0/131.7
13									80[b]		50.0/100.0/129.9
14											5.1/10.1/24.2
15		1.61 - 2.05							9[b]		21.3/42.5/43.4
16											10.0/40.0/40.1
17											2.5/5/5.7
20											1.8/2.9/2.7
21		2.49									200.0/400.0/401.8
22											3.8/7.6/30.9
23											2.5/5.0/8.3
24							0.97[c]				6.0/12.0/12.4
25	1.9	1.9		1.9	5.7		4.7[c]; 2.3-4.7[d]; 1.9[e]	5.7[e]	4.7-5.2[b]		3.0/6.0/6.1

Table 4.18. continued.

	ECETOC	ECHA	ECOTOX				Literature				The present study
ID	**NOEC**	**NOEC unless otherwise noted**	***EC*$_{10}$**	**NOEC**	**LOEC**	***EC*$_{25}$**	**NOEC**	**LOEC**	***EC*$_{10}$**	***IC*$_{20}$**	**NOEC/LOEC/*IC*$_{20}$**
26											2.0/4.0/5.6
27				0.696-1	1.39		<0.12[a]	0.12[a]	0.053[a];33-14[f]		2.5/5.0/3.1
28				1	10		0.99[a]	1.98[a]	2.2[a]; 18-31[f]		6.0/12.0/15.2
29				0.3-1	1.39-10		<0.15[a]	0.15[a]	0.07 [a];10[b];<1-3[f]		2.5/5.0/5.4
30				1-10	25		0.51[a]	1.02[a]	0.388[a];8[b];28-17[f]		2.5/5.0/13.1
31											1.0/2.0/1.9
32											<10.0/10.0/26.7
33											7.68/15.36/13.4
34					6.8						5.0/10.0/10.0
36											14.9/29.9/30.2
37											2.0/4.0/7.4
38				1-100	10				16[b]		<1.5/1.5/4.0
39											2.0/4.0/6.2
40											1.0/2.0/4.6
41											0.3/0.6/0.9
42											2.0/3.9/5.0
43											8.0/16.0/17.4
45		25									75.0/150.0/215.3
47									2[b]		5.0/10.0/15.0

Table 4.18. continued.

	ECETOC	ECHA	ECOTOX				Literature				The present study
ID	**NOEC**	**NOEC unless otherwise noted**	***EC*$_{10}$**	**NOEC**	**LOEC**	***EC*$_{25}$**	**NOEC**	**LOEC**	***EC*$_{10}$**	***IC*$_{20}$**	**NOEC/LOEC/*IC*$_{20}$**
49											5.0/10.0/4.7
51							<100[g]				12.8/25.6/37.9
52	28			28	54		28[e]	54[e]			20.5/41.0/55.5
54											<1.8/1.8/5.2
55											2.1/4.1/3.7
56											49.3/98.5/106.4
57											<9.9/9.9/29.1
59		2.1 (*EC*$_{10}$)									3.03/6.05/11.0
60											3.3/6.6/18.3
61											1.0/2.01/2.6
62											17.6/35.3/66.4
63			0.329-250	94.11	235.275		<8.48[a];120[h]	8.48[a];180[h]	3.89 [a]	193.7[h]	120.0/240.0/276.3
64				10			4.93[a];20[h]	12.3[a];40[h]	3.38 [a];42[b]	57.6[h]	15.0/30.0/55.4
66		5.8 (*IC*$_{10}$)		0.6-10			<5[a];10[h];13-1.7-5.8[d]	5[a];20[h]	4.21 [a];5.8[b]	34.5[h]	5.0/10.0/24.9
68		0.7; 0.35–1.09(*EC*$_{10}$); 0.69-1.76 (*IC*$_{10}$)		1			<0.97[a];5[h];0.67-3.6-6.3[j]; <0.73[i]	0.97 [a];10[h];0.73[i]	0.771[a]; 6.3[b];9.76[k]	10.7[h];6.31[i]	2.0/4.0/4.7
70							40[h]	60[h]		86.1[h]	<5.0/5.0/7.9
71											0.7/1.3/3.3
72				0.38-0.75	0.75-1.5		2[h]	4[h]		6.5[h]	0.8/1.6/1.6

Table 4.18. continued.

	ECETOC	ECHA	ECOTOX				Literature				The present study
ID	**NOEC**	**NOEC unless otherwise noted**	***EC*$_{10}$**	**NOEC**	**LOEC**	***EC*$_{25}$**	**NOEC**	**LOEC**	***EC*$_{10}$**	***IC*$_{20}$**	**NOEC/LOEC/*IC*$_{20}$**
75											1.0/2.0/3.1
76				0.3-1			1[h];0.1-0.24[l]	2[h]		4.2[h]	0.5/1.0/2.1
80			0.32				<0.100[a]	0.1[a]	0.0000265[a]		2.5/5.0/5.5
81				0.6			1[h]	2[h]		4.2[h]	2.0/4.0/4.7
82			0.094-8.65	0.005-2.66	0.0125-13.32	0.0315-0.0925	0.001[a];0.1[h]	0.002[a]; 0.2[h]	0.001[a];0.17[i]	0.31[h]	1.0/2.0/4.8
84		1.5-33 µg/L; 8.5-34 µg/L(*EC*$_{10}$)						10/20/60.3			10.0/20.0/60.3
88											1.5/3.0/5.0
90		47-97					67[m]				50.0/100.0/159.0
91											20.0/40.0/46.3
92											5.0/10.0/11.5

a: Chen et al., 2009; b: Kuhn and Pattard, 1990; c: Janus and Posthumus, 2002; d: Moermond and Heugens, 2009a; e: Urrestarazu Ramos et al., 1999; f: Madhavi et al., 1995; g: OECD SIDS Report, 2001; h: Ertürk and Saçan, 2012; i: Geiger et al., 2016; j: Moermond and Heugens, 2009b; k: Xing et al., 2012; l: Moermond and Heugens, 2009c; m: Tamura et al., 2013.

Table 4.19. The interspecies model results using *T. pyriformis* toxicity for the prediction of *C. vulgaris* toxicity.

ID	Experimental $pT_{T.pyriformis}$	Experimental $pT_{C.vulgaris}$	Predicted $pT_{C.vulgaris}$ from Equation 4.8	Hat Val.	Std.Res.
1*	-0.300	-0.080	-0.016	0.063	0.196
2	0.120	0.390	0.294	0.037	-0.291
3	0.070	0.440	0.257	0.040	-0.554
4	0.080	0.330	0.264	0.039	-0.199
6	0.120	0.580	0.294	0.037	-0.864
7*	0.110	0.510	0.286	0.038	-0.675
8*	0.210	0.240	0.360	0.033	0.362
9	-0.143	-0.290	0.100	0.052	1.186
13	-0.090	-0.190	0.139	0.049	0.999
14	0.360	0.510	0.471	0.028	-0.118
15	0.280	0.360	0.412	0.030	0.156
16	0.440	0.160	0.530	0.025	1.110
17	2.200	1.210	1.829	0.081	1.912
21*	-0.390	-0.680	-0.083	0.070	1.836
23	0.390	0.890	0.493	0.027	-1.192
24	0.700	0.860	0.722	0.021	-0.414
25	0.800	1.170	0.796	0.020	-1.121
26*	1.200	1.150	1.091	0.024	-0.178
27	0.670	1.120	0.700	0.021	-1.259
28*	0.506	0.700	0.579	0.024	-0.364
29	1.420	1.230	1.253	0.031	0.070
30	1.080	1.050	1.002	0.022	-0.143
31	0.950	1.810	0.906	0.020	-2.706
32	0.270	0.620	0.405	0.031	-0.649
33	0.610	0.650	0.655	0.022	0.016
34	1.730	1.030	1.482	0.046	1.371
35	0.570	1.230	0.626	0.022	-1.811
36	0.740	0.570	0.751	0.020	0.543
37	0.586	1.150	0.638	0.022	-1.535
38	1.720	1.400	1.475	0.046	0.226
40	1.590	1.470	1.379	0.039	-0.276
41	2.050	1.850	1.718	0.068	-0.405
43	0.630	0.960	0.670	0.022	-0.868
44	0.940	1.260	0.899	0.020	-1.081
45	-0.520	-0.680	-0.179	0.081	1.550
49	0.780	1.030	0.781	0.020	-0.746
62	-0.030	0.000	0.183	0.045	0.555
63	-0.210	-0.600	0.050	0.056	1.984
64	0.180	0.170	0.338	0.034	0.507

Table 4.19. continued.

ID	Experimental $pT_{T.pyriformis}$	Experimental $pT_{C.vulgaris}$	Predicted $pT_{C.vulgaris}$ from Equation 4.8	Hat Val.	Std.Res.
65*	0.871	0.360	0.848	0.020	1.461
66	0.550	0.460	0.611	0.023	0.453
67	1.276	1.090	1.147	0.026	0.171
68*	1.040	1.240	0.973	0.021	-0.801
69*	1.130	1.120	1.039	0.023	-0.242
70	0.740	0.880	0.751	0.020	-0.385
71	1.750	1.470	1.497	0.047	0.081
72	1.570	1.670	1.364	0.038	-0.925
74*	2.370	1.860	1.954	0.097	0.294
76	2.097	1.670	1.753	0.072	0.255
77*	1.410	1.530	1.246	0.031	-0.856
79	2.710	2.340	2.205	0.135	-0.430
80*	2.180	1.450	1.814	0.079	1.124
81	2.220	1.430	1.844	0.083	1.279
82	2.050	1.450	1.718	0.068	0.823
83	0.850	1.280	0.833	0.020	-1.340
84	0.470	0.020	0.552	0.025	1.596
85	1.260	1.130	1.135	0.026	0.015
86	2.110	1.460	1.762	0.073	0.931
87	0.750	0.270	0.759	0.020	1.463
88	1.060	1.130	0.988	0.022	-0.427
90	-0.652	-0.490	-0.276	0.093	0.666
91	0.125	0.270	0.298	0.037	0.083
92*	0.967	1.020	0.919	0.021	-0.303

* Test set of chemical

Equation 4.8 was developed from a linear regression analysis of the data set:

$$pT_{C.vulgaris} = 0.738\ (\pm 0.123)\ pT_{T.pyriformis} + 0.205(\pm 0.141) \quad (4.8)$$

n_{tr} =50, R^2_{tr} = 0.752, Q^2_{LOO}=0.731, $RMSE_{tr}$ =0.331, SE = 0.337, F =145.284

n_{test} = 13, R^2_{test} = 0.844, $RMSE_{test}$ = 0.276

Q^2_{F1}= 0.838, Q^2_{F2}= 0.838, Q^2_{F3}= 0.827, CCC_{test}= 0.908, r^2_m aver. = 0.720

The graph of predicted vs observed values and the Williams plot for Equation 4.8 are available in Figure 4.16 below. All chemicals had standardized residuals less than 3 units and leverages less than the critical hat value (h^*=0.120), except 2,3,4,5-tetrachlorophenol. This chemical belong to training set of chemicals and not considered as an outlier. Equation 4.8 could be used to predict toxicity of a chemical to *C. vulgaris* with the restrictions that if a chemical has a leverage value less than 0.120 and the *T. pyriformis* toxicity value is between -0.652 and 2.71.

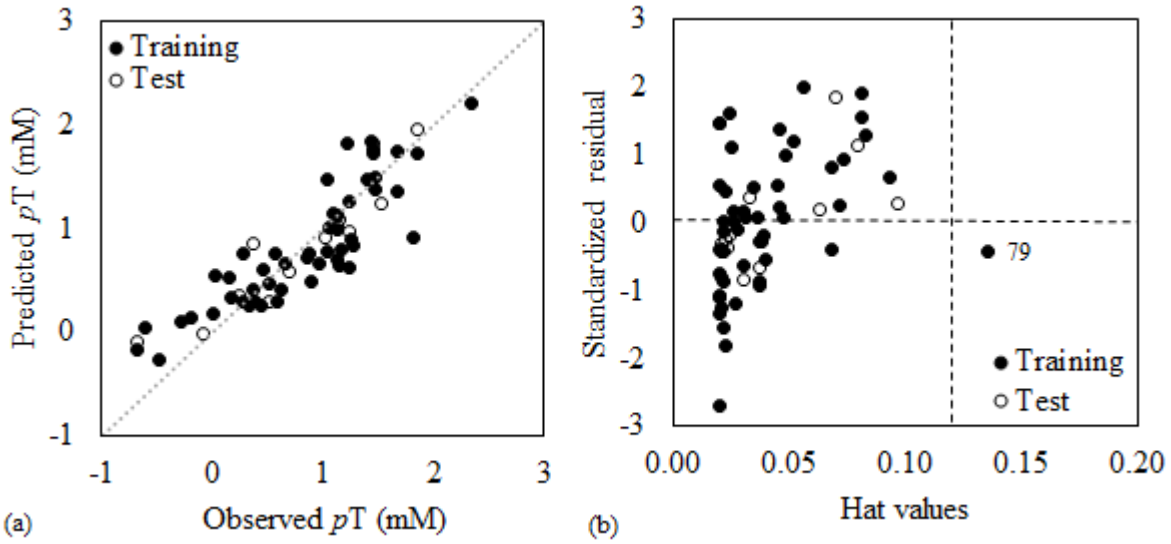


Figure. 4.16. Graphical representation of Eq. 4.8 for prediction of toxicity of *C. vulgaris* (a) Predicted vs. experimental toxicity values, (b) Williams plot for Eq. 4.8.

4.6.2. Algae-algae QTTR

72-h algal *(P.subcapitata*) toxicity data ($pT_{P.subcapitata}$=-log EC_{50}) were taken from Aruoja et al. (2011). There are 23 common chemicals between *C. vulgaris* and *P. subcapitata* data sets (Table 4.20).

Table 4.20. The interspecies model results using *P. subcapitata* toxicity for the prediction of *C. vulgaris* toxicity.

ID	Experimental $pT_{P.subcapitata}$	Experimental $pT_{C.vulgaris}$	Predicted $pT_{C.vulgaris}$ from Equation 4.9	Hat Val.	Std. Res.
1*	-0.070	-0.080	-0.312	0.239	-1.236
2	0.405	0.390	0.161	0.118	-1.136
3	0.802	0.440	0.563	0.067	0.592
4*	0.575	0.330	0.342	0.090	0.057
5	0.468	0.140	0.231	0.106	0.448
6	0.582	0.580	0.34	0.090	-1.162
7	0.653	0.510	0.412	0.081	-0.475
8	0.750	0.240	0.513	0.071	1.316
63	-0.321	-0.600	-0.563	0.329	0.209
64	0.395	0.170	0.151	0.120	-0.096
65	1.049	0.360	0.814	0.059	2.178
66*	0.612	0.460	0.372	0.086	-0.429
67*	1.175	1.090	0.935	0.061	-0.745
68	1.302	1.240	1.066	0.069	-0.841
69	1.646	1.120	1.417	0.114	1.470
70	1.005	0.880	0.774	0.059	-0.509
71	1.872	1.470	1.639	0.160	0.856
72	1.890	1.670	1.659	0.165	-0.058
73	1.676	1.640	1.448	0.119	-0.954
74	1.941	1.860	1.709	0.178	-0.775
75*	1.390	1.510	1.156	0.077	-1.714
76*	1.416	1.670	1.186	0.080	-2.347
77	1.544	1.530	1.307	0.096	-1.092

* Test set chemical

The high correlation (R=0.95) between these species led to a predictive QTTR. The interspecies QTTR for *C.vulgaris* and *P.subcapitata* is given in Equation 4.9:

$$pT_{C.vulgaris} = 1.005\ (\pm 0.175)\ pT_{P.subcapitata} - 0.241(\pm 0.213) \qquad (4.9)$$

n_{tr} = 17, R^2_{tr} = 0.909, Q^2_{LOO} = 0.889, $RMSE_{tr}$ = 0.202, SE= 0.215, F= 149.501

n_{test} = 6, R^2_{test} = 0.953, $RMSE_{test}$ =0.272,

Q^2_{F1}= 0.819, Q^2_{F2}= 0.818, Q^2_{F3}= 0.83, CCC_{test} = 0.900, r^2_m aver. = 0.862

The graph of predicted vs observed values and the Williams plot of Eq. 4.9 are available in Figure 4.17. All chemicals had standardized residuals less than 3 units and leverages less

than the critical hat value (h^*=0.350). Eq. 4.9 could be used to predict toxicity of a chemical to *C.vulgaris* with the restrictions that if a chemical has a leverage value less than 0.353 and *P.subcapitata* toxicity value is between -0.32 and 1.94.

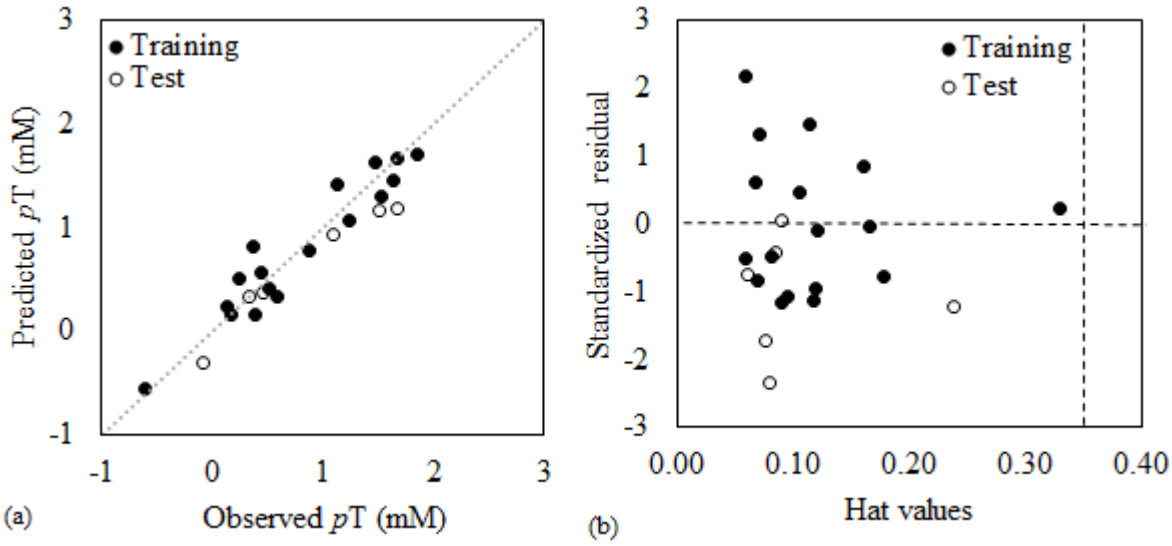


Figure 4.17. Graphical representation of Eq. 4.9 for the prediction of toxicity of *C.vulgaris* (a) Predicted vs. experimental toxicity values, (b) Williams plot for Eq. 4.9.

The relation between the modeled algal species was seen to be more correlated than the ciliate-algae one. This issue was reviewed by Kar et al. (2016). The correlations are more profound for close taxonomic groups than for more distant groups. In interspecies models, usually additional descriptors are needed. In the present data sets, the models did not need additional descriptors. This demonstrated that developed models are simple and easy to use. Since there is no need to add an extra descriptor to interspecies models, it is likely that MOAs of the chemicals and their uptakes of the correlated species may be similar.

5. CONCLUSIONS

Acute and chronic algal toxicity values of selected phenol and aniline derivatives were determined using standardized algal growth inhibition assays. A significant amount of chemicals' toxicity values was reported for the first time in the literature. With the transmission of new toxicity values, high quality and single source data generated have been contributed to the literature. REACH regulation requires ecotoxicity information of chemicals in order to protect the environment and human health. Considering the high number of chemicals to be tested, alternative methods to bioassays such as QSTRs are favorable. To this end, single source toxicity data is valuable for the risk assessment regulations.

Visual analysis of the data set by plotting the algal toxicity of phenols against their hydrophobicity revealed that chlorophenols, methylphenols, resorcinols, and nitroanilines fell within the domain of polar narcosis, whereas polyphenols with *ortho* and *para* substitution, dinitrophenols, and dinitroanilines displayed toxicity in excess of that predicted by hydrophobicity. The analysis also pointed out that classifying dinitrophenols, tetrachlorophenols, and pentachlorophenol in the same mode of toxic action might be misleading and a new classification scheme should be used to differentiate bulky chlorophenols and dinitrophenols. Although anilines appear to be polar narcosis, they were shown not to be significantly correlated with hydrophobicity.

Various toxic endpoints of 84 phenol and aniline derivatives were modeled. Considering our previous experience and reputation of hydrophobicity parameters in toxicity models, log *P* and log *D* were given priority in modelling part. Linear (MLR) and nonlinear (SVR and BPNN) models were developed, then their performances were compared using proper statistical parameters.

Six significant and validated QSTR models developed in this study highlighted the importance of hydrophobicity, electronic properties, and structure of molecules on the acute toxicity of phenols to *C. vulgaris*. All models applied on an external set of chemicals to

evaluate their predictive ability on untested chemicals. Application of models to predict *C. vulgaris* toxicities without experimental data revealed that toxicities of chemicals in the AD could be predicted confidently. Models were found to have high prediction coverage on this diverse data set consisting of emerging pollutants such as pesticides, pharmaceuticals, phthalates, and phenol and aniline derivatives.

Average ACRs of the data set were found to be in congruence with ECOSAR and ECETOC database average values and other literature studies. In addition, average MATC of the present study was comparable with ECOSAR average MATC. A factor of 10 for general estimation of NOEC for algae is protective for phenols and anilines, since IC_{50}/NOEC is less than 10 on the average.

Robust and predictive QSAR models for the estimation of NOEC and IC_{20} for *C. vulgaris* were developed and validated according to the OECD principles. Prediction of NOEC is constrained by the fact that it is a test design dependent parameter, i.e. the reported NOEC values do not represent the actual NOEC, but rather they are dependent on the concentrations tested. Prediction of IC_{20} is found to be more convenient than the prediction of NOEC. While models with theoretical descriptors had satisfactory statistics, the model with empirical variable (IC_{50}) was more predictive (better test set statistics). The models developed can be used in risk assessment by estimating the low-toxic-effects of new/untested chemicals regardless of their MOAs, within the AD of the developed models.

Interspecies models were developed and used to predict *C. vulgaris* toxicity with ciliate and another algae species, namely, *Pseudokirchneriella subcapitata* and *Tetrahymena pyriformis*, respectively. The generated QTTR models could be considered to fill *C. vulgaris* data gaps for chemicals with no experimental data, which fall in the AD of each model, as well as for prioritization and screening of chemicals.

REFERENCES

Abbasitabar, F., Zare-Shahabadi, V., 2017. In silico prediction of toxicity of phenols to *Tetrahymena pyriformis* by using genetic algorithm and decision tree-based modeling approach. Chemosphere, 172, 249-259.

ADMET Predictor v.8, Simulations Plus, 2016. http://www.simulations-plus.com/.

Ahlers, J., Riedhammer, C., Vogliano, M., Ebert, R.-U., Kühne, R., Schüürmann, G., 2006. Acute to chronic ratios in aquatic toxicity—variation across trophic levels and relationship with chemical structure. Environmental Toxicology and Chemistry, 25, 2937-2945.

Austin, T.J., Eadsforth C.V., 2014. Development of a chronic fish toxicity model for predicting sub-lethal NOEC values for non-polar narcotics. SAR and QSAR in Environmental Research, 25, 147–160.

APHA-AWWA-WEF, 2012. Algae. In: Standard Methods for the Determination of Water and Wasteware, Eds., A.E. Greenberg, 22nd ed. Prepared by American Public Health Association, American Water Works Association, and Water Environment Federation. American Public Health Association, Washington, DC, Part 8110, pp. 850-856.

Aptula A.O., Roberts, D.W., Cronin, M.T.D., Schultz, T.W., 2005. Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. Chemical Research in Toxicology, 18, 844-854.

Aruoja, V., Sihtmäe, M., Dubourguier, H.-C., Kahru, A., 2011. Toxicity of 58 substituted anilines and phenols to algae *Pseudokirchneriella subcapitata* and bacteria *Vibrio fischeri*: Comparison with published data and QSARs. Chemosphere, 84, 1310-1320.

Asadollahi-Baboli, M., 2012. Exploring QSTR analysis of the toxicity of phenols and thiophenols using machine learning methods. Environmental Toxicology and Pharmacology, 34, 826-831.

Ballabio, D., Consonni, V., Todeschini R., 2009. The Kohonen and CP-ANN toolbox: a collection of MATLAB modules for Self Organizing Maps and Counterpropagation Artificial Neural Networks. Chemometrics and Intelligent Laboratory Systems, 98, 115-122.

Ballabio, D., Vasighi, M., 2012. A MATLAB Toolbox for Self Organizing Maps and supervised neural network learning strategies. Chemometrics and Intelligent Laboratory Systems, 118, 24-32.

Bao, Y., Huang, Q., Li, Y., L., N., He, T., Feng, C., 2012. Prediction of nitrobenzene toxicity to the algae (*Scenedesmus obliguus*) by quantitative structure–toxicity relationship (QSTR) models with quantum chemical descriptors. Environmental Toxicology and Pharmacology, 33, 39-45.

Baskin, I.I., Ait, A.O., Halberstam, N.M., Palyulin, V.A. Zefirov, N.S., 2002. An approach to the interpretation of backpropagation neural network models in QSAR studies. SAR and QSAR in Environmental Research, 13, 35-41.

Baykal, T., Açıkgöz, İ., Udoh, A.U., Yıldız, K., 2011. Seasonal variations in phytoplankton composition and biomass in a small lowland river-lake system (Melen River, Turkey). Turkish Journal of Biology, 35, 485-501.

Beasley, A., Belanger, S.E., Brill, J.L., Otter, R.R., 2015. Evaluation and comparison of the relationship between NOEC and EC10 or EC20 values in chronic Daphnia toxicity testing. Environmental Toxicology and Chemistry, 34, 2378-84.

Bradbury, S.P., Lipnick, R. L., 1990. Introduction: structural properties for determining mechanisms of toxic action. Environmental Health Perspectives, 87, 181-182.

Burden, F. R., 2001. Quantitative Structure-Activity Relationship studies using Gaussian Processes. Journal of Chemical Information and Computer Sciences, 41, 830-835.

Caballero, J., Fernandez, M., 2006. Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. Journal of Molecular Modeling, 12,168-81.

Cai, X.Y., Ye, J., Sheng, G., Liu, W., 2009. Time-dependent degradation and toxicity of diclofop-methyl in algal suspensions. Environmental Science and Pollution Research, 16, 459-465.

Carlsson L., Helgee, E. A., Boyer, S., 2009. Interpretation of nonlinear QSAR models applied to ames mutagenicity data. Journal of Chemical Information and Modeling, 49, 2551–2558.

Cassotti, M., Grisoni, F., Todeschini, R., 2014. Reshaped Sequential Replacement algorithm: An efficient approach to variable selection. Chemometrics and Intelligent Laboratory Systems, 133, 136–148.

Chaminda Lakmal, H.H., Samarakoon, K. W. Jeon, Y.-J., 2015. Enzyme-Assisted Extraction to Prepare Bioactive Peptides from Microalgae. In: Marine Algae Extracts: Processes, Products, and Applications. Kim, S.-K., Chojnacka, K. (Eds) 310, Wiley-VCH, Weinheim, Germany.

Chen, Y.C., Wang, Y.J., Yang, C.F., 2009. Estimating low-toxic-effect concentrations in closed-system algal toxicity tests. Ecotoxicology and Environmental Safety, 72, 1514–1522.
Chen, X. H., Shan, Z. J., Zhai, H. L., 2017. QSAR models for predicting the toxicity of halogenated phenols to Tetrahymena. Toxicological and Environmental Chemistry, 99, 273-284.

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, 2320−2335.

Chirico, N., Gramatica, P., 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, 2044-2058.

Christensen, E.R., Kusk, K.O., Nyholm, N., 2009. Dose-response regressions for algal growth and similar continuous endpoints: calculation of effective concentrations. Environmental Toxicology and Chemistry, 28, 826-835.

Claeys, L., Iaccino, F., Janssen, C.R., Van Sprang, P., Verdonck, F., 2013. Development and validation of a quantitative structure-activity relationship for chronic narcosis to fish. Environmental Toxicology and Chemistry, 32, 2217-2225.

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

Consonni, V., Todeschini, R., 2010. Molecular Descriptors. In: Recent Advances in QSAR Studies: Methods and Applications. Puzyn, T., Leszczynski, J., Cronin, M. T. (Eds) Springer Science&Business Media B.V., Netherlands.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning, 20, 273-297.

Cronin, M.T.D., Aptula, A.O., Duffy, J.C., Netzeva, T.I., Rowe, P.H. , Valkova, I.V., Schultz, T.W., 2002. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere, 49, 1201–1221.

Cronin, M. T. D., Netzeva, T. I., Dearden, J. C., Edwards, R., Worgan, A. D. P., 2004. Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database. Chemical Research in Toxicology, 17, 545–554.

Cronin, M.T.D., 2010. Quantitative Structure-Activity Relationships (QSARs)-Applications and Methodology, in Recent Advances in QSAR Studies Methods and Applications, (Ed.) Puzyn, T., Leszczynski, J., Cronin, M. T. D., Springer.

Cronin, M.T.D., 2017. (Q)SARs to predict environmental toxicities: current status and future needs. Environmental Science: Processes & Impacts, 19, 213-220.

Danish (Q)SAR Database, Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark. http://qsar.food.dtu.dk. (accessed May 2017).

Devillers, J. (ed.), 1996. Neural Networks in QSAR and Drug Design, Academic Pres Inc., San Diego, CA.

Dieguez-Santana, K., Pham-The, H., Villegas-Aguilar, P.J., Le-Thi-Thu, H., Castillo-Garit, J.A., Casanola-Martin, G.M., 2016. Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for *Tetrahymena pyriformis* contaminant identification in a median-size database. Chemosphere, 165, 434-441.

Dimou, A.D., Sakellarides, T.M., Vosniakos, F.K., Giannoulis, N., Leneti, E., Albanis, T., 2006. Determination of phenolic compounds in the marine environment of Thermaikos Gulf, Northern Greece. International Journal of Environmental Analytical Chemistry, 86, 119–130.

Dobchev, D.A., Tulp, I., Karelson, G., Tamm, T., Tämm, K., Karelson, M., 2013. Subchronic oral and inhalation toxicities: a challenging attempt for modeling and prediction. Molecular Informatics, 32, 793-801.

DRAGON for Windows 6.0.38, Talete srl, 2014, Milan, http://www.talete.mi.it/.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V., 1996. Support Vector Regression Machines. In Mozer, M., Jordan, M.I., Petsche, T. (Eds), Advances in Neural Information Processing Systems 9, 155-161, MIT Press, Denver.

Doucet, J. P., Panaye, A., 2010. Three Dimensional QSAR, Applications in Pharmacology and Toxicology. CRC Press, FL.

EC, 2006. European Commission, Regulation No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). Official Journal of the European Union, L 396/1-849, Brussels, Belgium.ECETOC, 2003 (European Centre for Ecotoxicology and Toxicology of Chemicals), TR 091 – ECETOC Aquatic Toxicity (EAT) database.
http://www.ecetoc.org/publications/technical-reports. (accessed May 2017).

ECHA 2008, Guidance on information requirements and chemical safety assessment, Chapter R.6: QSARs and grouping of chemicals.
http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf.
(accessed May 2017)

ECHA Database. https://echa.europa.eu/information-on-chemicals. (accessed May 2017).

ECOSAR v1.10. The ECOSAR Class Program for Microsoft Windows U.S. Environmental Protection Agency, 2011.

ECOSAR Methodology Document v1.11, 2012.
https://www.epa.gov/sites/production/files/2015-09/documents/ecosartechfinal.pdf.
(accessed May 2017).

ECOTOX, v. 4. https://cfpub.epa.gov/ecotox/ecotox_home.cfm. (accessed May 2017).
Egan, W. J., Morgan, S. L., 1998. Outlier detection in multivariate analytical chemical data. Analytical Chemistry, 70, 2372-2379.

Environment Canada, 2007. Environmental Science and Technology Centre, EPS 1/RM/25 2nd edition, Canada.

EPA, 2002. Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms, 4th edition, EPA-821-R-02-013.

EPA, 2008. EPA White paper: Aquatic life criteria for contaminants of emerging concern, OW/ORD Emerging Contaminants Workgroup, July 2008a. http://water.epa.gov/scitech/swguidance/standards/upload/2008_06_03_criteria_sab-emergingconcerns.pdf (accessed May 2017).

EPA, 2012. Ecological Effects Test Guidelines OCSPP 850.4500: Algal Toxicity.

Eriksson, L., Johansson, E., 1996. Multivariate design and modeling in QSAR. Chemometrics and Intelligent Laboratory Systems 34, 1-19.

Eroglu, E., Melis, A., 2016. Microalgal hydrogen production research. International Journal of Hydrogen Energy, 41, 12772–12798.

Ertürk, M.D., Saçan, M.T., 2012. First toxicity data of chlorophenols on marine alga *Dunaliella tertiolecta*: correlation of marine algal toxicity with hydrophobicity and interspecies toxicity relationships. Environmental Toxicology and Chemistry, 31, 1113-1120.

Ertürk, M.D., Saçan, M.T., Novic, M., Minovski, N., 2012. Quantitative structure–activity relationships (QSARs) using the novel marine algal toxicity data of phenols. Journal of Molecular Graphics and Modelling, 38, 90-100.

Ertürk, M. D., 2013. Assessment and modelling of the toxicity of phenols: A comparative study with marine and freshwater algae, Ph.D. Thesis, Bogazici University.

Ertürk, M.D., Saçan, M.T., 2013. Assessment and modeling of the novel toxicity data set of phenols to *Chlorella vulgaris*. Ecotoxicology and Environmental Safety, 90, 61–68.

Estrada, E., 1997. Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. Journal of Chemical Information and Computer Sciences, 37, 320-328.

Fan, D., Liu, J., Wang, L., Yang, X., Zhang, S., Zhang, Y., Shi, L., 2016. Development of Quantitative Structure–Activity Relationship models for predicting chronic toxicity of substituted benzenes to *Daphnia magna*. Bulletin of Environmental Contamination and Toxicology, 96, 664–670.

Fang, K-T., Yin, H., Liang, Y.-Z., 2004. New approach by Kriging models to problems in QSAR. Journal of Chemical Information and Computer Sciences, 44, 2106-2113.

Fatemi, M. H., Malekzadeh, H., 2012. *In silico* prediction of dermal penetration rate of chemicals from their molecular structural descriptors. Environmental Toxicology and Pharmacology, 34, 297-306.

Fernandez, M., Caballero, J., Fernandez, L., Sarai, A., 2011. Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). Molecular Diversity, 15, 269–289.

Fu, L., Li, J.J., Wang, Y., Wang, X.H., Wen, Y., Qin, W.C., Su, L. M., Zhao, Y.H., 2015. Evaluation of toxicity data to green algae and relationship with hydrophobicity. Chemosphere, 120, 16-22.

Furuhama, A., Hasunuma, K., Aoki, Y., 2015. Interspecies quantitative structure–activity–activity relationships (QSAARs) for prediction of acute aquatic toxicity of aromatic amines and phenols. SAR and QSAR in Environmental Research, 26, 301-323.

Geiger, E., Hornek-Gausterer, R., Saçan, M.T., 2016. Single and mixture toxicity of pharmaceuticals and chlorophenols to freshwater algae *Chlorella vulgaris*. Ecotoxicology and Environmental Safety, 129, 189-198.

GaussView 3.0, 2000-2003. Semichem Inc., Gaussian Inc., Pittsburgh, USA.

Ghasemi, J. B., Salahinejad, M., Rofouei, M. K., 2013. Alignment independent 3D-QSAR modeling of fullerene (C_{60}) solubility in different organic solvents. Fullerenes, Nanotubes and Carbon Nanostructures, 21, 367-380.

Golbraikh, A., Tropsha, A., 2002. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection**.** Journal of Computer-Aided Molecular Design, 16, 357–369.

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

Gramatica, P., 2013. On the Development and Validation of QSAR Models. In Brad Reisfeld and Arthur N. Mayeno (Eds), Computational Toxicology: Volume II (Methods in Molecular Biology, vol. 930), 499-526, Springer Science+Business Media, LLC, N.Y. USA.

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

Gramatica, P., Cassani, S., Chirico, N., 2014. QSARINS-Chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. Journal of Computational Chemistry, 35, 1036–1044.

Gramatica, P., Sangion, A., 2016. A historical excursus on the statistical validation parameters for QSAR Models: A clarification concerning metrics and terminology. Journal of Chemical Information and Modeling, 56, 1127-1131.

Gouveia, L., Batista, A.P., Miranda, A., Empis, J., Raymundo, A., 2007. *Chlorella vulgaris* biomass used as colouring source in traditional butter cookies. Innovative Food Science & Emerging Technologies, 8, 433–436.

Grisoni, F., Cassotti, M., Todeschini, R., 2014. Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods. Journal of Chemometrics, 28, 249-259.

Hawe, G. I. Alkorta, I, Popelier, P. L. A., 2010. Prediction of the basicities of pyridines in the gas phase and in aqueous solution. Journal of Chemical Information and Modeling, 50, 87–96.

He, L., Jurs, C., 2005. Assessing the reliability of a QSAR model's predictions. Journal of Molecular Graphics and Modelling, 23, 503–523.

Hoenicke, R., Oros, D.R., Oram, J.J., Taberski, K.M., 2007. Adapting an ambient monitoring program to the challenge of managing emerging pollutants in the San Francisco Estuary. Environmental Monitoring and Assessment, 81, 15-25.

Hoff, D., Lehmann, W., Pease, A., Raimondo, S., Russom, C., Steeger, T., 2010. Predicting the Toxicities of Chemicals to Aquatic Animal Species. EPA.

Huang, C.P., Wang, Y.J., Chen, C.Y., 2007. Toxicity and quantitative structure–activity relationships of nitriles based on *Pseudokirchneriella subcapitata*. Ecotoxicology and Environmental Safety, 67, 439–446.

ISO Water quality - freshwater algal growth inhibition test with unicellular green algae, 2004. International Organization for Standardization (ISO), Geneva, Switzerland.

Iwasaki, Y., Kotani, K., Kashiwada, S., Masunaga, S., 2015. Does the choice of NOEC or EC10 affect the hazardous concentration for 5% of the species? Environmental Science and Technology, 49, 9326-30.

Jager, T., 2012. Bad habits die hard: the NOEC's persistence reflects poorly on ecotoxicology. Environmental Toxicology and Chemistry, 31, 228-229.

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

Janus, J.A., Posthumus, R., 2002. Environmental Risk Limits for 2-propanol, formaldehyde and 4-chloromethylphenols - updated proposals, RIVM Report 601501015/2002, Netherlands.

Janssen, C.R. and Heijerick, D.G., 2003. Algal Toxicity Tests for Environmental Risk Assessments of Metals. In Ware, G. W. (Ed), Reviews of Environmental Contamination and Toxicology, 23-52, Springer, New York.

Jing, G., Zhou, Z., Zhuo, J., 2012. Quantitative structure–activity relationship (QSAR) study of toxicity of quaternary ammonium compounds on *Chlorella pyrenoidosa* and *Scenedesmus quadricauda*. Chemosphere, 86, 76-82.

Kar, S., Roy, K., 2010. First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals. Chemosphere, 81, 738-747.

Kar, S., Das, R. N., Roy, K., Leszczynski, J., 2016. Can toxicity for different species be correlated?: The concept and emerging applications of interspecies Quantitative Structure-Toxicity Relationship (i-QSTR) modeling. International Journal of Quantitative Structure-Property Relationships, 1, 23-51.

Koller, G., Hungerbühler, K., Fent, K., 2000. Data ranges in aquatic toxicity of chemicals, consequences for environmental risk analysis. Environmental Science and Pollution Research, 7, 135-143.

Kuhn, R., Pattard, M., 1990. Results of the harmful effects of water pollutants to green algae (*Scenedesmus subspicatus*) in the cell multiplication inhibition test. Water Research, 24, 31-38.

Kumar, A., Batley, G.E., Nidumolu, B., Hutchinson, T.H., 2016. Derivation of water quality guidelines for priority pharmaceuticals. Environmental Toxicology and Chemistry, 35, 1815–1824.

Kuzmin, V. E., Artemenko, A. G., Muratov, E. N., Polischuk, P. G., Ognichenko, L. N., Liahovsky, A. V., Hromov, A. I., Varlamova, E. V., 2010. Virtual screening and molecular design based on hierarchical QSAR technology, in Recent Advances in QSAR Studies Methods and Applications, (Ed.) Puzyn, T., Leszczynski, J., Cronin, M. T. D., Springer.

Landis, W.G., Chapman, P.M., 2011. Well past time to stop using NOELs and LOELs. Integrated environmental assessment and management, 7, VI-VIII.

Liang, Y., Xu, Q.-S., Li, H.-D., Cao, D.-S., 2011. Support Vector Machines and QSAR/QSPR. In Support Vector Machines and Their Application in Chemistry and Biotechnology, 115-147, CRC Press, FL.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45, 255–268.

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

Lim, S.-L., Chu, W.-L., Phang, S.-M., 2010. Use of *Chlorella vulgaris* for bioremediation of textile wastewater. Bioresource Technology, 101, 7314–7322.

Lu, G-H., Wang, C., Guo, X-L., 2008. Prediction of toxicity of phenols and anilines to algae by quantitative structure-activity relationship. Biomedical and Environmental Sciences, 21, 193-196.

Ma, J., Xu, L., Wang, S. Zheng, R., Jin, S., Huang S., Huang Y., 2002. Toxicity of 40 herbicides to the green alga *Chlorella vulgaris*. Ecotoxicology and Environmental Safety, 51, 128-132.

Madhavi, D.R., Umamaheswari, A., Venkateswarlu, K., 1995. Effective concentrations toward growth yield of selected microalgae and cyanobacteria isolated from soil. Ecotoxicology and Environmental Safety, 32, 205-208.

Mallick, N., Mandal, S., Singh, A.K., Bishai, M., Dash, A., 2011. Green microalga *Chlorella vulgaris* as a potential feedstock for biodiesel. Journal of Chemical Technology and Biotechnology, 87, 137-145.

Mandaric, L., Celic, M., Marce, R., Petrovic, M., 2016. Introduction on Emerging Contaminants in Rivers and Their Environmental Risk. In Petrovic M., Sabater S., Elosegi A., Barcelo D. (Eds), Emerging Contaminants in River Ecosystems: Occurrence and Effects under Multiple Stress Conditions, 20-21, Springer, Switzerland.

Martincic, R., Kuzmanovski, I., Wagner, A., Novic, M., 2015. Development of models for prediction of the antioxidant activity of derivatives of natural compounds. Analytica Chimica Acta, 868, 23-35.

May, M., Drost, W., Germer, S., Juffernholz, T., Hahn, S., 2016. Evaluation of acute-to-chronic ratios of fish and *Daphnia* to predict acceptable no‑effect levels. Environmental Sciences Europe, 28, 1-9.

Mazzatorta, P., Smiesko, M., Lo Piparo, E., Benfenati, E., 2005. QSAR model for predicting pesticide aquatic toxicity. Journal of Chemical Information and Modeling, 45, 1767-1774.

McGrath, J.A., Parkerton, T.F., Di Toro, D.M., 2004. Application of the narcosis target lipid model to algal toxicity and deriving predicted-no-effect concentrations. Environmental Toxicology and Chemistry, 23, 222–236.

MDM (Molegro Data Modeller) 2.6.0, 2011. Molegro ApS., http://www.molegro.com.

Michalowicz, J., Duda, W., 2007. Phenols – Sources and toxicity. Polish Journal of Environmental Studies, 16, 347-362.

Michielan, L., Pireddu, L., Floris, M., Moro, S., 2010. Support Vector Machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals. Molecular Informatics, 29, 51-64.

Minovski, N., Zuperl, S., Drgan, V., Novic, M., 2013. Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study. Analytica Chimica Acta, 759, 28– 42.

Moermond, C.T.A. Heugens, E.H.W., 2009a. Environmental risk limits for monochlorophenols, 4-chloro-3-methylphenol and aminochlorophenol, The Netherlands. RIVM Report 601714006/2009

Moermond, C.T.A. Heugens, E.H.W., 2009b. Environmental risk limits for 2,4-dichlorophenol, The Netherlands. RIVM Report 601714007/2009

Moermond, C.T.A. Heugens, E.H.W., 2009c. Environmental risk limits for trichlorophenols, The Netherlands. RIVM Report 601714005/2009

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to Linear Regression Analysis, 4-5. John Wiley & Sons, New Jersey.

Morel, O.J.X., Christie, R.M., 2011. Current trends in the chemistry of permanent hair dyeing. Chemical Reviews, 111, 2537–2561.

Murkovski, A. and Skórska, E. 2010. Effect of (C6H5)3PbCl and (C6H5)3SnCl on delayed luminescence intensity, evolving oxygen and electron transport rate in photosystem II of *Chlorella vulgaris*. Bulletin of Environmental Contamination and Toxicology, 84, 157-160.

Netzeva, T.I., Pavan, M., Worth, A.P., 2008. Review of (Quantitative) Structure – Activity Relationships for acute aquatic toxicity. QSAR and Combinatorial Science, 27, 77-90.

Newman, M.C., 2008. "What exactly are you inferring?" A closer look at hypothesis testing. Environmental Toxicology and Chemistry, 27, 1013-1019.

Niculescu, S. P., 2003. Artificial neural networks and genetic algorithms in QSAR. Journal of Molecular Structure, 622, 71-83.

Norberg-King T. 1988. An interpolation estimate for chronic toxicity: The ICp approach. National Effluent Toxicity Assessment Center. Technical Report 05–88. U.S. Environmental Protection Agency, Duluth, MN, USA.

Obrezanova, O., Csanyi, G., Gola, J. M. R., Segall, M. D., 2007. Gaussian processes: A method for automatic QSAR modeling of ADME properties. Journal of Chemical Information and Modeling, 47, 1847-1857.

OECD, 1998. ENV/MC/CHEM(98)18, OECD Series on Testing and Assessment Number 10. Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data.
OECD SIDS, 2001. http://www.inchem.org/documents/sids/sids/nitroaniline.pdf. (accessed May 2017).

OECD, 2007. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69, ENV/JM/MONO(2007)2, Paris.
http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2007)2&docLanguage=En. (accessed May 2017).

OECD, 2011. OECD Guidelines for the Testing of Chemicals. Guideline 201, Alga, Growth Inhibition Test. Paris, France.

Önlü, S., Saçan, M.T., 2016. An in silico algal toxicity model with a wide applicability potential for industrial chemicals and pharmaceuticals. Environmental Toxicology and Chemistry, Vol. 999, 1-8.

Örücü, E., Tugcu, G., Saçan, M.T., 2014. Molecular structure–adsorption study on current textile dyes. SAR and QSAR in Environmental Research, 25, 983-998.

Panaye, A., Fan, B.T., Doucet, J.P., Yao, X.J., Zhang, R.S., Liu, M.C., Hu, Z.D., 2006. Quantitative structure-toxicity relationships (QSTRs): A comparative study of various non linear methods. General regression neural network, radial basis function neural network and support vector machine in predicting toxicity of nitro- and cyanoaromatics to *Tetrahymena pyriformis*. SAR and QSAR in Environmental Research, 17, 75-91.

Papa E., Dearden, J.C., Gramatica, P., 2007. Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors, Chemosphere, 6, 351-358.

Pramanik, S., Roy, K., 2014. Predictive modeling of chemical toxicity towards *Pseudokirchneriella subcapitata* using regression and classification based approaches. Ecotoxicology and Environmental Safety, 101, 184-190.

Priyadarshani, I., Rath, B., 2012. Commercial and industrial applications of micro algae – A review. Journal of Algal Biomass Utilization, 3, 89-100.

Raeesossadati, M.J., Ahmadzadeh, H., McHenry, M.P., Moheimani, N.R., 2014. CO_2 bioremediation by microalgae in photobioreactors: Impacts of biomass and CO_2 concentrations, light, and temperature. Algal Research, 6, 78-85.

Raevsky O. A., Liplavskaya, E. A., Yarkov, A. V., Raevskaya, O. E., Worth, A. P., 2011. Linear and nonlinear QSAR models of acute intravenous toxicity of organic chemicals for mice. Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry, Vol. 5, No. 3, 213–225.

Rai, U., Deshar, G., Rai, B., Bhattarai, K., Dhakal, R.P., Rai, S.K., 2013. Isolation and culture condition optimization of *Chlorella vulgaris*. Nepal Journal of Science and Technology, 14, 43-48.

Reynolds, C.S., 1984. The Ecology of Freshwater Phytoplankton, Cambridge University Press, UK.

Roex, E.W.M., Van Gestel, C.A.M., Van Wezel, A.P., Van Straalen, N.M., 2000. Ratios between acute aquatic toxicity and effects on population growth rates in relation to toxicant mode of action. Environmental Toxicology and Chemistry, 19, 685-693.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection, John Wiley & Sons, USA.

Roy, K., Kar, S., Das, R.N., 2015a. QSAR/QSPR Modeling: Introduction. In A Primer on QSAR/QSPR Modeling, Springer, London.

Roy, K., Kar, S., Ambure, P., 2015b. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems, 145, 22–29.

Roy, K., Ambure, P., Aher, R. B., 2017. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometrics and Intelligent Laboratory Systems, 162, 44-54.

Roy, K., Roy, P.P., 2009. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. European Journal of Medicinal Chemistry, 44, 2913-2922.

Saçan, M. T., Vracko, M., Tugcu, G., Modelling the Toxicity of Organic Compounds to Freshwater Algae Using Counter-Propagation Neural Networks and Linear Techniques 14th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences. 24-28 May 2010.

Safi, C., Zebib, B., Merah, O., Pontalier, P.-Y., Vaca-Garcia, C., 2014. Morphology, composition, production, processing and applications of *Chlorella vulgaris*: A review. Renewable and Sustainable Energy Reviews, 35, 265-278.

Sanderson, H., Johnson, D.J., Reitsma, T., Brain, R.A., Wilson, C.J., Solomon, K.R., 2004. Ranking and prioritization of environmental risks of pharmaceuticals in surface waters. Regulatory Toxicology and Pharmacology, 39, 158-183.

Schmitt, H., Altenburger, R., Jastorff, B., Schüürmann, G., 2000. Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds, Chemical Research in Toxicology, 13, 441-450.

Schultz, T.W., Sinks, G.D., Bearden, A.P., 1998. QSAR in aquatic toxicology: a mechanism of action approach comparing toxic potency to *Pimephales promelas*, *Tetrahymena pyriformis*, and *Vibrio fischeri*. in Comperative QSAR (J. Devillers ed.) pp. 51-109.

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.
Selassie, C., Verma, R.P., 2015. QSAR of toxicology of substituted phenols. Journal of Pesticide Science, 40, 1–12.

Sevcik, P., Cik, G., Sersen, F., 2009. Inhibition of toxic effects of chlorophenols on the growth of *Chlorella vulgaris* by modified TiO2 photocatalyst. Fresenius Environmental Bulletin, 18, 2165-2169.

Shieh, J.-N., Chao, M.-R., Chen, C-Y., 2001. Statistical comparisons of the no-observed-effect concentration and the effective concentration at 10% inhibition (EC10) in algal toxicity tests. Water Science and Technology, 43, 141-146.

Singh, K.P., Gupta, S., Kumar, A., Mohan, D., 2014. Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. Chemical Research in Toxicology, 27, 741-753.

SPARTAN v. 10, Wavefunction, Inc., 2011, Irvine, USA, http://wavefun.com.

SPSS Statistics 17.0 for Windows (Statistical Package for Social Scientists), 2008. SPSS Inc., USA.

Stauber, J., Franklin, N., Adams, M., 2005. Microalgal Toxicity Tests Using Flow Cytometry. In Blaise, C., Ferard, J.-F. (Eds), Small-scale Freshwater Toxicity Investigations, Vol.1: Toxicity Test Methods, Springer, Netherlands.

Staveley, J.P., Smrchek, J.C., 2005. Algal Toxicity Test. In Blaise, C., Ferard, J.-F. (Eds), Small-scale Freshwater Toxicity Investigations, Vol.1: Toxicity Test Methods, Springer, the Netherlands.

Sullivan, K.M., Manuppello, J.R., Willett, C.E., 2014. Building on a solid foundation: SAR and QSAR as a fundamental strategy to reduce animal testing. SAR and QSAR in Environmental Research, 25, 357–365.

Sun, L.W., Qu, M.M., Li, Y.Q., Wu, Y.L., Chen, Y.G., Kong, Z.M., Liu, Z.T., 2004. Toxic effects of aminophenols on aquatic life using the zebrafish embryo test and the comet assay. Bulletin of Environmental Contamination and Toxicology, 73, 628–634.

Tamhane, A.C., Dunlop, D.D., 2000. Statistics and Data Analysis, 427-428, Prentice-Hall, NJ.

Tamura, I., Kagota, K., Yasuda, Y., Yoneda, S., Morita, J., Nakada, N., Kameda, Y., Kimura, K., Tatarazako, N., Yamamoto, H., 2013. Ecotoxicity and screening level ecotoxicological risk assessment of five antimicrobial agents: triclosan, triclocarban, resorcinol, phenoxyethanol and *p*-thymol. Journal of Applied Toxicology, 33, 1222-1229.

Tas, B., Gonulol, A., 2007. An ecologic and taxonomic study on phytoplankton of a shallow lake, Turkey. Journal of Environmental Biology, 28, 439-445.

Technical guidance document (TGD) in support of Commission Directive 93/67/EEC on Risk assessment for new notified substances and Commission Regulation (EC) No 1488/94

on Risk assessment for existing substances and Commission Directive (EC) 98/8 on biocides, 2003. European Commission, 2nd edition, Luxembourg, Part 1, 2 and 3, 760.

Timofei, S., Kurunczi, L., Suzuki, T., Fabian, W. M. F., Mureşan, S., 1997. Multiple Linear Regression (MLR) and Neural Network (NN) Calculations of some disazo dye adsorption on cellulose. Dyes and Pigments, 34, 181-193.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics. Vol. 41, 2nd edition, Wiley-VCH, Weinheim.

Tokuşoglu, Ö., Ünal, M.K., 2003. Biomass nutrient profiles of three microalgae: *Spirulina platensis*, *Chlorella vulgaris*, and *Isochrisis galbana*. Journal of Food Science, 68, 1144-1148.

Toxcalc ver. 5.0.32, 2009. Tidepool Scientific Software, CA, USA.

Tugcu, G., Saçan, M.T., Vracko, M., Novic, M., Minovski, N., 2012. QSTR modeling of the acute toxicity of pharmaceuticals to fish. SAR QSAR in Environmental Research, 23, 297-310.

Tugcu, G., Yilmaz, H.B., Saçan, M.T., 2014. Comparative performance of descriptors in a multiple linear and Kriging models: a case study on the acute toxicity of organic chemicals to algae. Environmental Science and Pollution Research, 21, 11924-11932.

Tugcu, G., Ertürk, D.M., Saçan, M.T., 2017. On the aquatic toxicity of substituted phenols to *Chlorella vulgaris*: QSTR with an extended novel data set and interspecies models. Under review.

Urrestarazu Ramos, U., Vaes, W.H.J., Mayer, P., Hermens, J.L.M., 1999. Algal growth inhibition of *Chlorella pyrenoidosa* by polar narcotic pollutants: toxic cell concentrations and QSAR modeling. Aquatic Toxicology, 46, 1-10.

van Dam, R.A., Harford, A.J., Warne, M.S., 2012. Time to get off the fence: the need for definitive international guidance on statistical analysis of ecotoxicity data. Integrated environmental assessment and management, 8, 242-245.

Vanthoor-Koopmans, M., Cordoba-Matson, M.V., Arredondo-Vega, B.O., Lozano-Ramirez, C., Garcia-Trejo, J.F., Rodriguez-Palacio, M.C., 2014. Microalgae and Cyanobacteria Production for Feed and Food Supplements. In Biosystems Engineering: Biofactories for Food Production in the Century XXI. Guevara-Gonzalez, R., Torres-Pacheco, I. (Eds.), Springer, Switzerland.

Ventura, S.P.M., Gonçalves, A.M.M., Gonçalves, F., Coutinho, J.A.P., 2010. Assessing the toxicity on [C3mim][Tf2N] to aquatic organisms of different trophic levels. Aquatic Toxicology, 96, 290-297.

Vracko, M., 2005. Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. Current Computer-Aided Drug Design, 1, 73-78.

Vo, T.-S., Ngo, D.-H., Kim, S-K., 2012. Marine algae as a potential pharmaceutical source for anti-allergic therapeutics. Process Biochemistry, 47, 386–394.

Wang, C., Lu, G., Tang, Z., Guo, X., 2007. Quantitative structure-activity relationships for joint toxicity of substituted phenols and anilines to *Scenedesmus obliquus*. Journal of Environmental Sciences, 20, 115-119.

Xing, L., Liu, H., Giesy, J.P., Yu, H., 2012. pH-dependent aquatic criteria for 2,4-dichlorophenol, 2,4,6-trichlorophenol and pentachlorophenol. Science of the Total Environment, 15, 125-131.

Xu, X., Zhang, W., Huang, C., Li, Y., Yu, H., Wang, Y., 2012. A novel chemometric method for the prediction of human oral bioavailability. Int. J. Mol. Sci., 13, 6964-6982.

Xuan, S., Wu, Y., Chen, X. Liu, J., Yan, A., 2013. Prediction of bioactivity of HIV-1 integrase ST inhibitors by multilinear regression analysis and support vector machine. Bioorganic & Medicinal Chemistry Letters, 23, 1648–1655.

Yan, X.F., Xiao, H.M., Gong, X.D., Ju, X.H., 2005. Quantitative structure–activity relationships of nitroaromatics toxicity to the algae (*Scenedesmus obliguus*), Chemosphere, 59, 467-471.

Yee, L.C., Wei, Y.C., 2012. Current Modeling Methods Used in QSAR/QSPR. In Dehmer, M., Varmuza, K., Bonchev, D. (Eds), Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH, Germany.

Zhang, X.J., Qin, H.W., Su, L.M., Qin, W.C., Zou, M.Y., Sheng, L.X., Zhao, Y.H., Abraham, M.H., 2010. Interspecies correlations of toxicity to eight aquatic organisms: Theoretical considerations. Science of the Total Environment, 408, 4549-4555.

Zhou, W., Wu, S., Dai, Z., Chen, Y., Xiang, Y., Chen, J., Sun, C., Zhou, Q., Yuan Z., 2015. Nonlinear QSAR models with high-dimensional descriptor selection and SVR improve toxicity prediction and evaluation of phenols on *Photobacterium phosphoreum*. Chemometrics and Intelligent Laboratory Systems, 145, 30–38.

Zupan J., Gasteiger, J., 1999. Neural Networks in Chemistry and Drug Design, 273-274, Wiley-VCH, Weinheim.

APPENDIX A: FORMULATIONS USED IN VALIDATIONS

Table A.1. Parameters and criteria for validation of QSAR models.

Name	Formula	Definitions
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$	n is the number of compounds, y_i is observed and $\hat{y}_i$ is predicted toxicity value
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$	
Standardized residual	$SR_i = \frac{\hat{y}_i - y_i}{sd}$	SR_i is the standardized residual of the i^{th} chemical, $\hat{y}_i$ is the predicted and y_i is the observed toxicity value, and sd is the standard deviation of the errors
CCC for the test set	$CCC = \frac{2\sum_{i=1}^{n_{test}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{test}}(y_i - \bar{y})^2 + \sum_{i=1}^{n_{test}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{test}(\bar{y} - \bar{\hat{y}})^2}$	y_i is observed and $\hat{y}_i$ is predicted toxicity value, n_{test} is the number of chemicals in the test set, $\bar{y}$is the mean of the test set chemicals, $\bar{\hat{y}}$ is the mean of predicted values
r_m^2	$r_m^2 = r^2\left(1 - \sqrt{r^2 - r_0^2}\right)$	r^2 is coefficient of determination for test set, r_0^2 is coefficient of determination (without intercept) for test set

Table A.1. continued.

Name	**Formula**	**Definitions**
External Q^2 parameters	$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2}$	y_i is observed and $\hat{y}_i$ is predicted toxicity value, n_{test} is the number of chemicals in the test set and n_{tr} is the number of chemicals in the training set, $\bar{y}_{tr}$ is the mean of the training set chemicals and $\bar{y}_{test}$ is the mean of the test set chemicals
	$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{test})^2}$	
	$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2\right] / n_{test}}{\left[\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2\right] / n_{tr}}$	

APPENDIX B: CHARACTERISTICS OF STUDIED CHEMICALS

Table B.1. The tested chemicals, their purities, source, and physicochemical properties.

ID	Purity	Source	Water solubility [a] mg L^{-1}	Vapor pressure [a] mm Hg	Log *P* [b]
1	99.9%	Fluka	25,900	0.299	1.95
2	99.9%	Fluka	4,570	0.089	2.48
3	99.6%	Fluka	7,870	0.102	2.3
4	99.9%	Fluka	3,540	0.156	2.33
5	99.9%	Fluka	6,050	0.171	2.36
6	99.6%	Fluka	4,760	0.0356	2.23
7	99.8%	Fluka	4,880	0.0405	2.35
8	99.8%	Fluka	4,900	0.0372	2.58
9	99%	Aldrich	40,000	0.0083	1.58
10	98%	Aldrich	16,632	0.00617	0.9
11	99%	Aldrich	17,200	0.00307	1.15
12	97%	Aldrich	16,632	0.00617	1.16
13	99%	Aldrich	16,632	0.00914	1.64
14	99%	Aldrich	762	0.0038	3.155
15	99.9%	Fluka	1,200	0.0174	2.73
16	99%	Aldrich	221,940	0.0000534	0.55
17	98%	Aldrich	185,670	0.000489	0.47
18	≥98%	Fluka	104,260	0.000128	0.91
19	97%	Aldrich	11,373	0.000204	1.69
20	98%	Aldrich	20,828	0.000146	2.322
21	97%	Aldrich	104,260	0.0000733	1.58
22	97%	Aldrich	4,213	0.173	2.705
23	99.5%	Ehrenstorfer	4,213	0.138	2.9
24	99.8%	Fluka	4,000	0.024	2.78
25	99%	Aldrich	3,830	0.05	3.1
26	99%	Ehrenstorfer	250	0.0018	3.27
27	≥99%	Fluka	2,500	0.113	1.79
28	99%	Sigma-	13,500	0.00118	2
29	99.9%	Fluka	11,600	0.0000979	1.91
30	99.9%	Fluka	2,790	0.00039	1.67
31	97%	Fluka	385	0.000122	1.75
32	≥97%	Aldrich	14,508	0.00000812	1.73
33	99.5%	Ehrenstorfer	3,510	0.0372	2.29
34	99%	Supelco	1,190	0.000222	2.48
35	99%	Aldrich	3,510	0.0372	2.37
36	≥97.5%	Aldrich	8,951	0.000632	2.455
37	97%	Aldrich	272	0.02	2.31

Table B.1. continued.

ID	Purity	Source	Water solubility [a] mg L^{-1}	Vapor pressure [a] mm Hg	Log *P* [b]
38	99.9%	Supelco	198	0.00012	2.13
39	98%	Aldrich	2,915	0.0000319	3
40	99%	Ehrenstorfer	6,913	0.000296	2.55
41	≥97%	Aldrich	141	0.000328	2.553
42	98%	Aldrich	6,913	0.000296	2.553
43	98%	Aldrich	1,684	0.0000328	2.94
44	99%	Aldrich	20000	0.000501	0.62
45	98%	Aldrich	27,000	0.00186	0.21
46	99.9%	Fluka	16000	0.00004	0.04
47	97%	Aldrich	14,050	0.00506	1.16
48	98%	Aldrich	25,043	0.000869	0.07
49	97%	Aldrich	11,163	0.000257	1.81
50	99.5%	Ehrenstorfer	925	0.0000149	1.26
51	99.9%	Supelco	1,470	0.00277	1.85
52	99.4%	Sigma-	1,200	0.0000956	1.37
53	≥99%	Aldrich	728	0.0000032	1.39
54	99.9%	Fluka	995	0.00000785	1.84
55	97%	Aldrich	995	0.00000854	1.29
56	97%	Aldrich	613	0.000337	2.02
57	99.9%	Fluka	613	0.00235	2.02
58	99.3%	Ehrenstorfer	474	0.00034	2.14
59	99.5%	Ehrenstorfer	500	0.000276	2.72
60	97%	Aldrich	474	0.000383	2.12
61	97%	Aldrich	240	0.00000735	2.48
62	98%	Aldrich	10,102	0.00047	1.435

[a]: Danish (Q)SAR database, [b]: ECOSAR

Table B.2. Structures and spectra of the tested chemicals analyzed spectrophotometrically.

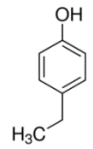
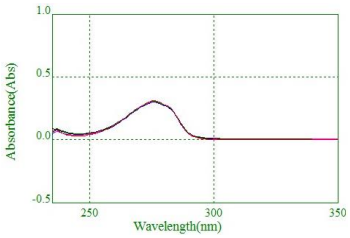
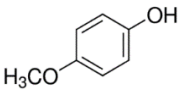
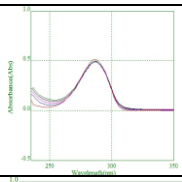
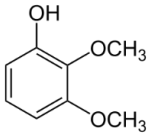
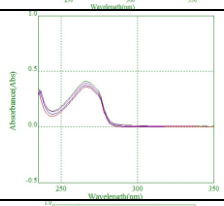
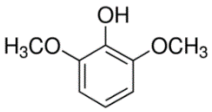
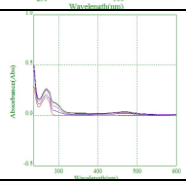
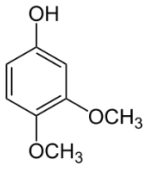
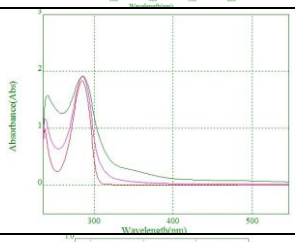
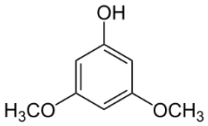
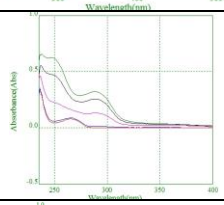
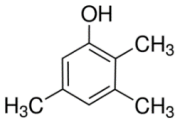
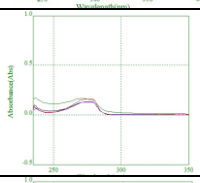
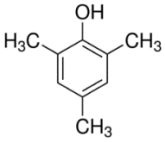
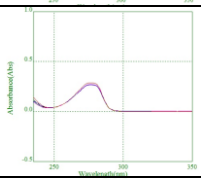
ID	Name	Structure	Spectrum
8	4-ethylphenol		
9	4-methoxyphenol		
10	2,3-dimethoxyphenol		
11	2,6-dimethoxyphenol		
12	3,4-dimethoxyphenol		
13	3,5-dimethoxyphenol		
14	2,3,5-trimethylphenol		
15	2,4,6-trimethylphenol		

Table B.2. continued.

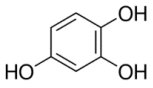
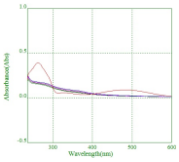
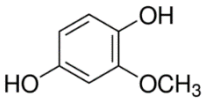
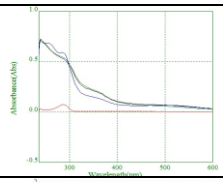
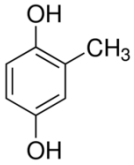
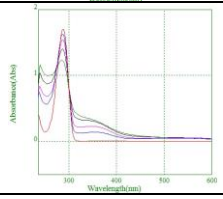
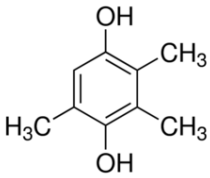
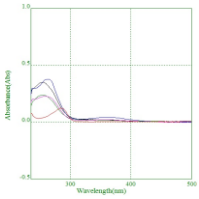
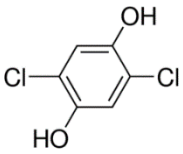
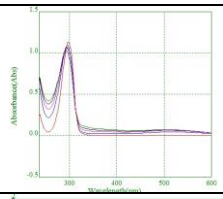
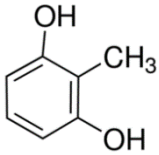
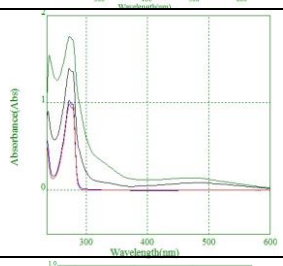
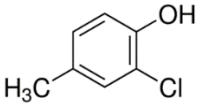
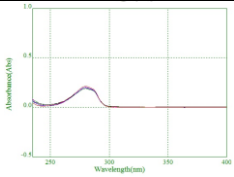
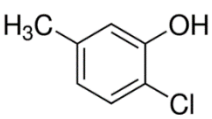
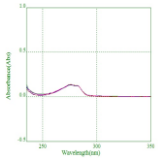
ID	Name	Structure	Spectrum
16	1,2,4-trihydroxybenzene (hydroxyhydroquinone)		
17	methoxyhydroquinone		
18	methylhydroquinone		
19	2,3,5-trimethylhydroquinone		
20	2,5-dichlorohydroquinone		
21	5-methylresorcinol (Orcinol)		
22	2-chloro-4-methylphenol		
23	2-chloro-5-methylphenol		

Table B.2. continued.

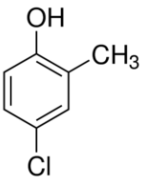
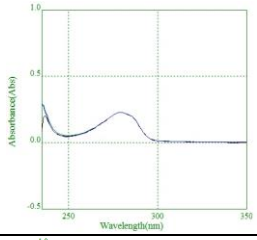
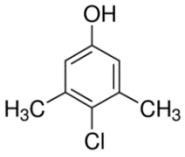
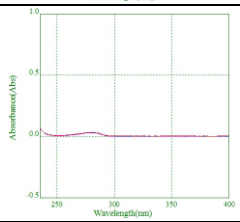
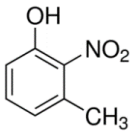
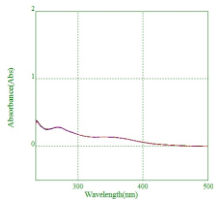
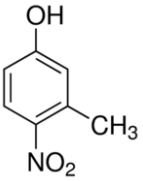
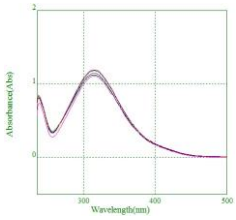
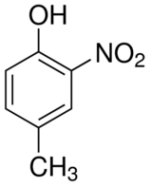
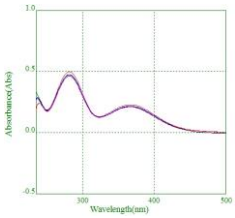
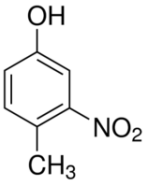
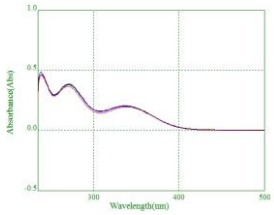
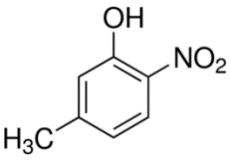
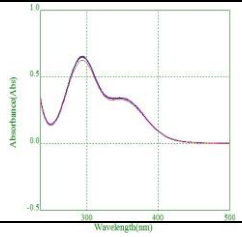
ID	Name	Structure	Spectrum
24	4-chloro-2-methylphenol		
26	4-chloro-3,5-dimethylphenol		
33	3-methyl-2-nitrophenol		
34	3-methyl-4-nitrophenol		
35	4-methyl-2-nitrophenol		
36	4-methyl-3-nitrophenol		
37	5-methyl-2-nitrophenol		

Table B.2. continued.

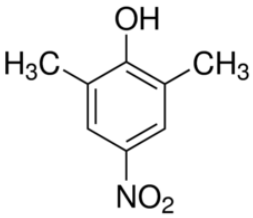
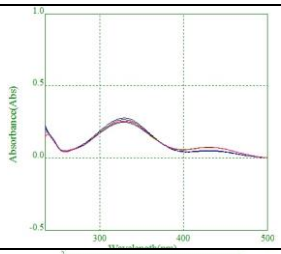
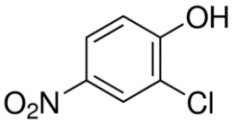
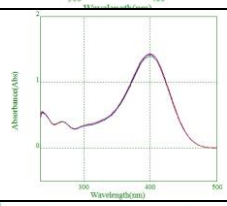
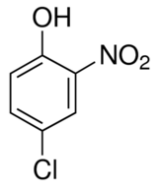
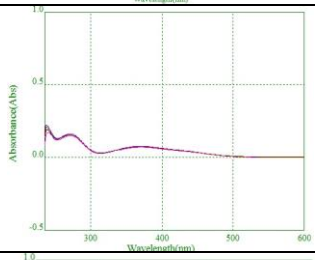
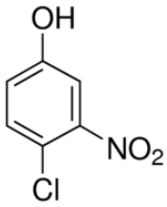
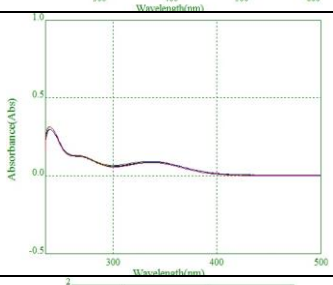
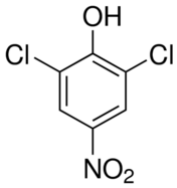
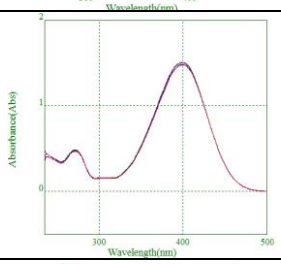
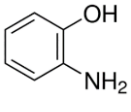
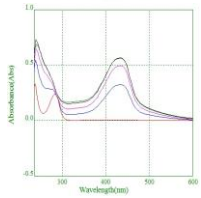
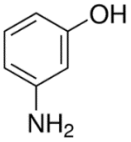
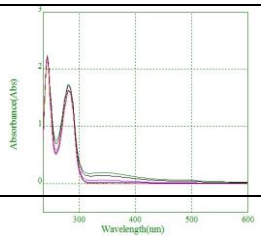
ID	Name	Structure	Spectrum
39	2,6-dimethyl-4-nitrophenol		
40	2-chloro-4-nitrophenol		
41	4-chloro-2-nitrophenol		
42	4-chloro-3-nitrophenol		
43	2,6-dichloro-4-nitrophenol		
44	2-aminophenol		
45	3-aminophenol		

Table B.2. continued.

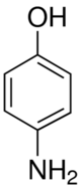
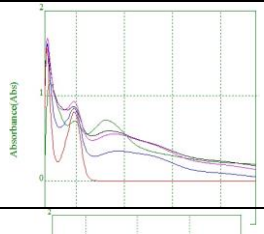
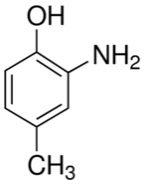
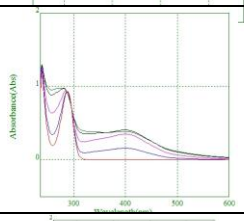
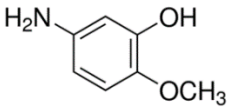
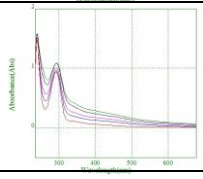
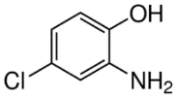
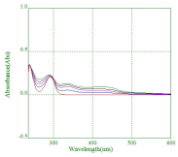
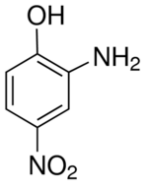
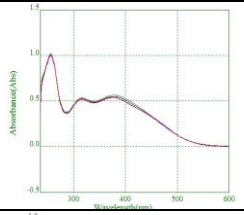
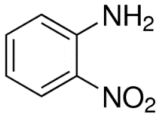
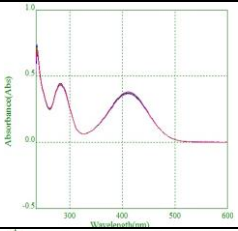
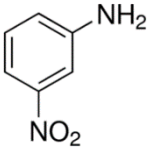
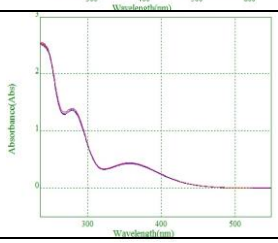
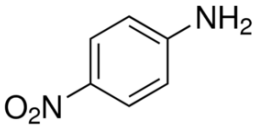
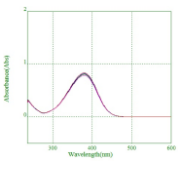
ID	Name	Structure	Spectrum
46	4-aminophenol		
47	2-amino-4-methylphenol		
48	5-amino-2-methoxyphenol		
49	2-amino-4-chlorophenol		
50	2-amino-4-nitrophenol		
51	2-nitroaniline		
52	3-nitroaniline		
53	4-nitroaniline		

Table B.2. continued.

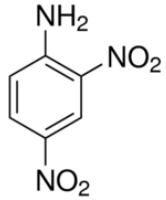
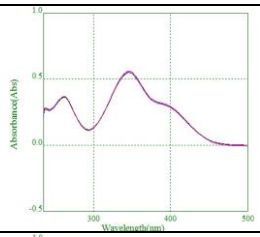
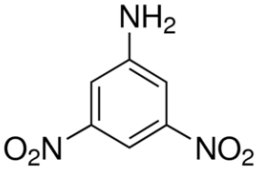
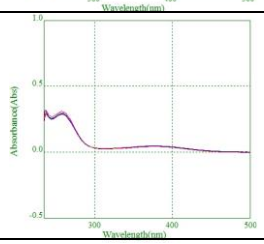
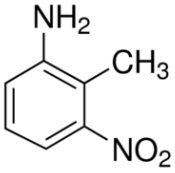
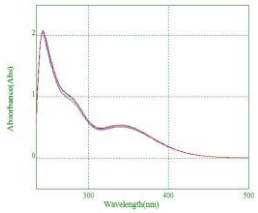
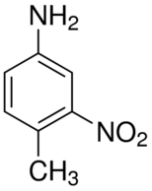
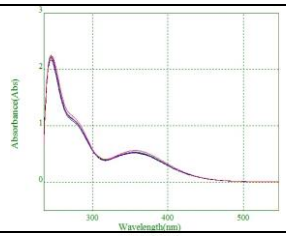
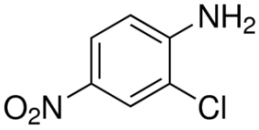
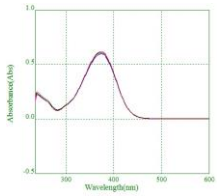
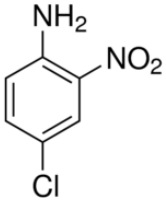
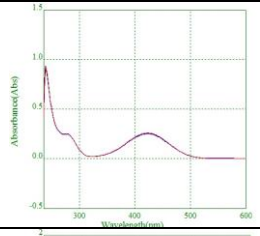
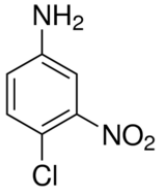
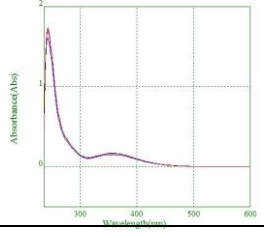
ID	Name	Structure	Spectrum
54	2,4-dinitroaniline		
55	3,5-dinitroaniline		
56	2-methyl-3-nitroaniline		
57	4-methyl-3-nitroaniline		
58	2-chloro-4-nitroaniline		
59	4-chloro-2-nitroaniline		
60	4-chloro-3-nitroaniline		

Table B.2. continued.

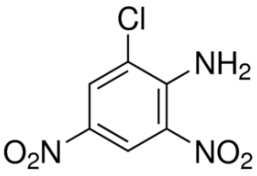
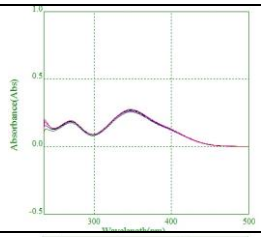
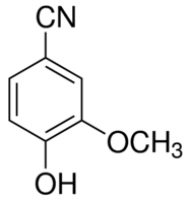
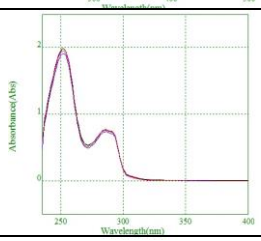
ID	Name	Structure	Spectrum
61	6-chloro-2,4-dinitroaniline		
62	4-hydroxy-3-methoxybenzonitrile		

Table B.3. Structures of the tested chemicals analyzed via GC.

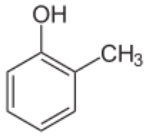
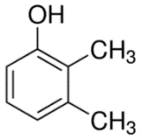
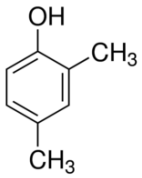
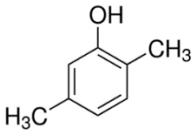
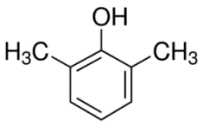
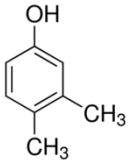
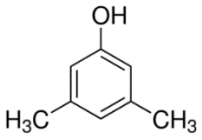
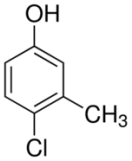
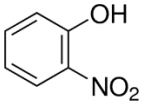
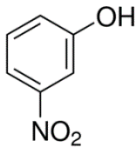
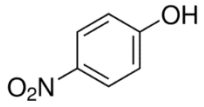
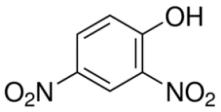
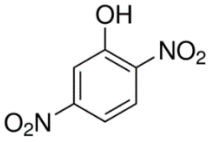
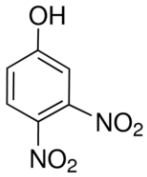
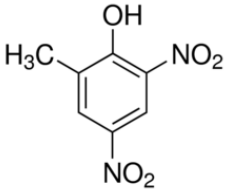
ID	Name	Structure
1	2-methylphenol	
2	2,3-dimethylphenol	
3	2,4-dimethylphenol	
4	2,5-dimethylphenol	
5	2,6-dimethylphenol	
6	3,4-dimethylphenol	
7	3,5-dimethylphenol	
25	4-chloro-3-methylphenol	
27	2-nitrophenol	

Table B.3. continued.

ID	Name	Structure
28	3-nitrophenol	
29	4-nitrophenol	
30	2,4-dinitrophenol	
31	2,5-dinitrophenol	
32	3,4-dinitrophenol	
38	2-methyl-4,6-dinitrophenol	

APPENDIX C: RELATIONSHIP BETWEEN ABSORBANCE AND ALGAL CELL COUNTS

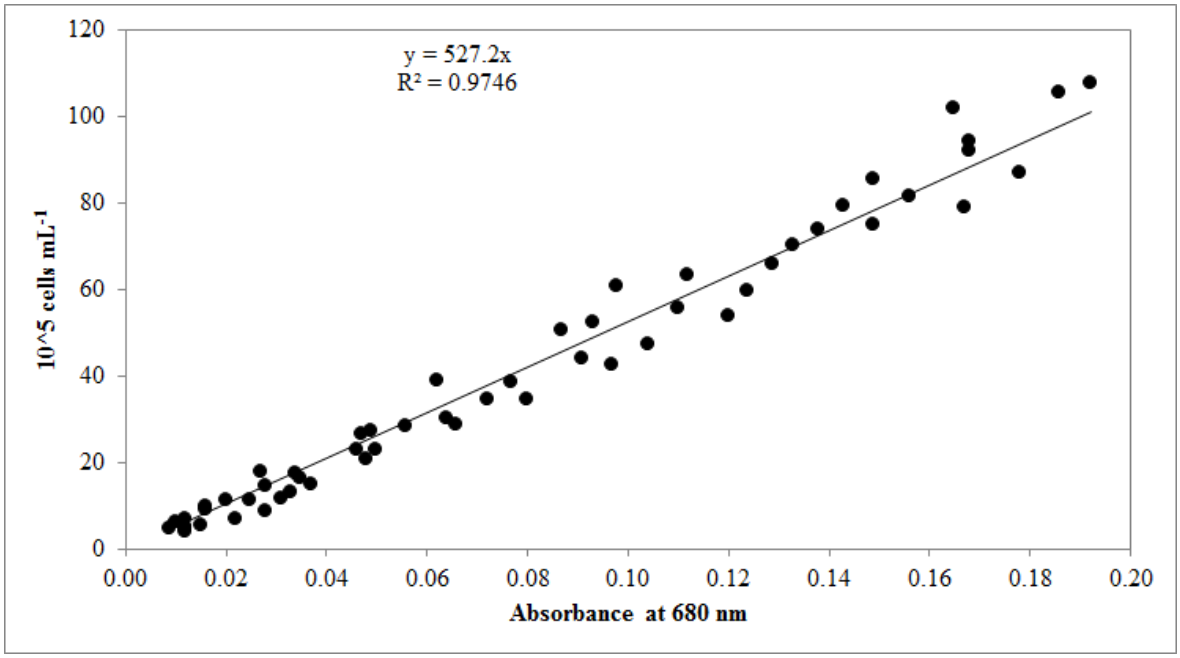


Figure C.1. Plot of absorbance versus algal cell counts.

APPENDIX D: DETAILS OF IC_{50} MODELS

Table D.1. The descriptor values used in IC_{50} models.

ID	ATSC3e	Admet_MlogP	PEoEDIa_3D	S+logD	EEM_XFon
1	0.040	1.859	0	1.770	0.360
2	0.047	2.193	0	2.179	0.371
3	0.046	2.193	0	2.337	0.372
4	0.046	2.193	0	2.324	0.371
5	0.037	2.193	0	2.200	0.377
6	0.056	2.193	0	2.310	0.366
7	0.055	2.193	0	2.360	0.365
8	0.068	2.193	0	2.372	0.369
9	0.122	1.246	0	1.558	0.354
13	0.201	1.012	0	1.584	0.364
14	0.053	2.510	0	2.743	0.380
15	0.045	2.510	0	2.771	0.387
16	0.334	0.836	0	0.058	0.320
17	0.288	1.190	1	0.544	0.344
18	0.102	1.246	0	1.106	0.349
20	0.425	2.116	2	2.331	0.214
21	0.116	1.246	0	1.174	0.343
22	0.152	2.461	0	2.640	0.287
23	0.152	2.461	0	2.696	0.286
24	0.092	2.461	0	2.632	0.285
25	0.096	2.461	0	2.647	0.280
26	0.051	2.778	0	3.026	0.297
27	0.289	1.385	3	0.753	1.635
28	0.291	0.874	1	0.176	1.634
29	0.291	0.874	1	0.193	1.634
30	0.516	0.958	4	-0.044	1.596
31	0.516	0.958	4	-0.035	1.597
32	0.533	0.447	3	-0.710	1.602
33	0.236	1.738	2	1.031	1.713
34	0.232	1.227	1	0.506	1.712
35	0.237	1.738	3	1.160	1.705
36	0.232	1.227	1	0.454	1.712
37	0.237	1.738	3	1.267	1.706
38	0.474	1.311	4	0.179	1.656
39	0.158	1.560	1	0.778	1.759
40	0.410	1.495	2	0.948	1.329
41	0.398	2.006	3	1.644	1.330

Table D.1. continued.

ID	ATSC3e	Admet_MlogP	PEoEDIa_3D	S+logD	EEM_XFon
42	0.402	1.495	2	0.889	1.340
43	0.486	2.096	2	1.585	1.120
44	0.102	1.404	1	0.524	1.282
45	0.079	0.893	0	0.311	1.281
47	0.088	1.757	1	1.038	1.349
49	0.191	2.025	1	1.533	1.024
51	0.181	1.385	2	0.088	1.583
52	0.180	0.874	1	-0.218	1.577
54	0.409	0.958	3	-0.665	1.558
55	0.405	0.447	2	-1.008	1.552
56	0.131	1.227	1	-0.026	1.655
57	0.134	1.227	1	0.068	1.655
58	0.286	1.495	1	0.508	1.305
59	0.288	2.006	2	0.985	1.309
60	0.286	1.495	2	0.526	1.315
61	0.479	1.579	3	0.057	1.332
62	0.210	1.433	1	1.391	0.627
63	0.052	1.506	0	1.391	0.338
64	0.184	2.127	0	2.080	0.264
65	0.148	2.127	0	2.265	0.257
66	0.148	2.127	0	2.277	0.258
67	0.300	2.729	0	2.665	0.213
68	0.288	2.729	0	2.900	0.214
69	0.288	2.729	0	2.870	0.213
70	0.307	2.729	0	2.111	0.218
71	0.281	2.729	0	3.023	0.208
72	0.268	2.729	0	3.031	0.207
73	0.394	3.314	0	3.339	0.178
74	0.390	3.314	0	3.147	0.178
75	0.398	3.314	0	2.314	0.183
76	0.390	3.314	0	3.445	0.179
77	0.394	3.314	0	2.544	0.184
78	0.386	3.314	0	3.651	0.174
79	0.451	3.617	0	3.571	0.153
80	0.451	3.617	0	2.571	0.157
81	0.451	3.617	0	2.338	0.157
82	0.455	3.909	0	2.612	0.138
83	0.368	0.836	2	0.299	0.320
84	0.160	0.893	0	0.733	0.328
85	0.306	1.514	0	1.562	0.262
86	0.528	2.736	0	1.644	0.161
87	0.210	1.404	1	0.652	0.328

Table D.1. continued.

ID	ATSC3e	Admet_MlogP	PEoEDIa_3D	S+logD	EEM_XFon
88	0.317	2.025	1	1.683	0.257
89	0.433	2.627	1	2.302	0.216
90	0.160	0.893	0	0.745	0.327
91	0.306	1.514	0	1.507	0.262
92	0.425	2.116	0	1.762	0.215

Table D.2. SVR1 Support vectors.

Index	Coefficient	Support Vector (normalized)
1	5	0.930328, -0.704792, 1
2	2.28236	-0.0327869, 0.318271, -1
3	-5	-0.971311, -0.388118, -1
4	-0.203268	0.663934, 0.831618, -1
5	-1.4798	0.680328, 1, -1
6	-4.31735	-0.32377, -0.446922, -0.5
7	-5	-0.442623, -0.458251, 0
8	-3.9937	0.930328, -0.704792, 1
9	-5	-0.32377, -0.4304, -0.5
10	-5	0.807377, -0.0470708, 0
11	5	-0.766393, -0.446922, -0.5
12	2.23872	0.778689, -0.345908, 0.5
13	-1.16601	-0.860656, -0.742131, -1
14	5	-0.528689, -0.742131, -1
15	-5	-0.00409836, -0.0993674, 0
16	5	0.590164, 0.259467, -0.5
17	4.35765	0.47541, -1, 0
18	1.53291	-0.213115, -0.254193, 0.5
19	5	-0.00409836, -0.570874, -0.5
20	5	-0.790984, 0.163446, -1
21	5	-0.233607, -0.549401, -0.5
22	-5	0.0696721, -0.383247, -1
23	3.05539	-0.959016, 0.00861998, -1
24	-5	0.184426, -0.774933, -1
25	5	-0.0860656, 0.318271, -1
26	-5	-1, 0.191873, -1
27	5	0.663934, 0.831618, -1
28	-5	-0.709016, -0.538072, -1
29	5	-0.807377, 0.163446, -1
30	-2.58988	0.045082, 0.318271, -1
31	1.02217	-0.97541, 0.346699, -1
32	4.2608	-0.82377, -0.242864, -0.5
33	5	0.0696721, -0.383247, -1
34	-5	-0.430328, -0.0292339, -1
35	-5	-0.0122951, -0.394576, 0
36	-5	-0.561475, 0.163446, -1
37	-5	0.663934, 0.831618, -1
38	5	0.446721, -0.0993674, 0.5
39	-5	1, -1, 0.5
40	5	0.00819672, -0.75346, -0.5

Table D.3. SVR2 Support vectors.

Index	Coefficient	Support Vector (normalized)
1	5	0.0182556, -0.475431, 0.845774
2	2.47754	0.501014, -0.145665, 0.469463
3	5	-0.0750507, 0.764141, -0.914867
4	-2.57776	0.68357, 0.581131, -1
5	-5	-0.221095, -0.361433, 0.942011
6	5	0.667343, 1, -0.981493
7	-5	-0.841785, -0.423892, 0.410241
8	-5	-0.979716, 0.650579, -0.692782
9	-5	0.809331, 0.132562, 0.211598
10	-5	-0.691684, -0.0469535, -0.74707
11	5	0.59432, 0.445731, -0.903763
12	5	-0.772819, 0.596418, -0.8248
13	-3.84916	0.561866, 0.209871, -0.904997
14	-0.916986	-0.667343, 0.120769, -0.733498
15	-5	-0.545639, 0.593361, -0.816163
16	2.37499	0.48073, -1, 0.744602
17	5	-0.200811, -0.0530684, 0.933374
18	-5	-0.204868, -0.109413, 0.943245
19	-3.59193	0.931034, -0.578947, 0.79889
20	-0.774099	-0.432049, -0.654946, 0.775447
21	5	-0.748479, -0.330858, 0.411474
22	-5	-0.415822, 0.348766, -0.84454
23	-5	0.192698, -0.534396, -0.775447
24	5	0.780933, -0.534833, 0.473165
25	5	0.123732, 0.175366, -0.853177
26	-5	0.419878, 0.944966, -0.949414
27	-4.31884	-0.513185, -0.234331, -0.766811
28	-5	0.0182556, -0.482857, 0.845774
29	5	0.931034, -0.575016, 0.800123
30	5	0.330629, -0.429133, -0.775447
31	5	-0.935091, 0.449225, -0.718692
32	2.64747	-0.955375, 0.761957, -0.803825
33	-5	0.0791075, 0.0984931, -0.847008
34	5	-0.521298, -0.219917, 1
35	5	-0.513185, -0.239572, -0.765577
36	-5	1, -0.869841, 0.806292
37	-1.47121	-0.630832, -0.571085, 0.871684
38	5	-0.310345, -0.274951, -0.765577

Table D.4. BPNN1 model function.

Function	// sigmoid(x) = 1/(1+exp(-x))
Input layer	IN_0 = ATSC3e*1.63934 + 0.0262295 IN_1 = Admet_MlogP*0.231083 - 0.00323984 IN_2 = PEoEDIa_3D*0.2 + 0.1
Hidden layer	HL_0 = sigmoid(-1.46588*IN_0 - 1.75817*IN_1 - 0.556879*IN_2 - 0.230559) HL_1 = sigmoid(2.62314*IN_0 + 4.72438*IN_1 + 5.18532*IN_2 - 1.38481) HL_2 = sigmoid(-1.61726*IN_0 - 2.10181*IN_1 + 0.0864572*IN_2 + 0.704068)
Output layer	OUT = (sigmoid(-1.98373*HL_0 + 4.24905*HL_1 - 2.14796*HL_2 - 2.92663) - 0.280132) / 0.264901

Table D.5. BPNN2 model function.

Function	// sigmoid(x) = 1/(1+exp(-x))
Input layer	IN_0 = ATSC3e*1.62272 + 0.0350913 IN_1 = S+logD"*0.174711 + 0.276108 IN_2 = EEM_XFon*0.493523 + 0.0318939
Hidden layer	HL_0 = sigmoid(-3.19593*IN_0 + 4.46547*IN_1 - 1.85729*IN_2 - 2.7643) HL_1 = sigmoid(-4.156*IN_0 - 1.49108*IN_1 - 2.86807*IN_2 + 2.89974) HL_2 = sigmoid(1.67337*IN_0 - 0.293578*IN_1 - 1.78175*IN_2 - 1.35071)
Output layer	OUT = (sigmoid(3.82807*HL_0 - 3.9077*HL_1 + 1.74793*HL_2 + 0.275415) - 0.280132) / 0.264901

Table D.6. External set for model validation CAS no, names, descriptor and *p*T values for available species. ID numbers are as appeared in the original paper*.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
26	000058-89-9	Lindane (γ-HCH)	0.911	4.092	3.830	0	2.07
27	000075-35-4	1,1-Dichloroethylene	0.100	1.672	2.105	1	-0.63
32	000079-01-6	Trichloroethylene	0.092	2.081	2.595	1	-0.53
39	000067-56-1	Methanol	0.016	-0.814	-0.701	0	-2.85
42	000067-63-0	2-propanol	0.114	0.347	0.091	0	-2.29
133	020679-58-7	Acetic acid, bromo-, 2-butene-1,4-diyl ester	0.592	1.124	2.040	2	3.99
141	000080-62-6	methyl methacrylate	0.032	0.482	0.929	1	-0.23
144	000097-88-1	n-butyl methacrylate	0.092	1.547	2.628	1	0.04
162	000335-67-1	perfluorooctanoic acid	2.428	3.764	1.428	21	0.30
173	000075-64-9	t-butylamine	0.069	0.800	-1.435	0	0.66
179	000108-91-8	Cyclohexylamine	0.075	1.195	-1.308	0	0.70
183	000124-40-3	Dimethylamine	0.012	-0.172	-2.304	0	0.86
184	000109-89-7	Diethylamine	0.052	0.800	-1.943	0	0.56
185	000108-18-9	Diisopropylamine	0.088	1.587	-1.478	0	0.70
186	000111-92-2	Dibutylamine	0.075	2.274	-0.464	0	0.83
226	000068-12-2	Dimethylformamide	0.010	-0.273	-0.630	1	-1.37
230	000062-75-9	Dimethylnitrosamine	0.098	-0.410	-0.413	1	1.27
231	000055-18-5	Diethylnitrosamine	0.148	0.562	0.422	1	0.96
236	000110-91-8	Morpholine	0.202	-0.473	-1.941	1	0.49
244	099129-21-2	Clethodim	0.272	2.956	3.245	1	1.20
262	000062-56-6	Thiourea	0.062	-1.434	-1.113	3	1.05
265	002212-67-1	Molinate	0.135	1.579	2.597	3	0.80
278	077182-82-2	Glufosinate	0.638	-3.758	-3.482	4	1.23
280	001071-83-6	Glyphosate	0.781	-4.581	-3.662	4	1.48

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
284	000126-72-7	Tris-(2,3-dibromopropyl)-phoshate	0.841	4.755	3.894	6	2.35
293	000115-29-7	Endosulfan	1.281	3.115	4.080	4	2.98
309	000108-90-7	Chlorobenzene	0.045	2.876	2.757	0	0.95
314	000095-50-1	1,2-Dichlorobenzene	0.161	3.478	3.443	0	1.82
316	000106-46-7	1,4-Dichlorobenzene	0.136	3.478	3.477	0	1.96
321	000087-61-6	1,2,3-Trichlorobenzene	0.271	4.063	4.003	0	2.30
323	000120-82-1	1,2,4-Trichlorobenzene	0.261	4.063	4.149	0	2.11
362	000131-11-3	Dimethyl phthalate	0.166	1.193	1.643	2	0.66
363	000084-66-2	Diethyl phthalate	0.295	1.765	2.474	2	0.39
366	000131-17-9	Diallyl phthalate	0.191	2.138	2.718	2	1.74
398	000094-74-6	2-methyl-4-chlorophenoxyacetic acid	0.486	1.680	-0.426	1	1.34
400	000882-09-7	Clofibrinic acid	0.565	1.972	-0.165	1	0.36
408	000108-39-4	3-cresol	0.052	1.859	1.886	0	-0.13
420	000644-35-9	2-n-propylphenol	0.071	2.510	2.798	0	0.75
427	000098-54-4	p-tert-butylphenol	0.077	2.813	3.275	0	0.95
437	001806-26-4	4-n-octylphenol	0.119	3.921	5.543	0	1.48
439	000104-40-5	4-n-Nonylphenol	0.126	4.177	6.075	0	1.56
474	002460-49-3	4,5-Dichloroguaiacol	0.387	2.944	2.924	1	1.80
475	002668-24-8	4,5,6-trichloroguaiacol	0.489	3.247	3.342	1	2.70
476	057057-83-7	3,4,5-trichloroguaiacol	0.489	3.247	3.298	1	2.48
477	002539-17-5	Tetrachloroguaiacol	0.548	3.539	3.499	1	2.82
478	002539-26-6	3,4,5-trichloro-2,6-dimethoxyphenol	0.598	2.703	3.102	2	2.48
488	001689-84-5	Bromoxynil	0.206	2.646	0.551	0	1.64
496	003428-24-8	4,5-dichlorocatechol	0.428	2.627	2.429	1	2.60
497	003978-67-4	3,4-Dichlorocatechol	0.438	2.627	2.280	1	2.85

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
500	056961-20-7	3,4,5-trichlorocatechol	0.507	2.944	2.733	1	2.92
501	032139-72-3	3,4,6-trichlorocatechol	0.509	2.944	2.396	1	3.04
502	001198-55-6	Tetrachlorocatechol	0.532	3.247	2.440	1	3.49
508	000062-53-3	Aniline	0.038	1.506	0.913	0	0.69
527	000106-47-8	4-Chloroaniline	0.070	2.127	1.870	0	1.73
529	000095-76-1	3,4-Dichloroaniline	0.189	2.729	2.653	0	1.41
549	000099-55-8	2-Amino-4-nitrotoluene	0.134	1.701	1.835	1	0.99
550	000119-32-4	4-Amino-2-nitrotoluene	0.134	1.701	1.717	1	1.01
551	000603-83-8	2-Amino-6-nitrotoluene	0.131	1.701	1.598	1	0.84
555	019406-51-0	4-Amino-2,6-dinitrotoluene	0.370	1.748	1.631	2	1.23
556	035572-78-2	2-Amino-4,6-dinitrotoluene	0.370	1.748	1.739	2	1.89
565	000095-80-7	2,4-Diaminotoluene	0.051	1.246	0.335	0	1.11
566	000823-40-5	2,6-Diaminotoluene	0.047	1.246	0.259	0	0.33
568	006629-29-4	2,4-Diamino-6-nitrotoluene	0.161	1.181	0.917	1	0.53
569	059229-75-3	2,6-Diamino-4-nitrotoluene	0.161	1.181	0.985	1	0.68
587	000100-00-5	4-Chloronitrobenzene	0.270	2.509	2.544	1	1.42
603	000056-75-7	Chloramphenicol	0.861	1.228	1.027	4	1.36
607	000121-14-2	2,4-dinitrotoluene	0.355	2.239	1.858	2	1.84
608	000606-20-2	2,6-Dinitrotoluene	0.357	2.239	1.707	2	1.04
621	040487-42-1	Pendimethalin	0.305	4.057	4.310	3	3.20
622	001582-09-8	Trifluralin	0.863	4.207	4.670	2	2.20
625	000118-96-7	2,4,6-Trinitrotoluene	0.578	2.359	1.662	3	2.60
629	001194-65-6	2,6-Dichlorobenzonitrile	0.158	2.956	2.752	0	1.80
631	001897-45-6	Tetrachloroisophthalonitrile	0.346	3.133	3.422	0	1.52
635	051218-45-2	Metolachlor	0.358	3.026	3.168	1	1.71

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
636	034256-82-1	Acetochlor	0.294	3.182	2.842	2	2.23
638	023184-66-9	Butachlor/N-(Butoxymethyl)-2-Chloro-2',6'-diethylacetanilide	0.331	3.913	4.430	2	3.17
639	051218-49-6	Pretilachlor/2-Chloro-2',6'-diethyl-N-(2-propoxyethyl)acetanilide <Pretilachlor>	0.428	3.507	4.276	1	3.42
641	057837-19-1	Metalaxyl/Methyl-(2-methoxyacetyl)-N-(2,6-xylyl)-DL-alaninate	0.431	1.501	1.880	2	1.57
648	034123-59-6	Isoproturon/1,1-dimethyl-3-(8-isopropylphenyl)-urea	0.102	2.797	2.543	3	4.18
649	015545-48-9	Chlorotoluron	0.096	2.784	2.477	3	5.40
651	000330-54-1	Diuron/1-(3,4 dichlorophenyl)-3,3 dimethyl urea	0.271	3.052	2.815	3	5.52
660	023564-05-8	Dimethyl 4,4'-(o-phenylente) bis(3-thioa(lophanate)	0.419	1.483	1.483	15	0.39
663	000057-67-0	Sulfaguanidine	0.269	0.325	-1.389	7	0.69
669	073231-34-2	Florfenicol	0.990	0.885	0.859	3	0.22
670	015318-45-3	Thiamphenicol	0.930	-0.016	-0.010	3	-0.56
672	000122-14-5	Fenitrothion	0.358	1.919	3.377	4	1.37
678	064249-01-0	Anilofos	0.338	2.300	4.183	4	1.58
682	022224-92-6	Fenamiphos	0.276	3.479	2.861	6	0.90
696	069377-81-7	Fluroxypyr	0.743	0.485	-0.890	3	1.64
700	000122-34-9	Simazine	0.229	1.867	2.357	9	2.43
701	001912-24-9	Atrazine	0.226	2.184	2.814	9	3.25
709	021725-46-2	Cyanazine	0.265	1.625	2.127	9	3.61
713	000834-12-8	Ametryn	0.172	1.916	3.046	9	4.26

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
714	007287-19-6	Prometryn	0.198	2.219	3.411	9	4.32
719	028159-98-0	Irgarol 1051/2-methylthio-4-tert-butylamino-6-cyclopropylamino-s-triazine	0.195	2.511	3.697	9	5.63
725	000059-87-0	Nitrofurazone	0.587	0.745	0.600	6	2.14
730	000061-82-5	3-Amino-1,2,4-triazole	0.110	-0.637	-1.079	2	0.32
734	034014-18-1	Tebuthiuron	0.084	1.242	1.844	7	3.27
749	000119-12-0	Pyridaphenthion	0.539	2.407	3.035	5	1.37
752	032809-16-8	Procymidone/N-(3,5-Dichlorophenyl)-1,2-dimethylcyclopropane-1,2-dicarboximide	0.287	3.240	3.168	2	2.61
754	024096-53-5	N-(3,5-Dichlorophenyl)succinidide	0.497	2.449	1.527	2	1.56
755	036734-19-7	3-(3,5-Dichlorophenyl)-N-isopropyl-2,4-dioxoimidazolidine-l-carboxamide	0.636	2.820	2.657	8	0.90
756	039807-15-3	Oxadiargyl	0.449	3.867	3.705	4	0.90
762	000096-09-3	Styrene-7,8-oxide	0.027	1.533	1.682	0	0.57
766	000080-05-7	Bisphenol A	0.107	3.306	3.644	0	0.93
772	051338-27-3	Diclofop-methyl/2-[4-(2,4-dichlorophenoxy)]-phenoxy propionate methyl ester	0.783	3.254	4.897	2	1.27
773	040843-25-2	Diclofop-P	0.844	3.017	1.393	2	1.81
775	040843-73-0	4-(2,4-dichlorophenoxy)-phenol	0.420	3.586	4.738	1	3.01
779	042874-03-3	Oxyfluorfen	1.074	4.121	4.841	2	1.86
781	068359-37-5	Beta-cyfluthrin	0.641	3.873	6.351	3	2.09
783	054910-89-3	Fluoxetine	0.372	4.153	1.701	0	3.89
787	022071-15-4	Ketoprofen	0.121	2.968	0.242	1	2.10
791	000085-68-7	Butylbenzyl phthalate	0.142	3.279	4.690	2	2.89
797	071626-11-4	R-(－)-benalaxyl/Rac-benalaxyl/S-(+)-benalaxyl	0.273	3.239	3.531	2	1.93

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
811	126833-17-8	Fenhexamid	0.386	3.438	4.444	1	2.76
813	072619-32-0	Haloxyfop-R	1.130	2.598	4.200	2	2.61
814	083066-88-0	Fluazifop-p	1.074	2.122	0.443	3	2.49
816	000738-70-5	Trimethoprim	0.385	0.855	0.820	5	0.54
817	083055-99-6	Bensulfuron-methyl	0.775	0.954	0.406	13	1.48
826	090982-32-4	Chlorimuron-ethyl	0.827	1.272	1.143	13	1.88
827	111991-09-4	Nicosulfuron	0.724	-0.012	-1.083	13	2.46
828	136849-15-5	Cyclosulfamuron	0.760	2.520	0.570	16	3.02
829	064902-72-3	Chlorsulfuron	0.612	2.119	0.155	15	3.27
831	074223-64-6	Metsulfuron-methyl	0.644	1.261	-0.618	15	1.19
833	106040-48-6	Tribenuron	0.643	1.261	-2.478	15	1.02
834	111353-84-5	Ethametsulfuron	0.821	1.356	-2.233	17	1.13
842	000723-46-6	sulfamethoxazole	0.428	0.565	-0.003	4	3.24
850	079319-85-0	N,N'-Methylene-di(2-amino-5-mercapto-1,3,4-thiodiazole)	0.206	-0.576	1.721	7	1.84
853	093697-74-6	Pyrazosulfuron-ethyl	0.902	-0.093	-0.302	13	1.57
861	000525-66-6	Propranolol	0.405	2.534	0.703	0	1.60
878	084087-01-4	Quinclorac	0.357	2.471	-0.051	1	1.42
885	052316-55-9	Carbendazim	0.229	0.983	1.366	4	1.18
887	017804-35-2	Methyl-l-(butylcarbamoyl)-2-benzimidazole carbamate	0.330	2.780	2.490	6	3.16
891	018691-97-9	Methabenzthiazuron	0.109	1.262	2.125	6	4.02
892	025059-80-7	Benazolin-ethyl	0.501	1.211	2.561	2	1.79
915	000260-94-6	Acridine	0.058	2.581	3.200	0	2.30
917	000298-46-4	Carbamazepine	0.102	3.139	2.404	3	0.53
938	079617-96-2	Sertraline	0.191	4.757	3.054	0	4.40

Table D.6. continued.

ID	CAS	Name	ATSC3e	MlogP	S+logD	PEoEDIa_3D	*p*T
939	000059-40-5	Sulfaquinoxaline	0.320	0.633	0.655	4	3.09
940	094051-08-8	Quizalofop-p	0.814	1.556	0.419	2	3.15
943	093106-60-6	Enrofloxacin	0.523	1.053	0.257	1	0.78
947	073250-68-7	Mefenacet	0.305	2.099	2.970	5	1.80
948	095617-09-7	Fenoxaprop	0.835	2.130	1.100	4	2.61
955	098967-40-9	Flumetsulam	0.807	1.909	-0.561	8	2.13
964	000139-91-3	Furaltadone	1.117	0.916	0.696	6	1.44
968	087818-31-3	Cinmethylin	0.421	3.367	4.988	0	5.79
969	125401-75-4	Bispyribac	0.734	1.461	-1.762	11	1.87
973	000564-25-0	deoxytetracycline	0.905	-0.875	-0.351	6	2.19
976	000057-62-5	chlorotetracycline	1.075	-0.392	-0.245	7	1.71
979	000060-54-8	Tetracycline	0.937	-0.875	-0.520	5	2.13
986	082419-36-1	Ofloxacin	0.685	0.064	-0.317	3	1.88
987	100986-85-4	Levofloxacin	0.685	0.064	-0.317	4	2.48

* Fu et al., 2015

APPENDIX E: DETAILS OF LOW-TOXIC-EFFECT-CONCENTRATION MODELS

Table E.1. Descriptors appear in chronic toxicity models.

ID	SM04_EA(bo)	E1m	T_Grav3	Mor09m
1	5.818	0.232	11.053	-0.561
2	6.034	0.330	11.462	-0.570
3	6.012	0.278	11.459	-0.498
4	6.012	0.214	11.459	-0.514
5	6.034	0.178	11.462	-0.533
8	5.833	0.404	11.568	-0.392
9	5.833	0.347	11.700	-0.673
13	6.057	0.201	12.616	-0.905
14	6.193	0.190	11.841	-0.518
15	6.193	0.259	11.841	-0.447
16	6.012	0.323	11.578	-0.763
17	6.046	0.326	12.120	-0.663
20	6.193	1.447	12.955	-0.368
21	5.990	0.305	11.516	-0.558
22	6.012	0.867	12.051	-0.348
23	6.012	0.950	12.051	-0.323
24	6.012	1.090	12.051	-0.263
25	6.012	0.755	12.051	-0.223
26	6.193	0.622	12.397	-0.295
27	6.239	0.566	12.197	-0.619
28	6.221	0.604	12.195	-0.563
29	6.221	0.671	12.195	-0.644
30	6.634	0.769	13.450	-0.735
31	6.634	0.842	13.450	-0.509
33	6.386	0.483	12.535	-0.599
34	6.370	0.611	12.533	-0.605
36	6.370	0.473	12.533	-0.557
37	6.370	0.574	12.533	-0.519
39	6.500	0.755	12.855	-0.641
40	6.370	0.835	13.034	-0.589
41	6.370	0.993	13.034	-0.397
42	6.370	0.669	13.034	-0.282
43	6.500	0.529	13.776	-0.687
45	5.791	0.275	11.083	-0.540
47	6.012	0.311	11.489	-0.468

Table E.1. continued.

ID	SM04_EA(bo)	E1m	T_Grav3	Mor09m
49	6.012	1.061	12.078	-0.211
51	6.239	0.555	12.170	-0.617
52	6.221	0.640	12.168	-0.536
55	6.622	0.747	13.426	-0.614
56	6.386	0.596	12.510	-0.741
59	6.370	0.891	13.010	-0.377
60	6.370	0.709	13.010	-0.290
61	6.735	0.457	14.131	-0.755
62	6.166	0.559	12.361	-0.540
63	5.542	0.311	10.612	-0.532
64	5.818	0.974	11.685	-0.373
66	5.791	1.035	11.683	-0.169
68	6.012	1.059	12.589	-0.234
71	6.012	1.025	12.589	0.017
72	5.990	1.072	12.587	-0.037
75	6.211	1.264	13.383	-0.352
76	6.193	1.404	13.381	-0.256
80	6.362	0.946	14.090	-0.626
81	6.362	1.332	14.090	-0.615
82	6.506	1.009	14.735	-1.066
84	5.791	0.418	11.115	-0.620
88	6.012	0.968	12.105	-0.298
90	5.791	0.381	11.115	-0.611
91	6.012	0.800	12.105	-0.362
92	6.193	0.893	12.955	-0.389
External set				
6	6.012	0.349	11.459	-0.487
7	5.990	0.351	11.457	-0.417
32	6.634	0.511	13.450	-1.031
38	6.735	0.763	13.730	-0.814
54	6.634	0.668	13.428	-0.709
57	6.370	0.490	12.508	-0.512
70	6.034	1.094	12.591	-0.356

[a] *p*NOEC: log(1/NOEC), [b] pIC_{20}: log(1/ IC_{20}), [a] pIC_{50}: log(1/ IC_{50})

Table E.2. Hat (leverage) values, and standardized residuals belong to each chronic toxicity model.

	Equation 4.4 (model 1)		Equation 4.5 (model 2)		Equation 4.6 (model 3)		Equation 4.7 (model 4)	
ID	**Hat Val.**	**Std.Res**	**Hat Val.**	**Std.Res.**	**Hat Val.**	**Std.Res.**	**Hat Val.**	**Std.Res.**
1	0.094	-0.419	0.057	-0.045	0.086	0.072	0.054	1.411
2	0.051	-0.144	0.028	0.922	0.057	-0.792	0.027	0.349
3	0.060	-0.368	0.026	1.197	0.049	-0.440	0.025	1.238
4	0.072	-0.008	0.030	1.787	0.051	-0.352	0.029	0.725
5	0.077	-0.294	0.040	0.312	0.052	-0.002	0.039	0.423
8	0.069	-0.785	0.035	-0.349	0.042	0.574	0.033	0.948
9	0.075	0.586	0.077	0.135	0.062	0.786	0.073	-0.184
13	0.072	0.591	0.066	0.188	0.108	1.089	0.063	-0.116
14	0.075	-1.380	0.024	-1.094	0.033	-0.372	0.024	0.139
15	0.062	0.386	0.029	1.303	0.030	0.528	0.028	1.154
16	0.053	-0.703	0.039	-1.211	0.095	-1.092	0.038	-0.900
17	0.051	-2.197	0.031	0.767	0.043	-2.092	0.031	0.792
20	0.144	1.025	0.057	1.485	0.040	-0.792	0.056	0.477
21	0.058	2.428	0.125	1.320	0.052	2.157	0.119	0.814
22	0.039	-0.250	0.024	-1.863	0.031	0.800	0.024	1.015
23	0.048	-0.445	0.022	-0.850	0.035	-0.537	0.022	-0.894
24	0.070	0.925	0.021	0.895	0.045	0.129	0.021	0.509
25	0.032	-0.827	0.029	0.954	0.054	-0.505	0.030	0.586
26	0.022	-1.216	0.029	-0.272	0.039	-0.432	0.029	-0.376
27	0.026	-0.873	0.027	0.354	0.033	-2.478	0.028	-2.857
28	0.024	0.116	0.021	0.190	0.027	-0.531	0.021	-0.047
29	0.022	-0.614	0.032	0.909	0.037	-1.962	0.032	0.793
30	0.090	0.629	0.025	-0.655	0.071	0.268	0.025	1.558
31	0.091	-0.155	0.076	1.164	0.052	-0.976	0.075	0.898
33	0.049	0.414	0.021	0.270	0.027	-0.392	0.021	-1.544
34	0.036	0.287	0.025	1.208	0.028	-0.729	0.025	1.027
36	0.048	1.050	0.023	1.312	0.024	0.645	0.022	1.190
37	0.039	-0.803	0.028	-0.235	0.022	-0.734	0.029	0.880
39	0.056	0.033	0.031	-0.108	0.035	-1.001	0.031	0.342
40	0.036	-0.877	0.046	-0.399	0.034	-0.891	0.046	1.520
41	0.048	-1.662	0.080	-1.085	0.039	-1.888	0.079	-1.605
42	0.034	-0.669	0.039	0.567	0.060	0.357	0.039	0.874
43	0.067	0.617	0.023	1.213	0.082	1.148	0.023	1.400
45	0.093	0.829	0.125	-0.672	0.080	1.033	0.119	-1.492
47	0.055	-1.460	0.026	-1.272	0.046	-1.135	0.025	-2.270
49	0.065	0.630	0.025	1.356	0.057	-0.696	0.025	-1.990
51	0.027	0.837	0.031	-0.129	0.034	0.232	0.030	-0.053
52	0.022	1.531	0.040	0.151	0.025	0.940	0.038	0.074

Table E.2. continued.

	Equation 4.4 (model 1)		Equation 4.5 (model 2)		Equation 4.6 (model 3)		Equation 4.7 (model 4)	
ID	**Hat Val.**	**Std.Res**	**Hat Val.**	**Std.Res.**	**Hat Val.**	**Std.Res.**	**Hat Val.**	**Std.Res.**
55	0.087	0.312	0.049	1.240	0.053	-0.693	0.048	0.698
56	0.040	2.744	0.048	1.297	0.054	1.312	0.045	1.327
59	0.039	0.456	0.021	-1.005	0.041	0.821	0.022	-0.803
60	0.034	0.008	0.021	-1.652	0.057	1.720	0.021	-0.261
61	0.149	-1.248	0.064	0.136	0.121	-0.728	0.063	0.253
62	0.024	0.889	0.050	-1.122	0.023	1.328	0.048	-1.025
63	0.165	0.887	0.114	1.213	0.125	0.794	0.108	1.260
64	0.090	1.102	0.038	-0.267	0.039	0.900	0.037	0.619
66	0.107	-0.002	0.025	-1.274	0.074	0.809	0.025	-0.016
68	0.064	-0.507	0.033	0.120	0.056	-0.144	0.033	-0.393
71	0.059	-1.802	0.046	-1.199	0.155	0.442	0.046	0.376
72	0.070	-1.533	0.062	0.220	0.128	-0.607	0.062	-0.839
75	0.091	-0.244	0.049	-0.511	0.065	-0.040	0.049	-0.365
76	0.130	-0.647	0.063	-1.252	0.088	-0.108	0.062	-0.433
80	0.043	0.077	0.045	0.856	0.102	0.472	0.044	0.818
81	0.111	1.027	0.044	0.280	0.102	0.338	0.043	-0.038
82	0.071	-0.439	0.045	-1.474	0.279	-0.594	0.045	-0.346
84	0.077	-0.953	0.048	-1.618	0.091	-0.655	0.046	0.090
88	0.050	-0.957	0.028	-0.801	0.038	-0.923	0.028	-0.703
90	0.080	0.693	0.100	-0.621	0.090	0.467	0.095	-0.871
91	0.034	1.316	0.033	0.576	0.029	1.253	0.032	0.253
92	0.030	0.426	0.024	0.841	0.037	0.703	0.025	0.914
External set								
6	0.050		0.023		0.049		0.022	
7	0.051		0.024		0.047		0.024	
32	0.106		0.022		0.174		0.022	
38	0.124		0.041		0.104		0.041	
54	0.093		0.033		0.065		0.033	
57	0.046		0.038		0.021		0.036	
70	0.068		0.022		0.031		0.022	